Statistical methods for causal inference from sequentially collected data and sequential decision making

by

Aurélien Florent Bibaut

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor in Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark J. van der Laan, Chair
Professor Maya Petersen
Professor Alan Hubbard
Professor Peter L. Bartlett

Spring 2021

Statistical methods for causal inference from sequentially collected data and sequential decision making

Abstract

Statistical methods for causal inference from sequentially collected data and sequential decision making

by

Aurélien Florent Bibaut

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Mark J. van der Laan, Chair

In my dissertation, I consider the type of statistical experiment commonly referred to as adaptive trials, in which the experimenter interacts with an individual or a set of individuals, and, sequentially, over time steps $t = 1, \ldots, T$, observes a vector of measurements $L_1(t)$ on the individual or individuals, then assigns treatment vector $A(t)$, and then observes a post-treatment vector of measurements $L_2(t)$. In an adaptive trial, the experimenter can update the treatment distribution at t based on previous observations.

This very general formulation covers many common settings such as dynamic treatment regimes, the stochastic contextual bandit model, the Markov Decision Process model in reinforcement learning. I consider two related types of learning tasks: causal inference from data collected under an adaptive trial, and sequential decision making with the objective of either maximizing the sample efficiency for an estimation task, or of minimizing some form of cumulative regret.

My primary concerns are to develop statistical methods and algorithms that use statistical models that assume no more than is known from domain knowledge (and therefore are nonparametric), and that are as sample efficient as possible.

Acknowledgements

This work is the result of collaborations with various great researchers.

Firstly, I want to thank my advisor Mark van der Laan for, among other things, being an example of someone who approaches mathematical and statistical problems frontally without being bound by existing literature, for setting an example with his inspiring drive to provide a comprehensive methodology for statistical inference under the form of the Targeted Learning framework, for having taught me with clarity what satisfactory statistical methodology should be, and for his great guidance throughout my doctoral studies.

I want to thank Antoine Chambaz for pushing me to understand in depth the proof techniques of empirical process and martingale process results, for the many hours spent reading bandit literature together, for inviting me to his lab various times, for our fruitful collaboration, and for our many pleasant chats on the rooftop of Paris Descartes University.

I'm grateful to Nikos Vlassis for introducing me to the Off-Policy Evaluation literature in reinforcement learning, for our fruitful collaboration, for investing a lot of time in me when I was interning at Netflix, and for many stimulating discussions.

I am fortunate to have worked with and learned from other great researchers. By alphabetical orders, they are: Guillaume Basse, Maria Dimakopoulou, Avi Feller, Cheng Ju, Alex Luedtke, Ivana Malenica, Maya Petersen, Aldo Pacchiano, and Sam Pimentel. I'm thankful to Iosif Pinelis, even though we have never met, as he is the person who answers most of my questions on MathOverflow, and as he always does so in an illuminating fashion.

Finally, I want to thank my family, and Lisa, for their support.

Contents

Chapter 1

Introduction

Throughout the course of my doctoral studies, the main focus of my research has been to develop methods and theory for causal inference from sequentially collected data, and for the related task of sequential decision making, under the constraint that the statistical models these rely on should only encode what we do know about nature. Causal inference from sequentially collected data and sequential causal inference are two narrowly connected tasks. Indeed, a widely used principle in sequential decision making algorithms is to estimate from the available data, the expected return (or upper or lower bounds on the expected return) of decision rules before carrying them out. Some of the learning problems I consider in my dissertation are off-policy evaluation in reinforcement learning, sequential decision making under the contextual bandit framework, nonparametric estimation from sequentially collected data, and inference and sequential testing in adaptive trials under network and temporal dependence.

Intended applications of the methods I developed include clinical trials, personalized medicine, public health decision making in the context of infectious diseases, ads placement and recommender systems.

From a technical point of view, the constraint that the statistical models should not make more assumptions than is known from domain knowledge implies in practice the statistical models I use are nonparametric. As a result, my work is heavily grounded in nonparametric statistics, and related techniques such as empirical process theory. Due to the sequential nature of the data collection in the problems I consider, other technical foundations of my work include martingale theory, and the theory of weakly dependent sequences of random variables, and in particular results on empirical processes induced by martignales or weakly dependent random variables. Whenever possible, I tried to provide efficient substitution estimators of the statistical parameters of interest, and therefore relied on semiparametric theory, efficiency theory and Targeted Learning theory [Van Der Laan and Rubin, 2006, van der Laan and Rose, 2011, 2018].

I organize this introduction chapter as follows. In section 1.1, I present a general formulation of sequential decision problems, or equivalently, of adaptive trials — the type of statistical experiment that these are equivalent to. In section 1.2, I present various settings, including dynamic treatment regimes, the stochastic contextual bandit model, and the Markov Decision Process model for reinforcement learning. In section 1.3, I detail various applications of my methodological and theoretical work. In section 1.4 I present the general orientation of my research goals, an in particular in which directions I strove to advance the state of the art. In particular, I formally characterize the requirements that ideal estimators for causal inference from sequentially collected data and sequential decision making algorithms should satisfy. Finally, in section 1.5, I give an overview of the chapters of my dissertation, and detail how they contribute to furthering the state of the art in the directions I outline in section 1.4.

1.1 General problem formulation

Data. The settings I consider in the various chapters of this dissertation are all special cases of the following general formulation. Suppose that an agent (or experimenter) interacts with nature (also referred to as the environment) over a succession of steps $t = 1, 2, \ldots$. At each step t, the agent observes a of vector $L_1(t)$ of variables, taking values in a set $\mathcal{L}_1$, which reflects the state of the environment, assigns a treatment vector (or action) $A(t)$ to the environment, and then observes a vector $L_2(t)$ of post-treatment variables taking values in a set $\mathcal{L}_2$. I assume that a time point specific outcome $Y(t)$ (or reward) can be defined for each t as a function of $L_2(t)$, that is $Y(t) = f_{Y(t)}(L_2(t))$ for some known function $f_{Y(t)}$. I denote $O(t) := (L_1(t), A(t), L_2(t))$ the observations collected in round t, and $\bar{O}(t) = (O(1), \ldots, O(t))$ the data collected up till time t. I suppose that $A(t)$ takes values in a set $\mathcal{A}$ which I will refer to as the action set. I denote $\mathcal{O} := \mathcal{L}_1 \times \mathcal{A} \times \mathcal{L}_2$ the domain of the observations $O(t)$, for $t = 1, 2, \ldots$. The domain of $\bar{O}(T)$ is then the Cartesian product $\mathcal{O}^T := \underbrace{\mathcal{O} \times \ldots \times \mathcal{O}}_{T \text{ times}}$.

Design. I suppose that at each time step t, the agent has access to $\bar{O}(t-1)$, and can adapt its rule for choosing $A(t)$ in response to the past. I denote g_t the treatment rule or policy that the agent follows in choosing $A(t)$. I view g_t as the conditional distribution of $A(t)$ given $\bar{O}(t-1)$, the data collected in previous rounds, and $\bar{L}_1(t)$ the pre-treatment variables vector. That is, for any t, $A(t) \mid \bar{O}(t-1), L_1(t) \sim g_t(\cdot \mid \bar{O}(t-1), L_1(t))$. I also refer to g_t as the design at time t, as it determines the way how data is collected at time t. I say that g_t is an *adaptive* design as the experimenter adjusts the distribution the the treatment based on the available data.

Adaptive trials. I refer to the statistical experiment that generates the data sequence $(\bar{O}(t))_{t \geq 1}$ as a *sequential adaptive trial.*

Components of the data-generating distribution. I suppose that the true data-generating distribution P_0^T admits a density w.r.t. some a known measure μ^T on $\mathcal{O}^T$. I denote $p_0^T := dP_0^T / d\mu^T$ the density of P_0^T w.r.t. μ^T. For any probability P^T distribution over $\mathcal{O}^T$ that is absolutely continuous w.r.t. μ^T, I denote $p^T := dP^T / d\mu^T$ its density w.r.t. μ^T.

From the chain rule, any such density p^T can be factored as

$$\begin{aligned} P^T(\bar{O}(T)) :=& \prod_{t=1}^T q_{1,t}(L_1(t) \mid \bar{O}(t-1)) \\ & \times \prod_{t=1}^T g_t(A(t) \mid \bar{O}(t-1), L_1(t)) \\ & \times \prod_{t=1}^T q_{2,t}(L_2(t) \mid A(t), L_1(t), \bar{O}(t-1)). \end{aligned}$$

Denoting $q_{1,1:T} := \prod_{t=1}^T q_{1,t}$, $q_{2,1:T} := \prod_{t=1}^T q_{2,t}$, $q_{1:T} = q_{1,1:T} q_{2,1:T}$, and $g_{1:T} = \prod_{t=1}^T g_t$, p^T can be written as $p^T = q_{1:T} g_{1:T}$. The factor $g_{1:T}$ is the collection of treatment rules or designs carried out by the experimenter along the trial. I refer to it as the *controlled* part of the data-generating density. The factor $q_{1:T}$ represents the dynamics of the environment, in response in particular to previous states and to the treatment history. It is a fact of nature, and is a priori unknown to the experimenter. I refer to it as the *uncontrolled* component of the data-generating density.

Causal model, causal targets and causal identifiability. I assume that the process generating the data $\bar{O}(T)$ can be represented as a Nonparametric Structural Equation Model (NPSEM). That is, I assume that there exists a collection of deterministic functions $(f_{L_1(t)}, f_{A(t)}, f_{L_2(t)} : t = 1, \ldots, T)$, and a collection $U(T) = (U_{L_1}(t), U_A(t), U_{L_2}(t))$ of exogenous random variables such that, for every t,

$$\begin{aligned} L_1(t) =& f_{L_1(t)}(\bar{O}(t-1), U_{L_1(t)}), \\ A(t) =& f_{A(t)}(\bar{O}(t-1), L_1(t), U_{A(t)}), \\ L_2(t) =& f_{L_2(t)}(\bar{O}(t-1), L_1(t), A(t), U_{L_2(t)}), \end{aligned}$$

where I set by convention $\bar{O}(0)$ to be a known constant. Consider now a counterfactual scenario in which I replace the treatment nodes in the above set of equations by a collection counterfactual interventions $(g^*_t : t \in [T])$. In this counterfactual scenario, I would observe data $\bar{O}^*(T) := (L^*_1(t), A^*(t), L^*_2(t) : t \in [T])$, which would obey the following set of equations:

$$\begin{aligned} L^*_1(t) =& f_{L_1(t)}(\bar{O}^*(t-1), U_{L_1(t)}), \\ A^*(t) \sim& g^*_t(A(t) \mid \bar{O}^*(t-1), L^*_1(t)), \\ L^*_2(t) =& f_{L_2(t)}(\bar{O}^*(t-1), L^*_1(t), A^*(t), U_{L_2(t)}). \end{aligned}$$

I use the notation P^T_F for any generic distribution over the domain of the full data, that is of the couple $(\bar{O}(T), U(T))$, under the observed collection of treatment rules $g = (g_t : t \in [T])$, and $P^{*,T}_F$ any generic distribution of the full data $(\bar{O}^*(T), U(T))$ under the counterfactual intervention. I refer to the distribution of $\bar{O}^*(T)$ as the *post-intervention distribution*. I denote it $P^{*,T}$. I denote $P^T_{0,F}$ the true full data-generating distribution, $P^{*,T}_{0,F}$ the corresponding distribution of the full data under the counterfactual intervention, and $P^{*,T}_0$ the true post-intervention distribution. Throughout my dissertation, I use the subscript "0" to indicate that I am referring to the true data-generating distribution or features thereof.

I define causal parameters as features of the post-intervention distribution, that is causal parameters can be written as $\Psi_F(P^{*,T})$ for some mapping Ψ_F. Under identifiability assumptions, which I state next, I can identify these causal parameters from P^T, the true distribution of the observed data. I now state the identifiability assumptions:

Assumption 1.1 (Sequential randomization)**.** *For any $t \geq 1$, and $\tau > t$ $A(t) \perp\!\!\!\perp O(\tau), O^*(\tau) \mid \bar{O}(t-1), L_1(t)$.*

Assumption 1.2 (Positivity)**.** *For any $t \geq 1$, and any $a \in \mathcal{A}$, and any $l_1(t), \bar{o}(t-1)$ that has positive likelihood under p^T_0,*

$$g_t(A(t) \mid l_1(t), \bar{o}(t-1)) > 0.$$

Under assumptions 1.1 and 1.2, the post-intervention distribution equals the G-computation formula $P^T_{g^*}$, defined as follows:

$$\frac{dP^T_{g^*}}{d\mu^T}(\bar{o}(T)) := \prod_{t=1}^T q_{1,t}(l_1(t) \mid \bar{o}(t-1))$$

$$\times \prod_{t=1}^{T} g_t^*(A(t) \mid \bar{O}(t-1), l_1(t))$$
$$\times \prod_{t=1}^{T} q_{2,t}(l_2(t) \mid \bar{o}(t-1), q(t), l_1(t)).$$

Since the factors $q_{1,t}$ and $q_{2,t}$ are known if one knows P^T, then the G-computation formula $P^T_{g^*}$, and thus the post intervention distribution $P^{*,T}$ it equals, can be obtained from P^T. As a result, under these assumptions, for any causal parameter of the form $\Psi_F(P^{*,T})$, there exists a mapping Ψ such that $\Psi_F(P^{*,T}) = \Psi(P^T_{g^*})$. Since we consider sequential trials in which treatment decisions at t are based on the observed past $L_1(t), \bar{O}(t-1)$, assumption 1.1 is always satisfied. I will also assume that assumption 1.2 is satisfied in the rest of this introduction chapter.

Examples of causal parameters of interest include the mean outcome at T, or the sum of outcomes across all time points in $[T]$, that one would observe if the experimenter had carried out interventions $g_1^*, \ldots, g_T^*$, which are respectively defined as

$$\Psi_{F,T}(P^{*,T}) := E_{P^{*,T}}[Y^*(T)] \qquad \text{and} \qquad \Psi_{F,1:T}(P^{*,T}) := E_{P^{*,T}}\left[\sum_{t=1}^{T} Y^*(t)\right],$$

and which equal the following parameters of the observed data distribution:

$$\Psi_T(P^T) := E_{P^T_{g^*}}[Y(T)] \qquad \text{and} \qquad \Psi_{1:T}(P^T) := E_{P^T_{g^*}}\left[\sum_{t=1}^{T} Y(t)\right].$$

Oracle designs. In various chapters of this dissertation I consider sequential adaptive designs that learn an oracle design, defined as a parameter of the uncontrolled component $q_{1:T}$ of the data-generating density. For this parameter to be univocally defined and learnable, I require that $q_{1:T}$ exhibit a repeated factor indexed by a parameter θ independent of T. A situation in which this is the case is for example when,

$$q_{2,t}(l_2(t) \mid \bar{o}(t-1), q(t), l_1(t)) = q_2(l_2(t) \mid q(t), l_1(t)).$$

for some conditional density q_2 independent of t, and therefore common across time points. In this situation, the aforementioned requirement holds with $\theta := q_2$. An oracle design can then be written as $g(\theta)$. In the adaptive designs I consider, g_t is an estimator of $g(\theta)$ computed from the vector $\bar{O}(t-1)$ of past observations.

Objective pursued by the experimenter. Different oracle designs are appropriate for different objectives. One goal the experimenter may pursue is to choose the sequence of designs $g_1, \ldots, g_T$ so as to maximize mean cumulative outcomes of the form $E_{P^T}[\sum_{t=1}^T Y(t)]$ over a fixed number of steps, or the asymptotic mean outcome, that is $\lim_{t \to \infty} E_{P^T}[Y(t)]$. In that case, a sensible choice of $g(\theta)$ is the optimal treatment rule $g^{\text{opt}}(\theta)$, or optimal policy, provided it is defined. Another reasonable choice of designs is a mixture distributions of the form $g_\epsilon(\theta) = (1-\epsilon) g^{\text{opt}}(\theta) + \epsilon g^{\text{unif}}$, where g^{unif} is the uniform distribution over the action set $\mathcal{A}$, and where $\epsilon \in (0, 1)$ represents an exploration rate. Another objective is to maximize the sample efficiency for a certain statistical parameter $\Psi(\theta)$ of the data-generating distribution, that is learning the design $g(\theta)$ that would minimize the variance of a certain class of estimators of $\Psi(\theta)$ if the experimenter had carried it out from the beginning of the experiment.

1.2 Settings covered by the general problem formulation

We consider various particular cases of the general formulation from the previous section. These particular cases correspond to various causal models, that is sets of distributions P_F^T over the domain of the full data $(\bar{O}(T), U(T))$ that we believe to contain the true full data distribution $P_{0,F}^T$. These causal models induce corresponding statistical models, that is sets of distributions P^T over the observed data domain $\mathcal{O}^T$, which we therefore believe to contain the true data-generating distribution P_0^T. In each of the following setting, we denote $\mathcal{M}_F^T$ the causal model, and $\mathcal{M}^T$ the statistical model.

1.2.1 Stochastic contextual bandit

Conditional independence assumptions, and homogeneity assumptions. The i.i.d. stochastic contextual bandit setting is a particular case of the setting I describe above, where I let $(L_1(t))_{t \geq 1}$ be an i.i.d. sequence of contexts, $A(t)$ be the action at time t, and I let $L_2(t) = Y(t)$ be the reward at time t. In the stochastic contextual bandit setting, for every t, $Y(t)$ depends on the past $\bar{O}(t-1), L_1(t), A(t)$ only through $L_1(t)$ and $A(t)$, and the conditional distribution of $Y(t)$ given $A(t)$ and $L_1(t)$ is the same for every t, that is, there exists a certain marginal density q_1 and a certain conditional density q_2, both independent of t, such that, for every $t \geq 1$,

$$\begin{aligned} q_{1,t}(l_1(t) \mid \bar{o}(t-1)) &= q_1(l_1(t)), \\ \text{and } q_{2,t}(y(t) \mid \bar{o}(t-1), l_1(t), q(t)) &= q_2(y(t) \mid l_1(t), q(t)). \end{aligned}$$

Optimal treatment rules, policies and policy class, and CB algorithm. As a result of theses conditional independence properties, the optimal action at each time point t depends only on the latest context $L_1(t)$. Denoting $\bar{q}_2(a, l_1) := E_{q_2}[Y(t) \mid A(t) = a, L_1(t) = l_1]$, the outcome regression function under q_1, the optimal treatment rule at each time step is the mapping d_{q_2} that maps any context l_1 to the action that would give the highest expected outcome given the context l_1, that is

$$d_{q_2}(l_1) = \arg\max_{a \in \mathcal{A}} \bar{q}_2(a, l_1),$$

where ties are broken arbitrarily. In the contextual bandit setting, I find it convenient to view the design at time t, or policy at time t as an $\bar{O}(t-1)$-measurable density probability distribution over $\mathcal{A}$ conditional on $L_1(t)$, that is

$$A(t) \mid L_1(t) \sim g_t(\cdot \mid L_1(t)),$$

where $g_t = \widetilde{g}_t(\bar{O}(t-1))$ for some deterministic $\widetilde{g}_t$ which maps the treatment history to a class $\mathcal{G}$ of conditional distributions. I refer to the class $\mathcal{G}$ as the policy class. The policy class $\mathcal{G}$ may or may not contain the optimal treatment rule d_{q_2}. In the former case, I say that the so-called *realizability* assumption is satisfied.

The deterministic sequence $(\widetilde{g}_t)_{t \geq 1}$ of mappings from trial history to stochastic treatment rules or policies fully specifies the contextual bandit algorithm. I will simply equate the two concepts and refer to $\widetilde{g} := (\widetilde{g}_t)_{t \geq 1}$ as the algorithm.

Distribution of the data induced by a CB algorithm, value of a policy, regret. We recall that we denote $\mathcal{M}^T$ the statistical model specified by the aforementioned restrictions on the data-generating distribution.

Any distribution P^T of $\bar{O}(T)$ in $\mathcal{M}^T$ is fully specified by the triple $(q_1, \widetilde{g}_{1:T}, q_2)$, as we have that its density w.r.t. μ^T satisfies:

$$p^T(\bar{O}(T)) = \prod_{t=1}^{T} q_1(L_1(t)) \times \widetilde{g}_t(\bar{O}(t-1))(A(t) \mid L_1(t)) \times q_2(L_2(t) \mid A(t), L_1(t)).$$

I then write the expectation operator under P^T as $E_{q_1, \widetilde{g}_{1:T}, q_2}$. For any fixed g, I let P_{q_1,g,q_2} be the distribution over $\mathcal{L}_1 \times \mathcal{A} \times \mathcal{L}_2$ defined by $P_{q_1,g,q_2} := q_1 g q_2$, and I let E_{q_1,g,q_2} be the expectation operator under P_{q_1,g,q_2}. Let $(L_1, A, Y) \sim P_{q_1,g,q_2}$. The value of a policy g under $q := (q_1, q_2)$ is defined as

$$\mathcal{V}_q(g) := E_{q_1,g,q_2}[Y],$$

The instantaneous regret under q of a fixed policy g w.r.t. policy class $\mathcal{G}$ is defined as

$$\mathrm{reg}_q(g, \mathcal{G}) := \sup_{g' \in \mathcal{G}} \mathcal{V}_q(g') - \mathcal{V}_q(g),$$

and the cumulative regret at T under q of algorithm $\widetilde{g}$ is defined as

$$\mathrm{Reg}_{q,T}(g, \mathcal{G}) := T \sup_{g' \in \mathcal{G}} \mathcal{V}_q(g') - E_{q_1, \widetilde{g}_{1:T}, q_2} \left[\sum_{t=1}^{T} Y(t) \right].$$

Learning goals. I consider two types of learning goals under the stochastic CB framework.

The first one is the cumulative regret minimization goal. The corresponding question asks, given a policy class $\mathcal{G}$, and maximum time point (or horizon) T, how to devise an algorithm $\widetilde{g}$ that makes $\mathrm{Reg}_{q,T}$ as small as possible.

The second learning goal is the question of how, given a known stochastic treatment rule g^*, to perform inference for the value $\mathcal{V}_q(g^*)$ of g^*. The statistical parameter $\mathcal{V}_q(g^*)$ has a causal interpretation as the mean counterfactual outcome we would observe if the experimenter carried out stochastic intervention g^*.

1.2.2 Dynamic treatment regimes.

Independence assumptions. The DTR literature considers data sets consisting of a certain number N of independent and identically distributed draws $\bar{O}(T,1), \ldots, \bar{O}(T,N)$, where, for each $i \in [N]$, $\bar{O}(T,i)$ is a longitudinal data structure over T time points, of the form

$$\bar{O}(T,i) := (L_1(1,i), A(1,i), L_1(1,i), \ldots, L_1(T,i), A(T), L_2(T,i)).$$

Oftentimes in practical applications, each of these draws correspond to observations collected on an individual that I follow along T time steps. I refer to $\bar{O}(T,i)$ as the trajectory of individual i. For any $t \geq 1$, and $X \in \{L_1, A, L_2, O, \bar{O}\}$, I let $X(t) := (X(t,i) : i \in [N])$. I use the notation $O^{T,N}$ for $\bar{O}(T)$, so as to make explicit in the notation the number of individuals, and I write $P^{T,N}$ instead of P^T for the distribution of $O^{T,N}$ This setting a special case of the general setting described in section 1.1, where due to the independence between trajectories, p^T can be further factorized as follows:

$$\begin{aligned} p^T(\bar{o}(T)) = p^{T,N}(o^{T,N}) = & \prod_{t=1}^{T} \prod_{i=1}^{N} q_{1,t}(l_1(t,i) \mid \bar{o}(t-1,i)) \\ & \times \prod_{t=1}^{T} \prod_{i=1}^{N} g_t(a(t,i) \mid l_1(t,i), \bar{o}(t-1,i)) \\ & \times \prod_{t=1}^{T} \prod_{i=1}^{N} q_{2,t}(a(t,i) \mid l_1(t,i), \bar{o}(t-1,i)). \end{aligned}$$

Observe that the above factorization makes no assumption on the dependence structure within a trajectory, which can be arbitrary.

Design adaptivity. This formulation of the DTR setting allows for the treatment decision $A(t,i)$ for individual i at time point t to depend on $\bar{O}(t-1,i), L_1(t,i)$, that is the past of i up till $L_1(t,i)$. Note also that the conditional distributions $g_1, \ldots, g_T$ are known in advance, and need to be the same across individuals.

As a result, this formulation limits the design adaptivity in several ways. First, it doesn't allow for the design at t to pool across trajectories the available information to inform the treatment rule at t. It also excludes the setting where the trajectories are observed one after the other – say j is observed after i if $j > i$ – and the treatment rules of trajectory j uses information from trajectory i.

Learning goals. Parameters of interest include

$$\Psi_T(P^T) := E_{P^T_{g^*}}[Y(T)] \qquad \text{and} \qquad \Psi_{1:T}(P^T) := E_{P^T_{g^*}}\left[\sum_{t=1}^{T} Y(t)\right],$$

which equal the following causal parameters: the mean counterfactual outcome at T under g^*, and the mean counterfactual sum of outcomes across time points, under g^*. A common learning goal is to perform inference for these statistical parameters.

1.2.3 Reinforcement learning in the MDP framework.

Dependence structure under the Markov Decision Process model. In reinforcement learning, it is standard to assume that the agent interacts with a system consisting of a unit, or a collection of units indexed by $i = 1, \dots, N$, where the trajectory of the unit along the experiment is modelled by a Markov Decision Process.

The homogeneous MDP model in the case of a single unit. In the case of a single unit, the trajectory of the system over T time steps is a data structure of the form $\bar{O}(T)$ as defined in section 1.1, where $L_1(t)$ is the state of the unit at the beginning of round t, $A(t)$ is the action or treatment assigned by the agent at time t, and $L_2(t)$ is the reward at time t collected by the agent. Saying that the trajectory of the unit is an MDP means that for any t, the condition distribution of the reward $L_2(t)$ and of the next state $L_2(t+1)$ given $\bar{O}(t-1), L_1(t), A(t)$ depends only on $L_1(t)$ and $A(t)$. The MDP is said to be homogeneous if these conditional distributions are identical across time points. For any distribution P^T over $\mathcal{O}^T$ that is absolutely continuous w.r.t. μ^T, its density $p^T := dP^T/d\mu^T$ factorizes as

$$p^T(\bar{o}(T)) := \prod_{t=1}^{T} q_1(l_1(t) \mid l_1(t-1), a(t-1)) g_t(a(t) \mid \bar{o}(t-1), l_1(t)) q_2(l_2(t) \mid a(t)),$$

for some conditional distributions q_1 and q_2 that do not depend on t. Note that this model does not place restrictions on the dependence of $A(t)$ on the past.

Concurrent homogeneous MDPs in the case of several units. Consider the situation where the agent interacts with N units at the same time. For every $i = 1, \dots, N$, I denote $\bar{O}(T, i)$ the trajectory of unit i over T time steps. As in the previous subsection, for any $X \in \{L_1, A, L_2, O, \bar{O}\}$, I let $X(t) := (X(t,i) : i \in [N])$. I dentote $O^{T,N} := \bar{O}(T)$ and $P^{T,N} := P^T$ so as to make explicit the number of units N.

A natural modelling assumption is to assume that the trajectory of each unit is an homogeneous MDP, and that at every $t \geq 1$, the reward $L_2(t,i)$ collected from unit i and the next state $L_1(t+1, i)$ of unit i depends on the past of the trial $\bar{O}(t-1), L_1(t), A(t)$ only through the the latest state $L_1(t,i)$ and the latest treatment $A(t,i)$ assigned to unit i. These modelling assumptions are equivalent to the following data-generating density factorization

$$\begin{aligned} p^{T,N}(o^{T,N}) = p^T(\bar{o}(T)) = & \prod_{t=1}^{N} \prod_{i=1}^{N} q_{1,t}(l_1(t,i) \mid l_1(t-1,i), a(t-1,i)) \\ & \times \prod_{t=1}^{T} g_t(a(t) \mid \bar{o}(t-1), \bar{l}_1(t)) \\ & \times \prod_{t=1}^{T} \prod_{i=1}^{N} q_{2,t}(l_2(t,i) \mid l_1(t,i), a(t,i)). \end{aligned}$$

Observe that the above described model for concurrent MDPs lets the treatment decision vector $A(t)$ depend on the past of the entire trial. As a result, this introduces dependence between the trajectories $\bar{O}(T,1),\dots,\bar{O}(T,N)$.

A further modelling assumption is to impose that the treatment decision at time t for individual i only depends on the past of individual i, and that the conditional distribution of $A(t,i)$ given the past of i is the same for every i. The model thus described is the set of distributions $P^{T,N}$ over the domain of $\bar{O}^{T,N}$ such that the the data-generating density $p^{T,N} := p^T := dP^T/d\mu^T$ factorizes as

$$
\begin{aligned}
p^{T,N}(o^{T,N}) := p^T(\bar{o}(T)) =& \prod_{t=1}^{N}\prod_{i=1}^{N} q_{1,t}(l_1(t,i) \mid l_1(t-1,i), a(t-1,i)) \\
&\times \prod_{t=1}^{N}\prod_{i=1}^{N} g_t(a(t,i) \mid \bar{o}(t-1,i), \bar{l}_1(t,i)) \\
&\times \prod_{t=1}^{T}\prod_{i=1}^{N} q_{2,t}(l_2(t,i) \mid l_1(t,i), a(t,i)).
\end{aligned}
$$

In this case, treatment trajectories of distinct units are independent and identically distributed. Observe that this model is a particular case of the DTR model presented in the previous subsection.

Optimal policies and regret. I now discuss the notions of optimal policies and regret in the case of a single MDP.

Optimal greedy policy. Since the reward $L_2(t)$ only depends on the past through $L_1(t)$ and $A(t)$, the treatment rule at t that maximizes the expectation of the next outcome $L_2(t)$ is a mapping that takes only $L_1(t)$ as input, and is defined as

$$
d_{q_2}^{\text{inst}}(l_1(t)) := \arg\max_{a\in\mathcal{A}} E_{q_2}[L_2(t) \mid A(t) = a, L_1(t) = l_1(t)].
$$

Optimal sequence of policies for finite horizon. Carrying out $d_{q_2}^{\text{inst}}$ at every time step might not be the strategy that maximizes the reward, since it may be that the action that has the largest instantaneous payoff makes the system transition to states from which it is subsequently harder to obtain good payoffs. The set of treatment rules $(d_{q,t} : t \in [T])$ that would maximize the expected cumulative reward $\sum_{t=1}^T L_2(t)$ is defined recursively as follows. Let

$$
\begin{aligned}
d_{q,T}(l_1(T)) :=& \arg\max_{a\in\mathcal{A}} E_{q_2}\left[L_2(T) \mid A(T) = a, L_1(T) = l_1(T)\right], \\
\text{and } V_{q,T}(l_1(T)) :=& E_{q_2,d_{q,T}}[L_2(T) \mid L_1(T) = l_1(T)] \\
=& E_{q_2}\left[L_2(T) \mid L_1(T) = l_1(T), A(T) = d_{q_2,T}(l_1(T))\right].
\end{aligned}
$$

and, for any $t < T$,

$$
d_{q,t}(l_1(t)) = \arg\max_{a\in\mathcal{A}} \{E_{q_2}\left[L_2(t) \mid A(t) = a, L_1(t) = l_1(t)\right]
$$

$$+ \gamma \, E_{q_1} \left[V_{q,t+1}(L_1(t+1)) \mid L_1(t) = l_1(t), A(t) = a \right] \right\},$$

$$\text{and } V_{q,t}(l_1(t)) := E_{q,d_{q,t:T}} \left[\sum_{s=t}^{T} L_2(s) \mid L_1(t) = l_1(t) \right].$$

In the MDP setting, I call policy a conditional density $(l_1, a) \mapsto g_t(a \mid l_1)$. An adaptive design scheme or reinforcement learning algorithm is specified by a deterministic sequence $\widetilde{g}_t$ such that, for any t, $\widetilde{g}_t$ maps the past $\bar{O}(t-1)$ to a policy. For a given fixed sequence $g = (g_t : t \in [T])$ of policies, I define the value of g as

$$\mathcal{V}_q(g) := E_{q,g_{1:T}} \left[\sum_{t=1}^{T} L_2(t) \right],$$

Similarly, for a reinforcement learning algorithm, I define the value of this algorithm as

$$\mathcal{V}_q(g) := E_{q,\widetilde{g}_{1:T},q_2} \left[\sum_{t=1}^{T} L_2(t) \right].$$

Given a certain policy class $\mathcal{G}$, I define the regret of a fixed sequence g or of an algorithm $\widetilde{g}$ as, respectively,

$$\mathrm{Reg}_{q,T}(g, \mathcal{G}) := \sup_{g' = (g'_1, \ldots, g'_T) \in \mathcal{G}^T} \mathcal{V}_q(g') - \mathcal{V}_q(g),$$

$$\text{and } \mathrm{Reg}_{q,T}(\widetilde{g}, \mathcal{G}) := \sup_{g' = (g'_1, \ldots, g'_T) \in \mathcal{G}^T} \mathcal{V}_q(g') - E_{q_1,\widetilde{g}_{1:T},q_2} \left[\sum_{t=1}^{T} L_2(t) \right].$$

Optimal policy in the discounted infinite horizon setting. The agent might also want to maximize the expected sum of the total discounted reward over an infinite horizon, defined as $\sum_{t=1}^{\infty} \gamma^t L_2(t)$, with $\gamma \in (0,1)$. In this case, it can be shown that the optimal intervention to carry out at each time point is identical across time points, and is given by

$$d_{q,\gamma}(l_1) := \arg\max_{a \in \mathcal{A}} \bar{q}_q(a, l_1),$$

where $\bar{q}_q$ is the so-called action value function and is defined as the solution of the Bellman equation:

$$\begin{aligned}\bar{q}_q(a, l_1) =& E_{q_2} \left[L_2(1) \mid A(1) = a, L_1(1) = l_1 \right] \\ &+ E_{q_1} \left[\max_{a \in \mathcal{A}} \bar{q}_q(a, L_1(2)) \mid A(1) = a, L_1(1) = l_1 \right].\end{aligned}$$

In this latter setting, since the optimal policy is the same across time points, it makes sense to define the value of a single policy g (as opposed to of a sequence of policies $g = (g_t : t \in [T])$):

$$\mathcal{V}_{q,\gamma}(g) := E_{q,g} \left[\sum_{t=1}^{\infty} \gamma^t L_2(t) \right],$$

where $E_{d,g}$ is the expectation under P^T defined by

$$dP^T/d\mu^T(o(T)) := \prod_{t=1}^{T} q_1(l_1(t) \mid l_1(t-1), a(t-1)) g(a(t) \mid l_1(t)) q_2(l_2(t) \mid a(t), l_1(t)).$$

The regret of a fixed policy g is then defined as

$$\mathrm{Reg}_{q,\gamma}(g, \mathcal{G}) := \sup_{g' \in \mathcal{G}} \mathcal{V}_{q,\gamma}(g') - \mathcal{V}_{q,\gamma}(g),$$

and the regret of an algorithm $(\widetilde{g}_t : t \geq 1)$ is defined as

$$\mathrm{Reg}_{q,\gamma}(\widetilde{g}, \mathcal{G}) := \sup_{g' \in \mathcal{G}} \mathcal{V}_{q,\gamma}(g') - E_{q,\widetilde{g}} \left[\sum_{t=1}^{\infty} \gamma^t L_2(t) \right].$$

Learning goals. A common learning goal is to perform inference for the mean outcome under a sequence of counterfactual stochastic treatment rules $g^* = (g_t^* : t \in [T])$ or $g^* = (g_t^* : t \geq 1)$, from passive data collected from an already finished trial, in which the treatment rules might have been adaptive. A related goal is to learn the optimal treatment rule from the same type of passive data. These goals are called off-policy evaluation (and inference) and off-policy policy learning in reinforcement learning.

Another goal is to learn and implement the optimal policy sequentially, by executing a certain algorithm $\widetilde{g}$, so as to maximize the total expected discounted reward $E_{q,\widetilde{g}}[\sum_{t=1} L_2(t)]$, or given a certain policy class $\mathcal{G}$, minimize the regret with respect to this policy class.

1.2.4 Time series of networks

Statistical model. I now present a statistical model of an adaptive trial over T time points with N units, where network dependence is allowed between units. As in the previous subsections, I denote the observed data set $O^{T,N}$, and its distribution $P^{T,N}$. I assume that there exists a collection of deterministic functions $\{c_{L_1(t,i)}, c_{L_2(t,i)} : t \in [T], i \in [N]\}$, where, for every t and i, $c_{L_1(t,i)}$ and $c_{L_2(t,i)}$ map $\bar{O}(t-1)$, and $(\bar{O}(t-1), L_1(t), A(t))$ into a set $\mathcal{C} \subset \mathbb{R}^d$, for some $d \geq 1$, such that the data-generating density $p^{T,N} := p^T := dP^T/d\mu^T$ can be written as

$$\begin{aligned} p^{T,N}(o^{T,N}) = & \prod_{t=1}^{T} \prod_{i=1}^{N} q_1(l_1(t,i) \mid c_{l_1(t,i)}(\bar{o}(t-1))) \\ & \times \prod_{t=1}^{T} g_t(a(t) \mid \bar{o}(t-1), l_1(t)) \\ & \times \prod_{t=1}^{T} \prod_{i=1}^{N} q_2(l_2(t,i) \mid c_{l_2(t,i)}(\bar{o}(t-1), l_1(t), a(t))). \end{aligned}$$

Let $C_{L_1}(t,i) := c_{L_1(t,i)}(\bar{O}(t-1))$ and $C_{L_2}(t,i) := c_{L_2(t,i)}(\bar{O}(t-1), L_1(t), A(t))$. Following existing terminology, I refer to $C_{L_1}(t,i)$, and $C_{L_2}(t,i)$ as the "contexts" for $L_1(t,i)$ and $L_2(t,i)$.

Learning goals. As in the previous subsection, one learning goal is to perform inference for the mean outcome at a certain time point, or cumulative mean outcome, from already collected data. Another learning goal is to sequentially learn and carry out a sequence of policies so as to maximize a cumulative mean outcome, or a mean final outcome.

1.3 Applications

1.3.1 Clinical trials.

Although adaptive designs are not the standard practice yet in clinical trials Bhatt and Mehta [2016], abundant literature exists on the subject and the theory is fully mature (see e.g. van der Laan [2008]).

In a trial aimed at conducting inference for the average treatment effect (ATE) of a certain drug, an adaptive design can learn the optimal oracle design, that is the one that maximizes efficiency for the statistical parameter of interest, in such a way that the asymptotic variance of an estimator of the ATE is the same as it would be had the optimal design been carried out since the beginning of the trial. van der Laan [2008] provides a comprehensive methodological framework for adaptive clinical trials.

When there is only one treatment node and one health outcome, and that patients are received sequentially, the observations collected the t-th are a triple $(L_1(t), A(t), Y(t))$, where $L_1(t)$ is the set of covariates, $A(t)$ is the treament received and $L_2(t)$ is the health outcome. When patients can be assumed to be i.i.d., the adaptive trial is an instance of the stochastic contextual bandit framework outlined above.

In practice, it is more common to consider group sequential adaptive designs, in which the outcomes of an entire batch of patient is observed before enrolling and assigning treatment to a new batch.

1.3.2 Mobile health and precision medicine

In mobile health applications [Steinhubl et al., 2013, Malvey and Slovensky, 2014, Istepanian and Woodward, 2017, Istepanian and Al-Anzi, 2018], data are collected on patients at a high frequency by some connected measuring devices, and treatments are assigned algorithmically based on these measurements, the observed response of the patient to past treatments, and potentially the history of other patients.

In mobile health, one goal is to adaptively learn and implement the intervention that maximizes the outcome at each time point.

One instance of application of mobile health is Just-In-Time Adaptive Interventions (JITAI) in which at each time point, the default action is to not intervene, unless the value of the patient's measurement vector makes it appear that it is especially worth it to do so [Spruijt-Metz and Nilsen, 2014]. One example is exercise encouragement systems, in which the patient receives a notification on their phone encouraging them to exercise, at the time when they are most likely to be receptive. The notification is triggered when the patient is likely to be most receptive or when it might be most beneficial to exercise (for example when it has been a long time since the last time the patient exercised).

In a mobile health application, if there is a reason to believe that the patient's measurement vectors are i.i.d. and that the at every time point depends only on the latest treatment and measurement vector, and that this dependence follows the same law across time points, an appropriate model for the adaptive trial is the stochastic contextual bandit model. I refer the reader to Tewari and Murphy [2017] for more details on the application of contextual bandits to mobile health.

In many settings, it is not realistic to assume such independence properties, and the trajectory of the patient might be better modelled by either an MDP, or a time series of the type presented in subsection 1.2.4, or by an unrestricted DTR model. In the latter case, one will need several patients to learn the optimal treatment rule, while in the former two cases, it is possible to learn it from one single patient's trajectory.

1.3.3 Online ads placement and online recommender systems

The ad selection problem or item recommendation problem in web applications can be modelled as follows. The web platform sequentially receives user sessions indexed by $t = 1, 2, \ldots$. At each session, the platform first observes a vector $L_1(t)$ of characteristics of the user and of the session, then chooses an action $A(t)$ and then observes an outcome. In an ads setting, a natural outcome is whether the user clicked on the ad that was presented to them.

1.3.4 Sequential trial for public health policy evaluation in an infectious disease setting

Consider the hypothetical situation where the experimenter follows the N inhabitants of a city over successive time steps $t = 1, 2, \ldots$, and where she cares about learning and evaluating public health interventions to limit the spread of an infectious disease. For example, at each time t, the experimenter or decision maker can choose to force individuals to stay at home for a certain period of time, or vaccinate individuals depending on some characteristics such as age, occupation and health history. Natural outcomes of interest include the infection status, and the mortality status. Using the notation presented earlier, I denote $L_1(t, i)$ a vector of measurements characterizing the state of individual i at time t, $A(t, i)$ the intervention assigned to i at t, and $L_2(t, i)$ the outcome for i at t.

Since individuals are interconnected through contagion effects, their trajectories are not independent, and therefore there is only one independent trajectory to learn from, the one of the entire city. If domain knowledge justifies assuming that $L_1(t, i)$ and $L_2(t, i)$ depends on the past through finite dimensional summary vectors, and that this dependence is identical across individuals and time points, then the model presented in subsection 1.2.4 is appropriate. As we show in chapter *[causal inference from a single time series of connected units]*, under this model, and under additional assumptions, the effective sample size at time point T of the trial is $N \times T$.

1.4 Research goals

The overarching goal of my research efforts is to design methods for causal inference from sequentially collected data and sequential decision making that rely only on models encoding the available domain knowledge. This is in contrast with methods that make parametric assumptions on unknown components of the data-generating distribution.

For such methods to have practical utility, they must satisfy a certain number of requirements. In the next subsections, I present the formal requirements that ideal estimators for causal inference problems and sequential decision making algorithms should satisfy. I distinguish the requirements that apply to off-policy inference and the ones that pertain to sequential decision making. As pointed out at the beginning of the chapter, these two tasks are narrowly connected, and as a result there is some interplay between the requirements that apply to either task.

1.4.1 Formal requirements for causal estimators

Statistical model for q. We believe that statistical models on the unknown and uncontrolled part q of the data-generating distribution should only encode what we know from domain knowledge. In some applications, such domain knowledge might tell us that the individual trajectories are independent, in some other we might in addition be founded to assume that successive observations of the same trajectory exhibit further conditional independence. While domain knowledge might justify conditional independence assumptions, or homogeneity assumptions (that is that a factor of the likelihood is constant across time points and or individuals), it seems rather implausible in most settings that it would justify assuming a parametric model for components of the likelihood. This is why we believe that, while the models we should be working with can impose restrictions on the dependence structure, and can impose repeated factors in the factorization of the likelihood, these factors should be modelled nonparametrically. To make things more concrete, take for instance the case of the single trajectory MDP model presented in subsection 1.2.3. This model assumes that under the data-generating distribution P^T, the likelihood of a trajectory $\bar{O}(T)$ over T time steps factorizes as

$$\begin{aligned} p^T(\bar{O}(T)) = & \prod_{t=1}^{T} q_1(L_1(t) \mid L_1(t-1), A(t-1)) \\ & \times \prod_{t=1}^{T} g_t(A(t) \mid \bar{O}(t-1), L_1(t)) \\ & \times \prod_{t=1}^{T} q_2(L_2(t) \mid L_1(t), A(t)), \end{aligned} \tag{1.1}$$

but makes no further assumption on the factors q_1 and q_2, that is these are assumed to be modelled fully nonparametrically, or in more formal terms, to belong to *saturated* nonparametric models $\mathcal{M}_{q_1}$ and $\mathcal{M}_{q_2}$. Let me now formally define the notion of a saturated nonparametric model. Take for example the model $\mathcal{M}_{q_1}$. We say that $\mathcal{M}_{q_1}$ is a saturated nonparametric model of conditional distributions of $L_1(t)$ given $A(t-1)$ and $L_1(t-1)$, if, for any $q_1 \in \mathcal{M}_{q_1}$, the tangent space of $\mathcal{M}_{q_1}$ at q_1 is equal to the Hilbert space

$$\Big\{(l_1, a, l_1') \mapsto s(l_1, a, l_1') : \forall l_1, a \int s^2(l_1, a, l_1') q(l_1' \mid a, l_1) dl_1' < \infty \\ \text{and} \int s(l_1, a, l_1') q(l_1' \mid a, l_1) dl_1' = 0\Big\}.$$

As a result, a more complete description of the model $\mathcal{M}^T$ for MDPs over T time steps is: the set of distributions that factorize as in (1.1), where q_1 and q_2 vary freely over the nonparametric saturated models $\mathcal{M}_{q_1}$ and $\mathcal{M}_{q_2}$.

Estimators of a given parameter $\Psi(q)$ often rely on intermediate estimators of infinite dimensional components $\eta_1(q), \ldots, \eta_p(q)$ of q, which are often referred to as *nuisance* parameters. Consider the nuisance parameter $\eta_1(q)$, and suppose for example that it is a d-variate real valued function. The saturated models $\mathcal{M}_{q_1}$ and $\mathcal{M}_{q_2}$ induce a nonparametric model for $\mathcal{M}_{\eta_1(q)}$. Existing nonparametric estimators of d-variate real valued functions usually learn functions in and have convergence guarantees over nonparametric classes of functions that are subsets of the model $\mathcal{M}_{\eta_1(q)}$ induced by the saturated model for q. Some of these nonparametric classes, or the union of a collection of such nonparametric classes can form a realistic model $\mathcal{M}'_{\eta_1(q)}$ for the nuisance parameter $\eta_1(q)$, even if $\mathcal{M}'_{\eta_1(q)}$ is a subset of the fully saturated model $\mathcal{M}_{\eta_1(q)}$. As a result, we will find it satisfactory enough to assume that the nuisance parameter $\eta_1(q)$ belongs to such a nonparametric model $\mathcal{M}'_{q_1} \in \mathcal{M}_{q_1}$, even though, as we further discuss in the next paragraph, we will be content with semiparametric efficiency of estimators with respect to the larger model $\mathcal{M}_{\eta_1(q)}$.

Semiparametric efficiency for estimators of pathwise differentiable parameters. Semiparametric efficiency theory tells us that, given a model $\mathcal{M}$, and a parameter $\Psi : \mathcal{M} \to \mathbb{R}$ satisfying a certain regularity condition, namely pathwise differentiablity at P w.r.t. $\mathcal{M}$, all "non-pathological"[1] estimators, must have asymptotic variance at least as large as a certain quantity, the *semiparametric efficiency bound* for Ψ at P w.r.t. $\mathcal{M}$, which I denote $\mathrm{EB}(\Psi, \mathcal{M}, P)$. The *semiparametric efficiency bound* is also referred to as the *generalized Cramer-Rao* lower bound. Under the model $\mathcal{M}$, an ideal estimator of $\Psi(P)$ should have asymototic variance equal to the semiparametric efficieny bound. Semiparametric efficient estimators have been an intense areas of research for many years. Early contributions include the one-step estimator and the estimating equation methodology. The targeted minimum loss estimation frameworks allows to derive a locally semiparametric efficient estimator for any pathwise differentiable target parameter. Observe that the generalized Cramer-Rao lower bound is an instance dependent bound (that is bound that depends on the actual data-generating distribution, as opposed to minimax lower bouds, which corresponds to a worst-case distribution not necessarily equal to P. Instance dependent lower bounds are generally less pessimistic than minimax bounds).

Robustness. Given statistical model $\mathcal{M}$, an ideal estimator of a pathwise differentiable target parameter $\Psi : \mathcal{M} \to \mathbb{R}$ should inherit the robustness properties of the canonical gradient of Ψ

[1] more rigorously, all estimators of $\Psi(P)$ that are *regular* at P w.r.t. $\mathcal{M}$

w.r.t. $\mathcal{M}$.

Coverage of confidence intervals. An ideal estimator of a pathwise differentiable target parameter should come with confidence intervals that should be at least asymptotically valid, and ideally valid in finite samples too.

Substitution estimators. We say that an estimator $\widehat{\Psi}$ of a parameter $\Psi : \mathcal{M} \to \mathbb{R}$ is a substitution estimator of $\Psi(P_0)$ if it can be written as $\Psi(\widehat{P}_n)$ where $\widehat{P}_n$ is an estimator of the components of the likelihood Ψ depends on. A substitution estimator respects the bounds of the paremeter space. For example, an substitution estimator of a probability always lies in $[0, 1]$.

Oracle efficient model selection for estimation of non-pathwise differentiable target parameters. Oftentimes, infinite dimensional parameters $\eta(q)$ such as optimal treatment rules or outcome models are non-pathwise differentiable. Various criteria can be applied to assess whether an estimator of such a parameter is satisfactory or not. Let me first review some candidate criteria, before presenting the one we find most satisfactory.

If one commits to a certain statistical model $\mathcal{M}_\eta$ for $\eta(q)$, a first candidate criterion would be *mimimax optimality*. A minimax optimal estimator over $\mathcal{M}_\eta$ achieve, up to $\log n$ factors, where n is the sample size, the mimimax estimation rate over $\mathcal{M}_\eta$. However, we find that minimax optimality over a fixed $\mathcal{M}_\eta$ isn't an entirely satisfactory notion for several reasons. Firstly, for it to be realistic that $\mathcal{M}_\eta$ contains the true parameter, one might have to consider a very large nonparametric model $\mathcal{M}_\eta$. Minimax rates over such large models can be very slow. An instance of such a very large model over which the minimax estimation rate is very slow is the class of d-variates functions that are 1-time continuously differentiable. The corresponding minimax estimation rate is $n^{-1/(d+1)}$.

Nevertheless, it might turn out that $\eta(q)$ lies in a small submodel $\mathcal{M}_\eta$ of $\mathcal{M}_\eta$ over which the minimax estimation rate is much faster. In our example where $\mathcal{M}_\eta$ is the class of 1-time continuously differentiable d-variate functions, it might turn out that $\eta(q)$ is actually β times continuously differentiable, with $\beta > 1$. The minimax rate over this latter smaller class of functions is $n^{-\beta/(d+1)}$. In this case, given a collection of submodels $\mathcal{M}_{J,\eta} \subset \ldots \subset \mathcal{M}_{1,\eta} = \mathcal{M}_\eta$, we would like to have a model selection procedure that outputs an estimator with guaranteed rate of convergence matching, up to $\log n$ factors, the minimax rate of the smallest submodel $\eta(q)$ belongs to. An alternative approach to realistic modelling of $\eta(q)$ is, instead of assuming a collection of nested models, to assume that $\eta(q)$ belongs to the union of a finite collection of not necessarily nested models $\mathcal{M}_1, \ldots, \mathcal{M}_J$. In this case, similarly to the nested case, it would be desirable to have a model selection procedure that outputs an estimator with guaranteed rate of convergence within a $\log n$ factor of fastest minimax rate among the models of this collection that contain $\eta(q)$. An model selection procedure that achieves this requirement is called *minimax adaptive.* Model selection literature has proposed several such procedures, such as for example Lepski's method in the case of nested models.

Still, we don't find minimax adaptivity to be an entirely satisfactory criterion for estimators of non-pathwise differentiable target parameters. The reason is that given a collection of estimators

$\widehat{\eta}_1, \ldots, \widehat{\eta}_J$ that are consistent over models $\mathcal{M}_{\eta,1}, \ldots, \mathcal{M}_{\eta,J}$, respectively, the estimator that performs best at a certain data-generating distribution P^T might not be the one that corresponds to the model with best minimax rates among the models that contain $\eta(q)$.

The discussion presented so far in this paragraph suggests two things. Firstly, that it is best to work with a model selection combining several estimators or several models, rather than committing to a certain model ahead of time. Secondly, that we would rather have this procedure return the estimator that works best under the actual data-generating distribution, as opposed to a minimax adaptive procedure. A notion that formalizes this latter requirement is that of *oracle optimality*, which applies to a procedure selecting among a collection of estimators. I present this notion next.

Consider a collection $\widehat{\eta}_{1,n}, \ldots, \widehat{\eta}_{J,n}$ of estimators of $\eta(q)$, and let $\eta' \mapsto \ell(\eta', \eta(P))$ be a loss, where I use the notation η' for a generic element of the parameter space. Consider a model selection procedure that returns the index $\widehat{j}_n$ of an estimator. We say that the model selection procedure is oracle efficient if, for any $\epsilon > 0$,

$$E_{P^T}\left[\ell(\widehat{\eta}_{\widehat{j}_T,T}, \eta(q))\right] \leq (1+\epsilon) \min_{j \in [J]} E_{P^T}\left[\ell(\widehat{\eta}_{j,T}, \eta(q))\right] + R(\epsilon, P^T, T),$$

where $R(\epsilon, P^T, T)$ is an error term that is negligible in front of the first term of the right-hand side above. As opposed to a minimax adaptive model selection procedure, an oracle efficient model selection procedure achieves the rate of the estimator that performs best at the actual P that generated the data. In that sense it is an instance-dependent guarantee. Note that an oracle efficient ensemble learning procedure that combines minimax efficient estimators over a collection of classes of models is minimax adaptive.

1.4.2 Requirements for sequential decision making algorithms

I distinguish two goals: pure learning goals, in which the objective is to maximize the statistical efficiency for a certain parameter such as the ATE, or identify the best treatment arm as fast as possible, and regret minimization goals, in which the objective is to obtain as high a cumulative or final outcome as possible.

Pure learning goals.

Equivalence with the oracle design. Suppose that the goal of the experimenter is to conduct inference for a certain pathwise differentiable parameter $\Psi(Q_0)$ of the uncontrolled component of the data-generating distribution. A natural parameter of interest is the average treatment effect in the case of a binary treatment. If we knew Q_0 from the onset, we could carry out from the beginning the oracle design $g(Q_0) := (g_t(Q_0) : t \in [T])$ that minimizes the efficiency bound for $\Psi(Q_0)$ w.r.t. $\mathcal{M}^T$ at P_0^T. An adaptive design is a sequence $(\widehat{g}_t : t \in [T])$, where the design $\widehat{g}_t$ at time t is fitted from the available data $\bar{O}(t-1)$ up till the previous time point. An optimal adaptive design is such that the asymptotic efficiency bound under that design is the same as that under the oracle optimal design $g(Q_0)$. We then say that such an adaptive design is asymptotically equivalent with the oracle design. van der Laan [2008] demonstrates the construction and analysis of adaptive designs that are asymptotically equivalent with the oracle design.

Best arm or best treatment rule identification. Considering a finite collection of treatment arms $\mathcal{A} := \{1, \ldots, K\}$ or a finite collection of treatment rules $g_1, \ldots, g_J$, a potential learning goal is to identify as quickly as possible, under a certain fixed confidence level the one that would yield the highest immediate, cumulative, or final outcome. An ideal design is one that, when combined with appropriate estimators of the value of each arm or of each stochastic treatment rule, allows to identify the best one as quickly as possible. There is abundant literature on the best arm identification problem in the bandit setting [Even-Dar et al., 2006, Gabillon et al., 2011, 2012, Kaufmann et al., 2016].

Regret minimization goals.

Nonparametric policy class. For essentially the same reasons as invoked in the previous subsection, an ideal sequential decision making procedure should learn policies in a nonparametric policy class, or perform some form of ensemble learning over a collection of nonparametric policy classes.

Minimax optimality. An ideal algorithm would be minimax optimal in regret w.r.t. its policy class. A minimax optimal algorithm achieves regret rate equal, up to $\log T$ factors, equal to the minimax regret rate over the policy class in which it learns its policies.

Instance dependent optimality. An ideal algorithm would have regret rate provably matching that of an instance dependent lower bound, that is a bound that depends on q.

Adaptivity. As it a priori unclear which nonparametric policy class contains the optimal treatment rule d_Q, an ideal regret minimization algorithm should be able to perform ensemble learning or model selection over a collection of algorithms each operating over a different policy class. As in the passive data setting, I distinguish two types of guarantees for model selection procedures: minimax adaptivity and oracle efficiency. First let me introduce the model selection setting. Consider a collection $\Pi_1, \ldots, \Pi_J$ of policy classes such that at least one contains the optimal treatment rule d_Q, and let $\widetilde{g}_1, \ldots, \widetilde{g}_J$ algorithms that achieve the minimax regret rate w.r.t. $\Pi_1, \ldots, \Pi_J$, respectively. Consider a model selection procedure, that at each t, computes from $\bar{O}(t-1)$ the index $\widehat{j}_t$ of one of the algorithms, and let $\widetilde{g}$ be the algorithm that assigns at t the action proposed by algorithm $\widetilde{g}$ Note that the policy that any of the base algorithms propose at t depends on how the algorithms share data. They can operate separately, in which case each algorithms uses only the data at the rounds it was chosen, or on the contrary share between them all of the available data. I discuss this in more detail in chapter *[model selection chapter]*

Minimax adaptivity. I say that the model selection procedure is said minimax adaptive if it achieves regret rate equal (up to log factors) to that of the fastest minimax rate among policies classes that contain d_Q.

Oracle efficiency. The model selection procedure is oracle efficient if, for any $\epsilon > 0$,

$$E_{\widetilde{g},q}\left[\sum_{t=1}^{T} -Y(t)\right] \leq (1+\epsilon) \min_{j \in [J]} E_{\widetilde{g}_j,q}\left[\sum_{t=1}^{T} -Y(t)\right] + R(\epsilon, q, T)$$

where $R(\epsilon, q, T)$ is an excess risk term that is negligible in front of the first term of the right-hand side as $T \to \infty$.

1.5 Contributions

Although there still do not exist algorithms and estimators that provably meet all of the requirements presented in the previous section, my contributions bring the state of the art closer to meeting some of these. Here is a quick summary of the different chapters of my dissertation, and a description of how each of them contributes in the directions outlined above.

Chapter 1: Regularized Targeted Maximum Likelihood Estimation for Off-Policy Evaluation in Reinforcement Learning. In this chapter, we consider the Off-Policy Evaluation problem in reinforcement learning. While we worked under the assumption that the data were i.i.d. trajectories of a Markov Decision Process, we worked under the larger model that allows the state and reward at one time point to depend on the entire past, that is we worked under the DTR model. The parameter of interest is $\Psi(P) = E_{q,g^*}\left[\sum_{t=1}^{T} Y(t)\right]$, that is the cumulative reward under the G-computation formula distribution $P_{g^*}^T$, which equals the mean outcome under counterfactual policy g^*. Theoretical contributions include the derivation of a representation of the EIF of Ψ w.r.t. the DTR model, and the derivation of a Targeted Maximum Likelihood Estimator for Ψ based on this EIF. We came up with several regularizations of the TML estimator so as reduce variance of the estimator, as the cost of added bias. We combined regularized estimators with a bootstrap version of the so-called MAGIC ensemble learning procedure (cite Thomas and Brunskill). While the unregularized estimator and the ensemble estimator are efficient w.r.t. a larger model than the MDP model we know to contain the data generating distribution, our methodology significantly outperformed the state of the art at the time in experiments (at the time, semiparametric estimators of Ψ in the RL literature were not yet efficient w.r.t. the MDP model, but rather were efficient w.r.t. the DTR model).

Chapter 2: Fast rates for empirical risk minimizers over cadlag functions with bounded Hardy-Krause variation. The class of d-variate cadlag functions with bounded sectional variation norm has received attention from researchers recently as it is a nonparametric class which can be used as a realistic statistical model in many settings and over which the rate of convergence of empirical risk minimizers can be shown to have mild dependence on the dimension. In this chapter, we give the first characterization of this class of functions. This allows us to show that under common losses used in regression settings, the rate of convergence of empirical risk minimizers over this class is $O(n^{-1/3}(\log n)^{2d-1})$. These guarantees hold for i.i.d. data. We extend them to dependent data in subsequent chapters.

Chapter 3: Generalized Policy Elimination: an efficient algorithm for nonparametric contextual bandits. In this chapter, we make progress towards making contextual bandit algorithms available for nonparametric policy classes, in the non-realizable case. This paper is a generalization of the papers Dudik et al. [2011] and also to some extent of Agarwal et al. [2014] to nonparametric policy classes. The algorithm we propose achieves the minimax regret up to log factors over policies classes with integrable sup norm entropy, and is efficient in the sense that it requires only a polynomial number of calls to some optimization oracles. It is the first such efficient algorithm that achieves regret optimality for what we referred to as "actual" nonparametric classes, that is classes with polynomial entropy, as opposed to classes with logarithmic entropy, such as VC classes.

Chapter 4: Nonparametric learning from sequentially collected data. In this chapter, we consider nonparametric learning of infinite dimensional components of q from sequentially collected data. We give high probability bounds on the excess risk of empirical risk minimizers fitted from such data. We then propose an extended version of the Super Learner for sequential cross-validation, which eliminates the need for some hard-to-check assumptions from the original sequential Super Learner article [Benkeser et al., 2018].

Chapter 5: Model selection for contextual bandits. In this paper, we present a method for model selection in the contextual bandit setting. Our procedure achieves minimax adaptivity for the rate in T of the regret (we treat other parameter of the problems such as the number of arms or the dimension of the contexts as constants), if the base algorithms are themselves minimax optimal.

Chapter 6. Sequential causal inference in a single world of connected units. In this chapter, we consider adaptive trials involving a set of N individuals we follow along T time steps. We allow for network dependence between individuals. We work under the time series of networks model presented in subsection 1.2.4. We give inference guarantees estimators of causal effects under adaptive desings under network dependence, and guarantees for adaptive stopping rules. Theoretical contributions include maximal inequalities and equicontinuity results for empirical processes induced by dependent data under mixing conditions. As a corollary of the maximal inequality, we provide guarantees for empirical risk minimizers under mixing conditions.

Chapter 7. Sufficient and insufficient conditions for the stochastic convergence of Cesaro means. Cesaro means of random variables arise naturally in several statistical problems that have a sequential aspect. For instance, the second order remainder term in the analysis of online one-step estimators is a Cesaro mean of products of differences between nuisance estimators and their targets. We provide sufficient conditions for the stochastic convergence of such Cesaro means, and we show that convergence in probability of the terms of the mean in not in general a sufficient condition.

Bibliography

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1638–1646, Bejing, China, 22–24 Jun 2014. PMLR.

D. Benkeser, C. Ju, S. Lendle, and M. J. van der Laan. Online cross-validation-based ensemble learning. *Statistics in Medicine*, 37(2):249–260, 2018.

Deepak L. Bhatt and Cyrus Mehta. Adaptive designs for clinical trials. *New England Journal of Medicine*, 375(1):65–74, 2016. doi: 10.1056/NEJMra1510061. PMID: 27406349.

Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, 2011.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(39):1079–1105, 2006.

Victor Gabillon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Sébastien Bubeck. Multi-bandit best arm identification. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24, pages 2222–2230. Curran Associates, Inc., 2011.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, pages 3212–3220. Curran Associates, Inc., 2012.

Robert Istepanian and Turki Al-Anzi. m-Health 2.0: New perspectives on mobile health, machine learning and big data analytics. *Methods*, 151:34–40, 12 2018.

Robert Istepanian and Bryan Woodward. *M-Health: Fundamentals and Applications: Fundamentals and Applications*. John Wiley-IEEE, 2017. ISBN 9781118496985.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models, 2016.

Donna Malvey and Donna J. Slovensky. *mHealth: Transforming Healthcare*. Springer Publishing Company, Incorporated, 2014. ISBN 1489974563, 9781489974563.

Donna Spruijt-Metz and Wendy Nilsen. Dynamic models of behavior for just-in-time adaptive interventions. *IEEE Pervasive Computing*, 13(3):13–17, 2014.

Steven R. Steinhubl, Evan D. Muse, and Eric J. Topol. Can Mobile Health Technologies Transform Health Care? *JAMA*, 310(22):2395–2396, 12 2013. ISSN 0098-7484. doi: 10.1001/jama.2013.281078.

Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer, 2017.

Mark van der Laan. The construction and analysis of adaptive group sequential designs. 2008.

Mark J van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

Mark J van der Laan and Sherri Rose. *Targeted learning in data science*. Springer, 2018.

Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. 2006.

Chapter 2

More efficient off-policy evaluation through regularized targeted learning

Aurélien Bibaut, Ivana Malenica, Nikos Vlassis,
Mark van der Laan

In this chapter, we study the problem of Off-Policy Evaluation (OPE) in Reinforcement Learning. The goal of OPE is to estimate, given data collected under a certain known policy, the value of a counterfactual policy, that is the mean cumulative (discounted) reward one would obtain under that counterfactual policy.

In the Reinforcement Learning literature, it is customary to assume that the trajectories of the system are observations of a Markov Decision Process (see subsection 1.2.3 in the introduction chapter). In this chapter, we consider the efficient influence function and efficient estimators of the value of a policy with respect to a larger model, namely the Dynamic Treatment Model presented in subsection 1.2.2. We give a representation of the EIF and of these estimators under the assumption that the data-generating distribution belongs to a MDP model. Whereas such representations had been given before our article was published [Jiang and Li, 2015, Thomas and Brunskill, 2016], our article was the first one to give a rigorous derivation of the representation of the EIF of this parameter w.r.t. the DTR model, under the MDP model assumptions. A concurrent [van der Laan et al., 2018] and a subsequent article [Kallus and Uehara, 2019] give the EIF w.r.t. the MDP model.

The other main contribution of this work is to propose a Targeted Maximum Likelihood Estimator (TMLE) of the value of counterfactual policy, regularized versions of this TMLE and an ensembling procedure to combine regularized TMLEs. The ensembling procedure is a variant of the MAGIC procedure introduced by Thomas and Brunskill [2016]. Our simulations demonstrate that our estimator dominates the state-of-the art at the time of publication in terms mean squared error, across various RL environments.

2.1 Introduction

Off-policy evaluation (OPE) is an increasingly important problem in reinforcement learning. Works on OPE address the pressing issue of evaluating the performance of a novel policy in a setting where actual enforcement might be too costly, infeasible, or even hazardous. This situation arises in many fields, including medicine, finance, advertising, and education, to name a few Murphy et al. [2001], Petersen et al. [2014], Theocharous et al. [2015], Hoiles and Van Der Schaar [2016]. The OPE problem can be treated as a counterfactual quantity estimation problem, as we inquire about the mean reward we would have accrued, had we, contrary to fact, implemented the policy π_e at the time of data-collection. Estimating and inferring such counterfactual quantities is a well studied problem in statistical causal inference, and has led to many methodological developments. One of the things we aim to do in this work is to further earlier efforts Dudik et al. [2011] in bridging the gap between the reinforcement learning and causal inference fields.

There are roughly two predominant classes of approaches to off-policy value evaluation in RL Jiang and Li [2015]. The first is the *direct method* (DM), analogous to the *G-computation* procedure in causal inference Robins et al. [1999, 2000]. The direct method first fits a model of the system's dynamics and then uses the learned fit in order to estimate the mean reward of the target policy (evaluation policy). The estimators produced by this approach usually exhibit low variance, but suffer from high bias when the model fit is misspecified or the sample size is small relative to the complexity of the function class of the model Mannor et al. [2007]. The second major avenue for off-policy value evaluation is *importance sampling* methods, also termed *inverse* *propensity score* methods in statistical causal inference Rosenbaum and Rubin [1983]. Importance sampling (IS) attempts to correct the mismatch between the distributions produced by the behavior and target policies Precup et al. [2000], Precup [2000]. IS estimators are unbiased under mild conditions, but their variance tends to be large when the evaluation and behavior policies differ significantly Farajtabar et al. [2018], and grows exponentially with the horizon, rendering them Farajtabar et al. [2018] impractical for many RL settings. A third class of estimators, *Doubly Robust* (DR) estimators, obtained by combining a DM estimator and an IS estimator, are becoming standard in OPE Farajtabar et al. [2018], Jiang and Li [2015], Thomas and Brunskill [2016]. These originate from the statistics literature Robins et al. [1994], Robins and Rotnitzky [1995], Bang and Robins [2005], van der Laan and Rubin [2006], van der Laan and Rose [2011, 2018], and were introduced in the RL literature by Dudik et al. [2011]. Combining a DM and an IS estimator under the form of a DR estimator leads to lower bias than DM alone, and lower variance than IS alone.

Our contribution to OPE in RL is multifold. First we adapt a doubly robust estimator from statistical causal inference, the Longitudinal Targeted Maximum Likelihood Estimator (LTMLE) to the OPE in RL setting. We show that our adapted estimator converges at rate $O_P(1/\sqrt{n})$ to the true policy value. Deriving the LTMLE requires us to identify a mathematical object known in semiparametric statistics as the *efficient influence function* (EIF) of the estimand (policy value). To the best of our knowledge, this article is the first one to explicitly derive the EIF of the policy value for the OPE problem in RL. Knowledge of the EIF allows us to prove that both our estimator (the LTMLE) and recently proposed DR estimators [Jiang and Li, 2015, Thomas and Brunskill, 2016] are optimal in the sense that they achieve the generalized Cramer-Rao lower bound.

Second, we introduce an idea from statistics to make better use of the data than prior OPE works [Jiang and Li, 2015, Thomas and Brunskill, 2016]. We noticed that most OPE papers, at least in theory, use sample splitting: the Q-function is fitted on a split of the data, while the DR estimator is obtained by evaluating the fitted Q-function on another split. We propose a cross-validation-based technique that allows to essentially average the Q-function over the entire sample, leading to a constant-factor gain in risk.

Finally, and most importantly for practice, we propose several regularization techniques for the LTMLE estimators, out of which some, but not all, apply to other DR estimators. Using the MAGIC ensemble method from Thomas and Brunskill [2016], we construct an estimator that combines various regularized LTMLEs. We call our estimator RLTMLE (TMLE for RL). Our experiments demonstrate that RLTMLE outperforms all considered competing off-policy methods, uniformly across multiple RL environments and levels of model misspecification.

2.2 Statistical Formulation of the Problem

2.2.1 Markov Decision Process

Consider a Markov Decision Process (MDP) defined as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, P_1, P, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, and $\gamma \in (0, 1]$ is a discount factor. A trajectory H is a succession of states S_t, actions A_t and rewards R_t, observed from $t = 1$ to the horizon $t = T$: $H = (S_1, A_1, R_1, ..., S_T, A_T, R_T)$. For all $(s, a, r, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{R} \times \mathcal{S}$, $P(s', r|s, a)$ is the probability of collecting reward r and transitioning to state s', conditional on starting in state s and taking action a, and $P_1(s)$ is the probability that the initial is s. A policy π is a sequence of conditional distributions $(\pi_1, \pi_2, ...)$ that stochastically map a state to an action: for all t, $A_t|S_t \sim \pi_t$.

Suppose we are given n i.i.d. T-step trajectories of the MDP, $D = (H_1, ..., H_n)$, collected under the behavior policy $\pi_b = (\pi_{b,1}, ..., \pi_{b,T})$. We assume all trajectories have the same initial state s_1, allowing for the data-generating mechanism to be fully characterized by (P, π_b).

2.2.2 Estimation Target

The goal of OPE is to estimate the average cumulative discounted reward we would have obtained by carrying out the target policy π_e instead of policy π_b. That is, we want to estimate the following counterfactual quantity:

$$V_1^{\pi_e}(s_1) := E_{P,\pi_e}\left[\sum_{t=1}^{T} \gamma^t R_t | S_1 = s_1\right]. \tag{2.1}$$

Consider the following common assumption from the causal inference literature.

Assumption 2.1 (Absolute continuity)**.** *For all* $s, a \in \mathcal{S} \times \mathcal{A}$*, if* $\pi_b(a|s) = 0$*, then* $\pi_e(a|s) = 0$ *too.*

Under assumption 2.1 and the Markov assumption of the MDP model, $V_1^{\pi_e}(s_1)$ can be written as an expectation under the data-generating mechanism (P, π_b):

$$V_1^{\pi_e}(s_1) = E_{P,\pi_b}\left[\prod_{t=1}^{T} \frac{\pi_{e,t}(A_t|S_t)}{\pi_{b,t}(A_t|S_t)} \sum_{t=1}^{T} \gamma^t R_t \middle| S_1 = s_1\right]. \tag{2.2}$$

For $t = 1, ..., T$, define $\bar{R}_{t:T} := \sum_{\tau=t}^{T} \gamma^{\tau-t} R_\tau$ as the total reward from step t to step T. For all $1 \leq t_1 \leq t_2 \leq T$, define $\rho_{t_1:t_2} := \prod_{\tau=t_1}^{t_2} \pi_{e,\tau}(A_\tau|S_\tau)/\pi_{b,\tau}(A_\tau|S_\tau)$. For all $t = 1, ..., T$, we will use the shortcut notation $\rho_t := \rho_{1:t}$. We use the convention that $\rho_0 = 0$. Denote $\bar{R}_{t:T}^{(i)}$, $\rho_t^{(i)}$, $\rho_{t_1:t_2}^{(i)}$ the corresponding quantities for a sample trajectory H_i. Consistently with (2.1) and (2.2), we define, for any $t = 1, ..., T$, and $s \in \mathcal{S}$, the value function (or reward-to-go) from time point t and state s, as

$$\begin{aligned} V_t^{\pi_e}(s) :&= E_{P,\pi_e}[\bar{R}_{t:T}|S_t = s] \\ &= E_{P,\pi_b}\left[\rho_{t:T}\bar{R}_{t:T}|S_t = s\right]. \end{aligned}$$

For every $t = 1, ..., T$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, we further define the action-value function from time step t as

$$\begin{aligned} Q_t^{\pi_e}(s, a) &:= E_{P,\pi_e}\left[\bar{R}_{t:T}|S_t = s, A_t = a\right] \\ &= E_{P,\pi_b}\left[\rho_{t:T}\bar{R}_{t:T}|S_t = s, A_t = a\right]. \end{aligned}$$

2.3 An existing state-of-the art approach

Our method can be seen as building upon and improving on Thomas and Brunskill [2016]. We believe it helps understanding our contribution to first briefly describe their estimators. For a detailed review of OPE methods, we refer the interested reader to the vast and excellent literature on the topic Precup et al. [2000], Thomas [2015], Jiang and Li [2015], Farajtabar et al. [2018].

2.3.1 Weighted Doubly Robust Estimator

Jiang and Li [2015] were the first authors to propose a doubly robust estimator for off-policy evaluation in the MDP setting. Thomas and Brunskill [2016] propose a stabilized version of the DR estimator of Jiang and Li [2015], termed Weighted Doubly Robust (WDR) estimator, which they obtain by replacing the importance sampling weights by stabilized importance sampling weights. The stabilized importance sampling weight for observation i at time step t is defined as $w_t^{(i)} = \rho_t^{(i)} / \sum_{i=1}^n \rho_t^{(i)}$. The WDR estimator is thus defined as

$$
\begin{aligned}
WDR := \sum_{i=1}^{n} \Bigg\{ & \frac{1}{n} V_1^{\pi_e}(S_1^{(i)}) \\
& + \sum_{t=1}^{T} \gamma^t w_t^{(i)} \left[R_t^{(i)} - Q_t^{\pi_e}(S_t^{(i)}, A_t^{(i)}) + \gamma V_{t+1}^{\pi_e}(S_{t+1}^{(i)}) \right] \Bigg\}.
\end{aligned}
\tag{2.3}
$$

2.3.2 MAGIC

While WDR has low bias and converges at rate $O_P(1/\sqrt{n})$ to the truth, its reliance on importance weights can make it highly variable. As a result, in some settings, especially if model misspecification is not too strong, DM estimators can beat WDR Thomas and Brunskill [2016]. This motivates the construction of an estimator that interpolates between DM and WDR, so as to benefit from the best of both worlds. Thomas and Brunskill [2016] propose the *partial importance sampling* estimators, which correspond to essentially cutting off the sum in (2.3) the terms with index $t \geq j$ for some $0 \leq j \leq T$. Formally, they define their partial importance sampling estimator as the average $g_j := \sum_{i=1}^n g_j^{(i)}$ of the so-called *off-policy j-step return*, that they define, for each trajectory i, as

$$
\begin{aligned}
g_i^{(j)} := & \sum_{t=1}^{j} \underbrace{\gamma^t w_t^i R_t^{(i)}}_{\text{a}} + \underbrace{\gamma^{j+1} w_j^i V_{j+1}^{\pi_e}(S_{j+1}^i)}_{\text{b}} \\
& - \sum_{t=1}^{j} \underbrace{\gamma^t [w_t^i Q_t^{\pi_e}(S_t^{(i)}, A_t^{(i)}) - w_{t-1}^i V_t^{\pi_e}(S_t^{(i)})]}_{\text{c}},
\end{aligned}
$$

Note that g_0 is equal to the DM estimator. Note that the last component, (c), represents the combined control variate for the importance sampling (a) and model based term (b). Hence, as j increases, we expect bias to decrease, at the expense of an increase in variance.

Thomas and Brunskill [2016]'s final estimator is a convex combination of the partial importance sampling estimators g_j. Ideally, we would like this convex combination to minimize mean squared error (MSE), that is we would like to use as estimator $(\mathbf{x}^*)^\top \boldsymbol{g}$, with $\boldsymbol{g} = (g_0, ..., g_T)$, where

$$\begin{aligned}\boldsymbol{x}^* &= \arg \min_{\substack{0 \leq \boldsymbol{x} \leq 1 \\ \sum_{j=0}^T x_j = 1}} \quad \text{MSE}(\boldsymbol{x}^\top \boldsymbol{g}, V_1^{\pi_e}) \\ &= \arg \min_{\substack{0 \leq \boldsymbol{x} \leq 1 \\ \sum_{j=0}^T x_j = 1}} \quad \Big\{ \text{Bias}^2(\boldsymbol{x}^\top \boldsymbol{g}, V_1^{\pi_e}) \\ &\qquad\qquad + \text{Var}(\boldsymbol{x}^\top \boldsymbol{g}) \Big\}.\end{aligned}$$

As we do not have access to the true variance and bias, Thomas and Brunskill [2016] propose to use as estimator $\hat{\boldsymbol{x}}^\top \boldsymbol{g}$, where $\hat{\boldsymbol{x}}$ is a minimizer, over the convex weights simplex, of an estimate of the MSE. The covariance matrix of $\boldsymbol{g}$, which we will denote $\boldsymbol{\Omega}_n$, can be estimated as the empirical covariance matrix $\hat{\boldsymbol{\Omega}}_n$ of the $\boldsymbol{g}^{(i)}$'s. Bias estimation is a more involved. For each $j = 1, ..., T$, Thomas and Brunskill [2016] estimate the bias of the partial importance sampling estimator g_j by its distance to a δ-confidence interval for g_T obtained by bootstrapping it, for some $\delta \in (0, 1)$. They named the resulting ensemble estimator MAGIC, standing for *model and guided importance sampling combining*. For further details, we refer the reader to the very clear presentation of their algorithm by Thomas and Brunskill [2016].

2.4 Longitudinal TMLE for MDPs

2.4.1 High level description

Our proposed estimator extends the longitudinal Targeted Maximum Likelihood Estimation methodology, initially developed in the statistics causal inference literature, to the MDP setting [van der Laan and Rubin, 2006, van der Laan and Gruber, 2011, van der Laan and Rose, 2011, 2018]. In order to build intuition on our estimator, we start with a high-level description. Targeted Maximum Likelihood Estimation is a general framework that allows to construct efficient nonparametric estimators of low-dimensional characteristics of the data-generating distribution, given machine learning based estimators of high-dimensional characteristics. Let us illustrate on an example what these low-dimensional and high-dimensional characteristics can be. Suppose we want to estimate an average treatment effect (ATE), and that we have pre-treatment covariates X, a treatment T and an outcome Y, with $(X, T, Y) \sim P$. In this situation, the low-dimensional characteristic is the ATE $E_P[E_P[Y|T = 1, X] - E_P[Y|T = 0, X]]$, while the high-dimensional characteristics of P are the outcome regression function $x, a \mapsto E_P[Y|A = a, X = x]$ and the propensity score function $x \mapsto E_P[T|X = x]$.

2.4.2 Simplified sample-splitting based algorithm

In the following sections we present a simplified version of the algorithm that constructs our Longitudinal Targeted Maximum Likelihood Estimator. The full-blown version of the algorithm is presented in the appendix, with the corresponding theoretical justifications.

Suppose we are provided with n i.i.d. trajectories, $D = (H_1, ..., H_n)$. Make two splits of the sample: for some $0 < p < 1$, let $D^{(0)} = (H_1, ..., H_{(1-p)n})$ and $D^{(1)} = (H_{(1-p)n+1}, ..., H_n)$. Use $D^{(0)}$ to fit estimators $\hat{Q}_1^{\pi_e}, \cdots, \hat{Q}_T^{\pi_e}$ of the action value functions $Q_1^{\pi_e}, \cdots, Q_T^{\pi_e}$ We will call $\hat{Q}_1^{\pi_e}, \cdots, \hat{Q}_T^{\pi_e}$ the *initial estimators*. Such estimators can be obtained for instance by fitting a model of the dynamics of the MDP, or by SARSA, among other methods Sutton and Barto [1998]. Estimators fitted in such a way tend to exhibit low variance but often suffer from misspecification bias. As mentioned in section 2.3, doubly-robust estimators take such initial estimators as input, and evaluate on $D^{(1)}$ and then average a certain function of them to produce an unbiased estimator of $V_1^{\pi_e}(s_1)$. These doubly-robust estimators rely on the addition of terms weighted by the importance sampling (IS) ratios $\rho_{i:t}^{(i)}, i = 1, \cdots, n, t = 1, \cdots, n$. The TMLE methodology takes another route: for each t, it defines, on top of the initial estimator fit, a parametric model, which we will call a *second-stage parametric model* $\hat{Q}_t^{\pi_e}$, and achieves bias reduction by fitting this parametric model by maximum likelihood, on the sample split $D^{(1)}$.

2.4.3 Formal presentation of the simplified algorithm

To formally describe our algorithm, it suffices to define the second-stage parametric models and describe the loss used for the fit. For all $x \in \mathbb{R}$, we define $\sigma(x) = 1/(1 + e^{-x})$ as the logistic function, and we denote σ^{-1} its inverse. Observe that bounding the range of rewards where $\forall t, R_t \in [r_{min}, r_{max}]$, implies that $\forall t$ and $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$, $Q_t(s, a) \in [-\Delta_t, \Delta_t]$ with $\Delta_t := \sum_{\tau=t}^{T} \gamma^{\tau-t} \max(r_{max}, |r_{min}|)$. We further denote $\tilde{Q}_t^{\pi_e}(s, a) := (\hat{Q}_t^{\pi_e} + \Delta_t)/(2\Delta_t)$ as the normalized initial estimator. In addition, $\forall \delta \in (0, 1/2)$ and $\forall (s, a)$, we define the following thresholded version of $\tilde{Q}_t^{\pi_e}$:

$$\tilde{Q}_t^{\pi_e,\delta}(s, a) := \begin{cases} 1 - \delta & \text{if } \tilde{Q}_t^{\pi_e}(s, a) > 1 - \delta, \\ \tilde{Q}_t^{\pi_e}(s, a) & \text{if } \tilde{Q}_t^{\pi_e}(s, a) \in [\delta, 1 - \delta], \\ \delta & \text{if } \tilde{Q}_t^{\pi_e}(s, a) < \delta. \end{cases}$$

For all $\epsilon \in \mathbb{R}$, we can now define the normalized version of our second-stage parametric model as:

$$\tilde{Q}_t^{\pi_e,\delta}(\epsilon)(s, a) := \sigma(\sigma^{-1}(\tilde{Q}_t^{\pi_e,\delta}(s, a)) + \epsilon).$$

Finally, we denote $\hat{Q}_t^{\pi_e,\delta}(\epsilon) = 2\Delta_t(\tilde{Q}_t^{\pi_e,\delta}(\epsilon) - 1/2)$ as the rescaled version of $\tilde{Q}_t^{\pi_e,\delta}(\epsilon)$.

The normalization, thresholding and rescaling steps in the definition of the parametric second-stage model ensure that (1) $\tilde{Q}_t^{\pi_e,\delta}(\epsilon) \in [\delta, 1 - \delta] \subset (0, 1)$ for all ϵ, and that (2) $\hat{Q}_t^{\pi_e,\delta}(\epsilon)$ always stays in the allowed range of rewards $[-\Delta_t, \Delta_t]$. The definition of $\tilde{Q}_t^{\pi_e,\delta}(\epsilon)$ as a logistic transform of ϵ that lies in $(0, 1)$ makes the fitting of ϵ possible through maximum likelihood for a logistic likelihood. For $t = T$, since $Q_T^{\pi_e}(s, a) = E_{P,\pi_b}[\rho_{1:T} R_T | S_T = s, A_T = a]$, it is natural to consider the log likelihood,

$$\begin{aligned}\mathcal{R}_{n,T}^{\delta}(\epsilon) = \frac{1}{n}\sum_{i=1}^{n} \rho_{1:T}^{(i)} \Bigg(& \tilde{U}_T^{(i)} \log(\tilde{Q}_T^{\pi_e,\delta}(\epsilon)(S_T^{(i)}, A_T^{(i)})) \\ & + (1 - \tilde{U}_T^{(i)}) \log(1 - \tilde{Q}_T^{\pi_e,\delta}(\epsilon)(S_T^{(i)}, A_T^{(i)}))\Bigg),\end{aligned} \tag{2.4}$$

where $\tilde{U}_T^{(i)} := (R_T^{(i)} + \Delta_T)/(2\Delta_T)$ is the normalized reward at time T. Normalization of the reward is necessary since we are using logistic regression to optimize ϵ, and to keep the definition of $\tilde{U}_T^{(i)}$ and $\tilde{Q}_T^{\pi_e,\delta}(s, a)$ consistent. The thresholding step that defines $\tilde{Q}_t^{\delta}(s, a)$ prevents the log likelihood from taking on non-finite values. In order to make the bias introduced by thresholding vanish as the sample size grows, we use a vanishing sequence $\delta_n \downarrow 0$ of thresholding values.

Let $\epsilon_{n,T}$ be the minimizer over $\mathbb{R}$ of the log likelihood $\mathcal{R}_{n,t}^{\delta}$ for step T. We fit the second-stage models for $t = T - 1, ..., 1$ by backward recursion, a procedure which we describe in more detail in this paragraph. Start with observing that for all $t = 1, ..., T$, and for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, $Q_t^{\pi_e}(s, a) = E_{\pi_b}[\rho_{1:t}(R_t + \gamma V_{t+1}^{\pi_e}(S_{t+1}))| S_t = s, A_t = a]$. This motivates defining, as outcome of the rescaled logistic regression model for time step t, the normalized reward-to-go:

$$\tilde{U}_{t,n}^{(i)} := (R_t^{(i)} + \gamma \hat{V}_{t+1}^{\pi_e}(\epsilon_{n,t+1})(S_{t+1}^{(i)}) + \Delta_t)/(2\Delta_t).$$

Define $\hat{V}_t^{\pi_e}(\epsilon)$ as the value function corresponding to the action-value function $\hat{Q}_t^{\pi_e,\delta_n}(\epsilon)$, that is, for all $s \in \mathcal{S}$, set $\hat{V}_t^{\pi_e}(\epsilon)(s) = \sum_{a' \in \mathcal{A}} \pi_e(a'|s) \hat{Q}^{\pi_e,\delta_n}(\epsilon)(s, a')$. We define the second-stage model log likelihood for each $t = T - 1, ..., 1$ as

$$\begin{aligned}\mathcal{R}_{t,n}^{\delta}(\epsilon) = \frac{1}{n}\sum_{i=1}^{n} \rho_{1:t}^{(i)} \Bigg(& \tilde{U}_t^{(i)} \log(\tilde{Q}_t^{\pi_e,\delta}(\epsilon)(S_t^{(i)}, A_t^{(i)})) \\ & + (1 - \tilde{U}_t^{(i)}) \log(1 - \tilde{Q}_t^{\pi_e,\delta}(\epsilon)(S_t^{(i)}, A_t^{(i)}))\Bigg).\end{aligned} \tag{2.5}$$

The fact that the outcome in the second-stage logistic model at time step t depends on the second-stage model fit at time step $t + 1$ is why we have to proceed backwards in time. This is why we say this procedure is a *backward recursion*.

Finally, once all of the T second-stage models have been fitted, we define the LTMLE estimator of $V_1^{\pi_e}(s_1)$ as follows:

$$\hat{V}_1^{\pi_e,LTMLE}(s_1) := \hat{V}_1^{\pi_e}(\epsilon_{n,1})(s_1).$$

This idea of backward recursion we just exposed was initially introduced in Bang and Robins [2005]. They called it *sequential regression*.

We present the pseudo-code of the procedure as Algorithm 2.1.

Algorithm 2.1 Longitudinal TMLE for MDPs

```
**Input:** Logged data split $D^{(1)}$, target policy $\pi_e$, initial estimators $\hat{Q}_1^{\pi_e}, ..., \hat{Q}_T^{\pi_e}$, discount factor $\gamma$.
Set $\Delta_T = 0$ and $\hat{V}_{T+1}^{\pi_e} = \mathbf{0}$.
**for** $t = T$ **to** $1$ **do**
    Set $\Delta_t = \max_{t,i} |R_t| + \gamma \Delta_t$.
    Set $\tilde{U}_t = (R_t + \gamma \hat{V}_{t+1}^{\pi_e} + \Delta_t)/2\Delta_t$.
    Set $\tilde{Q}_t^{\pi_e,\delta_n} =$ threshold$(\delta_n, (\hat{Q}_t^{\pi_e} + \Delta_t)/2\Delta_t)$.
    Compute $\epsilon_{n,t} = \arg\min_\epsilon \mathcal{R}_{n,t}^{\delta_n}(\epsilon)$.
    Set $\hat{Q}_t^{\pi_e,\delta_n} = 2\Delta_t(\tilde{Q}_t^{\pi_e,\delta_n} - 0.5)$.
    Set, for all $s \in \mathcal{S}$,
        $$\hat{V}_t^{\pi_e}(s) = \sum_{a' \in \mathcal{A}} \pi_e(a'|s) \hat{Q}_t^{\pi_e,\delta_n}(s, a').$$
**end for**
**return** $\hat{V}_1^{\pi_e}(\epsilon_{n,1})(s_1)$.
```

2.4.4 Guarantees and benefits

It might at first appear surprising that fitting the second-stage models, which amounts to simply fitting the intercept of a logistic regression model, suffices to fully remove the bias. We nevertheless prove that it does so in theorem 2.1 under mild assumptions. Theorem 2.1 requires assumption 2.1 stated in section 2.2 and assumptions 2-4 stated below.

Assumption 2.2. *For all $t = 1, ..., T$, $r_t \in [r_{min}, r_{max}]$ almost surely.*

Assumption 2.3. *For all $t = 1, ..., T$, the initial estimator $\hat{Q}_{t,n}^{\pi_e}$ converges in probability to some limit $Q_{t,\infty} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, that is $\|\hat{Q}_{t,n}^{\pi_e} - Q_{t,\infty}\|_{P,2} = o_P(1)$.*

Assumption 2.4. *For all $t = 1, ..., T$, let $Q_{t,\infty}$ be the limit as defined in Assumption 2.3. Assume there exists a (small) positive constant $\eta \in (0, 1/2)$ such that $\forall t$ and $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$, $Q_{t,\infty}(s, a) \in [\eta, 1 - \eta]$.*

Assumption 2.5. *Suppose there exists a finite positive constant M such that $\forall t$, $\rho_{1:t} \leq M$ almost surely.*

We can now state our main theoretical result, for the algorithm presented in section 2.4.3.

Theorem 2.1. *Suppose assumptions 2.2, 2.3, 2.4, and 2.5 hold. Then the LTMLE estimator has bias $o(1/\sqrt{n})$, that is*

$$E_{P,\pi_b}[\hat{V}_1^{\pi_e,LTMLE}(s_1)] - V_1^{\pi_e}(s_1) = o(1/\sqrt{n}).$$

In addition, the LTMLE estimator converges in probability at rate $\sqrt{n}$, that is

$$\hat{V}_1^{\pi_e,LTMLE}(s_1) - V_1^{\pi_e}(s_1) = O_P(1/\sqrt{n}).$$

With a little extra work, we can also characterize the asymptotic distribution and the asymptotic variance of the LTMLE estimator. In particular, we show in the appendix that, provided that $\hat{Q}^{\pi_e}$ is consistent, our estimator attains the generalized Cramer-Rao bound and is therefore *locally efficient*. We also argue that it is asymptotically equivalent with the doubly robust estimator Thomas and Brunskill [2016], Jiang and Li [2015].

2.5 RLTMLE

In this section, we (1) present regularizations that can be applied to the LTMLE estimator, and (2) describe our "final estimator", which we call RLTMLE (standing for *LTMLE for RL*), and which consists of a convex combination of regularized LTMLE estimators. The weights in the RLTMLE convex combination are obtained following a variant of the ensembling procedure of the MAGIC estimator.

2.5.1 Regularization and base estimators

We present three regularization techniques that allow to stabilize the variance of the LTMLE estimator. The first two have a clear WDR analogue, while the third one only applies to LTMLE.

1. **Weights softening.** For $\alpha \in [0,1]$, $x \in \mathbb{R}^d$, define $\texttt{soften}(x,\alpha) := (x_k^\alpha / \sum_{l=1}^d x_l^\alpha : k = 1,...,d)$. The LTMLE algorithm corresponding to softening level α is obtained by replacing, in the second-stage log likelihoods (2.4) and (2.5), the IS ratios $(\rho_{1:t}^{(i)} : i = 1,...,n)$ by $\texttt{soften}((\rho_{1:t}^{(i)} : i = 1,...,n), \alpha)$. The same operation can be applied as well to the importance weights of the WDR estimator.

2. **Partial horizon.** The LTMLE with partial horizon $\tau < T$ is obtained by setting to zero the coefficients $\epsilon_{n,\tau_1}, ..., \epsilon_{n,T}$ before fitting the other second-stage coefficients. This enforces that the importance sampling ratios $\rho_{1:t}$ for $t \geq j$ have no impact on the estimator. The WDR equivalent is to use the τ-step return g_τ.

3. **Penalization.** The penalized LTMLE is obtained by adding a penalty $\lambda|\epsilon_{n,t}|$ for some $\lambda \geq 0$ to the the log-likelihoods (2.4) and (2.5) of the second-stage models.

The three regularizations can be applied simultaneously. A regularized LTMLE estimator can therefore be indexed by a triple (α, τ, λ), where α, τ and λ denote the level of softening, the partial horizon, and the level of likelihood penalization.

2.5.2 Ensemble estimator

Our final estimator is an ensemble of a pool of regularized LTMLE estimators, which we will denote $g_1, ..., g_K$, that correspond to a sequence of triples $(\alpha_1, \tau_1, \lambda_1), ..., (\alpha_K, \tau_K, \lambda_K)$ of regularization levels. We set g_K to be the unregularized LTMLE, that is we set $(\alpha_K, \tau_K, \lambda_K) = (1, T, 0)$. We ensemble the regularized LTMLE estimators $g_1, ..., g_K$ by taking a convex combination of them that minimizes an estimate of MSE. The ensembling step closely follows that of the MAGIC procedure. We propose two variants of it, which we call RLTMLE 1 and RLTMLE 2, differing in how we estimate the covariance matrix $\boldsymbol{\Omega}_n$ (defined in section 2.3) of base estimators $g_1, ..., g_K$.

RLTMLE 1. In this variant of RLTMLE, covariance estimation relies on the following property of the LTMLE estimator. As we show in the appendix, the difference between a regularized LTMLE estimator with regularization parameters (α, τ, λ), and its asymptotic limit is given by $n^{-1} \sum_{i=1}^n \text{EIF}(\hat{\boldsymbol{Q}}, \alpha, \tau, \lambda)(H_i) + o_P(n^{-1/2})$, where EIF is the efficient influence function, presented in the appendix, whose expression is given by

$$\begin{aligned}&\text{EIF}(\hat{\boldsymbol{Q}}^{\pi_e}, \alpha, \lambda, \tau)(h)\\ &= \sum_{t=1}^{T} \gamma^t \rho_t \times \big(r_t + \gamma \hat{V}_{t+1}^{\pi_e}(\epsilon_{n,t+1})(s_{t+1})\\ &- \hat{Q}_t^{\pi_e}(\epsilon_{n,t})(s_t, a_t)\big),\end{aligned}$$

where, for all t, $\epsilon_{n,t}$ is the maximizer of the regularized version of the log-likelihood (2.5), that is expression (2.5) where ρ_t is replaced with soften(ρ_t, α) and penalized by $\lambda|\epsilon|$. Denote $\text{EIF}_k(h) = \text{EIF}(\hat{\boldsymbol{Q}}, \alpha_k, \lambda_k, \tau_k)(h)$, the EIF corresponding to estimator g_k. We use as estimate of the covariance matrix $\boldsymbol{\Omega}_n$ the empirical covariance matrix $\hat{\boldsymbol{\Omega}}_n$ of $(\text{EIF}_1(H), ..., \text{EIF}_K(H))$.

RLTMLE 2. In this variant of RLTMLE, an estimate of the covariance matrix $\boldsymbol{\Omega}_n$ of the base estimators $\boldsymbol{g} = (g_1, ..., g_K)$ is obtained by computing bootstrapped values $\boldsymbol{g}^{(1)}, ..., \boldsymbol{g}^{(B)}$, of $\boldsymbol{g}$, for a large enough number of bootstrap samples B, and computing the empirical covariance $\hat{\boldsymbol{\Omega}}_n$ matrix of $\boldsymbol{g}^{(1)}, ..., \boldsymbol{g}^{(B)}$.

Bias estimation. We follow closely Thomas and Brunskill [2016] for bias estimation. For $k = 1, ..., K$, denote $b_{n,k}$ the bias of estimator g_K, and $\boldsymbol{b}_n := (b_{n,1}, ..., b_{n,K})$. Denote CI$(\alpha)$ the α-percentile bootstrap confidence interval for the LTMLE estimator. In both RLTMLE 1 and RLTMLE 2, for each $k = 1, ..., K$, estimate the bias $b_{n,k}$ with $\hat{b}_{n,k} := \text{dist}(g_k, \text{CI}(\alpha))$. Denote $\hat{\boldsymbol{b}}_n := (\hat{b}_{n,1}, ..., \hat{b}_{n,K})$.

Because of space limitation, we only give a pseudo-code description of RLTMLE 2, which is our most performant algorithm, as we will see in the next section.

2.6 Experiments

In this section, we demonstrate the effectiveness of RLTMLE by comparing it with other state-of-the-art methods used for OPE problem in various RL benchmark environments. We used three main domains, with detailed description of each allocated to the Appendix. We implement the same behavior and evaluation policies as in previous work Thomas and Brunskill [2016], Farajtabar et al. [2018].

Algorithm 2.2 RLTMLE 2

```
Input: Logged data split $D^{(1)}$, target policy $\pi_e$, initial estimator $\hat{\boldsymbol{Q}}^{\pi_e} := (\hat{Q}_1^{\pi_e}, ..., \hat{Q}_T^{\pi_e})$, discount factor $\gamma$, triples of regularization levels $(\alpha_1, \tau_1, \lambda_1), ..., (\alpha_K, \tau_K, \lambda_K)$, number of bootstrap samples $B$.
for $b = 1$ to $B$ do
    Sample with replacement from $D^{(1)}$ a bootstrap sample $D^{*,(b)}$.
    for $k = 1$ to $K$ do
        Compute $g_k^{(b)}$ by running algorithm 2.1 with inputs $D^{*,(b)}$, $\hat{\boldsymbol{Q}}^{\pi_e}$, $\pi_e$, $\gamma$, using regularizations levels $(\alpha_k, \tau_k, \lambda_k)$.
    end for
end for
for $k = 1$ to $K$ do
    Compute $g_k$ by running algorithm 2.1 with inputs $D^{(1)}$, $\hat{\boldsymbol{Q}}^{\pi_e}$, $\pi_e$, $\gamma$, using regularizations levels $(\alpha_k, \tau_k, \lambda_k)$.
    for $l = 1$ to $K$ do
        $\hat{\boldsymbol{\Omega}}_{k,l} \leftarrow n^{-1} \sum_{b=1}^{B} g_k^{(b)} g_l^{(b)} - \left(n^{-1} \sum_{b=1}^{B} g_k^{(b)}\right) \left(n^{-1} \sum_{b=1}^{B} g_l^{(b)}\right)$.
    end for
    $\text{CI}(\alpha) \leftarrow \left[\text{percentile}(\{g_k^{(b)} : b\}, \alpha), \text{percentile}(\{g_k^{(b)} : b\}, 1 - \alpha)\right]$.
    $\hat{b}_{n,k} \leftarrow \text{distance}(g_k, \text{CI}(\alpha))$.
end for

    $\hat{\boldsymbol{x}} \leftarrow \arg \min_{\substack{\boldsymbol{0} \leq \boldsymbol{x} \leq \boldsymbol{1} \\ \boldsymbol{x}^\top \boldsymbol{1} = 1}} \frac{1}{n} \boldsymbol{x}^\top \hat{\boldsymbol{\Omega}}_n \boldsymbol{x} + (\boldsymbol{x}^\top \hat{\boldsymbol{b}}_n)^2.$

return $\hat{\boldsymbol{x}}^\top \boldsymbol{g}$.
```

1. **ModelFail**: a partially observable, deterministic domain with $T = 3$. Here the approximate model is incorrect, even asymptotically, due to three of the four states appearing identical to the agent.

2. **ModelWin**: a stochastic MDP with $T = 10$, where the approximate model can perfectly represent the MDP.

3. **GridWorld**: a 4×4 grid used for evaluating OPE methods, with an episode ending at $T = 100$ or when a final state ($s16$) is reached.

We omit benefits of RLTMLE over IS, PDIS (per-decision IS), WIS (weighted IS), CWPDIS (consistent weighted per-decision IS) and DR (doubly robust) estimators due to the extensive empirical studies performed by Thomas and Brunskill Thomas and Brunskill [2016]. Instead, we compare our estimator to WDR and MAGIC, as they demonstrate improved performance over all simulations in benchmark RL environments considered Thomas and Brunskill [2016].

In evaluating our estimator, we also explore how various degree of model misspecification and sample size can affect the performance of considered methods. We start with small amount of

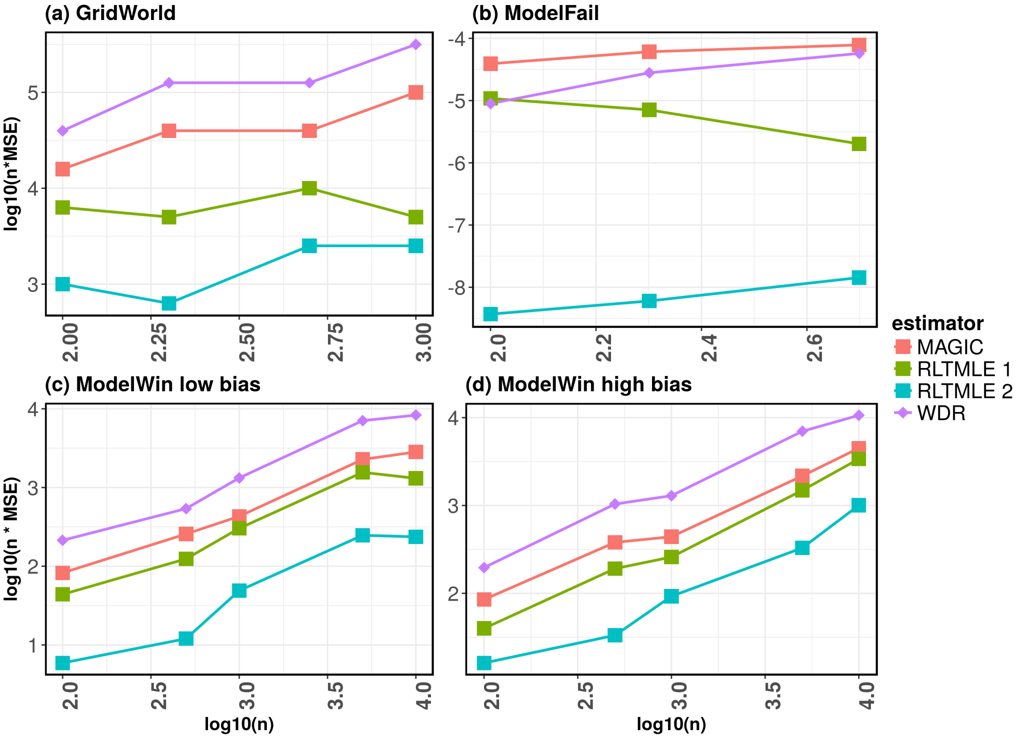


Figure 2.1: Empirical results for three different environments and varying level of model misspecification. **(a)** GridWorld MSE across varying sample size $n = (100, 200, 500, 1000)$ and bias equivalent to $b_0 = 0.005 \times \text{Normal}(0, 1)$ over 71 trials; **(b)** ModelFail MSE across varying sample size $n = (100, 200, 500, 1000)$ and bias equivalent to $b_0 = 0.005 \times \text{Normal}(0, 1)$ over 71 trials; **(c)** ModelWin MSE across varying sample size $n = (100, 500, 1000, 5000, 10000)$ and bias equivalent to $b_0 = 0.005 \times \text{Normal}(0, 1)$ over 63 trials; **(d)** ModelWin MSE across varying sample size $n = (100, 500, 1000, 5000, 10000)$ and bias equivalent to $b_0 = 0.05 \times \text{Normal}(0, 1)$ over 63 trials.

bias, $b_0 = 0.005 * \text{Normal}(0, 1)$, where most estimators should do well. Consequently, we increase model misspecification to $b_0 = 0.05 * \text{Normal}(0, 1)$ at the same sample size, and consider the performance of all estimators. In addition, we test sensitivity to the number of episodes in D with $n = \{100, 200, 500, 1000)$ for GridWorld and ModelFail, and $n = \{100, 500, 1000, 5000, 10000)$ for ModelWin.

In addition, we consider the benefits of adding few regularization techniques as opposed to all three described in subsection 2.5.1. In particular, we concentrate on RLTMLE with only weight softening and partial LTMLE (RLTMLE 1) as opposed to using penalized LTMLE as well (RLTMLE 2). The goal of these experiments was to demonstrate the improved performance of our estimator when fully exploiting all the variance reduction techniques in a clever way. The MSE across varying sample size and model misspecification for GridWorld, ModelFail and ModelWin can be found in Figure 2.1. We can see that RLTMLE 2 outperforms all other estimators for all RL environments and varying levels of model misspecification.

Finally, we compare WDR and LTMLE base estimators augmented with various regulariza-

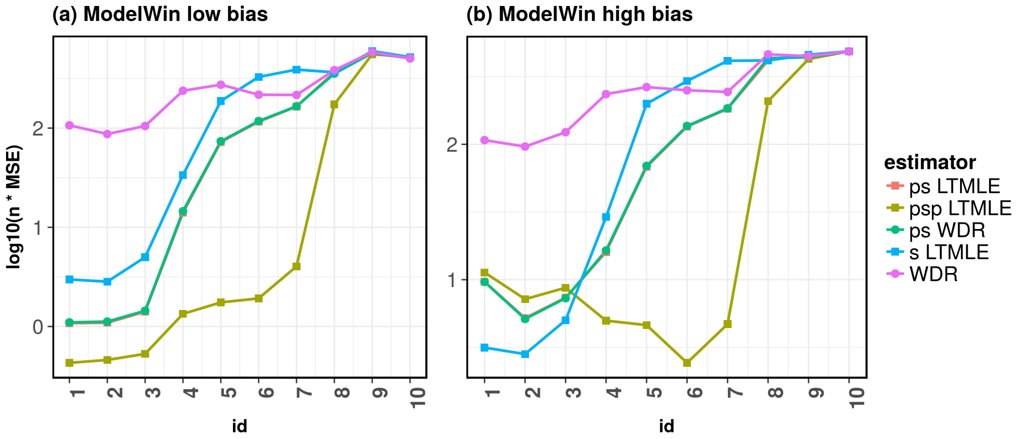


Figure 2.2: Comparison of WDR and LTMLE base estimators across various regularization methods in ModelWin at low ($b_0 = 0.005 \times \text{Normal}(0,1)$) and high ($b_0 = 0.05 \times \text{Normal}(0,1)$) model misspecification. Regularized base estimators include ps LTMLE (partial, softened LTMLE), ps WDR (partial, softened WDR), psp LTMLE (partial, softened, penalized LTMLE), s LTMLE (softened LTMLE) and WDR (no regularization). The x-axis indicates the id of the k^{th} estimator, corresponding to $(\alpha_k, \lambda_k, \tau_k)$. **(a)** ModelWin MSE for sample size $n = 1000$ and low bias over 315 trials; **(b)** ModelWin MSE for sample size $n = 1000$ and high bias over 315 trials.

tion methods before the ensemble step in Figure 2.2. In particular, for ModelWin, we look at the MSE of $\hat{V}_1^{\pi_e,j}(\epsilon_{n,1})(s_1)$ and g_k for each k, where the k^{th} estimator corresponds to regularization $(\alpha_k, \lambda_k, \tau_k)$. Regularized base estimators considered include ps LTMLE (partial, softened LTMLE), ps WDR (partial, softened WDR), psp LTMLE (partial, softened, penalized LTMLE), s LTMLE (softened LTMLE) and WDR (no regularization). We note the vast improvement of WDR just by adding weight softening across all base estimators, evident for both low and high model misspecification setting. For the low bias environment of ModelWin, psp LTMLE (RLTMLE 2) uniformly outperforms all competitors for all k. High bias setting loses to s LTMLE for low k, but still outperforms majority of the time, including having the best ensemble MSE. While uniform win over all k is not necessary, we note that this behavior stems from the fact that for $k < 3$, $(\alpha_k, \lambda_k, \tau_k)$ used had very small τ_k and α_k. As such, with no strong debiasing effect of LTMLE, minimizing variance becomes more effective with respect to minimizing MSE.

2.7 Conclusion

The contributions we make in this chapter are essentially two fold. Firstly, we derive the EIF of the value of a counterfactual policy, w.r.t. the DTR model, and we provide a representation of under the MDP model. We use this representation to derive the TMLE of the value of a counterfactual policy.

Secondly, we propose several regularized versions of this the TMLE. We combine them through a variant of the MAGIC [Thomas and Brunskill, 2016] ensemble learning procedure, in which unlike in the original version, we use the bootstrap to estimate the covariance matrix of the library of regularized estimators.

Our simulations show that the resulting estimator outperforms the existing methods at the time of publication of the original article, in particular the estimator proposed by Thomas and Brunskill [2016].

Other works [van der Laan et al., 2018], Kallus and Uehara [2019] (subsequent to the original article this chapter is based upon) investigate the efficient influence curve of the OPE target in RL w.r.t. the MDP model. Since this model is strictly contained in the DTR model, the efficiency bound is smaller. Furthermore, they show that the variance of the efficiency bound for the horizon T MDP model scales as T^{-1}, therefore allowing $\sqrt{T}$ asymptotic normality from a single trajectory. Note that the efficient influence function in the MDP model isn't double robust. van der Laan and Malenica [2018] have introduced an alternative target parameter for which it is possible to obtain robust estimators in the MDP model and $\sqrt{T}$ asymptotic normality. While this latter parameter has a clear interpretation, it might not be satisfactory for practitioners inquiring about the usual OPE target. It might be worthwhile for future work to investigate the impact on empirical performance of the loss of robustness due to working with the EIF in the MDP model as compared to using the estimators proposed in the present chapter. We expect that the comparison would certainly be nuanced, with the present estimators doing better in situation with relatively short horizons, high number of i.i.d. trajectories and high outcome model misspecification, while locally efficient estimators in the MDP model will certainly do best in the opposite situation.

We further discuss efficient estimators in the MDP model in chapter 7.

Acknowledgements

The authors would like to thank Philip Thomas and the reviewers for valuable comments.

Bibliography

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, December 2005.

Miroslav Dudik, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 1097–1104, USA, 2011. Omnipress. ISBN 978-1-4503-0619-5.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. *CoRR*, abs/1802.03493, 2018.

William Hoiles and Mihaela Van Der Schaar. Bounded off-policy evaluation with missing data for course recommendation and curriculum design. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1596–1604. JMLR.org, 2016.

Nan Jiang and Lihong Li. Doubly Robust Off-policy Value Evaluation for Reinforcement Learning. *arXiv e-prints*, art. arXiv:1511.03722, November 2015.

Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning, 2019.

Shie Mannor, Duncan Simester, Peng Sun, and John N. Tsitsiklis. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007. doi: 10.1287/mnsc.1060.0614.

S. A. Murphy, M. J. van der Laan, and J. M. Robins. Marginal Mean Models for Dynamic Regimes. *J Am Stat Assoc*, 96(456):1410–1423, December 2001.

Maya Petersen, Joshua Schwab, Susan Gruber, N Blaser, M Schomaker, and van der Laan M. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of Causal Inference*, 2(2):147–185, 2014. PMCID: PMC4405134.

Doina Precup. *Temporal abstraction in reinforcement learning*. PhD thesis, University of Massachusetts Amherst, 2000. https://scholarworks.umass.edu/dissertations/AAI9978540.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML ’00, pages 759–766, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1-55860-707-2.

J. M. Robins, M. A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, September 2000.

James M. Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995. ISSN 01621459.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. ISSN 01621459.

James M. Robins, Sander Greenland, and Fu-Chang Hu. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association*, 94(447):687–700, 1999. doi: 10.1080/01621459.1999.10474168.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41.

Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.

Georgios Theocharous, Philip S. Thomas, and Mohammad Ghavamzadeh. Personalized ad recommendation systems for life-time value optimization with guarantees. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 1806–1812. AAAI Press, 2015. ISBN 978-1-57735-738-4.

Philip Thomas. *Safe Reinforcement Learning*. PhD thesis, University of Massachusetts Amherst, 2015.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2139–2148, New York, New York, USA, June 2016. PMLR.

Mark J van der Laan and Susan Gruber. Targeted minimum loss based estimation of an intervention specific mean outcome. Technical report, U.C. Berkeley Division of Biostatistics Working Paper Series, https://biostats.bepress.com/ucbbiostat/paper290/, 2011.

Mark J. van der Laan and Ivana Malenica. Robust estimation of data-dependent causal effects based on observing a single time-series, 2018.

Mark J. van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data (Springer Series in Statistics)*. Springer, 2011.

Mark J. van der Laan and Sherri Rose. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer Science & Business Media, 2018.

Mark J. van der Laan and Daniel Rubin. Targeted maximum likelihood learning. Technical Report Working Paper 213, U.C. Berkeley Division of Biostatistics Working Paper Series, 10 2006.

Mark J. van der Laan, Antoine Chambaz, and Sam Lendle. *Online Targeted Learning for Time Series*, pages 317–346. Springer International Publishing, 2018.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes)*. Springer, 1996.

2.A Appendix organization

This appendix is organized as follows. In section 2.B, we prove theorem 2.1, which characterizes the statistical properties of the simplified algorithm presented in the main text of the article. Although we have also derived a more advanced and efficient version of the LTMLE estimator, which we introduce in section 2.D of this appendix, we choose to present the simplified version first, so as to convey the key ideas of the theoretical analysis without burdening our reader with too many technicalities.

In section B, we derive the EIF of the policy value, a necessary preliminary to establishing the semiparametric efficiency of DR estimators.

In section 2.D, we present a more advanced version of the LTMLE estimator, which makes better use of the data. This results in a constant factor speed-up of the convergence rate. This more advanced algorithm also relies on sample splitting, but fits each second stage model using the full sample, insteasd of just using a split of the full sample.

Note on notation. So as to lighten notation, we will drop the π_e superscript.

2.B Theoretical analysis of the simplified sample splitting based algorithm

In this section, we walk our reader through the theoretical analysis for the algorithm derived in section 2.4.3. We outline the steps of the proof in the proof sketch below. We then state the four main lemmas on which our proof relies, and then present the formal proof.

Proof sketch. The first fact underpinning our proof is that for any of candidate action-value $Q' = (Q'_1, ..., Q'_T)$ and corresponding value functions $V' = (V'_1, ..., V'_T)$, the difference between the candidate and the true value function at time point $t = 1$ can be decomposed as follows:

$$V'_1(s_1) - V(s_1) = -\int D(Q')(h) dP^{\pi_b}(h), \tag{2.6}$$

where $D(Q')(h) = \sum_{t=1}^{T} D_t(Q')(h)$, with $D_t(Q')(h) = \rho_{1:t}(h)(r_t + \gamma V'_{t+1}(s_{t+1}) - Q'_t(s_t, a_t))$. This is formally stated in lemma 2.1 below. For non-random functions Q' and V' note that the RHS of (2.6) is equal to $-E_{P,\pi_b}[D(Q')]$.

The second fact our proof relies on is that the estimators $\hat{Q}(\epsilon_n)$ resulting from the fitting of the parametric second stages verify the following equation:

$$\frac{1}{n} \sum_{i=1}^{n} D(\hat{Q}(\epsilon_n))(H_i) = 0. \tag{2.7}$$

This is formally stated in lemma 2.2 below. The argument in the proof of lemma 2.2 can be simply summarized as follows. For each t, $D_t(\hat{Q}(\epsilon_{n,t}))$ is the score function of the log likelihood of the second-stage logistic model for time point t.

The third fact we use in our proof is that ϵ_n converges in probability to some limit ϵ_∞. Heuristically, the reason why this is the case is that, due to the convergence of $\hat{Q}_n$ to Q_∞, the log likelihoods of the second stage models converge to a limit, which in turns implies that their arg min ϵ_n converge to the arg min of their limit. We make this rigorous in lemma 2.3 below.

Using the first two facts stated above, we obtain, by adding up equations (2.6) and (2.7), that the difference between our estimator $\hat{V}_1^{LTMLE}(\epsilon_n)(s_1)$ and the truth $V_1(s_1)$ is

$$\hat{V}_1^{LTMLE}(\epsilon_n)(s_1) - V_1(s_1)$$

$$= \frac{1}{n}\sum_{i=1}^{n} D(\hat{Q}(\epsilon_n))(H_i) - \int D(\hat{Q}(\epsilon_n))(h)dP^{\pi_b}(h).$$

Using the third fact stated above, that ϵ_n converges to some ϵ_∞, motivates rewriting the above display as

$$\begin{aligned}
&\hat{V}_1^{LTMLE}(\epsilon_n)(s_1) - V_1(s_1)\\
&= \frac{1}{n}\sum_{i=1}^{n} D(\hat{Q}(\epsilon_\infty))(H_i) - \int D(\hat{Q}(\epsilon_\infty))(h)dP^{\pi_b}(h)\\
&+ \frac{1}{n}\sum_{i=1}^{n} D(\hat{Q}(\epsilon_n))(H_i) - D(\hat{Q}(\epsilon_\infty))(H_i)\\
&- \int D(\hat{Q}(\epsilon_\infty))(h) - D(\hat{Q}(\epsilon_n))(h)dP^{\pi_b}(h). \qquad (2.8)
\end{aligned}$$

Denote $\mathcal{T}$ the sample split on which the initial estimators are fitted. Since $h \mapsto D(\hat{Q}(\epsilon_\infty))(h)$ is a non-random function conditional on $\mathcal{T}$, $\int D(\hat{Q}(\epsilon_\infty))(h)dP^{\pi_b}(h) = E_{P,\pi_b}[D(\hat{Q}(\epsilon_\infty))|\mathcal{T}]$. Therefore, applying the Central Limit theorem conditional on $\mathcal{T}$ gives us that the first line of the RHS in the above display is asymptotically normally distributed and is of order $O_P(1/\sqrt{n})$. As we will show in the formal proof, this also holds after marginilazing w.r.t. $\mathcal{T}$.

The term formed by the second and third lines in the RHS of the above display can be shown to be $o_P(1/\sqrt{n})$. This is formally stated in lemma **??** below. □

The following lemma gives a useful decomposition of the difference between any candidate state-value function V_1' and the true state-value function V_1.

Lemma 2.1 (First order expansion)**.** *Consider $Q' = (Q_1', ..., Q_T')$ a candidate vector of action-value functions $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for polict π_e, and let $V' = (V_1', ..., V_T')$ the corresponding vector of state-value functions under π_e, that is, for all t, $s \in \mathcal{S}$, $V_t'(s) = \sum_{a' \in \mathcal{A}} \pi_e(a'|s)Q_t'(s, a')$. Denote $Q = (Q_1, ..., Q_T)$ and $V = (V_1, ..., V_T)$ the true action-value and state value functions under π_e. For all t, for all $h \in \mathcal{H}$, denote $\rho_{1:t}'(h)$ an importance sampling ratio for time point t and trajectory h, not necessarily equal to the true importance sampling ratio. Denote $\rho = (\rho_1, ..., \rho_T)$ and $\rho' = (\rho_1', ..., \rho_T')$. We have that*

$$\begin{aligned}
V_1'(s_1) - V_1(s_1) = &- \int D(\rho', Q')(h)dP^{\pi_b}(h)\\
&- \int Rem(\rho, \rho', Q, Q')(h)dP^{\pi_b}(h),
\end{aligned}$$

where

$$D(\rho', Q')(h) = \sum_{t=1}^{T} D_t(\rho', Q')(h)$$

and

$$Rem(\rho, \rho', Q, Q')(h) = \sum_{t=1}^{T} Rem_t(\rho, \rho', Q, Q')(h)$$

with

$$\begin{aligned} D_t(\rho', Q')(h) = \gamma^{t-1} \rho'_{1:t}(h) \big(& r_t + \gamma V'_{t+1}(s_{t+1}) \\ & - Q'_t(s_t, a_t) \big), \end{aligned}$$

and

$$\begin{aligned} & Rem_t(\rho, \rho', Q, Q')(h) \\ & = \gamma^{t-1} \big(\rho_{1:t}(h) - \rho'_{1:t}(h) \big) \big(Q_t(s_t, a_t) - Q'_t(s_t, a_t) \\ & \qquad + (V_{t+1}(s_{t+1}) - V'_{t+1}(s_{t+1})) \big). \end{aligned}$$

From the expression in the RHS of the above display, it is immediately clear that

$$Rem_t(\rho, \rho', Q, Q')(h) = 0$$

if $\rho = \rho'$ or $Q = Q'$.

The lemma below shows that the maximum likelihood fits $\epsilon_{n,t}$ of the second-stage parametric models solve a certain equation, termed score equation in statistics.

Lemma 2.2 (Score equation)**.** *Consider the simplified LTMLE algorithm described in section 2.4.3. For each $t = 1, ..., T$, the maximum likelihood fit $\epsilon_{n,t}$ satisfies*

$$\sum_{i=1}^{n} D_t(\rho_{1:t}, \hat{Q}(\epsilon_{n,t}))(H_i) = 0.$$

The following lemma shows that the vector $\epsilon_n = (\epsilon_{n,1}, ..., \epsilon_{n,T})$ of the maximum likelihood fits of the second stage models converges in probability to a limit.

Lemma 2.3 (Convergence of ϵ_n)**.** *Make assumptions 2.2, 2.3, 2.4 and 2.5. Then, there exists $\epsilon_\infty \in \mathcal{R}^T$ such that*

$$\epsilon_n - \epsilon_\infty = o_P(1).$$

The following lemma allows to bound the last two lines of the RHS in (2.8) from the proof sketch above.

Lemma 2.4 (Equicontinuity)**.** *Denote, for all $h \in \mathcal{H}$, $\epsilon \in \mathbb{R}$, Q' and ρ'*

$$g_\epsilon(Q', \rho')(h) = D(Q'(\epsilon), \rho')(h),$$

where Q' and ρ' are possibly random. Suppose $H_1, ..., H_n$ are i.i.d. trajectories drawn from P^{π_b}. Suppose further that $H_1, ..., H_n$ are independent from the potentially random functions Q' and ρ'. Suppose $\epsilon'_n \xrightarrow{P} \epsilon'_\infty$ for some ϵ'_∞. Then

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} g_{\epsilon'_n}(Q',\rho')(H_i) - \int g_{\epsilon'_n}(Q',\rho')(h)dP^{\pi_b}(h)\\ &-\frac{1}{n}\sum_{i=1}^{n} g_{\epsilon'_\infty}(Q',\rho')(H_i) - \int g_{\epsilon'_\infty}(Q',\rho')(h)dP^{\pi_b}(h)\\ &= o_P\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

We now present the formal proof of theorem 2.1.

Proof. From lemma 2.1,

$$\hat{V}_1^{TMLE}(s_1) - V_1(s_1) = -P^{\pi_b}D(\hat{Q}_n(\epsilon_n),\rho).$$

Since from lemma 2.2 we have $P_n(D(\hat{Q}_n(\epsilon_n),\rho) = 0$, we can add this latter identity to the above display, which yields

$$\begin{aligned}&\hat{V}_1^{TMLE}(s_1) - V_1(s_1)\\ =&(P_n - P^{\pi_b})D(\hat{Q}_n(\epsilon_n),\rho)\\ =&(P_n - P^{\pi_b})D(Q_\infty(\epsilon_\infty),\rho)\\ &+(P_n - P^{\pi_b})(D(Q_\infty(\epsilon_\infty),\rho) - D(\hat{Q}_n(\epsilon_n),\rho)).\end{aligned} \tag{2.9}$$

From the Central Limit theorem applied conditionally on $\mathcal{T}$,

$$\begin{aligned}&\sqrt{n}((P_n - P^{\pi_b})D(Q_\infty(\epsilon_\infty),\rho))\\ &\xrightarrow{d}\mathcal{N}(0,\sigma^2(Q_\infty(\epsilon_\infty)),\end{aligned}$$

with

$$\sigma^2(Q_\infty(\epsilon_\infty)) := Var_{P^{\pi_b}}(D(Q_\infty(\epsilon_\infty),\rho)).$$

Using dominated convergence on the c.d.f. on the LHS,

$$\begin{aligned}&\sqrt{n}((P_n - P^{\pi_b})D(Q_\infty(\epsilon_\infty),\rho))\\ &\xrightarrow{d}\mathcal{N}(0,\sigma^2(Q_\infty(\epsilon_\infty))\end{aligned}$$

also holds true unconditionally. As proven in section B, the variance of the $D(Q_\infty(\epsilon_\infty),\rho)$ is the efficient variance from the Cramer-Rao lower bound, provided $Q_\infty(\epsilon_\infty) = Q$, that is provided the initial estimator's model is correctly specified. This is the notion of local efficiency from semiparametric statistics Robins and Rotnitzky [1995], van der Laan and Rubin [2006].

From lemmas 2.3 and 2.4, the line (2.9) is $o_P(1/\sqrt{n})$.

Therefore, we have that

$$\sqrt{n}(\hat{V}_1^{TMLE}(s_1) - V_1(s_1)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(Q_\infty(\epsilon_\infty)),$$

and that

$$E_{P^{\pi_b}}\left[\hat{V}_1^{TMLE}(s_1) - V_1(s_1)\right] = o(1/\sqrt{n}).$$

□

2.B.1 Proof of lemma 2.1

Proof. Let $H \sim P^{\pi_e}$. If Q', V' are random functions, further suppose, without loss of generality, that H is independent of Q' and V'. Denote $\mathcal{G}$ a σ-field such that Q', V' are $\mathcal{G}$-measurable.

Step 1. Observe that

$$P^{\pi_b}D(Q', \rho') = P^{\pi_b}D(Q', \rho) + P^{pi_b}(D(Q', \rho') - D(Q', \rho)).$$

Step 2: First order term. Observe that

$$P^{\pi_b}D(Q', \rho) = E_{P^{\pi_b}}[D(Q', \rho)(H)|\mathcal{G}].$$

For all $t \geq 1, ..., T$, denote $\mathcal{F}_t$ the σ-field induced by $S_1, A_1, R_1, ..., S_t, A_t, R_t$. Observe that

$$\begin{aligned}
&E_{P^{\pi_b}}[D_t(Q', \rho)(H)|S_t, A_t, \mathcal{F}_{t-1}, \mathcal{G}] \\
&= \gamma^{t-1} E_{P^{\pi_b}}[\rho_{1:t}(R_t + \gamma V'_{t+1}(S_{t+1}) \\
&\qquad - Q'_t(S_t, A_t))|S_t, A_t, \mathcal{F}_{t-1}, \mathcal{G}] \\
&= \gamma^{t-1}\rho_{1:t} E_P[R_t + \gamma V_{t+1}(S_{t+1}) \\
&\qquad - Q'_t(S_t, A_t)|S_t, A_t, \mathcal{G}] \\
&\ + \gamma^t \rho_{1:t} E_P[(V'_{t+1}(S_{t+1}) - V_{t+1}(S_{t+1}))|S_t, A_t, \mathcal{G}].
\end{aligned}$$

Recall that by definition of Q, we have that $E_P[R_t + \gamma V_{t+1}(S_{t+1})|S_t, A_t] = Q_t(S_t, A_t)$. Inserting this in the last line of the above display yields

$$\begin{aligned}
&E_{P^{\pi_b}}[D_t(Q', \rho)(H)|S_t, A_t, \mathcal{F}_{t-1}, \mathcal{G}] = \\
&\gamma^{t-1}\rho_{1:t}(Q_t(S_t, A_t) - Q'_t(S_t, A_t)) \\
&\ + \gamma^t \rho_{1:t} E_P[V'_{t+1}(S_{t+1}) - V_{t+1}(S_{t+1})|S_t, A_t, \mathcal{G}].
\end{aligned} \tag{2.10}$$

We take the expectation conditional on S_t, $\mathcal{F}_{t-1}$, $\mathcal{G}$ of the first term in the right-hand side of the above display:

$$E_{P^{\pi_b}}[\gamma^{t-1}\rho_{1:t}(Q_t(S_t, A_t) - Q'_t(S_t, A_t))|S_t, \mathcal{F}_{t-1}, \mathcal{G}]$$

$$
\begin{aligned}
&= \gamma^{t-1}\rho_{1:t-1}E_{P,\pi_b}[\rho_t(Q_t(S_t,A_t) - Q'_t(S_t,A_t)|S_t,\mathcal{G}]\\
&= \gamma^{t-1}\rho_{1:t-1}E_{P,\pi_e}[(Q_t(S_t,A_t) - Q'_t(S_t,A_t)|S_t,\mathcal{G}]\\
&= \gamma^{t-1}\rho_{1:t-1}(V_t(S_t) - V'_t(S_t)). \qquad (2.11)
\end{aligned}
$$

The second equality above uses that, for all $\mathcal{G}$-measurable function f,

$$E_{P,\pi_b}[\rho_t f(S_t,A_t)|S_t,\mathcal{G}] = E_{P,\pi_e}[f(S_t,A_t)|S_t,\mathcal{G}].$$

The third equality follows from the relationship between the value function and the action value function.

Using the law of iterated expectations, and identities (2.10) and (2.11) yields

$$
\begin{aligned}
&E_{P,\pi_b}[D_t(Q',\rho)(H)|\mathcal{G}]\\
&= E_{P,\pi_b}[E_{P,\pi_b}[D_t(Q',\rho)(H)|S_t,A_t,\mathcal{F}_{t-1},\mathcal{G}]|\mathcal{G}]\\
&= E_{P,\pi_b}[\gamma^{t-1}\rho_{1:t}(Q_t(S_t,A_t) - Q'_t(S_t,A_t))|\mathcal{G}]\\
&\quad + E_{P,\pi_b}[\gamma^{t}\rho_{1:t}E_P[V'_{t+1}(S_{t+1}) - V_{t+1}(S_{t+1})|S_t,A_t,\mathcal{G}]|\mathcal{G}]\\
&= E_{P,\pi_b}[E_{P,\pi_b}[\gamma^{t-1}\rho_{1:t}(Q_t(S_t,A_t)\\
&\qquad\qquad - Q'_t(S_t,A_t))|S_t,\mathcal{F}_{t-1},\mathcal{G}]|\mathcal{G}]\\
&\quad + E_{P,\pi_b}[\gamma^{t}\rho_{1:t}(V'_{t+1}(S_{t+1}) - V_{t+1}(S_{t+1}))|\mathcal{G}]\\
&= E_{P,\pi_b}[\gamma^{t-1}\rho_{1:t-1}(V_t(S_t) - V'_t(S_t))|\mathcal{G}]\\
&\quad + E_{P,\pi_b}[\gamma^{t}\rho_{1:t}(V'_{t+1}(S_{t+1}) - V_{t+1}(S_{t+1}))|\mathcal{G}]
\end{aligned}
$$

Using the above expression in the definition of $D(Q',V')$ yields

$$
\begin{aligned}
&E_{P,\pi_b}[D(Q',\rho)(H)|\mathcal{G}]\\
=&\sum_{t=1}^{T} E_{P,\pi_b}[\gamma^{t}\rho_{1:t}(V'_{t+1}(S_{t+1}) - V'_{t+1}(S_{t+1}))\\
&\qquad\qquad - \gamma^{t-1}\rho_{1:t-1}(V'_t(S_t) - V_t(S_t))|\mathcal{G}]\\
=&E_{P,\pi_b}[\gamma^{T}\rho_{1:T+1}(V'_{T+1}(S_{T+1}) - V_{T+1}(S_{T+1}))\\
&\quad - \rho_{1:0}(V'_1(s_1) - V_1(s_1))|\mathcal{G}]\\
=&-(V'_1(s_1) - V_1(s_1)),
\end{aligned}
$$

where we have used that by convention $V'_{T+1}(S_{T+1}) = V_{T+1}(S_{T+1}) = 0$ and $\rho_{1:0} = 1$.

Step 3: remainder term. We similarly show that $P^{\pi_b}(D'(Q',\rho) - D(Q',\rho)) = Rem_t(Q,Q',\rho,\rho')$. □

2.B.2 Proof of lemma 2.2

We present a proof sketch in this subsection. The complete formal proof is presented in the case of the full algorithm in section B.

Proof sketch. The result essentially follows from the following two facts:

- The score of the logistic likelihood of the second stage model for time point t is $P_n D_t(\hat{Q}, \rho)$,
- A maximum likelihood fit solves the empirical score equation.

□

2.B.3 Proof of lemma 2.3

We present a proof sketch in this subsection. The complete formal proof is presented in the case of the full algorithm in section B.

Proof sketch. The convergence of $\hat{Q}$ to Q_∞ implies the pointwise convergence of the log likelihood risk $\mathcal{R}_{n,t}$ to some asymptotic risk $\mathcal{R}_{\infty,t}$. The fact that $Q_{\infty,t} \in [\delta, 1-\delta] \subset (0,1)$ (in other words, that $Q_{\infty,t}$ is bounded away from 0 and 1) implies that the asymptotic log likelihood risk $\mathcal{R}_{\infty,t}$ is strongly convex. This implies it has a unique minimizer $\epsilon_{\infty,t}$. We then show in the formal proof that since $\mathcal{R}_{n,t}$ are a sequence of convex functions that converge pointwise in probability to a strongly convex function minimized by $\epsilon_{\infty,t}$, the sequence of their minimizers $\epsilon_{n,t}$ converges in probability to $\epsilon_{\infty,t}$ □

2.B.4 Proof of lemma 2.4

The proof of lemma 2.4 relies on the following three technical lemmas. Recall the following definition: for all Q' ρ', $h \in \mathcal{H}$, $\epsilon \in \mathbb{R}$,

$$g_\epsilon(Q', \rho')(h) = D(Q'(\epsilon), \rho')(h).$$

Lemma 2.5. *Assume that* $0 \leq \rho'_{1:t}(H) \leq M$ *almost surely for all* $t = 1, ..., T$. *Make assumption 2.2 on the range of the rewards. Then for all* $\epsilon \in \mathbb{R}^T$,

$$\|g_\epsilon(Q', \rho')\|_{L_\infty(P^{\pi_b})} \leq 3MT,$$

and for all $\epsilon_1, \epsilon_2 \in \mathbb{R}^T$

$$\|g_{\epsilon_1}(Q', \rho') - g_{\epsilon_2}(Q', \rho')\|_{L_\infty(P^{\pi_b})} \leq 2MT\|\epsilon_1 - \epsilon_2\|_\infty.$$

For any $\epsilon_0 \in \mathbb{R}$, and any $\xi > 0$, define the class of functions

$$\begin{aligned}&\mathcal{G}(Q', \rho')(\epsilon_0, \xi)\\ &:= \{g_\epsilon(Q', \rho') - g_{\epsilon_0}(Q', \rho') : \|\epsilon - \epsilon_0\|_\infty \leq \xi\}.\end{aligned}$$

The next lemma characterizes covering numbers of this class of functions. Covering numbers are a measure of geometric complexity whose definition we recall here (we reproduce the definition 2.1.6. from van der Vaart and Wellner [1996]).

Definition 2.1 (Covering number). *The covering number $N(\epsilon, \mathcal{F}, \| \cdot \|)$ is the minimal number of balls $\{g : \|f - g\| \leq \epsilon\}$ of radius ϵ needed to cover the set $\mathcal{F}$.*

Lemma 2.6. *For any $\alpha > 0$, for any probability distribution Λ on $\mathcal{H}$,*

$$N(\alpha, L_2(\Lambda), \mathcal{G}(\epsilon_0, \xi)) \leq \left(\frac{2\xi L}{\alpha} \right)^T,$$

with $L = 2MT$.

Proof. Consider the set

$$\left\{ \left(\epsilon_{0,1} + i_1 \frac{\alpha}{L}, ..., \epsilon_{0,T} + i_T \frac{\alpha}{L} \right) \right. \\ \left. : \forall t = 1, ...T, \ i_t \in \mathbb{Z} \cap \left[-\frac{\xi L}{\alpha}, \frac{\xi L}{\alpha} \right] \right\}.$$

Observe that for any $f_\epsilon := g_\epsilon(Q', \rho') - g_{\epsilon_0}(Q', \rho') \in \mathcal{G}(Q', \rho')(\epsilon_0, \xi)$, there exists an $f_{\epsilon'} := g_{\epsilon'}(Q', \rho') - g_{\epsilon_0}(Q', \rho')$ in the set above such that $\|\epsilon - \epsilon'\|_\infty \leq \alpha/L$. From the second claim in lemma 2.5, for all $h \in \mathcal{H}$, $|f_{\epsilon'}(h) - f_\epsilon(h)| \leq \alpha$. Therefore, for any probability distribution Λ over $\mathcal{H}$,

$$\begin{aligned} \|f_{\epsilon'} - f_\epsilon\|_{L_2(\Lambda)} = & \left(\int (f_{\epsilon'}(h) - f_\epsilon(h))^2 d\Lambda(h) \right)^{1/2} \\ \leq & \alpha. \end{aligned}$$

Therefore the set defined above is an α-cover of $\mathcal{G}(\epsilon_0, \xi))$ for the norm $L_2(\Lambda)$. Since this set has at most $(2\epsilon_L/\alpha)^T$ elements, this proves that

$$N(\alpha, L_2(\Lambda), \mathcal{G}(\epsilon_0, \xi)) \leq \left(\frac{2\xi L}{\alpha} \right)^T.$$

□

The covering numbers characterized in lemma 2.6 are the basis for another measure of geometric complexity of a class of function, the uniform entropy integral, whose definition we recall below (see also van der Vaart and Wellner [1996]).

Definition 2.2 (Uniform entropy integral). *Consider a class of functions $\mathcal{X} \to \mathbb{R}$. Let $F : \mathcal{X} \to \mathbb{R}$ be an envelope function for $\mathcal{F}$, that is a function such that for all $x \in \mathcal{X}$, $|f(x)| \leq F(x)$. The uniform entropy integral of $\mathcal{F}$, w.r.t. the envelope function F and L_2 norm is defined, for all $\beta > 0$ as*

$$\begin{aligned} & J_F(\beta, \mathcal{F}, L_2) \\ := & \int_0^\beta \sup_\Lambda \sqrt{\log(1 + N(\alpha \|F\|_{\Lambda,2}, L_2(\Lambda), \mathcal{F})} d\alpha, \end{aligned}$$

where the supremum is over all discrete probability distributions on $\mathcal{X}$.

The following lemma characterizes the uniform entropy integral of $\mathcal{G}(\epsilon_0, \xi)$.

Lemma 2.7. *Let $\beta > 0$. Denote $L = 2MT$. The function $F_\xi : h \mapsto L\xi$ is an envelope function for $\mathcal{G}(\epsilon_0, \xi)$. The uniform entropy integral of $\mathcal{G}(\epsilon_0, \xi)$ w.r.t. the envelope function F_ξ and for the L_2 norm is upper bounded as follows:*

$$J_{F_\xi}(\beta, \mathcal{G}(\epsilon_0, \xi), L_2) = O\left(T\beta\sqrt{\log(1/\beta)}\right).$$

Proof. For every probability distribution Λ on $\mathcal{H}$, $\|F_\xi\|_{\Lambda,2} = L\xi$. From lemma 2.6,

$$N(\alpha\|F_\xi\|_{2,\Lambda}, L_2(\Lambda), \mathcal{G}(\epsilon_0, \xi)) \leq (2/\alpha)^T.$$

Therefore,

$$\begin{aligned} J_{F_\xi}(\beta, \mathcal{G}(\epsilon_0, \xi), L_2) \leq & \int_0^\beta \sqrt{\log(1 + (2/\alpha)^T)} d\alpha \\ =& O\left(T\beta\sqrt{\log(1/\beta)}\right), \end{aligned}$$

where the second equality above follows from an integration by parts. □

Finally, we prove the lemma 2.4. The proof relies on a classical result in empirical process theory. We first introduce the relevant definitions and the relevant result before stating the proof of our lemma.

Definition 2.3 (Empirical process and empirical process notation)**.** *Consider $\mathcal{X}, \Sigma, P')$ a probability space and let $X_1, ..., X_n$ be n i.i.d. draws from P'. Let $\mathcal{F}$ be a class of functions $\mathcal{X} \to \mathbb{R}$. For all $f \in \mathcal{F}$, define the so-called "empirical process notation"*

$$P'f := \int f(h) dP'(h).$$

Denote $P_n := n^{-1}\sum_{i=1}^n \delta_{X_i}$ the empirical probability distribution associated to the sample $X_1, ...,$ X_n. Observe that using the empirical process notation defined above, we have that $P_n f = n^{-1}\sum_{i=1}^n f(X_i)$. The stochastic process

$$\{(P_n - P')f : f \in \mathcal{F}\}$$

is termed the empirical process associated to P' and n indexed by $\mathcal{F}$.

We restate here the classical empirical process result van der Vaart and Wellner [1996] we will use to prove lemma 2.4. (This is lemma 2.14.1 in van der Vaart and Wellner [1996], for $p = 1$ in their notation.)

Lemma 2.8 (Pollard's maximal inequality, vdV-Wellner 1996 2.14.1)**.** *Consider $\mathcal{X}, \Sigma, P')$ a probability space and let $X_1, ..., X_n$ be n i.i.d. draws from P'. Let $\mathcal{F}$ be a class of functions $\mathcal{X} \to \mathbb{R}$. Let $\mathcal{F}$ be a class of functions $\mathcal{X} \to \mathbb{R}$ with envelope function F. Then*

$$E_{P'}[\sup_{f \in \mathcal{F}} \sqrt{n}|(P_n - P')f|] \lesssim J_F(1, \mathcal{F}, L_2)\|F\|_{L_2(P')}.$$

We now have all the ingredients to prove lemma 2.4.

Proof of lemma 2.4. Recasting the claim of lemma 2.1 in terms of empirical process notation, we want to show that

$$\sqrt{n}(P_n - P^{\pi_b})(g_{\epsilon_n}(Q', \rho') - g_{\epsilon_\infty}(Q', \rho')) = o_P(1).$$

Let $\kappa > 0$, $\gamma \in (0, 1/2)$. Define, for all $\xi > 0$, the following two events:

$$\mathcal{E}_1(\xi) := \{\|\epsilon_n - \epsilon_\infty\|_\infty \leq \xi\}$$

and

$$\begin{aligned}&\mathcal{E}_2(\xi) :=\\ &\Bigg\{ \sup_{\substack{\epsilon \\ \|\epsilon - \epsilon_\infty\|_\infty \leq \xi}} \sqrt{n}|(P_n - P^{\pi_b})(g_\epsilon(Q', \rho') - g_{\epsilon_\infty}(Q', \rho'))|\\ &\qquad \leq \kappa \Bigg\}.\end{aligned}$$

The function $F_\xi : h \mapsto \xi L$ is an envelope function for $\mathcal{G}(\epsilon_0, \xi)$. By Markov's inequality and lemma 2.8 applied with the uniform entropy integral bound given in lemma 2.7, we have that

$$\begin{aligned}&1 - P^{\pi_b}[\mathcal{E}_2(\xi)]\\ =&P^{\pi_b}\left[\sqrt{n}|(P_n - P^{\pi_b})(g_{\epsilon_n}(Q', \rho') - g_{\epsilon_\infty}(Q', \rho'))| \geq \kappa\right]\\ \leq&\kappa^{-1}E_{P^{\pi_b}}\left[\sqrt{n}|(P_n - P^{\pi_b})(g_{\epsilon_n}(Q', \rho') - g_{\epsilon_\infty}(Q', \rho'))|\right]\\ \leq&\kappa^{-1}J_F(1, \mathcal{G}(\epsilon_0, \xi), L_2)\|F_\xi\|_{2,\Lambda}\\ \leq&K\kappa^{-1}\xi L,\end{aligned}$$

for some constant K. Set $\xi = \kappa\gamma/(2KL)$. Then, from the above display $P^{\pi_b}[\mathcal{E}_2(\kappa\gamma/(2KL))] \geq 1 - \gamma/2$.

Besides, since $\epsilon_n \overset{P}{\to} \epsilon_\infty$, there exists n_0 such that for all $n \geq n_0$, $P^{\pi_b}[\mathcal{E}_1(\kappa\gamma/(2KL))] \geq 1 - \gamma/2$. Observe that if $\mathcal{E}_1(\kappa\gamma/(2KL)) \cap \mathcal{E}_2(\kappa\gamma/(2KL))$ is realized, then

$$\sqrt{n}|(P_n - P^{\pi_b})(g_{\epsilon_n}(Q', \rho') - g_{\epsilon_\infty}(Q', \rho'))| \leq \kappa.$$

Using a union bound, we have that, for all $n \geq n_0$,

$$\begin{aligned}&P^{\pi_b}\left[\sqrt{n}|(P_n - P^{\pi_b})(g_{\epsilon_n}(Q', \rho') - g_{\epsilon_\infty}(Q', \rho'))| \leq \kappa\right]\\ \geq&1 - (1 - P^{\pi_b}[\mathcal{E}_1(\kappa\gamma/(2KL))])\\ &- (1 - P^{\pi_b}[\mathcal{E}_2(\kappa\gamma/(2KL))])\\ \geq&1 - \gamma.\end{aligned}$$

Recapitulating the above, we have proven that for all $\kappa > 0$, $\gamma \in (0, 1/2)$, there exists n_0 such that for all $n \geq n_0$,

$$\begin{aligned}&P^{\pi_b}\left[\sqrt{n}|(P_n - P^{\pi_b})(g_{\epsilon_n}(Q', \rho') - g_{\epsilon_\infty}(Q', \rho'))| \leq \kappa\right]\\ \geq & 1 - \gamma.\end{aligned}$$

In other words, we have thus proven that

$$\sqrt{n}|(P_n - P^{\pi_b})(g_{\epsilon_n}(Q', \rho') - g_{\epsilon_\infty}(Q', \rho'))| = o_P(1).$$

which concludes the proof. □

2.C Efficiency and efficient influence function derivation

In this section we show that our estimator is optimal in a certain sense. Specifically, we show that it is *locally semiparametric efficient*. We will introduce our reader to the notions from semiparametric statistics necessary to understand and prove *semiparametric efficiency* of an estimator.

In particular we will introduce the concept of *efficient influence function* (EIF). Deriving the EIF is the cornerstone of the efficiency analysis. It is also key to the construction of the estimator: in semiparametric statistics, looking for the EIF is typically the starting point for building an efficient estimator.

Deriving the EIF in the general MDP setting is one of the main contributions of this work.

Note that the presentation of the notions of semiparametric inference is heavily drawn from **?**, and entails no novel contribution of our part. We wrote it so as to make this appendix a self-contained document for the reader non-familiar with semiparametric statistics.

2.C.1 Introducing notions of optimality from semiparametric statistics

The notions of optimality we are about to introduce are relative to both the estimand and the statistical model. The statistical model $\mathcal{M}$ is the set of probability distributions we believe to contain the true data-generating mechanism, which we will denote P_0. We will typically denote P an arbitrary element of $\mathcal{M}$. The larger the model, the more realistic it is that it contains the truth, but also the larger is the variance of estimators over this model.

The first notion of optimality we introduce is the notion of *efficiency [cite Kosorok]*. An estimator is *efficient* at P, if, were the true data-generating mechanism to be P, it would have the lowest variance among a certain class of estimators, namely the class of estimators that are *regular* at P w.r.t $\mathcal{M}$. We define formally the notion of a regular estimator at P w.r.t. $\mathcal{M}$ below. The concepts of *regularity*, *efficiency* and *semiparametric efficiency* that we are about to introduce are defined relative to P and to $\mathcal{M}$, but they really only involve $\mathcal{M}$ through its geometry in a neighborohood around P. Even more so, these notions only involve $\mathcal{M}$ through its so-called *tangent space* at P, which we will denote $T_{\mathcal{M}}(P)$.

We now proceed to stating the formal definitions. In all of this section, we will assume for simplicity that all probability distributions are dominated by the same measure μ. Definitions will be stated accordingly, but our reader should be aware that some of them can be extended to the case where there is not a single dominating measure μ.

Definition 2.4 (Statistical model)**.** *A statistical model $\mathcal{M}$ is a collection of probability distributions $\{P \in \mathcal{M}\}$ on a sample space $\mathcal{X}$.*

Usually, we suppose that the true data-generating mechanism, that we will denote P_0 in this section, belongs to the statistical model.

In the reinforcement learning setting of this article, the statistical model is a collection of probability distributions over the space of trajectories $\mathcal{H}$. Any probability distribution P over $\mathcal{H}$ can be factored as follows:

$$P = \prod_{t=1}^{T} \tilde{Q}_t \prod_{t=1}^{T} \pi_t \equiv \tilde{Q}\pi,$$

with $\tilde{Q} \equiv \prod_{t=1}^{T} \tilde{Q}_t$ and $\pi \equiv \prod_{t=1}^{T} \pi_t$, and where $\tilde{Q}_t$ is the conditional distribution of R_t, S_{t+1} given S_t, A_t and π_t is the conditional distribution of A_t given S_t. Since we know the logging policy, our statistical model supposes that for any of its elements P, π is equal to the known value of the logging policy. We therefore write our statistical model as indexed by the known value of the logging policy:

$$\begin{aligned}\mathcal{M}(\pi) = & \\ & \left\{ P = \prod_{t=1}^{T} \tilde{Q}_t \prod_{t=1}^{T} \pi_t : \forall t = 1, ..., T,\ \tilde{Q}_t \in \mathcal{M}_{Q,t} \right\} \\ = & \left\{ P = \tilde{Q}\pi : \tilde{Q} \in \mathcal{M}_Q \right\},\end{aligned}$$

with $\mathcal{M}_Q = \mathcal{M}_{Q,1} \times ... \times \mathcal{M}_{Q,T}$. We suppose that the model is fully nonparametric, that is for every t, $\mathcal{M}_{Q,t}$ is equal to the set of all conditional probability distributions $P_{R_t, S_{t+1} | S_t, A_t}$.

Definition 2.5 (Estimand / target parameter)**.** *The target parameter mapping, which we denote Ψ, is a map defined on the statistical model $\mathcal{M}$, with values either in $\mathbb{R}^d$ for a certain d, or in some function space. The estimand or target parameter is this map evaluated at the data-generating distribution: $\Psi(P_0)$.*

In the setting of this article, for every probability distribution $P = \tilde{Q}\pi$ over $\mathcal{H}$, the target parameter mapping at P is defined as $\Psi(P) = E_{P' \equiv \tilde{Q}\pi_e}[\sum_{t=1}^{T} \gamma^t R_t]$. Note that this expression doesn't depend on π, so we can write $\Psi(P) = \tilde{\Psi}(\tilde{Q})$ for a mapping $\tilde{\Psi}$ thus defined.

Definition 2.6 (Estimator)**.** *For any sample size n, an estimator $\hat{\Psi}_n$ is a mapping of the sample space $\mathcal{H}^n$ to the space of the target parameter/estimand.*

Definition 2.7 (One-dimensional submodel)**.** *A one-dimensional submodel of $\mathcal{M}$ that passes through P in 0 is a subset of $\mathcal{M}$ of the form $\{P_\epsilon : \epsilon \in [-\eta, \eta]\}$, for some $\eta > 0$, such that $P_{\epsilon=0} = P$.*

Definition 2.8 (Score of a one-dimensional parametric model). *The score in ϵ_0, for any ϵ_0, of a one-dimensional parametric model $\{P_\epsilon : \epsilon\}$ is the derivative of the log-likelihood w.r.t. ϵ, evaluated at ϵ_0. Denoting it s, the score is the function defined, for all x in the sample space $\mathcal{X}$,*

$$s(x) = \left.\frac{d \log(dP_\epsilon/d\mu)(x)}{d\epsilon}\right|_{\epsilon=\epsilon_0}.$$

Definition 2.9 (Tangent space). *The tangent space of a statistical model $\mathcal{M}$ at P, which we denote $T_{\mathcal{M}}(P)$ is the linear closure of the set of score functions of all of the one-dimensional submodels of $\mathcal{M}$ that pass through P. Formally,*

$$T_{\mathcal{M}}(P) = \overline{\mathit{Span}}(\mathcal{S}(\mathcal{M}, P)),$$

with

$$\begin{gathered}\mathcal{S}(\mathcal{M}, P) \equiv \\ \left\{ \left.\frac{d \log(dP_\epsilon/d\mu)}{d\epsilon}\right|_{\epsilon=0} : \{P_\epsilon : \epsilon\} \text{ 1-dim. submodel of } \mathcal{M}, \right. \\ \left. P_{\epsilon=0} = P \right\}.\end{gathered}$$

Definition 2.10 (Regularity). *Suppose the data-generating distribution is P, for some $P \in \mathcal{M}$. An estimator is regular at P w.r.t. $\mathcal{M}$ if, for any one dimensional submodel $\{P_\epsilon : \epsilon\}$ of $\mathcal{M}$ such that $P_{\epsilon=0} = P$, the asymptotic distribution of*

$$\sqrt{n}(\hat{\Psi}_n - \Psi(P_{\epsilon=1/\sqrt{n}}))$$

is the same as the asymptotic distribution of

$$\sqrt{n}(\hat{\Psi}_n - \Psi(P)).$$

It is the understanding of the authors of this article that non-regular estimators at P correspond either to pathological estimators or pathological P's.

Definition 2.11 (Efficiency). *An estimator is a locally efficient estimator of $\Psi(P)$ at P w.r.t. $\mathcal{M}$ if it has smallest asymptotic variance among all regular estimators of $\Psi(P)$ at P w.r.t. $\mathcal{M}$.*

Definition 2.12 (Generalized Cramer-Rao lower bound). *Consider a one-dimensional model $\{P_\epsilon : \epsilon\}$ such that $P_{\epsilon=0} = P$. From classical parametric statistics theory, an regular estimator Ψ_n of $\Psi(P)$ w.r.t. $\{P_\epsilon : \epsilon\}$ has asymptotic variance greater than the Cramer-Rao lower bound $v_{CR}(\{P_\epsilon : \epsilon\})$:*

$$\lim_{n\to\infty} n \mathit{Var}_P(\hat{\Psi}_n) \geq v_{CR}(\{P_\epsilon : \epsilon\}) \equiv \frac{\left(\left.\frac{d\Psi(P_\epsilon)}{d\epsilon}\right|_{\epsilon=0} \right)^2}{\mathcal{I}(P)}$$

where $\mathcal{I}(P) = E_P \left[\frac{d^2 \log(dP/d\mu)(X)}{d\epsilon^2} \Big|_{\epsilon=0} \right]$ *is the Fisher information of the model* $\{P_\epsilon : \epsilon\}$ *at* $\epsilon = 0$.

For a statistical model $\mathcal{M}$*, the generalized Cramer-Rao lower bound* v_{GCR} *for* $\Psi(P)$ *w.r.t.* $\mathcal{M}$*, is the sup of the Cramer-Rao lower bound over the parametric submodels of* $\mathcal{M}$ *through* P*:*

$$v_{GCR}(\mathcal{M}) \equiv \sup_{\{P_\epsilon:\epsilon\} \subseteq \mathcal{M}, P_{\epsilon=0}=P} v_{CR}(\{P_\epsilon : \epsilon\}).$$

A parametric submodel whose Cramer-Rao lower bound is equal to the generalized Cramer-Rao lower bound is called a least-favorable parametric submodel.

Definition 2.13 (Semiparametric efficiency)**.** *An estimator* $\hat{\Psi}_n$ *is a locally semiparametric efficient of* $\Psi(P)$ *w.r.t.* $\mathcal{M}$ *if it is consistent for* $\Psi(P)$ *and if its asymptotic variance is equal to the generalized Cramer-Rao lower bound, that is if*

$$\lim_{n\to\infty} n Var_P(\hat{\Psi}_n) = v_{CGR}(\mathcal{M}).$$

If there exists a least-favorable parametric submodel, a semiparametric efficient estimator has the same asymptotic variance as the last-favorable parametric submodel.

2.C.2 Proving that an estimator is semiparametric efficient

In this section, we present a sufficient condition for an estimator to be locally semiparametric efficient. Checking this condition is a standard approach to proving that an estimator is locally semiparametric efficient.

The condition requires a certain characteristic of the estimator, the *influence function* (IF), and a certain characteristic of the estimand and the model, the *efficient influence function* (EIF) to be defined and equal. The IF of an estimator is defined if the estimator satisfies the *asymptotic linearity* property. The EIF at P w.r.t. $\mathcal{M}$ of the estimand $\Psi(P)$ is defined if the estimand is *pathwise differentiable* at P w.r.t. $\mathcal{M}$.

Definition 2.14 (Asymptotic linearity and IF)**.** *An estimator* $\hat{\Psi}_n : \mathcal{X}^n \to \mathbb{R}$*, based on i.i.d. sample* $X_1, ..., X_n$*, of a parameter* $\Psi(P)$ *is asympototically linear at* P*, with influence function* $D(P) : \mathcal{X} \to \mathbb{R}$ *if*

$$\hat{\Psi}_n - \Psi(P) = \frac{1}{n} \sum_{i=1}^{n} D(P)(X_i) + o_P(n^{-1/2}).$$

Definition 2.15 (Pathwise differentiability, gradient, and EIF)**.** *The target parameter mapping/the estimand* Ψ *is pathwise differentiable at* P*, w.r.t.* $\mathcal{M}$*, if there exists a function* $D^0(P) \in L_2^0(P)$*, (where* $L_2^0(P) = \{f \in L_2(P) : Pf = 0\}$*), such that, for all parametric submodel* $\{P_\epsilon : \epsilon\} \subseteq \mathcal{M}$*, with score function* S *at* $\epsilon = 0$ *such that* $P_{\epsilon=0} = P$*, we have that*

$$\frac{d\Psi(P_\epsilon)}{d\epsilon}\Big|_{\epsilon=0} = P\{D^0(P)S\}.$$

If it exists, $D^0(P)$ *is called a gradient of* Ψ *at* P *w.r.t.* $\mathcal{M}$*. The efficient influence function of* Ψ *at* P *w.r.t.* $\mathcal{M}$*, also called canonical gradient is the unique gradient of* Ψ *at* P *w.r.t.* $\mathcal{M}$ *that belongs to* $T_{\mathcal{M}}(P)$*.*

Proposition 2.1. *The EIF is the projection on $T_{\mathcal{M}}(P)$, for the $L^2(P)$ norm, of any gradient.*

Proposition 2.2. *Consider $P \in \mathcal{M}$ Suppose $\hat{\Psi}_n$ is a regular estimator of $\Psi(P)$ w.r.t. $\mathcal{M}$, and that it is asymptotically linear with influence function $D(P)$. Then Ψ is pathwise differentiable at P w.r.t. $\mathcal{M}$ and $D(P)$ is a gradient of Ψ at P w.r.t. $\mathcal{M}$.*

Theorem 2.2. *If a RAL estimator has IF the EIF of the target parameter at P, then it is locally semiparametric efficient at P for the target parameter.*

These results suggest the following strategy to find the efficient influence function of a target parameter: find a RAL estimator of the target, observe that its IF is a gradient, obtain the EIF by projecting the gradient onto the tangent space.

2.C.3 Explicit derivation of the EIF

Proof. We proceed in three steps.

Step 1: Finding a gradient. Denote $\Psi(P) \equiv V_1^{\pi_e}(s_1)$. Consider

$$\hat{\Psi}_n^0 \equiv \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\gamma^t \rho_{1:t}^{(i)} R_t^{(i)}.$$

Observe that

$$\hat{\Psi}_n^0 - \Psi(P) = \frac{1}{n}\sum_{i=1}^{T} D^0(P)(H_i),$$

where $D^0(P)(h) = \sum_{t=1}^{T}\gamma^t \rho_{1:t}(h) r_t - \Psi(P)$.

Therefore $D^0(P)(h)$ is the influence function of the estimator $\hat{\Psi}_n^0$ at P. It is straightforward to check that $\hat{\Psi}_n^0$ is regular. Therefore, from proposition 2.2, Ψ is pathwise differentiable at P w.r.t. our statistical model $\mathcal{M}(\pi)$ and $D^0(P)$ is a gradient of Ψ at P w.r.t. $\mathcal{M}(\pi)$

Step 2: Identifying the tangent space. Since we assumed that distributions in $\mathcal{M}(\pi)$ are dominated by a measure μ, every element $P \in \mathcal{M}(\pi)$ can be represented by its density w.r.t. μ, which we will denote p: for every $h \in \mathcal{H}$, denoting $\bar{h}_t \equiv (s_1, a_1, r_1, ..., s_t, a_t, r_t)$ the history of the trajectory up till time t, we have

$$\begin{aligned} p(h) =& \frac{dP}{d\mu}(h) \\ =& \prod_{t=1}^{T} \tilde{q}_t(s_{t+1}, r_t | s_t, a_t, \bar{h}_{t-1}) \prod_{t=1}^{T} \pi_b(a_t | s_t,). \end{aligned}$$

Consider a one-dimensional submodel $\{P_\epsilon : \epsilon\} \subseteq \mathcal{M}$ that passes through P in $\epsilon = 0$. Then

$$\begin{aligned} &\left.\frac{d \log p_\epsilon(h)}{d\epsilon}\right|_{\epsilon=0} \\ = \sum_{t=1}^{T} &\left.\frac{d \log \tilde{q}_t(s_{t+1}, r_t | s_t, a_t, \bar{h}_{t-1})}{d\epsilon}\right|_{\epsilon=0}. \end{aligned}$$

Since, for any t, $h \mapsto \left.\frac{d \log \tilde{q}_t(s_{t+1}, r_t | s_t, a_t, \bar{h}_{t-1})}{d\epsilon}\right|_{\epsilon=0}$ is a score function, it is in $L_2^0(P_{R_t, S_{t+1} | S_t, A_t, \bar{H}_{t-1}})$. Therefore,

$$T_{\mathcal{M}(\pi)}(P) \subseteq \sum_{t=1}^{T} L_2^0(P_{R_t, S_{t+1} | S_t, A_t, \bar{H}_{t-1}}).$$

Conversely, for any $(g_1, ..., g_T) \in L_2^0(P_{R_1, S_2 | S_1, A_1}) \times ... \times L_2^0(P_{R_T, S_{T+1} | S_T, A_T, \bar{H}_{T-1}})$, for $\eta > 0$ small enough

$$\{P_\epsilon : dP/d\epsilon = p_\epsilon, \epsilon \in [\eta, -\eta]\},$$

where p_ϵ is defined, for all $h \in \mathcal{H}$ as

$$\begin{aligned} p_\epsilon(h) = \prod_{t=1}^{T} &\tilde{q}_t(s_{t+1}, r_t | s_t, a_t) \\ &\times (1 + g_t(s_{t+1}, r_t, s_t, a_t))\pi_b(a_t | s_t), \end{aligned}$$

is a submodel of $\mathcal{M}$ that passes through P at $\epsilon = 0$. We have that, for all $h \in \mathcal{H}$,

$$\left.\frac{d \log p_\epsilon}{d\epsilon}\right|_{\epsilon=0} = \sum_{t=1}^{T} g_t(s_{t+1}, r_t, a_t, s_t, \bar{h}_{t-1}).$$

Since $\frac{d \log p_\epsilon}{d\epsilon}$ is in $T_{\mathcal{M}(\pi)}$ by definition of $T_{\mathcal{M}(\pi)}$, and that $\sum_{t=1}^{T} g_t$ is an arbitrary element of

$$\sum_{t=1}^{T} L_2^0(P_{R_t, S_{t+1} | S_t, A_t, \bar{H}_{t-1}}),$$

this shows that $\sum_{t=1}^{T} L_2^0(P_{R_t, S_{t+1} | S_t, A_t}) \subseteq T_{\mathcal{M}}(P)$ and therefore that

$$T_{\mathcal{M}(P)} = \sum_{t=1}^{T} L_2^0(P_{R_t, S_{t+1} | S_t, A_t, \bar{H}_{t-1}})$$

It is straightforward to check that the sum is direct and orthogonal.

Step 3: Projecting $D^0(P)$ on the tangent space. From proposition 2.1, the EIF is given by

$$\begin{aligned}
&\Pi(D^0(P)\big|T_{\mathcal{M}(\pi)}(P)) \\
&=\sum_{t=1}^{T}\Pi(D^0(P)\big|L_2^0(P_{R_t,S_{t+1}|S_t,A_t,\bar{H}_{t-1}})) \\
&=\sum_{t=1}^{T}\big(E_P[D^0(P)(H)|S_{t+1},R_t,S_t,A_t,\bar{H}_{t-1}] \\
&\qquad - E_P[D^0(P)(H)|S_t,A_t,\bar{H}_{t-1}]\big).
\end{aligned}$$

Observing that the terms that are deterministic conditional on $\bar{H}_{t-1}$ cancel out, we have that

$$\begin{aligned}
&E_P[D^0(P)(H)|S_{t+1},R_t,S_t,A_t,\bar{H}_{t-1}] \\
&- E_P[D^0(P)(H)|S_t,A_t,\bar{H}_{t-1}] \\
=&E_P\left[\sum_{\tau=t}^{T}\gamma^\tau\rho_{1:\tau}R_\tau|S_{t+1},R_t,S_t,A_t,\bar{H}_{t-1}\right] \\
&- E_P\left[\sum_{\tau=t}^{T}\gamma^\tau\rho_{1:\tau}R_\tau|S_t,A_t,\bar{H}_{t-1}\right] \\
=&\gamma^t\rho_{1:t}\Bigg(R_t \\
&+\gamma E_P\left[\sum_{\tau=t}^{T}\gamma^{\tau-t}\rho_{t:\tau}R_\tau\big|S_{t+1},R_t,S_t,A_t,\bar{H}_{t-1}\right]\Bigg) \\
&-\gamma^t\rho_{1:t}E_P\left[\sum_{\tau=t}^{T}\gamma^{\tau-t}\rho_{t:\tau}R_\tau\big|S_t,A_t,\bar{H}_{t-1}\right] \\
=&\gamma^t\rho_{1:t}\Big(R_t+\gamma V^{\pi_e}(S_{t+1})-Q^{\pi_e}(S_t,A_t)\Big)
\end{aligned}$$

Conclusion. The right-hand side of the last line above is equal to $D(P)$ from section 2.B. Since, as we see in the next section, our full-blown estimator has asymptotic variance equal to the variance of $D(P)$ which we just shown to be the EIF, it is semiparemetric efficient.

□

2.D Cross-validated LTMLE

We now present the full-blown version of our algorithm. The key difference between simplified version and the full-blown version is that the latter uses the entire dataset to fit the second-stage models, as opposed to just a split of the dataset. (Using just a split of the dataset is what the simplified algorithm does, along with other algorithms presented recently in the OPE literature.)

The standard error in the simplified version scales as $1/\sqrt{n'}$, where n' is the size of the sample split used to fit ϵ. With the full-blown algorithm, it scales as $1/\sqrt{n}$, where n is the size of the entire sample.

2.D.1 Algorithm description

Consider a sample $H_1, ..., H_n$ of n i.i.d. trajectories of the MDP. Observe that there is a one-to-one relationship between the sample $H_1, ..., H_n$ and the empirical probability distribution $P_n = n^{-1}\sum_{i=1}^{n} \delta_{H_i}$. Therefore, we will refer to the sample and to P_n interchangeably. Let $b_{1,n}, ...,$ $b_{V,n}$ be V vectors in $\{0,1\}^n$ representing splits of the sample: under a given $b_{v,n}$, the training set is given by $\{i : b_{v,n}(i) = 0\}$ and the test set is given by $\{i : b_{v,n}(i) = 1\}$. Let B_n a random vector uniformly distributed on the set $\{b_{v,n} : v = 1, ..., V\}$. Denote P^0_{n,B_n} and P^1_{n,B_n} the empirical distributions of the training set and the test set, respectively, under sample split B_n. Suppose that, for every t, we are given an estimator of $Q^{\pi_e}_t$, that is a mapping of any sample $H'_1, ..., H'_n$, or equivalently, of any probability distribution P'_n, to a model fit, which we will denote $\hat{Q}^{\pi_e}_t(P'_n)$. In practice, $\hat{Q}^{\pi_e}_t(P'_n)$ can be the estimator of $Q^{\pi_e}_t$ obtained under a model of the dynamics fitted from trajectories $H'_1, ..., H'_n$. Denote $\sigma(x) = 1/(1+e^{-x})$ the logistic function, and σ^{-1} its inverse. Observe that under assumption 1, the range of $\bar{R}_{t:T}$, and therefore of $Q^{\pi_e}_t$ and $V^{\pi_e}_t$ is $[-\Delta_t, \Delta_t]$ with $\Delta_t := \sum_{\tau=t}^{T} \gamma^{\tau-t}$. For all t, P'_n, define the scaled action-value function estimator as $\tilde{Q}^{\pi_e}_t(P'_n) = (\hat{Q}^{\pi_e}(P'_n) + \Delta_t)/(2\Delta_t)$. For $\delta \in (0, 1/2)$ and any $\tilde{Q}'_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, define

$$\tilde{Q}'^{\delta}_t(s_t, a_t) := \max(\delta, \min(1-\delta, \tilde{Q}'(s_t, a_t)),$$

the thresholded version of $\tilde{Q}'$ is always at least δ away from 0 and 1. Let $\delta_n \downarrow 0$. For any t, P'_n, introduce the perturbed scaled estimator

$$\tilde{Q}^{\pi_e}_t(P'_n)(\epsilon) = \sigma(\sigma^{-1}(\tilde{Q}^{\pi_e,\delta_n}_t(P'_n)) + \epsilon)$$

and let $\hat{Q}^{\pi_e}(P'_n)(\epsilon)$ be defined by $\tilde{Q}^{\pi_e}_t(P'_n)(\epsilon) = (\hat{Q}^{\pi_e}(P'_n)(\epsilon) + \Delta_t)/(2\Delta_t)$. The expression above defines a logistic regression model with a fixed offset and parameterized by an intercept ϵ, such that $\hat{Q}^{\pi_e}_t(P'_n)(0) = \hat{Q}^{\pi_e}_t(P'_n)$. The TML estimate is obtained by sequentially fitting such logistic models. Specifically, start at time point T and define the cross-validated risk

$$\begin{aligned}
&\mathcal{R}_{n,T}(\epsilon) = \\
&E_{B_n} E_{P^1_{n,B_n}} \Bigg[\rho_{1:T} \Bigg(R_T \log \left(\tilde{Q}^{\pi_e}_T (P^0_{n,B_n})(\epsilon)(S_T, A_T) \right) \\
&+ (1 - R_T) \log \left(1 - \tilde{Q}^{\pi_e}_T (P^0_{n,B_n})(\epsilon)(S_T, A_T) \right) \Bigg) \Bigg].
\end{aligned}$$

Denote $\epsilon_{n,T}$ the minimizer of $\mathcal{R}_{n,T}(\epsilon)$. The other models are fitted by backward recursion. Suppose that we have fitted $\epsilon_{n,T}, ..., \epsilon_{n,t+1}$. Define $\hat{V}^{\pi_e}_{t+1}(P'_n)(\epsilon)(s) := \sum_a \pi_e(a|s) \hat{Q}^{\pi_e}_{t+1}(P'_n)(\epsilon)(s, a)$, $U_{t,B_n} = R_t + \gamma \hat{V}^{\pi_e}_{t+1}(P^0_{n,B_n})(\epsilon_{n,t+1})(S_{t+1})$ and $\tilde{U}_{t,B_n} = (U_{t,B_n} + \Delta_t)/(2\Delta_t)$. Define the cross-validated risk

$$\mathcal{R}_{n,t}(\epsilon) =$$

$$E_{B_n} E_{P^1_{n,B_n}} \Bigg[\rho_{1:t} \Bigg(\tilde{U}_{t,B_n} \log \left(\tilde{Q}^{\pi_e}_t (P^0_{n,B_n})(\epsilon)(S_t, A_t) \right) \\ + (1 - \tilde{U}_{t,B_n}) \log \left(1 - \tilde{Q}^{\pi_e}_t (P^0_{n,B_n})(\epsilon)(S_t, A_t) \right) \Bigg) \Bigg].$$

The perturbation $\epsilon_{n,t}$ is defined as the minimizer of the above risk. The TML estimator of $V_1^{\pi_e}(s_1)$ is defined as

$$\hat{V}_1^{\pi_e, TMLE}(s_1) := E_{B_n} \hat{V}_1^{\pi_e}(P^0_{n,B_n})(\epsilon_1)(s_1).$$

Theorem 2.3. *Suppose assumptions 2.1, 2.2, 2.3, 2.4, 2.5 are satisfied. Then*

$$E_{P^{\pi_b}}[\hat{V}^{\pi_e, TMLE}(s_1) - V_1^{\pi_e}(s_1)] = o(1/\sqrt{n}),$$

and

$$\sqrt{n} \left(\hat{V}^{\pi_e, TMLE}(s_1) - V_1^{\pi_e}(s_1) = O_P(1/\sqrt{n}) \right) \\ \xrightarrow{d} \mathcal{N}(0, \sigma^2(Q_\infty(\epsilon_\infty, \rho)),$$

where, for all non-random Q' and ρ'

$$\sigma^2(Q', \rho') = Var_{P^{\pi_b}}(D(Q', \rho')).$$

It has been established in earlier works Jiang and Li [2015] that the DR estimator with initial estimator $\hat{Q}$ also have asymptotic variance $\sigma^2(Q_\infty)/n$, and that $\sigma^2(Q_\infty)/n$ is the efficient variance from the Cramer-Rao lower bound, provided $Q_\infty = Q$ (that is provided the initial estimator $\hat{Q}$ is asymptotically consistent.). If the initial estimator $\hat{Q}$ is consistent $\epsilon_\infty = 0$ and $Q_\infty(\epsilon_\infty) = Q_\infty = Q$, therefore the DR estimator and the LTMLE have the same asymptotic distribution and they achieve the Cramer-Rao lower bound.

2.D.2 Additional notation

For a given policy π and a transition probability P, denote P^π the corresponding probability distribution over a trajectory with fixed inital state, that is, for all $h = (s_1, a_1, r_1, ..., s_T, a_T, r_T)$,

$$P^{\pi_b}(H = h) = \prod_{t=1}^{T} P(r_t, s_{t+1} | s_t, a_t) \pi(a_t | s_t).$$

From now, we will denote $Q' := (Q'_1, ..., Q'_T)$ an arbitrary action-value function, and let $V' = (V'_1, ..., V'_T)$ be the corresponding value function under π_e, that is for all t, s_t, $V'_t(s_t) = \sum_{a'_t} \pi_e(a'_t | s_t) Q'(s_t, a'_t)$. We will also drop the π_e subscript whenever possible and denote $Q_t := Q_t^{\pi_e}$, the true action value function at time t, and similarly, we will denote $V_t := V_t^{\pi_e}$, the true value function at time t. Denote $Q = (Q_1, ..., Q_T)$ and $V = (V_1, ..., V_T)$.

We introduce the following notation for the perturbed estimators: denote

$$\begin{aligned}\hat{Q}^*_t(P^0_{n,B_n}) :=& \hat{Q}^{\pi_e}_t(P^0_{n,B_n})(\epsilon_{n,t}),\\ \tilde{Q}^*_t(P^0_{n,B_n}) :=& \tilde{Q}^{\pi_e}_t(P^0_{n,B_n})(\epsilon_{n,t}),\\ \hat{V}^*_t(P^0_{n,B_n})(\cdot) =& \sum_{a'_t} \pi_e(a'_t|\cdot)\tilde{Q}^*_t(P^0_{n,B_n})(a'_t|\cdot).\end{aligned}$$

Finally, define

$$\begin{aligned}\hat{Q}^*(P^0_{n,B_n}) :=& (\hat{Q}^*_1(P^0_{n,B_n}), ..., \hat{Q}^*_T(P^0_{n,B_n})),\\ \hat{V}^*(P^0_{n,B_n}) :=& (\hat{V}^*_1(P^0_{n,B_n}), ..., \hat{V}^*_{T+1}(P^0_{n,B_n})).\end{aligned}$$

2.D.3 The fits of the second-stage models solve a score equation

Lemma 2.9. *The perturbed estimators $\hat{Q}^*$, $\hat{V}^*$ given by the LTMLE algorithm satisfy*

$$E_{B_n} P^1_{n,B_n} D^*(\hat{Q}^*(P^0_{n,B_n}), \hat{V}^*(P^0_{n,B_n})) = 0.$$

Proof. Defining $\tilde{U}_{T,B_n} := R_T$, we have that, for all $t = 1, ..., T$,

$$l_{t,n}(\epsilon) = -E_{B_n} P^1_{n,B_n} f_{B_n}(\epsilon),$$

with

$$\begin{aligned}&f_{B_n}(\epsilon)(H) :=\\ &\rho_{1:t}\big(- \tilde{U}_{t,B_n} \log \sigma(a_{B_n} + \epsilon)\\ &\qquad - (1 - \tilde{U}_{t,B_n}) \log(1 - \sigma(a_{B_n} + \epsilon))\big),\end{aligned}$$

where $a_{B_n} := \sigma^{-1}(\tilde{Q}^{\pi_e}(P^0_{n,B_n})(S_t, A_t))$. Using the expression of σ, we rewrite $f_{B_n}(\epsilon)$ as

$$\begin{aligned}&f_{B_n}(\epsilon)(H) =\\ &\rho_{1:t}(\tilde{U}_{t,B_n} \log(1 + e^{-a_{B_n}-\epsilon})\\ &\qquad + (1 - \tilde{U}_{t,B_n}) \log(1 + e^{a_{B_n}+\epsilon})).\end{aligned}$$

We take the derivative of f_{B_n} w.r.t. ϵ:

$$\begin{aligned}&f'_{B_n}(\epsilon)(H)\\ =&\rho_{1:t}\left(-\tilde{U}_{t,B_n} \frac{e^{-a_{B_n}-\epsilon}}{1 + e^{-a_{B_n}-\epsilon}} + (1 - \tilde{U}_{t,B_n}) \frac{e^{a_{B_n}+\epsilon}}{1 + e^{a_{B_n}+\epsilon}}\right)\\ =&\rho_{1:t}\left(-\tilde{U}_{t,B_n}(1 - \sigma(a_{B_n} + \epsilon)) + (1 - \tilde{U}_{t,B_n})\sigma(a_{B_n} + \epsilon)\right)\\ =&\rho_{1:t}\left(\sigma(a_{B_n} + \epsilon) - \tilde{U}_{t,B_n}\right).\end{aligned}$$

Recalling the definitions of a_{B_n}, $\tilde{U}_{t,B_n}$, and $\tilde{Q}^{\pi_e}$, we rewrite the above expression as

$$\begin{aligned}
&f'_{B_n}(\epsilon)(H)\\
=&\rho_{1:t}\left(\tilde{Q}^{\pi_e}(P^0_{n,B_n})(\epsilon)(S_t,A_t)-\tilde{U}_{t,B_n}\right)\\
=&(2\Delta_t)^{-1}\rho_{1:t}\Bigg(\hat{Q}^{\pi_e}_t(P^0_{n,B_n})(\epsilon)(S_t,A_t)\\
&\qquad\qquad -R_t-\gamma\hat{V}^{\pi_e}_{t+1}(P^0_{n,B_n})(\epsilon_{t+1})(S_{t+1})\Bigg)
\end{aligned}$$

Since ϵ_t verifies $l'_{n,t}(\epsilon)=0$, we have that

$$\begin{aligned}
E_{B_n}E_{P^1_{n,B_n}}\Bigg[\rho_{1:t}\Bigg(&R_t+\gamma\hat{V}^{\pi_e}_{t+1}(P^0_{n,B_n})(\epsilon_{t+1})(S_{t+1})\\
&-\hat{Q}^{\pi_e}_t(P^0_{n,B_n})(\epsilon_t)(S_t,A_t)\Bigg)\Bigg]=0,
\end{aligned}$$

that is

$$E_{B_n}P^1_{n,B_n}D_t(\hat{Q}^*_t(P^0_{n,B_n}),\hat{V}^*_t(P^0_{n,B_n}))=0.$$

Summing over t yields the result. □

2.D.4 Proof of convergence of the perturbations

Lemma 2.10. *Define, for all $x\in(0,1)$, $\epsilon\in\mathbb{R}$,*

$$\begin{aligned}
\phi_1(\epsilon,x):=&\log(\sigma(\sigma^{-1}(x)+\epsilon))\\
\text{and } \phi_2(\epsilon,x):=&\log(1-\sigma(\sigma^{-1}(x)+\epsilon)).
\end{aligned}$$

It holds that, for all $x\in(0,1)$, $\epsilon\in\mathbb{R}$,

$$\begin{aligned}
\frac{\partial\phi_1}{\partial x}(\epsilon,x)=&\left(\frac{1}{x}+\frac{1}{1-x}\right)(1-\sigma(\sigma^{-1}(x)+\epsilon)),\\
\text{and } \frac{\partial\phi_2}{\partial x}(\epsilon,x)=&\left(\frac{1}{x}+\frac{1}{1-x}\right)\sigma(\sigma^{-1}(x)+\epsilon).
\end{aligned}$$

Therefore, if $x\in[\delta,1-\delta]$ for some $\delta\in(0,1/2)$ we have that for all $\epsilon\in\mathbb{R}$,

$$\left|\frac{\partial\phi_1}{\partial x}(\epsilon,x)\right|\leq 2\delta^{-1}\qquad\text{and}\qquad\left|\frac{\partial\phi_2}{\partial x}(\epsilon,x)\right|\leq 2\delta^{-1}.$$

Lemma 2.11. *Consider ϕ_1 and ϕ_2 as in lemma 2.10 above, and suppose that $x\in[\delta,1-\delta]$, for some $\delta\in(0,1/2)$. It holds that for all $\epsilon\in\mathbb{R}$*

$$\begin{aligned}
&|\phi_1(\epsilon,x)|\leq\log(1+\delta^{-1}e^{\epsilon}),\\
\text{and } &|\phi_2(\epsilon,x)|\leq\log(1+\delta^{-1}e^{\epsilon}).
\end{aligned}$$

Lemma 2.12. *Assume that for all $h \in \mathcal{H}$, for all $t = 1, .., T$, $0 \leq \rho_{1:t}(h) \leq M$ for some $M > 0$. Assume that for all $\|\hat{Q}_t(P_n') - Q_{\infty,t}\|_{P^{\pi_b},2} = o_P(\delta_n)$ for some $\delta_n \downarrow 0$. Assume that for all $t = 1, ..., T$, for all $_t, a_t \in \mathcal{S} \times \mathcal{A}$, $\tilde{Q}_{t,\infty}(s_t, a_t) \in [\delta, 1-\delta]$ for some $\delta \in (0, 1/2)$. Then, for all $\epsilon \in \mathbb{R}$,*

$$\mathcal{R}_{n,t}(\epsilon) - \mathcal{R}_{\infty,t}(\epsilon) = o_P(1).$$

Proof. Let $\epsilon \in \mathbb{R}$. We express the risk $\mathcal{R}_{n,t}$ as a cross-validated empirical mean of a loss, and the risk $\mathcal{R}_\infty$ as the population mean of a loss:

$$\begin{aligned} \mathcal{R}_{n,t}(\epsilon) =& E_{B_n} P^1_{n,B_n} l_t(\tilde{Q}_t^{\delta_n}(P^0_{n,B_n})(\epsilon)), \\ \text{and } \mathcal{R}_{\infty,t}(\epsilon) =& P^{\pi_b} l_t(\tilde{Q}_{\infty,t}(\epsilon)), \end{aligned}$$

where, for all $\tilde{Q}_t' : \mathcal{S} \times \mathcal{A} \to (0,1)$, for all $h \in \mathcal{H}$

$$\begin{aligned} & l_t(\tilde{Q}_t')(h) := \\ & \rho_{1:t}(h)\Bigg(\tilde{u}_{t,n,B_n} \log(\tilde{Q}_t'(s_t, a_t)) \\ & \qquad + (1 - \tilde{u}_{t,n,B_n}) \log(1 - \tilde{Q}_t'(s_t, a_t))\Bigg). \end{aligned}$$

From there, we are going to proceed in three steps: in the first steps, we will first decompose $\mathcal{R}_{n,t}(\epsilon) - \mathcal{R}_{\infty,t}(\epsilon)$ in two terms $A_{n,t}$ and $B_{n,t}$, that we will then each bound separately in the second and third step.

Step 1: decomposition of $\mathcal{R}_{n,t}(\epsilon) - \mathcal{R}_{\infty,t}(\epsilon)$. Observe that

$$\mathcal{R}_{n,t}(\epsilon) - \mathcal{R}_{\infty,t}(\epsilon) = A_{n,t} + B_{n,t},$$

with

$$A_{n,t} = E_{B_n}(P^1_{n,B_n} - P^{\pi_b}) l_t(\tilde{Q}_t^{\delta_n}(P^0_{n,B_n})(\epsilon)),$$

and

$$B_{n,t} = E_{B_n} P^{\pi_b} \left(l_t(\tilde{Q}_t^{\delta_n}(P^0_{n,B_n})(\epsilon)) - l_t(\tilde{Q}_{\infty,t}(\epsilon)) \right).$$

Step 2: bounding $A_{n,t}$. Let $n_0 = \lfloor np \rfloor$ and $n_1 = n - n_0$. Denote $H^0_{B_n,1}, ..., H^0_{B_n,n_0}$, the trajectories in the training set and $H^1_{B_n,1}, ..., H^1_{B_n,n_1}$ the trajectories in the test set corresponding to sample split B_n.

Since $\tilde{Q}_t^{\delta_n} \in [\delta_n, 1-\delta_n]$, lemma 2.11 shows that

$$\begin{aligned} |\log(\tilde{Q}_t^{\delta_n}(\epsilon)(H^1_{B_n,i}))| \leq & \log(1 + \delta_n^{-1} e^\epsilon) \\ \lesssim & \log(1/\delta_n), \end{aligned}$$

and

$$\begin{aligned} |\log(1 - \tilde{Q}_t^{\delta_n}(\epsilon)(H^1_{B_n,i}))| \leq & \log(1 + \delta_n^{-1} e^{\epsilon}) \\ \lesssim & \log(1/\delta_n). \end{aligned}$$

Recalling the expression of $l_{n,t}$, the fact that by assumption, for every $i = 1, ..., n_1$, $\rho_{1:t}(H^1_{B_n,i}) \leq M$ almost surely, and the fact that $\tilde{U}_{t,n,B_n}(H^1_{B_n,i}) \in [0,1]$, we can bound the loss as follows:

$$|l_t(\tilde{Q}_t^{\delta_n}(P^0_{n,B_n})(\epsilon))(H^1_{B_n,i})| \lesssim M \log(1/\delta_n),$$

almost surely, for every $i = 1, ..., n_1$. Conditional on $H^0_{B_n,1}, ..., H^0_{B_n,n_0}$, $l_t(\tilde{Q}_t^{\delta_n}(P^0_{n,B_n})(\epsilon))(H^1_{B_n,1}), ..., l_t(\tilde{Q}_t^{\delta_n}(P^0_{n,B_n})(\epsilon))(H^1_{B_n,n_1})$ are i.i.d. random variables upper bounded, up to a constant, by $M \log(1/\delta_n)$. Therefore, by Hoeffding's inequality, for every $x > 0$,

$$\begin{aligned} & P[|(P^1_{n,B_n} - P^{\pi_b}) l_t(\tilde{Q}_t^{\delta_n}(P^0_{n,B_n})(\epsilon))| > x] \\ \leq & 2 \exp\left(- \frac{n x^2}{2 \log(1/\delta_n)} \right). \end{aligned}$$

Therefore,

$$\begin{aligned} & (P^1_{n,B_n} - P^{\pi_b}) l_t(\tilde{Q}_t^{\delta_n}(P^0_{n,B_n})(\epsilon)) \\ = & O_P\left(\sqrt{\frac{\log(1/\delta_n)}{n}} \right). \end{aligned}$$

Since B_n takes finitely many values, and that $\log(1/\delta_n) = o(n)$, the above display implies that

$$E_{B_n}(P^1_{n,B_n} - P^{\pi_b}) l_t(\tilde{Q}_t^{\delta_n}(P^0_{n,B_n})(\epsilon)) = o_P(1).$$

Step 3: bounding $B_{n,t}$. Since $\rho_{1:t} \leq M$ for every t almost surely under P^{π_b}, there exists a subset $\bar{\mathcal{H}}$ of $\mathcal{H}$ such that $H \in \bar{\mathcal{H}}$ almost surely, and for all $h \in \bar{\mathcal{H}}$, $\rho_{1:t}(h) \leq M$ for every t. As far as integrals w.r.t. are concerned, P^{π_b}, it is enough to characterize the integrands on $\bar{\mathcal{H}}$. Let h be an arbitrary element of $\bar{\mathcal{H}}$.

$$\begin{aligned} & l_t(\tilde{Q}_t^{\delta_n}(P^0_{n,B_n})(\epsilon)) - l_t(\tilde{Q}_{\infty,t}(\epsilon)) \\ & = \rho_{1:t}(h) \Bigg\{ \tilde{u}_{t,n,B_n}(h) \Bigg(\log(\tilde{Q}_t^{\delta_n}(P^0_{n,B_n})(\epsilon)(h)) \\ & \qquad - \log(\tilde{Q}_{\infty,t}(\epsilon)(h)) \Bigg) \\ & + (1 - \tilde{u}_{t,n,B_n}(h)) \Bigg(\log(1 - \tilde{Q}_t^{\delta_n}(P^0_{n,B_n})(\epsilon)(h)) \\ & \qquad - \log(1 - \tilde{Q}_{\infty,t}(\epsilon)(h)) \Bigg) \Bigg\}. \end{aligned} \tag{2.12}$$

From lemma 2.10 and the mean value theorem, for all n such that $\delta_n \leq \delta$,

$$\begin{aligned}
&\big|\log(\tilde{Q}_t^{\delta_n}(P^0_{n,B_n})(\epsilon)(h)) - \log(\tilde{Q}_{\infty,t}(\epsilon)(h))\big| \\
&\leq 2\delta_n^{-1}\big|\tilde{Q}_t^{\delta_n}(P^0_{n,B_n})(h) - \tilde{Q}_{\infty,t}(h))\big| \\
&\leq 2\delta_n^{-1}\big|\tilde{Q}_t(P^0_{n,B_n})(h) - \tilde{Q}_{\infty,t}(h))\big| \\
&\leq 2\delta_n^{-1}(2\Delta_t)^{-1}\big|\hat{Q}_t(P^0_{n,B_n})(h) - Q_{\infty,t}(h))\big|.
\end{aligned} \tag{2.13}$$

The third line above follows from the fact that, for all $x \in [0,1]$, $y \in [\delta, 1-\delta]$, and n such that $\delta_n \leq \delta$, it holds that $|\max(\delta_n, \min(1-\delta_n, x)) - y| \leq |x - y|$. The same reasoning shows that

$$\begin{aligned}
&\big|\log(1 - \tilde{Q}_t^{\delta_n}(P^0_{n,B_n})(\epsilon)(h)) - \log(1 - \tilde{Q}_{\infty,t}(\epsilon)(h))\big| \\
&\leq 2\delta_n^{-1}(2\Delta_t)^{-1}\big|\hat{Q}_t(P^0_{n,B_n})(\epsilon)(h) - Q_{\infty,t}(\epsilon)(h))\big|.
\end{aligned} \tag{2.14}$$

Taking the absolute value of (2.12), using the triangle inequality, the fact that $0 \leq \rho_{1:t}(h) \leq M$, that $\tilde{u}_{t,n,B_n}(h) \in [0,1]$ and the upper bounds (2.13) and (2.14) yields

$$\begin{aligned}
&\big|l_t(\tilde{Q}_t^{\delta_n}(P^0_{n,B_n})(\epsilon)) - l_t(\tilde{Q}_{\infty,t}(\epsilon))\big| \\
&\leq M2\delta_n^{-1}\Delta_t^{-1}\big|\hat{Q}_t(P^0_{n,B_n})(h) - Q_{\infty,t}(h))\big|.
\end{aligned}$$

Therefore, using the triangle inequality and Cauchy-Schwartz, and the fact that B_n takes finitely many values,

$$\begin{aligned}
&\big|E_{B_n}P^{\pi_b}(l_t(\tilde{Q}_t^{\delta_n}(P^0_{n,B_n})(\epsilon)) - l_t(\tilde{Q}_{\infty,t}(\epsilon)))\big| \\
&\leq M2\delta_n^{-1}\Delta_t^{-1}\|\hat{Q}_t(P^0_{n,B_n}) - Q_{\infty,t}\|_{P^{\pi_b},2}.
\end{aligned}$$

Therefore, using the assumption that $\|\hat{Q}_t(P^0_{n,B_n}) - Q_{\infty,t}\|_{P^{\pi_b},2} = o_P(\delta_n)$, we have

$$E_{B_n}P^{\pi_b}(l_t(\tilde{Q}_t^{\delta_n}(P^0_{n,B_n})(\epsilon)) - l_t(\tilde{Q}_{\infty,t}(\epsilon))) = o_P(1).$$

Therefore, putting together that $A_{n,t} = o_P(1)$ and $B_{n,t} = o_P(1)$ and the fact that $\mathcal{R}_{n,t} - \mathcal{R}_{\infty,}(\epsilon) = A_{n,t} + B_{n.t}$ gives the desired result. □

Lemma 2.13. *Make the same assumptions as in lemma 2.12 above. Then*

- $\mathcal{R}_{\infty,t}$ *has a unique minimizer* $\epsilon_{\infty,t}$,
- $\epsilon_{n,t} - \epsilon_{\infty,t} = o_P(1)$.

Proof. Let $\eta > 0$ and $\kappa > 0$. The fact that $Q_{\infty,t} \in [\delta, 1-\delta]$, with $\delta \in (0, 1/2)$ implies that $\mathcal{R}_{\infty,t}$ is m-strongly convex for some $m > 0$. Therefore $\mathcal{R}_{n,t}$ has a unique minimizer on $\mathbb{R}$ that we will denote $\epsilon_{\infty,t}$. Denoting $\Delta := m\eta^2/2$, we have, from m-strong convexity, that

$$\mathcal{R}_{\infty,t}(\epsilon_{\infty,t} + \eta) \geq \mathcal{R}_{\infty,t}(\epsilon_{\infty,t}) + \Delta, \tag{2.15}$$

$$\text{and } \mathcal{R}_{\infty,t}(\epsilon_{\infty,t} - \eta) \geq \mathcal{R}_{\infty,t}(\epsilon_{\infty,t}) + \Delta. \tag{2.16}$$

Consider the following event:

$$\mathcal{E} := \Bigg\{ |\mathcal{R}_{n,t}(\epsilon) - \mathcal{R}_{\infty,t}(\epsilon)| \leq \frac{\Delta}{3}, \\ \forall \epsilon \in \{\epsilon_\infty, \epsilon_\infty - \eta, \epsilon_\infty + \eta\} \Bigg\}.$$

From the pointwise convergence in probability of $\mathcal{R}_{n,t}$, which is given to us by lemma 2.12 above, there exists n_0 such that for all $n \geq n_0$, $P[\mathcal{E}] \geq 1 - \kappa$. Assume $\mathcal{E}$ holds. Then, from (2.15) and (2.16), and the inequalities that define event $\mathcal{E}$, we have that

$$\mathcal{R}_{n,t}(\epsilon_\infty \pm \eta) \geq \mathcal{R}_{n,t}(\epsilon_\infty) + \frac{\Delta}{3}.$$

From convexity of $\mathcal{R}_{n,t}$, the above display implies that for all ϵ such that $|\epsilon - \epsilon_{\infty,t}| \geq \eta$, we have that

$$\mathcal{R}_{n,t}(\epsilon) \geq \mathcal{R}_{n,t}(\epsilon_{\infty,t}) + \frac{\Delta}{3}.$$

Since $\epsilon_{n,t}$ minimizes $\mathcal{R}_{n,t}$, we must have $\mathcal{R}_{n,t} \leq \mathcal{R}_{n,t}(\epsilon_{\infty,t}) < \mathcal{R}_{n,t}(\epsilon_{\infty,t}) + \Delta/3$. Therefore, if $\mathcal{E}$ is realized, $\epsilon_{n,t}$ must lie in $[\epsilon_{\infty,t} - \eta, \epsilon_{\infty,t} + \eta]$. Since $\mathcal{E}$ is realized with probability $1 - \kappa$, this concludes the proof. □

2.D.5 Proof of theorem 2

The proof relies on the following empirical process result.

Lemma 2.14. *Consider $\mathcal{F}_\eta(P^0_{n,B_n})$ as defined in the previous section. Consider $\eta_n = o_P(1)$. Then*

$$\sqrt{n} \sup_{f \in \mathcal{F}_{\eta_n}(P^0_{n,B_n})} |(P^1_{n,B_n} - P^{\pi_b})f| = o_P(1).$$

Proof. This is a direct corollary of lemma 2.4. □

Proof of theorem 2. Recall that $\hat{V}_1^{\pi_e,TMLE}(s_1) = E_{B_n}\hat{V}(P^0_{n,B_n})(\epsilon_{n,1})(s_1)$. Therefore, from lemma 2.1,

$$\begin{aligned}&\hat{V}_1^{\pi_e,TMLE}(s_1) - V^{\pi_e}(s_1) \\ =& - E_{B_n} P^{\pi_b} D(\hat{Q}(P^0_{n,B_n})(\epsilon_n), \hat{V}(P^0_{n,B_n})(\epsilon_n)).\end{aligned} \tag{2.17}$$

Recall that from lemma 2.9,

$$\begin{aligned}&\hat{V}_1^{\pi_e,TMLE}(s_1) - V^{\pi_e}(s_1) \\ =& E_{B_n} P^1_{n,B_n} D(\hat{Q}(P^0_{n,B_n})(\epsilon_n), \hat{V}(P^0_{n,B_n})(\epsilon_n)).\end{aligned} \tag{2.18}$$

Summing (2.17) and (2.18) yields

$$\hat{V}_1^{\pi_e,TMLE}(s_1) - V^{\pi_e}(s_1) =$$
$$E_{B_n}(P^1_{n,B_n} - P^{\pi_b})D(\hat{Q}(P^0_{n,B_n})(\epsilon_n), \hat{V}(P^0_{n,B_n})(\epsilon_n)).$$

Using the notation f_ϵ introduced in the previous section, we can rewrite the above expression as

$$\begin{aligned}&\hat{V}_1^{\pi_e,TMLE}(s_1) - V^{\pi_e}(s_1)\\ &= E_{B_n}(P^1_{n,B_n} - P^{\pi_b})f_{\epsilon_\infty}(P^0_{n,B_n})\\ &+ E_{B_n}(P^1_{n,B_n} - P^{\pi_b})(f_{\epsilon_n}(P^0_{n,B_n}) - f_{\epsilon_\infty}(P^0_{n,B_n})).\end{aligned}$$

By the central limit theorem for triangular arrays

$$\begin{aligned}&\sqrt{n}(P^1_{n,B_n} - P^{\pi_b})f_{\epsilon_\infty}(P^0_{n,B_n})\\ &\xrightarrow{d} \mathcal{N}(0, Var(D(Q_\infty(\epsilon_\infty), V_\infty(\epsilon_\infty)(H))).\end{aligned}$$

Therefore, $\sqrt{n}(P^1_{n,B_n} - P^{\pi_b})f_{\epsilon_\infty}(P^0_{n,B_n}) = O_P(n^{-1/2})$. The second term is the RHS of is $o_P(n^{-1/2})$ by lemma 2.14. Since B_n takes values on a finite support, this implies that

$$\hat{V}_1^{\pi_e,TMLE}(s_1) - V^{\pi_e}(s_1) = O_P(n^{-1/2}).$$

□

2.E Experiment Details

In this section, we provide full details of our experiments and utilized domains. In particular, we provide detailed descriptions of discrete-state domains ModelWin, ModelFail and Gridworld.

2.E.1 ModelWin

The ModelWin environment was constructed in order to simulate situations in which the approximate model of the MDP will converge quickly to the truth. On the other hand, importance-sampling based methods might suffer from high variance.

The ModelWin MDP consists of 3 states, and the agent always begins at state s_1. At s_1, the agent stochastically picks between two actions, a_1 and a_2. Under action a_1, the agent transitions to s_2 with probability 0.4 and s_3 with probability 0.6. On the other hand, under action a_2 the agent does the opposite- it transitions to s_2 and s_3 with probability 0.6 and 0.4, respectively. Under both actions, if the agent transitions to s_2, it gets a positive reward of +1. Consequently s_1 to s_3 transitions are penalized with -1 reward. In states s_3 and s_2, both actions a_1 and a_2 will take the agent back to s_1 with probability 1 and no reward. The horizon is set to $T = 20$.

The considered behavior policy takes action a_1 from s_1 with probability 0.73, and action a_2 with probability 0.27. The evaluation policy has the opposite behavior. Note that both the behavior and evaluation policies select actions uniformly at random while in states s_1 and s_2.

2.E.2 ModelFail

Unlike the ModelWin domain, the agent does not observe the true underlying states of the MDP in ModelFail. The purpose of this domain is to test environments are not known perfectly, and where the approximate model will fail to converge to the true MDP. ModelFail attempts to mimic partial observability, common in real applications.

The actual MDP consists of 4 states, 3 states and a final absorbing state, however the agent is not able to distinguish between them. The agent always starts at the same state, s_1, where it has two actions available. With actions a_1 it transitions into the upper state (s_2), whereas with action a_2 it goes to the lower state (s_3). No matter which state the agent transitioned to, both s_2 and s_3 lead to the terminal absorbing state s_4. However, s_2 to s_4 transition carries reward +1, whereas s_3 to s_4 leads to reward of -1. The horizon is $T = 2$.

The considered behavior policy takes action a_1 with probability 0.88, and action a_2 with probability 0.22. The evaluation policy has the opposite behavior.

2.E.3 Gridworld

The last discrete-state environment used is a 4×4 gridworld domain with 4 actions (up, down, left, right) developed by Thomas [2015]. As emphasized by Thomas and Brunskill [2016], this is a domain specifically developed for evaluation of OPE estimators. However, due to its deterministic nature, it will favor model-based approaches.

The horizon for GridWorld is $T = 100$, after which the episode ends unless the terminal state of s_{12} is reached before T. The reward is always -1, expect at states s_8 where it is +1, s_{12} with +10, and s_6 where the agent is penalized with -10 reward.

We used two different polices for the gridworld, as described in Thomas [2015]. In particular, policy π_1 selects each of the 4 actions with equal probability regardless of the observation. Intuitively this policy takes a long time to reach the goal, and potentially often visits the state with the maximum negative reward. In addition, we also considered the near-optimal+ policy π_5, which exemplifies a near-deterministic near-optimal policy that moves quickly to s_8 with reward +1, without visiting s_6 with -10 reward. At s_8 it chooses action down with high probability, collecting as many positive rewards as possible until the time limit runs out. Once it eventually chooses the right action, it moves almost deterministically to s_{12} where it collects its final reward and end the episode.

Chapter 3

Fast rates for empirical risk minimization over càdlàg functions with bounded sectional variation norm

AURÉLIEN BIBAUT, MARK VAN DER LAAN

In this chapter, we refine existing guarantees for empirical risk minimizers over the nonparametric function class of cadlag functions with bounded section variation norm (s.v.n., also referred to as Hardy-Krause variation). At the time of release of the original preprint [Bibaut and van der Laan, 2019], van der Laan [2016] had shown that that ERMs over this class of function had rate of convergence in L_2 norm $o(n^{-1/4})$, and Fang et al. [2019] had shown that, in regression settings under some restrictive conditions on the design, ERMs had L_2 norm convergence rate $\widetilde{O}(n^{-1/3})$, where n is the sample size. In this work, we show the $\widetilde{O}(n^{-1/3})$ under more general conditions than Fang et al. [2019].

Getting faster than $n^{-1/4}$ rates of convergence over realistic nonparametric models for inifinite dimensional statistical parameters such as regression functions is a paramount importance in semiparametric estimation problems. Indeed in such problems, estimators minus their true target can often be decomposed an empirical mean (or empirical process), which can be shown to converge to its limit distribution at speed $\sqrt{n}$, plus a remainder term. The remainder term often involves the product of the estimation error of two infinite dimensional parameters. When these errors are $o_P(n^{-1/4})$, the remainder can be shown using Cauchy-Schwartz to be $o_P(n^{-1/2})$, and thus to be negligible in front of the first empirical mean or empirical process term. This situation arises for example in average treatment effect estimation from observational data, where the remainder term involves the product of the estimation error for the propensity score and for the outcome model.

The key theoretical building block of this work is a new bound on the bracketing entropy of the class of cadlag function with bounded s.v.n., which we obtain from a bound on the bracketing entropy on the class of multivariate cumulative distribution functions by Gao [2013]. We show that the bracketing entropy of cadlag functions with s.v.n. no larger than a fixed constant is $O(\epsilon^{-1}\log(1/\epsilon)^{2(d-1)})$, where ϵ is the L_2 size of the brackets and d is the dimension of the domain. This enables us to obtain rate of convergence in L_2 norm for ERMs $O_P(n^{-1/3}(\log n)^{2(d-1)/3})$. Perhaps surprisingly, the dependence on the dimension only applies to the $logn$ factor, making these estimators good candidates for high dimensional problems. Simulations studies, in particular that of **?** confirm that these estimators have very strong practical performance.

In this work, we derived statistical guarantees only in the i.i.d. setting. As we see in subsequent chapters, the bracketing entropy bound can be reused in various dependent observations setting, allowing us to get convergence guarantees in particular under martingale conditions and also for weakly dependent data.

3.1 Introduction

Empirical risk minimization setting. We consider the empirical risk minimization setting over classes of real-valued, d-variate functions. Suppose that $O_1, ..., O_n$ are i.i.d. random vectors with common marginal distribution P_0, and taking values in a set Θ. Suppose that $\mathcal{O} \subseteq [0,1]^d \times \mathcal{Y}$, for some integer $d \geq 1$ and some set $\mathcal{Y} \subseteq \mathbb{R}$. Suppose that for all i, $O_i = (X_i, Y_i)$, where $X_i \in [0,1]^d$, $Y_i \in \mathcal{Y}$. We suppose that P_0 lies in a set of probability distributions over $\mathcal{O}$ that we denote $\mathcal{M}$, and which we call the statistical model. Consider a mapping θ from the statistical model to a set Θ of real-valued functions with domain $[0,1]^d$. We call Θ the parameter set. We want to estimate a parameter θ_0 of the data-generating distribution P_0 defined by $\theta_0 = \theta(P_0)$. Let $L : \Theta \to \mathbb{R}^{\mathcal{O}}$ be a loss mapping, that is for every $\theta \in \Theta$, $L(\theta) : \mathcal{O} \to \mathbb{R}$ is a loss function corresponding to parameter value θ. We suppose that L is a valid loss mapping for θ_0 in the sense that

$$\theta_0 = \arg\min_{\theta \in \Theta} P_0 L(\theta).$$

Statistical model, sieve, and estimator We define our statistical model implicitly by making a functional class assumption on the parameter set Θ. Specifically, we suppose that Θ is a subset of the class $\mathcal{F}_d$ of càdlàg functions over $[0,1]^d$ with bounded sectional variation norm [Gill et al., 1995]. We define now the notion of sectional variation norm. Denote $\mathbb{D}([0,1]^d)$ the set of real-value càdlàg functions with domain $[0,1]^d$. Consider a function $f \in \mathbb{D}([0,1]^d)$. For all subset $\emptyset \neq s \subseteq [d]$ and for all vector $x \in [0,1]^d$, define the vectors $x_s = (x_j : j \in s)$, $x_{-s} = (x_j : j \notin s)$, and the section f_s of f as the mapping $f_s(x_s) : x_s \mapsto f(x_s, 0_{-s})$. The sectional variation norm of f is defined as

$$\|f\|_v \equiv |f(0)| + \sum_{\emptyset \neq s \subseteq [d]} \int |f_s(dx_s)|,$$

where $[d]$ is a shorthand notation for $\{1, ..., d\}$ and $f_s(dx_s)$ is the signed measure generated by the càdlàg function f_s. Consider a sequence $(\Theta_n)_{n \geq 1}$ of subsets of Θ such that is non-decreasing for the inclusion. For any $n \geq 1$, we define our estimator $\hat{\theta}_n$ as the empirical risk minimizer over Θ_n, that is

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta_n} P_n L(\theta).$$

Rate of convergence results. Our main theoretical result states that the empirical risk minimizer $\hat{\theta}_n$ converges to θ_0 at least as fast as $O_P(n^{-1/3}(\log n)^{2(d-1)/3} a_n)$, where a_n depends on the rate of growth of Θ_n in terms of variation norm. The key to proving this result is a characterization of the bracketing entropy of the class of càdlàg functions with bounded sectional variation norm. A rate of convergence is then derived based on the famed "peeling" technique.

Tractable representation of the estimator. Fang et al. [2019] showed that if the parameter space is itself a set of càdlàg functions with bounded sectional variation norm, then the empirical risk minimizer $\hat{\theta}_n$ can be represented as a linear combination of a certain set of basis functions. (The number of basis functions grows with n and is no larger than $(ne/d)^d$). The empirical risk minimization problem then reduces to a LASSO problem.

Related work and contributions. van der Laan [2016] considered empirical risk minimization over sieves of $\mathcal{F}_d$, under the general bounded loss setting, and showed that it achieves a rate of convergence strictly faster than $n^{-1/4}$ in loss-based dissimilarity. Fang et al. [2019] consider nonparametric least-squares regression with Gaussian errors and a lattice design, over $\mathcal{F}_{d,M}$ for a certain $M > 0$, and show that the least-squares estimator achieves rate of convergence

$n^{-1/3}(\log n)^{C(d)}$ for a certain constant $C(d)$. In this article, we show that a similar rate of convergence $n^{-1/3}(\log n)^{2(d-1)/3}$ can be achieved under the general setting of empirical risk minimization with unimodal Lipschitz losses (defined formally in section 3.3). We show that this setting covers the case of nonparametric least-squares regression with a bounded dependent variable, and logistic regression, under no assumption on the design. We also consider the nonparametric regression with sub-exponential errors setting, and show that this $n^{-1/3}(\log n)^{2(d-1)/3}a_n$ rate is achieved by the least-squares estimator over a certain sieve of the set of càdlàg functions with bounded sectional variation norm.

3.2 Representation and entropy of the càdlàg functions with bounded sectional variation norm

As recently recalled by van der Laan [2016], Gill et al. [1995] showed that any càdlàg function on $[0,1]^d$ with bounded sectional variation norm can be represented as a sum of (2^d-1) signed measures of bounded variation. This readily implies that any such function can be written as a sum of (2^d-1) differences of scaled cumulative distribution functions, as formally stated in the following proposition.

Proposition 3.1. *Consider $f \in \mathbb{D}([0,1]^d)$ such that $\|f\|_v \leq M$, for some $M \geq 0$. For all subset $s \subseteq [d]$, and for all vector $x \in [0,1]^d$, define the vector $x_s = (x_j : j \in s)$. The function f can be represented as follows: for all $x \in [0,1]^d$,*

$$f(x) = f(0) + (M - |f(0)|) \sum_{\emptyset \neq s \subseteq [d]} \int_0^{x_s} \alpha_{s,1} g_{s,1}(dx_s) - \alpha_{s,2} g_{s,2}(dx_s),$$

where $g_{s,1}$ and $g_{s,2}$ are cumulative distribution functions on the hypercube $[0_s, 1_s]$, and $\boldsymbol{\alpha} = (\alpha_{s,i} : \emptyset \neq s \subseteq [d], i = 1,2\} \in \Delta^{2^{d+1}-2}$, where $\Delta^{2^{d+1}-2}$ is the $(2^{d+1}-2)$-standard simplex.

This and a recent result [Gao, 2013] on the bracketing entropy of distribution functions implies that the class $\mathcal{F}_{d,M}$ of càdlàg functions over $[0,1]^d$ with variation norm bounded by M has well-controlled entropy, as formalized by the following proposition.

Proposition 3.2. *Let $d \geq 2$ and $M > 0$. Denote $\mathcal{F}_{d,M}$ the class of càdlàg functions on $[0,1]^d$ with sectional variation norm smaller than M. Suppose that P_0 is such that, for all $1 \leq r < \infty$, for all real-valued function f on $[0,1]^d$, $\|f\|_{P_0,r} \leq c(r)\|f\|_{\mu,r}$, for some $c(r) > 0$, and where μ is the Lebesgue measure. Then for all $1 \leq r < \infty$ and all $0 < \epsilon < 1$, the bracketing entropy of $\mathcal{F}_{d,M}$ with respect to the $\|\cdot\|_{P_0,r}$ norm satisfies,*

$$\log N_{[]}(\epsilon, \mathcal{F}_{d,M}, \|\cdot\|_{P_0,r}) \lesssim C(r,d) M \epsilon^{-1} |\log(\epsilon/M)|^{2(d-1)},$$

where $C(r,d)$ is a constant that depends only on r and d. This implies the following bound on the bracketing entropy integral of $\mathcal{F}_{d,M}$ with respect to the $\|\cdot\|_{P_0,r}$ norm: for all $0 < \delta < 1$,

$$J_{[]}(\delta, \mathcal{F}_{d,M}, \|\cdot\|_{P_0,r}) \lesssim \sqrt{C(r,d)}\sqrt{M}\delta^{1/2}|\log(\delta/M)|^{d-1}.$$

3.3 Rate of convergence under unimodal Lipschitz losses

In this section, we present an upper bound on the rate of convergence of $\hat{\theta}_n$ under a general class of loss functions. Essentially, we require the loss to be unimodal and Lipschitz with respect to the parameter, in a pointwise sense. We formally state below the assumptions of our result. Let $(a_n)_{n \geq 1}$ be a non-decreasing sequence of positive numbers, that can potentially diverge to ∞.

Assumption 3.1 (Control of the variation norm of the sieve). *Suppose that for all $n \geq 1$,*

$$\Theta_n \subseteq \{\theta \in \mathbb{D}([0,1]^d : \|\theta\|_v \leq a_n\}.$$

Assumption 3.2 (Loss class). *There exists some $\tilde{L} : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ such that, for any n, for any $\theta \in \Theta_n$, and for any $o = (x, y) \in [0,1]^d \times \mathcal{Y}$,*

$$L(\theta)(x,y) = \tilde{L}(\theta(x), y).$$

Further assume that $\tilde{L}$ is such that, for any y, there is an u_y such that $u \mapsto \tilde{L}(u,y)$ is

- *non-increasing on $(-\infty, u_y]$, and non-decreasing on $[u_y, \infty)$,*
- *a_n-Lipschitz.*

We will express the rate of convergence in terms of loss-based dissimilarity, which we define now.

Definition 3.1 (Loss-based dissimilarity). *Let $n \geq 1$. Denote $\theta_n = \arg\min_{\theta \in \Theta_n} P_0 L(\theta)$. For all $\theta \in \Theta_n$, we define the square of the loss-based dissimilarity $d(\theta, \theta)$ between θ and θ_n as the discrepancy*

$$d^2(\theta, \theta_n) = P_0 L(\theta) - P_0 L(\theta_n).$$

The third main assumption of our theorem requires the loss L to be smooth with respect to the loss-based dissimilarity.

Assumption 3.3 (Smoothness). *For every n, it holds that*

$$\sup_{\theta \in \Theta_n} \|L(\theta) - L(\theta_n)\|_{P_0,2} \leq a_n d(\theta, \theta_n).$$

We can now state our theorem.

Theorem 3.1. *Consider Θ_n a sieve such that assumptions 3.1-3.3 hold for the sequence a_n considered here. Suppose that $a_n = O(n^p)$ for some $p > 0$. Consider our estimator $\hat{\theta}_n$, which, we recall, is defined as the empirical risk minimizer over Θ_n, that is*

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta_n} P_n L(\theta).$$

Suppose that

$$\theta_0 \in \Theta_\infty \equiv \{\theta \in \Theta : \exists n_0 \text{ such that } \forall n \geq n_0,\ \theta \in \Theta_n\}.$$

Then, we have the following upper bound on the rate of convergence of $\hat{\theta}_n$ to θ_0:

$$d(\hat{\theta}_n, \theta_0) = O_P(a_n n^{-1/3} (\log n)^{2(d-1)/3}).$$

The reason why we consider a growing sieve Θ_n is to ensure we don't have to know in advance an upper bound on the variation norm of the losses. The rate a_n impacts the asymptotic rate of convergence and finite sample performance. As the theorem makes clear, the slower we pick a_n, the better the speed of convergence. However, for too slow a_n, θ_0 might not be included in Θ_n even for reasonable sample sizes. Note that, if there are reasons to believe that $\|\theta_0\|_v \leq A$ for some $A > 0$, one can set $a_n = A$ and then the rate of convergence will be $O_P(n^{-1/3}(\log n)^{2(d-1)/3})$.

3.4 Applications of theorem 3.1

3.4.1 Least-squares regression with bounded dependent variable

Consider $\tilde{a}_n$ a non-decreasing sequence of positive numbers, that can potentially diverge to ∞. Let $O_1 = (X_1, Y_1), ..., (X_n, Y_n)$ be i.i.d. copies of a random vector $O = (X, Y)$ with distribution P_0. Suppose that X takes values in $[0,1]^d$ and Y takes values in $\mathcal{Y}_n = [-\tilde{a}_n, \tilde{a}_n]$. In the setting of least-squares regression, one wants to estimate the regression function $\theta_0 : x \in [0,1]^d \mapsto E_{P_0}[Y|X = x]$ using the square loss L defined, for all $\theta \in \Theta$ as $L(\theta) : (x, y) \mapsto (y - \theta(x))^2$. Let $\Theta_n = \{\theta \in \mathbb{D}([0,1]^d) : \|\theta\|_v \leq \tilde{a}_n\}$. We consider the least-squares estimator $\hat{\theta}_n$ over Θ_n, defined as

$$\hat{\theta}_n = \arg \min_{\theta \in \Theta_n} P_n L(\theta).$$

Proposition 3.3 and proposition 3.4 below justify that assumptions 3.2 and 3.3 of theorem 3.1 are satisfied.

Proposition 3.3. *Consider the setting of this subsection. We have, for all $n \geq 1$, $\theta \in \Theta_n$, $x \in [0,1]^d$, and $y \in \mathcal{Y}_n$, that*

$$L(\theta)((x, y)) = \tilde{L}(\theta(x), y)$$

where, $\tilde{L}(u, y) = (y - u)^2$ for all u, y.

Furthermore, for all $y \in \mathcal{Y}_n$, the mapping $u \mapsto \tilde{L}(u, y)$ is

- *non-increasing on $(-\infty, y]$ and non-decreasing on $[y, \infty)$,*
- *and $4\tilde{a}_n$-Lipschitz on $\{\theta(x) : \theta \in \Theta_n, x \in [0,1]^d\}$.*

Proposition 3.4. *Consider the setting of this subsection and recall the definition of the loss-based dissimilarity (see definition 3.1). For all $n \geq 1$, $\theta \in \Theta_n$, we have that*

$$\|L(\theta) - L(\theta)_n\|_{P_0,2} \leq 4\tilde{a}_n d_n(\theta, \theta_n).$$

Corollary 3.1. *Set $a_n = 4\tilde{a}_n$. Then,*

$$\|\theta - \theta_n\|_{P_0,2} = O_P(a_n n^{-1/3}(\log n)^{2(d-1)/3}).$$

3.4.2 Logistic regression

Consider $\tilde{a}_n$ a non-decreasing sequence of positive numbers that can potentially diverge to ∞. Let $O_1 = (X_1, Y_1), ..., O_n = (X_n, Y_n)$ be i.i.d. copies of a random vector $O = (X, Y)$, where X takes values in $[0, 1]^d$ and $Y \in \{0, 1\}$. Denote P_0 the distribution of O. We want to estimate

$$\theta_0 : x \mapsto \log \left(\frac{E_{P_0}[Y|X = x]}{1 - E_{P_0}[Y|X = x]} \right),$$

the conditional log-odds function. Let L be the negative log likelihood loss, that is, for all $\theta \in \Theta$, $x \in [0, 1]^d$, $y \in \{0, 1\}$, $L(\theta)(x, y) = y \log(1 + \exp(-\theta(x))) + (1 - y) \log(1 + \exp(\theta(x)))$. Denote $\Theta_n = \{\theta \in \mathbb{D}([0, 1]^d) : \|\theta\|_v \leq \tilde{a}_n\}$. We denote $\hat{\theta}_n$ the empirical risk minimizer over Θ_n, that is

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta_n} P_0 L(\theta).$$

Propositions 3.5 and 3.6 below justify that assumptions 3.2 and 3.3 of theorem 3.1 are satisfied.

Proposition 3.5. *Consider the setting of this subsection. We have, for all $n \geq 1$, $\theta \in \Theta_n$, $o = (x, y) \in [0, 1]^d \times \{0, 1\}$, that*

$$L(\theta)(o) = \tilde{L}(\theta(x), y),$$

where, for all u, y,

$$\tilde{L}(u, y) = y \log(1 + e^{-u}) + (1 - y) \log(1 + e^u).$$

Furthermore, for all $y \in \{0, 1\}$, the mapping $u \mapsto \tilde{L}(u, y)$ is

- *non-increasing on $\mathbb{R}$ if $y = 1$,*
- *non-decreasing on $\mathbb{R}$ if $y = 0$,*
- *1-Lipschitz on $\mathbb{R}$.*

Proposition 3.6. *Consider the setting of this subsection, and recall the definition of the loss-based dissimilarity. For all $n \geq 1$, we have that*

$$\|L(\theta) - L(\theta_n)\|_{P_0,2} \leq 2(1 + e^{\tilde{a}_n})^{1/2} d_n(\theta, \theta_n).$$

Corollary 3.2. *Set $a_n = 2(1 + e^{\tilde{a}_n})^{1/2}$. Then*

$$d_n(\theta, \theta_n) = O_P(a_n n^{-1/3} (\log n)^{2(d-1)/3}),$$

and

$$\|\theta - \theta_n\|_{P_0,2} = O_P(a_n^2 n^{-1/3} (\log n)^{2(d-1)/3}).$$

3.5 Least-squares regression with sub-exponential errors

In this section we consider a fairly general nonparametric regression setting, namely least-squares regression over a sieve of càdlàg functions with bounded sectional variation norm, under the assumption that the errors follow a subexponential distribution. Although this situation isn't covered by the hypothesis of theorem 3.1, our general bounded loss result, it is handled by fairly similar arguments. This is a setting of interest in the literature (see e.g. section 3.4.3.2 of van der Vaart and Wellner [1996]).

Suppose that we collect observations $(X_1, Y_1), ..., (X_n, Y_n)$, which are i.i.d. random variable with common marginal distribution P_0. Suppose that for all i, $X_i \in \mathcal{X} \equiv [0,1]^d$, $Y_i \in \mathcal{Y} \equiv \mathbb{R}$, and that

$$Y_i = \theta_0(X_i) + e_i,$$

where $\theta_0 \in \Theta \equiv \{\theta \in \mathbb{D}([0,1]^d) : \|\theta\|_v < \infty\}$, and $e_1, ...e_n$ are i.i.d. errors that follow a sub-exponential distribution with parameters (α, ν). Suppose that for all i, X_i and e_i are independent. Let a_n be a not-decreasing sequence of positive numbers that can diverge to ∞. Define, for all $n \geq 1$, $\Theta_n = \{\theta \in \Theta : \|\theta\|_v \leq a_n\}$.

The following theorem characterizes the rate of convergence of our least-squares estimators, which we explicitly define in the statement of the theorem.

Theorem 3.2. *Consider the setting of this section. Suppose that $\theta_0 \in \Theta$. Then, $\hat{\theta}_n$, the least-squares estimator over Θ_n, formally defined as*

$$\hat{\theta}_n = \arg \min_{\theta \in \Theta_n} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \theta(X_i))^2,$$

satisfies

$$\|\hat{\theta}_n - \theta_0\|_{P_0,2} = O_P(((\tilde{C}(\alpha, \nu) + 3)a_n + \|\theta_0\|_\infty) n^{-1/3} (\log n)^{2(d-1)/3}).$$

where the constant $\tilde{C}(\alpha, \nu)$ is defined in the appendix.

3.6 Discussion

In this chapter, we analyzed the bracketing entropy of the class of d-variate cadlag functions with bounded sectional variation norm. We use this bound on the bracketing entropy to provide bounds on the convergence rate in probability of empirical risk minimizers.

We studied in particular empirical risk minimizers corresponding to two losses: the square loss and the logistic loss (under the condition that the underlying regression function lies away from 0 and 1). We showed in these two cases that the L_2 norm of difference between the estimated function and the truth converges at rate $O_P(n^{-1/3}(\log n)^{2(d-1)/3})$. This is especially meaningful in practice as asymptotic linearity and efficiency of semiparametric estimators often relies on estimating nuisance parameter at L_2 norm rate $o_P(n^{-1/4})$.

It is relatively straightforward extend the results of this chapter in several directions: firstly giving high probability bounds instead of rates of convergence in probability, extending the results to other common losses such as the hinge loss or the negative likelihood loss, studying the effect of margin bounds. These are direct consequences of classical statistical learning theory results on empirical risk minimization (see e.g Bartlett et al. [2006]). Another further extension of the present work is to give guarantees for the cross-validated selection of the sectional variation norm via the Super Learner. We do several of these extensions in subsequent chapters. In particular, we provide high probability bounds under various losses for different types of non independent data in chapters 5 and 7.

Bibliography

P. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Aurélien F. Bibaut and Mark J. van der Laan. Fast rates for empirical risk minimization over càdlàg functions with bounded sectional variation norm, 2019.

Billy Fang, Adityanand Guntuboyina, and Bodhisattva Sen. Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and hardy-krause variation. 2019.

Fuchang Gao. Bracketing entropy of high dimensional distributions. In *High Dimensional Probability VI*, volume 66. Birkhäuser, 2013.

Richard D. Gill, Mark J. van der Laan, and Jon A. Wellner. Inefficient estimators of the bivariate survival function for three models. *Annales de l'Institut Henri Poincaré*, pages 545–597, 1995.

Mark J. van der Laan. A generally efficient tmle. *The International Journal of Biostatistics*, 2016.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

3.A Proof of the bracketing entropy bound (proposition 3.2)

The proof of proposition 3.2 relies on the representation of càdlàg functions with bounded sectional variation norm and on the the three results below. For all $d \geq 1$, $M > 0$, denote

$$\mathcal{F}_{d,M} = \{f \in \mathbb{D}([0,1]^d) : \|f\|_v \leq M\}.$$

The first result characterizes the bracketing entropy of the set of d-dimensional cumulative distribution functions.

Lemma 3.1 (Theorem 1.1 in Gao [2013]). *Let $\mathcal{G}_d$ be the set of probability distributions on $[0,1]^d$. For $1 \leq r < \infty$ and $d \geq 2$,*

$$\log N_{[]}(\epsilon, \mathcal{G}_d, \|\cdot\|_{\mu,r}) \leq C'(d,r)\epsilon^{-1}|\log \epsilon|^{2(d-1)}$$

for some constant $C'(d,r)$ that only depends on d and r, and where μ is the Lebesgue measure on $[0,1]^d$.

As an immediate corollary, the following result holds for bracketing numbers w.r.t. $\|\cdot\|_{P_0,2}$.

Corollary 3.3. *Let $1 \leq r < \infty$. Suppose there exists a constant $c(r)$ such that $\|\cdot\|_{P_0,r} \leq c(r)\|\cdot\|_{\mu,r}$. Then,*

$$\log N_{[]}(\epsilon, \mathcal{G}_d, \|\cdot\|_{P_0,r}) \leq \tilde{C}(d,r)\epsilon^{-1}|\log \epsilon|^{2(d-1)}$$

for some constant $\tilde{C}(r,d)$ that only depends on r and d.

The next lemma will be useful to bound the bracketing entropy integral.

Lemma 3.2. *For any $d \geq 0$ and any $0 < \delta \leq 1$, we have that*

$$\int_0^\delta \epsilon^{-1/2}(\log(1/\epsilon))^{d-1} d\epsilon \lesssim \delta^{1/2}(\log(1/\delta))^{d-1}.$$

Proof. The result is readily obtained by integration by parts. □

We can now present the proof of proposition 3.2.

Proof. We will first upper bound the $(\epsilon, \|\cdot\|_{P_0,r})$-bracketing number for $\mathcal{F}_{d,1}$. An upper bound on the $(\epsilon, \|\cdot\|_{P_0,r})$-bracketing number for $\mathcal{F}_{d,M}$ will then be obtained at the end of the proof by means of change of variable. Recall that any function in $\mathcal{F}_{d,1}$ can be written as

$$f = \sum_{s \subseteq [d]} \alpha_{s,1} g_{s,1} - \alpha_{s,2} g_{s,2},$$

with $g_{\emptyset,1} = g_{\emptyset,2} = 1$, and for all $\emptyset \neq s \subseteq [d]$, $g_{s,1}, g_{s,2} \in \mathcal{G}_s$, and $\boldsymbol{\alpha} = (\alpha_{s,i} : s \subseteq [d], i = 1,2) \in \Delta^{2^{d+1}}$, where $\Delta^{2^{d+1}}$ is the 2^{d+1}-standard simplex.

Let $\epsilon > 0$. Denote $N(\epsilon/2^{d+1}, \Delta^{2^{d+1}}, \|\cdot\|_\infty)$ the $(\epsilon/2^{d+1}, \|\cdot\|_\infty)$-covering number of $\Delta^{2^{d+1}}$. Let

$$\{\boldsymbol{\alpha}^{(j)} : j = 1, ..., N(\epsilon/2^{d+1}, \Delta^{2^{d+1}}, \|\cdot\|_\infty)\}$$

be an $(\epsilon/2^{d+1}, \|\cdot\|_\infty)$-covering of $\Delta^{2^{d+1}}$. For all $s \subseteq [d]$, denote $N_{[]}(\epsilon, \mathcal{G}_s, \|\cdot\|_{P_0,r})$ the $(\epsilon, \|\cdot\|_{P_0,r})$-bracketing number of $\mathcal{G}_s$, and let

$$\{(l_s^{(j)}, u_s^{(j)}) : j = 1, ..., N_{[]}(\epsilon, \mathcal{G}_s, \|\cdot\|_{P_0,r})\}$$

be an $(\epsilon, \|\cdot\|_{P,r})$-bracketing of $\mathcal{G}_s$.

Step 1: Construction of a bracket for $\mathcal{F}_{d,1}$. We now construct a bracket for f from the cover for $\Delta^{2^{d+1}}$ and the bracketings for $\mathcal{G}_s, \emptyset \neq s \subseteq [d]$, we just defined. By definition of an $(\epsilon/2^{d+1}, \|\cdot\|_\infty)$-cover, there exists $j_0 \in \{1, ..., N(\epsilon/2^{d+1}, \Delta^{2^{d+1}}, \|\cdot\|_{P_0,r})\}$ such that $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^{(j_0)}\|_\infty \leq \epsilon/2^{d+1}$. Consider $s \subseteq [d]$, $i \in \{1, 2\}$. By definition of an $(\epsilon, \|\cdot\|_{P_0,r})$-bracket exists $j_{s,i} \in \{1, ..., N_{[]}(\epsilon, \mathcal{G}_s, \|\cdot\|_{P,r})$ such that

$$l_s^{(j_{s,i})} \leq g_{s,i} \leq u_s^{(j_{s,i})}.$$

This and the fact that

$$\alpha_{s,i}^{(j_0)} - \epsilon/2^{d+1} \leq \alpha_{s,i} \leq \alpha_{s,i}^{(j_0)} + \epsilon/2^{d+1},$$

will allow us to construct a bracket for $\alpha_{s,i}g_{s,i}$. Some care has to be taken due to the fact $l_s^{j_{s,i}}$ can be negative (as bracketing functions do not necessarily belong to the class they bracket). Observe that, since $\alpha_{s,i} \geq 0$, we have

$$\alpha_{s,i} l_s^{(j_{s,i})} \leq \alpha_{s,i} g_{s,i} \leq \alpha_{s,i} u_s^{(j_{s,i})}.$$

Denoting $(l_s^{(j_{s,i})})^+$ and $(l_s^{(j_{s,i})})^-$ the positive and negative part of $l_s^{(j_{s,i})}$, we have that

$$\begin{aligned} (\alpha_{s,i}^{(j_0)} - \epsilon/2^{d+1})(l_s^{(j_{s,i})})^+ &\leq \alpha_{s,i} l_s^+ \\ \text{and } -(\alpha_{s,i}^{(j_0)} + \epsilon/2^{d+1})(l_s^{(j_{s,i})})^- &\leq -\alpha_{s,i} l_s^-. \end{aligned}$$

Therefore,

$$\alpha_{s,i}^{(j_0)} l_s^{(j_{s,i})} - \epsilon/2^{d+1}|l_s^{(j_{s,i})}| \leq \alpha_{s,i} l_s^{(j_{s,i})}.$$

Since $u_{s,i}^{(j_{s,i})} \geq 0$ (at it is above at least one cumulative distribution function from $\mathcal{G}_s$), and $\alpha_{s,i}^{(j_0)} + \epsilon/2^{d+1} \geq \alpha_{s,i}$, we have that

$$\alpha_{s,i} g_{s,i} \leq (\alpha_{s,i}^{(j_0)} + \epsilon/2^{d+1}) u_s^{(j_{s,i})}.$$

Therefore, we have shown that

$$\alpha_{s,i}^{(j_0)} l_s^{(j_{s,i})} - \epsilon/2^{d+1}|l_s^{(j_{s,i})}| \leq \alpha_{s,i} g_{s,i} \leq (\alpha_{s,i}^{(j_0)} + \epsilon/2^{d+1}) u_s^{(j_{s,i})}.$$

Summing over $s \subset \{1, ..., d\}$ and $i = 1, 2$, we have that

$$\Lambda_1 - \Gamma_2 \leq f \leq \Gamma_1 - \Lambda_2,$$

where, for $i = 1, 2$,

$$\Lambda_i = \sum_{s \subseteq [d]} \alpha_{s,i}^{(j_0)} l_s^{j_{s,i}} - \epsilon/2^{d+1} |l_s^{j_{s,i}}|,$$
$$\text{and } \Gamma_i = \sum_{s \subseteq [d]} (\alpha_{s,i}^{(j_0)} + \epsilon/2^{d+1}) u_s^{j_{s,i}}.$$

Step 2: Bounding the size of the brackets. For $i = 1, 2$,

$$0 \leq \Gamma_i - \Lambda_i = \sum_{s \subseteq [d]} \alpha_{s,i}^{(j_0)} (u_s^{j_{s,i}} - l_s^{j_{s,i}}) + \epsilon/2^{d+1} (u_s^{j_{s,i}} + |l_s^{j_{s,i}}|).$$

Since, for every $s \subseteq [d]$, $i = 1, 2$ $u_s^{j_{s,i}}$ and $l_s^{j_{s,i}}$ are at most ϵ-away in $\| \cdot \|_{P,r}$ norm from a cumulative distribution function, we have that $\|u_s^{j_{s,i}}\|_{P,r} \leq 1 + \epsilon$ and $\|l_s^{j_{s,i}}\|_{P,r} \leq 1 + \epsilon$. By definition, for all $s \subseteq [d]$, $i = 1, 2$, $\|u_s^{j_{s,i}} - l_s^{j_{s,i}}\|_{P,r} \leq \epsilon$. Therefore, from the triangle inequality,

$$\|\Gamma_i - \Lambda_i\|_{P,r} \leq \epsilon \sum_{s \in \subseteq [d]} \alpha_{s,i} + \epsilon(1 + \epsilon).$$

Therefore, using the triangle inequality one more time,

$$\begin{aligned} \|\Gamma_1 - \Lambda_2 - (\Lambda_1 - \Gamma_2)\|_{P,r} &\leq \epsilon \sum_{s \subseteq [d]} \alpha_{s,1} + \alpha_{s,2} + 2\epsilon(1 + \epsilon) \\ &\leq 3\epsilon + 2\epsilon^2. \end{aligned}$$

Since cumulative distribution functions have range $[0, 1]$, brackets never need to be of size larger than 1. Therefore, without loss of generality, we can assume that $\epsilon \leq 1$. Therefore, pursuing the above display, we get

$$|\Gamma_1 - \Lambda_2 - (\Lambda_1 - \Gamma_2)\|_{P,r} \leq 5\epsilon.$$

Step 3: Counting the brackets. Consider the set of brackets of the form $(\Gamma_1 - \Lambda_2, \Lambda_1 - \Gamma_2)$, where, for $i = 1, 2$,

$$\Lambda_i = \sum_{s \subseteq [d]} \alpha_{s,i}^{(j_0)} l_s^{j_{s,i}} - \epsilon/2^{d+1} |l_s^{j_{s,i}}|,$$
$$\text{and } \Gamma_i = \sum_{s \subseteq [d]} (\alpha_{s,i}^{(j_0)} + \epsilon/2^{d+1}) u_s^{j_{s,i}},$$

where $j_0 \in \{1, ..., N(\epsilon/2^{d+1}), \Delta^{2^{d+1}}, \| \cdot \|_\infty)\}$ and for any s, i $j_{s,i} \in \{1, ..., N_{[]}(\epsilon, \mathcal{G}_s, \| \cdot \|_{P_0,r})\}$. From step 1 and step 2, we know that this set of brackets is a $(5\epsilon, \| \cdot \|_{P_0,r})$-bracketing of $\mathcal{F}_1$. Its cardinality is no larger than the cardinality of its index set. Therefore

$$N_{[]}(5\epsilon, \mathcal{F}_1, \| \cdot \|_{P,r}) \leq N(\epsilon/2^{d+1}, \Delta^{2^{d+1}}, \| \cdot \|_\infty) \prod_{s \subseteq [d]} N_{[]}(\epsilon, \mathcal{G}_s, \| \cdot \|_{P_0,r})^2.$$

The covering number of the simplex can be bounded (crudely) as follows:

$$N(\epsilon/2^{d+1}, \Delta^{2^{d+1}}, \| \cdot \|_\infty) \leq \left(\frac{2^{d+1}}{\epsilon} \right)^d$$

therefore

$$\log N(\epsilon/2^{d+1}, \Delta^{2^{d+1}}, \| \cdot \|_\infty) \leq d \log(1/\epsilon) + d(d+1) \log 2.$$

From corollary 3.3,

$$\log N_{[]}(\epsilon, \mathcal{G}_s, \| \cdot \|_{P_0,r}) \leq C(r,d) \epsilon^{-1} | \log \epsilon |^{2(d-1)}.$$

Therefore,

$$\begin{aligned} \log N_{[]}(5\epsilon, \mathcal{F}_1, \| \cdot \|_{P_0,r}) &\leq \tilde{C}(r,d) 2^{d+2} \epsilon^{-1} | \log \epsilon |^{2(d-1)} + d \log(1/\epsilon) + d(d+1) \log 2 \\ &\lesssim \tilde{C}(r,d) 2^{d+2} \epsilon^{-1} | \log \epsilon |^{2(d-1)}. \end{aligned}$$

Therefore, doing a change of variable, (and for a different constant absorbed in the $\lesssim$ symbol),

$$\log N_{[]}(\epsilon, \mathcal{F}_M, \| \cdot \|_{P_0,r}) \lesssim \tilde{C}(r,d) 2^{d+2} M \epsilon^{-1} | \log(\epsilon/M) |^{2(d-1)}.$$

The wished claims hold for $C(r,d) = 2^{d+2} \tilde{C}(r,d)$. □

3.B Proofs of theorem 3.1 and preliminary results

3.B.1 Overview and preliminary lemmas

The proof of the theorem relies on theorem 3.4.1 in van der Vaart and Wellner [1996], which gives an upper bound on the rate of convergence of the estimator in terms of the “modulus of continuity” of an empirical process indexed by a difference in loss functions. We bound this “modulus of continuity” by using a maximal inequality for this empirical process. This maximal inequality is expressed in terms of the bracketing entropy integrals of the class of function $\mathcal{L}_n = \{L(\theta) - L(\theta)_n : \theta \in \Theta_n\}$. We link the bracketing entropy of $\mathcal{L}_n$ to the one of Θ_n through lemma 3.4.

We first restate here the theorem 3.4.1. in van der Vaart and Wellner [1996].

Theorem 3.3 (Theorem 3.4.1 in van der Vaart and Wellner [1996]). *For each n, let $\mathbb{M}_n$ and M_n be stochastic processes indexed by a set Θ. Let $\theta_n \in \Theta$ (possibly random) and $0 \leq \delta_n \leq \eta$ be arbitrary, and let $\theta \mapsto d_n(\theta, \theta_n)$ be an arbitrary map (possibly random) from Θ to $[0, \infty)$. Suppose that, for every n and $\delta_n \leq \delta \leq \eta$,*

$$\sup_{\substack{\theta \in \Theta_n \\ \delta/2 \leq d_n(\theta, \theta_n) \leq \delta}} M_n(\theta) - M_n(\theta_n) \leq -\delta^2,$$

$$E^* \sup_{\substack{\theta \in \Theta_n \\ \delta/2 \leq d_n(\theta, \theta_n) \leq \delta}} \sqrt{n}[(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_n)]^+ \lesssim \phi_n(\delta),$$

for functions ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing on (δ_n, η) for some $\alpha < 2$. Let $r_n \lesssim \delta_n^{-1}$ satisfy

$$r_n^2 \phi_n\left(\frac{1}{r_n}\right) \leq \sqrt{n}, \textit{ for every } n.$$

If the sequence $\hat{\theta}_n$ takes its values in Θ_n and satisfies

$$\mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_n) - O_P(r_n^{-2})$$

and $d_n(\theta, \theta_n)$ converges to zero in outer probability, then $r_n d_n(\hat{\theta}_n, \theta_n) = O_P^*(1)$. If the displayed conditions are valid for $\eta = \infty$, then the condition that θ_n is consistent is unnecessary.

The quantity $\phi_n(\delta)$ is the so-called "modulus of continuity" of the centered process $\sqrt{n}(\mathbb{M}_n - M_n)$ over $\Theta_n = \mathcal{F}_n$. Theorem 3.3 essentially teaches us that the rate of the modulus of continuity gives us the (an upper bound on) the rate of convergence of the estimator.

We now restate the maximal inequality that we will use to bound the modulus of continuity.

Lemma 3.3 (Lemma 3.4.2 in van der Vaart and Wellner [1996]). *Let $\mathcal{F}$ be a class of measurable functions such that $Pf^2 < \delta^2$ and $\|f\|_\infty \leq M$ for every $f \in \mathcal{F}$. Then*

$$E_P^* \sup_{f \in \mathcal{F}} \sqrt{n}|(P_n - P)f| \lesssim J_{[]}(\delta, \mathcal{F}, L_2(P)) \left(1 + \frac{J_{[]}(\delta, \mathcal{F}, L_2(P))}{\delta^2 \sqrt{n}} M\right).$$

Application of the above maximal inequality is what will allow us to bound the "modulus of continuity". The following lemma will be useful to upper bound the entropy integral of $\mathcal{L}_n = \{L(\theta) - L(\theta_n) : \theta \in \Theta_n\}$ in terms of the entropy integral of Θ_n.

Lemma 3.4. *Let $F : \mathbb{R} \times \mathcal{W} \to \mathbb{R}$ a mapping such that, for any $w \in \mathcal{W}$, there exists $a_w \in \mathbb{R}$ such that $a \mapsto F(a, w)$ is*

- *non-increasing on $(-\infty, a_w]$,*
- *non-decreasing on $[a_w, \infty)$,*

- *M-Lipschitz for some M that does not depend on w.*

Let $\mathcal{A}$ a set of real-valued functions defined on a set $\mathcal{V}$. Let

$$\mathcal{B} = \{(v, w) \in \mathcal{V} \times \mathcal{W} \mapsto F(a(v), w) : a \in \mathcal{A}\}.$$

Let $r \geq 1$, and let P a probability distribution over $\mathcal{V} \times \mathcal{W}$. Then, for any $\delta > 0$,

$$N_{[]}(\delta, \mathcal{B}, \| \cdot \|_{P_0,r}) \leq N_{[]}(\delta/M, \mathcal{A}, \| \cdot \|_{P_0,r}),$$

and

$$J_{[]}(\delta, \mathcal{B}, \| \cdot \|_{P_0,r}) \leq M J_{[]}(\delta/M, \mathcal{A}, \| \cdot \|_{P_0,r}).$$

(Note that the above quantities might not be finite.)

We defer the proof of this lemma to subsection 3.B.3. The following lemma shows that the variation norm dominates the supremum norm.

Lemma 3.5. *For all $f \in \mathbb{D}([0, 1]^d)$,*

$$\|f\|_\infty \leq \|f\|_v.$$

3.B.2 Proof of theorem 3.1

We now present the proof of theorem 3.1.

Proof of theorem 3.1. The proof essentially consists of checking the assumptions of theorem 3.3 for a certain choice of $\mathbb{M}_n$, M_n, d_n and r_n. Specifically, we set, for every $\theta \in \Theta_n$, and every n,

$$\begin{aligned}
\mathbb{M}_n(\theta) &= -P_n L(\theta), \\
M_n(\theta) &= -P_0 L(\theta), \\
\theta_n &= \arg\min_{\theta \in \Theta_n} P_0 L(\theta), \\
d_n^2(\theta, \theta_n) &= P_0 L(\theta) - P_0 L(\theta_n), \\
r_n &= C(r, d)^{-1/3} a_n^{-1} n^{1/3} (\log n)^{-2(d-1)/3}.
\end{aligned}$$

Further set $\eta = \infty$ and $\delta_n = 0$. From now, we proceed in three steps.

Step 1: Checking condition 3.3. By definition of M_n and by definition of the loss-based dissimilarity, we directly have, for every $\theta \in \Theta_n$,

$$M_n(\theta) - M_n(\theta_n) = -P_0(L(\theta) - L(\theta_n)) = -d_n^2(\theta, \theta_n).$$

Therefore, condition 3.3 holds.

Step 2: Bounding the modulus of continuity. We want to bound

$$\begin{aligned}
&E_{P_0} \sup_{\substack{\theta\in\Theta_n\\ d_n(\theta,\theta_n)\leq\delta}} |(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_n)| \\
&= E_{P_0} \sup_{\substack{\theta\in\Theta_n\\ d_n(\theta,\theta_n)\leq\delta}} |(P_n - P_0)(L(\theta) - L(\theta_n))| \\
&= E_{P_0} \sup_{g\in\mathcal{G}_n(\delta)} |(P_n - P_0)g|, \qquad (3.1)
\end{aligned}$$

where

$$\mathcal{G}_n(\delta) = \{L(\theta) - L(\theta_n) : \theta \in \Theta_n, d_n(\theta,\theta_n) \leq \delta\}.$$

We now further characterize the set $\mathcal{G}_n(\delta)$. From assumption 3.3, for all $\theta \in \Theta_n$, $\|L(\theta) - L(\theta_n)\|_{P_0,2} \leq a_n d_n(\theta,\theta_n)$. Therefore, denoting $\mathcal{L}_n = \{L(\theta) - L(\theta_n) : \theta \in \Theta_n\}$ and $\mathcal{L}_n(\delta) = \{g \in \mathcal{L} : \|g\|_{P_0,2} \leq \delta\}$, we have that $\mathcal{G}_n(\delta) \subseteq \mathcal{L}_n(a_n\delta)$. We now turn to bounding in supremum norm the class $\mathcal{L}_n$. From assumption 3.2, for all $\theta \in \Theta_n$, $\|L(\theta) - L(\theta_n)\|_\infty \leq a_n\|\theta - \theta_n\|_\infty$. From the definition of Θ_n and lemma 3.5, we have that, for all $\theta \in \Theta_n$, $\|\theta - \theta_n\|_\infty \leq 2a_n$, which implies that $\|L(\theta) - L(\theta_n)\|_\infty \leq 2a_n^2$. Therefore, from (3.1) and the maximal inequality of lemma 3.3, we have

$$\begin{aligned}
&E_{P_0} \sup_{\substack{\theta\in\Theta_n\\ d_n(\theta,\theta_n)\leq\delta}} |(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_n)| \\
&\leq E_{P_0} \sup_{g\in\mathcal{L}_n(a_n\delta)} |(P_n - P_0)g| \\
&\leq \frac{\phi_n(\delta)}{\sqrt{n}},
\end{aligned}$$

with

$$\phi_n(\delta) \equiv J_{[]}(a_n\delta, \mathcal{L}_n, \|\cdot\|_{P_0,2}) \left(1 + \frac{J_{[]}(a_n\delta, \mathcal{L}_n, \|\cdot\|_{P_0,2})}{(a_n\delta)^2\sqrt{n}} 2a_n^2\right).$$

Step 3: Checking the rate condition $r_n^2\phi_n(1/r_n) \leq \sqrt{n}$. From lemma 3.4, and then from proposition 3.2,

$$\begin{aligned}
J_{[]}(a_n\delta, \mathcal{L}_n, \|\cdot\|_{P_0,2}) &\lesssim a_n J_{[]}(\delta, \Theta_n, L_2(P_0)) \\
&\lesssim a_n C(r,d)^{1/2} a_n^{1/2} \delta^{1/2} (\log(a_n/\delta))^{d-1} \\
&\lesssim C(r,d)^{1/2} a_n^{3/2} \delta^{1/2} (\log(a_n/\delta))^{d-1}.
\end{aligned}$$

Recall that we set

$$r_n = C(r,d)^{-1/3} a_n^{-1} n^{1/3} (\log n)^{-2(d-1)/3}.$$

Since we supposed that $a_n = O(n^p)$ for some $p > 0$, we have that $\log(a_n r_n) \lesssim \log n$. Therefore,

$$\begin{aligned}
&r_n^2 \phi_n(1/r_n) \\
\lesssim & r_n^2 C(r,d)^{1/2} a_n^{3/2} r_n^{-1/2} (\log(a_n r_n))^{d-1} \left(1 + \frac{C(r,d)^{1/2} a_n^{3/2} r_n^{-1/2} (\log(a_n r_n))^{d-1}}{(a_n/r_n)^2 \sqrt{n}} 2a_n^2\right) \\
\lesssim & C(r,d)^{1/2} a_n^{3/2} r_n^{3/2} (\log n)^{d-1} \left(1 + 2\frac{C(r,d)^{1/2} a_n^{3/2} r_n^{3/2} (\log n)^{d-1}}{\sqrt{n}}\right) \\
\lesssim & 3\sqrt{n}.
\end{aligned}$$

□

3.B.3 Proof of technical lemmas 3.4 and 3.5

Proof of lemma 3.4. Let $[l, u]$ an $(\epsilon, \| \cdot \|_{P,r})$-bracket for $\mathcal{A}$ and let $a \in \mathcal{A}$ such that $a \in [l, u]$. Define, for all $(v, w) \in \mathcal{V} \times \mathcal{W}$,

$$\Lambda(v,w) = \begin{cases} F(a_w, w) & \text{if } l(v) \leq a_w \leq u(v), \\ F(l(v), w) \wedge F(u(v), w) & \text{otherwise,} \end{cases}$$

and

$$\Gamma(v,w) = F(l(v), w) \vee F(u(v), w).$$

We claim that (Λ, Γ) is an $(M\epsilon, \| \cdot \|_{P,r})$-bracket for $(u, v) \mapsto F(a(v), w)$. We distinguish three cases. Let $(u, v) \in \mathcal{V} \times \mathcal{W}$.

Case 1. Suppose that $l(v) \leq a_w \leq u(v)$. Then since $a \mapsto F(a, w)$ reaches its minimum in a_w, we have that $\Lambda(v, w) = F(a_w, w) \leq F(a(v), w)$. If $a(v) \in [a_w, u(v)]$, then, as $a \mapsto F(a, w)$ is non-decreasing on $[a_w, \infty)$, we have that $F(a(v), w) \leq F(u(v), w)$. If $a(v) \in [l(v), a_w]$, then, as $a \mapsto F(a, w)$ is non-increasing on $(-\infty, a_w]$, $F(a(v), w) \leq F(l(v), w)$. Thus $F(a(v), w) \leq F(l(v), w) \vee F(u(v), w) = \Gamma(v, w)$.

Observe that, under $[a_w \in [l(v), u(v)]$, we have that $|l(v) - a_w| \leq |u(v) - l(v)|$ and $|u(v) - a_w| \leq |u(v) - l(v)|$. Therefore, if $\Gamma(v, w) = F(u(v), w)$,

$$|\Gamma(v,w) - \Lambda(v,w)| = |F(u(v), w) - F(a_w, w)| \leq M|u(v) - a_w| \leq M|u(v) - l(v)|.$$

Case 2. Suppose that $a_w \leq l(v) \leq u(v)$. Then, as $a \mapsto F(a, v)$ is non-decreasing on $[a_w, \infty)$,

$$\Lambda(u,v) = F(l(v), w) \leq F(a(v), w) \leq F(u(v), w) = \Gamma(u,v),$$

and $|\Gamma(u, v) - \Lambda(u, v)| \leq M|u(v) - l(v)|$.

Case 3. Suppose that $l(v) \leq u(v) \leq a_w$. Then, as $a \mapsto F(a, v)$ is non-increasing on $(-\infty, a_w]$,

$$\Lambda(u, v) = F(u(v), w) \leq F(a(v), w) \leq F(l(v), w) = \Gamma(u, v),$$

and $|\Gamma(u, v) - \Lambda(u, v)| \leq M|u(v) - l(v)|$.

We have thus shown that, for all $(v, w) \in \mathcal{V} \times \mathcal{W}$,

$$\Lambda(v, w) \leq F(a(v), w) \leq \Gamma(v, w),$$

and

$$\Gamma(v, w) - \Lambda(v, w)| \leq M|u(v) - l(v)|.$$

By integration of the above display, we have that

$$\|\Gamma - \Lambda\|_{P,r} \leq M\|u - l\|_{P,r}.$$

Therefore, we have shown that an $(\epsilon, \| \cdot \|_{P,r})$-bracket for $\mathcal{A}$ induces an $(M\epsilon, \| \cdot \|_{P,r})$-bracket for $\mathcal{B}$. Therefore, for all $\epsilon > 0$,

$$N_{[]}(\epsilon, \mathcal{B}, \| \cdot \|_{P,r}) \leq N_{[]}(\epsilon/M, \mathcal{A}, \| \cdot \|_{P,r}),$$

and, for all $\delta > 0$,

$$\begin{aligned} J_{[]}(\delta, \mathcal{B}, \| \cdot \|_{P,r}) &\leq \int_0^\delta \sqrt{\log N_{[]}(\epsilon/M, \mathcal{A}, \| \cdot \|_{P,r})} d\epsilon \\ &\leq M \int_0^{\delta/M} \sqrt{\log N_{[]}(\zeta, \mathcal{A}, \| \cdot \|_{P,r})} d\zeta \\ &= M J_{[]}(\delta/M, \mathcal{A}, \| \cdot \|_{P,r}). \end{aligned}$$

□

Proof of lemma 3.5. Let $x \in [0, 1]^d$. From the representation formula in proposition 3.1,

$$f(x) = f(0) + \sum_{\emptyset \neq s \subseteq [d]} \int_{[0_s, x_s]} f(dx_s).$$

Therefore,

$$\begin{aligned} |f(x)| &\leq |f(0)| + \sum_{\emptyset \neq s \subseteq [d]} \int_{[0_s, x_s]} |f(dx_s)| \\ &\leq |f(0)| + \sum_{\emptyset \neq s \subseteq [d]} \int_{[0_s, 1_s]} |f(dx_s)| \\ &= \|f\|_v. \end{aligned}$$

By taking the sup with respect to x, we obtain the wished result. □

3.C Proof of propositions of section 3.4

3.C.1 Proof of results on least-squares with bounded dependent variable

Proof of proposition 3.3. Let $y \in [-\tilde{a}_n, \tilde{a}_n]$. It is clear that $u \mapsto \tilde{L}(u, y) = (y - u)^2$ is non-increasing on $(-\infty, y]$ and non-decreasing on $[y, \infty)$.

We now turn to showing the Lipschitz property claim. Observe that

$$\mathcal{U}_n \equiv \{\theta(x) : \theta \in \Theta_n, x \in [0,1]^d\} \subseteq [-\tilde{a}_n, \tilde{a}_n].$$

Let $u_1, u_2 \in \mathcal{U}_n$. We have that

$$\begin{aligned}|\tilde{L}(u_1, y) - \tilde{L}(u_2, y)| =& |(y - u_2)^2 - (y - u_1)^2| \\ =& |2y - u_1 - u_2||u_2 - u_1| \\ \leq& 4\tilde{a}_n |u_2 - u_1|,\end{aligned}$$

which is the wished claim. □

The proof of proposition 3.4 requires the following lemma.

Lemma 3.6. *Consider Θ_n, θ_n, θ_0, and d_n as defined in subsection 3.4.1. Then, for all $\theta \in \Theta$,*

$$d_n^2(\theta, \theta_n) = \|\theta - \theta_0\|_{P_0,2}^2 - \|\theta - \theta_n\|_{P_0,2}^2 \geq \|\theta - \theta_0\|_{P_0,2}^2.$$

Proof. It is straighforward to check that Θ_n is a closed convex set. Denote, for all θ_1, θ_2, $\langle \theta_1, \theta_2 \rangle = E_{P_0}[\theta_1(X)\theta_2(X)]$. Observe that, for all $\theta \in \Theta_n$, $\|\theta\|_{P_0,2}^2 \langle \theta, \theta \rangle$. Let $\theta \in \Theta_n$. We have that

$$\begin{aligned}d_n^2(\theta, \theta_n) =& E_{P_0}[(Y - \theta(X))^2] - E_{P_0}[(Y - \theta_n(X))^2] \\ =& E_{P_0}[(Y - \theta_0(X))^2] + E_{P_0}[(\theta_0(X) - \theta(X))^2] \\ & - \left\{ E_{P_0}[(Y - \theta_0(X))^2] + E_{P_0}[(\theta_0(X) - \theta_n(X))^2] \right\} \\ =& \|\theta - \theta_0\|_{P_0,2}^2 - \|\theta_n - \theta_0\|_{P_0,2}^2.\end{aligned}$$

Therefore,

$$\begin{aligned}d_n^2(\theta, \theta_n) - \|\theta - \theta_n\|_{P_0,2}^2 =& \|(\theta - \theta_n) + (\theta - \theta_n)\|_{P_0,2}^2 - \|\theta_n - \theta_0\|_{P_0,2}^2 - \|\theta - \theta_0\|_{P_0,2}^2 \\ =& -2\langle \theta - \theta_n, \theta_0 - \theta_n \rangle \\ \geq& 0.\end{aligned}$$

The last line follows from the fact that $\theta \in \Theta_n$ and that θ_n is the projection for the $\|\cdot\|_{P_0,2}$ of θ_0 onto the closed convex set Θ_n. □

We can now state the proof of proposition 3.4.

Proof of proposition 3.4. From proposition 3.3, for all $o = (x, y) \in [0,1]^d \times [-\tilde{a}_n, \tilde{a}_n]$, $|L(\theta)(o) - L(\theta)(o)| \leq |\theta(x) - \theta_n(x)|$. Therefore, by integration

$$\begin{aligned}\|L(\theta) - L(\theta_n)\|_{P_0,2} \leq& 4\tilde{a}_n \|\theta - \theta_n\|_{P_0,2} \\ \leq& 4\tilde{a}_n d_n(\theta, \theta_n),\end{aligned}$$

where the last line follows from lemma 3.6. □

3.C.2 Proofs of the results on logistic regression

Proof of proposition 3.5. Let $y \in \{0,1\}$. It is clear that $u \mapsto \tilde{L}(u,y)$ is non-increasing on $\mathbb{R}$ if $y = 1$ and non-decreasing on $\mathbb{R}$ if $y = 0$. Let's now turn to the Lipschitz property claim. For all $u \in \mathbb{R}$,

$$\frac{\partial \tilde{L}}{\partial u}(u,y) = \frac{1}{1+e^{-u}} - y.$$

Therefore, for all $u \in \mathbb{R}$, $y \in \{0,1\}$,

$$\left| \frac{\partial \tilde{L}}{\partial u}(u,y) \right| \leq 1,$$

which implies that $\tilde{L}$ is 1-Lipschitz in its first argument. □

Proof of proposition 3.6. For all x, denote $\eta_0(x) = E_{P_0}[Y|X = x] = (1+\exp(-\theta_0(x))^{-1}$, and $\eta_n(x) = (1+\exp(-\theta_n(x))^{-1}$. For all $p \in [0,1]$, $q \in \mathbb{R}$, denote

$$f_p(q) = p\log(1+e^{-q}) + (1-p)\log(1+e^{-q}).$$

Observe that, for all θ,

$$P_0 L(\theta) = E_{P_0}[f_{\eta_0(X)}(\theta(X))]. \tag{3.2}$$

For all $p \in [0,1]$, $q \in \mathbb{R}$, we have that

$$\begin{aligned} f_p'(q) =& \frac{1}{1+e^{-q}} - p, \\ \text{and } f_p''(q) =& \frac{1}{1+e^{-q}} \times \left(1 - \frac{1}{1+e^{-q}}\right) \\ \geq& \frac{1}{2} \times \min\left(\frac{1}{1+e^{-q}}, \frac{1}{1+e^{q}}\right). \end{aligned}$$

Therefore, for $p \in [0,1]$ and $q \in [\tilde{a}_n, \tilde{a}_n]$, we have that $f_p''(q) \geq 2^{-1}(1+e^{\tilde{a}_n})^{-1}$. From the above display, we have that, for all $x \in [0,1]^d$,

$$\begin{aligned} f_{\eta_0(x)}(\theta(x) - f_{\eta_0(x)}(\theta_n(x)) \geq& f_{\eta_0(x)}'(\theta_n(x))(\theta(x) - \theta_n(x)) + \frac{1}{4(1+e^{\tilde{a}_n})}(\theta(x) - \theta_n(x))^2 \\ =& (\eta_n(x) - \eta_0(x))(\theta(x) - \theta_n(x)) + \frac{1}{4(1+e^{\tilde{a}_n})}(\theta(x) - \theta_n(x))^2. \end{aligned}$$

Therefore, for any $\theta \in \Theta_n$, using (3.2),

$$\begin{aligned} d_n^2(\theta, \theta_n) =& P_0 L(\theta) - P_0 L(\theta_n) \\ \geq& E_{P_0}[(\eta_n(X) - \eta_0(X))(\theta(X) - \theta_n(X))] + \frac{1}{4(1+e^{\tilde{a}_n})}\|\theta - \theta_n\|_{P_0,2}^2. \end{aligned} \tag{3.3}$$

Let $\theta \in \Theta_n$. For all t, define $\tilde{\theta}(t) = \theta_n + t(\theta - \theta_n)$ and $g(t) = P_0 L(\tilde{\theta}(t))$. Since θ_n and θ are in Θ_n and that Θ_n is convex, for all $t \in [0,1]$, $\tilde{\theta}(t) \in \Theta_n$. Therefore, by definition of θ_n, for all $t \in [0,1]$, $g(t) \geq g(0)$. Thus, by taking the limit of $(g(t) - g(0))/t$ as $t \downarrow 0$, we obtain that $g'(0) \geq 0$. We now calculate $g'(0)$:

$$\begin{aligned}
g'(0) =& \frac{d}{dt}\Big\{ E_{P_0}\big[\eta_0(X) \log(1 + e^{-(\theta_n(X) + t(\theta(X) - \theta_n(X))} \\
&\qquad + (1 - \eta_0(X)) \log(1 + e^{\theta_n(X) + t(\theta(X) - \theta_n(X))})\big]\Big\}\Big|_{t=0} \\
=& E_{P_0}\big[- \eta_0(X) \frac{e^{-\theta_n(X)}}{1 + e^{-\theta_n(X)}} (\theta(X) - \theta_n(X)) \\
&\quad + (1 - \eta_0(X)) \frac{e^{\theta_n(X)}}{1 + e^{\theta_n(X)}} (\theta(X) - \theta_n(X))\big] \\
=& E_{P_0}[\{-\eta_0(X)(1 - \eta_n(X)) + (1 - \eta_0(X))\eta_n(X)\}(\theta(X) - \theta_n(X))] \\
=& E_{P_0}[(\eta_n(X) - \eta_0(X))(\theta(X) - \theta_n(X)],
\end{aligned}$$

which is equal to the first term in the right-hand side of (3.3). Therefore, as $g'(0) \geq 0$,

$$d_n^2(\theta, \theta_n) \geq \frac{1}{4(1 + e^{\tilde{a}_n})} \|\theta - \theta_n\|_{P_0,2}^2.$$

From proposition3.5, for all $o = (x, y) \in [0,1]^d \times \{0,1\}$, $|L(\theta)(o) - L(\theta_n)(o)| \leq |\theta(x) - \theta_n(x)|$, therefore, by integration,

$$\|L(\theta) - L(\theta_n)\|_{P_0,2} \leq \|\theta - \theta_n\|_{P_0,2} \leq 2(1 + e^{\tilde{a}_n})^{-1/2} d_n(\theta, \theta_n).$$

□

3.D Proof of the rate theorem for least-squares regression with sub-exponential errors

We first give an informal overview of the proof. We will proceed very similarly as in the case of the proof of the rate theorem under bounded losses, that is we will first identify $\mathbb{M}_n$, M_n, d_n that satisfy the hypothesis of theorem 3.3, and then we will bound the modulus of continuity of $\mathbb{M}_n - M_n$.

Observe that

$$\begin{aligned}
\hat{\theta}_n &= \arg\min_{\theta \in \Theta_n} \frac{1}{n} \sum_{i=1}^n (Y_i - \theta(X_i))^2 \\
&= \arg\min_{\theta \in \Theta_n} \frac{1}{n} \sum_{i=1}^n (\theta_0(X_i) - \theta(X_i) + e_i)^2
\end{aligned}$$

$$= \arg\max_{\theta \in \Theta_n} \frac{1}{n} \sum_{i=1}^{n} 2(\theta(X_i) - \theta_0(X_i))e_i - (\theta(X_i) - \theta_0(X_i))^2.$$

This motivates setting

$$\mathbb{M}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} 2(\theta - \theta_0)(X_i)e_i - (\theta - \theta_0)(X_i),$$

and, since $E_{P_0}[(\theta - \theta_0)(X_i)e_i] = 0$,

$$M_n(\theta) = -P_0(\theta - \theta_0)^2,$$

and introducing the loss-based dissimilarity d_n, defined, for all $\theta \in \Theta_n$, by

$$d_n^2(\theta, \theta_n) = -(M_n(\theta) - M_n(\theta_n))^2.$$

The main effort will then be to upper bound, for any $\delta > 0$, the quantity

$$E_{P_0} \sup_{\substack{\theta \in \Theta_n \\ d_n(\theta, \theta_n) \leq \delta}} |(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_n)|.$$

The proof relies on the following lemmas, whose proofs we defer to subsection 3.D.1.

Lemma 3.7. *For all $\theta \in \Theta_n$,*

$$\|\theta - \theta_n\|_{P_0,2} \leq d_n(\theta, \theta_n).$$

For any $\theta \in \Theta_n$, we introduce the functions $g_{1,n}(\theta)$ and $g_{2,n}(\theta)$, where, for all (x, e)

$$g_{1,n}(\theta)(x, e) = (\theta(x) - \theta_n(x))e,$$

and

$$g_{2,n}(\theta) = (\theta - \theta_n)(2\theta - \theta_n - \theta_0).$$

We will consider the following two sets:

$$\mathcal{G}_{1,n} = \{g_{1,n}(\theta) : \theta \in \Theta_n\}$$
$$\mathcal{G}_{2,n} = \{g_{2,n}(\theta) : \theta \in \Theta_n\}.$$

We will use the following version of the so-called Bernstein norm, defined for any $t > 0$ and for any function $g : (x, e) \mapsto g(x, e)$ as

$$\|g\|_{P_0,B,t}^2 = t^{-2} P\phi(tg),$$

where $\phi(x) = e^x - x - 1$. As for all i, e_i is sub-exponential with parameters (α, ν), $|e_i|$ is sub-exponential with parameters $(\alpha'(\alpha, \nu), \nu'(\alpha, \nu))$. We will shorten notations by denoting $\alpha' = \alpha'(\alpha, \nu)$ and $\nu' = \nu'(\alpha, \nu)$. The following lemma characterizes the Bernstein norm of a certain type of functions.

Lemma 3.8. *Let $f : \mathcal{X} \to \mathbb{R}$ such that $\|f\|_\infty \leq M$. Suppose that $M \geq 1$. Consider $g_1 : (x, e) \mapsto f(x)e$. Then, setting $t = (\alpha M)^{-1}$, we have*

$$\|g_1\|_{P_0,B,t} \leq \|f\|_{P_0,2} \alpha M e^{\nu^2/(4\alpha^2)}.$$

Similarly, now consider $g_2 : (x, e) \mapsto f(x)|e|$. Setting $t = (\alpha' M)^{-1}$, we have

$$\|g_2\|_{P_0,B,t} \leq \|f\|_{P_0,2} \alpha' M e^{\nu'^2/(4\alpha'^2)}.$$

This has the following immediate corollary for $g_{1,n}$. In this following result as well as in the rest of this section, we will denote $t_n = (2_a \alpha')^{-1}$.

Corollary 3.4. *We have that for all $\theta \in \Theta_n$,*

$$\|g_{1,n}(\theta)\|_{P_0,B,t_n} \leq C_n \|\theta - \theta_n\|_{P_0,2},$$

where $C_n = \tilde{C}(\alpha, \nu) a_n$, with $\tilde{C}(\alpha, \nu) = 2\alpha'(\alpha, \nu) e^{\nu'(\alpha,\nu)^2/(4\alpha'(\alpha,\nu)^2)}$.

The upcoming lemma relates the bracketing numbers in $\|\cdot\|_{P_0,B,t_n}$ norm of $\mathcal{G}_{1,n}$ to the bracketing numbers of Θ_n in $\|\cdot\|_{P_0,2}$ norm.

Lemma 3.9. *For any $\epsilon > 0$,*

$$N_{[]}(\epsilon, \mathcal{G}_{1,n}, \|\cdot\|_{P_0,B,t_n}) \leq N_{[]}(C_n^{-1}\epsilon, \Theta_n, \|\cdot\|_{P_0,2}),$$

and the bracketing entropy integral of $\mathcal{G}_{1,n}$ satisfies, for all $\delta > 0$,

$$J_{[]}(\delta, \mathcal{G}_{1,n}, \|\cdot\|_{P_0,B,t_n}) \leq C_n J_{[]}(C_n^{-1}\delta, \Theta_n, \|\cdot\|_{P_0,2}).$$

The upcoming lemma relates characterizes the $\|\cdot\|_{P_0,2}$ and the $\|\cdot\|_\infty$ norm of $g_{2,n}$ and the bracketing numbers in $\|\cdot\|_{P_0,2}$ norm of $\mathcal{G}_{2,n}$.

Lemma 3.10. *Consider $g_{2,n}$ defined above. For every $\theta \in \Theta_n$,*

$$\|g_{2,n}(\theta)\|_{P_0,2} \leq (\|\theta_0\|_\infty + 3a_n)\|\theta - \theta_n\|_{P_0,2}$$
$$\|g_{2,n}(\theta)\|_\infty \leq 2a_n(\|\theta_0\|_\infty + 3a_n),$$

and, for all $\epsilon > 0$,

$$N_{[]}(\epsilon, \mathcal{G}_{2,n}, \|\cdot\|_{P_0,2}) \leq N_{[]}((\|\theta_0\|_\infty + 3a_n)^{-1}\epsilon, \Theta_n, \|\cdot\|_{P_0,2}),$$

and, for all $\delta > 0$,

$$J_{[]}(\delta, \mathcal{G}_{2,n}, \|\cdot\|_{P_0,2}) \leq (\|\theta_0\|_\infty + 3a_n) J_{[]}((\|\theta_0\|_\infty^{-1} + 3a_n)^{-1}\delta, \Theta_n, \|\cdot\|_{P_0,2}).$$

In addition to lemma 3.3 (lemma 3.4.2 from van der Vaart and Wellner [1996]), we will use the maximal inequality of lemma 3.4.3 from van der Vaart and Wellner [1996], which we restate here.

Lemma 3.11 (Lemma 3.4.3 in van der Vaart and Wellner [1996])**.** *Let $\mathcal{F}$ be a class of measurable functions such that $\|f\|_{P,B} \leq \delta$ for every $f \in \mathcal{F}$. Then*

$$E_P^* \sup_{f \in \mathcal{F}} |\sqrt{n}(P_n - P_0) f| \leq J_{[]}(\delta, \mathcal{F}, \|\cdot\|_{P,B}) \left(1 + \frac{J_{[]}(\delta, \mathcal{F}, \|\cdot\|_{P,B})}{\delta^2 \sqrt{n}}\right).$$

We can now present the proof of theorem 3.2.

Proof. We will oragnize the proof in three steps

Step 1: Checking that $\mathbb{M}_n$, M_n, and d_n satisfy the conditions of theorem 3.3.

- By definition of d_n, for all $\theta \in \Theta_n$, $M_n(\theta) - M_n(\theta_n) = -d_n^2(\theta, \theta_n)$, therefore condition 3.3 is satisfied.
- By definition of $\hat{\theta}_n$, $\mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_n) - O_P(r_n^{-2})$.

We will apply the theorem with $\eta = \infty$.

Step 2: Bounding the modulus of continuity. We have that

$$
\begin{aligned}
& E_{P_0} \sup_{\substack{\theta \in \Theta_n \\ d_n(\theta, \theta_n) \leq \delta}} |(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_n)| \\
=& E_{P_0} \sup_{\substack{\theta \in \Theta_n \\ d_n(\theta, \theta_n) \leq \delta}} \left| \frac{2}{n} \sum_{i=1}^n (\theta(X_i) - \theta_n(X_i)) e_i + (P_n - P_0)((\theta - \theta_0)^2 - (\theta_n - \theta_0)^2) \right| \\
=& 2 E_{P_0} \sup_{\substack{\theta \in \Theta_n \\ d_n(\theta, \theta_n) \leq \delta}} |(P_n - P_0) g_{1,n}(\theta)| + E_{P_0} \sup_{\substack{\theta \in \Theta_n \\ d_n(\theta, \theta_n) \leq \delta}} |(P_n - P_0) g_{2,n}(\theta)|,
\end{aligned}
$$

with $g_{1,n}$ and $g_{2,n}$ as defined above. From lemma 3.7, for any $\theta \in \Theta_n$, $\|\theta - \theta_n\|_{P_0,2} \leq d(\theta, \theta_n)$, and from corollary 3.4 and lemma 3.10, that $\|\theta - \theta_n\|_{P_0,2} \leq \delta$ implies that $\|g_{1,n}(\theta)\|_{P_0,B,t_n} \leq C_n \delta$ and $(\|\theta_0\|_\infty + 3a_n)\delta$. Therefore, the right-hand side of the above display is upper-bounded by

$$
2 E_{P_0} \sup_{\substack{g \in \mathcal{G}_{1,n} \\ \|g\|_{P_0,B,t_n} \leq C_n \delta}} |(P_n - P_0) g| + E_{P_0} \sup_{\substack{g \in \mathcal{G}_{2,n} \\ \|g\|_{P_0,2} \leq (\|\theta_0\|_{P_0,2} + 3a_n)\delta}} |(P_n - P_0) g|,
$$

where $\mathcal{G}_{1,n}$ and $\mathcal{G}_{2,n}$ are as defined above.

From lemma 3.3 and lemma 3.11, we can bound the above display by

$$
\begin{aligned}
& J_{[]}(C_n \delta, \mathcal{G}_{1,n}, \|\cdot\|_{P,B,t_n}) \left(1 + \frac{J_{[]}(C_n \delta, \mathcal{G}_{1,n}, \|\cdot\|_{P,B,t_n})}{C_n^2 \delta^2 \sqrt{n}} \right) \\
& + J_{[]}((\|\theta_0\|_\infty + 3a_n)\delta, \mathcal{G}_{2,n}, \|\cdot\|_{P_0,2}) \\
& \times \left(1 + \frac{J_{[]}((\|\theta_0\|_\infty + 3a_n)\delta, \mathcal{G}_{2,n}, \|\cdot\|_{P_0,2}) 2a_n (\|\theta_0\|_\infty + 3a_n)}{(\|\theta_0\|_\infty + 3a_n)^2 \delta^2 \sqrt{n}} \right) \\
\leq & (J_{[]}(C_n \delta, \mathcal{G}_{1,n}, \|\cdot\|_{P,B,t_n}) + J_{[]}((\|\theta_0\|_\infty + 3a_n)\delta, \mathcal{G}_{2,n}, \|\cdot\|_{P_0,2})) \\
& \times \left(1 + \frac{J_{[]}(C_n \delta, \mathcal{G}_{1,n}, \|\cdot\|_{P,B,t_n})}{C_n^2 \delta^2 \sqrt{n}} + \frac{J_{[]}((\|\theta_0\|_\infty + 3a_n)\delta, \mathcal{G}_{2,n}, \|\cdot\|_{P_0,2}) 2a_n (\|\theta_0\|_\infty + 3a_n)}{(\|\theta_0\|_\infty + 3a_n)^2 \delta^2 \sqrt{n}} \right).
\end{aligned}
\tag{3.4}
$$

From lemma 3.9,

$$
J_{[]}(C_n \delta, \mathcal{G}_{1,n}, \|\cdot\|_{P,B,t_n}) \leq C_n J_{[]}(\delta, \Theta_n, \|\cdot\|_{P_0,2}).
$$

Therefore,

$$\frac{J_{[]}(C_n\delta, \mathcal{G}_{1,n}, \|\cdot\|_{P,B,t_n})}{C_n^2\delta^2\sqrt{n}} \leq C_n^{-1}\frac{J_{[]}(\delta, \Theta_n, \|\cdot\|_{P_0,2})}{\delta^2\sqrt{n}}.$$

From lemma 3.10,

$$J_{[]}((\|\theta_0\|_\infty + 3a_n)\delta, \mathcal{G}_{2,n}, \|\cdot\|_{P_0,2}) \leq (\|\theta_0\|_{P_0,2} + 3a_n)J_{[]}(\delta, \Theta_n, \|\cdot\|_{P_0,2}).$$

Therefore,

$$\frac{J_{[]}((\|\theta_0\|_\infty + 3a_n)\delta, \mathcal{G}_{2,n}, \|\cdot\|_{P_0,2})2a_n(\|\theta_0\|_\infty + 3a_n)}{(\|\theta_0\|_\infty + 3a_n)^2\delta^2\sqrt{n}} \leq (\|\theta_0\|_\infty + 3a_n)J_{[]}(\delta, \Theta_n, \|\cdot\|_{P_0,2}).$$

Therefore, we can bound (3.4) by

$$\begin{aligned}&(C_n + 3a_n + \|\theta_0\|_\infty)J_{[]}(\delta, \Theta_n, \|\cdot\|_{P_0,2})\left(1 + \frac{(C_n^{-1} + 3a_n + \|\theta_0\|_\infty)J_{[]}(\delta, \Theta_n, \|\cdot\|_{P_0,2})}{\delta^2\sqrt{n}}\right)\\ &\lesssim \phi_n(\delta),\end{aligned}$$

with

$$\begin{aligned}\phi_n(\delta) \equiv &((\tilde{C}(\alpha,\nu) + 3)a_n + \|\theta_0\|_\infty)J_{[]}(\delta, \Theta_n, \|\cdot\|_{P_0,2})\\ &\times\left(1 + \frac{(C_n^{-1} + 3a_n + \|\theta_0\|_\infty)J_{[]}(\delta, \Theta_n, \|\cdot\|_{P_0,2})}{\delta^2\sqrt{n}}\right).\end{aligned}$$

Step 3: Checking the rate condition. Recall that we set

$$r_n = C(r,d)^{-1/3}((\tilde{C}+3)a_n + \|\theta_0\|_\infty)^{-1}(\log n)^{-2(d-1)/3}n^{1/3}.$$

Therefore,

$$\begin{aligned}&r_n^2((\tilde{C}+3)a_n + \|\theta_0\|_\infty)J_{[]}(r_n^{-1}, \Theta_n, \|\cdot\|_{P_0,2})\\ &\lesssim r_n^2((\tilde{C}+3)a_n + \|\theta_0\|_\infty)C(r,d)^{1/2}a_n^{1/2}r_n^{-1/2}(\log(a_nr_n))^{d-1}\\ &\lesssim r_n^{3/2}((\tilde{C}+3)a_n + \|\theta_0\|_\infty)^{3/2}C(r,d)^{1/2}(\log n)^{d-1}\\ &\lesssim \sqrt{n},\end{aligned}$$

where, we used in the third line above, that since $a_n = O(n^p)$ for some $p > 0$, $\log(a_nr_n) = O(\log n)$, and in the fourth line, we replaced r_n with its expression. Therefore,

$$r_n^2\phi(1/r_n) \lesssim \sqrt{n},$$

which concludes the proof. □

3.D.1 Proofs of the technical lemmas

Proof of lemma 3.7. The proof follows easily from observing that Θ_n is convex and that θ_n is the projection on Θ_n of θ_0 for the $\|\cdot\|_{P_0,2}$ norm. □

Proof of lemma 3.8. By definition of the Bernstein norm, and using the power series expansion of ϕ, we have

$$\begin{aligned}
\|g_1\|^2_{P_0,B,t} =& t^{-2}\sum_{k=2}^{\infty} t^k \frac{P_0(f^k e^k)}{k!}\\
=& t^-2 \sum_{k=2}^{\infty} t^k \frac{P_0(f^k) P_0 e^k}{k!}\\
\leq& t^{-2}\|f\|^2_{P_0,2}\sum_{k=2}^{\infty}\frac{t^k M^{k-2} E_{P_0}[e^k]}{k!}\\
\leq& t^{-2}\|f\|^2_{P_0,2}\sum_{k=2}^{\infty}\frac{t^k M^{k} E_{P_0}[e^k]}{k!}\\
\leq& t^{-2}\|f\|^2_{P_0,2} E_{P_0}[e^{tMe}]\\
\leq& t^{-2}\|f\|^2_{P_0,2} e^{\frac{\nu^2}{2\alpha^2}}.
\end{aligned}$$

The second line in the above display follows from the fact that X and e are independent under P_0. The fourth line uses that $M \geq 1$, which implies that $M^{k-2} \leq M^k$. The sixth line uses that e is sub-exponential with parameters (α, ν). This proves the first claim.

The second claim follows by the exact same reasoning, by replacing e with $|e|$ in the above developments and using that for $t = (\alpha' M)^{-1}$, $E_{P_0}[e^{tM|e|}] \leq e^{\frac{\nu'^2}{2\alpha'^2}}$. □

Proof of lemma 3.9. . Let $\theta \in \Theta_n$ Consider $[l, u]$ an $(\epsilon, \|\cdot\|_{P_0,2})$-bracket for θ. By appropriately thresholding l and u, we can ensure that l, u have values in $[-a_n, a_n]$ while still preserving that $l \leq \theta \leq u$ and $\|l - u\|_{P_0,2} \leq \epsilon$. For all x, e, we have that

$$\Lambda(x, e) \leq (\theta(x) - \theta_n(x))e \leq \Gamma(x, e),$$

where

$$\Lambda(x, e) = (l - \theta_n)(x)e^+ + (u - \theta_n)(x)e^-,$$

and

$$\Gamma(x, e) = (u - \theta_n)(x)e^+ + (l - \theta_n)(x))e^-.$$

For all x, e,

$$\Gamma(x, e) - \Lambda(x, e) = (u - l)(x)|e|.$$

Set $t_n = (2a_n\alpha')$. From lemma 3.8, $\|\Gamma - \Lambda\|_{P_0,B,t_n} \leq 2\alpha' M e^{\nu'(\alpha,\nu)^2/(4\alpha'(\alpha,\nu)^2)}\epsilon$.

We have just shown that an $(\epsilon, \|\cdot\|_{P_0,2})$-bracketing of Θ_n induces a $(C_n\epsilon, \|\cdot\|_{P_0,B,t_n})$-bracketing of $\mathcal{G}_{1,n}$, which implies that

$$N_{[]}(\epsilon, \mathcal{G}_{1,n}, \|\cdot\|_{P_0,B,t_n}) \leq N_{[]}(C_n^{-1}\epsilon, \Theta_n, \|\cdot\|_{P_0,2}).$$

Therefore, using the above bound on the bracketing number of $\mathcal{G}_{1,n}$, and doing a change of variable in the integral, we obtain that

$$\begin{aligned} J_{[]}(\delta, \mathcal{G}_{1,n}, \|\cdot\|_{P_0,B,t_n}) &= \int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{G}_{1,n}, \|\cdot\|_{P_0,B,t_n})} d\epsilon \\ &\leq \int_0^\delta \sqrt{\log N_{[]}(C_n^{-1}\epsilon, \Theta_n, \|\cdot\|_{P_0,2})} d\epsilon \\ &\leq C_n \int_0^{C_n^{-1}\delta} \sqrt{\log N_{[]}(u, \Theta_n, \|\cdot\|_{P_0,2})} du \\ &\leq C_n J_{[]}(C_n^{-1}\delta, \Theta_n, \|\cdot\|_{P_0,2}). \end{aligned}$$

□

Proof of lemma 3.10. The first two claims are elementary.

We turn to the claim on the bracketing numbers. Let $[l, u]$ be an $(\epsilon, \|\cdot\|_{P_0,2})$-bracketing of Θ_n. Defining

$$\begin{aligned} \Lambda &= u(2\theta - \theta_0 - \theta_n)^+ - l(2\theta - \theta_0 - \theta_n)^- \\ \text{and } \Gamma &= l(2\theta - \theta_0 - \theta_n)^+ - u(2\theta - \theta_0 - \theta_n)^-, \end{aligned}$$

we have that $\Lambda \leq (\theta - \theta_n)(2\theta - \theta_0 - \theta_n) \leq \Gamma$. Observe that

$$\Gamma - \Lambda = (u - l)|2\theta - \theta_0 - \theta_n|.$$

Therefore $\|\Gamma - \Lambda\|_{P_0,2} \leq \epsilon(3a_n + \|\theta_0\|_\infty)$. This proves that an $(\epsilon, \|\cdot\|_{P_0,2})$-bracketing of Θ_n induces an $(\epsilon(\|\theta_0\|_\infty + 3a_n), \|\cdot\|_{P_0,2})$-bracketing of $\mathcal{G}_{2,n}$. From there, proceeding as in the proof of lemma 3.9 yields the claims on the bracketing number and the bracketing entropy integral. □

Chapter 4

Generalized Policy Elimination: an efficient algorithm for Nonparametric Contextual Bandits

Aurélien Bibaut, Antoine Chambaz, Mark van der Laan

In this chapter, we look into the problem of sequential decision making under the stochastic contextual bandit setting presented in subsection 1.2.1. Consistently with our objective of avoiding modelling assumptions not warranted by the available domain knowledge, we consider only nonparametric policy classes. In this chapter, we make progress toward computationally efficient contextual bandit algorithms that achieve the minimax regret rate over nonparametric policy classes.

Specifically, we propose the Generalized Policy Elimination (GPE) algorithm, which is the first to be regret optimal (up to logarithmic factors) for policy classes with integrable entropy, while only making a polynomial (in time) number of calls to optimization oracles.

For classes with larger entropy, we show that the core techniques used to analyze GPE can be used to design an ε-greedy algorithm with regret bound matching that of the best algorithms to date.

At a technical level, the key enabler of our regret analysis is a novel maximal inequality for importance sampling weighted martingale processes. On the computational side, we provide examples of nonparametric policy classes over which the relevant optimization oracles can be efficiently implemented.

4.1 Introduction

In the contextual bandit (CB) feedback model, an agent (the learner) sequentially observes a vector of covariates (the context), chooses an action among finitely many options, then receives a reward associated to the context and the chosen action. A CB algorithm is a procedure carried out by the learner, whose goal is to maximize the reward collected over time. Known as policies, functions that map any context to an action or to a distribution over actions play a key role in the CB literature. In particular, the performance of a CB algorithm is typically measured by the gap between the collected reward and the reward that would have been collected had the best policy in a certain class Π been exploited. This gap is the so-called *regret against policy class* Π. The class Π is called the *comparison class*.

The CB framework applies naturally to settings such as online recommender systems, mobile health and clinical trials, to name a few. Although the regret is defined relative to a given policy class, the goal in most settings is arguably to maximize the (expected cumulative) reward in an absolute sense. It is thus desirable to compete against large nonparametric policy classes, which are more likely to contain a policy close to the best measurable policy.

The complexity of a nonparametric class of functions can be measured by its covering numbers. The ϵ-covering number $N(\epsilon, \mathcal{F}, L_r(P))$ of a class $\mathcal{F}$ is the number of balls of radius $\epsilon > 0$ in $L_r(P)$ norm ($r \geq 1$) needed to cover $\mathcal{F}$. The ϵ-covering entropy is defined as $\log N(\epsilon, \mathcal{F}, L_r(P))$. Upper bounds on the covering entropy are well known for many classes of functions. For instance, the ϵ-covering entropy of a p-dimensional parametric class is $O(p \log(1/\epsilon))$ for all $r \geq 1$. In contrast, the ϵ-covering entropy of the class $\{f : [0,1]^d \to \mathbb{R} : \forall x, y, |f^{(\lfloor \alpha \rfloor)}(x) - f^{(\lfloor \alpha \rfloor)}(y)| \leq M\|x-y\|^{\alpha - \lfloor \alpha \rfloor}\}$[1]of d-variate Hölder functions is $O(\epsilon^{-d/\alpha})$ for $r = \infty$ (hence all $r \geq 1$) [van der Vaart and Wellner, 1996, Theorem 2.7.1]. Another popular measure of complexity is the Vapnik-

[1] $\lfloor \alpha \rfloor$ is the integer part; $f^{(m)}$ is the m-th derivative.

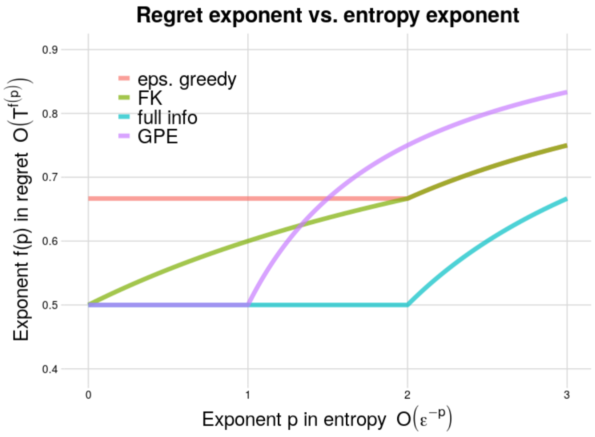


Figure 4.1: Exponent in regret upper bound (up to logarithmic factors) as a function of the exponent in the (supremum norm) covering entropy. *FK* is the theoretical upper bound of Foster and Krishnamurthy [2018]. *Full info* is the bound achieved by Empirical Risk Minimizers under full information feedback.

Chervonenkis (VC) dimension. Since the ϵ-covering entropy of a class of VC dimension V is $O(rV \log(1/\epsilon))$ for all $r \geq 1$ [van der Vaart and Wellner, 1996, Theorem 2.6.7], the complexity of a class with finite VC dimension is essentially the same as that of a parametric class.

We will consider classes Π of policies with either a *polynomial* or a *logarithmic* covering entropy, for which $\log N(\epsilon, \Pi, L_r(P))$ is either $O(\epsilon^{-p})$ for some $p > 0$ or $O(\log(1/\epsilon))$. The former are much bigger than the latter.

Efficient CB algorithms competing against classes of functions with polynomial covering entropy have been proposed [e.g. by Cesa-Bianchi et al., 2017, Foster and Krishnamurthy, 2018]. However, these algorithm are not regret-optimal in a minimax sense. In parallel, Dudik et al. [2011], Agarwal et al. [2014] have proposed efficient algorithms which are regret-optimal for finite policy classes, or for policy classes with finite VC dimension. Thus there seems to be a gap: as of today, no efficient algorithm has been proven to be regret-optimal for comparison classes with polynomial entropy (or with infinite VC dimension). In this article, we partially bridge this gap. We provide the first efficient algorithm to be regret-optimal (up to some logarithmic factors) for comparison classes with integrable entropy (that is, $\log N(\epsilon, \Pi, L_r(P)) = O(\epsilon^{-p})$ for $p \in (0, 1)$). Our main algorithm, that we name Generalized Policy Elimination (GPE) algorithm, is derived from the Policy Elimination algorithm of Dudik et al. [2011].

4.1.1 Previous work

Many contributions have been made to the area of nonparametric contextual bandits. Among others, one way to classify them is according to whether they rely on some version of the exponential weights algorithm, on optimization oracles, or on a discretization of the covariates space.

Exponential weights-based algorithms. The exponential weights algorithm has a long history in adversarial online learning, dating back to the seminal articles of Vovk [1990] and Littlestone and Warmuth [1994]. The Exp3 algorithm of Auer et al. [2002b] is the first instance of exponential weigthts for the adversarial multi-armed bandit problem. The Exp4 algorithm of Auer et al. [2002a] extends it to the contextual bandit setting. Infinite policy classes can be handled by running a version of the Exp4 algorithm on an ε-cover of the policy class. While the Exp4 algorithm enjoys optimal (in a minimax sense) regret guarantees, it requires maintaining a set of weights over all elements of the cover, and is thus intractable for most nonparametric classes, because their covering numbers typically grow exponentially in $1/\epsilon$. Cesa-Bianchi et al. [2017] proposed the first cover-based efficient online learning algorithm. Their algorithm relies on a hierarchical cover obtained by the celebrated chaining device of Dudley [1967]. It achieves the minimax regret under the full information feedback model but not under the bandit feedback model, although it yields rate improvements over past works for large nonparametric policy classes. Cesa-Bianchi et al. [2017]'s regret bounds are expressed in terms of an entropy integral. An alternative approach to nonparametric adversarial online learning is that of Chatterji et al. [2019], who proposed an efficient exponential-weights algorithm for a reproducing kernel Hilbert-space (RKHS) comparison class. They characterized the regret in terms of the eigen-decay of the kernel. They obtained optimal regret if the kernel has exponential eigen-decay.

Oracle efficient algorithms. The first oracle-based CB algorithm is the epoch-greedy algorithm of Langford and Zhang [2008]. Epoch-greedy allows to turn any supervised learning algorithm into a CB algorithm, making it practical and efficient (in terms of the number of calls to a supervised classification subroutine). Its regret can be characterized in a straighforward manner as a function of the sample complexity of the supervised learning algorithm, but is suboptimal. Dudik et al. [2011] introduced RandomizedUCB, the first regret-optimal efficient CB algorithm. Agarwal et al. [2014] improved on their work by requiring fewer calls to the oracle. [Foster et al., 2018a] pointed out that the aforementioned algorithms rely on cost-sensitive classification oracles, which are in general intractable (even though for some relatively natural classes there exist efficient algorithms). Foster et al. [2018a] proposed regret-optimal, regression oracles-based algorithms, motivated by the fact that regression oracles can in general be implement efficiently. Another way to make tractable these oracles is, in the case of cost-sensitive classification oracles, to use surrogate losses, as studied by Foster and Krishnamurthy [2018]. They gave regret upper bounds (see Figure **??**) and a nonconstructive proof of the existence of an algorithm that achieves them. They also proposed an epoch greedy-style algorithm that achieves the best regret guarantees to date for entropy $\log N(\epsilon, \Pi)$ of order ϵ^{-p} for some $p > 2$. The caveat of the surrogate loss-based approach is that guarantees are either in terms of so-called *margin-based regret*, or can be expressed in terms of the usual regret, but under the so-called realizability assumption. We refer the interested reader to Foster and Krishnamurthy [2018] for further details.

Covariate space discretization-based algorithms. A third way to design nonparametric CB algorithms consists in discretizing the context space into bins and running multi-armed bandit algorithms in each bin. This approach was pioneered by Rigollet and Zeevi [2010] and extended by Perchet and Rigollet [2013]. They take a relatively different perspective from the previously mentioned works, in the sense that the comparison class is defined in an implicit fashion: they assume that the expected reward of each action is a smooth (Hölder) function of the context, and they compete against the policy defined by the argmax over actions of the expected reward. Their regret guarantees are optimal in a minimax sense.

4.1.2 Our contributions

Primary contribution. In this article, we introduce the Generalized Policy Elimination algorithm, derived from the Policy Elimination algorithm of Dudik et al. [2011]. GPE is an oracle-efficient algorithm, of which the regret can be bounded in terms of the metric entropy of the policy class. In particular we show that if the entropy is integrable, then GPE has optimal regret, up to logarithmic factors. The key enabler of our results is a new maximal inequality for martingale processes (Theorem 4.5 in appendix 4.C), inspired by [van de Geer, 2000, van Handel, 2011]. Although our regret upper bounds for GPE are no longer optimal for policy classes with non-integrable entropy, we show that we can use the same type of martingale process techniques to design an ε-greedy type algorithm that matches the current best upper bounds.

Comparison to previous work. Earlier works on regret-optimal oracle-efficient algorithms [Dudik et al., 2011, Agarwal et al., 2014, Foster et al., 2018a, for instance] have in common that the regret analysis holds for a finite number of policies or for policy classes with finite VC dimension. GPE is the first oracle-efficient algorithm for which are proven regret optimality guarantees against a truly nonparametric policy classes (that is, larger than VC).

Secondary contributions. In addition to the nonparametric extension of policy elimination and analysis of ε-greedy in terms of (bracketing) entropy, we introduce several ideas that, to the best of our knowledge, have not appeared so far in the literature. In particular, we demonstrate the possibility of doing what we call *direct policy optimization*, that is of directly finding a maximizer $\widehat{\pi}$ of $\pi \mapsto \widehat{\mathcal{V}}(\pi)$ over Π where $\widehat{\mathcal{V}}(\pi)$ estimates the value $\mathcal{V}(\pi)$ of policy π. As far as we know, no example has been given yet of a nonparametric class Π for which $\widehat{\pi}$ can be efficiently computed, although some articles postulate the availability of $\widehat{\pi}$ [Luedtke and Chambaz, 2019, Athey and Wager, 2017]. Here, we exhibit several rich classes for which direct policy optimization can be efficiently implemented. Another secondary contribution is the first formal regret bounds for the ε-greedy algorithm, which follows from the same type of arguments as in the analysis of GPE. We were relatively surprised to see that unlike the epoch-greedy algorithm, the ε-greedy algorithm has not been formally analyzed yet, to the best of our knowledge. This may be due to the fact that doing so requires martingale process theory, which has only recently started to receive attention in the CB literature.

4.1.3 Setting

For each $m \geq 1$, denote $[m] \doteq \{1, \ldots, m\}$.

At time $t \geq 1$, the learner observes context $W_t \in \mathcal{W} \doteq [0,1]^d$, chooses an action $A_t \in [K]$, $K \geq 2$, and receives the outcome/reward $Y_t \in \{0,1\}$. We suppose that the contexts are i.i.d. and the rewards are conditionally independent given actions and contexts, with fixed conditional distributions across time points. We denote O_t the triple (W_t, A_t, Y_t), and P the distribution[2] of the infinite sequence $O_1, O_2, \ldots, O_t, \ldots$. Moreover, let $O^{\text{ref}} \doteq (W^{\text{ref}}, A^{\text{ref}}, Y^{\text{ref}})$ be a random variable such that $W^{\text{ref}} \sim W_1$, $A^{\text{ref}} | W^{\text{ref}} \sim \text{Unif}([K])$, $Y^{\text{ref}} | A^{\text{ref}}, W^{\text{ref}} \sim Y_1 | A_1, W_1$. We denote F_t the filtration induced by $O_1, \ldots, O_t$.

Generically denoted f or π, a policy is a mapping from $\mathcal{W} \times [K]$ to $\mathbb{R}_+$ such that, for all $w \in \mathcal{W}$, $\sum_{a \in [K]} f(a,w) = 1$. Thus, a policy can be viewed as mapping a context to a distribution over actions. We say the learner is carrying out policy π at time t if, for all $a \in [K]$, $w \in \mathcal{W}$, $P[A_t = a | W_t = w] = \pi(a,w)$. Owing to statistics terminology, we also call *design* the policy carried out at a given time point. The value $\mathcal{V}(\pi)$ of π writes as

$$\mathcal{V}(\pi) \doteq E_P \left[\sum_{a \in [K]} E_P[Y | A = a, W] \pi(a|W) \right].$$

For any two policies f and g, we denote

$$V(g,f) \doteq E_P \left[\sum_{a \in [K]} \frac{f(a|W)}{g(a|W)} \right]. \tag{4.1}$$

We call $V(g,f)$ the importance sampling (IS) ratio of f and g. The IS ratio drives the variance of IS estimators of $\mathcal{V}(f)$ had the data been collected under policy g.

4.2 Generalized Policy Elimination

Introduced by Dudik et al. [2011], the policy elimination algorithm relies on the following key fact. Let g_{ref} be the uniform distribution over actions used as a reference design/policy:

$$\forall (a,w) \in [K] \times \mathcal{W},\ g_{\text{ref}}(a,w) \doteq K^{-1}.$$

Proposition 4.1. *Let $\delta > 0$. For all compact and convex set $\mathcal{F}$ of policies, there exists a policy $g \in \mathcal{F}$ such that*

$$\sup_{f \in \mathcal{F}} V(\delta g_{\text{ref}} + (1-\delta) g, f) \leq 2K. \tag{4.2}$$

We refer to their article for a proof of this result. Proposition 4.1 has an important consequence for exploration. Suppose that at time t we have a set of candidate policies $\mathcal{F}_t$, and that the designs $g_1, \ldots, g_t$ satisfy (4.2) with $\mathcal{F}_t$ substituted for $\mathcal{F}$. We can then estimate the value of

[2] P is partly a fact of nature, through the marginal distribution of context and the conditional distributions of reward given context and action, and the result of the learner's decisions.

candidate policies with error uniformly small over $\mathcal{F}_t$. This in turn has an important implication for exploitation: we can eliminate from $\mathcal{F}_t$ all the policies that have value below some well-chosen threshold, yielding a new policy set $\mathcal{F}_{t+1}$, and choose the next exploration policy g_{t+1} in $\mathcal{F}_{t+1}$. This reasoning suggested to Dudik et al. [2011] their policy elimination algorithm: (1) initialize the set of candidate policies to the entire policy class, (2) choose an exploration policy that ensures small value estimation error uniformly over candidate policies, (3) eliminate low value policies, (4) repeat steps (2) and (3). We present formally our version of the policy algorithm as algorithm 4.1 below.

In this section, we show that under an entropy condition, and if we have access to a certain optimization oracle, our GPE algorithm is efficient and beats existing regret upper bounds in some nonparametric settings. Our contribution here is chiefly to extend the regret analysis of Dudik et al. [2011] to classes of functions characterized by their metric entropy in $L_\infty(P)$ norm. This requires us to prove a new chaining-based maximal inequality for martingale processes (Theorem 4.6 in appendix 4.C). On the computational side, our algorithm relies on having access to slightly more powerful oracles than that of Dudik et al. [2011]. We present them in subsection 4.2.2 and give several examples where these oracles can be implemented efficiently.

We now formally state our GPE algorithm. Consider a policy class $\mathcal{F}$. For any policy f, any $o = (w, a, y) \in \mathcal{W} \times [K] \times \{0, 1\}$, define the policy loss and its IS-weighted counterpart

$$\begin{aligned} \ell(f)(o) &\doteq f(a, w)(1 - y), \\ \ell_\tau(f)(o) &\doteq \frac{g_{\text{ref}}(a, w)}{g_\tau(a, w)} f(a, w)(1 - y), \end{aligned}$$

the corresponding risk $R(f) \doteq E[\ell(f)(O^{\text{ref}})] = E_P[\ell_\tau(f)(O_\tau)]$ and its empirical counterpart $\widehat{R}_t(f) \doteq t^{-1} \sum_{\tau=1}^{t} \ell_\tau(f)(O_\tau)$.

Algorithm 4.1 Generalized Policy Elimination

```
**Inputs:** policy class $\mathcal{F}$, $\epsilon > 0$, sequences $(\delta_t)_{t\geq 1}$, $(x_t)_{t\geq 1}$.
Initialize $\mathcal{F}_1$ as $\mathcal{F}$.
**for** $t \geq 1$ **do**
    Find $\widetilde{g}_t \in \mathcal{F}_t$ such that, for all $f \in \mathcal{F}_t$,
        $$\frac{1}{t-1} \sum_{\tau=1}^{t-1} \frac{f(a|W_\tau)}{(\delta_t g_{\text{ref}} + (1 - \delta_t)\widetilde{g}_t)(a|W_\tau)} \leq 2K. \quad (4.3)$$
    Define $g_t = \delta_t g_{\text{ref}} + (1 - \delta_t)\widetilde{g}_t$.
    Observe context $W_t$, sample action $A_t \sim g_t(\cdot|W_t)$, collect reward $Y_t$.
    Define $\mathcal{F}_{t+1}$ as
        $$\left\{ f \in \mathcal{F}_t : \widehat{R}_t(f) \leq \min_{f \in \mathcal{F}_t} \widehat{R}_t(f) + x_t \right\}. \quad (4.4)$$
**end for**
```

4.2.1 Regret analysis

Our regret analysis relies on the following assumption.

Assumption 4.1 (Entropy condition)**.** *There exist $c > 0$, $p > 0$ such that, for all $\epsilon > 0$,*

$$\log N(\epsilon, \mathcal{F}, L_\infty(P)) \leq c\epsilon^{-p}.$$

Defining $\mathcal{F}_{t+1} \subset \mathcal{F}_t$ as (4.4), the policy elimination step, consists in removing from $\mathcal{F}_t$ all the policies that are known to be suboptimal with high probability. The threshold x_t thus plays the role of the width of a uniform-over-$\mathcal{F}_t$ confidence interval. Set $\epsilon > 0$ arbitrarily. We will show that the following choice of $(\delta_\tau)_{\tau \geq 1}$ and $(x_\tau)_{\tau \geq 1}$ ensures that the confidence intervals hold with probability $1 - 6\epsilon$, uniformly both in time and over the successive $\mathcal{F}_\tau$'s: for all $\tau \geq 1$, $\delta_\tau \doteq \tau^{-(1/2 \wedge 1/(2p))}$ and

$$x_\tau \doteq x_\tau(\epsilon) \doteq \sqrt{v_\tau(\epsilon)} \Bigg\{ \frac{c_1}{\tau^{\frac{1}{2} \wedge \frac{1}{2p}}} + \frac{c_2 + c_5 \sqrt{v_\tau(\epsilon)}}{\sqrt{\tau}} \times \sqrt{\log\left(\frac{\tau(\tau+1)}{\epsilon}\right)} \\ + \frac{1}{\tau \delta_\tau}\left(c_3 + c_7 \log\left(\frac{\tau(\tau+1)}{\epsilon}\right)\right) \Bigg\}$$

— defined in appendix 4.D, $v_\tau(\epsilon)$ is a high probability upper bound on

$$\sup_{f \in \mathcal{F}_\tau} \mathrm{Var}_P(\ell_\tau(f)(O_\tau) | F_{\tau-1}).$$

It is constructed as follows. It can be shown that the conditional variance of $\ell_\tau(f)(O_\tau)$ given $F_{\tau-1}$ is driven by the expected IS ratio $E_P[\sum_{a \in [K]} f(a, W)/g_\tau(a, W) | F_{\tau-1}]$. Step 4.3 ensures that the empirical mean over past observations of the IS ratio is no greater than $2K$, uniformly over $\mathcal{F}_\tau$. The gap $(v_\tau(\epsilon) - 2K)$ is a bound on the supremum over $\mathcal{F}_\tau$ of the deviation between empirical IS ratios and the true IS ratios.

We now state our regret theorem for algorithm 4.1. Let $f^* \doteq \arg\min_{f \in \mathcal{F}}$ be the optimal policy in $\mathcal{F}$.

Theorem 4.1 (High probability regret bound for policy elimination)**.** *Consider algorithm 4.1. Suppose that Assumption 5.1 is met. Then, with probability at least $1 - 7\epsilon$, for all $t \geq 1$,*

$$\begin{aligned} &\sum_{\tau=1}^{t} (\mathcal{V}(f^*) - Y_\tau) \\ &\leq \sqrt{t \log\left(\frac{1}{\epsilon}\right)} + 2\sum_{\tau=1}^{t} x_\tau(\epsilon) + \sum_{\tau=1}^{t} \delta_\tau \\ &= \begin{cases} O\left(\sqrt{t}\left(\log(\frac{t}{\epsilon})\right)^{3/2}\right) & \textit{if } p \in (0,1) \\ O\left(t^{\frac{p-1/2}{p}}\left(\log(\frac{t}{\epsilon})\right)^{3/2}\right) & \textit{if } p > 1 \end{cases}. \end{aligned}$$

The proof of Theorem 4.1, presented in appendix 4.D, hinges on the three following facts.

1. Controlling the supremum w.r.t. $f \in \mathcal{F}_\tau$ of the empirical estimate of the IS ratio (see (4.3) in the first step of the loop in algorithm 4.1) allows to control the supremum w.r.t. f of the true IS ratio $V(g_\tau, f)$.

2. With the specification of $(x_t)_{t\geq 1}$ and $(\delta_t)_{t\geq 1}$ sketched above we can guarantee that, with probability at least $1 - 3\epsilon$, $f^* \in \mathcal{F}_t \subset \ldots \subset \mathcal{F}_1$.

3. If $f^* \in \mathcal{F}_t$ then we can prove that, with probability at least $1 - 5\epsilon$, for all $\tau \in [t]$,

$$R(\widetilde{g}_\tau) - R(f^*) \leq 2x_\tau(\epsilon).$$

This in turn yields a high probability bound on the cumulative regret of algorithm 4.1.

4.2.2 An efficient algorithm for the exploration policy search step

We show that the exploration policy search step can be performed in $O(\text{poly}(t))$ calls to two optimization oracles that we define below. The explicit algorithm and proof of the claim are presented in appendix 4.F.

Definition 4.1 (Linearly Constrained Least-Squares Oracle). *We call Linearly Constrained Least-Squares Oracle (LCLSO) over $\mathcal{F}$ a routine that, for any $t \geq 1$, $q \geq 1$, vector $w \in \mathbb{R}^{Kt}$, sequence of vectors $W_1, ..., W_t \in \mathcal{W}$, set of vectors $u_1, ..., u_q \in \mathbb{R}^{Kt}$, and scalars $b_1, ..., b_q$, returns, if there exists one, a solution to*

$$\min_{f \in \mathcal{F}} \sum_{\substack{a \in [K] \\ \tau \in [t]}} (w(a,\tau) - f(a, W_\tau))^2 \textit{ subject to}$$

$$\forall m \in [q], \sum_{\substack{a \in [K] \\ \tau \in [t]}} u_m(a,\tau) f(a, W_\tau) \leq b_\tau.$$

Definition 4.2 (Linearly Constrained Cost-Sensitive Classification Oracle). *We call Linearly Constrained Cost-Sensitive Classification Oracle (LCCSCO) over $\mathcal{F}$ a routine that, for any $t \geq 1$, $q \geq 1$, vector $C \in (\mathbb{R}_+)^{Kt}$, set of vectors $W_1, ..., W_t \in \mathcal{W}$, set of vectors $u_1, ..., u_q \in \mathbb{R}^{Kt}$, and set of scalars $b_1, ..., b_q \in \mathbb{R}$ returns, if there exists one, a solution to*

$$\min_{f \in \mathcal{F}} \sum_{\substack{a \in [K] \\ \tau \in [t]}} C(a,\tau) f(a, W_\tau) \textit{ subject to}$$

$$\forall m \in [q], \sum_{\substack{a \in [K] \\ \tau \in [t]}} u_m(a,\tau) f(a, W_\tau) \leq b_\tau.$$

The following theorem is our main result on the computational tractability of the policy search step.

Theorem 4.2 (Computational cost of exploration policy search)**.** *For every $t \geq 1$, exploration policy search at time t can be performed in $O((Kt)^2 \log t)$ calls to both LCLSO and LCCSCO.*

The proof of Theorem 4.2 builds upon the analysis of Dudik et al. [2011]. Like them, we use the famed ellipsoid algorithm as the core component. The general idea is as follows. We show that the exploration policy search step (4.3) boils down to finding a point $w \in \mathbb{R}^{Kt}$ that belongs to a certain convex set $\mathcal{U}$, and to identifying a $\widetilde{g}_t \in \mathcal{F}_t$ such that $\sum_{a,\tau}(f(a, W_\tau) - w(a, \tau))^2 \leq \Delta$ for a certain $\Delta > 0$. In section 4.F.1, we identify $\mathcal{U}$ and Δ. In section 4.F.2, we demonstrate how to find a point in $\mathcal{U}$ with the ellipsoid algorithm.

4.3 Finite sample guarantees for ε-greedy

In this section, we give regret guarantees for two variants of the ε-greedy algorithm competing against a policy class characterized by bracketing entropy, denoted thereon $\log N_{[\,]}$, and defined in the appendix[3]. Corresponding to two choices of an input argument ϕ, the two variants of algorithm 4.2 differ in whether they optimize w.r.t. the policy either an estimate of its value or an estimate of its hinge loss-based risk.

We formalize this as follows. We consider a class $\mathcal{F}_0$ of real-valued functions over $\mathcal{W}$ and derive from it two classes $\mathcal{F}^{\mathrm{Id}}$ and $\mathcal{F}^{\mathrm{hinge}}$ defined as

$$\begin{aligned}\mathcal{F}^{\mathrm{Id}} \doteq \big\{&(a, w) \mapsto f_a(w) : f_1, \ldots, f_K \in \mathcal{F}_0, \\ &\forall w \in \mathcal{W}, (f_1(w), ..., f_K(w)) \in \Delta(K)\big\},\end{aligned} \tag{4.5}$$

where $\Delta(K)$ is the K-dimensional probability simplex, and

$$\begin{aligned}\mathcal{F}^{\mathrm{hinge}} \doteq \big\{&(a, w) \mapsto f_a(w) : f_1, \ldots f_K \in \mathcal{F}_0, \\ &\forall w \in \mathcal{W}, \textstyle\sum_{a \in [K]} f_a(w) = 0\big\}.\end{aligned} \tag{4.6}$$

Let ϕ^{Id} be the identity mapping and ϕ^{hinge} be the hinge mapping $x \mapsto \max(0, 1 + x)$, both over $\mathbb{R}$. Following exisiting terminology [Foster and Krishnamurthy, 2018, for instance], an element of $\mathcal{F}$ is called a regressor. Each regressor f is mapped to a policy π through a *policy mapping*, either $\widetilde{\pi}^{\mathrm{Id}}$ if $f \in \mathcal{F}^{\mathrm{Id}}$ or $\widetilde{\pi}^{\mathrm{hinge}}$ if $f \in \mathcal{F}^{\mathrm{hinge}}$ where, for all $(a, w) \in [K] \times \mathcal{W}$,

$$\begin{aligned}\widetilde{\pi}^{\mathrm{Id}}(f)(a, w) &= f(a, w), \\ \widetilde{\pi}^{\mathrm{hinge}}(f)(a, w) &= \mathbf{1}\{a = \arg\max_{a' \in [K]} f(a', w)\}.\end{aligned}$$

For ϕ set either to ϕ^{Id} or ϕ^{hinge}, for any $f : [K] \times \mathcal{W} \to \mathbb{R}$, for every $o = (w, a, y) \in \mathcal{W} \times [K] \times \{0, 1\}$ and each $\tau \geq 1$, define

$$\ell^{\phi}(f)(o) \doteq \phi(f(a, w))(1 - y),$$

[3] It is known that $\log N(\epsilon, \mathcal{F}, L_r(P))$ is smaller than $\log N_{[\,]}(2\epsilon, \mathcal{F}, L_r(P))$ for all $\epsilon > 0$.

$$\ell_\tau^\phi(f) \doteq \frac{g_{\mathrm{ref}}(a,w)}{g_\tau(a,w)}\phi(f(a,w))(1-y),$$

the corresponding ϕ-risk $R^\phi(f) \doteq E[\ell^\phi(f)(O^{\mathrm{ref}})] = E_P[\ell_\tau^\phi(f)(O_\tau)]$ and its empirical counterpart $\widehat{R}_t(f) \doteq t^{-1}\sum_{\tau=1}^t \ell_\tau^\phi(f)(O_\tau)$. Finally, the *risk* of any policy π is defined as $R(\pi) \doteq R^\phi(\pi)$ with $\phi = \phi^{\mathrm{Id}}$ and the *hinge-risk* of any regressor $f \in \mathcal{F}^{\mathrm{hinge}}$ is defined as $R^{\mathrm{hinge}}(f) \doteq R^\phi(f)$ with $\phi = \phi^{\mathrm{hinge}}$.

We can now present the ε-greedy algorithm.

Algorithm 4.2 ε-greedy.

```
Input: convex surrogate $\phi$, regressor class $\mathcal{F}$, policy mapping $\widetilde{\pi}$, sequence $(\delta_t)_{t\geq 1}$.
Initialize $\widehat{\pi}_0$ as $g_{\mathrm{ref}}$
for $t \geq 1$ do
    Define policy as mixture between $g_{\mathrm{ref}}$ and $\widehat{\pi}_{t-1}$:
        $g_t = \delta_t g_{\mathrm{ref}} + (1-\delta_t)\widehat{\pi}_{t-1}$
    Observe context $W_t$, sample action $A_t \sim g_t(\cdot|W_t)$, collect reward $Y_t$.
    Compute optimal empirical regressor
        $\widehat{f}_t = \arg\min_{f\in\mathcal{F}} \frac{1}{t}\sum_{\tau=1}^{t} \ell_\tau^\phi(f)(O_\tau).$    (4.7)
    Compute optimal policy estimator $\widehat{\pi}_t = \widetilde{\pi}(\widehat{f}_t)$.
end for
```

We consider two instantiations of the algorithm: one corresponding to $(\phi^{\mathrm{Id}}, \mathcal{F}^{\mathrm{Id}}, \widetilde{\pi}^{\mathrm{Id}})$ and called *direct policy optimization*, the other corresponding to $(\phi^{\mathrm{hinge}}, \mathcal{F}^{\mathrm{hinge}}, \widetilde{\pi}^{\mathrm{hinge}})$ and called *hinge-risk optimization*.

Regret decomposition. Denote π_Π^* the optimal policy in $\Pi \doteq \widetilde{\pi}(\mathcal{F})$ and π^* any[4] optimal measurable policy. The key idea in the regret analysis of the ε-greedy algorithm is the following elementary decomposition (details in appendix 4.E):

$$Y_t - R(\pi^*) = \underbrace{Y_t - E_P[Y_t|F_{t-1}]}_{\text{reward noise}} + \underbrace{\delta_t(R(g_{\mathrm{ref}}) - R(\pi^*))}_{\text{exploration cost}} + (1-\delta_t)\underbrace{(R(\widehat{\pi}_{t-1}) - R(\pi^*))}_{\text{exploitation cost}}. \quad (4.8)$$

Control of the exploitation cost. In the direct policy optimization case, we can give exploitation cost guarantees under no assumption other than an entropy condition on $\mathcal{F}$. In the hinge-risk optimization case, we need a so-called realizability assumption. Denote $\mathbb{R}_{=0}^K \doteq \{x \in \mathbb{R}^K : \sum_{a\in[K]} x_a = 0\}$.

[4] There may exist more than one.

Assumption 4.2 (Hinge-realizability)**.** *Let*

$$f^* \doteq \operatorname*{arg\,min}_{f:[K]\times\mathcal{W}\to\mathbb{R}^K_{=0}} R^{\mathrm{hinge}}(f)$$

be the minimizer over all measurable regressors of the hinge-risk. We say that a regressor class $\mathcal{F}^{\mathrm{hinge}}$ satisfies the hinge-realizability assumption for the hinge-risk if $f^* \in \mathcal{F}^{\mathrm{hinge}}$.

Imported from the theory of classification calibration, Assumption 4.2 allows us to bound the risk of a policy $R(\widetilde{\pi}^{\mathrm{hinge}}(f))$ in terms of the hinge-risk of the regressor f. The proof relies on the following result:

Lemma 4.1 (Hinge-calibration)**.** *Consider a regressor class $\mathcal{F}^{\mathrm{hinge}}$. Let*

$$\pi^* \in \operatorname*{arg\,min}_{\pi:[K]\times\mathcal{W}\to\Delta(K)} R(\pi)$$

be an optimal measurable policy. It holds that $R(\pi^*) = R(\widetilde{\pi}^{\mathrm{hinge}}(f^*))$ and, for all $f \in \mathcal{F}^{\mathrm{hinge}}$,

$$R(\widetilde{\pi}^{\mathrm{hinge}}(f)) - R(\pi^*) \leq R^{\mathrm{hinge}}(f) - R^{\mathrm{hinge}}(f^*).$$

We refer the reader to Bartlett et al. [2006], Ávila Pires and Szepesvári [2016] for proofs, respectively when $K = 2$ and when $K \geq 2$. Under Assumption 4.2, Lemma 4.1 teaches us that we can bound the exploitation cost in terms of the excess hinge-risk $R^{\mathrm{hinge}}(f) - \min_{f' \in \mathcal{F}^{\mathrm{hinge}}} R^{\mathrm{hinge}}(f')$, a quantity that we can bound by standard arguments from the theory of empirical risk minimization. The fondamental building block of our exploitation cost analysis is therefore the following finite sample deviation bound for the empirical ϕ-risk minimizer.

Theorem 4.3 (ϕ-risk exponential deviation bound for the ε-greedy algorithm)**.** *Let ϕ and $\mathcal{F}$ be either ϕ^{Id} and $\mathcal{F}^{\mathrm{Id}}$ or ϕ^{hinge} and $\mathcal{F}^{\mathrm{hinge}}$. Suppose that $g_1, \ldots, g_t$ is a sequence of policies such that, for all $\tau \in [t]$, g_τ is $F_{\tau-1}$-measurable. Suppose that there exist $B, \delta > 0$ such that*

$$\sup_{f_1, f_2 \in \mathcal{F}} \sup_{a \in [K], w \in \mathcal{W}} |\phi(f_1(a, w)) - \phi(f_2(a, w))| \leq B,$$

$$\min_{\tau \in [t]} g(A_\tau, W_\tau) \geq \delta \text{ a.s.}$$

Define $f^*_{\mathcal{F}} \doteq \arg\min_{f \in \mathcal{F}} R^\phi(f)$, the $\mathcal{F}$-specific optimal regressor of the ϕ-risk, and let $\widehat{f}_t$ be the empirical ϕ-risk minimizer (4.7)*. Then, for all $x > 0$ and $\alpha \in (0, B)$,*

$$P\Bigg[R^\phi(\widehat{f}_t) - R^\phi(f^*_{\mathcal{F}}) \geq H_t\left(\alpha, \delta, B^2K/\delta, B\right) \\ + 160B\sqrt{Kx/\delta t} + 3B/\delta t x\Bigg] \leq 2e^{-x},$$

with $H_t(\alpha, \delta, v, B) \doteq \alpha + 160\sqrt{v/t}$

$$\times \int_{\alpha/2}^{B} \sqrt{\log(1 + N_{[]}(\epsilon, \mathcal{F}, L_2(P))} d\epsilon + \frac{3B}{\delta t} \log 2.$$

As a direct corollary, we can express rates of convergence for the ϕ-risk in terms of the bracketing entropy rate.

Corollary 4.1. *Suppose that* $\log(1 + N_{[]}(\epsilon, \mathcal{F}, L_2(P))) = O(\epsilon^{-p})$ *for some* $p \in (0, 1)$. *Then*

$$R^{\phi}(\widehat{f}_t) - R^{\phi}(f^*_{\mathcal{F}}) = O_P\left((\delta t)^{-\left(\frac{1}{2} \wedge \frac{1}{p}\right)}\right).$$

Control of the regret. The cumulative reward noise $\sum_{\tau=1}^{t}(Y_\tau - E_P[Y_\tau|F_{\tau-1}])$ can be bounded by the Azuma-Hoeffding inequality. From (4.16) and Corollary 4.1, δ_t controls the trade off between the exploration and exploitation costs. We must therefore choose a δ_t that minimizes the total of these two which, from the above, scales as $O(\delta_t + (t\delta_t)^{-\left(\frac{1}{2} \wedge \frac{1}{p}\right)})$. The optimal choice is $\delta_t \propto t^{-\left(\frac{1}{3} \wedge \frac{1}{p+1}\right)}$. The following theorem formalizes the regret guarantees under the form of a high-probability bound.

Theorem 4.4 (High probability regret bound for ε-greedy.)**.** *Suppose that the bracketing entropy of the regressor class* $\mathcal{F}$ *satisfies* $\log(1 + N_{[]}(\epsilon, \mathcal{F}, L_2(P)) = O(\epsilon^{-p})$ *for some* $p > 0$. *Set* $\delta_t = t^{-\left(\frac{1}{3} \vee \frac{p}{p+1}\right)}$ *for all* $t \geq 1$. *Suppose that*

- *either* $\phi = \phi^{\mathrm{Id}}$, $\mathcal{F}$ *is of the form* $\mathcal{F}^{\mathrm{Id}}$, $\widetilde{\pi} = \widetilde{\pi}^{\mathrm{Id}}$,
- *or* $\phi = \phi^{\mathrm{hinge}}$, $\mathcal{F}$ *is of the form* $\mathcal{F}^{\mathrm{hinge}}$, $\widetilde{\pi} = \widetilde{\pi}^{\mathrm{hinge}}$, *and* $\mathcal{F}$ *satisfies Assumption 4.2.*

Then, with probability $1 - \epsilon$,

$$\sum_{\tau=1}^{t}(\mathcal{V}(\pi^*) - Y_\tau) \leq \sqrt{t \log(2/\epsilon)} + t^{\frac{p}{p+1}}\sqrt{\log(2t(t+1)/\epsilon)}.$$

4.4 Examples of policy classes

4.4.1 A nonparametric additive model

We say that $a(\epsilon) = \widetilde{O}(b(\epsilon))$ if there exists $c > 0$ such that $a(\epsilon) = O(b(\epsilon)\log^c(1/\epsilon))$. We present a policy class that has entropy $\widetilde{O}(\epsilon^{-1})$, and over which the two optimization oracles presented in Definitions 4.1 and 4.2 reduce to linear programs. Let $\mathbb{D}([0, 1])$ be the set of càdlàg functions and let the variation norm $\|\cdot\|_v$ be given, for all $h \in \mathbb{D}([0, 1])$, by

$$\|h\|_v \doteq \sup_{m \geq 2} \sup_{x_1, \ldots, x_m} \sum_{i=1}^{m-1} |h(x_{i+1}) - h(x_i)|$$

where the right-hand side supremum is over the subdivisions of $[0, 1]$, that is over $\{(x_1, \ldots, x_m) : 0 \leq x_1 \leq \ldots \leq x_m \leq 1\}$. Set $C, M > 0$ then introduce

$$\mathcal{H} \doteq \{h \in \mathbb{D}([0, 1]) : \|h\|_v \leq M\}$$

and the additive nonparametric additive model derived from it by setting $\mathcal{F}_0 \doteq$

$$\big\{(a,w) \mapsto \sum_{l=1}^{d} \alpha_{a,l} h_l(w_l) : |\alpha_{a,l}| \leq C, h_{a,l} \in \mathcal{H}\big\}.$$

Let $\mathcal{F} = \mathcal{F}^{\mathrm{Id}}$ derived from $\mathcal{F}_0$ as in (4.5).

The following lemma formally bounds the entropy of the policy class.

Lemma 4.2. *There exists* $\epsilon_0 \in (0,1)$ *such that, for all* $\epsilon \in (0, \epsilon_0)$*,*

$$\begin{aligned}\log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq & K \log N_{[]}(\epsilon, \mathcal{F}_0, \|\cdot\|_\infty) \\ \leq & K c_0 \epsilon^{-1} \log(1/\epsilon).\end{aligned}$$

for some $c_0 > 0$ *depending on* (C, d, M)*.*

We now state a result that shows that LCLSO and LCCSCO reduce to linear programs over $\mathcal{F}$. We first need to state a definition.

Definition 4.3 (Grid induced by a set of points)**.** *Consider* d *subdivisions of* $[0,1]$ *of the form*

$$\begin{aligned} 0 =& w_{1,1} \leq w_{1,2} \leq \ldots \leq w_{1,q_1} = 1, \\ & \vdots \\ 0 =& w_{d,1} \leq w_{1,2} \leq \ldots \leq w_{d,q_d} = 1. \end{aligned}$$

The rectangular grid induced by these d subdivisions *is the set of points* $(w_{1,i_1}, w_{2,i_2}, \ldots, w_{i,i_d})$ *with* $i_1 \in [q_1]$, ..., $i_d \in [q_d]$*. We call a* rectangular grid *any rectangular grid induced by some set of* d *subdivisions of* $[0,1]$*.*

Consider a set of points $w_1, \ldots, w_n \in [0,1]^d$*. A minimal grid induced by* $w_1, \ldots w_n$ *is any rectangular grid that contains* $w_1, \ldots w_n$ *and that is of minimal cardinality. We denote by* $G(w_1, \ldots, w_n)$ *a minimal rectangular grid induced by* $w_1, \ldots w_n$ *chosen arbitrarily.*

Lemma 4.3. *Let* $w_0 = \mathbf{0}, w_1, \ldots, w_t \in [0,1]^d$*. For all* $l \in [d]$*, let* $\widetilde{\mathcal{H}}_{l,t} \doteq \widetilde{\mathcal{H}}_{l,t}(w_{0,l}, \ldots, w_{t,l}) \doteq$

$$\big\{x \mapsto \sum_{\tau=0}^{t} \beta_\tau 1\{x \geq w_{\tau,l}\} : \beta_\tau \in \mathbb{R}, \sum_{\tau=0}^{t} |\beta_\tau| \leq M\big\}$$

and $\widetilde{\mathcal{F}}_{0,t} \doteq$

$$\big\{(a,w) \mapsto \sum_{l=1}^{d} \alpha_{a,l} \widetilde{h}_{a,l}(w_l) : |\alpha_{a,l}| \leq B, \widetilde{h}_{a,l} \in \mathcal{H}_{l,t}\big\}.$$

Let $(u_{a,\tau})_{a \in [K], \tau \in [t]}$ *be a vector in* $\mathbb{R}^{Kt}$*. Let* $\widetilde{f}^*$ *be a solution to the following optimization problem* $(\mathcal{P}_2)$*:*

$$\max_{\widetilde{f} \in \widetilde{\mathcal{F}}_{0,t}} \sum_{a \in [K]} \sum_{\tau=1}^{t} u_{a,\tau} \widetilde{f}(a, W_\tau)$$

$$\textit{s.t. } \forall a \in [K],\ \forall w \in \mathcal{G}(w_0, \ldots, w_t),\ \widetilde{f}(a, w) \geq 0, \tag{4.9}$$

$$\forall w \in \mathcal{G}(w_0, \ldots, w_t),\ \sum_{a \in [K]} \widetilde{f}(a, w) = 1. \tag{4.10}$$

Then, $\widetilde{f}$ is a solution to the following optimization problem $(\mathcal{P}_1)$:

$$\max_{f \in \mathcal{F}_0} \sum_{a \in [K]} \sum_{\tau=1}^{t} u_{a,\tau} f(a, W_\tau)$$

$$\textit{s.t. } \forall a \in [K], \forall w \in [0,1]^d, f(a, w) \geq 0, \tag{4.11}$$

$$\forall w \in [0,1]^d,\ \sum_{a \in [K]} f(a, w) = 1. \tag{4.12}$$

4.4.2 Càdlàg policies with bounded sectional variation norm

The class of d-variate càdlàg functions with bounded sectional variation norm is a nonparametric function class with bracketing entropy bounded by $O(\epsilon^{-1} \log(1/\epsilon)^{2(d-1)})$, over which empirical risk minimization takes the form of a LASSO problem. It has received attention recently in the nonparametric statistics literature [van der Laan, 2016, Fang et al., 2019, Bibaut and van der Laan, 2019]. Empirical risk minimizers over this class of functions have been termed Highly Adaptive Lasso estimators by van der Laan [2016]. The experimental study of Benkeser and van der Laan [2016] suggests that Highly Adaptive Lasso estimators are competitive against supervised learning algorithms such as Gradient Boosting Machines and Random Forests.

Sectional variation norm. For a function $f : [0,1]^d \to \mathbb{R}$, and a non-empty subset s of $[d]$, we call the s-section of f and denote f_s the restriction of f to $\{x \in [0,1]^d : \forall i \in s, x_i = 0\}$. The sectional variation norm (svn) is defined based on the notion of Vitali variation. Defining the notion of Vitali variation in full generality requires introducing additional concepts. We thus relegate the full definition to appendix 4.H, and present it in a particular case. The Vitali variation of an m-times continuously differentiable function $g : [0,1]^m \to \mathbb{R}$ is defined as

$$V^{(m)}(g) \doteq \int_{[0,1]^m} \left| \frac{\partial^m g}{\partial x_1 \ldots \partial x_m} \right|.$$

For arbitrary real-valued càdlàg functions g on $[0,1]^m$ (non necessarily m times continuously differentiable), the Vitali variation $V^{(m)}(g)$ is defined in appendix 4.H. The svn of a function $f : [0,1]^d \to \mathbb{R}$ is defined as

$$\|f\|_v \doteq |f(0)| + \sum_{\emptyset \neq s \subset [d]} V^{(|s|)}(f_s),$$

that is the sum of its absolute value at the origin and the sum of the Vitali variation of its sections. Let $\mathbb{D}([0,1]^d)$ be the class of càdlàg functions with domain $[0,1]^d$ and, for some $M > 0$, let

$$\mathcal{F}_0 \doteq \left\{ f \in \mathbb{D}([0,1]^d) : \|f\|_v \leq M \right\} \tag{4.13}$$

be the class of càdlàg functions with svn smaller than M.

Entropy bound. The following result is taken from [Bibaut and van der Laan, 2019].

Lemma 4.4. *Consider $\mathcal{F}_0$ defined in* (4.13)*. Let P be a probability distribution over $[0,1]^d$ such that $\|\cdot\|_{P,2} \leq c_0\|\cdot\|_{\mu,2}$, with μ the Lebesgue measure and $c_0 > 0$. Then there exist $c_1 > 0, \epsilon_0 \in (0,1)$ such that, for all $\epsilon \in (0, \epsilon_0)$ and all distributions P over $[0,1]^d$,*

$$\log N_{[]}(\epsilon, \mathcal{F}_0, L_2(P)) \leq c_1 M \epsilon^{-1} \log(M/\epsilon)^{2d-1}.$$

Representation of ERM. We show that empirical risk minimization (ERM) reduces to linear programming in both our direct policy and hinge-risk optimization settings.

Lemma 4.5 (Representation of the ERM in the direct policy optimization setting)**.** *Consider a class of policies of the form $\mathcal{F}^{\mathrm{Id}}$* (4.5) *derived from $\mathcal{F}_0$* (4.13)*. Let $\phi = \phi^{\mathrm{hinge}}$. Suppose we have observed (W_1, A_1, Y_1), ..., (W_t, A_t, Y_t) and let $\widetilde{W}_1, \ldots, \widetilde{W}_m$ be the elements of $G(W_1, \ldots, W_t)$.*

Let $(\beta_j^a)_{a\in[K], j\in[m]}$ be a solution to

$$\begin{aligned} \min_{\beta\in\mathbb{R}^{Km}} \sum_{\tau=1}^{t} \sum_{a\in[K]} \Bigg\{ & \frac{1\{A_\tau = a\}}{g_\tau(A_\tau, W_\tau)}(1 - Y_\tau) \\ & \times \sum_{j=1}^{m} \beta_j^a 1\{W_\tau \geq \widetilde{W}_j\}\Bigg\} \\ \text{s.t. } \forall l \in [m], & \sum_{a\in[K]} \sum_{j=1}^{m} \beta_j^a 1\{\widetilde{W}_l \geq \widetilde{W}_j\} = 1, \\ \forall l \in [m], \forall a \in [K], & \sum_{j=1}^{m} \beta_j^a 1\{\widetilde{W}_l \geq \widetilde{W}_j\} \geq 0, \\ \forall a \in [K], & \sum_{j=1}^{m} |\beta_j^a| \leq M. \end{aligned} \tag{4.14}$$

Then $f : (a, w) \mapsto \sum_{j=1}^m \beta_j^a 1\{w \geq \widetilde{W}_j\}$ is a solution to $\min_{f\in\mathcal{F}^{\mathrm{Id}}} \sum_{\tau=1}^t \ell_\tau^\phi(f)(O_\tau)$.

We present a similar result for the hinge-risk setting in appendix 4.H. It is relatively easy to prove with the same techniques that ERM over $\mathcal{F}^{\mathrm{hinge}}$ also reduces to linear programming when $\mathcal{F}_0$ is an RKHS.

4.5 Conclusion

In this chapter, we proposed and analyzed a polynomial time algorithm for sequential decision making under the stochastic contextual bandit model. Our algorithm achieves the minimax (up to log factors) regret rate w.r.t. nonparametric policy classes with integrable entropy. For larger classes, the regret rate isn't minimax optimal. We believe this is due to our proof techniques, and that it is doable to design polynomial time algorithms for classes with non-integrable entropy as well. One notable feature of our setting is that we do not require the realizability assumption — a setting sometimes called the *policy setting*.

In our view, the main contribution of this work is to show that it is possible to design and analyze regret efficient polynomial time algorithms on what we call "real" nonparametric classes, in the policy setting. Our proposed solution is far from fully satisfactory, for several reasons. Firstly, as already mentioned, our algorithm isn't regret optimal for large entropy classes. Secondly, as mentioned in the introduction chapter, a practical contextual bandit algorithm should not rely on knowing a priori the complexity of the policy class, rather it should data-adaptively learn it. Finally, it would be desirable to improve the current work in the direction of more broadly tractable optimization oracles. This latter question has been in particular studied in Foster et al. [2018b] where the authors propose procedures relying on square loss regression oracles instead of in general intractable cost-sensitive classification oracles.

Bibliography

A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. E. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1638–1646, Beijing, China, 2014. PMLR.

S. Athey and S. Wager. Efficient policy learning, 2017. arXiv preprint arXiv:1702.02896v5.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002a.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

B. Ávila Pires and C. Szepesvári. Multiclass classification calibration functions, 2016. arXiv preprint arXiv:1609.06385v1.

P. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

D. Benkeser and M. J. van der Laan. The highly adaptive lasso estimator. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 689–696, 2016.

A. F. Bibaut and M. J. van der Laan. Fast rates for empirical risk minimization over càdlà functions with bounded sectional variation norm, 2019. arXiv preprint arXiv:1907.09244v2.

N. Cesa-Bianchi, P. Gaillard, C. Gentile, and S. Gerchinovitz. Algorithmic chaining and the role of partial feedback in online nonparametric learning. In S. Kale and O. Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 465–481, Amsterdam, Netherlands, 2017. PMLR.

N. Chatterji, A. Pacchiano, and P. Bartlett. Online learning with kernel losses. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 971–980, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 169—178, Arlington, Virginia, USA, 2011. AUAI Press.

R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1:290–330, 1967.

B. Fang, A. Guntuboyina, and B. Sen. Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and hardy-krause variation, 2019. arXiv preprint arXiv:1903.01395v2.

D. Foster, A. Agarwal, M. Dudik, H. Luo, and R. E. Schapire. Practical contextual bandits with regression oracles. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1539–1548, Stockholmsmässan, Stockholm Sweden, 2018a. PMLR.

D. J. Foster and A. Krishnamurthy. Contextual bandits with surrogate losses: Margin bounds and efficient algorithms. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2621–2632. Curran Associates, Inc., 2018.

Dylan Foster, Alekh Agarwal, Miroslav Dudik, Haipeng Luo, and Robert Schapire. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 1539–1548. PMLR, 2018b.

J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 817–824. Curran Associates, Inc., 2008.

N. Littlestone and M. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

A. R. Luedtke and A. Chambaz. Performance guarantees for policy learning. *Annales de l'Institut Henri Poincaré – Probabilité et Statistiques*, 0(0), 2019.

P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007.

V. Perchet and P. Rigollet. The multi-armed bandit problem with covariates. *Ann. Statist.*, 41(2): 693–721, 04 2013.

P. Rigollet and A. Zeevi. Nonparametric bandits with covariates. In A. Tauman Kalai and M. Mohri, editors, *COLT*, pages 54–66, Haifa, Israel, 2010. Ominipress.

S. A. van de Geer. *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000.

M. J. van der Laan. A generally efficient TMLE. *The International Journal of Biostatistics*, 1(1), 2016.

A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

R. van Handel. On the minimal penalty for Markov order estimation. *Probability Theory and Related Fields*, 150:709–738, 2011.

V. G. Vovk. Aggregating strategies. In M. A. Fulk and J. Case, editors, *Proceedings of the Third Annual Workshop on Computational Learning Theory, COLT 1990, University of Rochester, Rochester, NY, USA, August 6-8, 1990*, pages 371–386. Morgan Kaufmann, 1990.

4.A Additional comparisons with previous articles

Comparison with Chatterji et al. [2019]. Their regret-optimality claim holds only against RKHS classes for which the kernel has exponential eigendecay These are small classes: the corresponding policy classes have entropy $\log N(\epsilon) \leq C \log(1/\epsilon)^d$, for some constant $C > 0$, which is essentially a parametric complexity.

Comparison with the original Policy Elimination of Dudik. An essential difference is that the Policy Elimination (PE) algorithm is not implementable, as it requires to optimize an expectation with respect to the true distribution of contexts, which is unknown. In Dudik et al. [2011], the authors present the Policy Elimination algorithm primarily so as to introduce their main ideas, and then propose a substantially more complex algorithm inspired by PE, RandomizedUCB, which is actually implementable.

Our Generalized Policy Elimination algorithm replaces the step in PE that optimizes w.r.t. a true expectation with a step that optimizes w.r.t. an empirical expectation, which makes it implementable. Our implementable version of PE is an alternative to the rather complex RandomizedUCB. We however pay a price for this gained simplicity, as we have to rely on more powerful optimization oracles (our linearly constrained version of cost sensitive classification and least squares oracles).

One seemingly minor difference (but key to the regret analysis) is that our algorithm comes with different settings of the uniform exploration rate, and of the empirical regret threshold that defines which policy to eliminate. In our work, these are dependent on the entropy of the policy class.

4.B Notation

Set arbitrarily $n \geq 1$ and let ϕ be either ϕ^{Id} or ϕ^{hinge}. We denote by P_n the empirical distribution $n^{-1} \sum_{i=1}^n \mathrm{Dirac}(O_i)$. For all measurable $f : [K] \times \mathcal{W} \to \mathbb{R}$, we let $\ell_{1:n}^\phi(f)$ be the vector-valued random function $(\ell_1^\phi(f), \ldots, \ell_n^\phi(f))$ over $[K] \times \mathcal{W}$. In order to alleviate notation, we introduce the following empirical process theory-inspired notation. For any *fixed*, measurable function $f : [K] \times \mathcal{W} \to \mathbb{R}$,

$$\begin{aligned} P\ell_{1:n}(f) &\doteq \frac{1}{n} \sum_{i=1}^n E_P \left[\ell_i^\phi(f)(O_i) | F_{i-1} \right], \\ P_n \ell_{1:n}(f) &\doteq \frac{1}{n} \sum_{i=1}^n \ell_i^\phi(f)(O_i), \\ (P - P_n)\ell_{1:n}(f) &\doteq \frac{1}{n} \sum_{i=1}^n \left(E_P \left[\ell_i^\phi(f)(O_i) | F_{i-1} \right] - \ell_i^\phi(f)(O_i) \right). \end{aligned}$$

For a *random* measurable function $f : [K] \times \mathcal{W} \to \mathbb{R}$, we let $P\ell_{1:n}(f) \doteq P\ell_{1:n}(f')|_{f'=f}$, and $P_n \ell_{1:n}(f')|_{f'=f}$, $(P - P_n)\ell_{1:n}(f')|_{f'=f}$.

4.C Maximal inequalities

4.C.1 The basic maximal inequality for IS-weighted martingale processes

Definition 4.4 (Bracketing entropy, van der Vaart and Wellner [1996]). *Given two functions $l, u : \mathcal{X} \to \mathbb{R}$, the bracket $[l, u]$ is the set of all functions $f : \mathcal{X} \to \mathbb{R}$ such that, for all $x \in \mathcal{X}$, $l(x) \leq f(x) \leq u(x)$. The bracketing number $N_{[]}(\epsilon, \mathcal{F}, L_r(P))$ is the number of brackets $[l, u]$ such that $\|l - u\|_{P,r} \leq \epsilon$ needed to cover $\mathcal{F}$.*

The following proposition is a well-known result relating bracketing numbers and covering numbers [van der Vaart and Wellner, 1996, for instance].

Proposition 4.2. *For any probability distribution P, for all $\epsilon > 0$, $N(\epsilon, \mathcal{F}, L_r(P)) \leq N_{[]}(2\epsilon, \mathcal{F}, L_r(P))$ and $N(\epsilon, \mathcal{F}, \| \cdot \|_\infty) \leq N_{[]}(2\epsilon, \mathcal{F}, \| \cdot \|_\infty)$.*

In the statement of Theorem 4.1, the high-probability regret bound for GPE, we used the covering numbers in uniform norm. The previous lemma allows us to carry out the analysis in terms of bracketing numbers in uniform norm.

Theorem 4.5 (Maximal inequality for IS-weighted martingale processes). *Consider the setting of Section 4.3 in the main text. Specifically, suppose that for all $i \geq 1$, $A_i | W_i \sim g_i(\cdot | A_i)$ where g_i is F_{i-1}-measurable. Let $n \geq 1$, and $f_0 \in \mathcal{F}$. Suppose that*

- *there exists $\delta > 0$ such that, for every $i \in [n]$, $g_i(a, w) \geq \delta$;*

- *there exists* $B > 0$ *such that* $\sup_{f\in\mathcal{F}} \sup_{a,w\in[K]\times\mathcal{W}} |\phi(f(a,w)) - \phi(f_0(a,w))| \leq B$*;*
- *there exists* $v > 0$ *such that* $\sup_{f\in\mathcal{F}} \bar{V}_n(\phi(f) - \phi(f_0)) \leq v$*, where, for any pair* (f,g) *of functions* $[K]\times\mathcal{W}\to\mathbb{R}_+$*,* $\bar{V}_n(g) \doteq n^{-1}\sum_{i=1}^t V(g_i, f)$ *(the definition of* $V(g,f)$ *is given in* (4.1) *in the main text).*

Then, for all $\alpha \in [0, B]$*,*

$$P\left[\sup_{f\in\mathcal{F}} M_n(f) \geq H_n(\alpha,\delta,v,B) + 160\sqrt{\frac{vx}{n}} + 3\frac{Bx}{\delta n}\right] \leq 2e^{-x},$$

where

$$M_n(f) \doteq \frac{1}{n}\sum_{i=1}^n E\left[\ell_i^\phi(O_i) - \ell_i^\phi(f_0)(O_i)|F_{i-1}\right] - \left(\ell_i^\phi(O_i) - \ell_i^\phi(f_0)(O_i)\right), \tag{4.15}$$

and

$$H_n(\alpha,\delta,v,B) \doteq \alpha + 160\sqrt{\frac{v}{n}}\int_{\alpha/2}^B \sqrt{\log(1 + N_{[]}(\epsilon,\phi(\mathcal{F}),L_2(P))}d\epsilon + 3\frac{B}{\delta n}\log 2.$$

Proof of theorem 4.5. The proof follows closely the proof of [Theorem A.4 in van Handel, 2011].

From a conditional expectation bound to a deviation bound. Let $x > 0$ and let A be the event

$$A \doteq \left\{\sup_{f\in\mathcal{F}} M_n(f) \geq \psi(x)\right\},$$

with $\psi(x) \doteq H_n(\alpha,\delta,v,B) + \sqrt{vx/n} + Bx/(\delta n)$. Observe that, for any $x > 0$,

$$\psi(x) \leq E_P^A\left[\sup_{f\in\mathcal{F}} \mathbf{1}\{\bar{V}_n(f) \leq v\}(P - P_n)\ell_{1:n}(f)\right].$$

Therefore, to prove the claim, it suffices to prove that

$$E_P^A\left[\sup_{f\in\mathcal{F}} \mathbf{1}\{\bar{V}_n(f) \leq v\}(P - P_n)\ell_{1:n}(f)\right] \leq \psi\left(\log\left(1 + \frac{1}{P[A]}\right)\right),$$

as this would imply

$$\Psi(x) \leq \psi\left(\log\left(1 + \frac{1}{P[A]}\right)\right) \leq \psi\left(\log\left(\frac{2}{P[A]}\right)\right),$$

which, as ψ is increasing, implies $P[A] \leq 2e^{-x}$, which is the wished claim.

Setting up the notation. In this proof, we will denote

$$H \doteq \{\phi(f) - \phi(f_0) : f \in \mathcal{F}\}.$$

Observe that by assumption $\mathcal{H}$ has diameter in $\|\cdot\|_\infty$ norm (and thus in $L_2(P)$ norm) smaller than B. For all $j \geq 0$, let $\epsilon_j = B2^{-j}$, and let

$$\mathcal{B}_j \doteq \{(\underline{h}^{j,\rho}, \overline{h}^{j,\rho}) : \rho = 1, \ldots, N_j\}$$

be an ϵ_j-bracketing of $\mathcal{H}$ in $L_2(P)$ norm. Further suppose that $\mathcal{B}_j$ is a minimal bracketing, that is that $N_j = N_{[]}(\epsilon_j, \mathcal{H}, L_2(P))$. For all j, h, let $\rho(j,h)$ be the index of a bracket in $\mathcal{B}_j$ that contains h, that is $\rho(j,h)$ is such that

$$\underline{h}^{j,\rho(j,f)} \leq h \leq \overline{h}^{j,\rho(j,f)}.$$

For all $h \in \mathcal{H}$, $j \geq 0$, $i \in [n]$ let

$$\lambda^{j,h} \doteq \underline{h}^{j,\rho(j,h)},$$

and

$$\Delta_i^{j,h} \doteq (h - \lambda^{j,h})(A_i, W_i).$$

Adaptive chaining. The core idea of the proof is a so-called adaptive chaining device: for any h, and any $i \in [n]$, we write

$$\begin{aligned} h(A_i, W_i) =& h(A_i, W_i) - \lambda^{\tau_i^h, h}(A_i, W_i) \vee \lambda^{\tau_i^h - 1, h}(A_i, W_i) \\ &+ \lambda^{\tau_i^h, h}(A_i, W_i) \vee \lambda^{\tau_i^h - 1, h}(A_i, W_i) - \lambda^{\tau_i^h - 1, h}(A_i, W_i) \\ &+ \sum_{j=1}^{\tau_i^h - 1} \lambda^{j,h}(A_i, W_i) \vee \lambda^{j-1,h}(A_i, W_i) \\ &+ \lambda^{0,h}(A_i, W_i), \end{aligned}$$

for some $\tau_i^h \geq 0$ that plays the role of the depth of the chain. We choose the depth τ_i^h so as to control the supremum norm of the links of the chain. Specifically, we let

$$\tau_i^h \doteq \min\left\{ j \geq 0 : \Delta_i^{j,h} > a_j \right\} \wedge J,$$

for some $J \geq 1$, and a decreasing positive sequence a_j, which we will explicitly specify later in the proof. The chaining decomposition in 4.C.2 can be rewritten as follows:

$$\begin{aligned} h(A_i, W_i) =& \lambda^{0,h}(A_i, W_i) \\ &+ \sum_{j=0}^{J} \left\{ h(A_i, W_i) - \lambda^{j,h} \vee \lambda^{j-1,h}(A_i, W_i) \right\} \mathbf{1}\{\tau_i^h = j\} \end{aligned}$$

$$+\sum_{j=1}^{J}\Big\{\big(\lambda^{j,h}(A_i,W_i)\vee\lambda^{j-1,h}(A_i,W_i)-\lambda^{j-1,h}(A_i,W_i)\big)\mathbf{1}\{\tau_i^h=j)\} \\ +\big(\lambda^{j,h}(A_i,W_i)-\lambda^{j-1,h}(A_i,W_i)\big)\mathbf{1}\{\tau_i^h>j\}\Big\}$$

Denote $a_i^h \doteq \lambda^{0,h}(A_i,W_i)$,

$$b_i^{j,h}\doteq\big\{h(A_i,W_i)-\lambda^{j,h}\vee\lambda^{j-1,h}(A_i,W_i)\big\}\mathbf{1}\{\tau_i^h=j\},$$

and

$$c_i^{j,h}\doteq\big(\lambda^{j,h}(A_i,W_i)\vee\lambda^{j-1,h}(A_i,W_i)-\lambda^{j-1,h}(A_i,W_i)\big)\mathbf{1}\{\tau_i^h=j)\} \\ +\big(\lambda^{j,h}(A_i,W_i)-\lambda^{j-1,h}(A_i,W_i)\big)\mathbf{1}\{\tau_i^h>j\}.$$

Overloading the notation, we will denote, for every $i\in[n]$ and function $h:[K]\times\mathcal{W}\to\mathbb{R}$,

$$\ell_i(h)\doteq\frac{h(A_i,W_i)(1-Y_i)}{g_i(A_i,W_i)}.$$

From the linearity of $\ell_1,\ldots,\ell_n$, we have that

$$(P-P_n)\ell_{1:n}(h)=A_n^h+\sum_{j=0}^{J}B_n^{j,h}+\sum_{j=1}^{J}C_n^{j,h},$$

with

$$\begin{aligned}A_n^h\doteq&\frac{1}{n}\sum_{i=1}^{n}E[\ell_i(a_i^h)|F_{i-1}]-\ell_i(a_i^h),\\ B_n^{j,h}\doteq&\frac{1}{n}\sum_{i=1}^{n}E[\ell_i(b_i^{j,h})|F_{i-1}]-\ell_i(b_i^{j,h}),\\ C_n^{j,h}\doteq&\frac{1}{n}\sum_{i=1}^{n}E[\ell_i(c_i^{j,h}|F_{i-1}]-\ell_i(c_i^{j,h}).\end{aligned}$$

The terms A_n^h, $B_n^{j,h}$ and $C_n^{j,h}$ can be intepreted as follows. For any given h and chain corresponding to h:

- A_n^h represents the root, at the coarsest level, of the chain,
- if the chain goes deeper than depth j, $C_n^{j,h}$ is the link of the chain between depths $j-1$ and j,
- if the chain stops at depth j, $B_n^{j,h}$ is the tip of the chain.

We control each term separately.

Control of the roots. Observe that, for all $i \in [n]$,

$$
\begin{aligned}
E_P[\ell_i(a_i^h)^2|F_{i-1}] =& E_P\left[\frac{\lambda^{0,h}(A_i,W_i)^2}{g_i(A_i|W_i)^2}(1-Y_i)^2|F_{i-1}\right] \\
\leq & E_P\left[\sum_{a\in[K]}\frac{\left(\lambda^{0,h}(a,W_i)-h(a,W_i)+h(a,W_i)\right)^2}{g_i(a,W_i)}\middle|F_{i-1}\right] \\
\leq & 2\delta^{-1}\left\{E_P\left[\sum_{a\in[K]}\left(\lambda^{0,h}(a,W_i)-h(a,W_i)\right)^2\middle|F_{i-1}\right]\right. \\
& \left.+E_P\left[\sum_{a\in[K]}h(a,W_i)^2\middle|F_{i-1}\right]\right\} \\
\leq & 4K\delta^{-1}\epsilon_0^2.
\end{aligned}
$$

In the second line we have used that $(1-Y_i)\in[0,1]$. As $\|\lambda^{0,h}\|_\infty \leq B$ and $\inf_{a,w} g_i(a,w)\geq\delta$, we have that

$$
|\ell_i(a_i^h)| \leq B\delta^{-1}.
$$

Therefore, from lemma 4.6,

$$
E_P^A\left[\sup_{h\in\mathcal{H}} A_n^h\right] \leq 8\epsilon_0\sqrt{\frac{K}{\delta}\log\left(1+\frac{N_0}{P[A]}\right)}+\frac{8}{3}\frac{B}{\delta n}\log\left(1+\frac{N_0}{P[A]}\right).
$$

Control of the tips. As $\lambda_i^{j,h}$ is a lower bracket, $\ell_i(b_i^{j,h})\leq 0$ and thus

$$
\begin{aligned}
& E_P\left[\ell_i(b_i^{j,h})|F_{i-1}\right]-\ell_i(b_i^{j,h}) \\
\leq & E_P\left[\ell_i(b_i^{j,h})|F_{i-1}\right] \\
= & E_P\left[\frac{(h(A_i,W_i)-\lambda^{j,h}(A_i,W_i)\vee\lambda^{j,h}(A_i,W_i)}{g_i(A_i,W_i)}(1-Y_i)\mathbf{1}\{\tau_i^h=j\}\middle|F_{i-1}\right] \\
\leq & E_P\left[\frac{h(A_i,W_i)-\lambda^{j,h}(A_i,W_i)\vee\lambda^{j-1,h}(A_i,W_i)}{g_i(A_i,W_i)}\mathbf{1}\{\tau_i^h=j\}\middle|F_{i-1}\right] \\
\leq & E_P\left[\frac{\Delta_i^{j,h}\mathbf{1}\{\tau_i^h=j\}}{g_i(A_i,W_i)}\middle|F_{i-1}\right].
\end{aligned}
$$

We treat separately the case $j<J$ and the case $j=J$. We first start with the case $j<J$. If $\tau_i^h=j$, we must then have $\Delta_i^{j,h}>a_j$, which implies that

$$
E\left[\ell_i(b_i^{j,h})\middle|F_{i-1}\right]-\ell_i(b_i^{j,f}) = E_P\left[\frac{\Delta_i^{j,h}\mathbf{1}\{\tau_i^h=j\}}{g_i(A_i,W_i)}\middle|F_{i-1}\right]
$$

$$\begin{aligned}
&\leq \frac{1}{a_j} E\left[\frac{(\Delta_i^{j,h})^2}{g_i(A_i, W_i)}\middle| F_{i-1}\right] \\
&\leq \frac{1}{a_j} E\left[\sum_{a\in[K]} \big(h(a, W_i) - \lambda^{j,h}(a, W_i)\big)^2 \middle| F_{i-1}\right] \\
&\leq \frac{K\epsilon_j^2}{a_j}.
\end{aligned}$$

Therefore, for $j < J$,

$$E_P^A\left[\sup_{h\in\mathcal{F}} B_n^{j,f}\right] \leq \frac{K\epsilon_j^2}{a_j}.$$

Now consider the case $j = J$. We have that

$$\begin{aligned}
B_n^{J,h} &\leq \frac{1}{n}\sum_{i=1}^n E_P\left[\Delta_i^{J,h}\middle| F_{i-1}\right] \\
&= \sum_{i=1}^n E_P\left[\frac{h(A_i, W_i) - \lambda^{J,h}(A_i, W_i)}{g_i(A_i, W_i)}\middle| F_{i-1}\right] \\
&\leq \frac{1}{n}\sum_{i=1}^n E_P\left[\sum_{a\in[K]} h(a, W_i) - \lambda^{J,h}(a, W_i)\middle| F_{i-1}\right] \\
&\leq \frac{1}{n}\sqrt{n}\left(\sum_{i=1}^n E_P\left[\left(\sum_{a\in[K]} h(a, W_i) - \lambda^{J,h}(a, W_i)\right)^2 \middle| F_{i-1}\right]\right)^{1/2} \\
&\leq \left(\frac{K}{n}\sum_{i=1}^n E_P\left[\sum_{a\in[K]} (h(a, W_i) - \lambda^{J,h}(a, W_i))^2 | F_{i-1}\right]\right)^{1/2} \\
&\leq K\epsilon_J.
\end{aligned}$$

Therefore,

$$E_P^A\left[\sup_{h\in\mathcal{H}} B_n^{J,h}\right] \leq K\epsilon_J.$$

Control of the links. Observe that $\lambda^{j,} - \lambda^{j-1,h} = \lambda^{j,h} - h + h - \lambda^{j-1,h}$. Using that $\lambda^{j,h} \leq h$ and $\lambda^{j-1,h} \leq h$ the definitions of $\Delta_i^{j,h}$ and $\Delta^{j-1,h}$ yield

$$\begin{gathered}
-\Delta_i^{j,h} \leq (\lambda^{j,h} - h)(A_i, W_i)\mathbf{1}\{\tau_i^h > j\} \leq 0, \\
\text{and } 0 \leq (h - \lambda^{j-1,h})(A_i, W_i)\mathbf{1}\{\tau_i^h \geq j\} \leq \Delta_i^{j-1,h}.
\end{gathered}$$

Therefore, recalling the definition of $c_i^{j,h}$, we have that

$$-\Delta_i^{j,h}\mathbf{1}\{\tau_i^h > j\} \le c_i^{j,h} \le \Delta_i^{j-1,h}\mathbf{1}\{\tau_i^h \ge j\}.$$

Applying ℓ_i to $c_i^{j,h}$ amounts to multiplying it with a non-negative random variable. Therefore,

$$-\ell_i(\Delta_i^{j,h}\mathbf{1}\{\tau_i^h > j\} \le \ell_i(c_i^{j,h} \le) \le \ell_i(\Delta_i^{j,h}\mathbf{1}\{\tau_i^{j,h} \ge j\}),$$

and then

$$|\ell_i(c_i^{j,f})| \le \Delta_i^{j,f}\mathbf{1}\{\tau_i^f > j\} \vee \Delta_i^{j-1,f}\mathbf{1}\{\tau_i^f \ge j\}.$$

From the definition of $\tau_i^{j,f}$ and the fact that $(1 - Y_i) \in [0,1]$, we have that

$$|\ell_i(c_i^{j,f})| \le a_j \vee a_{j-1}.$$

Besides,

$$E_P\left[\ell_i(c_i^{j,f})^2 | F_{i-1}\right] \le 2\left\{E_P\left[\ell_i(\Delta_i^{j,f})^2 | F_{i-1}\right] + E_P\left[\ell_i(\Delta_i^{j-1,f})^2 | F_{i-1}\right]\right\}.$$

We have that, for all j,

$$\begin{aligned} E_P\left[\ell_i(\Delta_i^{j,h})^2 \middle| F_{i-1}\right] =& E_P\left[\frac{(f(A_i,W_i) - \lambda^{j,h}(A_i,W_i))^2(1-Y_i)^2}{g_i^2(A_i,W_i)}\middle| F_{i-1}\right] \\ \le& E_P\left[\sum_{a\in[K]} \frac{(f(a,W_i) - \lambda^{j,h}(a,W_i))^2}{g_i(a,W_i)}\middle| F_{i-1}\right] \\ \le& \delta^{-1}K\epsilon_j^2. \end{aligned}$$

Therefore, for all i, j,

$$E_P\left[(\ell_i(c_i^{j,h}))^2 | F_{i-1}\right] \le \delta^{-1}K(\epsilon_{j-1}^2 + \epsilon_j^2).$$

Observe that $C_n^{j,h}$ depends on h only through $\rho(0,h)$,…,$\rho(j,h)$. Therefore, as h varies over $\mathcal{H}$, $C_n^{j,h}$ varies over a collection of at most

$$\bar{N}_j \doteq \prod_{k=0}^{j} N_k$$

random variables. Therefore, from lemma 4.6,

$$E_P^A\left[\sup_{h\in\mathcal{H}} C_n^{j,h}\right] \le 4\sqrt{\frac{2K(\epsilon_j^2 + \epsilon_{j-1}^2)}{\delta n}\log\left(1 + \frac{\bar{N}_j}{P[A]}\right)} + \frac{8}{3}\frac{a_j \vee a_{j-1}}{\delta n}\log\left(1 + \frac{\bar{N}_j}{P[A]}\right)$$

End of the proof. Collecting the bounds on $E_P^A[\sup_{h\in\mathcal{H}} B_n^{j,h}]$, $E_P^A[\sup_{h\in\mathcal{H}} B_n^{j,h}]$ and $E_P^A[\sup_{h\in\mathcal{H}} C_n^{j,h}]$ yields

$$\begin{aligned}
E_P^A\left[\sup_{h\in\mathcal{H}}(P-P_n)\ell_{1:n}(h)\right] \leq & K\epsilon_J + \sum_{j=0}^{J-1}\frac{K\epsilon_j^2}{a_j} \\
& + 8\sqrt{\frac{K}{\delta n}\log\left(1+\frac{N_0}{P[A]}\right)} + \frac{8}{3}\frac{B}{\delta n}\log\left(1+\frac{N_0}{P[A]}\right) \\
& + \sum_{j=1}^{J} 8\sqrt{\frac{K}{\delta n}\log\left(1+\frac{\bar{N}_j}{P[A]}\right)} \\
& + \sum_{j=1}^{J}\frac{8}{3}\frac{a_{j-1}}{\delta n}\log\left(1+\frac{\bar{N}_j}{P[A]}\right).
\end{aligned}$$

Set

$$a_j = \epsilon_j\sqrt{\frac{\delta n}{K\log(1+\bar{N}_j/P[A])}}.$$

Replacing a_j in the previous display yields

$$\begin{aligned}
E_P^A\left[\sup_{f\in\mathcal{F}}(P-P_n)\ell_{1:n}(f)\right] \leq & K\epsilon_J + \frac{8}{3}\frac{B}{\delta n}\log\left(1+\frac{N_0}{P[A]}\right) \\
& + 20\sum_{j=0}^{J-1}\epsilon_j\sqrt{\frac{K}{\delta n}\log\left(1+\frac{\bar{N}_{j+1}}{P[A]}\right)}.
\end{aligned}$$

Since $(1+\bar{N}_j/P[A]) \leq (1+1/P[A])\prod_{k=0}^{j}(1+N_k)$, we have

$$\begin{aligned}
\sum_{j=1}^{J}\epsilon_{j-1}\sqrt{\log\left(1+\frac{\bar{N}_j}{P[A]}\right)} \leq & 2\sum_{j=0}^{J}\epsilon_j\sqrt{\log\left(1+\frac{1}{P[A]}\right)+\sum_{k=0}^{j}\log(1+N_k)} \\
\leq & 2\left(\sum_{j=0}^{J}\epsilon_j\right)\sqrt{\log\left(1+\frac{1}{P[A]}\right)} + 2\sum_{j=0}^{J}\epsilon_j\sum_{k=0}^{j}\sqrt{\log(1+N_k)}
\end{aligned}$$

We first look at the second term. We have that

$$\begin{aligned}
\sum_{j=0}^{J}\epsilon_j\sum_{k=0}^{j}\sqrt{\log(1+N_k)} = & \sum_{k=0}^{J}\sqrt{\log(1+N_k)}\sum_{j=k}^{J}2^{-j} \\
\leq & 2\sum_{k=j}^{J}2^{-k}\sqrt{\log(1+N_k)}
\end{aligned}$$

$$
\begin{aligned}
=&4\sum_{k=0}^{J}(\epsilon_k-\epsilon_{k+1})\sqrt{\log(1+N_k)}\\
\le&4\int_{\alpha/2}^{B}\sqrt{\log(1+N_{[]}(\epsilon,\mathcal{F},L_2(P)))}d\epsilon.
\end{aligned}
$$

Therefore, observing that $\sum_{j=0}^{J}\epsilon_j\le 2$, and gathering the previous bounds yields that

$$
\begin{aligned}
&E_P^A\left[\sup_{h\in\mathcal{H}}(P-P_n)\ell_{1:n}(h)\right]\\
\le&K\epsilon_J+160\sqrt{\frac{K}{\delta n}}\int_{\alpha/2}^{B}\sqrt{\log(1+N_{[]}(u^2,\mathcal{F},L_\infty(P))}du\\
&+\frac{8}{3}\frac{B}{\delta n}\log\left(1+N_{[]}(1,\mathcal{F},L_\infty(P))\right)\\
&+160\sqrt{\frac{K}{\delta n}}\sqrt{\log\left(1+\frac{1}{P[A]}\right)}+\frac{8}{3}\frac{B}{\delta n}\log\left(1+\frac{1}{P[A]}\right)\\
\le&H_n(v,\delta,\alpha)+160\sqrt{\frac{v}{n}}\sqrt{\log\left(1+\frac{1}{P[A]}\right)}+3\frac{B}{\delta n}\log\left(1+\frac{1}{P[A]}\right),
\end{aligned}
$$

with

$$
\begin{aligned}
H_n(v,\delta,\alpha)\doteq&K\alpha+160\sqrt{\frac{K}{\delta n}}\int_{\alpha/2}^{B}\sqrt{\log(1+N_{[]}(u^2,\mathcal{F},L_\infty(P)))}du\\
&+3\frac{B}{\delta n}\log\left(1+N_{[]}(1,\mathcal{F},L_\infty(P))\right).
\end{aligned}
$$

□

4.C.2 Maximal inequality for policy elimination

Theorem 4.6 (Maximal inequality under parameter-dependent IS ratio bound)**.** *Let $\mathcal{F}$ be a class of functions $\mathcal{A}\times\mathcal{W}\to[0,1]$. Suppose that we are under the contextual bandit setting described earlier, and that g_i is the F_{i-1}-measurable design at time point i. Let, for any $i\ge 1$, any $f\in\mathcal{F}$,*

$$V_i(f)\doteq E_P\left[\sum_{a\in[K]}\frac{f(a|W)}{g(a|W)}\right].$$

For any $n\ge 1$, $f\in\mathcal{F}$, denote

$$\bar{V}_n(f)\doteq\frac{1}{n}\sum_{i=1}^{n}V_i(f).$$

Let l be the direct policy optimization loss, and for all i, let ℓ_i be its importance-sampling weighted counterpart for time point i, that is, for all $f \in \mathcal{F}$, $o = (w, a, y) \in \mathcal{O}$,

$$\begin{aligned} l(f)(o) \doteq& \sum_{a \in [K]} C(a, W) f(a, W) \\ \text{and } \ell_i(f)(o) \doteq& \frac{f(a|w)}{g_i(a|w)}(1-y). \end{aligned}$$

Suppose that there exists $\delta > 0$ such, that for all $a, w \in \mathcal{A} \times \mathcal{W}$ and $i \in [n]$, $g_i(a|w) \geq \delta$.

Then, for all $x > 0$, $v > 0$, $\epsilon \in [0, 1]$

$$P\left[\sup_{f \in \mathcal{F}} \mathbf{1}\{\bar{V}_n(f) \leq v\}(P - P_n)\ell_{1:n}(f) \geq H_n(v, \delta, \epsilon) + 37\sqrt{\frac{vx}{n}} + 3\frac{x}{\delta n}\right] \leq 2e^{-x},$$

with

$$\begin{aligned} H_n(v, \delta, \epsilon) \doteq& \sqrt{v\epsilon} + 127\sqrt{\frac{v}{n}} \int_{\sqrt{\epsilon/2}}^1 \sqrt{\log(1 + N_{[]}(u^2, \mathcal{F}, L_\infty(P)))} du \\ &+ \frac{3}{\delta n} \log\left(1 + N_{[]}(1, \mathcal{F}, L_\infty(P))\right). \end{aligned}$$

The proof of the preceding theorem relies on the following lemma, which is a direct corollary of corollary A.8 in van Handel [2011].

Lemma 4.6 (Bernstein-like maximal inequality for finite sets)**.** *Let, for any $i \in [n], j \in [N]$, $X_{i,j}$ be an F_i-measurable random variable, and let, for any $j \in [N]$, $M_t^j \doteq \sum_{i=1}^n X_{i,j}$. Let for all $j \in [N]$,*

$$\sigma^2_{n,j} \doteq \frac{1}{n} \sum_{i=1}^n E_P[X^2_{i,j} | F_{i-1}].$$

Suppose that for all $i \in [n]$, $j \in [N]$, $|X_{i,j}| \leq b$ a.s. for some $b \geq 0$. Then, for any event $A \in \mathcal{F}$,

$$E^A\left[\max_{j \in [N]} \mathbf{1}\{\sigma^2_{n,j} \leq \sigma^2\} M_t^j\right] \leq 4\sigma \sqrt{\log\left(1 + \frac{N}{P[A]}\right)} + \frac{8}{3} b \log\left(1 + \frac{N}{P[A]}\right).$$

Proof of lemma 4.6. Observe that

$$\begin{aligned} \frac{2b^2}{n} \sum_{i=1}^n E\left[\phi\left(\frac{X_i}{b}\right) \middle| F_{i-1}\right] \leq& \frac{2b^2}{n} \sum_{i=1}^n \sum_{k \geq 2} \frac{b^{k-2}}{b^k k!} E[X_i^2 | F_{i-1}] \\ \leq& \frac{2}{n} \sum_{k \geq 2} \frac{1}{k!} \sum_{i=1}^n E[X_i^2 | F_{i-1}] \\ \leq& 2\phi(1)\sigma^2_{n,j} \\ \leq& 2\sigma^2_{n,j}. \end{aligned}$$

The conclusion follows from corollary A.8 in van Handel [2011]. □

Proof of theorem 4.6. The proof follows closely the proof of theorem A.4 in van Handel [2011]

From a conditional expectation bound to a deviation bound. Let $x > 0$ and let A be the event

$$A \doteq \left\{ \sup_{f \in \mathcal{F}} \mathbf{1}\{\bar{V}_n(f_{\leq} v\}(P - P_n)\ell_{1:n}(f) \geq \psi(x) \right\},$$

with

$$\psi(x) = H_n(v, \delta, \epsilon) + 37\sqrt{\frac{vx}{n}} + 3\frac{x}{\delta n}$$

Observe that for any $x > 0$,

$$\psi(x) \leq E_P^A \left[\sup_{f \in \mathcal{F}} \mathbf{1}\{\bar{V}_n(f) \leq v\}(P - P_n)\ell_{1:n}(f) \right].$$

Therefore, to prove the claim, it suffices to prove that

$$E_P^A \left[\sup_{f \in \mathcal{F}} \mathbf{1}\{\bar{V}_n(f) \leq v\}(P - P_n)\ell_{1:n}(f) \right] \leq \psi\left(\log\left(1 + \frac{1}{P[A]}\right)\right),$$

as this would imply

$$\Psi(x) \leq \psi\left(\log\left(1 + \frac{1}{P[A]}\right)\right) \leq \psi\left(\log\left(\frac{2}{P[A]}\right)\right),$$

which, as ψ is increasing, implies $P[A] \leq 2e^{-x}$, which is the wished claim.

Setting up the notation. For all $j \geq 0$, let $\epsilon_j = 2^{-j}$, and let

$$\mathcal{B}_j \doteq \{(\underline{f}^{j,\rho}, \overline{f}^{j,\rho}) : \rho = 1, \ldots, N_j\}$$

be an ϵ_j-bracketing of $\mathcal{F}$ in $L_\infty(P)$ norm. Further suppose that $\mathcal{B}_j$ is a minimal bracketing, that is that $N_j = N_{[\,]}(\epsilon_j, \mathcal{F}, L_\infty(P))$. For all j, f, let $\rho(j, f)$ be the index of a bracket of $\mathcal{B}_j$ that contains f, that is $\rho(j, f)$ is such that

$$\underline{f}^{j,\rho(j,f)} \leq f \leq \overline{f}^{j,\rho(j,f)}.$$

For all $f \in \mathcal{F}$, $j \geq 0$, $i \in [n]$ let

$$\lambda^{j,f} \doteq \underline{f}^{j,\rho(j,f)},$$

and

$$\Delta_i^{j,f} \doteq (f - \lambda^{j,f})(A_i, W_i).$$

Adaptive chaining. The core idea of the proof is a so-called adaptive chaining device: for any f, and any $i \in [n]$, we write

$$\begin{aligned} f(A_i, W_i) =& f(A_i, W_i) - \lambda^{\tau_i^f, f}(A_i, W_i) \vee \lambda^{\tau_i^f - 1, f}(A_i, W_i) \\ &+ \lambda^{\tau_i^f, f}(A_i, W_i) \vee \lambda^{\tau_i^f - 1, f}(A_i, W_i) - \lambda^{\tau_i^f - 1, f}(A_i, W_i) \\ &+ \sum_{j=1}^{\tau_i^f - 1} \lambda^{j,f}(A_i, W_i) \vee \lambda^{j-1,f}(A_i, W_i) \\ &+ \lambda^{0,f}(A_i, W_i), \end{aligned}$$

for some $\tau_i^f \geq 0$ that plays the role of the depth of the chain. We choose the depth τ_i^f so as to control the supremum norm of the links of the chain. Specifically, we let

$$\tau_i^f \doteq \min \left\{ j \geq 0 : \frac{\Delta_i^{j,f}}{g_i(A_i, W_i)} > a_j \right\} \wedge J,$$

for some $J \geq 1$, and a decreasing positive sequence a_j, which we will explicitly specify later in the proof. The chaining decomposition in 4.C.2 can be rewritten as follows:

$$\begin{aligned} f(A_i, W_i) =& \lambda^{0,f}(A_i, W_i) \\ &+ \sum_{j=0}^{J} \left\{ f(A_i, W_i) - \lambda^{j,f} \vee \lambda^{j-1,f}(A_i, W_i) \right\} \mathbf{1}\{\tau_i^f = j\} \\ &+ \sum_{j=1}^{J} \Big\{ \left(\lambda^{j,f}(A_i, W_i) \vee \lambda^{j-1,f}(A_i, W_i) - \lambda^{j-1,f}(A_i, W_i) \right) \mathbf{1}\{\tau_i^f = j)\} \\ &\qquad + \left(\lambda^{j,f}(A_i, W_i) - \lambda^{j-1,f}(A_i, W_i) \right) \mathbf{1}\{\tau_i^f > j\} \Big\} \end{aligned}$$

Denote $a_i^f \doteq \lambda^{0,f}(A_i, W_i)$,

$$b_i^{j,f} \doteq \left\{ f(A_i, W_i) - \lambda^{j,f} \vee \lambda^{j-1,f}(A_i, W_i) \right\} \mathbf{1}\{\tau_i^f = j\},$$

and

$$\begin{aligned} c_i^{j,f} \doteq& \left(\lambda^{j,f}(A_i, W_i) \vee \lambda^{j-1,f}(A_i, W_i) - \lambda^{j-1,f}(A_i, W_i) \right) \mathbf{1}\{\tau_i^f = j)\} \\ &+ \left(\lambda^{j,f}(A_i, W_i) - \lambda^{j-1,f}(A_i, W_i) \right) \mathbf{1}\{\tau_i^f > j\}. \end{aligned}$$

From the linearity of $\ell_1, \ldots, \ell_n$, we have that

$$(P - P_n)\ell_{1:n}(f) = A_n^f + \sum_{j=0}^{J} B_n^{j,f} + \sum_{j=1}^{J} C_n^{j,f},$$

with

$$\begin{aligned}
A_n^f \doteq& \frac{1}{n}\sum_{i=1}^n E[\ell_i(a_i^f)|F_{i-1}] - \ell_i(a_i^f),\\
B_n^{j,f} \doteq& \frac{1}{n}\sum_{i=1}^n E[\ell_i(b_i^{j,f})|F_{i-1}] - \ell_i(b_i^{j,f}),\\
C_n^{j,f} \doteq& \frac{1}{n}\sum_{i=1}^n E[\ell_i(c_i^{j,f}|F_{i-1}] - \ell_i(c_i^{j,f}).
\end{aligned}$$

The terms A_n^f, $B_n^{j,f}$ and $C_n^{j,f}$ can be intepreted as follows. For any given f and chain corresponding to f:

- A_n^f represents the root, at the coarsest level, of the chain,
- if the chain goes deeper than depth j, $C_n^{j,f}$ is the link of the chain between depths $j-1$ and j,
- if the chain stops at depth j, $B_n^{j,f}$ is the tip of the chain.

We control each term separately.

Control of the roots. Observe that, for all $i \in [n]$, $|\ell_i(a_i^f)| \leq \delta^{-1}$ a.s., and that

$$\begin{aligned}
E_P[\ell_i(a_i^f)^2|F_{i-1}] =& E_P\left[\frac{\lambda^{0,f}(A_i,W_i)^2}{g_i(A_i|W_i)^2}(1-Y_i)^2|F_{i-1}\right]\\
\leq& E_P\left[\sum_{a\in[K]}\frac{f(a,W_i)}{g_i(A_i|W_i)}|F_{i-1}\right]\\
=& V_i(f).
\end{aligned}$$

In the second line we have used that, $\lambda^{0,f}(A_i,W_i) \leq f(A_i,W_i)$, that $(1-Y_i) \in [0,1]$, and that $f(a,W_i) \in [0,1]$. Therefore, from lemma 4.6,

$$E_P^A\left[\sup_{f\in\mathcal{F}} \mathbf{1}\{\bar{V}_n(f) \leq v\} A_n^f\right] \leq 4\sqrt{\frac{v}{n}\log\left(1+\frac{N_0}{P[A]}\right)} + \frac{8}{3\delta n}\log\left(1+\frac{N_0}{P[A]}\right).$$

Control of the tips. As $\ell_i(b_i^{j,f}) \leq 0$, we have that

$$\begin{aligned}
&E_P[\ell_i(b_i^{j,f})|F_{i-1}] - \ell_i(b_i^{j,f})\\
\leq& E_P[\ell_i(b_i^{j,f})|F_{i-1}]\\
=& E_P\left[\frac{f(A_i,W_i) - \lambda^{j,f}(A_i,W_i) \vee \lambda^{j-1,f}(A_i,W_i)}{g_i(A_i,W_i)}(1-Y_i)\mathbf{1}\{\tau_i^f = j\}\middle| F_{i-1}\right]
\end{aligned}$$

$$\leq E_P \left[\frac{\Delta_i^{j,f}}{g_i(A_i, W_i)} \mathbf{1}\{\tau_i^f = j\} \middle| F_{i-1} \right]$$

We treat separately the case $j < J$ and the case $j = J$. We first start with the case $j < J$. If $\tau_i^f = j$, we must then have $\Delta_i^{j,f} / g_i(A_i, W_i) > a_j$, which implies that

$$\begin{aligned}
E\left[\ell_i(b_i^{j,f}) \middle| F_{i-1} \right] - \ell_i(b_i^{j,f}) \leq & \frac{1}{a_j} E\left[\frac{(\Delta_i^{j,f})^2}{g_i^2(A_i, W_i)} \middle| F_{i-1} \right] \\
\leq & \frac{1}{a_j} E\left[\sum_{a \in [K]} \frac{f(a, W_i)}{g_i(a, W_i)} (f(a, W_i) - \lambda^{j,f}(a, W_i)) \middle| F_{i-1} \right] \\
\leq & \frac{1}{a_j} V_i(f) \epsilon_j.
\end{aligned}$$

The second line above follows from the fact that $0 \leq f - \lambda^{j,f} \leq f$ since $0 \leq \lambda^{j,f} \leq f$. The third line above follows from the fact that $0 \leq (f - \lambda^{j,f})(a, W_i) \leq \|f - \lambda^{j,f}\|_\infty \leq \epsilon_j$. Therefore, for $j < J$,

$$E_P^A \left[\sup_{f \in \mathcal{F}} \mathbf{1}\{\bar{V}_n(f) \leq v\} B_n^{j,f} \right] \leq \frac{1}{a_j} v \epsilon_j.$$

Now consider the case $j = J$. We have that

$$\begin{aligned}
B_n^{J,f} \leq & \frac{1}{n} \sum_{i=1}^n E_P \left[\frac{(f - \lambda^{J,f})(A_i, W_i)}{g_i(A_i, W_i)} \middle| F_{i-1} \right] \\
\leq & \frac{1}{n} \sqrt{n} \left(\sum_{i=1}^n E_P \left[\frac{(f - \lambda^{J,f})^2 (A_i, W_i)}{g_i^2(A_i, W_i)} \middle| F_{i-1} \right] \right)^{1/2} \\
\leq & \left(\frac{1}{n} \sum_{i=1}^n E_P \left[\sum_{a \in [K]} \frac{f(a, W_i)}{g_i(a | W_i)} (f(a, W_i) - \lambda(a, W_i)) | F_{i-1} \right] \right)^{1/2} \\
\leq & \left(\frac{1}{n} \sum_{i=1}^n V_i(f) \epsilon_J \right)^{1/2} \\
\leq & \sqrt{v \epsilon_J}.
\end{aligned}$$

The second line follows from Cauchy-Schwartz and Jensen. The third line uses the same arguments as in the case $j < J$ treated before. Therefore,

$$E_P^A \left[\sup_{f \in \mathcal{F}} \mathbf{1}\{\bar{V}_n(f) \leq v\} B_n^{J,f} \right] \leq \sqrt{v \epsilon_J}.$$

Control of the links. Observe that $\lambda^{j,f} - \lambda^{j-1,f} = \lambda^{j,f} - f + f - \lambda^{j-1,f}$. Using that $\lambda^{j,f} \leq f$ and $\lambda^{j-1,f} \leq f$ the definitions of $\Delta_i^{j,f}$ and $\Delta^{j-1,f}$ yields

$$\begin{aligned} -\Delta_i^{j,f} \leq (\lambda^{j,f} - f)(A_i, W_i)\mathbf{1}\{\tau_i^f > j\} \leq 0, \\ \text{and } 0 \leq (f - \lambda^{j-1,f})(A_i, W_i)\mathbf{1}\{\tau_i^f \geq j\} \leq \Delta_i^{j-1,f}. \end{aligned}$$

Therefore, recalling the definition of $c_i^{j,f}$, we have that

$$-\Delta_i^{j,f}\mathbf{1}\{\tau_i^f > j\} \leq c_i^{j,f} \leq \Delta_i^{j-1,f}\mathbf{1}\{\tau_i^f \geq j\}.$$

Applying ℓ_i to $c_i^{j,f}$ amounts to multiplying it with a non-negative random variable. Therefore,

$$-\ell_i(\Delta_i^{j,f}\mathbf{1}\{\tau_i^f > j\} \leq \ell_i(c_i^{j,f} \leq) \leq \ell_i(\Delta_i^{j,f}\mathbf{1}\{\tau_i^{j,f} \geq j\}),$$

and then

$$|\ell_i(c_i^{j,f})| \leq \Delta_i^{j,f}\mathbf{1}\{\tau_i^f > j\} \vee \Delta_i^{j-1,f}\mathbf{1}\{\tau_i^f \geq j\}.$$

From the definition of $\tau_i^{j,f}$ and the fact that $(1 - Y_i) \in [0, 1]$, we have that

$$|\ell_i(c_i^{j,f})| \leq a_j \vee a_{j-1}.$$

Besides,

$$E_P\left[\ell_i(c_i^{j,f})^2 | F_{i-1}\right] \leq 2\left\{E_P\left[\ell_i(\Delta_i^{j,f})^2 | F_{i-1}\right] + E_P\left[\ell_i(\Delta_i^{j-1,f})^2 | F_{i-1}\right]\right\}.$$

We have that, for all j,

$$\begin{aligned} E_P\left[(\ell_i(\Delta_i^{j,f}))^2 | F_{i-1}\right] =& E_P\left[\frac{(f(A_i, W_i) - \lambda^{j,f}(A_i, W_i))^2}{g_i(A_i|W_i)^2}(1 - Y_i)^2 \middle| F_{i-1}\right] \\ \leq& E_P\left[\sum_{a \in [K]} \frac{f(a, W_i)}{g_i(a|W_i)}(f(a, W_i) - \lambda^{j,f}(a, W_i)) \middle| F_{i-1}\right] \\ \leq& V_i(f)\epsilon_j. \end{aligned}$$

Therefore, for all i, j,

$$E_P\left[(\ell_i(c_i^{j,f}))^2 | F_{i-1}\right] \leq V_i(f)(\epsilon_j + \epsilon_{j-1}).$$

Observe that $C_n^{j,f}$ depends on f only through $\rho(0, f)$,...,$\rho(j, f)$. Therefore, as f varies over $\mathcal{F}$, $C_n^{j,f}$ varies over a collection of at most

$$\bar{N}_j \doteq \prod_{k=0}^{j} N_k$$

random variables. Therefore, from lemma 4.6,

$$\begin{aligned} E_P^A\left[\sup_{f \in \mathcal{F}} \mathbf{1}\{\bar{V}_n(f) \leq v\} C_n^{j,f}\right] \leq& 4\sqrt{\frac{v(\epsilon_j + \epsilon_{j-1})}{n} \log\left(1 + \frac{\bar{N}_j}{P[A]}\right)} \\ &+ \frac{8}{3}\frac{a_j \vee a_{j-1}}{n} \log\left(1 + \frac{\bar{N}_j}{P[A]}\right) \end{aligned}$$

End of the proof. Collecting the bounds on $E_P^A[\sup_{f\in\mathcal{F}} \mathbf{1}\{\bar{V}_n(f) \leq v\}B_n^{j,f}]$, $E_P^A[\sup_{f\in\mathcal{F}} \mathbf{1}\{\bar{V}_n(f) \leq v\}B_n^{j,f}]$ and $E_P^A[\sup_{f\in\mathcal{F}} \mathbf{1}\{\bar{V}_n(f) \leq v\}C_n^{j,f}]$ yields

$$\begin{aligned}E_P^A\left[\sup_{f\in\mathcal{F}} \mathbf{1}\{\bar{V}_n(f) \leq v\}(P-P_n)\ell_{1:n}(f)\right] \leq & \sqrt{v\epsilon_J} + \sum_{j=0}^{J-1} \frac{v\epsilon^j}{a_j} \\ & + 4\sqrt{\frac{v}{n}\log\left(1+\frac{N_0}{P[A]}\right)} + \frac{8}{3\delta n}\log\left(1+\frac{N_0}{P[A]}\right) \\ & + \sum_{j=1}^{J} 4\sqrt{\frac{2\epsilon_{j-1}v}{n}\log\left(1+\frac{\bar{N}_j}{P[A]}\right)} \\ & + \sum_{j=1}^{J} \frac{8}{3}\frac{a_{j-1}}{n}\log\left(1+\frac{\bar{N}_j}{P[A]}\right).\end{aligned}$$

Set

$$a_j = \frac{3}{8}\sqrt{\frac{nv\epsilon_j}{\log(1+\bar{N}_{j+1}/P[A])}}.$$

Replacing a_j in the previous display yields

$$\begin{aligned}E_P^A\left[\sup_{f\in\mathcal{F}} \mathbf{1}\{\bar{V}_n(f) \leq v\}(P-P_n)\ell_{1:n}(f)\right] \leq & \sqrt{v\epsilon_J} + \frac{8}{3\delta n}\log\left(1+\frac{N_0}{P[A]}\right) \\ & + \sum_{j=0}^{J}(8+2\sqrt{2})\sqrt{\frac{v\epsilon_j}{n}\log\left(1+\frac{\bar{N}_j}{P[A]}\right)}.\end{aligned}$$

Since $(1+\bar{N}_j/P[A]) \leq (1+1/P[A])\prod_{k=0}^{j}(1+N_k)$, we have

$$\begin{aligned}\sum_{j=0}^{J}\sqrt{\epsilon_j\log\left(1+\frac{\bar{N}_j}{P[A]}\right)} \leq & \sum_{j=0}^{J}\sqrt{\epsilon_j}\sqrt{\log\left(1+\frac{1}{P[A]}\right)+\sum_{k=0}^{j}\log(1+N_k)} \\ \leq & \left(\sum_{j=0}^{J}\sqrt{\epsilon_j}\right)\sqrt{\log\left(1+\frac{1}{P[A]}\right)} + \sum_{j=0}^{J}\sqrt{\epsilon_j}\sum_{k=0}^{j}\sqrt{\log(1+N_k)}\end{aligned}$$

We first look at the second term. We have that

$$\begin{aligned}\sum_{j=0}^{J}\sqrt{\epsilon_j}\sum_{k=0}^{j}\sqrt{\log(1+N_k)} = & \sum_{k=0}^{J}\sqrt{\log(1+N_k)}\sum_{j=k}^{J}(\sqrt{2})^{-j} \\ \leq & \frac{\sqrt{2}}{\sqrt{2}-1}\sum_{k=0}^{J}(\sqrt{2})^{-k}\sqrt{\log(1+N_k)}\end{aligned}$$

$$= \left(\frac{\sqrt{2}}{\sqrt{2}-1} \right)^2 \sum_{k=0}^{J} (\epsilon_k - \epsilon_{k+1}) \sqrt{\log(1+N_k)}.$$

Letting $u_k = \sqrt{\epsilon_k}$, we have that $N_k = N_{[]}(u_k^2, \mathcal{F}, L_\infty(P))$ and thus

$$\sum_{j=0}^{J} \sqrt{\epsilon_j} \sum_{k=0}^{j} \sqrt{\log(1+N_k)} \leq \int_{u_{J+1}}^{u_1} \sqrt{\log(1 + N_{[]}(u^2, \mathcal{F}, L_\infty(P)))} du.$$

Therefore, observing that $\sum_{j=0}^{J} \sqrt{\epsilon_j} \leq \sqrt{2}/(\sqrt{2}-1)$, and gathering the previous bounds yields that

$$\begin{aligned}
& E_P^A \left[\sup_{f \in \mathcal{F}} \mathbf{1}\{\bar{V}_n(f) \leq v\} (P - P_n) \ell_{1:n}(f) \right] \\
\leq & \sqrt{v \epsilon_J} + (8 + 2\sqrt{2}) \left(\frac{\sqrt{2}}{\sqrt{2}-1} \right)^2 \sqrt{\frac{v}{n}} \int_{\sqrt{\epsilon_J / 2}}^{1} \sqrt{\log(1 + N_{[]}(u^2, \mathcal{F}, L_\infty(P))} du \\
& + \frac{8}{3} \frac{1}{\delta n} \log \left(1 + N_{[]}(1, \mathcal{F}, L_\infty(P)) \right) \\
& + \frac{\sqrt{2}}{\sqrt{2}-1} (8 + 2\sqrt{2}) \sqrt{\frac{v}{n}} \sqrt{\log \left(1 + \frac{1}{P[A]} \right)} + \frac{8}{3\delta n} \log \left(1 + \frac{1}{P[A]} \right) \\
\leq & H_n(v, \delta, \epsilon_J) + 37 \sqrt{\frac{v}{n}} \sqrt{\log \left(1 + \frac{1}{P[A]} \right)} + \frac{3}{\delta n} \log \left(1 + \frac{1}{P[A]} \right),
\end{aligned}$$

with

$$\begin{aligned}
H_n(v, \delta, \epsilon) \doteq & \sqrt{v \epsilon} + 127 \sqrt{\frac{v}{n}} \int_{\sqrt{\epsilon/2}}^{1} \sqrt{\log(1 + N_{[]}(u^2, \mathcal{F}, L_\infty(P)))} du \\
& + \frac{3}{\delta n} \log \left(1 + N_{[]}(1, \mathcal{F}, L_\infty(P)) \right).
\end{aligned}$$

□

4.D Regret analysis of the policy evaluation algorithm

4.D.1 Definition of v_τ and constants in the definition of x_τ

For all $\delta > 0$, $v > 0$, $p > 0$, $\tau \geq 1$, let

$$a_\tau(\epsilon, \delta, v, p) \doteq \sqrt{v} \left\{ \frac{c_1(c,p)}{\tau^{\frac{1}{2} \wedge \frac{1}{2p}}} + \frac{c_2}{\sqrt{\tau}} \sqrt{\log \left(\frac{\tau(\tau+1)}{\epsilon} \right)} + \frac{1}{\delta \tau} \left(c_3 + c_4 \log \left(\frac{\tau(\tau+1)}{\epsilon} \right) \right) \right\},$$

with

$$c_1(c,p) \doteq \begin{cases} \frac{127\sqrt{c}}{1-p} & \text{if } p \in (0,1) \\ 1 + \frac{127\sqrt{c}2^{\frac{p-1}{2}}}{p-1} & \text{if } p > 1, \end{cases}$$

$c_2 = 37$, $c_3 = 3\log 2$, and $c_4 = 3$. For all $\delta > 0$, $v > 0$, $\tau \geq 1$, let

$$b_\tau(\epsilon, \delta, v) \doteq c_5 \sqrt{\frac{v}{\tau} \log\left(\frac{\tau(\tau+1)}{\epsilon}\right)} + \frac{c_6}{\delta\tau} \log\left(\frac{\tau(\tau+1)}{\epsilon}\right),$$

with $c_5 = c_6 = 2$. For all $\delta > 0$, $v > 0$, $\tau \geq 1$, $p > 0$, let

$$x_\tau(\epsilon, \delta, v, p) \doteq 2(a_\tau(\epsilon, \delta, v, p) + b_\tau(\epsilon, \delta, v)).$$

For all $\delta > 0$, $\tau \geq 1$, let

$$v_\tau(\epsilon, \delta) \doteq 2K + \delta^{-1} \left\{ \frac{c_1'(c,p)}{\tau^{\frac{1}{2} \wedge \frac{1}{p}}} + \frac{32}{\sqrt{\tau}} \sqrt{\log\left(\frac{\tau(\tau+1)}{\epsilon}\right)} + \frac{16 \log 2}{\tau} + \frac{16}{\tau} \log\left(\frac{\tau(\tau+1)}{\epsilon}\right) \right\},$$

with

$$c_1'(c,p) \doteq \begin{cases} \frac{64\sqrt{c}}{1-p/2} & \text{if } p \in (0,2), \\ 1 + \frac{64 \times 2^{p/2-1}\sqrt{c}}{p/2-1} & \text{if } p > 1. \end{cases}$$

The quantity v_τ from the main text is defined as $v_\tau \doteq v_\tau(\epsilon, \delta_\tau)$.

We can now give the explicit definitions of the sequences (δ_t) and (x_t). For all $\tau \geq 1$, let

$$\delta_\tau \doteq \tau^{-\left(\frac{1}{2} \wedge \frac{1}{2p}\right)} \qquad \text{and} \qquad x_\tau \doteq x_\tau(\epsilon, \delta_\tau, v_\tau(\epsilon, \delta_\tau), p).$$

The constant c_7 in the main text is defined as $c_7 \doteq c_4 + c_6$.

4.D.2 Proofs

Lemma 4.7 (Bound in the max IS ratio in terms of max empirical IS ratio)**.** . *Consider a class of policies $\mathcal{F}$ as in the current section. Suppose that $g : \mathcal{A} \times \mathcal{W} \to [0,1]$ is such that g is uniformly lower bounded by some $\delta > 0$, that is, for all $a, w \in \mathcal{A} \times \mathcal{W}, g(a,w) \geq \delta$.*

*Suppose that assumption **A1** holds. Then, for all $\epsilon > 0$,*

$$P\left[\sup_{f \in \mathcal{F}} (P - P_n) \left\{ \sum_{a \in [K]} \frac{f(a|W)}{g(a|W)} \right\} \geq v_n(\epsilon, \delta) - 2K \right] \leq 2\frac{\epsilon}{n(n+1)}.$$

The proof of lemma 4.7 relies on the following result, which is a slighlty modified version of corollary 6.9 in Massart [2007]. The only differences are that

- we state it with lower bound of the entropy integral $\alpha/2 > 0$, instead of 0, which makes appear an approximation error term α,
- we state it for i.i.d. random variables instead of independent random variables, we set to 1 the value of ϵ in the original statement of the theorem.

Proposition 4.3. *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$. Let $X_1, \ldots, X_n$ be i.i.d. random variables with domain $\mathcal{X}$ and common marginal distribution P. Suppose that there exists σ and b such that, for all $f \in \mathcal{F}$, for any $k \geq 2$,*

$$E_P[|f(X)|^k] \leq \frac{k!}{2}\sigma^2 b^{k-2}.$$

Assume that for all $\epsilon > 0$, there exists a set of brackets $\mathcal{B}(\epsilon, b)$ covering $\mathcal{F}$ such that, for all bracket $[l, u]$ in $\mathcal{B}(\epsilon, \delta)$,

$$E[((u-l)(X))^k] \leq \frac{k!}{2}\epsilon^2 b^{k-2}.$$

We call such a $\mathcal{B}(\epsilon, \delta)$ an (ϵ, b) bracketing of $\mathcal{F}$, and we denote $\mathcal{N}_{[]}(\epsilon, b\mathcal{F})$ the minimal cardinality of such an $\mathcal{B}(\epsilon, b)$.

Then, for all $\alpha \in (0, \sigma)$, and for all $x > 0$,

$$P\left[\sup_{\in\mathcal{F}}(P - P_n)f \geq H_n(\alpha, \sigma, b) + 10\sigma\sqrt{\frac{x}{n}} + 2bx\right] \leq e^{-x},$$

where

$$H_n(\alpha, \sigma, b) \doteq \alpha + \frac{27}{\sqrt{n}}\int_{\alpha/2}^{\sigma}\sqrt{\log \mathcal{N}_{[]}(\epsilon, b, \mathcal{F})}d\epsilon + \frac{2(\sigma + b)}{n}\log \mathcal{N}_{[]}(\sigma, b, \mathcal{F}).$$

Proof of proposition 4.3. It suffices to choose J in the proof of corollary 6.9 in Massart [2007] such that $\alpha/2 \leq \epsilon_J < \alpha$, and not let it go to ∞ at the end of the proof. □

Proof of lemma 4.7. Let

$$\mathcal{H} \doteq \left\{ h : w \mapsto \sum_{a \in [K]} \frac{f(a, w)}{g(a, w)} : f \in \mathcal{F} \right\}.$$

Observe that, for all $h \in \mathcal{H}$, $h(W)| \leq \delta^{-1}$, as $g \geq \delta$, and thus $E_P[h^2(W)] \leq \delta^{-2}$. Observe that an ϵ-bracketing of $\mathcal{F}$ in $L_2(P)$ induces a $(\sqrt{K}\epsilon\delta^{-1}, b)$ bracketing of $\mathcal{H}$ in the sense of proposition 4.3. Therefore, from proposition 4.3

$$P\left[\sup_{f \in \mathcal{F}}(P - P_n)f \geq v_n(\epsilon, \delta) - 2K\right] \leq \frac{\epsilon}{n(n+1)}.$$

□

The following lemma shows that, with high probability, the policy elimination algorithm doesn't eliminate the optimal policy.

Lemma 4.8. *Suppose that **A1** holds. Suppose $(x_t(\epsilon))$ is as specified in subsection 4.2.1. Then, for all $t \geq 1$,*

$$P[f^* \in \mathcal{F}] \geq 1 - 3\epsilon.$$

Proof. Denote $\hat{f}_\tau \doteq \arg\min_{f \in \mathcal{F}_\tau} \hat{R}_\tau(f)$. We have that

$$\begin{aligned}
\hat{R}_\tau(f^*) - \hat{R}_\tau(\hat{f}_\tau) \leq & R(f^*) - R(\hat{f}_\tau) \\
& + \hat{R}_\tau(f^*) - R(f^*) \\
& + R(\hat{f}_\tau) - \hat{R}_\tau(\hat{f}_\tau) \\
\leq & \hat{R}_\tau(f^*) - R(f^*) \\
& + \sup_{f \in \mathcal{F}_\tau} R(f) - \hat{R}_\tau(f).
\end{aligned}$$

Define the event

$$\mathcal{E}_{1,t} \doteq \left\{ \forall \tau \in [t] : \sup_{f \in \mathcal{F}_\tau} V(g_\tau, f) \leq v_\tau(\epsilon, \delta_\tau) \right\},$$

where $v_\tau(\epsilon, \delta_\tau)$ is defined in subsection 4.2.1. From lemma 4.7,

$$P[\mathcal{E}_{1,t}] \geq 1 - 2\epsilon.$$

For all $\tau \in [t]$, define the event

$$\mathcal{E}_{2,t} \doteq \left\{ \max_{\tau \in [t]} \sup_{f \in \mathcal{F}_\tau} R(f) - \hat{R}_\tau(f) \leq a_\tau(\epsilon, \delta_\tau, v_\tau(\epsilon, \delta_\tau), p) \right\},$$

where a_τ is defined in subsection 4.2.1. From theorem 4.6,

$$P[\mathcal{E}_{2,t}^c, \mathcal{E}_{1,t}] \leq \epsilon.$$

We now turn to controlling $\hat{R}_\tau(f^*) - R(f^*)$. So as to be able to obtain a high probability bound scaling as $\sqrt{v_\tau(\epsilon, \delta_\tau)/\tau}$, we need f^* to be in $\mathcal{F}_\tau$. As we are about to show, if the desired bound holds, that $\mathcal{E}_{1,t} \cap \mathcal{E}_{2,t}$ holds, and that $f^* \in \mathcal{F}_\tau$, them we will have that $f^* \in \mathcal{F}_{\tau+1}$. This motivates a reasoning by induction.

Let, for all $\tau \in [t]$,

$$\mathcal{E}_{3,\tau} \doteq \left\{ \hat{R}_\tau(f^*) - R(f^*) \leq b_\tau(\epsilon, \delta_\tau, v(\epsilon, \delta_\tau)) \right\},$$

where b_τ is defined in subsection 4.2.1. We are going to show by induction that for all $\tau \in [t]$,

$$P\left[\mathcal{E}_{3,t}^c, \mathcal{E}_{1,t}, \mathcal{E}_{2,t}\right] \leq \sum_{s=1}^{\tau} \frac{\epsilon}{s(s+1)}.$$

By convention, we let $\mathcal{E}_{3,0} \doteq \{f^* \in \mathcal{F}\}$. and $\sum_{s=1}^{0} 1/(s(s+1)) = 0$. The induction claim thus trivially holds at $\tau = 0$. Consider $\tau \in [t]$. Suppose that

$$P[\mathcal{E}_{3,\tau-1}^c, \mathcal{E}_{1,t}, \mathcal{E}_{2,t}] \leq \sum_{s=1}^{\tau-1} \frac{\epsilon}{s(s+1)}.$$

Observe that $\mathcal{E}_{\tau-1} \cap \mathcal{E}_{1,t} \cap \mathcal{E}_{2,t}$ implies $f^* \in \mathcal{F}_\tau$ as we then have

$$\begin{aligned} \hat{R}_\tau(f^*) - \hat{R}_\tau(\hat{f}_\tau) \leq & a_{\tau-1}(\epsilon, \delta_{\tau-1}, v_{\tau-1}(\epsilon, \delta_{\tau-1}), p) + b_{\tau-1}(\epsilon, \delta_{\tau-1}, v_{\tau-1}(\epsilon, \delta_{\tau-1})) \\ < & x_{\tau-1}(\epsilon, \delta_{\tau-1}, v_{\tau-1}(\epsilon, \delta_{\tau-1})). \end{aligned}$$

Using this fact, distinguishing the cases $\mathcal{E}_{3,\tau-1}$ and $\mathcal{E}_{3,\tau-1}^c$, and using the induction hypothesis yields

$$\begin{aligned} P[\mathcal{E}_{3,\tau}^c, \mathcal{E}_{1,t}, \mathcal{E}_{2,t}] \leq & P[\mathcal{E}_{3,\tau}^c, \mathcal{E}_{3,\tau-1}, \mathcal{E}_{1,t}, \mathcal{E}_{2,t}] + P[\mathcal{E}_{3,\tau-1}^c, \mathcal{E}_{1,t}, \mathcal{E}_{2,t}] \\ \leq & P[\mathcal{E}_{3,\tau}^c, f^* \in \mathcal{F}, \mathcal{E}_{2,t}] + \sum_{s=1}^{\tau-1} \frac{\epsilon}{s(s+1)}. \end{aligned}$$

Observe that under $\{f^* \in \mathcal{F}_\tau\} \cap \mathcal{E}_{2,t}$, we have that $V(g_\tau, f^*) \leq v_\tau(\epsilon, \delta_\tau)$ and thus

$$E[(\ell_\tau(f^*)(O_\tau))^2 | F_{\tau-1}] \leq K v_\tau(\epsilon, \delta_\tau).$$

Besides, $|\ell_\tau(f^*)(O_\tau) - E[\ell_\tau(f^*)(O_\tau)|F_{\tau-1}] \leq \delta_\tau^{-1}$. Therefore, from Bernstein's inequality for martingales

$$P[\mathcal{E}_{3,\tau}^c, f^* \in \mathcal{F}, \mathcal{E}_{2,t}] \leq \frac{\epsilon}{\tau(\tau+1)}.$$

Therefore,

$$P[\mathcal{E}_{3,\tau}^c, \mathcal{E}_{1,t}, \mathcal{E}_{2,t}] \leq \sum_{s=1}^{\tau} \frac{\epsilon}{s(s+1)}.$$

We have thus shown that, for all $\tau \in [t]$,

$$P[\mathcal{E}_{3,\tau}^c, \mathcal{E}_{1,t}, \mathcal{E}_{2,t}] \leq \sum_{s=1}^{\tau} \frac{\epsilon}{s(s+1)}.$$

Therefore,

$$\begin{aligned} P[\mathcal{E}_{3,t}, \mathcal{E}_{1,t}, \mathcal{E}_{2,t}] = & P[\mathcal{E}_{1,t}, \mathcal{E}_{2,t}] - P[\mathcal{E}_{3,t}^c, \mathcal{E}_{1,t}, \mathcal{E}_{2,t}] \\ = & P[\mathcal{E}_{1,t}] - P[\mathcal{E}_{1,t}, \mathcal{E}_{2,t}^c] - P[\mathcal{E}_{3,t}^c, \mathcal{E}_{1,t}, \mathcal{E}_{2,t}] \\ = & 1 - P[\mathcal{E}_{1,t}^c] - P[\mathcal{E}_{1,t}, \mathcal{E}_{2,t}^c] - P[\mathcal{E}_{1,t}, \mathcal{E}_{2,t}, \mathcal{E}_{3,t}^c] \\ \geq & 1 - 4\epsilon. \end{aligned}$$

□

The following lemma gives a bound on $\sup_{f \in \mathcal{F}_\tau} R(f) - R(f^*)$ which holds uniformly in time with high probability.

Lemma 4.9. *Consider algorithm 4.1. Make assumption **A1**. Then, with probability* $1 - 4\epsilon$, *we have that, for all* $\tau \in [t]$,

$$\sup_{f \in \mathcal{F}_\tau} R(f) - R(f^*) \leq 2x_\tau.$$

Proof. Observe that, for all $f \in \mathcal{F}$,

$$\begin{aligned} R(f) - R(f^*) =& \hat{R}_\tau(f) - \hat{R}_\tau(f^*) \\ & + R(f) - \hat{R}_\tau(f) \\ & - R(f^*) - \hat{R}_\tau(f^*)) \\ \leq& \hat{R}_\tau(f) - \hat{R}_\tau(\hat{f}_\tau) \\ & + \sup_{f \in \mathcal{F}_\tau} (R(f) - \hat{R}_t(f)) \\ & - (R(f^*) - \hat{R}_t(f^*)) \\ \leq& x_\tau \\ & + \sup_{f \in \mathcal{F}_\tau} (R(f) - \hat{R}_t(f)) \\ & - (R(f^*) - \hat{R}_t(f^*)). \end{aligned}$$

Define the events

$$\begin{aligned} \mathcal{E}_{1,t} \doteq & \left\{ \forall \tau \in [t], \sup_{f \in \mathcal{F}_\tau} V(g_\tau, f) \leq v_\tau(\epsilon, \delta_\tau) \right\}, \\ \mathcal{E}_{2,t} \doteq & \left\{ f^* \in \mathcal{F}_t \right\}. \end{aligned}$$

From lemma 4.8,

$$P[\mathcal{E}_{1,t}] \geq 1 - 4\epsilon.$$

Under $\mathcal{E}_{1,t}$, we have that, for all $f \in \mathcal{F}_\tau$,

$$E\left[(\ell_\tau(f)(O_\tau))^2 | F_{\tau-1} \right] \leq K v_\tau(\epsilon, \delta_\tau).$$

Therefore, using also that $|\ell_\tau(f)(O_\tau)| \leq \delta_\tau^{-1}$, theorem 4.6 gives us that, for all $\tau \in [t]$,

$$P\left[\sup_{f \in \mathcal{F}_\tau} R(f) - \hat{R}_\tau(f) \geq a_\tau(\epsilon, v_\tau(\epsilon, \delta_\tau), \delta_\tau, p), \mathcal{E}_{1,t} \right] \leq \frac{\epsilon}{\tau(\tau+1)},$$

which, by a union bound gives us that

$$P\left[\mathcal{E}_{3,t}^c, \mathcal{E}_{1,t} \right] \leq \epsilon,$$

with

$$\mathcal{E}_{3,t} \doteq \left\{ \forall \tau \in [t], \sup_{f \in \mathcal{F}_\tau} R(f) - \hat{R}_\tau(f) \leq a_\tau(\epsilon, v_\tau(\epsilon, \delta_\tau), \delta_\tau, p) \right\}.$$

We now consider the term $\hat{R}_\tau(f) - R(f^*)$. We have that

$$\hat{R}_\tau(f^*) - R(f^*) = \frac{1}{t} \sum_{\tau=1}^{t} \ell_\tau(f^*)(O_\tau) - E[\ell_\tau(f^*)(O_\tau) | F_{\tau-1}]$$

Under $\mathcal{E}_{1,t} \cap \mathcal{E}_{2,t}$, each term in the sum satisfies

$$E_P \left[(\ell_\tau(f^*)(O_\tau))^2 | F_{\tau-1} \right] \leq K v_\tau(\epsilon, \delta_\tau)$$

and

$$|\ell_\tau(f^*)(O_\tau) - E_P \left[\ell_\tau(f^*)(O_\tau) | F_{\tau-1} \right] \leq \delta_\tau^{-1}.$$

Therefore, from Bernstein's inequality and a union bound, letting

$$\mathcal{E}_{4,t} \doteq \left\{ \forall \tau \in [t], \hat{R}_\tau(f^*) - R(f^*) \leq b_\tau(\delta, v_\tau(\epsilon, \delta_\tau), \delta_\tau) \right\},$$

we have that

$$P[\mathcal{E}_{4,t}^c, \mathcal{E}_{1,t}, \mathcal{E}_{2,t}] \leq \epsilon.$$

Observe that under $\mathcal{E}_{3,t} \cap \mathcal{E}_{4,t}$ it holds that

$$\forall \tau \in [t] \sup_{f \in \mathcal{F}_\tau} R(f) - R(f^*) \leq x_\tau.$$

Therefore, to conclude the proof, it suffices to bound $P[\mathcal{E}_{3,t}, \mathcal{E}_{4,t}]$. We have that

$$\begin{aligned} P[(\mathcal{E}_{3,t} \cap \mathcal{E}_{4,t})^c] \leq & P[\mathcal{E}_{1,t}, \mathcal{E}_{2,t}, (\mathcal{E}_{3,t} \cap \mathcal{E}_{4,t})^c] + P[\mathcal{E}_{1,t}^c] + P[\mathcal{E}_{2,t}^c] \\ \leq & P[\mathcal{E}_{1,t}, \mathcal{E}_{2,t}, \mathcal{E}_{3,t}^c] \\ & + P[\mathcal{E}_{1,t}, \mathcal{E}_{2,t}, \mathcal{E}_{4,t}^c] + P[\mathcal{E}_{1,t}^c] + P[\mathcal{E}_{2,t}^c] \\ \leq & 6\epsilon, \end{aligned}$$

which yields the wished claim. □

We can now prove theorem 4.1.

Proof of theorem 4.1. Observe that

$$\sum_{\tau=1}^{t} \mathcal{V}(f^*) - Y_\tau = \sum_{\tau=1}^{t} (1 - Y_\tau) - E_P[(1 - Y_\tau) | F_{\tau-1}]$$

$$+\sum_{\tau=1}^{t} E_P[(1-Y_\tau)|F_{\tau-1}] - R(f^*).$$

Since $(1-Y_\tau) \in [0,1]$, from Azuma-Hoeffding, we have that, with probability at least $1-\epsilon$,

$$\sum_{\tau=1}^{t}(1-Y_\tau) - E_P[(1-Y_\tau)|F_{\tau-1}] \leq \sqrt{t \log\left(\frac{1}{\epsilon}\right)}.$$

Observe that

$$E_P\left[(1-Y_\tau)|F_{\tau-1}\right] = R(g_\tau) = \delta_\tau R(g_{ref}) + (1-\delta_\tau)R(\tilde{g}_\tau),$$

where $\tilde{g}_\tau \in \mathcal{F}_\tau$. Therefore,

$$\begin{aligned} E_P\left[(1-Y_\tau)|F_{\tau-1}\right] - R(f^*) \leq & \delta_\tau(R(g_{ref}) - R(f^*)) + (1-\delta_\tau)(R(\tilde{g}_\tau) - R(f^*)) \\ \leq & \delta_\tau + (R(\tilde{g}_\tau) - R(f^*)) \end{aligned}$$

From lemma 4.9, with probability $1-6\epsilon$, for all $\tau \in [t]$,

$$R(\tilde{g}_\tau) - R(f^*) \leq x_\tau.$$

Therefore, with probability at least $1-7\epsilon$, we have the wished bound. □

4.E Regret analysis of the ε-greedy algorithm

4.E.1 Regret decomposition

Using in particular the linearity of $\pi \mapsto R(\pi)$ and the definition of g_t, we have that

$$\begin{aligned} & Y_t - R(\pi^*) \\ =& Y_t - E[Y_t|F_{t-1}] + E[Y_t|F_{t-1}] - R(\pi^*) \\ =& Y_t - E[Y_t|F_{t-1}] + R(g_t) - R(\pi^*) \\ =& \underbrace{Y_t - E[Y_t|F_{t-1}]}_{\text{reward noise}} + \underbrace{\delta_t(R(g_{ref}) - R(\pi^*))}_{\text{exploration cost}} \\ & + (1-\delta_t)\underbrace{(R(\hat{\pi}_{t-1}) - R(\pi^*))}_{\text{exploitation cost}}. \end{aligned} \tag{4.16}$$

4.E.2 Proof of deviations inequalities

Proof of theorem 4.3. Observe that

$$R^\phi(\hat{f}_t) - R^\phi(f^*_{\mathcal{F}}) = \frac{1}{t}\sum_{\tau=1}^{t} E\left[\ell_\tau(f)(O_\tau) - \ell_\tau(f^*_{\mathcal{F}})(O_\tau)|F_{\tau-1}\right]\big|_{f=\hat{f}_t}$$

$$
\begin{aligned}
=&\frac{1}{t}\sum_{\tau=1}^{t} \ell_\tau^\phi(\hat{f}_t)(O_\tau) - \ell_\tau^\phi(f^*_{\mathcal{F}})(O_\tau) \\
&+ M_t(\hat{f}_t)
\end{aligned}
$$

with $M_t(f)$ as defined in (4.15) and where we take $f_0 = f^*_{\mathcal{F}}$ in the definition of M_t. Since $\hat{f}_t$ is the empirical ϕ-risk minimizer, line 4.E.2 is non-positive, and thus

$$
R^\phi(\hat{f}_t) - R^\phi(f^*_{\mathcal{F}}) \leq M_t(\hat{f}_t). \tag{4.17}
$$

Observe that, for all $f \in \mathcal{F}$,

$$
|\ell_\tau^\phi(f) - \ell_\tau^\phi(f^*_{\mathcal{F}})| \leq \frac{B}{\delta}
$$

and

$$
\begin{aligned}
&E_P\left[\left((\ell_\tau^\phi(f)(O_\tau) - \ell_\tau^\phi(f^*_{\mathcal{F}})(O_\tau)\right)^2 \middle| F_{\tau-1}\right] \\
=&E_P\left[\frac{\left(\phi(f(A_\tau, W_\tau) - \phi(f^*_{\mathcal{F}}(A_\tau, W_\tau))\right)^2}{g_\tau(A_\tau, W_\tau)^2}(1-Y_\tau)^2 \big| F_{\tau-1}\right] \\
\leq&E\left[\sum_{a\in[K]} \frac{\left(\phi(f(a, W_\tau) - \phi(f^*_{\mathcal{F}}(a, W_\tau))\right)^2}{g_\tau(a, W_\tau)} \big| F_{\tau-1}\right] \\
\leq&\frac{KB^2}{\delta}.
\end{aligned}
$$

Therefore, using (4.17) and theorem 4.5, we have that

$$
P\left[R^\phi(\hat{f}_t) - R^\phi(f^*_{\mathcal{F}}) \geq H_t\left(\alpha, \delta, B\sqrt{\frac{K}{\delta}}, B\right) + 160B\sqrt{\frac{Kx}{\delta t}} + 3\frac{Bx}{\delta t}\right] \leq 2e^{-x},
$$

with

$$
H_t(\alpha, \delta, v, B) = \alpha + 160\sqrt{\frac{v}{t}}\int_{\alpha/2}^{B} \sqrt{\log(1 + N_{[]}(\epsilon, \mathcal{F}, L_2(P)))}d\epsilon + 3\frac{B}{\delta t}\log 2.
$$

□

Proof of theorem 4. For any $p \in (0,2) \cup (2, \infty)$,

$$
\int_{\alpha/2}^{B} \sqrt{\log(1 + N_{[]}(\epsilon, \mathcal{F}, L_2(P))} \leq \frac{\sqrt{c_0}}{1-p/2}\left(B^{1-p/2} - \left(\frac{\alpha}{2}\right)^{1-p/2}\right).
$$

We set

$$
\alpha = \begin{cases} 0 & \text{for } p \in (0,2) \\ B^{2/p}\left(\frac{K}{\delta\tau}\right)^{\frac{1}{p}} & \text{for } p > 2. \end{cases}
$$

Then, we have

$$H_\tau(\alpha, \delta, v, B) \leq \begin{cases} B\sqrt{\frac{K}{\delta\tau}} \frac{\sqrt{c_0}}{1-p/2} B^{1-p/2} + \frac{3B\log 2}{\delta\tau} & \text{for } p \in (0,2), \\ B^{2/p}\left(\frac{K}{\delta\tau}\right)^{1/p}\left(1 + \frac{\sqrt{c_0} 2^{1/p} 2^{p/2-1}}{1-p/2}\right) + \frac{3B}{\delta\tau}\log 2 & \text{for } p > 2. \end{cases}$$

Therefore, for

$$\begin{aligned} & x_\tau(\epsilon, K, \delta, B, p) \\ \doteq & \begin{cases} B\sqrt{\frac{K}{\delta\tau}}\left(\frac{\sqrt{c_0}}{1-p/2} B^{1-p/2} + 160\sqrt{\log(2/\epsilon)}\right) + \frac{3B}{\delta\tau}\log(4/\epsilon) & \text{if } p \in (0,2) \\ B^{2/p}\left(\frac{K}{\delta\tau}\right)^{1/p}\left(1 + \frac{\sqrt{c_0} 2^{p/2-1}}{1-p/2}\right) + B\sqrt{\frac{K}{\delta\tau}\log(2/\epsilon)} + \frac{3B}{\delta\tau}\log(4/\epsilon) & \text{if } p > 2. \end{cases} \end{aligned}$$

Theorem 4.3 gives that

$$P\left[R^\phi(\hat{f}_t) - R^\phi(f^*_{\mathcal{F}}) \geq x_\tau(\epsilon, K, \delta, \delta, B, p)\right] \leq \epsilon.$$

Observe that

$$\begin{aligned} \sum_{\tau=1}^t \mathcal{V}(\pi^*_\Pi) - Y_\tau = & \sum_{\tau=1}^t \mathcal{V}(\pi^*_\Pi) - E_P\left[Y_\tau | F_{\tau-1}\right] + \sum_{\tau=1}^t E_P[Y_\tau | F_{\tau-1}] \\ \leq & \sum_{\tau=1}^t \delta_\tau(R(g_{ref}) - R(\pi^*_\Pi)) + (1-\delta_\tau) R(\tilde{\pi}(\hat{f}_{\tau-1}) - R(\pi^*_\Pi) \\ & + \sum_{\tau=1}^t E_P[Y_\tau | F_{\tau-1}] - Y_\tau \\ \leq & \sum_{\tau=1}^t \delta_\tau \\ & + \sum_{\tau=1}^t \left(R^\phi(\hat{f}_{\tau-1}) - R^\phi(f^*_{\mathcal{F}})\right) \\ & + \sum_{\tau=1}^t E_P[Y_\tau | F_{\tau-1}]. \end{aligned}$$

By a union bound, with probability at least $1 - \epsilon/2$,

$$\sum_{\tau=1}^t R^\phi(\hat{f}_{\tau-1}) - R^\phi(f^*_{\mathcal{F}}) \leq \sum_{\tau=1}^t x_\tau\left(\frac{\epsilon}{\tau(\tau+1)}, K, \delta, B, p\right).$$

By Azuma-Hoeffding, with probability at least $1 - \epsilon/2$,

$$\sum_{\tau=1}^t E_P[Y_\tau | F_{\tau-1}] - Y_\tau \leq \sqrt{2\log(2/\epsilon)}.$$

Therefore, with probability at least $1-\epsilon$,

$$\begin{aligned}\sum_{\tau=1}^{t} \mathcal{V}(\pi_{\Pi}^{*}) - Y_\tau \leq & \sum_{\tau=1}^{t} \delta_\tau + x_\tau \left(\frac{\epsilon}{2\tau(\tau+1)}, K, \delta_\tau, B, p \right) \\ \lesssim & t^{\frac{2}{3} \vee \frac{p}{p+1}} \sqrt{\log(t/\epsilon)}.\end{aligned}$$

□

4.F Results on efficient algorithm for policy search in GPE

4.F.1 Casting exploration policy search as a convex feasibility problem

For any $M > 0$, denote $\mathcal{P}_t(M)$ the following feasibility problem.

$$\text{Find } \tilde{g}_t \in \mathcal{F}_t \text{ such that } \frac{1}{t-1} \sum_{\substack{a \in [K] \\ \tau \in [t-1]}} \frac{f(a, W_\tau)}{\delta_t / K + (1-\delta_t)\tilde{g}_t(a|W_\tau)} \leq M.$$

For all $f \in \mathcal{F}_t$, let

$$w_{t,f} \doteq (f(a, W_\tau) : a \in [K], \tau \in [t]).$$

For any given $f \in \mathcal{F}$, observe that

$$\begin{aligned} f \in \mathcal{F}_t \iff & \forall \tau \in [t-1], \hat{R}_\tau(f) \leq \min_{f \in \mathcal{F}_\tau} \hat{R}_\tau(f) + \epsilon_\tau \doteq b_\tau \\ \iff & \forall \tau \in [t-1], u_{t-1,\tau}^\top w_{t-1,f} \leq b_\tau, \end{aligned} \tag{4.18}$$

where

$$u_{t,\tau} \doteq \left(\mathbf{1}\{s \leq \tau\} \frac{\mathbf{1}\{A_s = a\}(1 - Y_s)}{g_\tau(a|W_s)} : a \in [K], s \in [t] \right).$$

Introduce the set

$$\mathcal{C}_t \doteq \{ w_{f,t} : f \in \mathcal{F}_t \},$$

which, by (4.18) can be rewritten as

$$\mathcal{C}_t \doteq \{ w_{f,t} : f \in \mathcal{F}_t, \forall \tau \in [t], u_{t,f} w_{f,t} \leq b_\tau \}. \tag{4.19}$$

Based on (4.18) and (4.19), we can thus rewrite $\mathcal{P}_t(M)$ as the following two-step problem.

1. Find $w \in \mathcal{C}_t$ such that $\forall z \in \mathcal{C}_t, \dfrac{1}{t-1} \displaystyle\sum_{a \in [K], \tau \in [t-1]} \dfrac{z_{a,\tau}}{\delta_t/K + (1-\delta_t) w_{a,\tau}} \leq M.$

2. Find $f \in \mathcal{F}_t$ such that $w_{f,t} = w$.

As $\mathcal{F}$ is convex, that functions in f have range in $[0,1]$, and that for all $z \in \mathbb{R}^{Kt}$,

$$w \mapsto \frac{1}{t-1} \sum_{\substack{a\in[K]\\ \tau\in[t]}} \frac{z_{a,\tau}}{\delta_t/K + (1-\delta_t)w_{a,\tau}}$$

is a convex mapping, the set

$$\mathcal{D}_t(M) \doteq \mathcal{C}_t \cap \left\{ w \in \mathbb{R}^{Kt} : \forall z \in \mathcal{C}_t, \frac{1}{t-1} \sum_{\substack{a\in[K]\\ \tau\in[t-1]}} \frac{z_{a,\tau}}{\delta_t/K + (1-\delta_t)w_{a,\tau}} \le M \right\}$$

is a convex set. The following lemma ensures it is not empty.

Lemma 4.10. *Let $\mathcal{C}$ be a compact convex subset of $\mathbb{R}^{K(t-1)}$. Set arbitrary $\delta \in (0,1)$ and $w \in \mathcal{C}$. Then*

$$\max_{z\in\mathcal{C}} \frac{1}{t-1} \sum_{\substack{a\in[K]\\ \tau\in[t-1]}} \frac{z_{a,\tau}}{\delta/K + (1-\delta)w_{a,\tau}} \le \frac{4}{3}K.$$

As we will recall precisely in the next subsection, so as to be able to give gaurantees on the number of iterations needed by the ellipsoid algorithm to find a point in a convex set, we need a lower bound on the volume of the set. As we can make the volume of $\mathcal{D}_t$ arbitrarily small in some cases, similarly to [Dudik et al., 2011], we will consider a slightly enlarged version of $\mathcal{D}_t$ whose volume we can explicitly lower bound. The following lemma informs how to construct such an enlarged set. Before stating the lemma, we introduce the following notation:

$$h_{t,\delta} \doteq \frac{1}{t} \sum_{\substack{a\in[K]\\ \tau\in[t]}} \frac{z_{a,\tau}}{\delta/K + (1-\delta)w_{a,\tau}}$$

Lemma 4.11. *Let $w \in (\mathbb{R}_+)^{Kt}$, $\delta \in (0,1)$, $\Delta \in (0, \delta/2)$. Then, for all $u \in B_{Kt}(0,1)$, $z \in [0,1]^{Kt}$,*

$$|h_{\delta,t}(w + \Delta u, z) - h_{\delta,t}(w,z)| \le \xi_{t,\delta}(\Delta),$$

with $\xi_{t,\delta}(\Delta) \doteq 2\Delta\delta^{-2}\sqrt{K/t}$.

For all $\Delta > 0$, let

$$C_{t,\Delta} = \left\{ w \in \mathbb{R}^{Kt} : d(w, \mathcal{C}_t \le \Delta \right\}.$$

From the above lemma, if $w \in \mathcal{D}_t(M)$, every point $w' \in B(w,\Delta)$ satisfies

$$\max_{z\in\mathcal{C}_t} h_{t,\delta}(w', z) \le M + \xi_{t,\delta}(\Delta).$$

Therefore, provided $\mathcal{D}_t$ contains at least one point, say w, the set

$$\mathcal{D}_{t,\Delta} \doteq \{w \in \mathcal{C}t, \Delta : \forall z \in \mathcal{C}_t, h_{t,\delta}(w,z) \leq M + \xi_{t,\delta}(\Delta)\}$$

contains $B(w,\Delta)$. Finally, suppose that $w \in \mathcal{D}_{t,\Delta}(M)$. Then, by definition of $\mathcal{D}_{t,\Delta}(M)$, there exists a $w' \in \mathcal{C}_t$ such that $d(w',w') \leq \Delta$, and thus by lemma 4.11,

$$\max_{z \in \mathcal{C}_t} h_{t,\delta}(w,z) \leq M + 2\xi_{t,\delta}(\Delta).$$

By lemma 4.10, we can pick $M = 4K/3$ while still ensuring that $\mathcal{D}_t(M)$ is non-empty. Them setting Δ such that $\xi_{t,\delta_t}(\Delta) = K/3$, that is setting it to $\Delta_t \doteq \delta^2\sqrt{(t-1)/K}$ ensures that $\mathcal{D}_{t,\Delta_t}$ contains a ball od radius Δ_t and that $M + 2\xi_{t,\delta_t}(\Delta_t) \leq 2K$. Therefore, the exploration policy search problem (4.3) is equivalent to the two-step process

1. Find $w \in \mathcal{D}_{t,\Delta_t}$
2. Find $f \in \mathcal{F}$ such that $\|w_{f,t} - w\|_2 \leq \Delta_t$.

4.F.2 Finding an element of $\mathcal{U}$ using the ellipsoid algorithm

Finding an element of a convex set of non-negligilble volume such as $\mathcal{D}_{t,\Delta_t}(4K/3)$ can be performed in polynomial time with the ellipsoid algorithm. The ellipsoid algorithm requires having access to a separation oracle.

Definition 4.5 (Separation oracle)**.** *Let $\mathcal{C} \subseteq \mathbb{R}^n$, $n \geq 1$ be a convex set. A separation oracle for $\mathcal{C}$ is a routine that, for any $w \in \mathbb{R}^n$ outputs whether $w \in \mathcal{C}$, and if $w \neq \mathcal{C}$, returns an hyperplane separating w and $\mathcal{C}$.*

We will not recall here the ellipsoid algorithm as it is standard, but we restate a know lemma on its runtime.

Lemma 4.12 (Runtime of the ellipsoid algorithm)**.** *Let $\mathcal{C}$ be a convex set. Suppose we know an $R > 0$ such that $\mathcal{C} \subseteq B_n(0,R)$, and that there exists a point $w \in \mathcal{C}$ and $\Delta > 0$ such that $B(w,\Delta) \subseteq \mathcal{C}$. Then the ellipsoid algorithm finds a point in $\mathcal{C}$ in no more than*

$$O\left(n^2 \log\left(\frac{R}{\Delta}\right)\right)$$

calls to a separation oracle for $\mathcal{C}$.

Therefore, to construct an efficient algorithm that finds the exploration policy at time t, we just need to find how to implement a separation oracle for $\mathcal{D}_{t,\Delta_t}$. Observe that we can rewrite $\mathcal{D}_{t,\Delta_t}$ as the intersection of two convex sets:

$$\mathcal{D}_{t,\Delta_t} \doteq \mathcal{C}_{t,\Delta_t} \cap \left\{w \in \mathbb{R}^{Kt} :\ \forall z \in \mathcal{C}_t h_{t,\delta_t}(w,z) \leq \frac{5}{3}K\right\}.$$

A separation oracle for $\mathcal{D}_{t,\Delta_t}$ can thus be built from a separation oracle for $\mathcal{C}_{t,\Delta}$ and a separation oracle for $\{w \in \mathbb{R}^{Kt} :\ \forall z \in \mathcal{C}_t h_{t,\delta_t}(w,z) \leq 5K/3\}$.

The following lemma shows how to implement a separation oracle for $\mathcal{C}_{t,\Delta}$ using one call to LCLSO.

Lemma 4.13 (Separation oracle for $\mathcal{C}_t$)**.** *Let $w \in \mathcal{C}^{Kt}$. Let*

$$\tilde{w} \doteq \arg\min_{w' \in \mathcal{C}_t} \|w - w'\|.$$

If $\|w - \tilde{w}\| \leq \Delta$, then $w \in \mathcal{C}_{t,\Delta}$. If not, then

$$\mathcal{H} \doteq \left\{ z \in \mathbb{R}^{Kt} : \langle z - w, w - \tilde{w} \rangle = 0 \right\}$$

is an hyperplane that separates w from $\mathcal{C}_{t,\Delta}$.

Proof. It suffices to show that $\forall z \in \mathcal{H}$, $d(z, \mathcal{C}_{t,\Delta}) > 0$, or equivalently that $d(z, \mathcal{C}_t) > \Delta$. Observe that since $w \in \mathcal{C}_{t,\Delta}$, we must have that $d(w, \mathcal{C}_t) > \Delta$. Therefore, it will be enough to show that

$$\forall z \in \mathcal{H},\ d(z, \mathcal{C}_t) \geq d(w, \mathcal{C}_t).$$

We first show that for all $\tilde{z} \in \mathcal{C}_t$, $\langle \tilde{z} - \tilde{w}, w - \tilde{w} \rangle > 0$. Then, for all $\lambda \in (0, 1)$,

$$\begin{aligned} \|w - (\lambda\tilde{z} + (1-\lambda)\tilde{w})\|_2^2 =& \|(w - \tilde{w}) - \lambda(\tilde{z} - \tilde{w})\|_2^2 \\ =& \|w - \tilde{w}\|_2^2 + \lambda^2 \|z - \tilde{w}\|_2^2 - 2\lambda \langle \tilde{z} - \tilde{w}, w - \tilde{w} \rangle. \end{aligned}$$

Therefore, for $\lambda \in (0, 1)$ small enough,

$$\|w - (\lambda\tilde{z} + (1 - \lambda\tilde{w})\|_2^2 \leq \|w - \tilde{w}\|_2^2.$$

Since, by convexity of $\mathcal{C}_t$, $\lambda\tilde{z} + (1-\lambda)\tilde{w} \in \mathcal{C}_t$, this contradicts that $\tilde{w}$ is the projection of w on $\mathcal{C}_t$. Therefore, we must have that

$$\langle \tilde{z} - \tilde{w}, w - \tilde{w} \rangle \leq 0 \tag{4.20}$$

for all $\tilde{z} \in \mathcal{C}_t$.

We can now use this property to show the wished claim. Let $z \in \mathcal{H}$, and let $\tilde{z} \in \mathcal{C}_t$. We have that

$$\begin{aligned} \|z - \tilde{z}\|_2^2 =& \|(z - w) + (w - \tilde{w}) + (\tilde{w} - \tilde{z})\|_2^2 \\ =& \|z - w\|_2^2 + \|w - \tilde{w}\|_2^2 + \|\tilde{w} - \tilde{z}\|_2^2 \\ &+ 2 \underbrace{\langle z - w, w - \tilde{w} \rangle}_{=0 \text{ by definition of } \mathcal{H}} \\ &+ 2 \underbrace{\langle w - \tilde{w}, \tilde{w} - \tilde{z} \rangle}_{\geq 0 \text{ from (4.20)}} \\ &+ 2 \underbrace{\langle z - w, \tilde{w} - \tilde{z} \rangle}_{\substack{\geq -\|z - w\| \|\tilde{w} - \tilde{z}\| \\ \text{by Cauchy-Schwartz}}} \\ \geq& (\|z - w\| - \|w - \tilde{w}\|)^2 + \|\tilde{w} - \tilde{z}\|_2^2 \\ \geq& d(w, \mathcal{C}_t), \end{aligned}$$

which concludes the proof. □

The next lemma shows how to implement a separation oracle for

$$\mathcal{L}_t \doteq \left\{ w \in \mathbb{R}^{Kt} : \forall z \in \mathcal{C}_t, h_{t,\delta_t}(w,z) \leq \frac{5}{3}K \right\}.$$

using one call to LCCSCO.

Lemma 4.14 (Separation oracle for $\mathcal{L}_t$)**.** *Let $w \in \mathbb{R}^{Kt}$. Let*

$$z^* \doteq \arg\max_{z \in \mathcal{C}_t} h_{t,\delta_t}(w,z).$$

z^* can be found in one call to LCCSCO. If $h_{t,\delta_t}(w,z^*) \leq 5K/3$, then $w \in \mathcal{L}_t$. If not, then $w \notin \mathcal{L}_t$ and

$$\mathcal{H} \doteq \left\{ w' : h_{t,\delta_t}(w,z^*) + (\nabla_w h_t)(w,z^*)^\top (w'-w) = 0 \right\}$$

separates w and $\mathcal{L}_t$.

We restate below for self-containdness lemma 10 from Dudik et al. [2011], which will be useful in the rest of the section.

Lemma 4.15 (Lemma 10 in [Dudik et al., 2011])**.** *For $x \in \mathbb{R}^n$, let $f(x)$ be a convex function of x, and consider the convex set K defined by $K = \{x : f(x) \leq 0\}$. Suppose we have a point y such that $f(y) > 0$. Let $\nabla f(y)$ be a subgradient of f at y. Then the hyperplane $f(y) + \nabla f(y)^\top (x-y) = 0$ separates y from K.*

Proof. Observe that

$$h_{t,\delta_t}(w,z) \doteq \frac{1}{t-1} \sum_{\substack{a \in [K] \\ \tau \in [t-1]}} u_{a,\tau} z_{a,\tau},$$

with

$$u_{a,\tau} \doteq \frac{1}{\delta_t/K + (1-\delta_t) w_{a,\tau}} \geq 0.$$

Therefore, $\arg\max_{z \in \mathcal{C}_t} h_{t,\delta}(w,z) = w_{t,f^*}$, where

$$f^* \doteq \arg\max_{f \in \mathcal{F}} \frac{1}{t-1} \sum_{\substack{a \in [K] \\ \tau \in [t-1]}} u_{a,\tau} f(a, W_\tau) \text{ subject to } \forall \tau \in [t], \hat{R}_\tau(f) \leq \max_{f \in \mathcal{F}} \hat{R}_\tau(f) + \epsilon_\tau.$$

As

$$\hat{R}_\tau(f) = \frac{1}{\tau} \sum_{\substack{a \in [K] \\ s \in [\tau]}} \frac{\mathbf{1}\{A_s = a\}(1-Y_s)}{g(a|W_s)} f(a, W_s),$$

the constraint $\hat{R}_\tau(f) \leq \max_{f \in \mathcal{F}} \hat{R}_\tau(f) + \epsilon_\tau$. is a linear constraint, and therefore, f^* can be obtained with one call to LCCSCO.

From lemma 4.15, if $h_{t,\delta_t}(w, z^*) - 5K/3 > 0$,

$$\mathcal{H} \doteq \left\{ w' : h_{t,\delta_t}(w, z^*) + (\nabla_w h_t)(w, z^*)^\top (w' - w) = 0 \right\}$$

separates w from

$$\left\{ w' \in \mathbb{R}^{Kt} : h_{t,\delta_t}(w', z^*) - \frac{5}{3} K \leq 0 \right\},$$

and thus from $\mathcal{L}_t$, which concludes the proof. □

4.G Proof of the results on the additive model policy class

4.G.1 Proof of lemma 4.2

The following result is the fundamental building block of the proof.

Lemma 4.16 (Bracketing entropy of univariate distribution functions)**.** *Let $\mathcal{G}$ the set of cumulative distribution functions on $[0, 1]$. There exist $c_0 > 0$, $\epsilon_0 \in (0, 1)$ such that, for all $\epsilon \in (0, \epsilon_0)$,*

$$\log N_{[]}(\epsilon, \mathcal{G}, \| \cdot \|_\infty) \leq c_0 \epsilon^{-1} \log(1/\epsilon).$$

We first state an intermediate result.

Lemma 4.17 (Bracketing entropy of linear combinations)**.** *Let $\mathcal{H}$ be a class of functions and let*

$$\mathcal{F} \doteq \left\{ \sum_{j=1}^{J} a_j h_j : a_1, \ldots, a_J \in [-B, B],\ h_1, \ldots, h_J \in \mathcal{H} \right\}.$$

Suppose that for all $h \in \mathcal{H}$, $\|h\|_\infty \leq M$. Then, for all $\epsilon > 0$,

$$\log N_{[]}(\epsilon, \mathcal{F}, \| \cdot \|_\infty) \leq J \log N_{[]} \left(\frac{\epsilon}{2JB}, \mathcal{H}, \| \cdot \|_\infty \right) + J \log \left(\frac{4JBM}{\epsilon} \right).$$

Proof of lemma 4.17. Let

$$\mathcal{B} = \{ (l_k, u_k) : k \in [N] \}$$

be an ϵ bracketing in $\| \cdot \|_\infty$ norm of $\mathcal{H}$. For all m, let $\alpha_m = m(B/M)\epsilon$. For all $f = \sum_{j=1}^J a_j h_j$, there exist $k_1, \ldots, k_J$ and $m_1, \ldots, m_J$ such that $\forall j \in [J]$,

$$l_{k_j} \leq h_j \leq u_{k_j},$$

and

$$\alpha_{m_{j-1}} \leq a_j \leq \alpha_{m_j}.$$

Therefore,

$$\Lambda(k_1, \dots, k_J, m_1, \dots, m_J) \leq f \leq \Upsilon(k_1, \dots, k_J, m_1, \dots, m_J),$$

with

$$\Lambda(k_1, \dots, k_J, m_1, \dots, m_J) \doteq \sum_{j=1}^{J} \alpha_{m_j-1}(l_{i_j})^+ + \alpha_{m_j}(l_{i_j})^-,$$

$$\text{and } \Upsilon(k_1, \dots, k_J, m_1, \dots, m_J) \doteq \sum_{j=1}^{J} \alpha_{m_j-1}(u_{i_j})^+ + \alpha_{m_j}(u_{i_j})^-.$$

Therefore, we have that

$$\begin{aligned}
&|\Upsilon(k_1, \dots, k_J, m_1, \dots, m_J) - \Lambda(k_1, \dots, k_J, m_1, \dots, m_J)| \\
=&\left| \sum_{j=1}^{J} \alpha_{m_j-1}(u_{i_j} - l_{i_j}) + \sum_{j=1}^{J} (\alpha_{m_j} - \alpha_{m_j-1})((u_{m_j})^+ - (l_{m_j})^-) \right| \\
\leq& JB\epsilon + \frac{JB\epsilon}{M} M \\
=& 2JB\epsilon.
\end{aligned}$$

Therefore,

$$N_{[]}(2JB\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq N_{[]}(\epsilon, \mathcal{H}, \|\cdot\|_\infty) \times \left(\frac{2M}{\epsilon}\right)^J,$$

hence the claim. □

We can now prove lemma 4.2

Proof of lemma 4.2. Let $\epsilon > 0$. Let

$$\mathcal{B} \doteq \{(\ell_i, u_i) : i \in [N]\},$$

be an ϵ-bracketing in $\|\cdot\|_\infty$ the set of distribution functions on $[0, 1]$, which we will denote $\mathcal{G}$. Let $h \in \mathcal{H}$. There exist $a \in [-B, B]$, $b \in [0, B]$, $h_1, h_2 \in$ such that $h = a + b(h_1 - h_2)$, and there exists $i_1, i_2 \in [N]$ such that

$$l_{i_1} \leq h_1 \leq u_{i_1} \qquad \text{and} \qquad l_{i_2} \leq h_2 \leq u_{i_2},$$

and $i_3 \in [-1/\epsilon, 1/\epsilon]$ such that $a \in [\alpha_{i_3-1}, \alpha_{i_3}]$ with

$$\alpha_{i_3} \doteq i_3 M\epsilon,$$

and $i_4 \in [0, 1/\epsilon]$ such that $b \in [\beta_{i_4 - 1}, \beta_{i_4}]$ with

$$\beta_{i_4} \doteq i_4 M \epsilon.$$

Therefore, we have that

$$\Lambda(i_1, i_2, i_3, i_4) \le h \le \Upsilon(i_1, i_2, i_3, i_4),$$

with

$$\begin{aligned} \Lambda(i_1, i_2, i_3, i_4) \doteq & \alpha_{i_3 - 1} + \beta_{i_3 - 1}(l_{i_1} - u_{i_2})^+ + \beta_{i_3}(l_{i_1} - u_{i_2})^- \\ \text{and } \Upsilon(i_1, i_2, i_3, i_4) \doteq & \alpha_{i_3} + \beta_{i_3}(u_{i_1} - l_{i_2})^+ + \beta_{i_3 - 1}(u_{i_1} - l_{i_2})^-. \end{aligned}$$

Note that

$$\begin{aligned} 0 \le \Upsilon(i_1, i_2, i_3, i_4) - \Lambda(i_1, i_2, i_3, i_4) = & \alpha_{i_3} - \alpha_{i_3 - 1} \\ & + \beta_{i_3 - 1}(u_{i_1} - l_{i_1} + u_{i_2} - l_{i_2}) \\ & + (\beta_{i_3} - \beta_{i_3 - 1})((u_{i_1} - l_{i_2})^+ - (l_{i_1} - u_{i_2})^-) \\ \le & M\epsilon + 2M\epsilon + M_\epsilon(u_{i_1} - l_{i_2})^+ \\ \le & 4M\epsilon. \end{aligned}$$

Therefore,

$$\mathcal{B}' \doteq \{(\Lambda(i_1, i_2, i_3, i_4), \Upsilon(i_1, i_2, i_3, i_4)) : i_1, i_2 \in [N], i_3 \in [-1/\epsilon, 1/\epsilon], i_4 \in [0, 1/\epsilon]\}$$

is an $4M\epsilon$-bracket in $\| \cdot \|_\infty$ norm of $\mathcal{H}$. Thus

$$N_{[]}(4M\epsilon, \mathcal{H}, \| \cdot \|_\infty) \le \frac{2}{\epsilon} \times \frac{1}{\epsilon} \times N_{[]}(\epsilon, \mathcal{G}, \| \cdot \|_\infty)^2$$

That is

$$\begin{aligned} \log N_{[]}(\epsilon, \mathcal{H}, \| \cdot \|_\infty) \le & 2 \log \left(\frac{8M^2}{\epsilon^2} \right) + 2 \log N_{[]} \left(\frac{\epsilon}{4M}, \mathcal{G}, \| \cdot \|_\infty \right) \\ = & 2 \log \left(\frac{\sqrt{8}M}{\epsilon} \right) + 2 \log N_{[]}(\epsilon, \mathcal{G}, \| \cdot \|_\infty). \end{aligned}$$

Therefore, from lemma 4.17 and lemma 4.16,

$$\begin{aligned} \log N_{[]}(\epsilon, \mathcal{F}, \| \cdot \|_\infty) \le & J \log \left(\frac{4JBM}{\epsilon} \right) + 2J \log \left(\frac{\sqrt{8}MJBM}{\epsilon} \right) \\ & + 2J \log N_{[]} \left(\frac{\epsilon}{2JB}, \mathcal{G}, \| \cdot \|_\infty \right) \\ \le & (2Jc_0 + 1)\epsilon^{-1} \log \left(\frac{4\sqrt{8}J(M \vee M^2)(B \vee 1)}{\epsilon} \right), \end{aligned}$$

for all $\epsilon \in (0, 2JB\epsilon_0)$. □

4.G.2 Proof of lemma 4.3

Proof of lemma 4.3. We decompose the proof in three steps. We will denote feas$(\mathcal{P}_1)$ and feas$(\mathcal{P}_2)$ the feasible sets of $\mathcal{P}_1$ and $\mathcal{P}_2$.

Step 1: The feasible set of $\mathcal{P}_2$ is contained in the feasible set of $\mathcal{P}_1$. First, observe that for any $h : x \mapsto \sum_{\tau=1}^{t} \beta_\tau \mathbf{1}\{x \geq x_\tau\}$, $\|h\|_v = \sum_{\tau=0}^{t} |\beta_\tau|$. Therefore, for every l, $\tilde{\mathcal{H}}_{l,t} \subseteq \mathcal{H}$ and thus $\tilde{F}_t \subseteq \mathcal{F}$.

Second, observe that for any $w \in [0,1]^d$, there exists $\tilde{w} \in \mathcal{G}(w_0, \ldots, w_n)$ such that $\tilde{f}(w) = \tilde{f}(\tilde{w})$. Therefore, if $\tilde{f}$ satisfies (4.9) and (4.10) at every $(a, w) \in [K] \times \mathcal{G}(w_0, \ldots, w_t)$, it satisfies them everywhere. Therefore, this proves that the feasible set of $\mathcal{P}_2$ is contained in the feasible set of $\mathcal{P}_1$.

Step 2: For any f in the feasible set of $\mathcal{P}_1$, there is an $\tilde{f}$ in the feasible set of $\mathcal{P}_2$ that achieves the same value of the objective function. Let $f : (a, w) \mapsto \sum_{l=1}^{d} \alpha_{a,l} h_{a,l}(w_l)$ be an element of the feasible set of $\mathcal{P}_1$. Observe that for all a, l, there exists $\tilde{h}_{a,l}$ of the form $\tilde{h}_{a,l} : x \mapsto \sum_{\tau=1}^{t} \beta_{a,l,\tau} \mathbf{1}\{x \geq w_{\tau,l}\}$ such that for all $\tau \in \{0, \ldots, t\}$, $\tilde{h}_{a,l}(w_{\tau,l}) = h_{a,l}(w_{\tau,l})$. As f and $\tilde{f}$ coincide at every $(a, w) \in [K] \times \mathcal{G}(w_0, \ldots, w_n)$, constraints (4.11) and (4.12) are satisfied at every $(a, w) \in [K] \times \mathcal{G}(w_0, \ldots, w_n)$, and f and $\tilde{f}$ achieve the same value of the objective function. To prove that $\tilde{f}$ is in the feasible set of $\mathcal{P}_2$, it remains to show that the functions $(\tilde{h}_{a,l})_{a \in [K], l \in [d]}$ are in $\mathcal{H}_{l,t}$, that is that for all a, l, $\sum_{\tau=0}^{t} |\beta_{a,l,\tau}| \leq M$. We have that

$$\begin{aligned}
\sum_{\tau=0}^{t} |\beta_{a,l,\tau}| =& |\tilde{h}_{a,l}(0)| + \sum_{\tau=1}^{t} |\tilde{h}_{a,l}(w_{\tau,l}) - \tilde{h}_{a,l}(w_{\tau,l})| \\
=& |h_{a,l}(0)| + \sum_{\tau=1}^{t} |h_{a,l}(w_{\tau,l}) - h_{a,l}(w_{\tau-1,l})| \\
\leq& |h_{a,l}(0)| + \sum_{\substack{m \in \mathbb{N} \\ 0 \leq x_1 \leq \ldots \leq x_m \leq 1}} |h_{a,l}(x_{m+1}) - h_{a,l}(x_m)| \\
\leq& \|h_{a,l}\|_v \\
\leq& M.
\end{aligned}$$

Step 3: End of the proof. Let f^* be a solution to $\mathcal{P}_1$. Let $\tilde{f}^*$ be a function in the feasible set of $\mathcal{P}_2$ such that $f^* = \tilde{f}^*$ on $[K] \times \mathcal{G}(w_0, \ldots, w_t)$. From step 2, such a function exists. The objective function evaluated at $\tilde{f}^*$ is equal to the objective function evaluated at f^*. Since, from step 1, feas$(\mathcal{P}_1) \subseteq$ feas$(\mathcal{P}_2)$, and f^* is a maximizer over feas$(\mathcal{P}_1)$, $\tilde{f}^*$ must be a maximizer over both $\mathcal{P}_1$ and $\mathcal{P}_2$. □

4.H Representation results for the ERM over cadlag functions with bounded sectional variation norm

4.H.1 Empirical risk minimization in the hinge case

The following result shows that empirical risk minimization over $\mathcal{F}^{hinge}$, with $\mathcal{F}_0$ the class of cadlag functions with bounded sectional variation norm.

Lemma 4.18 (Representation of the ERM in the hinge case)**.** *Consider a class of policies of the form $\mathcal{F}^{hinge}$, as defined in* (4.6)*, derived from $\mathcal{F}_0$, as defined in* (4.13)*. Let $\phi = \phi^{hinge}$. Suppose we have observed $(W_1, A_1, Y_t), \ldots, W_t, A_t, Y_t)$ and let $\tilde{W}_1, \ldots, \tilde{W}_m$ be the elements of $G(W_1, \ldots, W_t)$.*

Let $(\beta_j^a)_{a \in [K], j \in [m]}$ be a solution to

$$\begin{aligned} \min_{\beta \in \mathbb{R}^{Km}} \sum_{\tau=1}^{t} \sum_{a \in [K]} & \left\{ \frac{\mathbf{1}\{A_\tau = a\}}{g_\tau(A_\tau, W_\tau)} (1 - Y_\tau) \right. \\ & \left. \times \max \left(0, 1 + \sum_{j=1}^{m} \beta_j^a \mathbf{1}\{W_\tau \geq \tilde{W}_j\} \right) \right\} \\ \textit{s.t. } \forall l \in [m], & \sum_{a \in [K]} \sum_{j=1}^{m} \beta_j^a \mathbf{1}\{\tilde{W}_l \geq \tilde{W}_j\} = 0, \\ \forall a \in [K], & \sum_{j=1}^{m} |\beta_j^a| \leq M. \end{aligned}$$

Then $f : (a, w) \mapsto \sum_{j=1}^{m} \beta_j^a \mathbf{1}\{w \geq \tilde{W}_j\}$ is a solution to $\min_{f \in \mathcal{F}^{Id}} \sum_{\tau=1}^{t} \ell_\tau^\phi(f)(O_\tau)$.

4.H.2 Formal definition of the Vitali variation and the sectional variation norm

We now present in full generality the definitions of the notions Vitali variation, Hardy-Krause variation and sectional variation norm. This requires introducing some prelimiary definitions. This section is heavily inspired from the excellent presentation of Fang et al. [2019], and we write it instead of directly referring to their work mostly for self-containdness, and so as to ensure matching notation.

Definition 4.6 (Rectangular split, rectangular partition and rectangular grid)**.** *For any d subdivisions*

$$0 = w_{k,1} \leq w_{k,2} \leq \ldots \leq w_{k,q_k} = 1, \; k = 1, \ldots, d,$$

of $[0, 1]$, let

- *$\mathcal{P}$ be the collection of all closed rectangles of the form* $[w_{1,i_1}, w_{1,i_1+1}] \times \ldots \times [w_{d,i_d}, w_{d,i_d+1}]$*,*
- *$\mathcal{P}^*$ be the collection of all open rectangles of the form* $[w_{1,i_1}, w_{1,i_1+1}) \times \ldots \times [w_{d,i_d}, w_{d,i_d+1})$*.*
- *$\mathcal{G}$ the collection of all points of the form* $(w_{i_1}, \ldots, w_{i_d})$*.*

Any collection of the form $\mathcal{P}$ is called a rectangular split of $[0,1]^d$, any collection of the form $\mathcal{P}^*$ is called a rectangular partition of $[0,1]^d$ and any set of points of the form $\mathcal{G}$ is called a rectangular grid on $[0,1]^d$.

Definition 4.7 (Minimum rectangular split, partition and grid)**.** *Let $w_1, \ldots, w_n$ be n points of $[0,1]^d$. We call minimum rectangular split induced by $w_1, \ldots, w_n$, and we denote $\mathcal{P}(w_1, \ldots, w_n)$, the rectangular split of minimum cardinality such that $w_1, \ldots, w_n$ are all corners of rectangles in $\mathcal{P}(w_1, \ldots, w_n)$. We define similarly the minimum rectangular parition induced by $w_1, \ldots w_n$. We denote it $\mathcal{P}^*(w_1, \ldots, w_n)$. We define the minimum rectangular grid induced by $w_1, \ldots, w_n$, which we denote $\mathcal{G}(w_1, \ldots, w_n)$, as the smallest cardinality rectangular grid that contains $w_1, \ldots, w_n$.*

Definition 4.8 (Section of a function)**.** *Let $s \in [d]$, $s \neq \emptyset$, and consider $f \in \mathbb{D}([0,1]^d)$. We call the s-section of f, and denote f_s, the restriction of f to the set*

$$\{(w_1, \ldots, w_d) \in [0,1]^d : \forall j \in s, w_j = 0\}.$$

Observe that the above set is a face of the cube $[0,1]^d$ and that f_s is a cadlag function with domain $[0,1]^{|s|}$.

Definition 4.9 (Vitali variation)**.** *For any $d \geq 1$ and any rectangle R of the form $[w_{1,1}, w_{2,1}] \times \ldots \times [w_{1,d}, w_{2,d}]$ or $[w_{1,1}, w_{2,1}) \times \ldots \times [w_{1,d}, w_{2,d})$, such that for all $k = 1, \ldots, d$, $w_{k,1} \leq w_{k,2}$, let*

$$\Delta^{(d)}(f, R) = \sum_{j_1=0}^{J_1} \ldots \sum_{j_d=0}^{J_d} (-1)^{j_1+\ldots+j_d} f(w_{2,1} + j_1(w_{1,1} - w_{2,1}), \ldots, w_{2,d} + j_d(w_{1,d} - w_{2,d})),$$

where, for all $k = 1, \ldots, d$, $J_k = I(w_{2,d} \neq w_{1,d})$. The quantity $\Delta^{(d)}(f, R)$ is called the quasi volume ascribed to R by f. The Vitali variation of f on $[0,1]^d$ is defined as

$$V^{(d)}(f, [0,1]^d) = \sup_{\mathcal{P}} \sum_{R \in \mathcal{P}} |\Delta^{(d)}(f, R)|,$$

where the sup *is over all the rectangular partitions of $[0,1]^d$.*

Definition 4.10 (Hardy-Krause variation and sectional variation norm)**.** *The Hardy-Krause variation anchored at the origin of a function $f \in \mathbb{D}([0,1]^d)$ is defined as the sum of the Vitali variation of its sections, that is it is defined as the quantity*

$$V_{HK,\mathbf{0}}(f) = \sum_{\emptyset \neq s \subseteq [d]} V^{(d)}(f_s, [0,1]^{|s|}).$$

The sectional variation norm of f is defined as follows:

$$\|f\|_v = |f(0)| + V_{HK,\mathbf{0}}(f).$$

4.H.3 Proof of lemmas 4.5 and 4.18

The proof of lemmas 4.5 and 4.18 will easily follow from the following two results.

Lemma 4.19. *Let $f \in \mathcal{F}_0$. Let $x_1, \dots, x_n \in [0,1]^d$. Denote $\tilde{x}_1, \dots, \tilde{x}_m$ the elements of $G(x_1, \dots, x_n)$. Let*

$$\tilde{\mathcal{F}}_0(x_1, \dots, x_n) \doteq \left\{ x \mapsto \sum_{j=1}^{m} \beta_j \mathbf{1}\{x \geq \tilde{x}_j\} : \sum_{j=1}^{m} |\beta_j| \leq M \right\}.$$

Then

- *$\tilde{\mathcal{F}}_0(x_1, \dots, x_n) \subseteq \mathcal{F}$,*
- *there exists $\tilde{f} \in \tilde{\mathcal{F}}_0(x_1, \dots, x_n)$ such that $\tilde{f}$ and f coincide on $G(x_1, \dots, x_m)$ and $\|\tilde{f}\|_v \leq \|f\|_v$.*

Lemma 4.20. *Let $\tilde{f}_1, \dots, \tilde{f}_q \in \tilde{\mathcal{F}}_0(x_1, \dots, x_n)$. Let $\alpha_1, \dots, \alpha_q, \beta \in \mathbb{R}$. Consider the inequality constraint*

$$\sum_{l=1}^{q} \alpha_l \tilde{f}_l \leq \beta.$$

The following are equivalent.

1. *$\tilde{f}_1, \dots, \tilde{f}_q$ satisfy the inequality constraint everywhere on $[0,1]^d$.*
2. *$\tilde{f}_1, \dots, \tilde{f}_q$ satisfy the inequality constraint everywhere at every point of $G(x_1, \dots, x_n)$.*

We relegate the proofs of the two above lemmas further down in this section. We can now state the proof of lemmas 4.5 and 4.18.

Proof of lemmas 4.5 and 4.18. The following arguments apply similarly to lemma 4.5 and lemma 4.18. We present the proof in the direct policy optimization case. We proceed in two steps.

Step 1: the feasible set of (4.14) **contains a solution the ERM problem over** $\mathcal{F}^{Id}$ Let f be a solution to

$$\min_{f \in \mathcal{F}^{Id}} \sum_{\tau=1}^{t} \ell_\tau^{Id}(f)(O_\tau). \tag{4.21}$$

There exists $f_1, \dots, f_K \in \mathcal{F}_0$ such that $\forall a \in [K]$, $f(a, \cdot) = f_a(\cdot)$. From lemma 4.19, there exists $\tilde{f}_1, \dots, \tilde{f}_K$ that coincide with $f_1, \dots, f_K$ on $G(x_1, \dots, x_n)$. Then the function $\tilde{f} : (x, a) \mapsto \tilde{f}_a(x)$ achieves the same value of the objective in (4.21) as f.

Since $\tilde{f}_1, \dots, \tilde{f}_K$ coincide with $f_1, \dots, f_K$ on $G(x_1, \dots, x_n)$, they satisfy the same inequality constraints as $f_1, \dots, f_K$ (that is non-negativity, and summing up to 1) on $G(x_1, \dots, x_n)$. From lemma 4.20, $\tilde{f}_1, \dots, \tilde{f}_K$ must satisfy these constraints everywhere.

That $\tilde{f}_1, \dots, \tilde{f}_K$ are in $\mathcal{F}_0$, satisfy the positivity constraint, and sum to 1 everywhere, imply that that $\tilde{f}$ defined above is in $\mathcal{F}^{Id}$.

Step 2: The feasible set of (4.14) **is included in $\mathcal{F}^{Id}$.** This follows directly from lemmas 4.19 and 4.20. □

Proof of lemma 4.19. Let $\tilde{f}$ be of the form $x \mapsto \sum_{j=1}^m \beta_j \mathbf{1}\{x \geq \tilde{x}_j\}$ such that for every $j \in [m]$, $\tilde{f}(\tilde{x}_j) = f(\tilde{x}_j)$. Let us show that $\|\tilde{f}\|_v \leq \|f\|_v \leq M$. We have that

$$\begin{aligned}
V(f, [0,1]^d) =& \sup_{\mathcal{P}} \sum_{R \in \mathcal{P}} |\Delta(f, R)| \\
=& \sup_{\substack{\mathcal{P}' = \mathcal{P} \cap \mathcal{P}(x_1,\ldots,x_n) \\ \mathcal{P} \text{ rect. split}}} \sum_{R \in \mathcal{P}} |\Delta(f, R)| \\
\geq& \sup_{R \in \mathcal{P}(x_1,\ldots,x_n)} |\Delta(f, R)| \\
=& \sup_{R \in \mathcal{P}(x_1,\ldots,x_n)} |\Delta(\tilde{f}, R)| \\
=& \sum_{R \in \mathcal{P}(x_1,\ldots,x_n)} |\Delta(\tilde{f}, R)| \\
=& V(\tilde{f}, [0,1]^d).
\end{aligned}$$

The second line in the above display follows from corollary 4.2. The third line follows from lemma 4.21. The fourth line follows from the fact that, as $|\Delta(f, R)|$ only depends on f through its values at the corners of R, which, for R in $\mathcal{P}(x_1, \ldots, x_n)$, are points of $G(x_1, \ldots, x_n)$, at which f and $\tilde{f}$ coincide. The last line follows from corollary 4.3.

The above implies that $M \geq \|f\|_v \geq \|\tilde{f}\|_v = \sum_{j=1}^m |\beta_j|$, where the last equality follows from lemma 4.22.

We have thus shown that for every $f \in \mathcal{F}_0$, we can find an $\tilde{f} \in \tilde{F}_0(x_1, \ldots, x_n)$ that coincides with $G(x_1, \ldots, x_n)$.

It remains to show that $\tilde{\mathcal{F}}_0(x_1, \ldots, x_n) \subseteq \mathcal{F}_0$. It is clear that the elements of $\tilde{F}_0(x_1, \ldots, x_n)$ are cadlag. From lemma 4.22, the definition of $\tilde{\mathcal{F}}_0(x_1, \ldots, x_n)$ implies that its elements have sectional variation norm smaller than M. Therefore, $\tilde{\mathcal{F}}_0(x_1, \ldots, x_n) \subseteq \mathcal{F}_0$. □

4.H.4 Technical lemmas on splits and Vitali variation

Effect on Vitali variation and absolute pseudo-volume of taking finer splits

The following lemma says that the sum over a split of the absolute pseudo-volume ascribed by f increases as one refines the split.

Lemma 4.21. *Let $f : [0,1]^d \to \mathbb{R}$. Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two rectangular splits of $[0,1]^d$. Define*

$$\mathcal{P}_1 \cap \mathcal{P}_2 \doteq \{R_1 \cap R_2 : R_1 \in \mathcal{P}_1,\ R_2 \in \mathcal{P}_2\}.$$

It holds that

$$\sum_{R \in \mathcal{P}_1} |\Delta^{(d)}(f, R)| \leq \sum_{R' \in \mathcal{P}_1 \cap \mathcal{P}_2} |\Delta(f, R')|.$$

We relegate the proof at the end of this section. The following lemma has the following corollary.

Corollary 4.2. *For any function $f : [0,1]^d \to \mathbb{R}$ and any rectangular split $\mathcal{P}_0$ of $[0,1]^d$, the Vitali variation of f, which we recall is defined as $V^{(d)}(f) \doteq \sup_{\mathcal{P} \text{ rect. split}} \sum_{R \in \mathcal{P}} |\Delta(f, R)|$ can actually be written as*

$$V^{(d)} = \sup_{\substack{\mathcal{P}' = \mathcal{P} \cap \mathcal{P}_0 \\ \mathcal{P} \text{ rect. split}}} \sum_{R \in \mathcal{P}'} |\Delta(f, R')|.$$

Proof of corollary 4.2. Observe that the set of rectangular splits $\{\mathcal{P} \cap \mathcal{P}_0 : \mathcal{P} \text{ rect. split}\}$ is included in the set of all rectangular splits. Therefore,

$$\sup_{\substack{\mathcal{P}' = \mathcal{P} \cap \mathcal{P}_0 \\ \mathcal{P}' \text{ rect. split}}} \sum_{R \in \mathcal{P}'} |\Delta(f, R)| \leq \sum_{\mathcal{P} \text{split}} |\Delta(f, R)|.$$

Lemma 4.21 implies the converse inequality:

$$\sup_{\substack{\mathcal{P}' = \mathcal{P} \cap \mathcal{P}_0 \\ \mathcal{P}' \text{ rect. split}}} \sum_{R \in \mathcal{P}'} |\Delta(f, R)| \geq \sum_{\mathcal{P} \text{split}} |\Delta(f, R)|.$$

We therefore have the wished equality. □

Vitali variation of piecewise constant functions

The following lemma characterizes the sum over a rectangular split of the absolute pseudo-volumes of a function that is piecewise constant on the rectangles of that split.

Lemma 4.22. *Let $x_1, \ldots, x_n \in [0,1]^d$ and let $\tilde{x}_1, \ldots, \tilde{x}_m$ be the elements of $G(x_1, \ldots, x_n)$. Consider a function f of the form*

$$f : x \mapsto \sum_{j=1}^{m} \beta_j \mathbf{1}\{x \geq x_j\},$$

It holds that

$$V_{HK,\mathbf{0}}(f) = \sum_{j=1}^{m} |\beta_j|.$$

Corollary 4.3 (Vitali variation of rectangular piecewise constant function)**.** *Let $x_1, \ldots, x_n \in [0,1]^d$, let $\tilde{x}_1, \ldots, \tilde{x}_m$ be the elements of $G(x_1, \ldots, x_m)$, and consider a function f of the form*

$$f : x \mapsto \sum_{j=1}^{J} \beta_j \mathbf{1}\{x \geq \tilde{x}_j\}.$$

Then

$$\sup_{\mathcal{P} \text{ rect. split}} |\Delta(f, R)| = \sum_{R \in \mathcal{P}(x_1,\ldots,x_n)} |\Delta(f, R)|,$$

where $\mathcal{P}(x_1, \ldots, x_n)$ is a minimal rectangular split induced by $x_1, \ldots, x_m$.

Proof of lemma 4.21. Consider a rectangle $R \in \mathcal{P}(x_1, \ldots, x_n)$. There exist $k, l \in [m]$ such that $R = [\tilde{x}_k, \tilde{x}_l]$. (Since $\mathcal{P}(x_1, \ldots, x_n)$ is a minimal split, we must have $\tilde{x}_k < \tilde{x}_l$ as otherwise the corresponding minimum grid would have duplicate points and would therefore not be minimal). Observe that

$$\begin{aligned}\Delta(f, [\tilde{x}_k, \tilde{x}_l]) =& \Delta\left(\sum_{j=1}^{m} \beta_j \mathbf{1}\{\cdot \geq \tilde{x}_j\}, [\tilde{x}_k, \tilde{x}_l]\right) \\ =& \beta_j \sum_{j=1}^{m} \Delta\left(\mathbf{1}\{\cdot \geq \tilde{x}_j\}, [\tilde{x}_k, \tilde{x}_l]\right),\end{aligned}$$

as the operator $f' \mapsto \Delta(f', [\tilde{x}_k, \tilde{x}_l])$ is linear. Let us calculate $\Delta(1\{\cdot \geq \tilde{x}_j\}, [\tilde{x}_k, \tilde{x}_l])$ for every $j \in [m]$. We have that

$$\begin{aligned}&\Delta(\mathbf{1}\{\cdot \geq \tilde{x}_j\}, [\tilde{x}_k, \tilde{x}_l]) \\ =& \sum_{j_1,\ldots,j_d \in \{0,1\}} (-1)^{j_1+\ldots+j_d} 1\left\{\tilde{x}_{l,1} + j_1(\tilde{x}_{k,1} - \tilde{x}_{l,1}) \geq \tilde{x}_{j,1}, \ldots, \tilde{x}_{l,d} + j_d(\tilde{x}_{k,d} - \tilde{x}_{l,d}) \geq \tilde{x}_{j,d}\right\}.\end{aligned} \tag{4.22}$$

From there, we distinguish three cases.

Case 1: There exists $i \in [d]$ such that $\tilde{x}_{j,i} > \tilde{x}_{l,i}$. Then, all terms in (4.22) are zero and thus $\Delta(\mathbf{1}\{\cdot \geq \tilde{x}_j\}, [\tilde{x}_k, \tilde{x}_l]) = 0$.

Case 2: $\tilde{x}_j = \tilde{x}_l$. Then, only the term in (4.22) corresponding to $j_1 = \ldots = j_d = 0$ is non-zero and thus $\Delta(\mathbf{1}\{\cdot \geq \tilde{x}_j\}, [\tilde{x}_k, \tilde{x}_l]) = 1$.

Case 3: $\tilde{x}_j \leq \tilde{x}_l$ and $\tilde{x}_j \neq \tilde{x}_l$. Then denote

$$\begin{aligned} I =& \{i \in [d] : \tilde{x}_{j,i} = \tilde{x}_{l,i}\} \\ \text{and } I^c =& [d] \backslash I.\end{aligned}$$

As $\tilde{x}_j \neq \tilde{x}_l$, $I^c \neq \emptyset$. Denote $i_1, \ldots, i_q$ the elements of I^c, where $q = |I^c|$. Then

$$\begin{aligned}&\Delta(\mathbf{1}\{\cdot \geq \tilde{w}_j\}, [\tilde{x}_k, \tilde{x}_l]) \\ =& \sum_{j_1,\ldots,j_d \in \{0,1\}} (-1)^{j_1+\ldots+j_d} 1\left\{\tilde{x}_{l1} + j_1(\tilde{x}_{k,1} - \tilde{x}_{l,1}) \geq \tilde{x}_{j,1}, \ldots, \tilde{x}_{l,d} + j_d(\tilde{x}_{k,d} - \tilde{x}_{l,d}) \geq \tilde{x}_{j,d}\right\}\end{aligned}$$

$$\begin{aligned}
&= \sum_{j_1,\dots j_d \in \{0,1\}} (-1)^{j_1+\dots+j_d} \mathbf{1}\{\forall i \in I, j_i = 0\} \\
&= \sum_{j_{i_1},\dots,j_{i_q} \in \{0,1\}} (-1)^{j_{i_1}+\dots+j_{i_q}} \\
&= \sum_{j_{i_1}=0}^{1} (-1)^{j_{i_1}} \sum_{j_{i_2}=0}^{1} (-1)^{j_{i_2}} \dots \sum_{j_{i_q}=0}^{1} (-1)^{j_{i_q}} \\
&=0.
\end{aligned}$$

Therefore, we have shown that, for all $j = 1, \dots, m$,

$$\Delta(\mathbf{1}\{\cdot \geq \tilde{x}_j\}, [\tilde{x}_k, \tilde{x}_l]) = \begin{cases} 1 & \text{if } \tilde{x}_j = \tilde{x}_k, \\ 0 & \text{otherwise.} \end{cases}$$

This implies that

$$|\Delta(f, [\tilde{w}_k, \tilde{w}_l]) = |\beta_k|.$$

which concludes the proof. □

Proof of corollary 4.3. From lemma 4.21,

$$\sup_{\mathcal{P} \text{ split}} \sum_{R \in \mathcal{P}} |\Delta(f, R)| = \sup_{\substack{\mathcal{P}' = \mathcal{P} \cap \mathcal{P}(x_1,\dots,x_n) \\ \mathcal{P}' \text{rect. split}}} \sum_{R \in \mathcal{P}'} \sum_{R \in \mathcal{P}'} |\Delta(f, R)|.$$

Consider a split $\mathcal{P}'$ of the form $\mathcal{P} \cap \mathcal{P}(x_1, \dots, x_n)$. We can write the corresponding rectangular grid as $\tilde{x}_1, \dots, \tilde{x}_m, \tilde{x}_{m+1}, \dots, x_{m'}$ where $\tilde{x}_1, \dots, \tilde{x}_m$ are the points of $G(x_1, \dots, x_n)$. We can rewrite f as

$$f : x \mapsto \sum_{j=1}^{m} \beta_j \mathbf{1}\{x \geq \tilde{x}_j\},$$

with $\beta_{m+1} = \dots = \beta_{m'} = 0$. From lemma 4.22, we have that

$$\sum_{R \in \mathcal{P}'} |\Delta(f, R)| = \sum_{j=1}^{m'} |\beta_j| = \sum_{j=1}^{m} |\beta_j| = \sum_{R \in \mathcal{P}(x_1,\dots,x_n)} |\Delta(f, R)|.$$

Therefore

$$\sup_{\substack{\mathcal{P}' = \mathcal{P} \cap \mathcal{P}(x_1,\dots,x_n) \\ \mathcal{P} \text{ rect. split}}} \sum_{R \in \mathcal{P}'} |\Delta(f, R)| = \sum_{R \in \mathcal{P}(x_1,\dots,x_n)} |\Delta(f, R)|.$$

□

Chapter 5

Nonparametric learning from sequentially collected data

AURÉLIEN BIBAUT, ANTOINE CHAMBAZ, MARK VAN DER LAAN

In this chapter, we give guarantees on estimators of infinite dimensional parameters of the uncontrolled part q of the data-generating distribution in a sequential trial. Our contributions are two-fold.

First, we give high probability deviations bounds for empirical risk minimizers over nonparametric function classes. We give applications to outcome model learning and policy learning in the contextual bandit setting, and on sequential learning of the transition density in the Markov Decision Process setting.

Second, we give new guarantees on the sequential Super Learner. Our new theorem improves on the original theorem from Benkeser et al. [2018] by avoiding a condition that is hard to check.

5.1 Introduction

5.1.1 Data and statistical experiment

Suppose that an experimenter interacts with an environment along T time steps. At time step t, the experimenter observes a vector of pre-treatment covariates $L_1(t)$ lying in a Euclidean set $\mathcal{L}_1$ characterizing the state of the environmnet, then assigns treatment $A(t) \in \mathcal{A} := [K]$, for some $K \geq 2$, and then observes a vector of post-treatment covariates $L_2(t) \in \mathcal{L}_2$. We denote $O(t) := (L_1(t), A(t), L_2(t))$ the data observed at step t, and we let $\bar{O}(t) := (O(1), \ldots, O(t))$. We adopt the convention that $\bar{O}(0)$ is a known constant, so that conditioning on it leaves unchanged a probability distribution.

5.1.2 Statistical models

Denote P_0^T the true distribution of $\bar{O}(T)$, and let $\mathcal{M}^T$ the statistical model which we assume to contain P_0^T. We denote P^T a generic element of $\mathcal{M}^T$. We suppose that there exists a dominating measure μ^T such that every element P^T admits a density w.r.t. μ. For any P^T, we denote p^T this density. For any $\bar{o}(T) \in \mathcal{O}^T$, any such density can be factored as

$$p^T(\bar{o}(T)) = \prod_{t=1}^{T} q_1(l_1(t) \mid \bar{o}(t-1)) \widetilde{g}_t(a(t) \mid \bar{o}(t-1), l_1(t)) q_2(l_2(t) \mid \bar{o}(t-1), l_1(t), a(t)).$$

We consider two statistical models: the stochastic contextual bandit model, and the Markov Decision Process model, as defined in the introduction chapter. We recall these models below.

Stochastic contextual bandit. In the stochastic contextual bandit model, we assume that $L_1(1)$, $L_1(2), \ldots$ form an i.i.d. sequence of random variables that we refer to as contexts. For every t, we assume that $L_2(t)$ is a reward that lies in $[0, 1]$ and that depends on the past $\bar{O}(t-1), L_1(t), A(t)$ only through $L_1(t), A(t)$. We suppose that the distribution of $L_2(t)$ given $A(t), L_1(t)$ is the same for every t. As a result, for any distribution P^T in the model $\mathcal{M}^T$, the corresponding density can be factorized as

$$p^T(\bar{o}(T)) = \prod_{t=1}^{T} q_1(l_1(t))\widetilde{g}_t(a(t) \mid \bar{o}(t-1), l_1(t))q_2(l_2(t) \mid l_1(t), a(t)).$$

Under this model, the action at t that maximizes the expected reward depends only on $l_1(t)$. We call a policy a conditional distribution $(l_1, a) \in \mathcal{L}_1 \times \mathcal{A} \mapsto g(a \mid l_1)$. We can write the likelihood of the data $\bar{O}(T)$ as follows:

$$p^T(\bar{O}(T)) = \prod_{t=1}^{T} q_1(L_1(t))g_t(A(t) \mid L_1(t))q_2(L_2(t) \mid L_1(t), A(t)),$$

with $g_t(a(t) \mid l_1(t)) := \widetilde{g}_t(a(t) \mid l_1(t), \bar{O}(t-1))$. Note that the policy g_t is an $\bar{O}(t-1)$ random function, unlike $\widetilde{g}_t$, which is a fixed function.

Markov Decision Process. We let $L_2(t) = 0$ for every $t \geq 1$. The Markov Decision Process model assumes that $L_1(t)$ depends on the past $\bar{O}(t-1)$ only through the latest state $L_1(t-1)$ and the latest action $A(t-1)$, and that the conditional distribution of $L_1(t)$ given $L_1(t-1)$ and $A(t-1)$ is the same across time points. As a result, for any distribution P^T in the model $\mathcal{M}^T$, the corresponding density p^T can be factorized as

$$p^T(\bar{o}(T)) = \prod_{t=1}^{T} q(l_1(t) \mid l_1(t-1), a(t-1))\widetilde{g}_t(a(t) \mid \bar{o}(t-1), l_1(t)).$$

Under the MDP model too, the action that maximizes the reward to go at t depends only on $l_1(t)$, which motivates introduce the notion of policies defined as conditional distributions over $\mathcal{A}$ given a value $l_1 \in \mathcal{L}_1$ of the latest state. As in the CB setting, we can write the likelihood of $\bar{O}(t-1)$ as

$$p^T(\bar{O}(T)) = \prod_{t=1}^{T} q_1(L_1(t) \mid L_1(t-1), A(t-1))g_t(A(t) \mid L_1(t))q_2(L_2(t) \mid L_1(t), A(t)),$$

with $g_t(a(t) \mid l_1(t)) := \widetilde{g}_t(a(t) \mid l_1(t), \bar{O}(t-1))$. Here too g_t is an $\mathcal{O}(t-1)$ measurable object.

5.1.3 Target parameters

In the contextual bandit model, parameters of interest include the outcome model

$$(l_1, a) \mapsto \bar{Q}(a, l_1) := E_{q_2}\left[L_2(t) \mid L_1(t), A(t)\right],$$

and, given a policy class $\mathcal{G}$, an optimal policy $g^* \in \mathcal{G}$, that is any element of $\arg\max \mathcal{V}(q, g)$, where $\mathcal{V}(q, g)$ is the value of g under q, as defined in the introduction chapter.

In the MDP setting, a parameter of interest is the conditional density q_1. Note that if one knows the true transition density $q_{0,1}$, one can find an optimal policy, that is a policy that maximizes the expected cumulative reward (or the expected cumulated discounted reward).

5.2 High probability bounds for empirical risk minimizers

In this section, we give high probability bounds on the excess risk of empirical risk minimizers over a nonparametric class under a bracketing entropy assumption. Let us first introduce the setting. Let $\mathcal{F}$ be a class of functions $\mathcal{O} \times \mathbb{R}$ (or $\mathcal{L}_1 \times \mathcal{A} \times \mathcal{L}_1 \to \mathbb{R}$). Let ℓ be a mapping such that, for any $f : \mathcal{O} \times \mathbb{R}$ or any $f : \mathcal{L}_1 \times \mathcal{A} \times \mathcal{L}_1 \to \mathbb{R}$, $\ell(f)$ is a function $\mathcal{O} \to \mathbb{R}$ or $\mathcal{L}_1 \times \mathcal{A} \times \mathcal{L}_1 \to \mathbb{R}$. We will interpret ℓ as a loss function. Let

$$R_t(f) := E_{q_1, g_{\mathrm{ref}}, q_2} \left[\ell(f)(O(t)) \mid \bar{O}(t-1) \right]$$

and

$$\bar{R}_T(f) := \frac{1}{T} \sum_{t=1}^{T} R_t(f)$$

be the conditional population risk, and the average conditional population risk. Let ℓ_t be the importance sampling weighted loss, defined for any $f \in \mathcal{F}$ as $\ell_t(f) := (g_{\mathrm{ref}}/g_t)\ell(f)$, and let

$$\widehat{R}_T(f) := \frac{1}{T} \sum_{t=1}^{T} \ell_t(f)(O(t)),$$

be the empirical (importance-sampling weighted) risk, where $g_{\mathrm{ref}}(a \mid l_1) := K^{-1}$. We denote $R_{0,t}$ and $\bar{R}_{0,T}$ the corresponding risks under $q_0 := (q_{0,1}, q_{0,2})$.

Let $f_1 \in \mathcal{F}$ be a fixed function, and let $\widehat{f}_T$ be an empirical risk minimizer over $\mathcal{F}$, that is a function $f \in \mathcal{F}$ such that

$$\widehat{R}_T(\widehat{f}_T) := \inf_{f \in \mathcal{F}} \widehat{R}_T(f).$$

For any $f : \mathcal{O} \to \mathbb{R}$, let

$$\sigma_T^2(f) := \frac{1}{T} \sum_{t=1}^{T} E_{q_1, g_{\mathrm{ref}}, q_2} \left[f(O(t))^2 \mid \bar{O}(t-1) \right].$$

It is straightforward to observe that σ_T is a norm. In the upcoming theorem, we give a high probability bound on the population risk difference $R_T(\widehat{f}_T) - R_T(f^*)$ of of the empirical risk minimizer. Our theorem relies on the following assumptions.

Assumption 5.1 (Entropy)**.** *There exists $p > 0$ such that*

$$\log N_{[]}(\epsilon, \ell(\mathcal{F}), \sigma_T) \lesssim \epsilon^{-p},$$

where $\ell(\mathcal{F}) := \{\ell(f) : f \in \mathcal{F}\}$.

The following assumption is a so-called *variance bound*. It is a common condition in the study of ERMs.

Assumption 5.2 (Variance bound)**.** *There exists* $\alpha > 0$ *such that, for any* $f \in \mathcal{F}$*,*

$$\sigma_T^2(\ell(f) - \ell(f_1)) \lesssim (R_T(f) - R_T(f_1)))^\alpha .$$

The following last assumption is a bound on the importance sampling ratios g_{ref}/g_t.

Assumption 5.3 (Importance sampling ratios bound)**.** *There exists* $\delta > 0$ *such that*

$$\sup_{t \in [T]} \|g_{\mathrm{ref}}/g_t\|_\infty \leq \delta^{-1}$$

almost surely.

Assumption 5.4 (Radius of $\ell(\mathcal{F})$)**.** *There exists* $B > 0$ *and* $r_0 > 0$ *such that*

$$\sup_{f \in \mathcal{F}} \|f\|_\infty \leq B \qquad \textit{and} \qquad \sup_{f \in \mathcal{F}} \sigma_T(f) \leq r_0.$$

Assumption 5.5 (Convexity)**.** *The class* $\mathcal{F}$ *is convex and the mapping* ℓ *is convex over* $\mathcal{F}$*.*

Theorem 5.1 (High probability bound for ERMs)**.** *Suppose that assumptions 5.1-5.5 hold. Then, with probability at least* $1 - 2e^{-x}$*,*

$$R_T(\widehat{f}_T) - R_T(f_1) \lesssim (\delta T)^{-\frac{1}{2-\alpha+p\alpha/2}} + \left(\frac{x}{\delta T}\right)^{\frac{1}{2-\alpha}} + \frac{Bx}{\delta T}$$

if $p \in (0, 2)$*, and*

$$R_T(\widehat{f}_T) - R_T(f_1) \lesssim (\delta T)^{-\frac{1}{p}} + r_0\sqrt{\frac{x}{\delta T}} + \frac{Bx}{\delta T}$$

if p *¿ 2.*

5.3 ERM for policy learning in the CB setting

5.3.1 Two approaches to policy learning

Policy learning objective. In the contextual bandit setting, one learning goal is to estimate an optimal policy, that is a to find a policy g that maximizes the value $\mathcal{V}(q, g)$, as defined in the introduction chapter, over a certain policy class $\mathcal{G}$. Recall that

$$\mathcal{V}(q, g) := E_{q_1}\left[\sum_{a=1}^{K} \bar{Q}(a, L_1) g(a \mid L_1)\right].$$

We define the risk of a policy g as its negative value:

$$R(q, g) := -\mathcal{V}(q, g).$$

Let $O^{\mathrm{ref}} := (L_1^{\mathrm{ref}}, A^{\mathrm{ref}}, L_2^{\mathrm{ref}}) \sim P^{\mathrm{ref}}$ where $dP^{\mathrm{ref}}/d\mu = q_1 g_{\mathrm{ref}} q_2$. We introduce a loss ℓ such that for any policy g, and any (l_1, a, l_2) $\ell(f)(l_1, a, l_2) = -\frac{1}{g_{\mathrm{ref}}(a|l_1)} l_2 g(a \mid l_1)$, and, for any $t \geq 1$ let $\ell_t(f) := (g_{\mathrm{ref}}/g_t)\ell(f)$. Observe that

$$R(q, g) = Eq_1, g_{\mathrm{ref}}, q_2\left[\ell(f)(O^{\mathrm{ref}})\right] = E_{q_1, g_t, q_2}\left[\ell_t(O(t)) \mid \bar{O}(t-1)\right].$$

Reduction to cost sensitive classification. The policy optimization problem in the contextual bandit setting can be recast as a cost-sensitive classification problem in which $l_1 \mapsto g(\cdot \mid l_1)$ is a classifier that maps a vector of predictors l_1 to a probability distribution over labels $1, \ldots, K$, and where $-\bar{Q}(a, l_1)$ is the cost of predicting label a under l_1. The connection between policy learning and cost-sensitive classification has been made in numerous articles in the past (see e.g. Zhao et al. [2012]). It is a useful connection as it allows in particular to reuse the theory of classification calibration (see in particular Bartlett et al. [2006]).

We consider two approaches to policy learning.

Direct policy optimization. Let

$$\widehat{R}_T(g) := \frac{1}{T} \sum_{t=1}^{T} \ell_t(g)(O(t)),$$

the (importance sampling weighted) empirical risk of policy g. The first approach to policy learning that we consider is to directly minimize w.r.t. g the empirical risk $\widehat{R}_T(g)$. We denote $\widehat{g}_T$ a minimizer over the policy class $\mathcal{G}$ of $\widehat{R}_T$. We refer to this approach as direct policy optimization. One major caveat of this approach, is that for most natural policy classes, optimizing $\widehat{R}_T(g)$ is computationally intractable. We give an example of a class over which it is tractable in chapter 4.

Optimizing a convex surrogate risk. In the second approach, we consider policies of the form

$$g_f(a \mid l_1) = \mathbf{1}\{a = \arg\max f(a, l_1)\}$$

for functions living in a certain class $\mathcal{F}$ such that $\sum_{a=1}^{K} f(a, l_1) = 0$ for every l_1. We define a convex surrogate loss as

$$\ell^{\phi}(f)(o) := -y\phi(f(a, l_1)),$$

where ϕ is a convex function. Common examples for ϕ include the hinge surrogate $\phi^{\text{hinge}} : x \mapsto (1 + x)^+$ and the truncated square surrogate $\phi^{\text{tsq}} : x \mapsto (1 + x)_+^2$. We define the importance sampling ϕ-loss as $\ell_t^{\phi}(f) := (g_{\text{ref}}/g_t)\ell(f)$, and the ϕ-risk as

$$R^{\phi}(q, f) := E_{q_1, g_{\text{ref}}, q_2}\left[\ell^{\phi}(f)(O^{\text{ref}})\right] = E_{q_1, g_t, q_2}\left[\ell_t(f)(O(t)) \mid \bar{O}(t-1)\right].$$

In the surrogate setting, we optimize the ϕ-risk w.r.t. f over $\mathcal{F}$. We define the empirical ϕ-risk as

$$\widehat{R}_T(f) := \frac{1}{T} \sum_{t=1}^{T} \ell(t)(O(t-1)).$$

Let $\widehat{f}_T$ be an empirical ϕ-risk minimizer over $\mathcal{F}$. Our fitted policy is then $g_{\widehat{f}_T}$. We define the excess risk $R(q, g_{\widehat{f}_T}) - \inf_g R(q, g)$ and terms of the excess ϕ-risk $R^{\phi}(q, \widehat{f}_T) - \inf_f R^{\phi}(q, f)$, where in the first expression, the infimum is over measurable policies, and in the second over measurable functions $f : \mathcal{L}_1 \times \mathcal{A} \to \mathbb{R}$ such that $\sum_{a=1}^{K} f(a, l_1) = 0$ for every l_1. Classification calibration theory allows us to bound the former in terms of the latter, under the so-called *realizability* assumption, which we present next.

Assumption 5.6 (Realizability)**.** *The function class $\mathcal{F}$ contains a minimizer of $R^{\phi}(q, f)$ over the set of measurable functions $f : \mathcal{L}_1 \times \mathcal{A} \to \mathbb{R}$ such that $\sum_{a=1}^{K} f(a, l_1) = 0$ for all $l_1 \in \mathcal{L}_1$.*

Obtaining guarantees on the excess risk. Consider either of the two approaches. In the surrogate setting, we denote $\widehat{g}_T = g_{\widehat{f}_T}$ so as to use the same notation for the fitted policy in both setting. We care about obtaining guarantees on $R(q, \widehat{g}_T) - \inf_{g \in \mathcal{G}} R(q, g)$, the excess risk relative to policy class $\mathcal{G}$. Note that under the realizability assumption, the excess risk relative to $\mathcal{G}$ coincides with the excess risk relative to all measurable policies.

In the direct policy optimization setting, we obtain guarantees directly by applying theorem 5.1.

In the surrogate setting, we obtain guarantees by first applying theorem 5.1 to bound the excess ϕ-risk, and we then use classification calibration theory to bound the excess risk in terms of the excess ϕ-risk.

5.3.2 Variance bounds

While guarantees can be obtained in the absence of a variance bound, that is when assumption 5.2 holds only for $\alpha = 0$ (in which case the bound is vacuous, as the bound always hold for $\alpha = 0$ if ℓ, or ℓ^{ϕ} is bounded over $\mathcal{F}$), we can obtain tighter bounds in the presence of such a variance bound. We consider two situations under which we can obtain a variance bound.

Variance bounds under non-zero modulus of convexity

The first situation only applies to the surrogate case, when the surrogate ϕ has non-zero *modulus of convexity*. We restate below the definition from Bartlett et al. [2006] of the modulus of convexity of a convex function.

Definition 5.1 (Modulus of convexity)**.** *Given a pseudo-metric d on a vector space S, and a convex function $f : S \to \mathbb{R}$, the modulus of convexity of f is defined as a function $\delta : [0, \infty) \to [0, \infty]$ such that*

$$\delta(\epsilon) = \inf \left\{ \frac{f(x_1) + f(x_2)}{2} - f\left(\frac{x_1 + x_2}{2}\right) : x_1, x_2 \in S, d(x_1, x_2) \geq \epsilon \right\}.$$

The following lemma gives a variance bound that holds when the modulus of convexity is quadratic.

Lemma 5.1 (Variance bound from modulus of convexity)**.** *Suppose that ϕ has modulus of convexity δ such that $\delta(\epsilon) \gtrsim \epsilon^2$. Let f^* be a population ϕ-risk minimizer, that is an element of* $\arg\min_{f \in \mathcal{F}} R^{\phi}(q, f)$

$$\left\| \ell^{\phi}(f) - \ell^{\phi}(f^*) \right\|^2_{q_1, g_{\mathrm{ref}}, q_2} \lesssim R(q, f) - R(q, f^*).$$

The proof of the above lemma is a straightforward extension to the cost-sensitive setting of lemmas 15 and 16 in Bartlett et al. [2006].

Variance bounds under Tsybakov noise assumption and realizability

In the direct policy optimization setting, and in the surrogate setting under surrogates that have zero modulus of convexity, variance bounds can still be obtained under the conjuction of the realizability assumption, and of the Tsybakov noise assumption, which we present next. We present it in the case $K = 2$. Extensions to the case $K > 2$ exist, but deriving the corresponding variance bounds appears to tedious. We leave it to future work. The version of the Tsybakov noise assumption that we present is a generalization to the cost-sensitive classification setting. One instance of this CSC Tysbakov noise assumption can be found in Zhao et al. [2012].

Assumption 5.7 (Tsybakov noise assumption)**.** *Let $\eta : l_1 \mapsto \bar{Q}(2, l_1)/(\bar{Q}(1, l_1) + \bar{Q}(2, l_2))$. We say that q satisfies the Tsybakov noise assumption if there exists $\nu > 0$ such that*

$$\Pr_q\left[|2\eta(L_1(1)) - 1| \geq t\right] \lesssim t^\nu$$

for every $t > 0$.

The following lemma gives a variance bound under the Tsybakov noise assumption and the realizability assumption.

Lemma 5.2 (Variance bounds under Tsybakov and realizability)**.** *Suppose that assumptions 5.6 and 5.7 hold. Then, for every $f \in \mathcal{F}$. Let f^* be a population ϕ-risk minimizer, that is an element of $\arg\min_{f \in \mathcal{F}} R^\phi(q, f)$. (Note that under since we are making the realizability assumption, f^* minimizes the ϕ-risk over measurable functions such that $\sum_{a=1}^K f(a, l_1) = 0$ for all l_1.). Let $\phi \in \{\phi^{\text{hinge}}, \phi^{\text{Id}}\}$, where $\phi^{\text{Id}} : u \mapsto u$. Observe that the loss ℓ we use in the direct policy optimization setting equals $\ell^{\phi^{\text{Id}}}$.*

We then have that, for any $f \in \mathcal{F}$,

$$\left\|\ell^\phi(f) - \ell^\phi(f^*)\right\|^2_{q_1, g_{\text{ref}}, q_2} \lesssim \left(R^\phi(q, f) - R^\phi(q, f^*)\right)^\alpha,$$

with $\alpha := \nu/(\nu + 1)$.

The proof in the hinge case is a straightforward extension to the CSC setting of lemma 6.1 in Steinwart and Scovel [2007]. The proof in the case $\phi = \phi^{\text{Id}}$ can be found in various existing articles such as Chambaz et al. [2017].

5.3.3 Classification calibration bounds

Let $R^*(q) := \inf_g R(q, g)$ and let $R^{*,\phi} \inf_f R^{*,\phi}(q) := \inf_f R^\phi(q, f)$, where the suprema are over measurable policies and measurable functions $f : \mathcal{L}_1 \times \mathcal{A} \to \mathbb{R}$ such that $\sum_{a=1}^K f(a, l_1) = 0$ for all $l_1 \in \mathcal{L}_1$. Under the realizability assumption and the Tsybakov noise assumption with exponent ν (no Tsybakov noise assumption corresponds to $\nu = 0$), a straightforward extension of theorem 10 in Bartlett et al. [2006] gives that, for every f

$$(R(q, g_f) - R^*(q))^\alpha \omega\left((R(q, g_f) - R^*(q))^{1-\alpha}\right) \lesssim R^\phi(q, f) - R^{*,\phi}(q),$$

for some function $\omega : \mathbb{R} \to \mathbb{R}$ that depends on ϕ. For $\phi = \phi^{\text{tsq}}$, $\omega(u) = u^2$, and for $\phi = \phi^{\text{hinge}}$, $\omega(u) = u$.

5.3.4 Theorem statement

Theorem 5.2. *Suppose that the entropy assumption (assumption 5.1), the importance sampling ratio bound assumption (assumption 5.3), the convexity assumption (assumption 5.5), the assumption on the radius of the loss class (assumption 5.4), and the Tsybakov noise assumption (assumption 5.7) hold. Let $\alpha := \nu/(\nu+1)$. Suppose that $\phi \in \{\phi^{Id}, \phi^{\text{hinge}}, \phi^{\text{tquad}}, \}$.*

We consider various situations depending on ϕ, the values of the entropy exponent p and of the Tsybakov noise exponent ν, and of whether the realizability assumption 5.6 holds.

Then table 5.1 gives the cases where an inequality of the type

$$\Pr_q \left[R(q, \widehat{g}_t) - \inf_{g \in \mathcal{G}} R(q,g) \gtrsim (T\delta)^{-\beta} + x \right] \lesssim \exp\left(-CT\delta(x^\gamma \wedge x)\right),$$

holds provide and explicit values for the exponents β and γ.

Table 5.1: Coefficients β and γ of the high probability bound on $R(q, \widehat{g}_t) - \inf_{g \in \mathcal{G}} R(q,g)$

ϕ	Realizability	p	β	γ
Id	No	$\in (0,2)$	$\frac{1}{2}$	2
		> 2	$\frac{1}{p}$	2
	Yes	$\in (0,2)$	$\frac{1}{2-\alpha+p\alpha/2}$	$2-\alpha$
		> 2	$\frac{1}{p}$	$2-\alpha$
hinge	No	No consistency result		
	Yes	$\in (0,2)$	$\frac{1}{2-\alpha+p\alpha/2}$	$2-\alpha$
		> 2	$\frac{1}{p}$	$2-\alpha$
tquad	No	No consistency result		
	Yes	$\in (0,2)$	$\frac{1}{(2-\alpha)(1+p/2)}$	$2-\alpha$
		> 2	$\frac{1}{p(2-\alpha)}$	$2-\alpha$

5.4 ERM for transition density learning in the MDP model

Since in our formulation of the MDP model we do not use the variables $L_2(t)$, we simplify notation by using $L(t)$ for $L_1(t)$ and q for q_1.

Consider the negative log likelihood loss, defined for any q as

$$\ell(q)(l, a, l') = -\log q(l' \mid a, l).$$

Under the MDP model, the population risk defined earlier can be written as

$$R_T(q) := \frac{1}{T}\sum_{t=1}^{T} E_q\left[\ell(q)(L(t), A(t), L(t+1)) \mid O(t)\right]$$

and the empirical risk as

$$\widehat{R}_T(q) := \frac{1}{T}\sum_{t=1}^{T} \ell(q)(L(t), A(t), L(t+1)).$$

Let q_1 be a fixed element of $\mathcal{Q}$, and let $\widehat{q}_T \in \arg\min_{q\in\mathcal{Q}} \widehat{R}_T(q)$. Introduce the alternative loss $\widetilde{l}$, defined for any q as

$$\widetilde{l}(q) = -\log\frac{q+q_1}{2q_1},$$

and let

$$\widetilde{R}_T(q) = \frac{1}{T}\sum_{t=1}^{T} E_q\left[\ell(q)(L(t), A(t), L(t+1)) \mid O(t)\right]$$

and

$$\widehat{\widetilde{R}}_T(q) = \frac{1}{T}\sum_{t=1}^{T} \frac{g_{\mathrm{ref}}}{g_t}(A(t) \mid L(t))\widetilde{l}(q)(L(t), A(t), L(t+1)).$$

Since $\widetilde{l}$ is convex, and from the definition of $\widehat{q}_T$ we have that $\widehat{\widetilde{R}}_T(\widehat{q}_T) - \widehat{\widetilde{R}}_T(q_1) \leq 0$.

We will apply theorem 5.1 with the loss $\widetilde{l}$. The key step is to show a variance bound. In the upcoming lemma, we give one under the following assumption.

Assumption 5.8 (Lower bound on q_1)**.** *There exists $\eta > 0$ such that, $\inf_{l,a,l'} q_1(l' \mid l, a) \geq \eta$.*

Lemma 5.3. *Suppose that assumption 5.8 holds. Then, for every $q \in \mathcal{Q}$,*

$$\sigma_T(\widetilde{\ell}(q) - \widetilde{\ell}(q_1)) \lesssim \widetilde{R}_T(q) - \widetilde{R}_T(q_1).$$

Proof. For any two $q, q' \in \mathcal{Q}$, and $o \in \mathcal{O}$, let

$$h(q, q' \mid o) = \left(\int (\sqrt{q} - \sqrt{q'})^2 (l' \mid o) dl'\right)^{1/2},$$

the conditional Hellinger distance, and for any $f : \mathcal{O} \times \mathcal{L} \to \mathbb{R}$, let

$$\|f(\cdot \mid o)\|_{2,q,o} = \left(f^2(l' \mid o) q(l' \mid o) dl'\right)^{1/2},$$

the conditional L^2 norm under q, and

$$\|f(\cdot \mid o)\|_{B,q,o} = \left(\sum_{k \geq 2} \frac{1}{k!} \int |f|^k(l' \mid o) q(l' \mid o) dl' \right)^{1/2},$$

the conditional Bernstein norm w.r.t. q.

Following the arguments of section 3.4.1, we obtain that

$$\begin{aligned}
&\left\| (\widetilde{l}(q) - \widetilde{l}(q_1))(\cdot \mid O(t)) \right\|^2_{2,q,O(t)} \\
\lesssim &\left\| (\widetilde{l}(q) - \widetilde{l}(q_1))(\cdot \mid O(t)) \right\|^2_{B,q,O(t)} \\
\lesssim & h^2(q, q_1 \mid O(t)) \\
\lesssim & E\left[(\widetilde{l}(q) - \widetilde{l}(q_1))(O(t), L(t+1)) \mid O(t) \right].
\end{aligned}$$

Therefore, averaging over $t = 1, \ldots, T$, we obtain

$$\sigma_T^2((\widetilde{l}(q) - \widetilde{l}(q_1))) \lesssim \widetilde{R}_T(q) - \widetilde{R}_T(q_1).$$

□

As a corollary of the above variance bound lemma and of theorem 5.1, we have the following result on the ERM $\widehat{q}_T$.

Theorem 5.3. *Suppose that the entropy assumption (assumption 5.1) and assumption 5.4 hold for $\widetilde{l}(\mathcal{F})$, and that assumption 5.5 holds for $\widetilde{\ell}$. Suppose further that assumptions 5.3 and assumption 5.8. Then, with probability at least $1 - 2e^{-x}$, we have that, if $p \in (0, 2)$,*

$$h_T(\widehat{q}_T, q_1) \lesssim (\delta T)^{-\frac{1}{2+p}} + \sqrt{\frac{x(1+B)}{\delta T}},$$

and, if $p > 2$,

$$h_T(\widehat{q}_T, q_1) \lesssim (\delta T)^{-\frac{1}{2p}} + \sqrt{\frac{x(1+B)}{\delta T}},$$

where $h_T(q, q_1') := T^{-1} \sum_{t=1}^T h(q, q' \mid O(t))$.

5.5 Stronger guarantees for the sequential Super Learner

We consider the sequential Super Learner, a cross-validation model selector for sequentially collected data. The sequential Super Learner was initially proposed in Benkeser et al. [2018]. In section, we propose a new version of the guarantees of the sequential Super Learner, in which we remove assumption A3 of the original paper, which is hard to check.

We work in the general setting presented in section 5.1. We consider a loss ℓ over functions $f : \mathcal{O} \to \mathbb{R}$, such that, for any such f, $\ell(l(f)$ is a function $\mathcal{O} \to \mathbb{R}$. We let ℓ_t be the importance sampling weighted loss defined as $\ell_t(f) := (g_{\text{ref}}/g_t)\ell(f)$. Let $(\widehat{f}_{1,t})_{t\geq 1}, \ldots, (\widehat{f}_{1,J})_{t\geq 1}$ be J sequences of random functions such that for every $j \in [J]$, $t \geq 1$, $\widehat{f}_{j,t}$ is $\bar{O}(t-1)$ measurable. Let f_0 be a fixed function.

For a fixed function f, we define average conditional risk, and its IS-weighted empirical risk as

$$\begin{aligned} R_T(q, f) :=& \frac{1}{T}\sum_{t=1}^{T} E_{q_1, g_{\text{ref}}, q_2}\left[\ell(f)(O(t)) \mid \bar{O}(t-1)\right] \\ \text{and } \widehat{R}_T(f) :=& \frac{1}{T}\sum_{t=1}^{T} \ell_t(f)(O(t)). \end{aligned}$$

and for any $j \in [J]$, we define the average conditional risk and the IS-weighted risk of the sequence $(\widehat{f}_{j,t})$ as

$$\begin{aligned} R_{j,T}(q) :=& \frac{1}{T}\sum_{t=1}^{T} E_{q_1, g_{\text{ref}}, q_2}\left[\ell(\widehat{f}_{j,t})(O(t)) \mid \bar{O}(t-1)\right], \\ \text{and } \widehat{R}_{j,T} :=& \frac{1}{T}\sum_{t=1}^{T} \ell_t(\widehat{f}_{j,t})(O(t)). \end{aligned}$$

Let

$$\widetilde{j}_T := \arg\min_{j\in[J]} R_{j,T}(q),$$

be the oracle selector and let

$$\widehat{j}_T := \arg\min_{j\in[J]} \widehat{R}_{j,T}$$

be the Super Learner selector.

Although our results hold in full generality, we will think of these sequences of random functions as sequences of estimators of a feature f_0 of the uncontrolled component q_0 of the data-generating density.

We make the following assumptions. Unlike in the analysis of the empirical risk minimizers, instead of making assumptions on the unweighted losses ℓ and on the importance sampling ratios g_{ref}/g_t, we directly make assumptions on the IS-weighted losses ℓ_t. This covers the former case and allows for more generality.

Assumption 5.9. *The fixed function f_0 is such that, for any $j \in [J]$, $t \geq 1$,*

$$R_{j,T}(q) - R_T(q, f_0) \leq 0.$$

Let us give an example of a common situation where assumption 5.9 holds. Observe that in the CB setting, we have that $R_T(q,f) = E_{q_1,g_{\mathrm{ref}},q_2}\left[\ell(f)(O(1))\right]$, and as a result $R_T(q,f)$ is non-random. In this case, if f_0 is a minimizer over the set of measurable functions of $R_T(q,f)$, f_0 is a non-random function that satisfies assumption 5.9.

Assumption 5.10 (Supremum norm bound)**.** *There exists $M_1 > 0$ such that, for all $t \geq 1$ $\sup_f \|\ell_t(f) - \ell_t(f_0)\|_\infty \leq M_1$.*

Assumption 5.11 (Variance bound)**.** *There exists $\alpha > 0$, $M_2 > 0$, such that, for all $t \geq 1$, and any $j \in [J]$,*

$$\begin{aligned} &E\left[\left(\ell_t(\widehat{f}_{j,t})(O(t)) - \ell_t(f_0)(O(t))\right)^2 \middle| \bar{O}(t-1)\right] \\ \leq &M_2 \left(E\left[\ell_t(\widehat{f}_{j,t})(O(t)) - \ell_t(f_0)(O(t)) \middle| \bar{O}(t-1)\right]\right)^\alpha. \end{aligned}$$

Assumption 5.12 (Global variance bound)**.** *There exists $M_3 > 0$ such that, for all $j \in [J]$ and $t \geq 1$,*

$$E\left[\left(\ell_t(\widehat{f}_{j,t})(O(t)) - \ell_t(f_0)(O(t))\right)^2 \middle| \bar{O}(t-1)\right] \leq M_3.$$

We can now state our oracle inequality result.

Theorem 5.4 (High probability oracle inequality)**.** *Consider the setting of the current section, and suppose that assumptions 5.10-5.12 hold. Consider an integer $N \geq 1$ and a number $a > 0$, and let $\underline{x}(a,N,\alpha) := a(2^{-(N+1)}M_3/M_2)^{1/\alpha}$. Then, for all $x > \underline{x}(a,N,\alpha)$, it holds that*

$$\begin{aligned} &\mathrm{Pr}_q\left[R_{\widehat{j}_T,T}(q) - R_T(q,f_0) \geq (1+2a)\left(R_{\widetilde{j}_T,T}(q) - R_T(q,f_0)\right) + x\right] \\ \leq &2J(N+1)\left(\exp\left(-\frac{tx^{2-\alpha}}{C_1(M_2,a)}\right) + \exp\left(-\frac{tx}{C_2(M_1,a)}\right)\right), \end{aligned}$$

with $C_1(M_2,a) := 8 \times 2^{2-\alpha}(1+a)^2 M_2/a^\alpha$, and $C_2(M_1,a) := 16(1+a)M_1/3$.

As corollary of theorem 5.4, we can obtain the following oracle inequality in expectation.

Corollary 5.1 (Oracle inequality for the expected risk)**.** *Suppose that assumptions 5.10-5.12 hold, and let $a > 0$. Then*

$$\begin{aligned} &E\left[R_{\widehat{j}_T,T}(q) - R_T(q,f_0) - (1+2a)\left(R_{\widetilde{j}_T,T}(q) - R_T(q,f_0)\right)\right] \\ \leq &\left(\frac{C_1(M_2,a)}{t}\right)^{\frac{1}{2-\alpha}} (1 + 4\log J(N+1))^{\frac{1}{2-\alpha}} + 2\frac{C_2(M_1,a)}{t}\left(1 + \log(J(N+1))\right), \end{aligned}$$

where

$$C_1(M_2,a) = 8(1+a)^2 M_2/a^\alpha \qquad \textit{and} \qquad C_2(M_1,a) = 8(1+a)M_1/3,$$

$$N + 1 = \log\left(a^\alpha \frac{M_3}{M_2}\left(\frac{t}{C_1}\right)^{\frac{\alpha}{2-\alpha}}\right) / \log 2.$$

Bibliography

P. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

D. Benkeser, C. Ju, S. Lendle, and M. J. van der Laan. Online cross-validation-based ensemble learning. *Statistics in Medicine*, 37(2):249–260, 2018.

Antoine Chambaz, Wenjing Zheng, and Mark J. van der Laan. Targeted sequential design for targeted learning inference of the optimal treatment rule and its mean reward. *Ann. Statist.*, 45 (6):2537–2564, 12 2017.

S. Dudoit and M. J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154, 2005.

Ivana Malenica, Aurelien Bibaut, and Mark J. van der Laan. Adaptive sequential design for a single time-series, 2021.

I. Steinwart and C. Scovel. Fast rates for support vector machines using gaussian kernels. *Ann. Statist.*, 35(2):575–607, 04 2007.

Y. Zhao, D. Zeng, Rush J. A., and Kosorok M. R. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

5.A Conclusion

The contributions we make in this chapter are two-fold. First, we have given guarantees for empirical risk minimizers fitted on contextual bandit data, and on MDP data. In both situations we gave high probability bounds under bracketing entropy conditions, and we required a certain rate of uniform exploration — that is we assumed that actions were sampled from an ε-greedy design. Our second contribution is a new high probability oracle inequality for the Super Learner for sequentially collected data.

These results are important in the sense that they allow for nuisance parameter fitting in causal estimators and off-policy policy learning from sequentially collected data. We use variants of these results in chapter 7, and in Malenica et al. [2021].

In chapter 4, we give study policy learning under a more efficient (from the point of view of regret minimization) design than ε-greedy. It would be an interesting direction for future work to generalize the high probability excess risk bounds to that design and to more general designs, under minimal assumptions.

5.B Proof of theorem 5.1

The proof of theorem 1 relies on a maximal inequality for importance-sampling weighted martingale processes, which we present next. Consider a class of functions $\mathcal{F} : \mathcal{O} \to \mathbb{R}$. Let, for every $f \in \mathcal{F}$,

$$\begin{aligned} M_T(f) :=& \frac{1}{T} \sum_{t=1}^{T} E_{q_1, g_{\text{ref}}, q_2} \left[f(O(t)) \mid O(t) \right] - \frac{g_{\text{ref}}}{g_t}(A(t) \mid L_1(t)) f(O(t)) \\ =& \frac{1}{T} \sum_{t=1}^{T} E_{q_1, g_t, q_2} \left[\frac{g_{\text{ref}}}{g_t}(A(t) \mid L_1(t)) f(O(t)) \mid O(t) \right] - \frac{g_{\text{ref}}}{g_t}(A(t) \mid L_1(t)) f(O(t)). \end{aligned}$$

Observe that the terms in the sum in $M_T(f)$ form a martingale difference sequence. The process $\{M_T(f) : f \in \mathcal{F}\}$ is a so-called martingale process. We refer to the particular type of processes of the form of $\{M_T(f) : f \in \mathcal{F}\}$ as *importance sampling weighted* martingale processes.

Our maximal inequality relies on the following assumptions.

Assumption 5.13 (Entropy)**.** *There exists* $p > 0$ *such that*

$$\log N_{[]}(\epsilon, \mathcal{F}, \sigma_T) \lesssim \epsilon^{-p}.$$

Assumption 5.14 (Radius of $\mathcal{F}$)**.** *There exists* $r_0 > 0$ *and* $B > 0$ *such that*

$$\sup_{f \in \mathcal{F}} \sigma_T(f) \leq r_0 \qquad \textit{and} \qquad \sup_{f \in \mathcal{F}} \|f\|_\infty \leq B.$$

Theorem 5.5 (Maximal inequality for IS-weighted martingale processes)**.** *Make assumptions 5.13 and 5.14. It holds with probability at least* $1 - 2e^{-x}$ *that, for any* $r \in [0, r_0/2]$*, that*

$$\sup_{f \in \mathcal{F}} M_T(f) \lesssim r_- + \mathcal{H}_T(\delta, r_0, r_-, B) + r_0 \sqrt{\frac{x}{\delta T}} + \frac{Bx}{\delta T},$$

where

$$\mathcal{H}_T(\delta, r_0, r_-, B) := \frac{1}{\delta T} \int_{r_-}^{r_0} \sqrt{\log(1 + N_{[]}(\epsilon, \mathcal{F}, \sigma_T)} d\epsilon + \frac{B}{\delta T} \log(1 + N_{[]}(r_0, \mathcal{F}, \sigma_T)$$

The proof is almost identical to that of theorem 4.5 in chapter 4.

We can now prove theorem 5.1. We distinguish the cases $p \in (0, 2)$ and $p > 2$. The proof in the case $p \in (0, 2)$ is a straightforward adaptation of lemma 13 in Bartlett et al. [2006].

Proof of theorem 1, case $p \in (0, 2)$. We proceed in three steps. In the first step, we identify a set of upper bounds on $R_T(\widehat{f}_T) - R_T(f_1)$ that hold with probability at least $1 - 2e^{-x}$. In the second step, we give a sufficient condition for the condition that defines the set of upper bounds. In the third step, we give an explicit upper bound.

Step 1: A set of high probability upper bounds. We will use the shorthand notation $R_T(f)$ and $\widehat{R}_T(f)$ for $R_T(q, f)$ and $\widehat{R}_T(q, f)$, and observe that $M_T(f) = R_T(f) - \widehat{R}_T(f)$. Let $r > 0$. From the convexity assumption (assumption 5.5) we have the following implication:

$$\begin{aligned} &\exists f \in \mathcal{F}, \widehat{R}_T(f) - \widehat{R}_T(f_1) \leq 0 \text{ and } R_T(f) - R_T(f_1) \geq r^2 \\ \Longrightarrow\ &\exists f \in \mathcal{F}, \widehat{R}_T(f) - \widehat{R}_T(f_1) \leq 0 \text{ and } R_T(f) - R_T(f_1) = r^2. \end{aligned}$$

Therefore,

$$\begin{aligned} &\Pr_q\left[\widehat{R}_T(\widehat{f}_T) - \widehat{R}_T(f_1) \text{ and } R_T(\widehat{f}_T) - R_T(f_1) \geq r^2\right] \\ \leq &\Pr_q\left[\widehat{R}_T(\widehat{f}_T) - \widehat{R}_T(f_1) \text{ and } R_T(\widehat{f}_T) - R_T(f_1) = r^2\right] \\ \leq &\Pr_q\left[\sup_{\substack{f \in \mathcal{F} \\ R_T(f) - R_T(f_1) \leq r^2}} M_T(\ell(f) - \ell(f_1) \geq r^2\right] \\ \leq &\Pr_q\left[\sup_{\substack{f \in \mathcal{F} \\ \sigma_T(\ell(f) - \ell(f_1)) \leq r^\alpha}} M_T(\ell(f) - \ell(f_1) \geq r^2\right], \end{aligned}$$

where the last line follows from the variance bound assumption (assumption 5.2).

Let r be such that

$$r^2 \gtrsim \mathcal{H}_T(\delta, r^\alpha, 0, B) + r^\alpha \sqrt{\frac{x}{BT}} + \frac{Bx}{\delta T}. \tag{5.1}$$

Then, from theorem 5.5

$$\Pr_q\left[\sup_{\substack{f \in \mathcal{F} \\ \sigma_T(\ell(f) - \ell(f_1) \leq r^\alpha}} M_T(\ell(f) - \ell(f_1) \geq r^2\right] \leq 2e^{-x},$$

and thus

$$R_T(\widehat{f}) - R_T(f_1) \lesssim r^2$$

with probability at least $1 - 2e^{-x}$.

Step 2: a sufficient condition to be an upper bound. It is straightforward to check that since $p < 2$,

$$r \mapsto \frac{\mathcal{H}_T(\delta, r, 0, B)}{r}$$

is a decreasing function, and this this implies that if r^* is a solution to $(r^*)^2 = \mathcal{H}_T(\delta, r^*, 0, B)$, then, if $r \geq r^*$, then $r^2 \geq \mathcal{H}_T(\delta, r, 0, B)$. As a result, if

$$r^2 \geq \max\left\{(r^*)^2, \left(\frac{x}{\delta T}\right)^{\frac{1}{2-\alpha}}, \frac{Bx}{\delta T}\right\},$$

then r satisfies condition (5.1).

Step 3: an explicit upper bound. Treating B as a constant that we absorb in the "$\lesssim$" symbol, we have that

$$\mathcal{H}(\delta, r, 0, B) \lesssim \frac{r^{1-p/2}}{\sqrt{\delta T}} + \frac{r^{-p}}{\delta T}.$$

It is straightforward to chcek that

$$r^* = (\delta T)^{-\frac{1}{2-\alpha+p\alpha/2}}$$

is such that

$$(r^*)^2 \asymp \mathcal{H}_T(\delta, r^*, 0, B).$$

□

Proof of theorem 5.1, case $p > 2$. Observe that since $\widehat{R}_T(\widehat{f}_T) \leq \widehat{R}_T(f_1)$,

$$\begin{aligned} R_T(q, \widehat{f}) - R_T(q, f_1) \leq & M_T(\ell(\widehat{f}) - \ell(f_1)) \\ \leq & \sup_{f \in \mathcal{F}} M_T(\ell(f) - \ell(f_1)). \end{aligned}$$

The result then follows by applying the maximal inequality 5.5 and optimizing the lower bound r_- of the entropy integral. □

5.C Proofs of the Super Learner results

Proof of theorem 5.4. The proof can be decomposed in three steps.

Step 1: an algrebraic decomposition. Denote $\widetilde{H}_{j,T} := R_{j,T}(q) - R_T(q, f_0)$ and $\widehat{H}_{j,T} := \widehat{R}_{j,T} - \widehat{R}_T(f_0)$, where, for all f, $\widehat{R}_t(f) := T^{-1} \sum_{t=1}^T \ell_t(\theta)(O(t))$. An algebraic decomposition proven in past works on the Super Learner (see e.g. Dudoit and van der Laan [2005], Benkeser et al. [2018]) shows, that for any $a > 0$, the (average cumulative) excess risk of the Super Learner's choice can be bounded by $(1 + 2a)$ times the (average cumulative) excess risk of the oracle choice, plus some remainder terms, as follows:

$$\begin{aligned} R_{\widehat{j}_T,T}(q) - R_T(q, f_0) \leq & (1 + 2a) \left(R_{\widetilde{j}_T,T}(q) - R_T(q, f_0) \right) + A_{\widehat{j}_T,T}(a) + B_{\widetilde{j}_T,T}(a) \\ \leq & (1 + 2a) \left(R_{\widetilde{j}_T,T}(q) - R_T(q, f_0) \right) + \max_{j \in [J]} A_{j,T}(a) + \max_{j \in [J]} B_{j,T}(a) \end{aligned} \quad (5.2)$$

where

$$\begin{aligned} & A_{j,T}(a) := (1 + a) \left(\widetilde{H}_{j,T} - \widehat{H}_{j,T} \right) - a \widetilde{H}_{j,T} \\ \text{and} \qquad & B_{j,T}(a) := (1 + a) \left(\widehat{H}_{j,T} - \widetilde{H}_{j,T} \right) - a \widetilde{H}_{j,T}. \end{aligned}$$

The first terms in both $A_{j,T}(a)$ and $B_{j,T}$ is a centered mean of a martingale difference sequence (MDS). From Bernstein's inequality we expect that their probability of being larger than x to behave, for small $x > 0$, like $\exp(-Ctx^2)$ for some constant C. Observe that by definition of f_0, $\widetilde{H}_{j,T} \leq 0$ for every j. As we will see in the next step, the negative shifting by $a\widetilde{H}_{j,T}$ in $A_{j,T}(a)$ and $B_{j,T}(a)$ allows to get bounds tighter than $\exp(-Ctx^2)$ for $P[A_{j,T}(a) \geq x]$ and $P[B_{j,T}(a) \geq x]$.

Step 2: Bounding positive deviations of $A_{j,T}(a)$ and $B_{j,T}(a)$. The analysis of $A_{j,T}(a)$ and $B_{j,T}(a)$ is identical. We present it only for $A_{j,T}(a)$. Denote $U_{j,t} := \ell_t(\widehat{f}_{j,t})(O(t)) - \ell_t(f_0)(O(t))$. Observe that $\widehat{H}_{j,T} = t^{-1}\sum_{t=1}^T U_{j,t}$ and that $\widetilde{H}_{j,T} := t^{-1}\sum_{t=1}^T E[U_{j,t} \mid \bar{O}(t-1)]$. Introduce $V_{j,T} := t^{-1}\sum_{t=1}^T E[U_{j,t}^2 \mid \bar{O}(t-1)]$ the mean of conditional second moments of $U_{j,1}, \ldots, U_{j,T}$.

The general idea of this step is to use Bernstein's inequality for martingales to bound $P[A_{j,T}(a) \geq x]$. The first step is to show that $\{A_{j,T}(a) \geq x\}$ implies that the martingale $\widetilde{H}_{j,T} - \widehat{H}_{j,T}$ is greater than some quantity. We indeed have that $A_{j,T}(a) \geq x$ is equivalent to $\widetilde{H}_{j,T} - \widehat{H}_{j,T} \geq (1+a)^{-1}(x + a\widetilde{H}_{j,T})$. From assumption 5.11, $\widetilde{H}_{j,T} \geq (V_{j,T}/M_2)^{1/\alpha}$. Therefore, $A_{j,T}(a) \geq x$ implies that $\widetilde{H}_{j,T} - \widehat{H}_{j,T} \geq (1+a)^{-1}(x + a(V_{j,T}/M_2)^{1/\alpha})$. Bernstein's inequality for martingales for $\widetilde{H}_{j,T} - \widehat{H}_{j,T}$ reads, for any fixed $y > 0$, as

$$P\left[\widetilde{H}_{j,T} - \widehat{H}_{j,T} \geq y, V_{j,T} \leq v\right] \leq \exp\left(-\frac{1}{2}\frac{ty^2}{v + \frac{2}{3}M_1 y}\right).$$

In order to be able to apply it to bound $P[A_{j,T}(a) \geq x]$, we thus need a fixed lower bound on $(1+a)^{-1}(x + a(V_{j,T}/M_2)^{1/\alpha})$, that is we need a lower bound on $V_{j,T}$, while in the meantime the above form of Bernstein's inequality also requires an upper bound on $V_{j,T}$. From assumption 5.12, $V_{j,T} \leq M_3$. A trivial lower bound is 0. However, if we use this trivial lower bound, we are essentially forgetting about and loosing the benefit of the negative shifting by $a\widetilde{H}_{j,T}$. We can do better if we can ensure that the lower bound is within a constant factor of the upper bound. This motivates using a peeling device: we consider a dyadic partition $\sqcup_{i=0}^N (v_{i-1}^N, v_i^N]$ of $[0, M_3]$, and we exploit the fact that, from a union bound, $P[A_{j,T}(a) \geq x] \leq \sum_{i=0}^N P[A_{j,T}(a) \geq x, V_{j,T} \in [v_{i-1}^N, v_i^N]]$. This leads to dealing with more terms than if we were not using the peeling device, but we will show that, for an appropriate choice of the size of the partition N, these are sufficiently small so that it ends up being beneficial. Let us now give the precise definition of the dyadic partition: for all $i \in \{0, \ldots, N\}$, we let $v_i^N := 2^{i-N}M_3$, and we set $v_{-1}^N = 0$ by convention. We then have that

$$\begin{aligned}
P[A_{j,T}(a) \geq x] \leq & P\left[\widetilde{H}_{j,T} - \widehat{H}_{j,T} \geq \frac{1}{1+a}\left(x + a(V_{j,T}/M_2)^{1/\alpha}\right)\right] \\
= & \sum_{i=0}^N P\left[\widetilde{H}_{j,T} - \widehat{H}_{j,T} \geq \frac{1}{1+a}\left(x + a(V_{j,T}/M_2)^{1/\alpha}\right), V_{j,T} \in [v_{i-1}^N, v_i^N]\right] \\
\leq & \sum_{i=0}^N P\left[\widetilde{H}_{j,T} - \widehat{H}_{j,T} \geq \frac{1}{1+a}\left(x + a(v_{i-1}^N/M_2)^{1/\alpha}\right), V_{j,T} \leq v_i^N\right]
\end{aligned}$$

$$\leq \sum_{i=0}^{N} \exp\left(-\frac{1}{2}\frac{t}{(1+a)^2} D_i(x)\right)$$

with

$$D_i(x) = \frac{\left(x + a(v_{i-1}/M_2)^{1/\alpha}\right)^2}{v_i + \frac{2}{3}\frac{M_1}{1+a}\left(x + a(v_{i-1}/M_2)^{1/\alpha})\right)},$$

where we have dropped the N exponent in the v_i's so as to lighten notation. We now derive lower bounds on the quantity $D_i(x)$. We distinguish two regimes in x, depending on whether x is small or large. Specifically, for $x \leq \widetilde{x}_i := v_i 3(1+a)/(2M_1) - a(v_{i-1}/M_2)^{1/\alpha}$, we have $v_i \geq (2M_1/(3(1+a))(x + a(v_{i-1}/M_2)^{1/\alpha})$ and thus

$$D_i(x) \geq \frac{\left(x + a(v_{i-1}/M_2)^{1/\alpha}\right)^2}{2v_i} = \frac{\left(x + a(v_{i-1}/M_2)^{1/\alpha}\right)^{2-\alpha}}{\frac{2v_i}{\left(x+a(v_{i-1}/M_2)^{1/\alpha}\right)^{\alpha}}} \geq \frac{x^{2-\alpha}}{\frac{2v_i}{\left(x+a(v_{i-1}/M_2)^{1/\alpha}\right)^{\alpha}}}.$$

Suppose that $i \geq 1$. As $x \geq 0$, the denominator in the right hand side of the above expression is at least as large as $2M_2 v_i/(v_{i-1}a^\alpha)$, and, since $v_i/v_{i-1} = 2$, we thus have

$$D_i(x) \geq \frac{x^{2-\alpha}}{\frac{4M_2}{a^\alpha}}.$$

Now consider the case $i = 0$. For the same lower bound to hold, we need

$$\frac{2v_0}{\left(x + a(v_{-1}/M_2)^{1/\alpha}\right)^{\alpha}} \leq \frac{4M_2}{a^\alpha} \iff \frac{2v_0}{x^\alpha} \leq \frac{4M_2}{a^\alpha} \iff x \geq a\left(2^{-(N+1)}\frac{M_3}{M_2}\right)^{\frac{1}{\alpha}} := \underline{\mathrm{x}}(a, N, \alpha).$$

We now consider the case $x \geq \widetilde{x}_i$. Then

$$D_i(x) \geq \frac{\left(x + a(v_{i-1}/M_2)^{1/\alpha}\right)^2}{\frac{4}{3}\frac{M_1}{1+a}\left(x + a(v_{i-1}/M_2)^{1/\alpha}\right)} \geq \frac{x}{\frac{4}{3}\frac{M_1}{1+a}}.$$

Therefore, denoting $\widetilde{C}_1(M_2, a) := 8(1+a)^2 M_2/a^\alpha$ and $\widetilde{C}_2(M_1, a) := 8(1+a)M_1/3$, we have that, for $x \geq \underline{\mathrm{x}}(a, N, \alpha)$,

$$\begin{aligned} P[A_{j,T}(a) \geq x] \leq& \sum_{i=0}^{N} \mathbf{1}\{x \leq \widetilde{x}_i\} \exp\left(-\frac{tx^{2-\alpha}}{\widetilde{C}_1(M_2,a)}\right) + \mathbf{1}\{x > \widetilde{x}_i\} \exp\left(-\frac{tx}{\widetilde{C}_2(M_1,a)}\right) \\ \leq& (N+1)\exp\left(-\frac{tx^{2-\alpha}}{\widetilde{C}_1(M_2,a)}\right) + (N+1)\exp\left(-\frac{tx}{\widetilde{C}_2(M_1,a)}\right). \end{aligned}$$

Step 3: end of the proof. From bound (5.2) from the first step and a union bound, we have that

$$
\begin{aligned}
&P\left[R_{\widehat{j}_T,T}(q) - R_T(q,f_0) \geq (1+2a)\left(R_{\widetilde{j}_T,T}(q) - R_T(q,f_0)\right) + x\right] \\
\leq &P\left[\max_{j\in[J]} A_{j,T}(a) + \max_{j\in[J]} B_{j,T}(a) \geq x\right] \\
\leq &\sum_{j=1}^{J} P\left[A_{j,T}(a) \geq \frac{x}{2}\right] + P\left[B_{j,T}(a) \geq \frac{x}{2}\right] \\
\leq &2J(N+1)\exp\left(-\frac{tx^{2-\alpha}}{C_1(M_2,a)}\right) + 2J(N+1)\exp\left(-\frac{tx}{C_2(M_1,a)}\right).
\end{aligned}
$$

□

We now present the proof of the oracle inequality in expectation (corollary 5.1).

Proof of corollary 5.1. Observe that

$$
\begin{aligned}
&E\left[R_{\widehat{j}_T,T}(q) - R_T(q,f_0) - (1+2a)\left(R_{\widetilde{j}_T,T}(q) - R_T(q,f_0)\right)\right] \leq \\
&\int_0^\infty P\left[R_{\widehat{j}_T,T}(q) - R_T(q,f_0) - (1+2a)\left(R_{\widetilde{j}_T,T}(q) - R_T(q,f_0)\right) \geq x\right] dx
\end{aligned}
$$

In what follows, we use C_1 for $C_1(M_2,a)$ and C_2 for $C_2(M_1,a)$. From theorem 5.4,

$$
\begin{aligned}
&\int_0^\infty P\left[R_{\widehat{j}_T,T}(q) - R_T(q,f_0) - (1+2a)\left(R_{\widetilde{j}_T,T}(q) - R_T(q,f_0)\right) \geq x\right] dx \\
=&\int_0^{\underline{x}(N,a)} P\left[R_{\widehat{j}_T,T}(q) - R_T(q,f_0) - (1+2a)\left(R_{\widetilde{j}_T,T}(q) - R_T(q,f_0)\right) \geq x\right] dx \\
&+\int_{\underline{x}(N,a)}^\infty P\left[R_{\widehat{j}_T,T}(q) - R_T(q,f_0) - (1+2a)\left(R_{\widetilde{j}_T,T}(q) - R_T(q,f_0)\right) \geq x\right] dx \\
\leq&\underline{x}(N,a) + 2\int_{\underline{x}(N,a)}^\infty \min\left(J(N+1)\exp\left(\frac{-tx^{2-\alpha}}{C_1}\right), 1\right) dx \\
&\qquad + 2\int_{\underline{x}(N,a)}^\infty \min\left(J(N+1)\exp\left(\frac{-tx}{C_2}\right), 1\right) dx \\
\leq&\underline{x}(N,a) + \underbrace{2\int_0^\infty \min\left(J(N+1)\exp\left(\frac{-tx^{2-\alpha}}{C_1}\right), 1\right) dx}_{I} \\
&\qquad + \underbrace{2\int_0^\infty \min\left(J(N+1)\exp\left(\frac{-tx}{C_2}\right), 1\right) dx}_{II}
\end{aligned}
$$

We choose N such that $\underline{x}(N,a) = (C_1/t)^{1/(2-\alpha)}$, that is we set N such that

$$N+1 = \log\left(a^{\alpha}\frac{M_3}{M_2}\left(\frac{t}{C_1}\right)^{\frac{\alpha}{2-\alpha}}\right)/\log 2.$$

We now turn to the terms I and II.

We start with I. Let x_1 such that $J(N+1)\exp(-tx_1^{2-\alpha}/C_1) = 1$, that is

$$x_1 = \left(\frac{C_1\log(J(N+1))}{t}\right)^{\frac{1}{2-\alpha}}.$$

We have that

$$\begin{aligned}
&\int_0^{\infty}\min\left(J(N+1)\exp\left(-\frac{tx^{2-\alpha}}{C_1}\right),1\right)dx\\
=&x_1 + J(N+1)\int_{x_1}^{\infty}\exp\left(-\frac{tx^{2-\alpha}}{C_1}\right)dx\\
=&x_1 + J(N+1)\left(\frac{C_1}{t}\right)^{\frac{1}{2-\alpha}}\int_{\log J(N+1)}^{\infty}\exp(-u)u^{\frac{1}{2-\alpha}-1}du\\
\leq&\left(\frac{C_1\log J(N+1)}{t}\right)^{\frac{1}{2-\alpha}}\left(1+\frac{1}{\log J(N+1)}\right)\\
\leq&2\left(\frac{C_1\log J(N+1)}{t}\right)^{\frac{1}{2-\alpha}}.
\end{aligned}$$

The third line in the above display follows from the change of variable $u = tx^{2-\alpha}/C_1$. The fourth line follows from the fact that $u \mapsto u^{1/(2-\alpha)-1}$ is non-increasing, as $\alpha \in [0,1]$. The fourth line uses that $\log J(N+1) \geq 1$.

We now turn to II. Let x_2 such that $J(N+1)\exp(-tx_2/C_2) = 1$, that is $x_2 = \frac{C_2\log J(N+1)}{t}$. We have that

$$\begin{aligned}
&\int_0^{\infty}\min\left(J(N+1)\exp\left(-\frac{tx}{C_2}\right),1\right)dx\\
=&x_1 + J(N+1)\int_{x_1}^{\infty}\exp\left(-\frac{tx}{C_2}\right)dx\\
\leq&x_1 + J(N+1)\frac{C_2}{t}\exp\left(-\frac{tx_1}{C_1}\right)\\
\leq&\frac{C_2}{t}\left(1+\log J(N+1)\right)
\end{aligned}$$

Therefore,

$$E\left[R_{\widehat{\jmath}_T,T}(q) - R_T(q,f_0) - (1+2a)\left(R_{\widetilde{\jmath}_T,T}(q) - R_T(q,f_0)\right)\right]$$

$$\leq \left(\frac{C_1}{t}\right)^{\frac{1}{2-\alpha}} (1 + 4 \log J(N+1))^{\frac{1}{2-\alpha}} + 2\frac{C_2}{t} (1 + \log(J(N+1))) .$$

□

5.D Proof of theorem 5.2

The proof of theorem 5.2 relies on the following two technical lemmas.

Lemma 5.4. *Suppose that*

$$\Pr\left[X_T \geq C_0(\delta T)^{-\widetilde{\beta}} + \widetilde{C}_1 \left(\frac{x}{\delta T}\right)^{\frac{1}{2-\nu}} + \widetilde{C}_2 \frac{x}{\delta T}\right] \leq 2e^{-x}.$$

Then,

$$\Pr\left[X_T \geq C_0(\delta T)^{-\widetilde{\beta}} + x\right] \leq 2e^{-\widetilde{C}_3(x^{\widetilde{\gamma}} \wedge x)},$$

with $\widetilde{C}_3 := (\widetilde{C}_1 + \widetilde{C}_2)^{-\widetilde{\gamma}} \wedge (\widetilde{C}_1 + \widetilde{C}_2)^{-1}$ *and* $\widetilde{\gamma} = 2 - \nu$*.*

The following lemma allows to obtain a high probability bound on the excess risk of a policy $g_{\widehat{f}_T}$ from a high probability bound on the excess ϕ-risk of $\widehat{f}_T$.

Lemma 5.5. *Suppose that* $R(q, g_{\widehat{f}_T}) - R(q, g_1) \leq \widetilde{c}(R^\phi(q, \widehat{f}_T) - R^\phi(f_1))^{1/(2-\nu)}$ *for some* g_1*,* f_1 *and* $\nu \in [0, 1]$*, and that*

$$\Pr\left[R^\phi(\widehat{f}_T) - R^\phi(f_1) \geq \widetilde{C}_0(\delta T)^{\widetilde{\beta}} + x\right] \leq 2 \exp\left(-\widetilde{C}_2 \delta T(x^{\widetilde{\gamma}} \wedge x)\right),$$

for $\widetilde{\beta} > 0$*,* $\widetilde{\gamma} \geq 1$*. Then*

$$\Pr\left[R(q, g_{\widehat{f}_T}) - R(q, g_1) \geq C_0(\delta T)^{-\beta} + x\right] \leq 2 \exp\left(-C_2 \delta T(x^\gamma \vee x)\right),$$

with $C_0 := \widetilde{c}\widetilde{C}_0^{1/(2-\nu)}$*,* $\gamma := (2-\nu)\widetilde{\gamma}$*,* $\widetilde{\beta} := \beta/(2-\nu)$*, and* $C_2 := \widetilde{C}_2(\widetilde{c}^{-\gamma} \wedge \widetilde{c}^{-1})$*.*

We can now present the proof of theorem 5.2.

Proof. We treat separately the different settings presented in table 5.1. Whenever we refer to the "constants of the problem" in this proof, we mean B and the constants absorbed in the "$\lesssim''$ and " $\gtrsim''$ symbols in the statements of the Tsybakov noise assumption, calibration bounds, the entropy assumption, and of the modulus of convexity bounds.

Direct policy optimization case. In this case $g_{\widehat{f}_T} = \widehat{f}_T$ and $R(q, g_{\widehat{f}_T}) - \inf_{g \in \mathcal{F}} R(q, g_f) = R^\phi(q, \widehat{f}_t) - \inf_{f \in \mathcal{F}} R^\phi(q, f)$. We distinguish two cases depending on whether or not the realizability assumption (assumption 5.6) holds. Let $f^* \in \arg\min_{f \in \mathcal{F}} R(q, f)$.

Without realizability. Using the above identity corollary **??** and lemma 5.4 gives that

$$\Pr\left[R(q, g_{\widehat{f}_T}) - R(q, g_{f^*}) \geq C_0(\delta T)^{-\beta} + x\right] \leq 2e^{-C_2(x^2 \wedge x)},$$

where $\beta = 1/2$ for $p \in (0, 2)$ and $\beta = 1/p$ for $p > 2$, and C_0 and C_2 depend on the constants of the problem.

With realizability. Under realizability, lemma 5.2 gives us that, for all $f \in \mathcal{F}$, $\|\ell(f) - \ell(f^*)\|_{q_1, g_{\mathrm{ref}}, q_2} \lesssim (R(q, f) - R(q, f^*))^\alpha$ with $\alpha = \nu/(\nu + 1)$. Then applying theorem 5.1 and lemma 5.4 yields

$$\Pr\left[R(q, \widehat{f}_T) - R(q, f^*) \geq C_0(\delta T)^{-\beta} + x\right] \leq 2e^{-C_2(x^{2-\alpha} \wedge x)},$$

with $\beta = 1/(2 - \alpha + p\alpha/2)$ if $p \in (0, 2)$ and $\beta = 1/p$ if $p > 2$.

Hinge surrogate case, under realizability. The theory of classification calibration relates the excess risk to the excess ϕ-risk, where these excess quantities are defined relative to the optimal measurable policy and the optimal measurable function $\mathcal{A} \times \mathcal{L}_1 \to \mathbb{R}$ such that $\sum_{a=1}^K f(a, l_1) = 0$ for all $l_1 \in \mathcal{L}_1$. We do not know of equivalent results w.r.t. minimizers in smaller classes. We thus only present results under the realizability assumption. We first start by characterizing the ϕ-risk. Lemma 5.2 gives us that, for all $f \in \mathcal{F}$, $\|\ell(f) - \ell(f^*)\|_{q_1, g_{\mathrm{ref}}, q_2} \lesssim (R^\phi(q, f) - R^\phi(q, f^*))^\alpha$. Applying theorem 5.1 and lemma 5.5 then yields

$$\Pr\left[R^\phi(q, \widehat{f}_t) - R^\phi(f^*) \geq C_0(\delta T)^{-\beta} + x\right] \leq 2e^{-C_2(x^{2-\alpha} \wedge x)},$$

for $C_0, C_2 > 0$ depending on the constants of the problem, and with $\beta = 1/(2 - \alpha + p\alpha/2)$ for $p \in (0, 2)$ and $\beta = 1/p$ for $p > 2$. The calibration bound gives that $R(q, g_{\widehat{f}_T}) - R(q, g_{f^*}) \leq R^\phi(q, \widehat{f}_t) - R^\phi(q, f^*)$, and therefore, the above bound also holds for the excess risk:

$$\Pr\left[R(q, g_{\widehat{f}_T}) - R(q, g_{f^*}) \geq C_0(\delta T)^{-\beta} + x\right] \leq 2e^{-C_2(x^{2-\alpha} \wedge x)}.$$

Quadratic surrogate case with realizability. Similarly to the previous case, as we rely on calibration results, we only present guarantees under the realizability assumption. The variance bound is obtained by using that ϕ^{tquad} has quadratic modulus of convexity. Lemma 5.1 then yields that $\|\ell(f) - \ell(f^*)\|_{q_1, g_{\mathrm{ref}}, q_2} \lesssim R^\phi(q, f) - R^\phi(q, f^*)$ for all $f \in \mathcal{F}$. Theorem 5.1 (applied with $\alpha = 1$) and lemma 5.4 then give that

$$P\left[R^\phi(\widehat{f}_T) - R^\phi(f^*) \geq \widetilde{C_0}(\delta T)^{-\widetilde{\beta}} + x\right] \leq 2e^{-\widetilde{C}_2 \delta T x},$$

with $\widetilde{\beta} = 1/(1 + p/2)$ if $p \in (0, 2)$ and $\beta = 1/p$ if $p > 2$, and where $\widetilde{C}_0$, and $\widetilde{C}_2$ depend on the constants of the problem.

We now relate the excess risk to the excess ϕ-risk with calibration results. Using the calibration bound under the Tsybakov noise assumption and then applying lemma 5.5 then gives us that

$$\Pr\left[R(g_{\widehat{f}_T}) - R(g_{f^*}) \geq \widetilde{C}_0(\delta T)^{-\beta} + x\right] \leq 2e^{-C_2\delta T x^{2-\alpha}},$$

with $\beta = 1/((2-\alpha)(1+p/2))$ for $p \in (0,2)$ and $\beta = 1/((2-\alpha)p)$ for $p > 2$, and where C_0 and C_2 depend on the constants of the problem.

□

Proof of lemma 5.4. Let $u(x) := (\widetilde{C}_1 + \widetilde{C}_2)(x/(\delta t))^{1/(2-\nu)} \vee (x/(\delta t))$. We have that

$$\Pr\left[X_T \geq C_0(\delta T)^{-\widetilde{\beta}} + u(x)\right] \leq \Pr\left[X_T \geq C_0(\delta T)^{-\widetilde{\beta}} + \widetilde{C}_1\left(\frac{x}{\delta T}\right)^{\frac{1}{2-\nu}} + \widetilde{C}_2\frac{x}{\delta T}\right] \leq 2e^{-x}.$$

Observe that, for any $y \geq 0$,

$$\begin{aligned} u^{-1}(y) =& \delta T\left(\left(\frac{y}{\widetilde{C}_1 + \widetilde{C}_2}\right)^{2-\nu} \wedge \frac{y}{\widetilde{C}_1 + \widetilde{C}_2}\right) \\ \geq& \widetilde{C}_3(y^{2-\nu} \wedge y), \end{aligned}$$

where $\widetilde{C}_3 := (\widetilde{C}_1 + \widetilde{C}_2)^{-\widetilde{\gamma}} \wedge (\widetilde{C}_1 + \widetilde{C}_2)^{-1}$. Therefore,

$$\Pr[X_T \geq C_0(\delta T)^{-\widetilde{\beta}} + y] \leq 2\exp(-\widetilde{C}_3\delta T(y^{\gamma} \wedge y)).$$

□

Proof of lemma 5.5. We have that, for any $a, x > 0$,

$$\begin{aligned} &\Pr\left[R(q, g_{\widehat{f}_T}) - R(q, g_1) \geq a + x\right] \\ \leq&\Pr\left[\widetilde{c}\left(R^{\phi}(q, \widehat{f}_T) - R^{\phi}(q, f_1)\right)^{\frac{1}{2-\nu}} \geq a + x\right] \\ =&\Pr\left[R^{\phi}(q, \widehat{f}_T) - R^{\phi}(q, f_1) \geq \left(\frac{a}{\widetilde{c}} + \frac{x}{\widetilde{c}}\right)^{2-\nu}\right] \\ \leq&\Pr\left[R^{\phi}(q, \widehat{f}_T) - R^{\phi}(q, f_1) \geq \left(\frac{a}{\widetilde{c}}\right)^{2-\nu} + \left(\frac{x}{\widetilde{c}}\right)^{2-\nu}\right], \end{aligned}$$

where the last line follows from the fact that, for any convex function ψ such that $\psi(0) = 0$, and any $x_1, x_2 > 0$, $\psi(x_1 + x_2) \geq \psi(x_1) + \psi(x_1)$. (To see it, observe that, as $x_1 + x_2 \geq x_1 \vee x_2$, $\frac{\psi(x_1+x_2)}{x_1+x_2} \geq \frac{\psi(x_1)}{x_1}$, and $\frac{\psi(x_1+x_2)}{x_1+x_2} \geq \frac{\psi(x_2)}{x_2}$, and thus $\psi(x_1+x_2) = \frac{x_1}{x_1+x_2}\psi(x_1+x_2) + \frac{x_2}{x_1+x_2}\psi(x_1+x_2) \geq \psi(x_1) + \psi(x_1)$). Set a such that $(a/\widetilde{c})^{2-\nu} = C_0(\delta t)^{-\widetilde{\beta}}$, that is $a = \widetilde{c}C_0^{1/(2-\nu)}(\delta t)^{-\widetilde{\beta}/(2-\nu)}$. Let $\widetilde{C}_0 := \widetilde{c}C_0^{1/(2-\nu)}$, $\beta = \widetilde{\beta}/(2-\nu)$. Then, we obtain,

$$\Pr\left[R(q, g_{\widehat{f}_T}) - R(q, g_1) \geq \widetilde{C}_0(\delta T)^{-\beta} + x\right]$$

$$
\begin{aligned}
&\leq \Pr\left[R^{\phi}(q,\widehat{f}_T) - R^{\phi}(q,f_1) \geq C_0(\delta T)^{-\widetilde{\beta}} + \left(\frac{x}{\delta T}\right)^{2-\nu}\right] \\
&\leq 2\exp\left(-\widetilde{C}_2 \delta T\left(\left(\frac{x}{\widetilde{c}}\right)^{\widetilde{\gamma}(2-\nu} \wedge \left(\frac{x}{\widetilde{c}}\right)^{2-\nu}\right)\right)
\end{aligned}
$$

Then, as $\widetilde{\gamma} \geq 1$, and $2-\nu \geq 1$, $1 \leq 2-\nu \leq \widetilde{\gamma}(2-\nu)$, and thus

$$
\left(\frac{x}{\widetilde{c}}\right)^{2-\nu} \in \left[\left(\frac{x}{\widetilde{c}}\right)^{\widetilde{\gamma}(2-\nu)} \wedge \frac{x}{\widetilde{c}}, \left(\frac{x}{\widetilde{c}}\right)^{\widetilde{\gamma}(2-\nu)} \vee \frac{x}{\widetilde{c}}\right].
$$

Therefore,

$$
\Pr\left[R(q,g_{\widehat{f}_T}) - R(q,g_1) \geq \widetilde{C}_0(\delta T)^{-\beta} + x\right] \leq 2\exp(-C_2\delta T(x^{\gamma} \wedge x)),
$$

with $C_2 := \widetilde{C}_2(\widetilde{c}^{-\gamma} \wedge \widetilde{c}^{-1})$, and $\gamma = \widetilde{\gamma}(2-\nu)$. □

Chapter 6

Model selection for contextual bandits

AURÉLIEN BIBAUT, ANTOINE CHAMBAZ, MARK VAN DER LAAN

As outlined in subsection 1.4.2 of the introduction chapter, as the underlying complexity of the environment instance is unknown to the learner a priori, an ideal sequential decision making algorithm should be able to adaptively choose the complexity of its policy class. In this chapter we consider model selection in the stochastic contextual setting (see subsection 1.2.1 of the introduction chapter).

We address the following generic model selection problem. Suppose we are given a collection of black box contextual bandit algorithms for which we know an upper bound on the minimax regret (these bounds are allowed to hold only under the realizability condition). Can we, without prior knowledge on the environment achieve the best minimax regret among algorithms for which the realizability condition holds?

We design a procedure that achieves this goal, up to logarithmic factor in time. We point out nevertheless that in this work we treat the number of arms and the dimension of the context space (or of feature space in linear bandit algorithms) as constants. Under these restrictions, our procedure is the first one to yield a meta algorithm that has adaptive rate of regret in time.

Our proposed procedure is relatively simple: for a well chosen sequence of probabilities $(p_t)_{t\geq 1}$, at each round t, it either chooses at random which candidate base algorithm to follow (with probability p_t) or compares, at the same internal sample size for each candidate, the cumulative reward of each, and selects the one that wins the comparison (with probability $1 - p_t$).

We demonstrate the effectiveness of our method with simulation studies.

6.1 Introduction

Contexual bandit (CB) learning is the repetition of the following steps, carried out by a an agent $\mathcal{A}$ and an environment $\mathcal{E}$.

1. the environment presents the agent a context $X \in \mathcal{X}$,
2. the agent chooses an action $A \in \{1, \dots, K\}$,
3. the environment presents the learner the reward Y corresponding to action A.

The goal of the agent is to accumulate the highest possible cumulative reward over a certain number of rounds T. The relative performance of existing CB algorithms depends on the environment $\mathcal{E}$: for instance some algorithms are best suited for settings where the reward structure is linear (LinUCB), but can be outperformed by greedy algorithms when the reward structure is more complex. It would therefore be desirable to have a procedure that is able to identify, in a data-driven fashion, which one of a pool of base CB algorithms is best suited for the environment at hand. This task is referred to as model selection. In batch settings and online full information settings, model selection is a mature field, with developments spanning several decades [Stone, 1974, Lepski, 1990, 1991, Gyorfi et al., 2002, Dudoit and van der Laan, 2005, Massart, 2007, Benkeser et al., 2018]. Cross-validation is now the standard approach used in practice, and it enjoys solid theoretical foundations [Devroye and Lugosi, 2001, Gyorfi et al., 2002, Dudoit and van der Laan, 2005, Benkeser et al., 2018].

Literature on model selection in online learning under bandit feedback is more recent and sparser. This owes to challenges specific to the bandit setting. Firstly, the bandit feedback structure implies that at any round, only the loss (here the negative reward) corresponding to one action can be observed, which implies that the loss can be observed only for a subset of the candidate learners (those which proposed the action eventually chosen). Any model selection procedure must therefore address the question of how to decide which base learner to follow at each round (the *allocation challenge*), and how to pass feedback to the base learners (the *feedback challenge*). A tempting approach to decide how to allocate rounds to different base learners is to use a standard multi-armed bandit (MAB) algorithm as a meta-learner, and treat the base learners as arms. This approach fails because, unlike in the usual MAB setting, the reward distribution of the arms changes with the number of times they get played: the more a base learner gets chosen, the more data it receives, and the better its proposed policy (and therefore expected reward) becomes. This exemplifies the *comparability challenge*: how to compare the candidate learners based on the available data at any given time?

Existing approaches solve these challenges in differents ways. We saw essentially two types of solutions in the existing literature, represented on the one hand by the OSOM algorithm of Chatterji et al. [2019b] and the ModCB algorithm of Foster et al. [2019], and on the other hand by the CORRAL algorithm of Agarwal et al. [2017], and the stochastic CORRAL algorithm, an improved version thereof introduced by Pacchiano et al. [2020a].

In OSOM and ModCB, the base learners learn policies in policy classes that form a nested sequence, which can be ordered from least complex to most complex. Their solution to the *allocation challenge* is to start by using the least complex algorithm, and move irreversibly to the next one if a goodness-of-fit test indicates its superiority. The goodness-of-fit tests uses all the data available to compute fits of the current and next policy, and compares them. This describes their solution to the *feedback challenge* and the *comparability challenge*.

CORRAL variants take another route. They use an Online Mirror Descent (OMD) based master algorithm that samples alternatively which base learner to follow, and gradually phases out the suboptimal ones. In that sense, their allocation strategy resembles the one of a MAB algorithm. The *comparability issue* arises naturally in the context of an OMD meta-learner, which can be understood easily with an example. Suppose that we have two base algorithms $\mathcal{A}(1)$ and $\mathcal{A}(2)$, and that $\mathcal{A}(1)$ has better asymptotic regret thant $\mathcal{A}(2)$. It can happen that either by chance ($\mathcal{A}(1)$ plays unlucky rounds) or by design (e.g. $\mathcal{A}(1)$ explores a lot in early rounds), $\mathcal{A}(1)$ fares worse than $\mathcal{A}(2)$ initially. As a result, the master would initially give a lesser weight to $\mathcal{A}(1)$ than to $\mathcal{A}(2)$, with the result that at some time t, the policy proposed by $\mathcal{A}(1)$ is based on a much smaller internal sample size than the policy proposed by $\mathcal{A}(2)$. As a result, at t, even though $\mathcal{A}(1)$ is asymptotically better than $\mathcal{A}(2)$, the losses of $\mathcal{A}(1)$ are worse than the losses of $\mathcal{A}(2)$, which accentuates the data-starvation of $\mathcal{A}(1)$ and can lead to $\mathcal{A}(1)$ never recovering from its early underperformance. The issue described here is that the losses used for the OMD weights update are not comparable across candidates, as they are based on policies informed by significantly different internal sample sizes. CORRAL can be viewed as the solution to the *comparability challenge* in the context of an OMD master: by using gentle weight updates (as opposed to the more aggressive weight updates of Exp3 for instance) and by regularly increasing the learning rate of base learners of which the weight drops too low, CORRAL prevents the base algorithm data-starvation phenomenon. The two CORRAL variants differ in their solution to the *feedback challenge*. The original CORRAL algorithm [Agarwal et al., 2017] passes, at each round, importance weighted losses to the master and to all base learners. In contrast, Pacchiano et al. [2020a]'s stochastic CORRAL passes unweighted losses to each base algorithm, but only at the time they get selected.

Guarantees in Chatterji et al. [2019a] and Foster et al. [2019] rely on the so-called *realizability* assumption, which states that at least one of the candidate policy classes contains $\pi_0(\mathcal{E})$, the optimal measurable policy under the current environment. Chatterji et al. [2019a] show that their approach achieves the minimax regret rate for the smallest policy class that contains $\pi_0(\mathcal{E})$. Foster et al. [2019] consider linear policy classes and show that their algorithm achieves regret no larger than $\widetilde{\mathcal{O}}(T^{2/3}d_*^{1/3})$ and $\widetilde{\mathcal{O}}(T^{3/4} + \sqrt{Td_*})$ where d^* is the dimension of the smallest policy class that contains $\pi_0(\mathcal{E})$. This is optimal if $d_* \geq \sqrt{T}$. In CORRAL variants, if one the the J base algorihtms has regret $O(T^\alpha)$, the master achieves regret $\widetilde{\mathcal{O}}(J/T + T\eta + T\eta^{(1-\alpha)/\alpha})$, with η the initial learning rate of the master. The learning rate η can be optimized so that this regret bound becomes $\widetilde{\mathcal{O}}(J^{1-\alpha}T^\alpha)$, that is, up to log factors, the upper bound on regret of that base algorithm. As pointed out by Agarwal et al. [2017], and as can be seen from the regret bound restated here, CORRAL presents an important caveat: the learning rate must be tuned to the rate of the base algorithm one wishes to compete with. This is not an issue when working with a collection of algorithms with same regret upper bound, and in that case CORRAL offers protection against model misspecification. However when base learners have different regret rates, CORRAL fails to adapt to the rate of the optimal algorithm.

In this article, we propose a master algorithm that allows to work with general off-the-shelf (contextual) bandit algorithhms, and achieves the same regret rate as the best of them. Our theoretical guarantees improve upon OSOM [Chatterji et al., 2019b] and ModCB [Foster et al., 2019] in the sense that our algorithm allows to work with a general collection of bandit algorithms, as opposed to a collection of algorithms based on a nested sequence of parametric reward models. It improves upon CORRAL variants in the sense that it is rate-adaptive. Our master algorithm can be described as follows: for a well chosen sequence $(p_t)_{t\geq 1}$ of exploration probabilities, at each time t, the master either samples a base algorithm uniformly at random and follows its proposal (with probability p_t), or it picks the base algorithm that maximizes a certain criterion based on past performance (with an exploitation probability of $1 - p_t$). Each algorithm receives feedback only if it gets played by the master. The crucial idea is to compare the performance of base algorithms at the same internal time. At global time t, the J algorithms are at internal times $n(1,t), \ldots, n(J,t)$ (with $n(1,t) + \ldots + n(J,t) = t$). We compare them based on their $\underline{\mathrm{n}}(t) := \min_{j\in[J]} n(j,t)$ first rounds, thus ensuring a fair comparison.

We organize the article as follows. In section 6.2, we formalize the setting consisting of a master algorithm allocating rounds to base algorithms. In section 6.3, we present our master algorithm, EnsBFC (Ensembling Bandits by Fair Comparison). We present its theoretical guarantees in section 6.4. We show in section 6.5 that many well-known existing bandit algorithms satisfy the assumption of our main theorem. We give experimental validation of our claims in section 6.6.

6.2 Problem setting

6.2.1 Master data and base algorithms internal data

A master algorithm $\mathcal{M}$ has access to J base contextual bandit algorithms $\mathcal{A}(1), \ldots, \mathcal{A}(J)$. At any time t, the master observes a context vector $X(t) \in \mathcal{X} \subset \mathbb{R}^d$, selects the index $\widehat{J}(t)$ of a base algorithm, and draws an action $A(t) \in [K] := \{1, \ldots, K\}$, following the policy of the selected base algorithm. The environment presents the reward $Y(t)$ corresponding to action $A(t)$. We distinguish two types of rounds for the master algorithm: exploration rounds and exploitation rounds. We will cover in more detail further down the definition of each type of round. We let $D(t)$ be the indicator of the event that round t is an exploration round. The data collected at time t by the master algorithm is $Z(t) := (D(t), \widehat{J}(t), X(t), A(t), Y(t))$. We denote $O(t) := (X(t), A(t), Y(t))$ the subvector of $Z(t)$ corresponding to the triple context, action, reward at time t. We denote $\mathcal{F}(t) := \sigma(Z(1), \ldots, Z(t))$, the filtration induced by the first t observations. We suppose that contexts are independent and identically distributed (i.i.d.) and that the conditional distribution of rewards given actions and contexts is fixed across time points.

After each round t, the master passes the triple $(X(t), A(t), Y(t))$ to base algorithm $\widehat{J}(t)$, which increments the internal time $n(\widehat{J}(t), t)$ of algorithm $\widehat{J}(t)$ by 1, and leaves unchanged the internal time of the other algorithms. For any $j \in [J]$, $n \geq 1$, we denote $\widetilde{O}(j,n) = (\widetilde{X}(j,n), \widetilde{A}(j,n), \widetilde{Y}(j,n))$ the triple collected by base algorithm j at its internal time n. Making this more formal, we define the internal time of j at global time t as $n(j,t) := \sum_{\tau=1}^{t} \mathbf{1}(\widehat{J}(\tau) = j)$, that is the number of times j has been selected by the master up till global time t. We define the reciprocal of $n(j,t)$ as $t(j,n) := \min\{t \geq 1 : n(j,t) = n\}$, that is the global time at which the internal time of j was updated from $n-1$ to n. We can then formally define $\widetilde{O}(j,n)$ as $\widetilde{O}(j,n) := (\widetilde{X}(j,n), \widetilde{A}(j,n), \widetilde{Y}(j,n)) := (X(t(j,n)), A(t(j,n)), Y(t(j,n))$. We denote $\widetilde{\mathcal{F}}(j,n) := \sigma(\widetilde{O}(j,1), \ldots, \widetilde{O}(j,n))$ the filtration induced by the first n observations of algorithm $\mathcal{A}(j)$.

Let $n^{\mathrm{xplr}}(j,t) := \sum_{\tau=1}^{t} \mathbf{1}(\widehat{J}(\tau) = j, D(\tau) = 1)$ and $n^{\mathrm{xplt}}(j,t) := \sum_{\tau=1}^{t} \mathbf{1}(\widehat{J}(\tau) = j, D(\tau) = 0)$, the number of exploration and exploitation rounds j was selected up till global time t. Note that $n(j,t) = n^{\mathrm{xplr}}(j,t) + n^{\mathrm{xplt}}(j,t)$. Define $\underline{\mathrm{n}}(t) := \min_{j \in [J]} n(j,t)$, $\underline{\mathrm{n}}^{\mathrm{xplr}}(t) := \min_{j \in [J]} n^{\mathrm{xplr}}(j,t)$, and $\underline{\mathrm{n}}^{\mathrm{xplt}}(t) := \min_{j \in [J]} n^{\mathrm{xplt}}(j,t)$.

6.2.2 Policies and base algorithm regret

A policy $\pi : [K] \times \mathcal{X} \to [0,1]$ is a conditional distribution over actions given a context, or otherwise stated, a mapping from contexts to a distribution over actions. So as to define the value and the risk of a policy, we introduce an triple of reference $(X^{\mathrm{ref}}, A^{\mathrm{ref}}, Y^{\mathrm{ref}})$ such that X^{ref} has same distribution as any context $X(t)$, $Y^{\mathrm{ref}} | A^{\mathrm{ref}}, X^{\mathrm{ref}}$ has same law as $Y(t) | A(t), X(t)$ for any t, and $A^{\mathrm{ref}} | X^{\mathrm{ref}} \sim \pi^{\mathrm{ref}}(\cdot, X^{\mathrm{ref}})$, where $\pi^{\mathrm{ref}}(a,x) := 1/K$ for every a and x. We introduce what we call the value loss ℓ, defined for any policy π and triple $o \in \mathcal{X} \times [K] \times \mathbb{R}$ as $\ell(\pi)(o) := -y\pi(a,w)/\pi^{\mathrm{ref}}(a,w)$. We then define the risk of π as $R(\pi) := E[\ell(\pi)(O^{\mathrm{ref}})]$. We will use that $-R(\pi) = E[Y^{\mathrm{ref}} \pi(A^{\mathrm{ref}}, X^{\mathrm{ref}}) / \pi^{\mathrm{ref}}(A^{\mathrm{ref}}, X^{\mathrm{ref}})] = E[\sum_{a=1}^{K} \pi(a | X^{\mathrm{ref}}) E[Y^{\mathrm{ref}} | A^{\mathrm{ref}} = a, X^{\mathrm{ref}}]]$, where the latter quantity is the value of π, that is the expected reward per round one would get if one carried out π under environment $\mathcal{E}$. We denote it $\mathcal{V}(\pi, \mathcal{E})$.

We denote $\pi(j,n)$ the policy proposed by $\mathcal{A}(j)$ at its internal time n. For any $x \in \mathcal{X}$, $\pi(j,n)(\cdot, x)$ is an $\widetilde{\mathcal{F}}(j, n-1)$-measurable distribution over $[K]$. We suppose that each algorithm $\mathcal{A}(j)$ operates over a policy class Π_j. The regret of $\mathcal{A}(j)$ over its first n rounds is defined as $\mathrm{Reg}(j,n) := \sum_{\tau=1}^{n}(\mathcal{V}_j^*(\mathcal{E}) - \widetilde{Y}(j,\tau))$, with $\mathcal{V}_j^*(\mathcal{E}) := \sup_{\pi \in \Pi_j} \mathcal{V}(\pi, \mathcal{E})$. We define the cumulative conditional regret as $\mathrm{CondReg}(j,n) := \sum_{\tau=1}^{n}(\mathcal{V}_j^*(\mathcal{E}) - E[\widetilde{Y}(j,\tau)|\widetilde{\mathcal{F}}_{\tau-1}]) = n(\overline{R}(j,n) - R_j^*)$, with $R_j^* = -\mathcal{V}_j^*(\mathcal{E})$ and $\overline{R}(j,n) = n^{-1}\sum_{\tau=1}^{n} R(\pi(j,\tau))$, where the identity follows from the fact that $E[\widetilde{Y}(j,\tau)|\widetilde{\mathcal{F}}(j,\tau-1)] = \mathcal{V}(\pi(j,\tau), \mathcal{E}) = -R(\pi(j,\tau))$. We define the pseudo regret as $\mathrm{pseudoReg}(j,n) := E[\mathrm{Reg}(j,n)]$.

6.2.3 Master regret and rate adaptivity

We let $\mathcal{V}^*(\mathcal{E}) := \max_{j \in [J]} \mathcal{V}_j^*(\mathcal{E})$, the optimal value across all policy classes $\Pi_1, \ldots, \Pi_J$, and similarly, we denote $R^* := \min_{j \in [J]} R_j^*$, the optimal risk across $\Pi_1, \ldots, \Pi_J$. We define the regret of the master as $\mathrm{Reg}(t) := \sum_{\tau=1}^{t} \mathcal{V}^*(\mathcal{E}) - Y(t)$, and the conditional regret as $\mathrm{CondReg}(t) := \sum_{\tau=1}^{t} \mathcal{V}^*(\mathcal{E}) - E[Y(\tau)|\mathcal{F}(\tau-1)]$.

The bandit literature gives upper bounds on either $\mathrm{Reg}(j,n)$ or $\mathrm{CondReg}(j,n)$ where the dependence in n is of the form $\widetilde{\mathcal{O}}(n^{1-\beta_j})$, for some $\beta_j \in (0,1)$. (We denote $a_n = \widetilde{\mathcal{O}}(b_n)$ if $a_n = \mathcal{O}(b_n (\log n)^\gamma)$ for some $\gamma > 0$.) While β_j is known, it is not the case for $\mathcal{V}_j^*(\mathcal{E})$, the asymptotic value of (the policy proposed by) $\mathcal{A}(j)$.

As a necessary requirement, a successful meta-learner should achieve asymptotic value $\mathcal{V}^*(\mathcal{E})$. A second natural requirement is that it should have as good regret guarantees as the best algorithm in the subset $\mathcal{J} := \{j \in [J] : \mathcal{V}_j^*(\mathcal{E}) = \mathcal{V}^*(\mathcal{E})\}$ of algorithms with optimal asymptotic value. We say that a master algorithm is *rate-adaptive* if it achieves these two requirements.

Definition 6.1 (Rate adaptivity). *Suppose that base algorithms have known regret (or conditional regret, or pseudo regret) upper bounds $\widetilde{\mathcal{O}}(n^{1-\beta_1}), \ldots, \widetilde{\mathcal{O}}(n^{1-\beta_J})$. Let $\beta(1) = \max_{j \in \mathcal{J}} \beta_j$, the rate exponent corresponding to the fastest upper bound rate among algorithms with optimal limit value $\mathcal{V}^*(\mathcal{E})$.*

We say that the master is rate-adaptive in regret (or conditional regret, or pseudo regret), up to logarithmic factors, if it holds that $\mathrm{Reg}(t) = \widetilde{\mathcal{O}}(t^{1-\beta(1)})$ *(or* $\mathrm{CondReg}(t) = \widetilde{\mathcal{O}}(t^{1-\beta(1)})$*, or* $\mathrm{pseudoReg}(t) = \widetilde{\mathcal{O}}(t^{1-\beta(1)})$*).*

Remark 6.1. *A natural setting where several base algorithms converge to the same value $\mathcal{V}^*(\mathcal{E})$ is when several of the candidate policy classes contain the optimal measurable policy $\pi_0(\mathcal{E})$, that is when the realizability assumption is satisfied for several base policy classes.*

Remark 6.2. *Suppose that rates $\widetilde{\mathcal{O}}(n^{1-\beta_1}), \ldots, \widetilde{\mathcal{O}}(n^{1-\beta_J})$ are minimax optimal (up to logarithmic factors) for the policy classes $\Pi_1, \ldots, \Pi_J$, and that at least one class contains $\pi_0(\mathcal{E})$. Then, in this context, rate adaptivity means that the master achieve the best minimax rate among classes that contain $\pi_0(\mathcal{E})$. In this context,* rate-adaptivity *coincides with the notion of* minimax adaptivity *from statistics' model selection literature (see e.g. Massart [2007], Giné and Nickl [2015]).*

Remark 6.3. *OSOM [Chatterji et al., 2019b] and ModCB Foster et al. [2019] are minimax adaptive (and thus rate-adaptive) under the condition that π_0 belongs to at least one of the policy classes (that is under the realizability assumption). CORRAL and stochastic CORRAL are not rate-adaptive.*

6.3 Algorithm description

Our master algorithm $\mathcal{M}$ can be described as follows. At each global time $t \geq 1$, $\mathcal{M}$ selects a base algorithm index $\widehat{J}(t)$ based on past data, observes the context $X(t)$, draws an action $A(t)$ conditional on $X(t)$ following the policy $\pi(\widehat{J}(t), n(\widehat{J}(t), t-1))$ proposed by $\mathcal{A}(\widehat{J}(t))$ at its current internal time, carries out action $A(t)$ and collects reward $Y(t)$. At the end of round t, $\mathcal{M}$ passes the triple $(X(t), A(t), Y(t))$ to $\mathcal{A}(\widehat{J}(t)))$, which then increments its internal time and updates its policy proposal based on the new datapoint.

To fully characterize $\mathcal{M}$ it remains to describe the mechanism that produces $\widehat{J}(t)$. We distinguish exploration rounds and exploitation rounds. We determine if round t is to be an exploration round by drawing, independently from the past $\mathcal{F}(t-1)$, the exploration round indicator $D(t)$ from a Bernoulli law with probability p_t, which we will define further down. During an exploration round (if $D(t) = 1$), we draw $\hat{J}(t)$ independently of $\mathcal{F}(t-1)$, from a uniform distribution over $[J]$. During an exploitation round (if $D(t) = 0$), we draw $\widehat{J}(t)$ based on a criterion depending on the past rewards of base algorithms. Let us define this criterion.

Let $\widehat{R}(j, n) := -n^{-1} \sum_{\tau=1}^{n} Y(j, \tau)$, the mean of negative rewards collected by algorithm j up till its internal time n. For any $n \geq 1$, define the algorithm selector $\widehat{j}(n, \widehat{R}(1, n), \ldots \widehat{R}(J, n), c_1) := \arg\min\{\widehat{R}(j, n) + c_1 n^{-\beta_j} : j \in [J]\}$, with $c_1 > 0$ a tuning parameter. When there is no ambiguity, we will use the shorthand notation $\widehat{j}(n)$. The selector $\widehat{j}(n)$ compares every base algorithm at the same internal time, and picks the one that minimizes the sum of the estimated risk at internal time n plus the theoretical regret upper bound rate $n^{-\beta_j}$. If $D(t) = 0$, we let $\hat{J}(t) := \widehat{j}(\underline{n}^{\text{xplr}}(t))$, that is we compare the base algorithms at a common internal time equal to the highest common number of exploration rounds each base has been called until t.

If any base algorithm j has average risk converging to some $R_j^* > R^*$, the regret of an exploration step is $O(1)$ in expectation. If we want the regret of the master with respect to (w.r.t.) R^* to be $\mathcal{O}(t^{-\beta(1)})$, we need the exploration probability p_t to be $\mathcal{O}(t^{-\beta(1)})$. Because $\beta(1)$ is unknown (it depends on $\mathcal{J}$ hence on $\mathcal{E}$ too), we make a conservative choice and we set $p_t := c_2 t^{-\overline{\beta}}$, with $\overline{\beta} := \max_{j \in [J]} \beta_j$ (a quantity available to us), where $c_2 > 0$ is a tuning parameter.

We give the pseudo code of the master algorithm $\mathcal{M}$ as algorithm 6.1 below.

6.4 Regret guarantees of the master algorithm

Our main result shows that the expected regret of the master satisfies the same theoretical upper bound with respect to R^* as the best base algorithm. The main assumption is that each base algorithm satisfies its conditional regret bound $\mathcal{O}(n^{1-\beta_j})$ with high probability. We state this requirement formally as an exponential deviation bound.

Algorithm 6.1 Ensembling Bandits by Fair Comparison (EnsBFC)

```
Input: base algorithms $\mathcal{A}(1), \ldots, \mathcal{A}(J)$, theoretical regret per round exponents $\beta_1, \ldots, \beta_j$, tuning parameters $c_1, c_2$.
  Initialize risk estimators: $\widehat{R}(j, 0) \leftarrow 0$ for every $j \in [J]$.
  for $t \geq 1$ do
    Draw exploration round indicator $D(t) \sim \text{Bernoulli}(p_t)$.
    if $D(t) = 1$ then
      Draw $\widehat{J}(t) \sim \text{Unif}([J])$.
    else
      Set $\widehat{J}(t) \leftarrow \widehat{j}(\underline{n}^{\text{xplr}}(t), \widehat{R}(1, \underline{n}^{\text{xplr}}(t)), \ldots \widehat{R}(J, \underline{n}^{\text{xplr}}(t)), c_1)$.
    end if
    Observe context $X(t)$.
    Sample action $A(t)$ following the policy proposed by $\mathcal{A}(\widehat{J}(t))$ at its current internal time:
        $$A(t)|X(t) \sim \pi(\widehat{J}(t), n(\widehat{J}(t), t-1))(\cdot, X(t)).$$
    Collect reward $Y(t)$.
    Pass the triple $(X(t), A(t), Y(t))$ to $\mathcal{A}(\widehat{J}(t))$, which then updates its policy proposal and increments its internal time by 1.
  end for.
```

Assumption 6.1 (Concentration). *There exists $C_0 \geq 0$, $C_1, C_2 > 0$, $\beta_1, \ldots \beta_J \leq 1/2$, $\nu_1, \ldots, \nu_J > 0$ such that, for any $n \geq 1$, $j \in [J]$ and $x \in [0, 1]$,*

$$P\left[\overline{R}(j, n) - R_j^* \geq C_0 n^{-\beta_j} + x\right] \leq C_1 \exp\left(-C_2 \times (n x^{1/\beta_j})^{\nu_j}\right), \tag{6.1}$$

and $\overline{R}(j, n) - R_j^* \geq 0$.

We also require that the rewards be conditionally sub-Gaussian given the past. Without loss of generality, we require that they be conditionally 1-sub-Gaussian.

Assumption 6.2. *For all $\lambda \in \mathbb{R}$, and every $t \geq 1$, $E[\exp(\lambda(Y_t - E[Y_t|\mathcal{F}_{t-1}])|\mathcal{F}_{t-1}] \leq \exp(\lambda^2/2)$.*

We show in the next section that the high probability regret bounds available in the literature for many well-known CB algorithms can be reformulated as an exponential deviation bound of the form (6.1). We can now state our main result.

Theorem 6.1 (Expected regret for the master). *Suppose that assumptions 6.1 and 6.2 hold, and recall the definition of $\beta(1)$ from subsection 6.2.3. Then,* EnsBFC *is rate-adaptive in pseudo-regret, that is,*

$$E\left[\sum_{t=1}^{T} (\mathcal{V}^*(\mathcal{E}) - Y(t))\right] \leq C T^{1-\beta(1)},$$

for some $C > 0$ depending only on the constants of the problem. If, in addition, the regret upper bounds satisfied by the base algorithms are minimax for their respective policy classes, then EnsBFC is minimax adaptive in pseudo regret.

Remark 6.4. *Assumption 6.1 is met for many well-known algorithms, as we show in the following section.*

Remark 6.5. *The $c_1 n^{-\beta_j}$ term in the criterion $\widehat{R}(j,n) + c_1 n^{-\beta_j}$ that $\widehat{j}(n)$ minimizes across $[J]$ ensures that $\widehat{R}(j,n) + c_1 n^{-\beta_j} - R^*$ is, in expectation, lower bounded by $c_1 n^{-\beta_j}$. It may be the case that, among the base algorithms that have optimal limit value $\mathcal{V}^*(\mathcal{E})$ (that is those in $\mathcal{J}$), the one that performs best in a given environment is not the one that has best regret rate upper bound $\widetilde{\mathcal{O}}(n^{1-\beta(1)})$. Enforcing this lower bound on the criterion ensures that the master picks an algorithm with optimal regret upper bound $\widetilde{\mathcal{O}}(n^{1-\beta(1)})$. We further discuss the need for such a lower bound in appendix 6.D.*

Remark 6.6. *The rate of pseudo-regret of EnsBFC is not impacted by the specific values of the tuning parameters c_1 and c_2 (as long as they are set to constants independent of T), but the finite performance is. We found in our simulations that setting $c_1 = 0.5$ and $c_2 = 10$ works fine. We leave to future work the task of designing a data-driven rule of thumb to select c_1 and c_2.*

In the next subsection, we take a step back to put our results in perspective with the broader model selection literature.

6.4.1 Comments on the nature of the result: minimax adaptivity vs. oracle equivalence

Results in the model selection literature are essentially of two types: minimax adaptivity guarantees and oracle inequalities.

Given a collection of statistical models, a model selection procedure is said to be minimax adaptive if it achieves the minimax risk of any model that contains the "truth". In our setting, the statistical models are policy classes and the "truth" is the optimal measurable policy $\pi_0(\mathcal{E})$. A notable example of minimax adaptive model selection procedure is Lepski's method [Lepski, 1990, 1991].

Consider a collection of estimators $\widehat{\theta}_1, \ldots, \widehat{\theta}_J$, and a data-generating distribution P, and denote $\mathcal{R}(\widehat{\theta}, P)$ the risk of any estimator $\widehat{\theta}$ under P. In our context, one should think of the estimators as the policies computed by the base algorithms, and of specifying P as specifying $\mathcal{E}$. We say that an estimator $\widehat{\theta}$ satisfies an oracle inequality w.r.t. $\widehat{\theta}_1, \ldots, \widehat{\theta}_J$ if $\mathcal{R}(\widehat{\theta}, P) \leq (1+\epsilon) \min_{j \in [J]} \mathcal{R}(\widehat{\theta}_j, P) + \mathrm{Err}$, with $\epsilon > 0$ and Err an error term. Moreover, we say that the estimator $\widehat{\theta}$ is oracle equivalent if $\mathcal{R}(\widehat{\theta}, P) / \min_{j \in [J]} \mathcal{R}(\widehat{\theta}_j, P) \to 1$. Being oracle equivalent means performing as well as the best instance-dependent (that is P-dependent) estimator. Multi-fold cross validation yields an oracle-equivalent estimator [Devroye and Lugosi, 2001, Gyorfi et al., 2002, Dudoit and van der Laan, 2005].

Our guarantees are closer to the notion of minimax adaptivity than to that of oracle equivalence, and, as we pointed out earlier, coincide with it if the base algorithms are minimax w.r.t. their policy classes. Minimax adaptivity is the property satisfied by the OSOM [Chatterji et al., 2019b] and ModCB Foster et al. [2019]. Minimax adaptivity is a worst-case (over each base model) statement, which represents a step in the right direction. We nevertheless argue that what practioners are looking for in a model selection procedure is to get the same performance as the base learner that performs best under the environment at hand, that is oracle equivalence, like the guarantee offered by multi-fold cross-validation.

6.5 High probability regret bound for some existing CB algorithms

In this section, we recast regret guarantees for well-known CB algorithms under the form the exponential bound (6.1) from our concentration assumption (assumption 6.1).

Recall the definitions of Reg, $\mathrm{CondReg}$ and $\mathrm{pseudoReg}$ from section 6.2. Observe that our concentration assumption is a high probability bound on $\overline{R}(n) - R^* = \mathrm{CondReg}(n)/n$, the average of the conditional instantaneous regret. Although some articles provide high probability bounds directly on $\mathrm{CondReg}(n)$ (e.g. Abbasi-Yadkori et al. [2011]), most works give high probability bounds on $\mathrm{Reg}(n)$. Fortunately, under the assumption that rewards are conditionally sub-Gaussian (assumption 6.2), we can recover a high probability regret bound on $\mathrm{CondReg}(n)$ from a high probability regret bound on $\mathrm{Reg}(n)$ using the Azuma-Hoeffding inequality.

(In the following paragraphs, we suppose, to keep notation consistent, that j is a base learner of the type considered in the paragraph).

UCB. [Lemma 4.9 in Pacchiano et al., 2020a], itself a corollary of [theorem 7 in Abbasi-Yadkori et al., 2011] states that if the rewards are conditionally 1-sub-Gaussian, the regret of UCB over n rounds is $\mathcal{O}(\sqrt{n \log(n/\delta)})$.

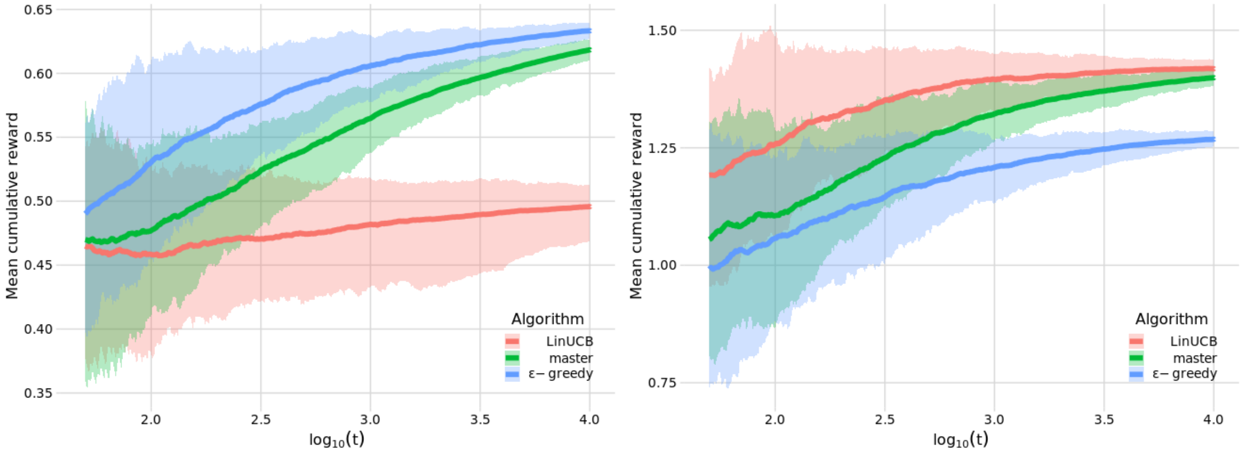
Corollary 6.1 (Exponential deviation bound for UCB)**.** *Suppose that assumption 6.2 holds. Then, there exist $C_0, C_1, C_2 > 0$ such that, for all $x \geq 0$, $P\left[\overline{R}(j,n) - R_j^* \geq C_0 n^{-1/2} (\log n)^{1/2} + x\right] \leq C_1 \exp(-C_2 n x^2)$.*

ε-greedy. Bibaut et al. [2020] consider the ε-greedy algorithm over a nonparametric policy class. The following result is a direct consequence of an intermediate claim in the proof [thereom 4 in Bibaut et al., 2020].

Lemma 6.1 (Exponential deviation bound for ε-greedy)**.** *Consider the ε-greedy algorithm over a nonparametric policy class Π. Suppose that the metric entropy in $\|\cdot\|_\infty$ norm of Π satisfies $\log N(\rho, \Pi, \|\cdot\|_\infty) = \mathcal{O}(\rho^{-p})$ for some $p > 0$, and that the exploration rate at t is $\epsilon_t \propto t^{-(\frac{1}{3} \vee \frac{p}{p+1})}$. Then, there exist $C_0, C_1, C_2 > 0$ such that, for all $x \geq 0$, $P\left[\overline{R}(j,n) - R_j^* \geq C_0 t^{-\beta} + x\right] \leq C_1 \exp(-C_2 \times (n x^{1/\beta})^{2\beta})$, with $\beta = \frac{1}{3} \vee \frac{p}{p+1}$.*

LinUCB. [Theorem 3 in Abbasi-Yadkori et al., 2011] states that LinUCB satisfies $\mathrm{CondReg}(n) = \mathcal{O}(\sqrt{n} \log(1/\delta))$ with probability at least $1 - \delta$. We recast their bound as follows.

Corollary 6.2. *Under the conditions of [theorem 3 in Abbasi-Yadkori et al., 2011], there exists $C_2 > 0$ such that, for all $x > 0$, $P\left[\overline{R}(j,n) - R_j^* \geq x\right] \leq \exp(-C_2 (n x^2)^{1/2})$*



(a) Environment 1

(b) Environment 2

Figure 6.1: Mean cumulative reward of the master and base algorithms over 100 runs, with (10%,90%) quantile bands

ILOVETOCONBANDITS. [Theorem 2 in Agarwal et al., 2014] et al. states that $\mathrm{Reg}(n) = \mathcal{O}(\sqrt{n}\log(n/\delta) + \log(n/\delta))$ with probability at least $1-\delta$. (The proof of their lemma actually states as an intermediate claim a $(1-\delta)$-probability bound on $\mathrm{CondReg}(n)$ which can easily be shown to be $\mathcal{O}(\sqrt{n}\log(n/\delta) + \log(n/\delta))$ as well). We recast their bound as follows.

Corollary 6.3 (Exponential deviation bound for ILOVETOCONBANDITS)**.** *Suppose that assumption 6.2 holds. Then, there exist $C_0 > 1$, $C_2 > 0$ such that, for any $x \geq 0$,*

$$P\left[\overline{R}(j,n) - R_j^* \geq C_0 n^{-1/2}\log n + x\right] \leq \exp(-C_2 n x^2).$$

6.6 Simulation study

We implemented EnsBFC using LinUCB and an ε-greedy algorihtm as base learners, and we evaluated it under two toy environments. We considered the setting $K = 2$. We chose environments $\mathcal{E}_1$ and $\mathcal{E}_2$, and the specifications of the two base algorithms such that:

- the ε-greedy has regret $\mathcal{O}(T^{2/3})$ w.r.t. the value $\mathcal{V}_0(\mathcal{E}_1)$ of the optimal measurable policy under $\mathcal{E}_1$, while LinUCB has linear regret lower bound $\Omega(T)$ w.r.t. $\mathcal{V}_0(\mathcal{E}_1)$,

- LinUCB has regret $\mathcal{O}(\sqrt{T})$ w.r.t. $\mathcal{V}_0(\mathcal{E}_2)$ while the ε-greedy algorithm has linear regret lower bound $\Omega(T)$ w.r.t. $\mathcal{V}_0(\mathcal{E}_2)$.

We present the mean cumulative reward results in figure 6.1. We demonstrate the behavior of the algorithm on a single run in figure 6.2 in appendix 6.E. We provide additional details about the experimental setting in appendix 6.E.

6.7 Discussion

In this chapter, we provide a uniform exploration based approach to model selection in contextual bandits. We showed that, by tuning adequately the rate of uniform exploration, the algorithm is guaranteed to achieve the rate in T of the best high probability regret bound among all algorithms for which the realizability assumption is satisfied. A caveat of our method is that it does not offer in its its current form adaptivity to other problem constants, such as the dimensionality of the feature space in linear bandit problems.

Other recent proposals [Abbasi-Yadkori et al., 2020, Pacchiano et al., 2020b] offers adaptivity to further problem constants such as the dimensionality of the feature space in linear bandit classes. Their analysis shares some similiarities with ours in the sense that it relies on high probability regret bounds and their guarantee is that their procedure achieve the best upper bound among algorithms for which realizability holds. The [Pacchiano et al., 2020a] only requires each base algorithm to come with a candidate upper bound (that may require on realizability for instance), and achieves the best regret bound among the base algorithms for which the candidate regret bound holds.

We conjecture that the analysis of our procedure can actually be extended in a straightforward fashion to achieve the same guarantees.

Broader Impact

Our work concerns the design of model selection / ensemble learning methods for contextual bandits. As it has the potential to improve the learning performance of any system relying on contextual bandits, it can impact essentially any setting where contextual bandits are used.

Contextual bandits are used or envisioned in settings as diverse as clinical trials, personalized medicine, ads placement and recommender systems. We therefore believe the broader impact of our work is positive inasmuch as these applications benefit to society.

Bibliography

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. pages 2312–2320, 2011.

Yasin Abbasi-Yadkori, Aldo Pacchiano, and My Phan. Regret balancing for bandit and rl model selection. *arXiv preprint arXiv:2006.05491*, 2020.

A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. E. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1638–1646, Beijing, China, 2014. PMLR.

A. Agarwal, Luo H., B Neyshabur, and R. E. Schapire. Corralling a band of bandit algorithms. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning*

Theory, volume 65 of *Proceedings of Machine Learning Research*, pages 12–38, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

D. Benkeser, C. Ju, S. Lendle, and M. J. van der Laan. Online cross-validation-based ensemble learning. *Statistics in Medicine*, 37(2):249–260, 2018.

A. F. Bibaut, A. Chambaz, and M. J. van der Laan. Generalized policy elimination: an efficient algorithm for nonparametric contextual bandits, 2020.

N. Chatterji, A. Pacchiano, and P. Bartlett. Online learning with kernel losses. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 971–980, Long Beach, California, USA, 09–15 Jun 2019a. PMLR.

N. S. Chatterji, V. Muthukumar, and P. L. Bartlett. Osom: A simultaneously optimal algorithm for multi-armed and linear contextual bandits, 2019b.

L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer-Verlag, New York, 2001.

S. Dudoit and M. J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2:131–154, July 2005.

D. J. Foster, A. Krishnamurthy, and H. Luo. Model selection for contextual bandits. In *Advances in Neural Information Processing Systems 32*, pages 14741–14752. Curran Associates, Inc., 2019.

E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015.

L. Gyorfi, M. Kohler, Krzyżak A., and Walk H. *A Distribution-free Theory of Nonparametric Regression*. Springer-Verlag, New York, 2002.

O. V. Lepski. A problem of adaptive estimation in gaussian white noise. *Theory of Probability and its Applications*, 35:454–470, 1990.

O.V. Lepski. Asymptotically minimax adaptive estimation i: Upper bounds. optimally adaptive estimates. *Theory of Probability and its Applications*, 36:682–697, 1991.

P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007.

A. Pacchiano, Phan M., Y. Abbasi-Yadkori, Rao A., J. Zimmert, T. Lattimore, and C. Szepesvari. Model selection in contextual stochastic bandit problems, 2020a.

Aldo Pacchiano, Christoph Dann, Claudio Gentile, and Peter Bartlett. Regret bound balancing and elimination for model selection in bandits and rl, 2020b.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974.

6.A Proof of theorem 6.1

We can without loss of generality assume that the tuning parameters c_1 and c_2 are set to 1. The proof of theorem 6.1 relies on the following lemmas.

Lemma 6.2. *For any $j \in [J]$, and $n, t \geq$, $\underline{n}^{\text{xplr}}(t)$ and $\widetilde{O}(j,n)$ are independent.*

The following lemma tells us that the probability of selecting j outside of the set $\mathcal{J}(1)$ of optimal candidates decrease exponentially with the common internal time of candidates.

Lemma 6.3 (Probability of selecting a suboptimal candidate)**.** *For all $n \geq 1$ and all $j \in [J] \backslash \mathcal{J}(1)$,*

$$P\left[\widehat{j}(n) = j\right] \leq C_{3,j} \exp\left(-C_{4,j} n^{\kappa_j}\right),$$

with $C_{3,j}, C_{4,j} > 0$ depending only on the constants of the problem, and $\kappa_j \in [0,1]$.

Proof. Suppose that $\widehat{j}(n) = j \in [J] \backslash \mathcal{J}(1)$. Then

$$\widehat{R}(j^*, n) + n^{-\beta(1)} \geq \widehat{R}(j,n) + n^{-\beta_j},$$

which we can rewrite as

$$\begin{aligned} &\left(\overline{R}(j^*,n) - R^*\right) + \left(\widehat{R}(j^*,n) - \overline{R}(j^*,n)\right) + \left(\overline{R}(j,n) - \widehat{R}(j,n)\right) \\ &\geq \left(\overline{R}(j,n) - R_j^*\right) + \left(R_j^* - R^*\right) + n^{-\beta_j} - n^{-\beta(1)}. \end{aligned}$$

Using that $\overline{R}(j,n) - R_j^* \geq 0$, we must then have

$$\begin{aligned} &\left(\overline{R}(j^*,n) - R^*\right) + \left(\widehat{R}(j^*,n) - \overline{R}(j^*,n)\right) + \left(\overline{R}(j,n) - \widehat{R}(j,n)\right) \\ &\geq \left(R_j^* - R^*\right) + n^{-\beta_j} - n^{-\beta(1)}. \end{aligned} \tag{6.2}$$

We distinguish two cases.

Case 1: $j \notin \mathcal{J}$. Then, $R_j^* - R^* \geq \Delta := \min_{j \notin \mathcal{J}} R_j^* - R^*$, which is strictly positive by definition of $\mathcal{J}$. Denote $\gamma(1) := \max\{\gamma_j : j \in \mathcal{J}(1)\}$ Therefore, for $n \geq n_0$ for some n_0 depending only of Δ and $n^{-\beta(1)}$, we can lower bound the right-hand side of (6.2) by $\Delta/2$, and we then have that for $n \geq n_0$,

$$\begin{aligned} P\left[\widehat{j}(n) = j\right] \leq & P\left[\left(\overline{R}(j^*,n) - R^*\right) + \left(\widehat{R}(j^*,n) - \overline{R}(j^*,n)\right) + \left(\overline{R}(j,n) - \widehat{R}(j,n)\right) \geq \frac{\Delta}{2}\right] \\ \leq & P\left[\overline{R}(j^*,n) - R^* \geq \frac{\Delta}{6}\right] + P\left[\widehat{R}(j^*,n) - \overline{R}(j^*,n) \geq \frac{\Delta}{6}\right] \\ & + P\left[\overline{R}(j,n) - \widehat{R}(j,n) \geq \frac{\Delta}{6}\right]. \end{aligned}$$

From assumption 6.1, the first term can be bounded as follows:

$$P\left[\overline{R}(j^*,n)-R^*\geq\frac{\Delta}{6}\right]\leq C_1\exp\left(-C_2\left(n\left(\frac{\Delta}{6}-C_0n^{-\beta(1)}(\log n)^{\gamma(1)}\right)^{1/\beta(1)}\right)^{\nu_{j^*}}\right)$$
$$\leq\widetilde{C}_{3,j}\exp\left(-\widetilde{C}_{4,j}n^{\nu_{j^*}}\right),$$

for some $\widetilde{C}_{3,j}>0$ and $\widetilde{C}_{4,j}>0$ that depend only on the constants of the problem.

The other two terms can be upper bounded using Azuma-Hoeffding: observing that for all j', $(\widehat{R}(j',\tau)-\overline{R}(j^*,\tau))_{\tau\geq 1}$ is a martingale difference sequence, and that from assumption 6.2, each of its term is 1-subGaussian conditionally on the past, we have that

$$P\left[\widehat{R}(j^*,n)-\overline{R}(j^*,n)\geq\frac{\Delta}{6}\right]\leq\exp\left(-n\frac{\Delta^2}{36}\right)$$
$$\text{and}\qquad P\left[\overline{R}(j,n)-\widehat{R}(j,n)\geq\frac{\Delta}{6}\right]\leq\exp\left(-n\frac{\Delta^2}{36}\right).$$

Therefore, $P[\widehat{j}(n)=j]\leq C_{3,j}\exp(-C_{4,j}n^{\kappa_j})$, for some $C_{3,j},C_{4,j}>0$ that only depend on the constants of the problem, and $\kappa_j:=\min(\nu_{j^*},1)$.

Case 2: $j\in\mathcal{J}\backslash\mathcal{J}(1)$**.** Then $R_j^*-R^*=0$. As $j\in\mathcal{J}\backslash\mathcal{J}(1)$, we have $\beta_j<\beta(1)$, and therefore, for $n\geq n_{1,j}$ that depends only on $\beta(1)$ and β_j, we have $n^{-\beta_j}-n^{-\beta(1)}\geq n^{-\beta_j}/2$. For any $n\geq n_{1,j}$, we can then lower bound the right-hand side of (6.1) by $n^{-\beta_j}/2$, and therefore, reasoning as in the previous step, we have that

$$P\left[\widehat{j}(n)=j\right]$$
$$\leq P\left[\overline{R}(j^*,n)-R^*\geq\frac{n^{-\beta_j}}{6}\right]+P\left[\widehat{R}(j^*,n)-\overline{R}(j^*,n)\geq\frac{n^{-\beta_j}}{6}\right]+P\left[\overline{R}(j,n)-\widehat{R}(j,n)\geq\frac{n^{-\beta_j}}{6}\right]$$

From assumption 6.1, the first term can be bounded as follows:

$$P\left[\overline{R}(j^*,n)-R^*\geq\frac{1}{6}n^{-\beta_j}\right]\leq C_1\exp\left(-C_2\left(n\left(\frac{1}{6}n^{-\beta_j}-C_0n^{-\beta(1)}(\log n)^{\gamma(1)}\right)^{1/\beta(1)}\right)^{\nu_j}\right).$$

For n large enough, $n^{-\beta_j}/6-C_0n^{-\beta(1)}(\log n)^{\gamma(1)}\geq n^{-\beta_j}/12$, and therefore, there exists $\widetilde{C}_{3,j}$ and $\widetilde{C}_{4,j}>0$ that depends only on the constants of the problem such that

$$P\left[\overline{R}(j^*,n)-R^*\geq\frac{1}{6}n^{-\beta_j}\right]\leq\widetilde{C}_{3,j}\exp\left(-\widetilde{C}_{4,j}n^{(1-\beta_j/\beta(1))\nu_j}\right).$$

Using Azuma-Hoeffding as in case 1 yields that

$$P\left[\widehat{R}(j^*, n) - \overline{R}(j^*, n) \geq \frac{n^{-\beta_j}}{6}\right] \leq \exp\left(-\frac{n^{1-2\beta_j}}{36}\right),$$
$$\text{and } P\left[\overline{R}(j, n) - \widehat{R}(j, n) \geq \frac{n^{-\beta_j}}{6}\right] \leq \exp\left(-\frac{n^{1-2\beta_j}}{36}\right).$$

Therefore, $P[\widehat{j}(n) = j] \leq C_{3,j} \exp(-C_{4,j} n^{\kappa_j})$, with $\widetilde{C}_{3,j}, \widetilde{C}_{4,j} > 0$ depending only on the constants of the problem, and $\kappa_j := \min(1 - 2\beta_j, (1 - \beta_j/\beta(1))\nu_j)$. Observe that $\kappa_j > 0$ as $\beta_j < \beta(1) \leq 1/2$ and $\nu_j > 0$. □

We can now prove theorem 6.1.

Proof of theorem 6.1. Observe that the regret at time of the master w.r.t. R^* can be decomposed as

$$\begin{aligned}
\text{Reg}(t) :=& E\left[\frac{1}{t}\sum_{\tau=1}^{t} Y(\tau)\right] - R^* \\
=& E\left[\sum_{j \in \mathcal{J}(1)} \frac{n(j,t)}{t}\left(\overline{R}(j, n(j,t)) - R^*\right)\right] + E\left[\sum_{j \notin \mathcal{J}(1)} \frac{n(j,t)}{t}\left(\overline{R}(j, n(j,t)) - R^*\right)\right].
\end{aligned}$$

Observe that for all $1 \leq n \leq t$, $n\overline{R}(j,n) - R^* = \sum_{\tau=1}^{n} R(\pi(j,\tau) - R^* \leq \sum_{\tau=1}^{t} R(\pi(j,\tau)) - R^* = t\overline{R}(j,t) - R^*$, since the terms in the sums are non-negative. Also, note that $\overline{R}(j, n(j,t)) - R^* \leq 1$ for all t and j. Therefore,

$$\text{Reg}(t) \leq \sum_{j \in \mathcal{J}(1)} E\left[\overline{R}(j,t) - R^*\right] + \sum_{j \notin \mathcal{J}(1)} \frac{E[n(j,t)]}{t}.$$

Recall that $n(j,t) = n^{\text{xplr}}(j,t) + n^{\text{xplt}}(j,t)$. It is straightforward to check that $E[n(j,t)^{\text{xplr}}(t)] \leq t^{-\overline{\beta}}/(1 - \overline{\beta})$. We now turn to $E[n(j,t)^{\text{xplt}}(t)]$. We have that

$$\begin{aligned}
E[n(j,t)^{\text{xplt}}(t)] =& E\left[\sum_{\tau=1}^{t} \mathbf{1}(\widehat{J}(t) = j, D(\tau) = 1)\right] \\
=& E\left[\sum_{\tau=1}^{t} \mathbf{1}(\widehat{j}(\underline{n}^{\text{xplr}}(\tau) = j, D(\tau) = 1)\right] \\
\leq& E\left[\sum_{\tau=1}^{t} \mathbf{1}(\widehat{j}(\underline{n}^{\text{xplr}}(\tau) = j)\right] \\
=& \sum_{\tau=1}^{t} E\left[P\left[\widehat{j}(\underline{n}^{\text{xplr}}(\tau) = j \middle| \underline{n}^{\text{xplr}}(\tau)\right]\right] \\
=& \sum_{\tau=1}^{t} E\left[\sum_{n=1}^{\tau} P\left[\widehat{j}(n) = j \middle| \underline{n}^{\text{xplr}}(\tau) = n\right] \mathbf{1}(\underline{n}^{\text{xplr}}(\tau) = n)\right]
\end{aligned}$$

From lemma 6.2, $\widehat{j}(n)$ is independent of $\underline{n}^{\text{xplr}}(\tau)$, and therefore, $P[\widehat{j}(n) = j|\underline{n}^{\text{xplr}}(\tau) = n] = P[\widehat{j}(n) = j]$. Therefore, using this fact and the bound on $P[\widehat{j}(n) = j]$ from lemma

$$\begin{aligned}E[n(j,t)^{\text{xplt}}(t)] \leq& \sum_{\tau=1}^{t}\sum_{n=1}^{\tau} P\left[\widehat{j}(n) = j\right] P\left[\underline{n}^{\text{xplr}}(\tau) = n\right]\\ \leq& \sum_{\tau=1}^{t}\sum_{n=1}^{\tau} C_{3,j}\exp\left(-C_{4,j}n^{\kappa_j}\right) P\left[\underline{n}^{\text{xplr}(\tau)} = n\right]\\ =& \sum_{\tau=1}^{t} C_{3,j}E\left[\exp\left(-C_{4,j}\underline{n}^{\text{xplr}}(\tau)\right)\right].\end{aligned}$$

It is straightforward to check that if $\kappa \in [0,1]$, $x \mapsto \exp(-x^\kappa)$ is convex. Therefore, from Jensen's inequality

$$\begin{aligned}E[n(j,t)^{\text{xplt}}(t)] \leq& C_{3,j}\sum_{\tau=1}^{t}\exp\left(-C_{4,j}E\left[\underline{n}^{\text{xplr}}(\tau)\right]\right)\\ \leq& C_{3,j}\sum_{\tau=1}^{t}\exp\left(-\frac{C_{4,j}}{J(1-\overline{\beta})}\tau^{-\overline{\beta}}\right)\\ \leq& C_{5,j},\end{aligned}$$

with $C_{5,j} := C_{3,j}\int_0^\infty \exp\left(-\frac{C_{4,j}}{J(1-\overline{\beta})}\tau^{-\overline{\beta}}\right)d\tau < \infty$.

Therefore, adding up the bounds on the expected number of exploration and exploitation rounds, we obtain

$$E[n(j,t)] \leq \left(C_{5,j} + \frac{1}{J(1-\overline{\beta})}t^{1-\overline{\beta}}\right) \leq C_{6,j}t^{1-\overline{\beta}},$$

for some $C_{6,j} > 0$ that depends only on the constants of the problem.

From assumption 6.1, for any $j \in \mathcal{J}(1)$, we have $E[\overline{R}(j,t) - R^*] \leq C_7 t^{-\beta(1)}$ for some $C_6 > 0$. Therefore,

$$\begin{aligned}\text{Reg}(t) \leq& JC_6t^{-\beta(1)} + \sum_{j=1}^{J} C_{7,j}t^{-\overline{\beta}}\\ \leq& Ct^{-\beta(1)},\end{aligned}$$

for some $C > 0$ that depends only on the constants of the problem. □

6.B Proof of the independence lemma

We start by stating a more general result of which lemma 6.2 is a corollary.

Lemma 6.4. *Consider some* $j \in [J]$*. Let, for all* $t \geq 1$*,* $U(t) := \mathbf{1}(D(t) = 1, J(t) = j)$*. Then, for every* $n, t \geq 1$*,* $U(t) \perp\!\!\!\perp \widetilde{\mathcal{F}}(j, n)$*.*

We can now prove lemma 6.2. We relegate the proof of lemma 6.4 after the one of lemma 6.2.

Proof of lemma 6.2. Observe that for every $j \in [J]$, $t \geq 1$, $n(j,t) := \sum_{\tau=1}^{t} \mathbf{1}(D(t) = 1, J(t) = j)$. Lemma 6.4 then immediately gives the wished claim. □

Proof of lemma 6.4. Let for all $t \geq 1$,$\mathcal{F}^{-}(t) := \sigma(\mathcal{F}(t-1), D(t), J(t))$. The hypothesis in the third bullet point can be rephrased as $Y(t)|\mathcal{F}^{-}(t) \overset{d}{=} \widetilde{Y}(j, n(j,t))|\widetilde{\mathcal{F}}(j, n(j,t)-1)$.

Fix j and t. We denote $U(t) := \mathbf{1}(D(t) = 1, J(t) = j)$. Observe that $U(t)$ is $\mathcal{F}^{-}(t)$-measurable, and that from the first and second conditions, $U(t) \perp\!\!\!\perp \mathcal{F}(t-1)$.

We show by induction that for all $n \geq 1$, $U(t) \perp\!\!\!\perp \widetilde{\mathcal{F}}(j,n)$. We treat the base case at the end of the proof. Suppose that for some $n \geq 1$, $U(t) \perp\!\!\!\perp \widetilde{\mathcal{F}}(j,n)$. Let us show that $U(t) \perp\!\!\!\perp \widetilde{\mathcal{F}}(j,n+1)$. It suffices to show that $U(t) \perp\!\!\!\perp \widetilde{Y}(j,n+1)|\widetilde{\mathcal{F}}(j,n)$. Observe that

$$\begin{aligned}
&P\left[\widetilde{Y}(j,n+1) = y, U(t) = u \middle| \widetilde{\mathcal{F}}(j,n)\right] \\
=&P\left[\widetilde{Y}(j,n+1) = y, U(t) = u, t(j,n+1) < t \middle| \widetilde{\mathcal{F}}(j,n)\right] \\
&+ P\left[\widetilde{Y}(j,n+1) = y, U(t) = u, t(j,n+1) \geq t \middle| \widetilde{\mathcal{F}}(j,n)\right].
\end{aligned}$$

We start with the first term. We have that

$$\begin{aligned}
&P\left[\widetilde{Y}(j,n+1) = y, U(t) = u, t(j,n+1) < t \middle| \widetilde{\mathcal{F}}(j,n)\right] \\
=&P\left[U(t) = u \middle| \widetilde{Y}(j,n+1) = y, t(j,n+1) < t \middle| \widetilde{\mathcal{F}}(j,n)\right] \\
&\times P\left[\widetilde{Y}(j,n+1) = y, t(j,n+1) < t, \widetilde{\mathcal{F}}(j,n)\right] \\
=&P\left[U(t) = u\right] P\left[\widetilde{Y}(j,n+1) = y, t(j,n+1) < t \middle| \widetilde{\mathcal{F}}(j,n)\right]
\end{aligned}$$

since $\{\widetilde{Y}(j,n) = y, t(j,n+1) < t\} \cap \widetilde{\mathcal{F}}(j,n)$ is $\mathcal{F}(t-1)$ measurable and $U(t) \perp\!\!\!\perp \mathcal{F}(t-1)$. Moreover, observe that $\{t(j,n+1) < t\} \cap \widetilde{\mathcal{F}}(j,n)$ is $\mathcal{F}^{-}(t(j,n+1))$-measurable, and therefore,

$$\begin{aligned}
&P\left[\widetilde{Y}(j,n+1) = y \middle| t(j,n+1) < t, \widetilde{\mathcal{F}}(j,n)\right] \\
=&E\left[P\left[Y(t(j,n+1)) = y \middle| \mathcal{F}^{-}(t(j,n+1))\right] \middle| t(j,n+1) < t, \widetilde{\mathcal{F}}(j,n)\right] \\
=&E\left[P\left[\widetilde{Y}(j,n+1) \middle| \widetilde{\mathcal{F}}(j,n)\right] \middle| t(j,n+1) < t, \widetilde{\mathcal{F}}(j,n)\right] \\
=&P\left[\widetilde{Y}(j,n+1) \middle| \widetilde{\mathcal{F}}(j,n)\right].
\end{aligned}$$

Therefore,

$$P\left[\widetilde{Y}(j,n+1) = y, U(t) = u, t(j,n+1) < t \middle| \widetilde{\mathcal{F}}(j,n)\right]$$

$$
\begin{aligned}
=&P\left[\widetilde{Y}(j,n+1)\middle|\widetilde{\mathcal{F}}(j,n)\right]P\left[U(t)=u\right]P\left[t(j,n+1)<t\middle|\widetilde{\mathcal{F}}(j,n)\right]\\
=&P\left[\widetilde{Y}(j,n+1)\middle|\widetilde{\mathcal{F}}(j,n)\right]P\left[U(t)=u,t(j,n+1)<t\middle|\widetilde{\mathcal{F}}(j,n)\right],
\end{aligned}
$$

since $U(t)\perp\!\!\!\perp\{t(j,n+1)<t\}$, as $\{t(j,n+1)<t\}$ is $\mathcal{F}(t-1)$-measurable, and $U(t)\perp\!\!\!\perp\widetilde{\mathcal{F}}(j,n)$ by induction hypothesis, which imply that $U(t)\perp\!\!\!\perp\{t(j,n+1)<t\}|\widetilde{\mathcal{F}}(j,n)$.

We now turn to the second term. Observe that $\{U(t)=u,t(j,n+1)\geq t\}\cap\mathcal{F}(j,n)$ is $\mathcal{F}^{-}(t(j,n+1))$-measurable. Therefore,

$$
\begin{aligned}
&P\left[\widetilde{Y}(j,n+1)=y\middle|U(t)=u,t(j,n+1)\geq t,\widetilde{\mathcal{F}}(j,n)\right]\\
=&E\left[P\left[Y(t(j,n+1))=y\middle|\mathcal{F}^{-}(t(j,n+1))\right]\middle|U(t)=u,t(j,n+1)\geq t,\widetilde{\mathcal{F}}(j,n)\right]\\
=&E\left[P\left[\widetilde{Y}(j,n+1)=y\middle|\widetilde{\mathcal{F}}(j,n)\right]\middle|U(t)=u,t(j,n+1)\geq t,\widetilde{\mathcal{F}}(j,n+1)\right]\\
=&P\left[\widetilde{Y}(j,n+1)=y\middle|\widetilde{\mathcal{F}}(j,n)\right].
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&P\left[\widetilde{Y}(j,n+1)=y,U(t)=u,t(j,n+1)\geq t\middle|\widetilde{\mathcal{F}}(j,n)\right]\\
=&P\left[\widetilde{Y}(j,n+1)=y\middle|\widetilde{\mathcal{F}}(j,n)\right]P\left[U(t)=u,t(j,n+1)\geq t\middle|\widetilde{\mathcal{F}}(j,n)\right].
\end{aligned}
$$

Therefore, adding up the identities for the two terms, we have

$$
P\left[\widetilde{Y}(j,n+1)=y,U(t)=u\middle|\widetilde{\mathcal{F}}(j,n)\right]=P\left[\widetilde{Y}(j,n+1)=y\middle|\widetilde{\mathcal{F}}(j,n)\right]P\left[U(t)=u\middle|\widetilde{\mathcal{F}}(j,n)\right].
$$

We have thus shown that $\widetilde{Y}(j,n+1)\perp\!\!\!\perp U(t)|\widetilde{\mathcal{F}}(j,n)$, which implies that $U(t)\perp\!\!\!\perp\widetilde{\mathcal{F}}(j,n+1)$.

The base case can be treated with the same arguments. □

6.C Proofs of reformulations of regret bounds for known base algorithms

Proof of corollary 6.1. As $\widetilde{Y}_\tau - E[\widetilde{Y}_\tau|\widetilde{F}_{\tau-1}]$ is conditionally 1-sub-Gaussian with probability at least $1-\delta/2$,

$$
\mathrm{CondReg}(n)\leq\mathrm{Reg}(n)+\sqrt{n\log(2/\delta)},
$$

and thus, using the high-probability regret bound from [Pacchiano et al., 2020a, lemma 4.9 in], there exists $C>0$ such that, with probability at least $1-\delta$,

$$
\mathrm{CondReg}(n)\leq C\sqrt{n\log(2n/\delta)}+\sqrt{n\log(2/\delta)}
$$

$$\begin{aligned}&\leq C\sqrt{n\log(2n)} + (C+1)\sqrt{n\log(2/\delta)}\\&\leq C'\sqrt{n\log n} + C'\sqrt{n\log(1/\delta)},\end{aligned}$$

for some $C' > C + 1$. Let $x = C\sqrt{n\log(1/\delta)}$, that is $\delta = \exp(-(C')^{-2}nx^2)$. Recalling that $\overline{R}(n) - R^* = \mathrm{CondReg}(n)/n$, we thus have that

$$P\left[\overline{R}(n) - R^* \geq C'n^{-1/2}(\log n)^{1/2} + x\right] \leq \exp\left(-(C')^{-2}nx^2\right).$$

□

Proof of lemma 6.1. It suffices to observe that

1. The bracketing entropy in any L_p norm is always dominated by the covering entropy in $\|\cdot\|_\infty$ norm.

2. The proof of [theorem 2 in Bibaut et al., 2020] gives the desired bound on $\overline{R}(n) - R^*$ as an intermediate result (right before relating it to the regret by using Azuma-Hoeffding).

□

Proof of corollary 6.2. [Theorem 3 in Abbasi-Yadkori et al., 2011] gives that there exists $C > 0$ such that $\overline{R}(n) - R^* = \mathrm{CondReg}(n)/n \leq Cn^{-1/2}\log(1/\delta)$ with probability at least $1-\delta$. Setting $x = Cn^{-1/2}\log(1/\delta)$, that is $\delta := \exp(-C^{-1}\sqrt{n}x)$, we have that

$$P\left[\overline{R}(n) - R^* \geq x\right] \leq \exp\left(-C^{-1}\sqrt{n}x\right),$$

which is the wished claim. □

Proof of corollary 6.3. As $\widehat{Y}_\tau - E[\widehat{Y}_\tau | \widehat{\mathcal{F}}_{\tau-1}]$ is conditionally 1-sub-Gaussian, Azuma-Hoeffding gives us that, with probability at least $1 - \delta/2$,

$$\mathrm{CondReg}(n) \leq \mathrm{Reg}(n) + \sqrt{n\log(2/\delta)}.$$

Therefore, combining this with the claim of [theorem 2 in Agarwal et al., 2014], there exists $C > 0$, such that, with probability $1 - \delta$

$$\begin{aligned}\overline{R}(n) - R^* &\leq Cn^{-1/2}\sqrt{\log(2n/\delta)} + Cn^{-1}\log(2n/\delta) + n^{-1/2}\sqrt{\log(2/\delta)}\\&\leq C\left(n^{-1/2}\sqrt{\log n} + n^{-1}\log n\right) + Cn^{-1}\log(2/\delta) + (C+1)n^{-1/2}\sqrt{\log(2/\delta)}\\&\leq C'\left(n^{-1/2}\sqrt{\log n} + n^{-1/2}\log(1/\delta)\right),\end{aligned}$$

for some $C' > C$. Letting $x = C'n^{-1/2}\log(1/\delta)$, this is equivalent with

$$P\left[\overline{R}(n) - R^* \geq C'n^{-1/2}(\log n)^{1/2} + x\right] \leq \exp\left(-C'\sqrt{n}x\right),$$

which is the wished claim. □

6.D Comment on the need to enforce a lower bound on the estimated risk

Unlike model selection methods such as Lepski's method and cross-validation, our method relies on explicit identification of the index of the best model. It is our understanding that such index identification tasks usually require the existence of a lower bound on the risk of each alternative, so as to ensure a gap in performance between the best and second best learner. Consider for instance the situation where one wants to adaptively estimate in L_∞ norm a density belonging to the union of a collection of Holder balls: $\mathcal{M}_s = H(s, B)$, where $H(s, M) := \{f : \mathbb{R}^d \to \mathbb{R} : |f(x) - f(y)| \leq M|x - y|^{s - \lfloor s \rfloor}\}$. It is well known, that while Lepski 's method is a minimax adaptive procedure with respect to $\{\mathcal{M}_s : s \in \mathcal{S}\}$, identification of the index s of the smallest Holder class that contains the truth is impossible without additional assumptions that enforce risk lower bounds [Giné and Nickl, 2015].

A parallel can perhaps be drawn with the best arm identification problem in multi-armed bandit settings: the analysis relies on the gap in mean reward between the best and second best arm.

Lower bounds of the sort we enforce are intrinsically tied to the minimax framework: they require the knowledge of a rate associated to the model class. Moving beyond the minimax framework to design a meta-learner that performs as well as the best instance-dependent base learner therefore seems to imply that such a procedure must not rely on identifying the index of the best model.

6.E Experimental details

In both environment, contexts are i.i.d. draws from $\mathcal{N}(\mathbf{0}, \mathbf{I}_4)$.

In environment 1, the rewards Bernoulli conditional on A and X, with conditional means specified as follows: for all $x = (x_1, x_2, x_3, x_4) \in \mathbb{R}^4$,

$$E[Y|A = 1, X = x] = \begin{cases} 0.1 \text{ if } x_1 < 0 \text{ and } x_2 < 0, \\ 0.5 \text{ if } x_1 < 0 \text{ and } x_1 \geq 0, \\ 0.7 \text{ if } x_1 \geq 0 \text{ and } x_2 < 0, \\ 0.45 \text{ otherwise} \end{cases},$$

and

$$E[Y|A = 1, X = x] = \begin{cases} 0.8 \text{ if } x_1 < 0 \text{ and } x_2 < 0, \\ 0.1 \text{ if } x_1 < 0 \text{ and } x_2 \geq 0, \\ 0.3 \text{ if } x_1 \geq 0 \text{ and } x_2 < 0, \\ 0.6 \text{ otherwise} \end{cases}.$$

In environment 2, for each $a \in \{1, 2\}$, rewards are normally distributed conditional on X: $E[Y|A = a, X = x] = \mu_a(x)\rangle + \eta$, with $\eta \sim \mathcal{N}(0, 1)$, with

$$\mu_1(x) = 0.9 + 0.5x_1 + 0.3x_2 - 0.9x_3 - 0.2x_4$$

$$\mu_2(x) = 0.9 - 0.5x_1 + 0.1x_2 - 0.7x_3 + 0.6x_4.$$

The ε-greedy learner uses as expected reward model the set of functions of the form $(a, x) \mapsto \beta_{a,0} + \beta_{a,1}\mathbf{1}(x_1 < 0, x_2 < 0) + \beta_{a,2}\mathbf{1}(x_1 < 0, x_2 \geq 0) + \beta_{a,3}\mathbf{1}(x_1 \geq 0, x_2 \geq 0)$. The reward learner therefore converges at a parametric rate to the truth under environment 1. Therefore, by setting the exploration rate to $t^{-1/3}$ at each t, regret under environment 1 is $\mathcal{O}(T^{2/3})$ over T rounds. However, this reward model does not contain the truth under environment 2,which implies that the ε-greedy algorithm incurs linear regret w.r.t. $\mathcal{V}_0(\mathcal{E}_1)$.

We use LinUCB with a linear model including all four components of x and an intercept, which implies that the realizability assumption is satisfied under environment 2, and therefore that the regret of LinUCB w.r.t. $\mathcal{V}_0(\mathcal{E}_2)$ is $\mathcal{O}(\sqrt{T})$ over T rounds.

Figure 6.2 demonstrates the master algorithm and its two base learners on a single run.

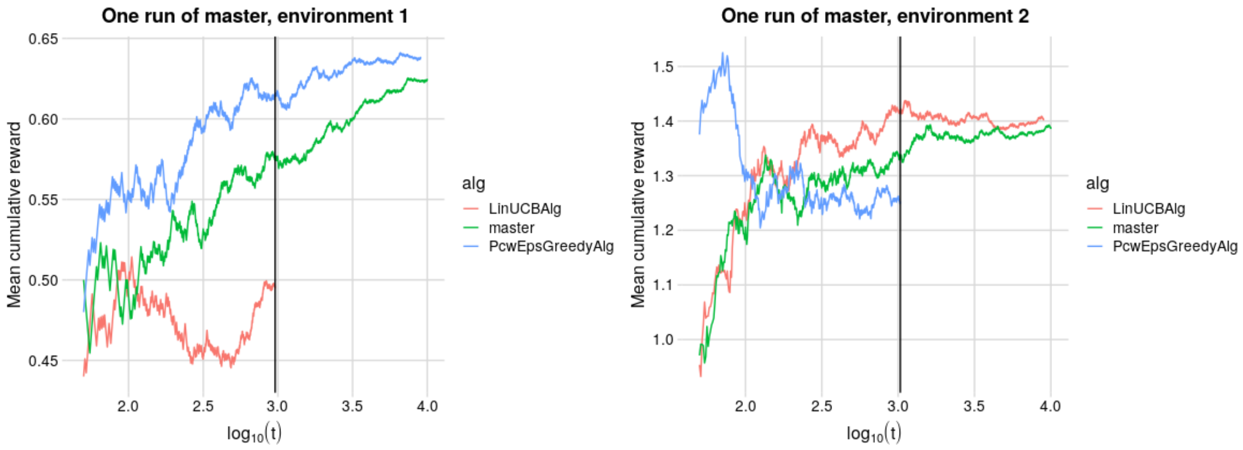


Figure 6.2: Mean cumulative reward of the master and its two base algorithms over 1 run. The vertical black line indicates $\underline{n}^{\text{xplr}}(T)$, with T the final global time in the simulation.

We tried the following values for the hyperparameters: $(c_1, c_2) \in \{0.1, 0.5, 1\} \times \{1, 10\}$. All specifications lead to the master appearing to converge to the performance of the optimal algorithm, but some values degrade a the performance in earlier rounds. As pointed out earlier, the specific constant values of c_1 and c_2 have no impact on the asymptotics.

The results of figure 6.1 were generated using an AWS EC2 instance of type r4.8xlarge, with 32 cores and 244 GiB of memory. Each plot takes about 30 minutes to compute.

The results of figure 6.2 were generated on a personal laptop and take less than 5 minutes to compute.

Chapter 7

Sequential causal inference in a single world of connected units

Aurélien Bibaut, Maya Petersen, Nikos Vlassis,
Maria Dimakopoulou, Mark van der Laan

In this chapter, we consider sequential decision making and causal inference under perhaps the most challenging setting we considered in this dissertation: we consider N units that exhibit network dependence that we follow along T time steps, that is we work under the statistical and causal models presented in subsection 1.2.4 of the introduction chapter.

We suppose that we are in a trial setting, that is the experimenter controls the treatment assignment of each unit at each time point, is allowed to choose treatments based on available history of all N units. We define causal quantities of interest under our causal model, show identifiability of these from the observed data distribution. We derive the canonical gradient and a targeted maximum likelihood estimator of these. The main technical challenge is to deal with the temporal and network dependence of observations. We derive a novel maximal inequality for empirical processes under mixing conditions, from which we derive an equicontinuity result and high probability risk bounds for empirical risk minimizers.

7.1 Introduction

We consider the setting in which, given a set of N individuals, a decision maker (or experimenter) alternatively, over a sequence of time points $t = 1, 2, \ldots$, assigns to each individual i a treatment $A(t, i)$ and then collects a vector of measurements $L(t, i)$ on this individual. We consider individual and time point specific outcomes $Y(t, i)$ that can be defined from $L(t, i)$. We also suppose that the decision maker can adapt the treatment assignment rule in response to past observations.

In these situations, it is often natural to define the performance of a treatment rule in terms of the expectation average outcomes of the form $N^{-1} \sum_{i=1}^{N} Y(\tau, i)$ at some time point τ, or as a function of such averages at different time points $\tau_1, \tau_2, \ldots$. Natural objectives the decision makers may want to pursue include learning as fast as possible the optimal treatment rule (a so-called *pure exploration* goal), or to ensure that over a certain time period, the individuals experience outcomes as high as possible (a so-called *regret minimization* objective).

This setting can arise in particular in business and public health applications. We consider two motivating examples.

First motivating example. Suppose an infectious disease is circulating in a country, and that public health officials can access, for each inhabitant i, at each time point t, a vector of measurements $L(t, i)$ including the infection status, which we define as the outcome $Y(t, i)$, demographic characteristics and the set of people i has been in contact with in the recent past. Suppose the government can assign to each individual i, at every time step t, a treatment $A(t, i)$ consistent of a certain set of restrictions on their daily activities. A question of interest is, given two candidates treatment rules, how to learn as quickly as possible which one is the most efficient.

Second motivating example. Suppose that the administrators of a web platform wonder which of two versions of the user interface users like best in the long run. They select a set of N users among all of the users of the platform, and assign from the beginning half of them to version 1 ($A(t, i) = 1$ for every $i = 1, \ldots, N/2$ and every $t \geq 1$) and the other half to version 2 ($A(t, i) = 2$

for every $i = N/2+1, \ldots, N$ and every $t \geq 1$). For each user i, at each time t, they collect of vector of measurements $L(t, i)$ on the user, which contains in particular measures of engagement with the platform, from which they define an outcome $Y(t, i)$. Given an arbitrary time point τ, a question of interest is: how to find out as quickly as possible which user interface would have maximized the expected average outcome $N^{-1}\sum_{i=1}^{N} Y(\tau, i)$ at τ?

While traditionally the causal inference and sequential decision problems literature have focused on single time point interventions or multiple time points interventions on independents units, the data collect in many real world situations involve network dependence between individuals. In the infectious disease setting presented above, the network dependence arises from contagion effects. In the web platform example, adjusting the treatment rule of individuals at time t as a function of the observed history of every individual up to time $t-1$ induces dependence between the trajectories of distinct individuals. In other business applications, there can be similar network effect arising due to word of mouth between socially connected users of the same service. Another source of association between units can arise from spillover effects, that is the effect of the treatment assignment of one individual on other individuals.

In this work, we define causal effects defined under temporal and network dependence, and we propose a methodology to design and analyze adaptive trials aiming at learning these causal effects. We propose a method to construct adaptive stopping rules for sequential hypothesis testing. We work under the key modelling assumption that the conditional distribution of the measurement vectors $L(t, i)$ given the past is constant across i and t. This assumption is comparable to an homogeneity assumption in a Markov Decision Process setting. We will see further down that as a key consequence of this homogeneity assumption, the error rates of our estimators over this model are a function of $T \times N$, which acts as the effective sample size, even though under temporal and network dependence, we only observe a single independent draw of the data-generating distribution.

7.1.1 Existing work

The setting we study conjugates three topics often treated separately: causal inference under temporal and network dependence, and adaptive experimentation.

Causal inference with temporal dependence. Temporal dependence in causal inference arises in longitudinal studies, and in particular in the dynamic treatment regime (DTR) literature for health applications, where patients are monitored across multiple time points, and a final outcome is measured at the end. In the DTR literature, dependence on the past is arbitrary and not identical across time points (as opposed to the homogeneous Markov Decision Process setting we discuss next), and convergence guarantees are stated in terms of the number of individuals N enrolled in the study. In another strand of causal inference for longitudinal settings, authors assume that the trajectory of each individual can be modelled as an homogeneous Markov Decision Process (MDP) (see Kallus and Uehara [2019] for the MDP model, and van der Laan et al. [2018] for a class of statistical models which include the MDP model), and convergence guarantees are stated in terms of the number of time points T if one follows only one single individual, or in terms of $T \times N$, if one follows the trajectories of N individuals over T time steps. These works can be categorized in the Off-Policy Evaluation (OPE) sub-field of Reinforcement Learning (RL), where it is standard to model the trajectory of the system by an MDP.

Causal inference in networks. Causal inference in networks has been studied by numerous authors (see e.g. Hudgens and Halloran [2008], Tchetgen and VanderWeele [2012], van der Laan [2013], Basse and Airoldi [2018], Basse et al. [2019], Ogburn et al. [2020]).

In network causal models, potential outcomes of an individual do not only depend on its own treatment history, but can also depend on the treatment history of other individuals it is connected to.

Applications include the study of infectious diseases and vaccines, spillover effects of advertising campaigns on social network platforms, and spillover effects in public policy interventions (see e.g. the aforementioned Basse et al. [2019]).

Adaptive experimentation. Adaptive experimentation, a sub-field of sequential decision problems has been a very active field of study for more than 75 years, with seminal contributions dating back to the work of Wald on sequential probability ratio tests [Wald, 1945], and the seminal multi-armed bandit paper of Robbins [Robbins, 1952]. The development of bandit algorithms was initially motivated by clinical trials with the goal of making these faster, and minimizing the opportunity cost of patients subjected to suboptimal treatments. Objective pursued in adaptive experimentation / sequential decision problems include (1) minimizing the cumulative regret, that is, over a fixed number of rounds, maximizing the sum of rewards collected / outcomes observed, and (2) inferential goals, such as identifying the best treatment arm with a certain predetermined level of confidence in as few rounds as possible (top arm identification in the fixed confidence setting), or to identify the best arm with as high a confidence level as possible, under a fixed number of rounds (best arm identification in the fixed confidence setting). In the bandit literature, it is usually assumed that rewards (and contexts in the contextual bandit setting) are independent of the past (this covers both the stochastic i.i.d. setting and the oblivious adversarial setting). In that sense, usual bandit methods aren't appropriate to deal with the case of trajectories of individuals with temporal dependence, which is the setting which interests us.

Bandit problems are a special case of reinforcement learning problems with trajectories of length $T = 1$. In the general case however, reinforcement learning is concerned with trajectories of the system over multiple time points, and states and outcomes at one time point can depend on the state, outcomes, and treatment at previous time points. Note that under the MDP model, which is the standard model considered in RL, dependency on the past is fully captured by the latest state and treatment. The reinforcement literature literature is concerned with learning optimal policies, either in a sequential fashion, or in an off-policy fashion, and with evaluating policies from off-policy data.

Sequential adaptive experimentation in the statistics literature. We see mainly two directions in which an experiment can be made adaptive to the past, and which the statistics literature has considered. The first one is the stopping rule. Contributions on adaptive stopping rules date back to the aforementioned work of Wald on sequential probability ratio tests [Wald, 1945]. The other component that can be chosen adaptively is the design, that is the stochastic rule that the experimenter uses to assign treatment in response to past observations and current covariates. An optimal design is defined with respect to a given statistical parameter and is the design that maximizes efficiency for that parameter, that is that leads to the smallest asymptotic variance of estimators of that parameter, and the highest power of tests of hypothesis defined from that parameter. van der Laan [2008] proposed a comprehensive methodology for sequential adaptive trials in the single and multiple time point settings, for independent individuals (as opposed to the network interference setting)

7.1.2 Contributions and comparison with past work

Our theoretical contributions are the following. We present our causal model, and we define our causal parameter as a mean outcome under a post-intervention distribution under this causal model. Under identifiability conditions, this post-intervention distribution equals a G-computation formula. We thus define formally our statistical parameter as the corresponding mean outcome under the G-computation formula.

We derive its efficient influence function (EIF), and thereby the semiparametric efficiency bound. Our statistical model subsumes in particular the homogeneous MDP model with independent trajectories. To the best of our knowledge, this work is the first to provide a formal derivation of the EIF of mean outcomes under a G-computation formula under the homogeneous MDP model.

We provide a Targeted Maximum Likelihood Estimator (TMLE) and a one-step estimator of our target parameter for generic sequential adaptive designs. We show under certain conditions that a certain process obtained by rescaling in time the sequence of estimates converges weakly to a Wiener process. In particular, this gives us the asymptotic distribution of our TMLE and one-step estimator. The conditions include in particular conditions on the design. We show that, for designs that converge to a fixed limit design, our estimator has asymptotic variance equal to the variance of the canonical gradient of the target parameter, which we conjecture equals the semiparametric efficiency bound. We use the results on the time-rescaled sequence of estimates to design a method to construct adaptive stopping rules for sequential hypothesis testing.

As mentioned earlier, some technical challenges arise from the fact that observations of different individuals at different time points are a priori dependent, and from the fact that we need to characterize the joint distribution of the sequence of estimates so as to be able to design an adaptive stopping rule. In particular the dependence between individuals implies that we cannot use the usual sample splitting techniques, which in the construction of semiparametric estimators allow to circumvent Donsker conditions [Klaassen, 1987, Zheng and van der Laan, 2011, Kallus and Uehara, 2020]. We therefore had to derive an almost sure equicontinuity result for empirical processes generated by weakly dependent data. The almost sure convergence aspect is key to obtaining guarantees on the joint distribution of the sequence of estimators. For the latter purpose, we also needed uniform-in-time convergence guarantees for nuisance estimators. We derived an exponential deviation bound for empirical risk minimizers fitted on weakly dependent data, which allows us to control these uniformly in time. Both the equicontinuity results and the results on empirical risk minimizers stem from a maximal inequality for empirical processes generated by weakly dependent data (that is from sequences that satisfy a certain mixing condition), which we obtain by applying an adaptive chaining device, a classical empirical process technique pioneered by [Ossiander, 1987], to a Bernstein inequality for mixing sequences, proven by Merlevède et al. [2009].

7.1.3 Paper organization

In section 7.2, we describe our causal model, our causal parameter, the statistical model, the statistical target parameter parameter, and the class of adaptive designs we consider. In section 7.3, we derive the efficient influence function (canonical gradient) of our target w.r.t. our statistical model, and we study the robustness properties of the EIF. In section 7.4, we provide a Targeted Maximum Likelihood Estimator and a one-step estimator of our parameter of interest, and we derive their convergence guarantees. In section 7.5, we introduce a class of functions that we use for nuisance modelling, and for modelling the EIF, and we give guarantees for empirical risk minimizers over it. In section 7.6, we show how to construct an adaptive stopping rule for sequential hypothesis testing. In section 7.7, we discuss adaptive learning of the optimal design.

7.2 Problem formulation

7.2.1 Observed data

An experimenter interacts with an environment consisting of N individuals indexed by $i = 1, \ldots,$ N, over rounds $t = 1, \ldots, T$. At each round t, the experimenter first assigns the treatment vector $A(t) := (A(t,1), \ldots, A(t,N))$ to the N individuals, where $A(t,i)$ is the treatment assigned to individual i at time t, and then observes for each individual i, a vector $L(t,i)$ of time-varying covariates and outcomes. Let $L(t) := (L(t,1), \ldots, L(t,N))$. We denote $O(t,i) := (A(t,i), L(t,i))$ the data observed for individual i at time t, and $O(t) = (A(t), L(t))$ the data observed for the all of the N individuals at round t, $\bar{O}(t) := (O(1), \ldots, O(t))$ the data available at round t, $L^-(t) := \bar{O}(t-1)$, the data observed before $L(t)$, $A^-(t) := (\bar{O}(t-1), L(t))$, the data observed before $A(t)$, $L^-(t,i) = (L^-(t), L(t,1), \ldots, L(t,i-1))$, and $A^-(t,i) = (A^-(t), A(t,1), \ldots, A(t,i-1))$.

Under this notation, the data observed throughout the course of the trial is thus $\bar{O}(T)$, which we will also denote $O^{T,N}$ to make explicit the dependence on the number of individuals N.

7.2.2 Causal model

Formal definition of the causal model. We suppose that there exists a set of deterministic functions $\{f_{A(t,i)}, f_{L(t,i)} : t \in [T], i \in [N]\}$, and a set of unobserved random variables $U := (U^A, U^L)$, with $U^L := (U^L(t,i) : t \in [T], i \in [N])$ and $U^A := (U^A(t,i) : t \in [T], i \in [N])$, such that, for every $(t,i) \in [T] \times [N]$,

$$A(t,i) = f_{A(t,i)}(\bar{L}(t-1), \bar{A}(t-1), U^A(t,i)), \tag{7.1}$$

$$L(t,i) = f_{L(t,i)}(\bar{A}(t), \bar{L}(t-1), U^L(t,i)). \tag{7.2}$$

We place no restriction at this point on the functional form of the functions $f_{A(t,i)}$, $f_{L(t,i)}$. The set of equations (7.1)-(7.2) form a so-called *Nonparametric Structural Equation Model* (NPSEM) (see e.g Pearl [2009]). Let

$$\mathcal{M}_F^{T,N} := \left\{ P_F^{T,N} : P_U, f_{A(t,i)}, f_{L(t,i)}, \widetilde{c}_{A(t,i)}, \widetilde{c}_{L(t,i)} : t \in [T], i \in [N] \right\},$$

be the set of probability distributions $P_F^{T,N}$ over the domains of $(O^{T,N}, U)$ induced by the NPSEM as the distribution P_U of the unmeasured variable vector U, and the functions $f_{A(t,i)}, f_{L(t,i)}, \widetilde{c}_{A(t,i)}$, $\widetilde{c}_{L(t,i)}$ vary freely. The set $\mathcal{M}_F^{T,N}$ is our so-called *causal model*.

We denote $P_{0,F}^{T,N}$ the true distribution of $(O^{T,N}, U)$. In the remainder of the article, we will use the subscript "0" to indicate true probability distributions or components thereof. Note that the full data distribution fully determines the observed data distribution.

Counterfactual data and post-intervention distribution. We now describe a counterfactual scenario in which the connectivity of the nodes and the intervention assigned to them is not as under the NPSEM above.

Let $\{g^*_{s,j} : s \in [\tau], j \in [N]\}$ be a collection of stochastic interventions at nodes $\{A(s,j), s \in [T], j \in [N]\}$, that is, for every (s,j), $g^*_{s,j}$ is a distribution over treatment arms conditional on $\bar{A}(t-1), \bar{L}(t-1)$.

Let $O^{*,T,N} := (O^*(t,i) : t \in [T], i \in [N])$, with $O^*(s,j) := (A^*(s,g), L^*(s,j))$ be the counterfactual data set generated from U by the following NPSEM, obtained from the NPSEM (7.1)-(7.2) by replacing the intervention nodes by the counterfactual interventions $g^*_{s,j}$:

$$\begin{aligned} A^*(s,j) \sim & g^*_{s,j}(\cdot \mid \bar{L}(s-1), \bar{A}(s-1)), \\ L^*(s,j) = & f_{L(s,j)}(\bar{A}(s-1), \bar{L}(s-1), U^L(s,j)). \end{aligned}$$

The distribution of $(O^{*,T,N}, U)$ is the so-called *post-intervention distribution* of the full data. We use the notation $P_F^{*,T,N}$ for the post-intervention distribution of the full data.

Causal target parameter. Let $\tau \geq 1$ be an arbitrary time point. We define as our causal target parameter as a certain mean outcome at time point τ, under the post-intervention distribution:

$$\Psi_\tau^F(P_F^{T,N}) = E_{P_F^{*,T,N}}\left[Y^*(\tau)\right],$$

where $Y^*(\tau) := N^{-1} \sum_{i=1}^N Y^*(\tau, i)$, where $Y^*(\tau, i)$ is a unit-specific outcome at time τ, defined as $Y^*(\tau, i) := f_Y(L^*(\tau, i))$ for a certain function f_Y. In words, $\Psi_\tau^F(P_F^{T,N})$ is the mean outcome at time τ, in the counterfactual scenario where the treatment mechanism is g^*.

Identifiability. We use the notation $P^{T,N}$ for a generic distribution over the domain of the observed data $O^{T,N}$ and we denote $P_0^{T,N}$ the true distribution of the observed data. As mentioned above, any distribution $P_F^{T,N}$ on the domain of the full data fully determines a corresponding distribution $P^{T,N}$ on the domain of the observed data. For any $P_F^{T,N}$, we say that the causal parameter

$\Psi^F_\tau(P^{T,N}_F)$ is identifiable if we can write it as a function of the corresponding observed data distribution $P^{T,N}$.

The parameter $\Psi^F_\tau(P^{T,N}_F)$ is identifiable under the following two assumptions.

Assumption 7.1 (Sequential randomization)**.** *For any* (t,i)*,* $A(t,i) \perp\!\!\!\perp L^*(t,i) \mid A(t,i)^-$.

Assumption 7.2 (Positivity)**.** *For any* (t,i)*,* $a \in \mathcal{A}$*, and* $l(t,i)^-$ *such that* $P_0\left[L(t,i)^- = l(t,i)^-\right] > 0$,

$$P_0\left[A(t,i) = a \mid L(t,i)^- = l(t,i)^-\right] > 0.$$

It can be shown under assumptions 7.1 and 7.2 that the post-intervention distribution of O^* under $P^{*,T,N}_F$ equals the following G-computation formula:

$$P^{T,N}_{g^*}(O^{T,N}) := \prod_{s=1}^{\tau}\prod_{j=1}^{N} P(L(s,j) \mid L(s,j)^-) g^*_{s,j}(A(s,j) \mid A(s,j)^-).$$

Note that the factors of $P^{T,N}_{g^*}$ are the conditional distributions $P(L(t,i) \mid L(t,i)^-)$, which are known if we know $P^{T,N}$, and the known counterfactual intervention g^*. Therefore, under assumptions 7.1 and 7.2, there exists a mapping Ψ_τ such that

$$\Psi^F_\tau(P^{T,N}_F) = \Psi_\tau(P^{T,N}).$$

The quantity $\Psi_\tau(P^{T,N}_0)$ is then our statistical target parameter. For $\gamma \in (0,1)$, which we interpret as a *discount factor*, we also define the following other target parameter, derived from Ψ_τ:

$$\Psi_{\tau,\gamma}(P^{T,N}) := \sum_{\tau' \geq \tau} \gamma^{\tau'-\tau} \Psi_{\tau'}(P^{T,N}).$$

This parameter is the typical target (in particular in the case $\tau = 1$) in the Off-Policy Evaluation problem in reinforcement learning (see e.g. Kallus and Uehara [2020, 2019]). As the analysis of $\Psi_{\tau,\gamma}$ follows from the analysis of Ψ_τ, in the rest of the paper, we treat Ψ_τ as our main object of interest, and we will simply denote it Ψ when there is no ambiguity about τ.

7.2.3 Statistical model

The statistical model is the set of distributions we believe to contain the true data-generating distribution $P^{T,N}_0$. We denote it $\mathcal{M}^{T,N}$. We suppose that all of the elements of $\mathcal{M}^{T,N}$ admit a density w.r.t. a common dominating measure μ. For any $P^{T,N}$, we denote $p^{T,N} := dP^{T,N}/d\mu$.

From the chain rule, any such $p^{T,N}$ can be factorized as a product of conditional densities as follows:

$$p^{T,N}(o^{T,N}) := \prod_{t=1}^{T}\prod_{i=1}^{N} g_{t,i}(a(t,i) \mid a(t,i)^-) q_{t,i}(l(t,i) \mid l(t,i)^-).$$

The conditional densities $q_{t,i}$ are a fact of nature and represent how each $L(t,i)$ responds to past interventions $A(s,j)$, and depends on the vectors $L(s,j)$ of past time-varying covariates and outcomes of for each individual $j \in [N]$, for every time point $s < t$. We refer to it as the uncontrolled part of the data-generating process of the trial, as the experimenter does not have control over it. The factors $g_{t,i}$, in our randomized experimental setting, are in control of the experimenter, and represent the set of stochastic decision rules she follows to assign treatment at each time step.

The above decomposition places no restriction on the temporal and network dependence in the observed data. We make an assumption on the complexity of the dependence allowed by supposing that each $L(t,i)$ can depend on the past of the trial only through a summary measure of the history of a fixed number of individuals.

Assumption 7.3 (Conditional independence given summary measure)**.** *There exists a Euclidean set $\mathcal{C}_L \subset \mathbb{R}^{d_1}$, for some $d_1 \geq 1$, and a set of deterministic functions $c_{L(t,i)}$, $t \in [T]$, $i \in [N]$, with image included in $\mathcal{C}$, such that, for every $t \in [T]$, $i \in [N]$,*

$$q_{t,i}(l(t,i) \mid l(t,i)^-) = q_{t,i}(l(t,i) \mid c_{L(t,i)}(l(t,i)^-)).$$

We denote $C_L(t,i) := c_{L(t,i)}(L(t,i)^-)$.

The vector $C_L(t,i)$, which lies in the Euclidean set $\mathcal{C}_L$, plays the role of a finite dimensional summary measure of the past, which is such that $L(t,i)$ is independent of its past when conditioning on this summary measure. Following terminology used in existing works [Boruvka et al., 2017, van der Laan et al., 2018], we will refer to it as the "context" preceding the node $L(t,i)$.

Without any further assumptions, it is a priori not possible to obtain consistent estimators from a single draw of $O^{T,N}$. For consistent estimation to be possible, we need the likelihood to exhibit a repeated factor. We therefore make the following assumption.

Assumption 7.4 (Homogeneity)**.** *The factors $q_{t,i}$ are constant across values of i and t, that is there exists a common conditional density q such that $q_{t,i} = q$ for every t and i.*

We can now state a formal definition of our statistical model.

Definition 7.1 (Statistical model)**.** *Fix μ a summary measure on the domain of $O^{T,N}$. We define our statistical model $\mathcal{M}^{T,N}$ as the set of distributions $P^{T,N}$ over the domain of $O^{T,N}$ that satisfy assumptions 7.3 and 7.4.*

Remark 7.1. *We emphasize that, as we are in the setting of a controlled trial, the factors $g_{t,i}$ are known.*

Remark 7.2. *Observe that any distribution $P^{T,N}$ in our statistical model $\mathcal{M}^{T,N}$ is fully determined by q, and $\mathcal{M}^{T,N}$ is indexed by the set of conditional densities $\mathcal{M}_q$ that we believe to contain q_0. Here, we assume that $\mathcal{M}_q$ is a saturated nonparametric model, that is for any $c \in \mathcal{C}_L$, the tangent space of $\{l \mapsto q(\cdot \mid c) : q \in \mathcal{M}_q\}$ is equal to the Hilbert space*

$$L^2_{0,c}(q) := \left\{ (l,c) \to f(l,c) : \int f^2(l,c) q(l \mid c) dl < \infty \text{ and } \int f(l,c) q(l \mid c) dl = 0 \right\}.$$

Remark 7.3. *Under the homogeneity assumption, the target parameter $\Psi(P^{T,N})$ depends on $P^{T,N}$ only through the common conditional density $q = q(P^{T,N})$. Therefore, there exists a mapping $\Psi^{(1)}$ such that $\Psi(P^{T,N}) = \Psi^{(1)}(q(P^{T,N}))$.*

In constructing and analyzing our estimators, we will require the following assumption.

Assumption 7.5. *For any $s \in [\tau]$, $j, k \in [N]$,*

$$E_{q,g^*}\left[Y(k) \mid L(s,j), C_L(s,j)\right] = E_{q,g^*}\left[Y(k) \mid L(s,j), L(s,j)^-\right],$$

where E_{q,g^*} is the expectation operator under the G-computation formula $P^{\tau,N}_{g^*}$.

Making assumption 7.5 on top of the previous two assumptions 7.3 and 7.4 defines a new statistical model $\mathcal{M}^{T,N}_1$, which is a subset of our previously defined statistical model $\mathcal{M}^{T,N}$. Note that this statistical model a priori depends on g^*. We want to emphasize the following: while we will derive in the next section the canonical gradient $D(P^{T,N})$ of our target parameter Ψ w.r.t. the larger model $\mathcal{M}_{T,N}$, we will use the assumptions defining the smaller model $\mathcal{M}^{T,N}_1$ to derive a more tractable representation $D_1(P^{T,N})$ of this canonical gradient $D(P^{T,N})$, which we will use to build our estimators. As a result, estimators that achieve asymptotic variance equal to the variance of $D_1(P^{T,N})$ can only be locally efficient w.r.t. the model $\mathcal{M}^{T,N}$: they can achieve the efficiency bound for $\mathcal{M}^{T,N}$ at $P^{T,N}$ only if $P^{T,N} \in \mathcal{M}^{T,N}_1$.

Finally, in some special cases that we discuss next it might be realistic to make the following set of three assumptions.

Assumption 7.6 (Context decomposition)**.** *For any $t \in [T]$, $i \in [N]$, the context summary mapping $c_{L(t,i)}$ can be decomposed as*

$$c_{L(t,i)}(l(t,i)^-) = (a(t,i), c_A^{g,g^*}(t,i)),$$

where $c_A^{g,g^*}(t,i)$ is a context for the node $A(t,i)$ of the form

$$c_A^{g,g^*}(t,i) = c_{A(t,i)}^{g,g^*}(a(t,i)^-),$$

where $c_{A(t,i)}^{g,g^*}$ is a known deterministic function with image in a Euclidean set $\mathcal{C}_A \subset \mathbb{R}^{d_2}$, such that, for any $a(t,i)$, $a(t,i)^-$,

$$\begin{aligned} g_{t,i}(a(t,i) \mid a(t,i)^-) =& g_{t,i}(a(t,i) \mid c_A^{g,g^*}(t,i)) \\ \textit{and } g^*_{t,i}(a(t,i) \mid a(t,i)^-) =& g^*_{t,i}(a(t,i) \mid c_A^{g,g^*}(t,i)). \end{aligned}$$

Assumption 7.7 (Individual outcomes independent from other trajectories)**.** *For any $q \in \mathcal{M}_q$, it holds under the corresponding G-computation formula $P^{\tau,N}_{g^*}$ that $Y(k) \perp\!\!\!\perp O(s,j)$ for any $s \in [\tau]$ and $k \neq j$.*

Assumption 7.8 (Observed treatment homogeneity)**.** *The treatment mechanisms $g_{t,i}$ are constant across t and i, that is, there exists a conditional density g such that $g_{t,i} = g$ for every t and i.*

Making assumptions 7.6, 7.7 and 7.8 on top of assumptions 7.3, 7.4 and 7.5 defines a new statistical model $\mathcal{M}_2^{T,N}$ such that $\mathcal{M}_2^{T,N} \subset \mathcal{M}_1^{T,N} \subset \mathcal{M}^{T,N}$. Here too, we emphasize that we will use these additional assumptions to obtain a simplified representation $D_2(P^{T,N})$ of the canonical gradient of Ψ w.r.t. the larger model $\mathcal{M}^{T,N}$, but that we won't derive the canonical gradient w.r.t. the smaller model $\mathcal{M}_2^{T,N}$. As a result, estimators achieving asymptotic variance equal to the variance of $D_2(P^{T,N})$ can only be efficient w.r.t. $\mathcal{M}^{T,N}$ if $P^{T,N} \in \mathcal{M}_2^{T,N}$.

7.2.4 Network structures covered by the statistical models considered in this article

Network structures covered by the larger model $\mathcal{M}^{T,N}$

Note that assumption 7.4 does not restrict the network structure. The network structures covered by model $\mathcal{M}^{T,N}$ are therefore those that satisfy assumption 7.3.

Example 1: finite memory, bounded number of contacts. Consider the setting where $L(t,i)$ is allowed to depend on $L(t,i)$ only through a summary measure of the history over last t_0 steps of a set $F_L(t,i)$ of at most N_0 friends. Then, if we allow the dimension of $\mathcal{C}_L$ to be as large as $(2t_0+1)N_0$, assumption 7.3 holds for the summary measure

$$\begin{aligned} c_{L(t,i)}(l(s,i)^-) := ((&a(s,j) : s = t-t_0,\dots,t,\ j \in F_L(t,i)), \\ &(l(s,j) : s = t-t_0,\dots,t-1,\ j \in F_L(t,i)))\,. \end{aligned}$$

Example 2: finite memory, dependence on aggregate measures only. Consider the set of distributions $P^{T,N}$ such that $L(t,i)$ depends on a finite set of aggregate measures of the trial's history observed before the nodes $L(t)$. Consider for example, in the infectious disease example, and suppose that the intervention $A(t,i)$ is whether the i wears a mask at t. Such aggregate measures could include summaries of $\bar{L}(t-1)$ such as the average infection rate across the entire population at time steps $t-t_0,\dots,t-1$, the average infection rate among individuals i has been in contact with at time steps $t_0,\dots,t-1$. Aggregate summary measures of $\bar{A}(t)$ could include the fraction of people wearing masks in the population at $t = t-t_0,\dots,t$, and the number of individuals at time steps $t = t-t_0,\dots,t$ not wearing masks that i has been in contact with. Note that in this setting, we can build a summary function mapping histories into a fixed dimensional set $\mathcal{C}_L$ without imposing restrictions on the number of contacts of each individuals.

Networks structures covered by the model $\mathcal{M}_1^{T,N}$

Example 3: disjoint independent clusters modelled by MDPs. Suppose that $H_1,\dots,H_n$ form a partition of $[N]$, and that there exists a constant N_0 such that $|H_k| \leq N_0$ for every k. We say that each H_k is a *cluster*. For any cluster H, denote $A(t,H) := (A(t,i) : i \in H)$, $L(t,H) := (L(t,i) : i \in H)$, $O(t,H) := O(t,i) : i \in H)$. For any i, let $k(i)$ be the cluster i belongs to. Suppose that

$$q(L(t,i) \mid L(t,i)^-) = q(L(t,i) \mid A(t,H_{k(i)}), L(t-1,H_{k(i)})),$$

that is, $L(t,i)$ depends on the nodes preceding it only through the latest treatment vector $A(t, H_{k(i)})$ of the individuals in the same cluster, and on the latest measurement vector $L(t, H_{k(i)})$ of individuals in the cluster. Suppose that

$$g^*_{t,i}(a(t,i) \mid a(t,i)^-) = g^*_{t,i}(a(t,i) \mid l(t-1, H_{k(i)})),$$

that is under the counterfactual intervention, $A(t,i)$ depends only on the latest vector of measurement vector $L(t-1, H_{k(i)})$ of the individuals in the same cluster as i. Let

$$c_{L(t,i)}(l(t,i)^-) = \left((l(t,j) : j < i, j \in H_{k(i)}), a(t, H_{k(i)}), l(t-1, H_{k(i)})\right).$$

Then it is immediate that assumption 7.3 holds. Since in general $A(t, H_{k(i)})$, $L(t, H_{k(i)})$ do not block the dependence between the nodes $\{L(t,j) : j < i, j \in H_{k(i)}\}$ and $L(\tau, k)$, for $k \in H_{k(i)}$, $\tau > t$, we include these in the context summary measure to ensure that $L(\tau,k) \perp\!\!\!\perp L(t,i)^- \mid C^L(t,i)$. It is then straightforward to check that assumption 7.5 holds. As a result, the class of network structures described in this example is covered by model $\mathcal{M}_1^{T,N}$.

Note that the network structure under the observed treatment mechanism g and the counterfactual treatment mechanism g^* do not need to be the same. Note that the assumptions defining model $\mathcal{M}_1^{T,N}$ do not place any restriction on how $A(t,i)$ might depend on the past under g.

Example 4: treatment limits social interactions, g^* forces individuals to stay in clusters. Let $H_1, \ldots, H_n$ be disjoints clusters of at most N_0 individuals forming a partition of $[N]$. In our infectious disease example, we take these clusters to be households. We define the treatment as follows: $A(t,i) = 1$ if individual i can meet with people outside of her household at time t, and $A(t,i) = 0$ if not. Regardless of treatment status, we suppose that $L(t,i)$ depends on the nodes $L(t,i)^-$ preceding it only through the history over the latest t_0 time steps of a set $F_L(t,i)$ of a most $N_1 \geq N_0$ individuals. We further suppose that $L(t,i)$ can only depend on the history over the last t_0 time steps of individuals i is allowed to meet. Define the censoring indicator

$$\Delta(t,i,j) = \mathbf{1}\left(\{a(t,i) = 1 \text{ and } j \in F_L(t,i)\} \text{ or } \left\{a(t,i) = 0 \text{ and } j \in H_{k(i)}\right\}\right),$$

and the history summary mapping

$$\begin{aligned} c_L(t,i)(l(t,i)^-) := (&((l(t,j)\Delta(t,i,j), \Delta(t,i,j)) : j < i, j \in F_L(t,i)), \\ &((a(s,j)\Delta(s,i,j), \Delta(s,i,j)) : j \in F_L(t,i), s = t - t_0, \ldots, t), \\ &((l(s,j)\Delta(s,i,j), \Delta(s,i,j)) : j \in F_L(t,i), s = t - t_0, \ldots, t-1)). \end{aligned}$$

Under the intervention g^* that deterministically assigns $A(s,j) = 0$ to every individual $j \in [N]$ at every time point $s \in [\tau]$, it is straightforward to check that the above defined finite dimensional summary measure mapping verifies assumptions 7.3 and 7.5.

Network structures covered by the model $\mathcal{M}_2^{T,N}$

Example 5: N independent MDPs under g^*. Consider our second motivating example in which the administrators of a web platform want to identify the treatment arm a that has highest long term outcome. Formally, let $\tau \geq 1$ be a time point at which we deem the outcome to be a "long-term" outcome, and for each arm $a = 1, 2$, let $g^{*,a}_{t,i}(a(t,i) \mid a(t,i)^-) := \mathbf{1}(a(t,i) = a)$, the intervention that always assigns deterministically arm a. We define the long terms outcomes of each arm as $\Psi(a)(P^{T,N}) := E_{q,g^{*,a}}[Y(\tau)]$. Suppose that

$$\begin{aligned} q(l(t,i) \mid l(t,i)^-) =& q(l(t,i) \mid a(t,i), l(t-1,i), \\ \text{and } g^*_{t,i}(a(t,i) \mid a(t,i)^-) =& g^*_{t,i}(a(t,i) \mid l(t,i)), \end{aligned}$$

that is, under g^*, individual trajectories are independent MDPs. Suppose further that the observed treatment mechanism satisfies

$$g_{t,i}(a(t,i) \mid a(t,i)^-) = g_{t,i}(a(t,i) \mid l(t,i), \theta(t)),$$

where $\theta(t) \in \mathbb{R}^{d_3}$ is a summary measure of the entire trial's history $\bar{o}(t-1)$ which contains the parameters of the design. In this case, assumption 7.6 is satisfied for $c^{g,g^*}_A(t,i) := (a(t,i), \theta(t))$.

It is straightforward to check that assumptions 7.3, 7.5 and 7.7 then hold.

We now discuss what type of adaptive designs can satisfy the constraint expressed in the previous display.

If the goal of the experimenter is to minimize regret, appropriate adaptive designs might include some type of variant of UCB, or some type of ε-greedy design. In the UCB case, so as to parameterize the design at time t, it suffices for $\theta(t)$ to contain estimates $(\widehat{\Psi}_t(a) : a = 1, 2)$ of the long term outcomes under each arms, and of the standard deviations, which we denote $(\widehat{\sigma}_t(a) : a = 1, 2)$ of these estimates. In the case of an ε-greedy design, $\theta(t)$ needs only to contain $(\widehat{\Psi}_t(a) : a = 1, 2)$. If the goal

If the goal of the experimenter is to maximize the efficiency of an estimator of the contrast $\Psi(2)(P^{T,N}_0) - \Psi(1)(P^{T,N}_0)$, an appropriate design might some type of Neyman allocation design (see e.g. van der Laan [2008]). Such a design can be defined based on estimates $(\widehat{\sigma}_t(a) : a = 1, 2)$ of the standard deviations of the canonical gradients of $\Psi(1)$ and $\Psi(2)$.

7.2.5 Comparison with the statistical model studied in past works

van der Laan et al. [2018] and Kallus and Uehara [2019] consider single individual trajectories or multiple independent single trajectories, that is they work in the case $N = 1$. The model studied in van der Laan et al. [2018] is $\mathcal{M}^{T,N}$ under $N = 1$. The homogeneous MDP model studied in Kallus and Uehara [2020] is a special case of the model $\mathcal{M}^{T,N}_2$, in the case $N = 1$.

We point out that neither of these two works provide a formal proof of the derivation of the canonical gradient of their target parameters w.r.t. the statistical models they consider. These can be obtained from the results of this article.

van der Laan [2013] and Ogburn et al. [2020] study a more general setting where g is unknown and $q_{t,i}$ is not assumed to be constant across time points. This means that their statistical model contains $\mathcal{M}^{T,N}$. Note that the the canonical gradient of Ψ w.r.t. their larger model is not equal to the canonical gradient of Ψ w.r.t. $\mathcal{M}^{T,N}$. We point out nevertheless that the derivation of the canonical gradient of Ψ w.r.t. $\mathcal{M}^{T,N}$ follows from a straightforward adaptation of the proof technique of van der Laan [2013].

7.3 Structural properties of our target paremeter

7.3.1 Efficient influence function

In the upcoming theorem, we present the canonical gradient D of Ψ w.r.t. $\mathcal{M}^{T,N}$. We also provide two simplified representations of D when $P^{T,N}$ is in $\mathcal{M}_1^{T,N}$, and in $\mathcal{M}_2^{T,N}$, respectively. As pointed out in the previous section, we stress out that these are simplified representation of the canonical gradient w.r.t. $\mathcal{M}^{T,N}$ when $P^{T,N}$ belongs to submodels of $\mathcal{M}^{T,N}$, and not the expressions of the canonical gradient w.r.t. these submodels.

Let $h_{t,i}^L$ and $h_{t,i}^{*,L}$, be the marginal densities of $C^L(t,i)$ under $P^{T,N}$ and the corresponding G-computation formula $P_{g^*}^{\tau,N}$, and let $\bar{h}_{T,N}^L := (TN)^{-1}\sum_{t=1}^N\sum_{i=1}^N h_{t,i}^L$ Under assumption 7.6, the contexts are $C_A^{g,g^*}(t,i)$ is defined. We then denote $h_{t,i}^A$ and $\bar{h}_{t,i}^*$ the marginal densities of $C_A^{g,g^*}(t,i)$ under $P^{T,N}$ and $P_{g^*}^{\tau,N}$, respectively, and we let $\bar{h}_{T,N}^A := (TN)^{-1}\sum_{t=1}^T\sum_{i=1}^N h_{t,i}^A$. Since we will refer more often to $h_{t,i}^{*,L}$ and $\bar{h}_{T,N}^L$, than to the other marginal densities, we will often simply denote them $h_{t,i}^*$ and $\bar{h}_{T,N}$.

Theorem 7.1 (Representation of the canonical gradient)**.** *The canonical gradient of Ψ w.r.t. $\mathcal{M}^{T,N}$ at $P^{T,N}$ is given by*

$$D(q)(o^{T,N}) = \frac{1}{TN}\sum_{t=1}^T\sum_{i=1}^N \bar{D}_{T,N}(q)(c^L(t,i), l(t,i)),$$

where, for any c^L and l,

$$\bar{D}_{T,N}(c^L, l) = \sum_{s=1}^{\tau}\sum_{j=1}^N \frac{h_{t,i}^L(c^L)}{\bar{h}_{T,N}^L(c^L)}\left\{E_{q,g}\left[Yg^*/g \mid L(t,i) = l, C^L(t,i) = c^L\right]\right.$$
$$\left. -E_{q,g}\left[Yg^*/g \mid C^L(t,i) = c^L\right]\right\}.$$

If $P^{T,N} \in \mathcal{M}_1^{T,N}$, then we can represent $\bar{D}_{T,N}$ as follows:

$$\bar{D}_{T,N}(c^L, l) = \sum_{s=1}^{\tau}\sum_{j=1}^N \frac{h_{t,i}^{*,L}(c^L)}{\bar{h}_{T,N}^L(c^L)}\left\{E_{q,g^*}\left[Y \mid L(t,i) = l, C^L(t,i) = c^L\right]\right.$$
$$\left. -E_{q,g^*}\left[Y \mid C^L(t,i) = c^L\right]\right\}. \tag{7.3}$$

Furthermore, if $P^{T,N} \in \mathcal{M}_2^{T,N}$, then the following representation of $\bar{D}_{T,N}$ holds:

$$\bar{D}_{T,N}(q)(c^L, l) = \sum_{s=1}^{\tau}\sum_{j=1}^N \omega_{s,j}(c^A)\eta_{s,j}(a \mid c^A)\left\{E_{q,g^*}[Y \mid L(t,i) = l, A(t,i) = a, C^A(t,i) = c^A]\right.$$

$$\left. -E_{q,g^*}[Y \mid A(t,i) = a, C^A(t,i) = c^A]\right\}, \tag{7.4}$$

where $\omega_{s,j} := h_{t,i}^{*,A}/\bar{h}_{T,N}^A$, and $\eta_{s,j} := g_{s,j}^*/g_{s,j} = g_{s,j}^*/g_{s,j}$ (since under assumption 7.8, $g_{s,j} = g$ for some g common across values of s and j).

7.3.2 First order expansion and robustness properties

Let $P^{T,N} \in \mathcal{M}^{T,N}$. Denote $q = q(P^{T,N})$ and $q_0 = q(P_0^{T,N})$. Let

$$\begin{aligned} R(q, q_0) :=& \Psi(P^{T,N}) - \Psi(P_0^{T,N}) + E_{P_0^{T,N}} \left[D(q)(O^{T,N}) \right] \\ =& \Psi^{(1)}(q) - \Psi^{(1)}(q_0) + E_{P_0^{T,N}} \left[D(q)(O^{T,N}) \right]. \end{aligned}$$

We like to view the equivalent representation

$$\Psi(P^{T,N}) - \Psi(P_0^{T,N}) = -E_{P_0^{T,N}} \left[D(q)(O^{T,N}) \right] + R(q, q_0)$$

as a functional first order Taylor expansion of the difference $\Psi(P^{T,N}) - \Psi(P_0^{T,N})$, in which we view $R(q, q_0)$ as a remainder term, which we will show is second order. We say that a remainder term $R'(q, q_0)$ is second order if it can be written as a sum of terms such that every term has a factor of the form $\prod_k (\eta_k(q) - \eta_p(q))^{\alpha_k}$, with $\sum_k \alpha_k \geq 2$. In the usual sense, we say that a remainder term $R'(q, q_0)$ is robust (or equivalently we say that the canonical gradient from which it is formed is robust) if it can be rewritten as $R'_1((\eta_1(q), \ldots, \eta_p(q)), (\eta_1(q_0), \ldots, \eta_p(q_0))$, with $\eta_1(q), \ldots, \eta_p(q)$ variation independent nuisance parameters, and is equal to zero is $\eta_i(q) = \eta_i(q_0)$ for every i in a subset $\mathcal{I} \subset [p]$, $\mathcal{I} \neq [p]$. Note that if a remainder term is second order w.r.t. variation independent parameters, then it is robust in the usual sense.

Unfortunately, the canonical gradient of Ψ w.r.t. $\mathcal{M}^{T,N}$ is not robust in the usual sense, but we can show that it is second order and robust in a weaker sense, in which the nuisance $\eta_1(q), \ldots, \eta_p(q)$ are not variation independent.

We give two results that show that the remainder term R is second order and robust in this weaker sense. These two results correspond to respectively representations (7.3) and (7.4) of the canonical gradient.

For pairs of indices (t, i) and (s, j), we write $(s, j) < (t, i)$ (resp. $(s, j) > (t, i)$) if (s, j) comes strictly before (resp. strictly after) (t, i) in the column ordering of indices. For any q, we denote

$$q_{t,i} : o^{T,N} \mapsto q(l(t,i) \mid c^L(t,i)), \qquad q_{-(t,i)} : o^{T,N} \mapsto \prod_{(s,j) \neq (t,i)} q(l(s,j) \mid c^L(s,j))$$

$$q_{(t,i)-} : o^{T,N} \mapsto \prod_{(s,j)<(t,i)} q(l(s,j) \mid c^L(s,j)),$$

$$\text{and } q_{(t,i)+} : o^{T,N} \mapsto \prod_{(s,j)>(t,i)} q(l(s,j) \mid c^L(s,j)).$$

Theorem 7.2. *Suppose that $P \in \mathcal{M}_1^{T,M}$. Then, we can rewrite $R(q, q_0)$ as $R_1(\bar{h}_{T,N}, \bar{h}_{0,T,N}, q, q_0)$ where $R_1(\bar{h}_{T,N}, \bar{h}_{0,T,N}, q, q_0)$ satisfies*

$$\begin{aligned} R_1(\bar{h}_{T,N}, \bar{h}_{0,T,N}, q, q_0) =& R_1(\bar{h}_{T,N}, \bar{h}_{0,T,N}, q, q_0) - R_1(\bar{h}_{0,T,N}, \bar{h}_{0,T,N}, q, q_0) \\ &+ R_1(\bar{h}_{0,T,N}, \bar{h}_{0,T,N}, q, q_0), \end{aligned}$$

where,

$$R_1(\bar{h}_{T,N}, \bar{h}_{0,T,N}, q, q_0) - R_1(\bar{h}_{0,T,N}, \bar{h}_{0,T,N}, q, q_0)$$

$$= \sum_{s=1}^{\tau} \sum_{j=1}^{N} \int \left(h^*_{s,j} \frac{\bar{h}_{0,T,N} - \bar{h}_{T,N}}{\bar{h}_{T,N}} \right) (c)(q_0 - q)(l \mid c) \times E_{q,g^*}[Y \mid L(t,i) = l, C(t,i) = c] dldc,$$

and

$$R_1(\bar{h}_{0,T,N}, \bar{h}_{0,T,N}, q, q_0) = \sum_{s=1}^{\tau} \sum_{j=1}^{N} E_{q_{(s,j)-},(q-q_0)_{(s,j)},q_{0,(s,j)+}-q_{(s,j)+},g^*} Y.$$

If $P^{T,N} \in \mathcal{M}_2^{T,N}$, we can further simplify the representation of the remainder term, as the following theorem shows.

Theorem 7.3. *Suppose that $P \in \mathcal{M}_2^{T,N}$. Denote $\omega = (\omega_{s,j} : s \in [\tau], j \in [N])$. We can then rewrite $R(q, q_0)$ as $R_2(\omega, \omega_0, q, q_0)$ where the latter satisfies that*

$$R_2(\omega, \omega_0, q, q_0) = \sum_{s=1}^{\tau} \frac{1}{N} \sum_{j=1}^{N} \int \bar{h}^A_{0,T,N}(c^A) g^*_{s,j}(a \mid c^A)(\omega_{s,j} - \omega_{0,s,j})(c^A)(q - q_0)(l \mid a, c^A)$$
$$\times E_{q,g^*} \left[Y(j) \mid L(s,j) = l, A(s,j) = a, C^A(s,j) = c^A \right] dldadc^A.$$

From the expression above, $R_2(\omega, \omega_0, q, q_0)$ if $\omega = \omega_0$ or $q = q_0$.

Remark 7.4. *In the above theorem, ω and q are not variation independent components of $P^{T,N}$. In fact, since we know g, ω is fully determined by q.*

Remark 7.5. *The proof of theorem 7.3 relies on the fact that we know the treatment mechanism g since we are in a controlled trial, while the proof of theorem 7.2 does not.*

7.4 Construction and analysis of our estimators

Let $\widehat{q}_{T,N}$, be an estimator of q_0.

TMLE estimator. Let $\widehat{q}^*_{T,N}$ be an estimator of q_0 obtained from $\widehat{q}_{T,N}$ by the TMLE targeting step such that it solves approximately the EIF equation:

$$\frac{1}{TN} \sum_{t=1}^{T} \sum_{j=1}^{N} \bar{D}_{T,N}(\widehat{q}^*_{T,N})(L(t,i), C(t,i)) = o((TN)^{-1/2}).$$

We refer the reader to the Targeted Learning methodology papers and books [van der Laan and Rubin, 28 Dec. 2006, Van der Laan and Rose, 2011, van der Laan and Gruber, 2016, Van der Laan and Rose, 2018] for details on the TMLE targeting steps.

We define our TMLE estimator as the following plug-in estimator:

$$\widehat{\Psi}^{\text{TMLE}}_{T,N} := \Psi(\widehat{q}^*_{T,N})$$

One-step estimator. The one-step estimator is defined as

$$\widehat{\Psi}_{T,N}^{1-\text{step}} := \Psi(\widehat{q}_{T,N}) + \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\bar{D}_{T,N}(\widehat{q}_{T,N})(L(t,i),C(t,i)).$$

In what follows, we restrict our analysis to the TMLE estimator since the analysis for the 1-step estimator is identical. We will just denote $\widehat{\Psi}_{T,N} := \widehat{\Psi}_{T,N}^{\text{TMLE}}$. The following theorem gives a decomposition of the difference between $\widehat{\Psi}_{T,N}$ and its target $\Psi(q_0)$.

Theorem 7.4 (TMLE expansion)**.** *We have that*

$$\widehat{\Psi} - \Psi(q_0) :=M_{1,T,N}(q_0) + M_{2,T,N}(\widehat{q}^*_{T,N}, q_0) + R(\widehat{q}^*_{T,N}, q_0),$$

with

$$\begin{aligned}M_{1,T,N}(q_0) =&\frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\bar{D}_{T,N}(q_0)(L(t,i),C(t,i))\\ M_{2,T,N}(q,q_0) =&\frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}(\delta_{L(t,i),C(t,i)} - P_{q_0,h_{0,t,i}})\left(\bar{D}_{T,N}(q) - \bar{D}_{T,N}(q_0)\right),\end{aligned}$$

and $R(q,q_0)$ is as defined in section 7.3 above.

The first term is the sum of a martingale difference sequence, and the process

$$\{x\sqrt{TN}M_{1,xT,N}(q_0) : x \in [0,1]\}$$

can be shown, using a functional central limit theorem for martingale triangular arrays, to converge weakly, as $T \to \infty$ and under fixed N, to a Wiener process $\sigma^2_{0,\infty,N}W$, with W a standard Wiener process and $\sigma^2_{0,\infty,N}$ the limit of the variance under a certain limit distribution, of $\lim_{T\to\infty}\bar{D}_{T,N}(q_0)$ (we make precise these limits further down). Note that, as mentioned above, it is not immediately clear that the variance of $\bar{D}_{T,N}(q_0)$ doesn't diverge as $N \to \infty$. We provide in section 7.4.1 below conditions under which $\bar{D}_{T,N}(q_0)$ remains finite as $N \to \infty$.

The second term can be bounded by the supremum of the process $\{M_{2,T,N}(q,q_0) : q \in \mathcal{M}_Q\}$. This process is an empirical process generated by the sequence $(X(t,i))_{t,i}$, where we define that $X(t,i) := (C(t,i), L(t,i))$. So as to analyze this term, we prove a maximal inequality for such processes which holds under mixing conditions (here on the sequence $(X(t,i))_{t,i}$) and entropy conditions. This maximal inequality will allow us to show the negligibility of $M_{2,T,N}(\widehat{q}^*_{T,N}, q)$ in front of the first term.

We will discuss the negligibility of the remainder term $R(\widehat{q}^*_{T,N}, q_0)$ under convergence rate conditions on $\widehat{q}^*_{T,N}$

7.4.1 Boundedness of the canonical gradient

We can rewrite $\bar{D}_{T,N}(q_0)$ as

$$\bar{D}_{T,N}(q_0)(l,c) = \sum_{s=1}^{\tau} \frac{1}{N} \sum_{j=1}^{N} \frac{h^*_{0,s,j}}{\bar{h}_{0,T,N}}(c) \widetilde{D}_{s,j,N}(l,c),$$

with

$$\widetilde{D}_{s,j,N}(q)(l,c) := \sum_{k=1}^{N} E_{q,g^*}\left[Y(k) \mid L(s,j) = l, C(s,j) = c\right] - E_{q,g^*}\left[Y(k) \mid C(s,j) = c\right].$$

A sufficient condition for $\bar{D}_{T,N}(q_0)$ to remain bounded as $N \to \infty$ is that the terms $\widetilde{D}_{s,j,N}(q)$ themselves remain bounded as $N \to \infty$. It is immediate to observe that for $s = \tau$, since $L(s,j) \perp\!\!\!\perp L(\tau,k) \mid C(s,j)$, and since $Y(k)$ is a a component of $L(\tau,j)$, we must have that

$$E_{q,g^*}\left[Y(k) \mid L(s,j) = l, C(s,j) = c\right] - E_{q,g^*}\left[Y(k) \mid C(s,j) = c\right] = 0$$

for every $j \neq k$, and therefore $\|\widetilde{D}_{\tau,j,N}(q)\|_\infty \leq 1$.

If we don't make any assumption on g^*, and that we just assume that under g^*, $A(s,1), \ldots,$ $A(s,N)$ are conditionally independent given $A(t)^-$, but can a priori depend on the entire past $A(t)^-$, then, if $j \neq k$, we don't have the same kind of conditional independence between $Y(k)$ and $L(s,j)$, for $s \leq \tau - 1$ as we have in the case $s = \tau$. As a result, we don't have the same cancellations as in the case $s = \tau$. Intuitively, for $\|\widetilde{D}_{s,j,N}(q)\|_\infty$ not to diverge as $N \to \infty$, we need some measure of association between $Y(k)$ and the nodes $O(s,1), \ldots, O(s,N)$ to remain controlled in some sense. A natural measure of association that can be used to formulate rigorously this requirement is the classical notion of φ-mixing coefficient (see Bradley [2005] for a survey of usual mixing coefficients), which we restate here in terms of densities.

Definition 7.2 (φ-mixing)**.** *For any two random variables $(X,Y) \sim P$, we define the φ-mixing coefficient between X and Y as*

$$\varphi_P(X,Y) := \sup\{|p_{Y|X}(y \mid x) - p_X(x)| : y, x, \text{ such that } p_X(x) > 0\},$$

where $p_{Y|X}$ and p_X are the conditional densities of Y given X and the marginal density of X w.r.t. an appropriate known dominating measure.

We now provide a generic condition under which $\|\widetilde{D}_{s,j,N}(q)\|_\infty$, and therefore $\|\bar{D}_{T,N}(q)\|_\infty$ are controlled. We introduce the short-hand notation φ_{q,g^*} for $\varphi_{P^{\tau,N}_{g^*}}$, where we recall that $P^{\tau,N}_{g^*}$ is the G-computation formula obtained from $P^{T,N}$.

Assumption 7.9. *Suppose that there exists $\boldsymbol{\varphi} < \infty$ such that, for any $s \in [\tau]$ and $k \in [N]$,*

$$\sum_{j=1}^{N} \varphi_{q,g^*}(Y(k) \mid O(s,j)) \leq \boldsymbol{\varphi}.$$

Under assumption 7.9, it is easy to show the following result.

Lemma 7.1. *Suppose that assumption 7.9 holds. Then,* $\|\widetilde{D}_{s,j,k}(q)\|_\infty \leq 2\boldsymbol{\varphi}$.

A sufficient condition for $\bar{D}_{T,N}(q_0)$ to be bounded is then a bound on the marginal density ratios.

Assumption 7.10 (Marginal density ratios bound). *Suppose that there exists* $B > 0$ *such that*

$$\left\| h^*_{0,s,j} / \bar{h}_{0,T,N} \right\|_\infty \leq B$$

for every $s \in [\tau]$ *and every* $j \in [N]$.

Lemma 7.2. *Suppose that assumptions 7.9 and 7.10 hold. Then* $\|\bar{D}_{T,N}(q_0)\|_\infty \leq 2\tau B\boldsymbol{\varphi}$.

While it might be hard to check that assumption 7.9 holds in practice, we see the value of it and of lemmas 7.1 and 7.2 in that they show explicitly the nature of a condition that is sufficient for $\bar{D}_{T,N}$ to remain bounded, that is a mixing condition controlling the level of association within the graph, across time points and individuals. We now discuss a few concrete examples where we can directly show that assumption 7.9 holds, or where we think it is reasonable to suppose that it holds.

Example 1. If $P_0^{T,N} \in \mathcal{M}_2^{T,N}$, then, under the G-computation formula distribution $P_{0,g^*}^{\tau,N}$, any two distinct trajectories $\bar{O}(\tau, i)$ and $\bar{O}(\tau, j)$ are independent.

Therefore $\sum_{j=1}^N \varphi_{q,g^*}(Y(k) \mid O(s,j)) = \varphi_{q,g^*}(Y(k) \mid O(s,k)) \leq 1$, and thus $\|\widetilde{D}_{s,j,N}\|_\infty \leq 1$. (We can also directly check that all terms except the k-th one cancel out in $\widetilde{D}_{s,j,N}$).

Example 2. Suppose now that $g^*_{s,j}(A(s,j) \mid A(s)^-) = g^*_{s,j}(A(s,j) \mid C^*_A(s,j))$, with $C^*_A(s,j) := c^*_{A(s,j)}(\{\bar{O}(s,j) : j \in F_A(s,j)\})$, where $F_A(s,j)$ is a set of at most N_0, individuals, including j itself, and where $c^*_{A(s,j)}$ is a known summary function. In words, we are supposing here that the treatment decision under g^* for individual j at time s depends only on the history up to $s-1$ of j and of a set of at most N_0 individuals. Then any $Y(k)$ is associated with at most N_0 nodes from time point $\tau - 1$, which are then in turn each associated with at most N_0 nodes from time point $\tau - 2$, and so on. Therefore, any $Y(k)$ is associated with at most $N_0^{\tau - s}$ nodes from time point s, and therefore $\sum_{j=1}^N \varphi_{q,g^*}(Y(k) \mid O(s,j))$ has at most $N_0^{\tau-s}$ non zero terms, which implies that $\|\widetilde{D}_{s,j,N}(q_0)\|_\infty \leq N_0 \tau^{\tau - s}$, and thus $\|\bar{D}_{T,N}(q)\|_\infty \leq B\tau N_0^\tau$.

While this upper bound can quickly explode as τ gets large, this shows that for fixed τ, the variance does not depend on N, and that therefore, under mixing conditions on the sequence $(O(t,i))_{t,i}$ that we study in the following subsection, the asymptotic variance of our estimators can scale as N^{-1}.

Example 3. Suppose now that $g^*_{s,j}(A(s,j) \mid A(s)^-) = g^*_{s,j}(A(s,j) \mid \theta_N(s-1))$, where $\theta_N(s-1) = \frac{1}{N}\sum_{j=1}^N f(L(s,j))$ for a certain f. As $\theta_N(s-1)$ concentrates, and should be almost constant for large N, we expect that since treatment decisions depend on the past only through this almost constant $\theta_N(s-1)$, treatment assignment dependence on the past should not introduce too much dependence between units. In this situation, we conjecture that most of the dependence within the graph happens through the dependence of nodes $L(s,j)$ on the contacts $F_L(s,j)$ of j. We have seen in the previous example that if this is the main source of dependence, we should have $\|\bar{D}_{T,N}(q_0)\| \lesssim \tau B N_0^\tau$, provided that $|F_L(s,j)| \leq N_0$ for every s and j.

In our infectious disease example, the setting described here can model the situation where the intervention g^* is to restrict, depending on the global infection rate $\theta_N(s)$, the set of individuals any individual j is allowed to meet

7.4.2 Weak invariance principle for the martingale term

It is immediate to observe that $E_{q_0,h_{0,t,i}}[\bar{D}_{T,N}(q_0)(L(t,i), C(t,i) \mid C(t,i)] = 0$, and therefore, $xTM_{1,xT,N}(q_0)$ is the sum of a martingale difference sequence. We will analyze the weak convergence properties of the process $\{M_{1,xT,N}(q_0) : x \in [0,1]\}$ via a classic functional central limit theorem for martingale triangular arrays, which we recall below.

Theorem 7.5 (Theorem 3.2 McLeish [1974]). *Suppose that $\{X_{n,i} : 1 \leq i \leq n\}$ is a martingale difference array, and (k_n) is a sequence of non-decreasing, right continuous, integer valued functions, such that for every n, $k_n(0) = 0$. Let, for any $x \in [0,1]$ $W_n(x) := \sum_{i=1}^{k_n(x)} X_{n,i}$. Suppose that, for every $x \in [0,1]$*

$$\max_{i \leq k_n(x)} |X_{n,i}| \xrightarrow{L_2} 0, \tag{7.5}$$

and

$$\sum_{i=1}^{k_n(x)} X_{n,i}^2 \xrightarrow{P} x. \tag{7.6}$$

Then $W_n \xrightarrow{d} W$ in $\mathbb{D}([0,1])$.

We will apply the above result by rewriting $M_{1,xT,N}(q_0)$ as the sum of a martingale difference triangular array, as we will make explicit in the proof. A key step ahead of applying theorem 7.5 is to show that the variance under $P_{q_0,h_{0,t,i}}$ of $\bar{D}_{T,N}(L(t,i), C(t,i))$ stabilizes as $T, t \to \infty$. We provide below a set of conditions under which it is the case.

Assumption 7.11. *For every N, there exists $h_{0,\infty,N}$ such that, for every t, i, $\|\bar{h}^{-1}_{0,T,N} - h^{-1}_{0,\infty,N}\|_{2,h_{0,t,i}} \to 0$ as $T \to 0$, and there exists $B > 0$ such that $\|h^*_{0,s,j}/h_{0,\infty,N}\|_\infty \leq B$.*

Assumption 7.12. *For every i, $C(t,i) \xrightarrow{d} C_\infty \sim h_{0,\infty,N}$ as $t \to \infty$.*

The first part of assumption 7.11, and assumption 7.12 are ergodicity/mixing conditions. Under these assumptions, we can show the following result on the limit of

$$\mathrm{Var}_{q_0,h_{0,t,i}}(\bar{D}_{T,N}(q_0)(L(t,i),C(t,i)).$$

Lemma 7.3 (Stabilization of the variance of the main term of the EIF). *Suppose that assumptions 7.9, 7.10, 7.11 and 7.12 hold. Denote*

$$\bar{D}_{0,\infty,N}(l,c) := \sum_{s=1}^{\tau} \frac{1}{N} \sum_{j=1}^{N} \frac{h^*_{0,s,j}}{h_{0,\infty,N}}(c)\widetilde{D}_{s,j,N}(l,c),$$

and

$$\sigma^2_{0,\infty,N} := \mathrm{Var}_{q_0,h_{0,\infty,N}}\left(\bar{D}_{0,\infty,N}(L_\infty,C_\infty)\right),$$

Then $\sigma_{0,\infty,N} < \infty$, and

$$\mathrm{Var}_{q_0,h_{0,t,i}}\left(\bar{D}_{T,N}(q_0)(L(t,i),(C(t,i))\right) \to \sigma^2_{0,\infty,N},\ \textit{as}\ T,t \to \infty.$$

For any t, i, let $X(t,i) := (C_L(t,i), L(t,i))$. A key requirement for our analysis of the process $\{M_{1,xT,N}(q_0) : x \in (0,1]\}$ is an α-mixing condition on the sequence $(X(t,i))_{t,i}$. We first recall the notion of α-mixing. We give here a definition based on theorem 4.4 in Bradley [2007].

Definition 7.3 (α-mixing). *Consider a couple of random variables $(X,Y) \sim P$, with marginals P_X and P_Y and domains $\mathcal{X}$ and $\mathcal{Y}$. The α-mixing coefficient between X and Y is defined as*

$$\alpha_P(X,Y) := \sup\left\{\frac{\mathrm{Cov}(f(X),g(Y))}{\|f\|_\infty \|g\|_\infty}, f : \mathcal{X} \to \mathbb{R}, g : \mathcal{Y} \to \mathbb{R}, \|f\|_\infty < \infty, \|g\|_\infty < \infty\right\}.$$

We state our α-mixing condition below.

Assumption 7.13 (α-mixing). *It holds that*

$$\sum_{t_1,t_2\in[T]} \sum_{i_1,i_2\in[N]} \alpha_P(X(t_1,i_1),X(t_2,i_2)) = o(TN).$$

Theorem 7.6 (Weak convergence of the martingale term M_1). *Suppose that assumptions 7.9, 7.10, 7.11, 7.12 and 7.13 hold. Then, for any fixed N, as $T \to \infty$,*

$$\{M_{1,xT,N}(q_0) : x \in [0,1]\} \xrightarrow{d} \sigma_{0,\infty,N} W$$

on the set $\mathbb{D}([0,1])$ of cadlag functions on $[0,1]$, where $\sigma_{0,\infty,N} \leq C$ for some $0 < C < \infty$ that does not depend on N, and where W is a standard Wiener process.

7.4.3 Analysis of the empirical process term under mixing conditions

Recall that we defined $X(t,i) := (C_L(t,i), L(t,i))$. The process $\{M_{2,T,N}(q,q_0) : q \in \mathcal{Q}\}$ is an empirical process generated by the sequence of dependent observations $(X(t,i))_{t,i}$.

For there to be a hope of controlling the deviations of this process from its mean, we must impose conditions on the amount of dependence between terms of the sequence $(X(t,i))_{t,i}$. As in the analysis of the term $M_{1,xT,N}$, we impose mixing conditions. Let $(\widetilde{X}(k))_k$ be the single-index sequence obtained by reordering the terms of the double-index sequence $(X_{t,i})_{t,i}$ in colunm order, that is $(\widetilde{X}_k)_k$ is the sequence

$$X(1,1), \dots, X(1,N), \dots, X(T,1), \dots, X(T,N).$$

We define $(\widetilde{C}_L(k))_k$ and $(\widetilde{L}(k))_k$ similarly. We state our mixing conditions in terms of the sequence $(\widetilde{X}(k))_k$.

Assumption 7.14 (Geometric α-mixing)**.** *There exists $c > 0$ such that the α-mixing coefficients of $(\widetilde{X}(k))_{k\geq 1}$ satisfy $\alpha(n) \leq \exp(-cn)$.*

The next assumption is a ρ-mixing condition on the sequence $(\widetilde{X}(k))$. We state below the definition of ρ-mixing. We refer the reader to Bradley [2005] for more details on mixing coefficients.

Definition 7.4 (ρ-mixing)**.** *Consider a couple of random variables $(X,Y) \sim P$, with marginals P_X and P_Y The maximum correlation coefficient between X and Y is defined as $\rho_P(X,Y) := \sup\{\mathrm{Corr}(f(X), g(Y)) : f \in L_2(P_X), g \in L_2(P_Y)\}$.*

Assumption 7.15 (ρ-mixing)**.** *The ρ-mixing coefficients of $(\widetilde{X}(k))$ have finite sum, that is $\sum_{n=1}^{\infty} \rho(n) := \boldsymbol{\rho} < \infty$.*

The main result of this section is an almost sure equicontinuity result, which will give us that $\sqrt{NT} M_{2,N,T}(q_{N,T}, q_0) = o(1)$ almost surely. As for similar equicontinuity results (see e.g. van der Vaart and Wellner [1996]) for i.i.d. empirical processes, we require two types of conditions: (1) we need that the individual terms of $M_{2,T,N}(q_{N,T}, q_0)$ converge to zero, in some sense to be made precise further down, and (2) we need a Donsker-like condition on the complexity of the class $\{\bar{D}_{T,N}(q,q_0) : q \in \mathcal{Q}\}$.

While equicontinuity results for empirical processes usually give a convergence in probability guarantee, we prove an almost sure convergence result. Almost sure convergence offers control over the entire realization of the sequence $(M_{2,T,N}(q_{N,T}, q_0))_{N,T}$, which we need in section 7.6 to design an adaptive stopping rule. As we work in a more challenging setting (mixing sequences v.s. i.i.d. sequences), and as we prove stochastic convergence in a stronger sense, we need stronger versions of the Donsker condition than in the classical equicontinuity results for empirical processes (those of van der Vaart and Wellner [1996] for example). In particular, while the classical results require convergence of some type of L_2 norm of the difference $\bar{D}_{T,N}(\widehat{q}_{T,N}) - \bar{D}_{T,N}(q_0)$, we require convergence of this difference in $\|\cdot\|_\infty$ norm. Furthermore, while in the classical results the type of stochastic convergence required is convergence in probability, here we need a form of stochastic convergence slightly stronger than almost sure convergence.

We formulate precisely our convergence requirement in the following assumption.

Assumption 7.16. *There exists a sequence of positive numbers (a_T) satisfying $a_T^{-2}(\log T)^2/\sqrt{T} = o(1)$, and $a_T^{\nu} \log T = o(1)$ for any $\nu > 0$, such that*

$$\forall \epsilon > 0,\ \exists C(\epsilon) > 0,\ P\left[\forall n \geq 1, \left\|\bar{D}_{T,N}(\widehat{q}_{T,N}) - \bar{D}_{T,N}(q_0)\right\|_\infty \leq C(\epsilon) a_n\right] \geq 1 - \epsilon.$$

We introduce in section 7.5 a large nonparametric class of d-variate functions $\mathcal{F}_d$, which is such that, in our dependent data setting, any empirical risk minimizer $\widehat{q}_{T,N}$ over any $\mathcal{Q} \subseteq \mathcal{F}_d$ satisfies an exponential deviation bound of the form $P\left[\|\widehat{q}_{T,N} - q_*\|_\infty \gtrsim n^{-\beta} + x\right] \lesssim \exp(-C(nx^\gamma)^\nu)$, with $\beta, \gamma, \nu > 0$, $1 - \gamma\beta > 0$, where q_* is a population risk minimizer over $\mathcal{Q}$. If the true transition density q_0 lies in our nonparametric class $\mathcal{Q}$, and if $\left\|\bar{D}_{T,N}(q) - \bar{D}_{T,N}(q_0)\right\|_\infty \lesssim \|q - q_0\|_\infty$, then it is straightforward to show that assumption 7.16 holds.

We now present our Donsker-like condition. Suppose that, for any k, the distribution of $\widetilde{C}(k)$ admits density $\widetilde{h}_k$ w.r.t. the Lebesgue measure. The density w.r.t the Lebesgue measure of $\widetilde{X}(k)$ is then $q_0\widetilde{h}_k$. Let $\mathcal{X}$ be such that, for any t and any i, $X(t,i)$ takes values in $\mathcal{X}$. Let σ be the norm defined, for any $f : \mathcal{X} \to \mathbb{R}$ by

$$\sigma(f) := \sup_{i \geq 1} \|f\|_{2,q_0,\widetilde{h}_i} \sqrt{1 + 2\boldsymbol{\rho}}.$$

Our Donsker-like condition is a bound on the bracketing entropy in σ norm of the canonical gradient class.

Assumption 7.17 (Donsker condition for the canonical gradient class)**.** *Let $\mathcal{D}_{T,N} := \{\bar{D}_{T,N}(q) : q \in \mathcal{Q}\}$. There exists $p \in (0,2)$ such that*

$$\log N_{[]}(\epsilon, \mathcal{D}_{T,N}, \sigma) \lesssim \epsilon^{-p}.$$

We show in section 7.5 that the nonparametric function class $\mathcal{F}_d$ we mentioned above satisfies $\log N_{[]}(\epsilon, \mathcal{F}, \sigma) \lesssim \epsilon^{-1}|\log(\epsilon)|^{2d-1}$ under mild conditions. Therefore, if the canonical gradient class $\mathcal{D}_{T,N}$ is included in $\mathcal{F}_{d'}$ for some $d' \geq 1$, then assumption 7.17 holds under the same mild conditions.

We can now state our equicontinuity result.

Theorem 7.7 (Asymptotic equicontinuity of the canonical gradient process)**.** *Suppose that assumptions 7.14, 7.15, 7.16 and 7.17 hold. Then*

$$\sqrt{NT} M_{2,T,N}(\widehat{q}_{T,N}, q_0) = o(1) \text{ a.s. as } T, N \to \infty.$$

Proof of theorem 7.7. The proof is a direct consequence of our generic equicontinuity result, theorem 7.12 in the appendix. □

Remark 7.6. *It might seem surprising to the reader familiar with proofs of equicontinuity results and maximal inequalities for empirical processes that, while the Donsker condition only requires control of the entropy w.r.t. the norm σ, which is an L_2 norm, we need convergence of $\|\bar{D}_{T,N}(\widehat{q}_{T,N}) - \bar{D}_{T,N}(q_0)\|_\infty$ in a norm a strong as the sup norm. Indeed, in the usual case where $Z_1, \ldots, Z_n$ are i.i.d. random draws from a distribution P taking values in a set $\mathcal{Z}$, if $\mathcal{F}$ has* *square integrable bracketing entropy w.r.t. $L_2(P)$, for a process of the form $n^{-1}\sum_{i=1}^n f_n(Z_i) - \int f_n(z)dP(z)$ to be $o_P(n^{-1/2})$, it suffices that the $L_2(P)$ norm of f_n converges to zero in probability.*

We discuss in subsection 7.C.2 in the appendix why, unlike in the i.i.d. setting, in the weakly dependent case we consider here, convergence in L_2 norm wouldn't suffice given the technical tools that we have, and why we do need convergence in $\| \cdot \|_\infty$ norm.

7.4.4 Weak invariance principle for our TMLE

Theorem 7.8 (Weak invariance principle for our TMLE)**.** *Suppose that the assumptions of theorems 7.6 and 7.7 are satisfied, and that $R(\widehat{q}_{T,N}, q_0) = o((NT)^{-1/2})$ almost surely. We then have that the process*

$$\left\{ t\sqrt{TN}\sigma_{0,\infty,N}^{-1} \left(\widehat{\Psi}_{tT,N} - \Psi(P_0^{tT,N}) \right) : t \in [0,1] \right\}$$

converges weakly in $\mathbb{D}([0,1])$ to a Wiener process W.

7.5 A nonparametric function class, nuisance estimation, and the canonical gradient class

We first present a generic function class, which we then use as a statistical model for nuisance estimation and canonical gradient modelling.

7.5.1 The function class

Consider a bounded Euclidean set $\mathcal{X}$. Without loss of generality, we will suppose that $\mathcal{X} = [0,1]^{d_1}$, the unit hypercube in $\mathbb{R}^{d_1}$. For $M > 0$, let $\mathcal{F}_{0,M}$ be the class of real-valued cadlag functions on $\mathcal{X}$, with sectional variation norm (also called Hardy-Krause variation) no larger than M, and, for $L > 0$, let $\mathcal{F}_{1,M,L}$ be the class of functions in $\mathcal{F}_{0,M}$ that are L-Lipschitz. The classes $\mathcal{F}_{0,M}$ and $\cup_{M>0}\mathcal{F}_{0,M}$ have been proposed as statistical models in several past articles [van der Laan, 2017, Fang et al., 2020, Bibaut and van der Laan, 2019]. We refer to these works for the rigorous definition of the notion of sectional variation norm. For the present purpose, it will suffice to say that the sectional variation norm is a multivariate extension of the 1-dimensional notion of total variation of a real-valued function.

Statistical properties of $\mathcal{F}_{0,M}$. This class of functions present several attractive properties as a statistical model. Bibaut and van der Laan [2019] have shown that its bracketing entropy is well controlled, which will prove useful in our problem. We recall here the formal result on bracketing entropy from Bibaut and van der Laan [2019].

Proposition 7.1 (Proposition 2 in Bibaut and van der Laan [2019])**.** *Consider $\mathcal{F}_{0,M}$ as defined above. For any $r \geq 1$, $\epsilon > 0$, it holds that*

$$\log N_{[]}(\epsilon, \mathcal{F}_{0,M}, L_r(\mu)) \lesssim M\epsilon^{-1} \left|\log(M/\epsilon)\right|^{2(d_1-1)},$$

where we have absorbed a constant depending on the dimension d_1 and on r in the " $\lesssim''$ notation, and where μ is the Lebesgue measure.

Notice that the entropy depends on the dimension only through the log factor. As a result, even in high dimensions this class remains Donsker, and rates of convergence of empirical risk minimizers (ERMs) over it remain relatively fast. Unlike other popular nonparametric function classes such as Holder classes, $\mathcal{F}_{0,M}$ doesn't impose local smoothness restrictions. Rather, it only places a bound on a global measure of variation, the sectional variation norm, thus allowing for function having different degrees of smoothness or roughness at different regions of their domains. As a result, from a bias-variance trade-off perspective, when one increases M by some amount, an ERM estimator will "spend" the additional allowed variation in the areas of the domain where it most improves the fit, while only impacting the entropy loglinearly. While this might not be a perfectly rigorous comparison, note that, Holder classes $H(M, \beta)$ have entropy depending on ϵ as $\epsilon^{d/\beta}$, and therefore decreasing β so as to reduce bias has a steep entropy price.

We believe that since the nonparametric model $\cup_{M>0}\mathcal{F}_{0,M}$ only assumes a form of piecewise continuity and that the sectional variation norm is not infinite, using it a statistical model for components of the data-generating distributing amounts to a mild assumption. Our guess is that functions that do not satisfy these requirements are essentially pathological functions $x \mapsto f(x)$ that oscillate increasingly fast as x approaches some value or region. Benkeser and Van Der Laan [2016] have shown with extensive simulations that ERMs over $\mathcal{F}_{0,\widehat{M}}$, with $\widehat{M}$ chosen by cross-validation, perform on par with Random Forests and Gradient Boosting Machines, thereby confirming that $\cup_{M>0}\mathcal{F}_{0,M}$ is a realistic statistical model in most practical settings.

Computational properties. Fang et al. [2020] have shown that ERMs over $\mathcal{F}_{0,M}$ can be computed as the solution of a LASSO problem over at most $(ne/d)^d$ distinct basis functions, where n is the sample size. van der Laan [2017] has proposed an alternative set of basis functions of cardinality $n2^d$, which, although it can be shown to not always be sufficient to represent the ERM, leads to very good practical performance.

Properties of $\mathcal{F}_{1,M,L}$. Introducing the additional assumption that the functions in our model are Lipschitz allows to bound the supremum norm of a function in terms of its $L_19(\mu)$ norm, as shown by the following lemma. We owe this result to Iosif Pinelis, who proved it as an answer to a question of the first author on MathOverflow [Pinelis, 2020].

Lemma 7.4. *There exists $\eta(d, L) > 0$ and $C(d, L) > 0$ such that, for any f is a d-variate, real-valued L-Lipschitz function such that $\|f\|_{1,\mu} \leq \eta(d, L)$, we have $\|f\|_\infty \leq C(d, L)\|f\|_{1,\mu}^{1/(d+1)}$, with μ the Lebesgue measure.*

Unlike in the i.i.d. setting, in our mixing data sequence setting, we will need to be able to show that the supremum norm of some functions converge to zero at a certain rate. We refer the interested reader to the proofs of the results of the next two subsections for more detail on these technical questions.

7.5.2 Nuisance estimation

The efficient influence function expression makes appear the nuisance parameters q, $(\phi_{s,j})$, $(h^*_{s,j})$ and $\bar{h}_{N,T}$. The latter are functions of q and can be computed from an estimate thereof via Monte-Carlo integration, as discussed in van der Laan et al. [2018] in the case $N = 1$. The key statistical challenge is then the estimation of the true conditional density q_0.

We propose to estimate q_0 by a maximum likelihood estimator over the subset of functions of $\mathcal{F}_{1,M,L}(\mathcal{X})$, with $\mathcal{X} := \mathcal{C} \times \mathcal{O}$, that are conditional probability density functions $(c, o) \mapsto q(o \mid c)$, that is over the set

$$\mathcal{Q}_{M,L} := \left\{ q \in \mathcal{F}_{1,M,L} : \forall c \int q(o \mid c) do = 1 \text{ and } q(\cdot \mid c) \geq 0 \right\}.$$

In practice, M and L should be chosen by cross validation. As there will be no ambiguity in the rest of this section, we use the notation $\mathcal{Q}$ instead of $\mathcal{Q}_{M,L}$. In this section too, we work with the reordered single-indexed sequence $(\widetilde{O}(k))$ as defined in the previous section.

For any conditional density $q : (o, c) \mapsto q(o \mid c)$, let $\ell(q)(c, o) := -\log q(o \mid c)$ be the log-likelihood loss for q, and let

$$\widehat{R}_n(q) := \frac{1}{n} \sum_{i=1}^{n} \ell(q)(\widetilde{C}(k), \widetilde{O}(k)) \qquad \text{and} \qquad R_{0,n}(q) := \frac{1}{n} \sum_{i=1}^{n} E[\ell(q)(\widetilde{C}(k), \widetilde{O}(k))].$$

Let $\widehat{q}_n \in \arg\min_{q \in \mathcal{Q}} \widehat{R}_n(q)$ and $q_n \in \arg\min_{q \in \mathcal{Q}} R_{0,n}(q)$ be a maximum likelihood estimator, and a maximizer over $\mathcal{Q}$ of the population log likelihood. We analyze $\widehat{q}_n$ using our generic result for ERMs under mixing sequences, theorem 7.13 in the appendix. We need the following assumptions.

Assumption 7.18 (Lower bound on the population MLE). *There exists δ independent of n such that* $\inf_{c,o \in \mathcal{C} \times \mathcal{O}} q_n(o \mid c) \geq \delta$.

Assumption 7.19. *There exists $M_1 > 0$ independent of n such that $\|q_0/q_n\|_\infty \leq M_1$.*

Assumption 7.20 (Uniform boundedness of $(\widetilde{h}_i)_{i \geq 1}$). *There exists $M_2 > 0$ such that*

$$\sup_{i \geq 1} \|\widetilde{h}_i\|_\infty \leq M_2.$$

Assumption 7.21. *Denote $\bar{\widetilde{h}}_n := n^{-1} \sum_{i=1}^{n} \widetilde{h}_i$. There exists $M_3 > 0$ independent of n such that*

$$\sup_{i \geq 1} \left\| \widetilde{h}_i / \bar{\widetilde{h}}_n \right\|_\infty \leq M_3.$$

Theorem 7.9 (High probability bound on the MLE of q_0)**.** *Suppose that assumptions 7.14, 7.15, 7.18, 7.19, 7.20 and 7.21 hold. Then, letting* $\alpha := 1/(d+1)$*, it holds that, for every* $x > 0$*, with probability at least* $1 - 2e^{-x}$*, that*

$$\sigma\left(\widehat{q}_n - q_n\right) \lesssim n^{-\frac{1}{4-2\alpha}} + \log n \sqrt{\frac{x}{n}} + (\log n)^{\frac{2}{2-\alpha}} \left(\frac{x}{n}\right)^{\frac{1}{2-\alpha}}.$$

7.5.3 Canonical gradient

As argued in subsection 7.5, we think that assuming that components of the data-generating distribution $P_0^{T,N}$ lie in $\mathcal{F}_{1,M,L}$ for some $M, L > 0$ is a relatively mild modelling assumption. We therefore assume that

$$\left\{\bar{D}_{T,N}(q) : q \in \mathcal{Q}\right\} \subset \mathcal{F}_{1,M,L}. \tag{7.7}$$

We conjecture that this actually automatically follows if $\mathcal{Q} \subset \mathcal{F}_{1,M,L}$ and we think that one could prove this using the usual arguments to prove bracketing numbers preservation results. However, this appears to be tedious, so we leave it to future work. Under (7.7), assumption 7.17 holds. If we further assume that $\|\bar{D}_{T,N}(q_2) - \bar{D}_{T,N}(q_1)\|_\infty \lesssim \|q_2 - q_1\|_\infty$, lemma 7.4 and 7.9 then imply that assumption 7.16 holds.

7.6 Adaptive stopping rules

In this section, we present an adaptive stopping rule for the test of the hypothesis $H_0 : \Psi(P_0^{T,N}) = 0$. In practice, it is natural to consider a parameter of the form $\Psi(P_0^{T,N}) = E_{P_0^{T,N}}[Y^{g_1^*} - Y^{g_2^*}]$, for which the analysis follows trivially from the the analysis of the individual terms of the difference we have presented so far. (Note that H_0 doesn't actually depend on T and N, since $\Psi(P_0^{T,N})$ can be written as $\Psi^{(1)}(q_0)$, as pointed out in section 7.2). An adaptive stopping rule allows to reject the null hypothesis as soon as sufficient evidence has been collected, without the need to wait for a pre-specified sample size to be met. Since an adaptive stopping rule checks a a criterion at every time step, multiple testing considerations must be taken into account so as to make sure the type I error remains controlled.

A typical approach to design a valid adaptive stopping rule is as follows. Say we want to ensure that type I error is no larger than $1 - \alpha$. The key step is to construct a uniform-in-time $(1-\alpha)$-probability confidence band, that is sequence of confidence intervals$([\pm a_{\alpha,N}(T)])_{T \geq 1}$, such that, with probability $1 - \alpha$, $\Psi(P_0^{T,N}) \in [\pm a_{\alpha,N}(T)]$ for every T. Then a natural stopping rule is to reject the null hypothesis at the earliest time T such that $0 \notin [\pm a_{\alpha,N}(T)]$. A uniform-in-time confidence band is a feature of the joint distribution of the sequence of estimators $(\widehat{\Psi}_{N,T})_{T \geq 1}$. Theorem 7.8 characterizes in an asymptotic sense the joint distribution of a process obtain from the finite sequence $(\widehat{\Psi}_{N,T})_{t=1}^T$ by rescaling it in time and in range: specifically, it shows that $\{t\sqrt{TN}\sigma_{0,\infty,N}^{-1}(\widehat{\Psi}_{N,T} - \Psi(P_0^{T,N})) : t \in [0,1]\}$ converges weakly to a Wiener process. Since confidence bands for the Wiener process are well documented, we will be able to use this to construct an adaptive stopping rule.

Since our results on the joint distribution of the (rescaled process built from the) sequence of estimates are asymptotic, our procedure requires a certain burn-in period, that is we must enforce a minimum time point before which the procedure cannot reject. We now present formally our type I error guarantees for the procedure we described.

Theorem 7.10 (Type I error of adaptive stopping)**.** *Let* $(a_\alpha(t) : t \in [0,1])$ *be such that* $P[\forall t \in [0,1], W(t) \in [\pm a_\alpha(t)]] \geq 1$. *Let* $T_{\max}$ *be the maximum number of time steps the experimenter is willing to run the trial. Let* $t_0 \in [0,1]$ *be such that* $T_0 := t_0 T_{\max}$ *is the duration of the burn-in period.*

Let

$$\tau(T_{\max}, t_0) := \min \left\{ T \geq T_0, \widehat{\Psi}_{T,N} \notin \left[\pm \sigma_{0,\infty,N} \frac{\sqrt{T_{\max}/T} a(T/T_{\max})}{\sqrt{NT}} \right] \right\}.$$

Suppose that the assumptions of theorem 7.8 are satisfied. Then, under the null hypothesis $H_0 : \Psi(P_0^{T,N}) = 0$, *it holds that*

$$\lim_{T_{\max} \to \infty} P_0 \left[\tau(T_{\max}, t_0) \leq T_{\max} \right] \geq 1 - \alpha,$$

that is the probability that the procedure rejects under the null is asymptotically no larger than the nominal level α.

In practice, theorem 7.10 teaches us that for reasonably large horizon $T_{\max}$ and burn-in period T_0, the procedure has type-I error approximately no larger than $1 - \alpha$.

An alternative direction to construct an adaptive stopping rule would be to analyze the deviations of our estimator with uniform-in-time concentration bounds, such as the ones presented in Howard et al. [2018], instead of using a limit theorem. We leave this direction for future research. We nevertheless point out that exact confidence bands/intervals obtained from concentration inequalities tend to be larger than approximate confidence bands/intervals obtained from FCLTs/CLTs. As a result, we conjecture that controlling exactly, rather than approximately the type I error by using concentration inequalities rather than limit theorems might cost a signicant loss of power.

7.7 Learning the optimal design along the trial

Consider a target parameter of the form $\Psi_\tau(q) := E_{q,g_2^*} Y - E_{q,g_1^*} Y$, where Y is an outcome at time τ, as defined earlier, and where g_1^* and g_2^* are known and fixed stochastic interventions. In the best arm identification example in the case where there are two arms, we would have $g_1^*(a \mid c^A) = \mathbf{1}(a = 1)$ and $g_2^*(a \mid c^A) = \mathbf{1}(a = 2)$, that is g_1^* and g_2^* are the deterministic interventions that always assign the same treatment. In the infectious disease example, g_1^* and g_2^* would be two different public health interventions, such as imposing that individuals wear a mask, or that they stay at home for except for a certain set of allowed activities.

Suppose that we have a collection of candidate designs $g_1(q), \ldots, g_J(q)$ that are indexed by q. We would like to achieve the same asymptotic variance as we would if we had carried out the best design among $g_1(q_0), \ldots, g_J(q_0)$ from the beginning. Let us make this more formal. Making explicit that $h_{\infty,N}$ depend on q and g, we will write $h_{\infty,N}(q,g)$. For every k, let

$$\chi_k(q) := \mathrm{Var}_{h_{\infty,N}(q,g),q}\left((\bar{D}^{g_1^*}_{T,N}(q,g) - \bar{D}^{g_2^*}_{T,N}(q,g))(C_\infty, L_\infty)\right)$$

be the asymptotic variance of the EIF under g_k.

Given an estimator $\widehat{q}_{T,N}$ of q_0, we can compute (approximately by Monte-Carlo simulation for example) $\chi_k(\widehat{q}_{T,N})$, the plug-in estimator of $\chi_k(q_0)$. Let $k(T) := \arg\min_{k \in [J]} \chi_k(\widehat{q}_{T,N})$. We define our adaptive design at T as $g_{k(T-1)}(\widehat{q}_{T-1,N})$.

We now study heuristically the conditions under which this adaptive design is such that the TMLE of $\Psi_\tau(q_0)$ achieves the asymptotic variance $\chi_{k^*}(q_0)$, with $k^* = \arg\min_{k\in[J]} \chi_k(q_0)$, that is the optimal asymptotic variance among the J designs considered.

Suppose that $\widehat{q}_{T,N}$ converges almost surely to q_0 and that $\chi_1(q_0), \ldots, \chi_J(q_0)$ are distinct. Then $\chi_k(\widehat{q}_{T,N})$ converges a.s. to $\chi_k(q_0)$, and therefore, with probability 1, $k(T) \neq k^*$ only a finite number of times. Therefore, we expect that $\|\bar{h}_{0,T,N} - h_{\infty,N}(q_0, g_{k^*}(q_0))\|_1 = o(1)$, which in turns, if $\bar{h}_{0,T,N}$ and $h_{\infty,N}(q_0, g_{k^*}(q_0))$ are lower bounded away from zero implies that $\|\bar{h}^{-1}_{0,T,N} - h^{-1}_{\infty,N}(q_0, g_{k^*}(q_0))\|_1 = o(1)$. Therefore, under the assumptions of lemma 7.3, the variance of the the terms $\bar{D}^{g_1^*}_{T,N} - \bar{D}^{g_2^*}_{T,N}$ of the EIF should stabilize to $\chi_{k^*}(q_0)$, which under the assumptions of theorem 7.8, implies that under the adaptive design, the asymptotic variance of the TMLE must be $\chi_{k^*}(q_0)$.

Examples of candidate designs in the best arm identification example. In the best arm identification example, in the case where there are only two arms and where $\tau = 1$ (that is the target is the ATE after one time step, starting from a known distribution of contexts), it is known that the optimal design is the so-called Neyman allocation design, defined as follows:

$$g(q_0)(a \mid c^A) := \frac{\sigma_{q_0}(a, c^A)}{\sigma_{q_0}(1, c^A) + \sigma_{q_0}(2, c^A)},$$

where

$$\sigma^2_q(a, c^A) := \mathrm{Var}_q(Y(1) \mid A(1,1) = a, C^A(1,1) = c^A).$$

In words, the Neyman allocation designs assigns treatment a with probability proportional to the standard deviation of the outcome conditional on a and c^A.

While we don't know whether this design is optimal design among all possible designs in the case $\tau > 1$, we conjecture it should be more efficient that the uniform design over treatment arms. In practice, we recommend considering a finite library of candidate designs including the Neyman allocation design. Other possible designs are the constant design with fixed probabilities for each arm.

7.8 Conclusion

In this chapter, we have studied the questions of inference and sequential decision making in a trial involving N individuals that we enroll at the same time in the trial and then follow for T steps. We allow for network dependence. The statistical model is the one presented in subsection 1.2.4 of the introduction chapter. We defined causal parameters that we interpret as the discounted cumulative effect and the long term effect of a counterfactual stochastic intervention, and we showed that these are identifiable from the observed data distribution.

At a high level, our analysis shows that statistical inference is possible if we impose restrictions on the amount of dependence within and across trajectories. We formalize this constraint by imposing mixing conditions. The main technical enablers of this work are our novel maximal inequality for empirical processes over weakly dependent sequences and our high probability bound for empirical risk minimizers.

One limitation of our work is that our estimators require that the asymptotic variance of the terms of the canonical gradient stabilize as a function of T. This in particular rules out designs that that let the probability of assignment of certain treatment arms converge to zero. A direction for future work would be to alleviate this requirement so as to allow for designs that gradually phase out suboptimal arms.

Bibliography

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

G W Basse, A Feller, and P Toulis. Randomization tests of causal effects under interference. *Biometrika*, 106(2):487–494, 02 2019.

Guillaume W Basse and Edoardo M Airoldi. Model-assisted design of experiments in the presence of network-correlated outcomes. *Biometrika*, 105(4):849–858, 08 2018.

D. Benkeser and M. Van Der Laan. The highly adaptive lasso estimator. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 689–696, 2016.

Aurelien Bibaut, Antoine Chambaz, and Mark van der Laan. Generalized policy elimination: an efficient algorithm for nonparametric contextual bandits. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 1099–1108, Virtual, 03–06 Aug 2020. PMLR.

Aurélien F. Bibaut and Mark J. van der Laan. Fast rates for empirical risk minimization over càdlàg functions with bounded sectional variation norm, 2019.

A. Boruvka, D. Almirall, K. Witkiewitz, and S. A. Murphy. Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association*, 0(ja):0–0, 2017.

R. C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probab. Surveys*, 2:107–144, 2005.

R. C. Bradley. *Introduction to strong mixing conditions*, volume 1. Kendrick Press, 2007.

Billy Fang, Adityanand Guntuboyina, and Bodhisattva Sen. Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and hardy-krause variation, 2020.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences, 2018.

Michael G. Hudgens and M. Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning, 2019.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

Chris A. J. Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.*, 15(4):1548–1562, 12 1987.

D. L. McLeish. Dependent central limit theorems and invariance principles. *Ann. Probab.*, 2(4): 620–628, 08 1974.

Florence Merlevède, Magda Peligrad, and Emmanuel Rio. *Bernstein inequality and moderate deviations under strong mixing conditions*, volume Volume 5 of *Collections*, pages 273–292. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2009.

Elizabeth L. Ogburn, Oleg Sofrygin, Ivan Diaz, and Mark J. van der Laan. Causal inference for social network data, 2020.

Mina Ossiander. A central limit theorem under metric entropy with l_2 bracketing. *Ann. Probab.*, 15(3):897–919, 07 1987.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Iosif Pinelis. Bounding supremum norm of lipschitz function by l1 norm. MathOverflow, 2020. URL:https://mathoverflow.net/q/379490 (version: 2020-12-22).

Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58(5):527–535, 09 1952.

Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75, 2012.

M. van der Laan. A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. *Int J Biostat*, 2, October 2017.

Mark van der Laan. The construction and analysis of adaptive group sequential designs. 2008.

Mark van der Laan. Causal inference for a population of causally connected units. *Journal of Causal Inference*, 2, 2013.

Mark van der Laan and Susan Gruber. One-step targeted minimum loss-based estimation based on universal least favorable one-dimensional submodels. *The international journal of biostatistics*, 12(1):351–378, 2016.

Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

Mark J Van der Laan and Sherri Rose. *Targeted learning in data science*. Springer, 2018.

Mark J. van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 28 Dec. 2006. doi: https://doi.org/10.2202/1557-4679.1043.

Mark J. van der Laan, Antoine Chambaz, and Sam Lendle. *Online Targeted Learning for Time Series*, pages 317–346. Springer International Publishing, 2018.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag New York, 03 1996. ISBN 9781475725452.

Ramon van Handel. On the minimal penalty for markov order estimation. *Probability Theory and Related Fields*, 150(3-4):709–738, April 2010. ISSN 1432-2064. doi: 10.1007/s00440-010-0290-y.

A. Wald. Sequential tests of statistical hypotheses. *Ann. Math. Statist.*, 16(2):117–186, 06 1945. doi: 10.1214/aoms/1177731118.

Wenjing Zheng and Mark J. van der Laan. *Cross-Validated Targeted Minimum-Loss-Based Estimation*, pages 459–474. Springer New York, New York, NY, 2011.

7.A Notation

7.A.1 Notation relative to the data

$$\begin{aligned}
A(t,i) &: \text{treatment assigned to individual } i \text{ at } t,\\
L(t,i) &: \text{time varying covariates and outcomes of individual } i \text{ at } t,\\
O(t,i) :=&(A(t,i), L(t,i),\\
A(t) :=&(A(t,i) : i \in [N])\\
L(t) :=&(L(t,i) : i \in [N])\\
\bar{A}(t) =&(A(1), \ldots, A(t)),\\
\bar{L}(t) :=&(L(1), \ldots, L(t)),
\end{aligned}$$

$$\begin{aligned}
\bar{O}(t) :=& (O(1), \dots, O(t)),\\
\bar{A}(t,i) :=& (A(s,i) : s \in [t]),\\
\bar{L}(t,i) :=& (L(s,i) : s \in [t]),\\
\bar{O}(t,i) :=& (O(s,i) : s \in [t]),\\
O^{T,N} :=& (O(t,i) : t \in [T], i \in [N]).
\end{aligned}$$

Observe that $O^{T,N} = \bar{O}(T)$.

Data is observed in the order $A(1), L(1), \dots, A(t), L(t)$. Within time points, we arbitrarily order data points by increasing index i, that is, we order individual observations as:

$$A(1,1), \dots, A(1,N), L(1,1), \dots, L(1,N), \dots, A(T,1), \dots, A(T,N), L(T,1), \dots, L(T,N).$$

We refer to this ordering as the column ordering. We let $A(t,i)^-$ and $L(t,i)^-$ be the vectors of all observations that come before $A(t,i)$ and $L(t,i)$ in the column ordering, that is

$$\begin{aligned}
A(t,i)^- :=& (\bar{O}(t-1), A(t,1), \dots, A(t,i-1))\\
L(t,i)^- :=& (\bar{O}(t-1), A(t), L(t,1), \dots, L(t,i-1)).
\end{aligned}$$

We let $F_L(t,i)$ be the set of individuals i is in contact with at t, and we let $F_A(t,i)$ be the set of individuals upon whose history the experimenter decides the treatment assignment of individual i at time t. We define the context functions $C^A(t,i)$ and $C^L(t,i)$ as

$$\begin{aligned}
C^A(t,i) &:= c_{A(t,i)}(A(t,i)^-, F_A(t,i))\\
C^L(t,i) &:= c_{L(t,i)}(L(t,i)^-, F_L(t,i)).
\end{aligned}$$

When there is no ambiguity, we drop the "L'' superscript and we use $C(t,i)$ for $C^L(t,i)$. We let

$$X(t,i) := (C^L(t,i), L(t,i)).$$

We denote $\widetilde{A}(k)$, $\widetilde{L}(k)$, $\widetilde{O}(k)$ and $\widetilde{X}(k)$ the k-th element in the sequences $(A(t,i))_{t,i}$, $(L(t,i))_{t,i}$, $(O(t,i))_{t,i}$, and $(X(t,i))_{t,i}$, respectively.

7.A.2 Notation relative to probability distributions, their components, and the target parameters

$P_F^{T,N}$: probability distribution of the full data $(O^{T,N}, U)$,

$P_F^{*,T,N}$: post-intervention distribution of the full data $(O^{*,T,N}, U)$,

$P^{T,N}$: probability distribution of the observed data $O^{T,N}$,

$P_{g^*}^{T,N}$: G-computation formula

We denote $\mathcal{M}_F^{T,N}$ the causal model, that is the set of possible full data distributions $P_F^{T,N}$, and $\mathcal{M}^{T,N}$ the statistical model that is the set of possible observed data distributions $P^{T,N}$. We now present notation for components of these distributions:

$$\begin{aligned}
q &: \text{conditional density of } L(t,i) \text{ given } L(t,i)^- \text{ (or given } C^L(t,i))\\
g = \prod_{t=1}^{T}\prod_{i=1}^{N} g_{t,i} &: \text{treatment mechanism },\\
g^* = \prod_{s=1}^{\tau}\prod_{j=1}^{N} g^*_{s,j} &: \text{counterfactual treatment mechanism},\\
h^A_{t,i}(q,g) \text{ and } h^L_{t,i}(q,g) &: \text{marginal densities of } C^A(t,i) \text{ and } C^L(t,i) \text{ under } P^{T,N} = \prod_{t,i} g_{t,i} q,\\
h^A_{t,i} \text{ and } h^L_{t,i} &: \text{shorthand for } h^A_{t,i(q,g)} \text{ and } h^L_{t,i}(q,g)\\
h^{*,A}_{s,j} \text{ and } h^{*,L}_{s,j} &: \text{shorthand for } h^A_{s,j}(q,g^*) \text{ and } h^L_{s,j}(q,g^*),\\
\bar{h}^A_{T,N} :=& (TN)^{-1}\sum_{t=1}^{T}\sum_{i=1}^{N} h^A_{t,i},\\
\bar{h}^L_{T,N} :=& (TN)^{-1}\sum_{t=1}^{T}\sum_{i=1}^{N} h^L_{t,i},\\
\omega_{s,j} :=& h^A_{s,j}/\bar{h}^A_{T,N} \text{ and } \eta_{s,j} := g^*_{s,j}/g_{s,j}.
\end{aligned}$$

When there is no ambiguity, we use $h_{t,i}$, $h^*_{s,j}$, $\bar{h}_{T,N}$ for $h^L_{t,i}$, $h^{*,L}_{s,j}$, $\bar{h}^L_{T,N}$. We denote $\widetilde{h}_k$ the k-th element of the column ordered sequence $(h_{t,i})_{t,i}$. We now recall the definition of the causal parameter and the statistical target parameter:

$$\begin{aligned}
\Psi^F_\tau(P_F^{T,N}) :=& E_{P_F^{*,T,N}}[Y^*(\tau)] \text{ (causal parameter)},\\
\Psi_\tau(P^{T,N}) :=& E_{P_{q,g^*}}[Y(\tau)] \text{ (statistical target parameter)}.
\end{aligned}$$

7.B Proofs of the structural results

7.B.1 Derivation of the efficient influence function

The proof of theorem 7.1 relies on the following lemma, which is a straightforward extension of lemma 1 in van der Laan [2013].

Lemma 7.5 (Projection onto tangent space.). *The tangent space of the statistical model $\mathcal{M}$ at $P^{T,N}$ is given by*

$$T(q) := \left\{ o^{T,N} \mapsto \sum_{t,i} s(l(t,i) \mid c^L(t,i)) : s : \mathcal{L}\times\mathcal{C} \to \mathbb{R},\ \forall c, \int s(l,c) q(l \mid c) dl = 0 \right\}.$$

The projection of any function $o^{T,N} \mapsto D^*(o^{T,N})$ *on* $T(q)$ *is*

$$\bar{D} : o^{T,N} \mapsto \frac{1}{TN} \sum_{t=1}^{N} \sum_{i=1}^{N} \frac{h_{t,i}(c^L)}{\bar{h}_{T,N}(c^L)} D_{t,i}(l \mid c),$$

with

$$D_{t,i}(l,c) := E\left[D^*(O^{T,N} \mid L(t,i) = l, C^L(t,i) = c^L\right] - E\left[D^*(O^{T,N} \mid C^L(t,i) = c^L\right].$$

The proof is almost identical to that of lemma 1 in van der Laan [2013]. We refer the interested reader to this work.

We now present the proof of theorem 7.1.

Proof of theorem 7.1. We start with the case $t = \tau$. We use the following classical strategy to find the canonical gradient: we first find a gradient of Ψ at q w.r.t. $T(q)$, and we then project it onto $T(q)$, which gives the canonical gradient.

Finding a gradient. Consider a one-dimensional sub-model of $\mathcal{M}$ of the form

$$\left\{ P_\epsilon^{T,N} : \epsilon \in [\pm\epsilon_{\max}], \frac{dP_\epsilon^{T,N}}{d\mu}(o^{T,N}) = \prod_{t,i} q_\epsilon(l(t,i) \mid c^L(t,i)) g^*(a(t,i) \mid c^A(t,i)) \right\},$$

such that $P_{\epsilon=0}^{T,N} = P^{T,N}$. We have that

$$\Psi(P_\epsilon^{T,N}) = \Psi(q_\epsilon) = \int y \prod_{t,i} q_\epsilon(l(t,i) \mid c^L(t,i)) g^*(a(t,i) \mid c^A(t,i)) do^{T,N}.$$

Therefore

$$\begin{aligned}
\left.\frac{d\Psi(q_\epsilon)}{d\epsilon}\right|_{\epsilon=0} &= \int y \frac{d}{d\epsilon} \prod_{t,i} q_\epsilon \Big|_{\epsilon=0} \prod_{t,i} g_{t,i}^* \\
&= \int \prod_{t,i} q g_{t,i} \left\{ \prod_{t,i} \frac{g_{t,i}^*}{g_{t,i}} y \right\} \frac{d}{d\epsilon} \log \prod_{t,i} q_\epsilon \Big|_{\epsilon=0} \\
&= \int \prod_{t,i} q g_{t,i} \left\{ \prod_{t,i} \frac{g_{t,i}^*}{g_{t,i}} y - \Psi(q) \right\} \frac{d}{d\epsilon} \log \prod_{t,i} q_\epsilon \Big|_{\epsilon=0} \\
&= E_{q,g} \left[\left(\prod_{t,i} \frac{g_{t,i}^*}{g_{t,i}} Y - \Psi(q) \right) \sum_{t,i} s(L(t,i), C^L(t,i)) \right],
\end{aligned}$$

where $s(l, c^L) := (d \log q_\epsilon / d\epsilon)|_{\epsilon=0}(l, c^L)$, and therefore $\sum_{t,i} s(l(t,i), c^L(t,i))$ is the score of the parametric submodel at $\epsilon = 0$.

Therefore

$$o^{T,N} \mapsto D^*(o^{T,N}) = \prod_{t,i} \frac{g_{t,i}^*}{g_{t,i}} y - \Psi(q)$$

is a gradient of Ψ at $P^{T,N}$ w.r.t. $T(P)$.

Projecting D^0 onto the tangent space. From lemma 7.5, the projection of D^0 onto $T(q)$ is

$$D(q)(o^{T,N}) = \frac{1}{TN} \sum_{t,i} \bar{D}_{T,N}(q)(c^L(t,i), l(t,i)),$$

with

$$\begin{aligned}\bar{D}_{T,N}(q)(c^L, l) = \sum_{s,j} \frac{h_{s,j}(c^L)}{\bar{h}_{T,N}(c^L)} \left\{ E_{q,g} \left[Y g^*/g \mid L(t,i) = l, C^L(t,i) = c^L \right] \right. \\ \left. - E_{q,g} \left[Y g^*/g \mid C^L(t,i) = c^L \right] \right\}.\end{aligned}$$

Second representation. Suppose that assumption 7.5 holds. We have that

$$\begin{aligned}
&E_{q,g} \left[Y g^*/g \mid L(t,i) = l, C^L(t,i) = c^L \right] \\
=&\frac{1}{h_{t,i}(c^L)} \int E_{q,g} \left[Y g^*/g \mid L(t,i) = l, L(t,i)^- = l(t,i)^- \right] \\
&\qquad \times \mathbf{1}(c^L(l(t,i)^-) = c^L) \prod_{(s,j)<(t,i)} q(l(s,j) \mid l(s,j)^-) g_{s,j}(a(s,j) \mid a(s,j)^-) do^{T,N} \\
=&\frac{1}{h_{t,i}(c^L)} \int E_{q,g^*} \left[Y \mid L(t,i) = l, L(t,i)^- = l(t,i)^- \right] \\
&\qquad \times \mathbf{1}(c^L(l(t,i)^-) = c^L) \prod_{(s,j)<(t,i)} q(l(s,j) \mid l(s,j)^-) g^*_{s,j}(a(s,j) \mid a(s,j)^-) do^{T,N} \\
=&\frac{1}{h_{t,i}(c^L)} E_{q,g^*} \left[Y \mid L(t,i) = l, c^L(L(t,i)^-) = c^L \right] \\
&\times \int \mathbf{1}(c^L(l(t,i)^-) = c^L) \prod_{(s,j)<(t,i)} q(l(s,j) \mid l(s,j)^-) g^*_{s,j}(a(s,j) \mid a(s,j)^-) do^{T,N} \\
=&\frac{h^*_{t,i}(c^L)}{h_{t,i}(c^L)} E_{q,g^*} \left[Y \mid L(t,i) = l, C^L(t,i) = c^L \right].
\end{aligned}$$

Similarly,

$$E_{q,g} \left[Y g^*/g \mid C^L(t,i) = c^L \right] = \frac{h^*_{t,i}(c^L)}{h_{t,i}(c^L)} E_{q,g^*} \left[Y \mid C^L(t,i) = c^L \right].$$

Replacing these expression in the expression of the canonical gradient gives the wished representation.

Third representation. Under assumption 7.6, the third representation follows immediately from the second one. □

7.B.2 Proofs of the results on the remainder term

Proof of theorem 7.2. Suppose $\bar{h}_{T,N} = \bar{h}_{0,T,N}$. We have that

$$\begin{aligned}
&E_{P_0^{T,N}}\left[D(q)(O^{T,N})\right] \\
=&E_{P_0^{T,N}}\left[\frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\bar{D}_{T,N}(C(t,i),L(t,i))\right] \\
=&\int \bar{h}_{0,T,N}(c)q_0(l \mid c)\bar{D}_{T,N}(c,l)dcdl \\
=&\sum_{s=1}^{\tau}\sum_{j=1}^{N}\left\{\int h^*_{s,j}(c)q_0(l \mid c)E_{q,g^*}\left[Y \mid L(s,j)=l, C(s,j)=c\right]dldc\right. \\
&\qquad\qquad \left.-\int h^*_{s,j}(c)E_{q,g^*}\left[Y \mid C(s,j)=c\right]dc\right\}
\end{aligned}$$

We have that

$$\begin{aligned}
&\int h^*_{s,j}(c)q_0(l \mid c)E_{q,g^*}\left[Y \mid L(s,j)=l, C(s,j)=c\right] \\
=&\int q_0(l \mid c)E_{q,g^*}\left[Y \mid L(s,j)=l, L(s,j)^- = l(s,j)^-\right] \\
&\qquad \times \mathbf{1}(c_{L(s,j)}(l(s,j)^-)=c)\prod_{(s',j')<(s,j)} q_{s',j'}g^*_{s,j}dldcd(l(s,j)^-) \\
=&\int E_{q,g^*}\left[Y \mid L(s,j)=l, L(s,j)^- = l(s,j)^-\right]q_{0,s,j}\prod_{(s',j')<(s,j)} q_{s',j'}g^*_{s,j} \\
=&E_{q_{(s,j)-},q_{0,s,j}q_{(s,j)+}}Y, \qquad (7.8)
\end{aligned}$$

where the last line was obtained by using Fubini's theorem and integrating out the indicator. □

The same arguments show that

$$\int h^*_{s,j}(c)E_{q,g^*}\left[Y \mid C(s,j)=c\right] = E_{q,g^*}[Y].$$

Therefore,

$$E_{P_0^{T,N}}\left[D(q)(O^{T,N})\right] = \sum_{s=1}^{\tau}\sum_{j=1}^{N}E_{q_{(s,j)-},(q_0-q)_{(s,j)},q^+_{(s,j)},g^*}Y.$$

Using the telescoping sum formula for the product difference $\prod_{s,j} q_{s,j} - \prod_{s,j} q_{0,s,j}$, we have that

$$\Psi(q) - \Psi(q_0) = \sum_{s,j}E_{q_{(s,j)-},q_{(s,j)}-q_{0,(s,j)},q_{0,(s,j)},g^*}Y. \qquad (7.9)$$

Therefore, putting (7.8) and (7.9) together gives the wished expression for $R(\bar{h}_{0,T,N},q)$.

The derivation of the expression $R(\bar{h}_{0,T,N},q) - R(\bar{h}_{T,N},q)$ is immediate.

Proof of theorem 7.3. We have that

$$\begin{aligned}
&E_{P_0^{T,N}}\left[D(q)(O^{T,N})\right]\\
=&E_{P_0^{T,N}}\left[\frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\bar{D}_{T,N}(q)(C^A(t,i),A(t,i),L(t,i)\right]\\
=&\sum_{s=1}^{\tau}\sum_{j=1}^{N}\int \bar{h}^A_{0,T,N}(c^A)g_{s,j}(a\mid c^A)q_0(l\mid a,c^A)\omega_{s,j}(c^A)\eta_{s,j}(a\mid^A)\\
&\qquad\times\left\{E_{q,g^*}\left[Y(j)\mid L(s,j)=l,A(s,j)=a,C^A(s,j)=c^A\right]\right.\\
&\qquad\left.-E_{q,g^*}\left[Y(j)\mid A(s,j)=a,C^A(s,j)=c^A\right]\right\}dldadc^A,
\end{aligned}$$

where we have used assumption 7.7 in the last line above.

Case $\omega=\omega_0$**.** We show that in this case, for any given j and s, the second term of the (s,j)-th term in the sum above cancels out with the first term of the $(s-1,j)$-th term.

We start with rewriting the second term of the (s,j)-th term:

$$\begin{aligned}
&\int \bar{h}^A_{0,T,N}(c^A)g_{s,j}(a\mid c^A)q_0(l\mid a,c^A)\omega_{0,s,j}(c^A)\eta_{s,j}(a\mid c^A)\\
&\qquad\times E_{q,g^*}\left[Y(j)\mid A(s,j)=a,C^A(s,j)=c^A\right]dldadc^A\\
=&\int \bar{h}^{*,A}_{0,s,j}(c^A)g^*_{s,j}(a\mid c^A)E_{q,g^*}\left[Y(j)\mid A(s,j)=a,C^A(s,j)=c^A\right]dadc^A\\
=&\int\prod_{t=1}^{s-1}g^*_{t,j}q_{0,t,j}g^*_{s,j}E_{q,g^*}\left[Y(j)\mid A(s,j)=a,\bar{O}(s-1,j)=\bar{o}(s-1,j)\right]dad\bar{o}(s-1,j)\\
=&E_{q_{0,1:s-1},q_{s:\tau},g^*}Y.
\end{aligned}$$

The third line above follows from assumption 7.5. We now show that the first term of the $(s-1,j)$-th term is equal to the above quantity:

$$\begin{aligned}
&\int \bar{h}^A_{0,T,N}(c^A)g_{s,j}(a\mid c^A)q_0(l\mid a,c^A)\omega_{s,j}(c^A)\eta_{s,j}(a\mid^A)\\
&\qquad\times E_{q,g^*}\left[Y(j)\mid L(s-1,j)=l,A(s-1,j)=a,C^A(s-1,j)=c^A\right]dldadc^A\\
=&\int \bar{h}^{*,A}_{0,s,j}(c^A)g^*_{s,j}(a\mid c^A)q_0(l\mid a,c^A)\\
&\qquad\times E_{q,g^*}\left[Y(j)\mid L(s-1,j)=l,A(s-1,j)=a,C^A(s-1,j)=c^A\right]dldadc^A\\
=&\int\prod_{t=1}^{s-2}g^*_{t,j}q_{0,t,j}g^*_{s-1,j}q_{0,s-1}E_{q,g^*}\left[Y(j)\mid\bar{O}(s-1,j)=\bar{o}(s-1,j)\right]d\bar{o}(s-1,j)\\
=&E_{q_{0,1:s-1},q_{s:\tau},g^*}Y.
\end{aligned}$$

Thus, by telescoping, $E_{P_0^{T,N}}[D(q)(O^{T,N})]=\Psi(q)-\Psi(q_0)$, and therefore $R(\omega_0,q)=0$.

At a high level, the reason why this cross-terms cancellation happens is the first term of the $(s-1,j)$-th term is obtained by integration against $g^*_{s,j}$ of the second term of (s,j)-th term. Applying the operator $E_{P_0^{T,N}}$ boils down to successively, in the backwards direction, integrating with respect to the factors of $P_0^{T,N}$. The first step in this process applied to the second term of (s,j) is to integrate w.r.t. $g^*_{s,j}$, which gives the first term of $(s-1,j)$. The subsequent steps being the same for both terms, the resulting quantities are the same.

Case $q = q_0$**.** In this case, it is immediate to observe that the cancellation happens within each term of the terms of the sum over (s,j). Therefore, $R(\omega, q_0) = 0$.

Therefore, we can write $R(\omega, q) = R(\omega, q) - R(\omega_0, q)$, which makes appear the wished product of differences structure. □

7.C Results on empirical process induced by weakly dependent sequences

Let $(X_n)_{n\geq 1}$ be a sequence of random variables taking values in a set $\mathcal{X}$, and let $\mathcal{F}$ be a class of functions with domain $\mathcal{X}$. In this section, we present a several novel results on empirical processes of the form

$$\{M_n(f) : f \in \mathcal{F}\} \qquad \text{where} \qquad M_n(f) := \frac{1}{n}\sum_{i=1}^{n} f(X_i) - E[f(X_i)].$$

We present three types of results: a maximal inequality over $\mathcal{F}$ (or over the intersection of $\mathcal{F}$ with a ball of controlled radius), an equicontinuity result, and an exponential risk bound for empircal risk mimizers over $\mathcal{F}$. The latter two are a consequence of the former.

We do not make independence nor stationarity assumptions on the sequence $(X_n)_{n\geq 1}$. Rather, we consider sequences $(X_n)_{n\geq 1}$ that satisfy only the following mixing conditions.

Assumption 7.22 (α-mixing)**.** *The uniform α-mixing coefficients of the sequence $(X_i)_{i\geq 1}$ satisfy*

$$\alpha(n) \leq \exp(-2cn), \textit{ for some } c > 0.$$

Assumption 7.23 (ρ-mixing)**.** *The uniform ρ-mixing coefficients of the sequence $(X_i)_{n\geq 1}$ have finite sum, that is $\sum_{n\geq 1} \rho(n) < \infty$. We denote $\boldsymbol{\rho} := \sum_{n\geq 1} \rho(n)$.*

We suppose that $\mathcal{X} \subset \mathbb{R}^d$ for some $d \geq 1$ and that, for every $i \geq 1$, the marginal distribution of X_i admits a density w.r.t. the Lebesgue measure that we denote h_i. Supposing that assumption 7.23 holds, we define the following mapping $\mathcal{F} \to \mathbb{R}$:

$$\sigma(f) := \sqrt{1 + 2\boldsymbol{\rho}} \sup_{i\geq 1} \|f\|_{2,h_i}.$$

It is straightforward to check that σ is a norm. Our results apply to classes of functions that are bounded in supremum norm.

Assumption 7.24 (Uniform boundedness)**.** *There exists $M \in (0,\infty)$ such that $\sup_{f\in\mathcal{F}} \|f\|_\infty \leq M$.*

7.C.1 A local maximal inequality

The result of this subsection is a *local* maximal inequality in the sense that it bounds the supremum of $M_n(f)$ over a σ-ball included in $\mathcal{F}$. We state below the corresponding assumption.

Assumption 7.25 (σ norm boundedness)**.** *There exists $r > 0$ such that $\sup_{f \in \mathcal{F}} \sigma(f) \leq r$.*

We can now state our result.

Theorem 7.11. *Suppose that assumptions 7.22, 7.23, 7.24 hold. Suppose that $r \geq 2Mn^{-1/2}$. Then, for any $r^{\in}[Mn^{-1/2}, r]$, it holds with probability at least $1 - 2e^{-x}$ that,*

$$\begin{aligned} \sup_{f \in \mathcal{F}} M_n(f) \lesssim & r^- + \frac{\log n}{\sqrt{n}} \int_{r^-}^{r} \sqrt{\log(1 + N_{[]}(\epsilon, \mathcal{F}, \sigma))} d\epsilon + \frac{M(\log n)^2}{n} \log(1 + N_{[]}(r, \mathcal{F}, \sigma)) \\ & + r\sqrt{\frac{(\log n)x}{n}} + \frac{M(\log n)^2}{n} x. \end{aligned}$$

The proof relies on the following lemma, which is a corollary of lemma A.7 in van Handel [2010] and theorem 2 in Merlevède et al. [2009].

Lemma 7.6. *Suppose that $f_1, \ldots, f_N \in \mathcal{F}$, and that conditions on the preceding theorem hold. Then, for any event A defined on the same probability space as $(X_i)_{n \geq 1}$,*

$$\begin{aligned} E\left[\max_{i \in [N]} M_n(f_i) \mid A\right] \lesssim & \frac{1}{\sqrt{n}} \left(\max_{i \in N} \sigma(f_i) + \frac{M}{\sqrt{n}} \right) \sqrt{\log\left(1 + \frac{N}{P[A]}\right)} \\ & + \frac{M(\log n)^2}{n} \log\left(1 + \frac{N}{P[A]}\right). \end{aligned}$$

Proof of theorem 7.11. The result will follow if we show that for $A := \left\{\sup_{f \in \mathcal{F}} M_n(f) \geq \Psi(x)\right\}$, with

$$\begin{aligned} \Psi(x) := & r^- + \frac{\log n}{\sqrt{n}} \int_{r^-}^{r} \sqrt{\log(1 + N_{[]}(\epsilon, \mathcal{F}, \sigma))} d\epsilon + \frac{M(\log n)^2}{n} \log(1 + N_{[]}(r, \mathcal{F}, \sigma)) \\ & + r\sqrt{\frac{(\log n)x}{n}} + \frac{M(\log n)^2}{n} x, \end{aligned}$$

it holds that

$$E\left[\sup_{f \in \mathcal{F}} M_n(f) \mid A\right] \leq \Psi\left(\log\left(1 + \frac{1}{P[A]}\right)\right)$$

Setting up the notation. Let $\epsilon_j := r2^{-j}$, and let $J \geq 1$ such that $\epsilon_J \leq r^- < \epsilon_{J-1}$. For every j, let

$$\mathcal{B}_j := \{(\lambda_k^j, \upsilon_k^j) : k \in [N_j]\}$$

be an ϵ_j-bracketing of $\mathcal{F}$ in σ norm. For every $f \in \mathcal{F}$ and every j, let $k(j,f)$ be such that

$$\lambda^j_{k(j,f)} \leq f \leq \upsilon^j_{k(j,f)}.$$

For every j and f, define the function $\Delta^j_f := \upsilon^j_{k(j,f)} - \lambda^j_{k(j,f)}$. Let a_j be a decreasing sequence of positive numbers, such that a_{j-1} and a_j are within constant factors of each other for every j. We introduce, for every f, the function

$$\tau(f) : x \mapsto \left(\min\{j \geq 0 : \Delta^j_f(x) > a_j\}\right) \wedge J.$$

Chaining decomposition. We write f as a telescoping sum using an adaptive chaining device. Adaptive chaining is a standard empirical process technique introduced by Ossiander [1987] for the analysis of empirical processes under bracketing entropy conditions, in which the depth of a chain is a function of the form of $\tau(f)$ that is chosen so as to control the supremum norm of the links of the chain. We have, in a pointwise sense, that, for every $f \in \mathcal{F}$,

$$\begin{aligned} f =& \lambda^0_{k(0,f)} + \sum_{j=1}^{\tau(f)} \left(\lambda^j_{k(j,f)} - \lambda^{j-1}_{k(j-1,f)}\right) + \left(f - \lambda^{\tau(f)}_{k(\tau(f),f)}\right) \\ =& \lambda^0_{k(0,f)} + \sum_{j=1}^{J} \left(\lambda^j_{k(j,f)} - \lambda^{j-1}_{k(j-1,f)}\right) \mathbf{1}(\tau(f) < j) \\ &+ \sum_{j=1}^{J} \left(f - \lambda^j_{k(j,f)}\right) \mathbf{1}(\tau(f) = j). \end{aligned}$$

The first term represents the root of the chain. The second term is the sum across depth levels j of the links of the chain. The third term is the tip of the chain.

Control of the tips. We treat separately the case $j < J$ and the case $j = J$.

Case $j < J$. We will use the fact that, for $j < J$, we must have that if $\tau(f) = j$, then $\Delta^j_f > a_j$. From non-negativity of $f - \lambda^j_{k(j,f)}$,

$$\begin{aligned} M_n((f - \lambda^j_{k(j,f)})\mathbf{1}(\tau(f) = j)) \leq & E\left[\frac{1}{n}\sum_{i=1}^n (f - \lambda^j_{k(j,f)})(X_i)\mathbf{1}(\tau(f)(X_i) = j)\right] \\ \leq & E\left[\frac{1}{n}\sum_{i=1}^n \Delta^j_f(X_i)\mathbf{1}(\tau(f)(X_i) = j)\right]. \end{aligned}$$

As $\Delta^j_f \mathbf{1}(\tau(f) = j) > a_j \mathbf{1}(\tau(f) = j)$, we have that $\Delta^j_f \mathbf{1}(\tau(f) = j) \leq a_j^{-1}(\Delta^j_f)^2 \mathbf{1}(\tau(f) = j)$, and therefore

$$E\left[\frac{1}{n}\sum_{i=1}^n \left(\Delta^j_f \mathbf{1}(\tau(f) = j)\right)(X_i)\right] \leq \frac{1}{a_j}\frac{1}{n}\sum_{i=1}^n E\left[\left(\Delta^j_f(X_i)\right)^2\right]$$

$$
\begin{aligned}
&= \frac{1}{a_j}\frac{1}{n}\sum_{i=1}^{n}\|\Delta_f^j\|_{2,h_i}^2 \\
&\leq \frac{1}{a_j}\sigma^2(\Delta_f^j) \\
&\leq \frac{\epsilon_j^2}{a_j}.
\end{aligned}
$$

Therefore

$$E\left[\sup_{f\in\mathcal{F}} M_n\left((f-\lambda_{k(j,f)}^j)\mathbf{1}(\tau(f)=j)\right)\right] \leq \frac{\epsilon_j^2}{a_j}.$$

Case $j = J$. We have that

$$
\begin{aligned}
M_n\left((f-\lambda_{k(j,f)}^j)\mathbf{1}(\tau(f)=J)\right) &\leq \frac{1}{n}\sum_{i=1}^{n}\|\Delta_f^j\|_{1,h_i} \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\|\Delta_f^j\|_{2,h_i} \\
&\leq \sigma(\Delta_f^j) \\
&\leq \epsilon_J.
\end{aligned}
$$

Control of the links. From the triange inequality

$$
\begin{aligned}
\sigma\left(\lambda_{k(j,f)}^j - \lambda_{k(j-1,f)}^{j-1}\right) &\leq \sigma\left(f-\lambda_{k(j,f)}^j\right) + \sigma\left(f-\lambda_{k(j-1,f)}^{j-1}\right) \\
&\leq \sigma\left(\Delta_f^j\right) + \sigma\left(\Delta_f^{j-1}\right) \\
&\leq \epsilon_j + \epsilon_{j-1} \\
&\lesssim \epsilon_j.
\end{aligned}
$$

Similarly

$$
\begin{aligned}
\left\|\left(\lambda_{k(j,f)}^j - \lambda_{k(j-1,f)}^{j-1}\right)\mathbf{1}(\tau(f)<j)\right\|_\infty &\leq \left\|\Delta_{k(j,f0}^j\mathbf{1}(\tau(f)<j)\right\|_\infty + \left\|\Delta_{k(j-1,f0}^{j-1}\mathbf{1}(\tau(f)<j)\right\|_\infty \\
&\leq a_j + a_{j-1} \\
&\lesssim a_j.
\end{aligned}
$$

When f varies over $\mathcal{F}$, $\lambda_{k(j,f)}^j - \lambda_{k(j-1,f)}^{j-1}$ varies over a collection of at most $\bar{N}_j := \prod_{l=0}^j N_l$ functions. From lemma 7.6, we thus have that

$$E\left[M_n\left(\lambda_{k(j,f)}^j - \lambda_{k(j-1,f)}^{j-1}\right) \mid A\right] \lesssim \frac{1}{\sqrt{n}}\left(\epsilon_j + \frac{M}{\sqrt{n}}\right)\sqrt{\log\left(1+\frac{\bar{N}_j}{P[A]}\right)}$$

$$
\begin{aligned}
&+ \frac{a_j (\log n)^2}{n} \log\left(1 + \frac{\bar{N}_j}{P[A]}\right) \\
\lesssim & \frac{1}{\sqrt{n}} \epsilon_j \sqrt{\log\left(1 + \frac{\bar{N}_j}{P[A]}\right)} + \frac{a_j (\log n)^2}{n} \log\left(1 + \frac{\bar{N}_j}{P[A]}\right),
\end{aligned}
$$

as $\epsilon_j \geq r^- \geq M n^{-1/2}$.

Control of the root. From the triangle inequality,

$$
\begin{aligned}
\sigma(\lambda^0_{k(j,f)}) \leq & \sigma(f - \lambda^0_{k(0,f)}) + \sigma(f) \\
\leq & 2\epsilon_0.
\end{aligned}
$$

In addition, we have that $\|\lambda^0_{k(j,f)}\|_\infty \leq M$ (if not, we can always, without loss of generality, truncate it to $[-M, M]$ without altering its bracketing properties).

Therefore,

$$
E\left[\sup_{f \in \mathcal{F}} M_n(f) \mid A\right] \lesssim \frac{\epsilon_0}{\sqrt{n}} \sqrt{\log\left(1 + \frac{N_0}{P[A]}\right)} + \frac{M (\log n)^2}{n} \log\left(1 + \frac{N_0}{P[A]}\right).
$$

Adding up the bounds. We obtain

$$
\begin{aligned}
E\left[\sup_{f \in \mathcal{F}} M_n(f) \mid A\right] \lesssim & \frac{\epsilon_0}{\sqrt{n}} \sqrt{\log\left(1 + \frac{\bar{N}_0}{P[A}\right)} + \frac{M (\log n)^2}{n} \log\left(1 + \frac{\bar{N}_0}{P[A]}\right) \\
& + \sum_{j=1}^{J} \frac{\epsilon_j}{\sqrt{n}} \sqrt{\log\left(1 + \frac{\bar{N}_j}{P[A]}\right)} + \frac{a_j (\log n)^2}{n} \log\left(1 + \frac{\bar{N}_0}{P[A]}\right) \\
& + \sum_{j=0}^{J-1} \frac{\epsilon_j^2}{a_j} + \epsilon_J.
\end{aligned}
$$

Set $a_j := \sqrt{n} (\log n)^{-1} (\log(1 + \bar{N}_j / P[A]))^{-1/2}$. Then above bound then becomes

$$
E\left[\sup_{f \in \mathcal{F}} M_n(f) \mid A\right] \lesssim \frac{\epsilon_J}{\sqrt{n}} + \frac{\log n}{\sqrt{n}} \sum_{j=0}^{J} \epsilon_j \sqrt{\log\left(1 + \frac{\bar{N}_j}{P[A]}\right)} + \frac{(\log n)^2}{n} M \log\left(1 + \frac{N_0}{P[A]}\right).
$$

Using the same arguments as at the end of the proof of theorem 5 in Bibaut et al. [2020], we have that

$$
\sum_{j=0}^{J} \epsilon_j \sqrt{\log\left(1 + \frac{\bar{N}_j}{P[A]}\right)} \lesssim \int_{\epsilon_J}^{\epsilon_0} \sqrt{\log\left(1 + N_{[]}(\epsilon, \mathcal{F}, \sigma)\right)} d\epsilon + \epsilon_0 \sqrt{\log\left(1 + \frac{1}{P[A]}\right)}.
$$

The above and the fact that $\log(1 + N_0 / P[A]) \leq \log(1 + N_0) + \log(1 + 1/P[A])$ yield the wished claim. □

7.C.2 Equicontinuity

Theorem 7.12. *Consider a class of functions $\mathcal{F}$ and a sequence (f_n) of elements of $\mathcal{F}$.*

Suppose that conditions 7.22, 7.23 and 7.24 hold.

Suppose that there exists a deterministic sequence of positive numbers (a_n) such that

$$a_n^{-2}(\log n)^2/\sqrt{n} = o(1) \qquad \textit{and} \qquad a_n^{\nu} \log n = o(1) \textit{ for every } \nu > 0,$$

and, for every $\epsilon > 0$, there exists $C_n(\epsilon) > 0$ such that

$$P\left[\forall n \geq 1, \|f_n\|_\infty \leq C(\epsilon) a_n\right] \geq 1 - \epsilon..$$

Suppose further that there

$$\log N_{[]}(u, \mathcal{F}, \sigma) \lesssim u^p \textit{ for some } p > 0.$$

Then

$$\sqrt{n} M_n(f) = o(1) \textit{ a.s.}$$

Proof of theorem 7.12. Let $\epsilon > 0$, and let $C(\epsilon)$ as in the conditions of the theorem, and introduce the event

$$\mathcal{E}_1(\epsilon) := \left\{\forall n \geq 1, \|f_n\|_\infty \leq C(\epsilon) a_n\right\}.$$

Observe that under $\mathcal{E}_1(\epsilon)$, for every $n \geq 1$,

$$\sqrt{n} M_n(f_n) \leq \sup\left\{\sqrt{n} M_n(f) : f \in \mathcal{F}, \|f_n\|_\infty \leq C(\epsilon) a_n\right\}.$$

We now bound with high probability the supremum in the right-hand side above, for every n. Let $x_n := \log(\epsilon/(n(n+1)))$. From theorem 7.11, with probability at least $1 - \epsilon/(n(n+1))$,

$$\sup\left\{\sqrt{n} M_n(f) : f \in \mathcal{F}, \|f_n\|_\infty \leq C(\epsilon) a_n\right\} \lesssim \psi_n(\epsilon, a_n, x_n),$$

with

$$\begin{aligned}\psi_n(\epsilon, a_n, x_n) :=& C(\epsilon) a_n + \log n \int_{\frac{C(\epsilon) a_n}{\sqrt{n}}}^{C(\epsilon) a_n} u^{-p/2} du + (C(\epsilon) a_n)^{-p} \frac{(\log n)^2}{\sqrt{n}} \\ &+ C(\epsilon) a_n \sqrt{\log(n(n+1)/\epsilon) \log n} + \frac{C(\epsilon) a_n}{\sqrt{n}} \log(n(n+1)/\epsilon)(\log n)^2.\end{aligned}$$

Therefore, from a union bound,

$$\begin{aligned}P\left[\exists n \geq 1 \sqrt{n} M_n(f_n) \geq \psi_n(\epsilon, a_n, x_n)\right] \leq& P[\mathcal{E}_1(\epsilon)^c] + \sum_{n=1}^{\infty} \frac{\epsilon}{n(n+1)} \\ \leq& 2\epsilon.\end{aligned}$$

Since, for every $\epsilon > 0$, condition 7.12 implies that $\psi_n(\epsilon, a_n, x_n) = o(1)$, the above implies that,

$$\forall \epsilon > 0, \; P\left[\lim_{n \to \infty} \sqrt{n} M_n(f) = 0\right] \leq 2\epsilon,$$

which, by letting $\epsilon \to 0$, implies the wished claim. □

Discussion of the supremum norm convergence requirement. As we pointed out in the main text, it might appear surprising at first that even though our Donsker condition involves the entropy w.r.t. the σ norm, which is an L_2 norm, we do need convergence in sup norm of $(f_n)_{n\geq 1}$.

The reason why this is the case can be understood from the expression and conditions of our maximal inequality for weakly dependent empirical processes, theorem 7.11 from the previous subsection. Recall that this result tells us that, under mixing conditions, given a class of functions $\mathcal{F}$ such that $\sup_{f\in\mathcal{F}} \sigma(f) \leq r$, and $\sup_{f\in\mathcal{F}} \|f\|_\infty \leq M$, for some $M, r > 0$, then for any $r^- \geq M/\sqrt{n}$, it holds with probability at least $1 - 2e^{-x}$ that

$$\sup_{f\in\mathcal{F}} M_n(f) \leq r^- + \frac{\log n}{\sqrt{n}} \int_{r^-}^{r} \sqrt{\log(1 + N_{[]}(\epsilon, \mathcal{F}, \sigma)} d\epsilon + r \log n \sqrt{\frac{x}{n}} + \frac{(\log n)^2 M x}{n}, \tag{7.10}$$

where $M_n(f) = n^{-1} \sum_{i=1}^n f(X_i) - Ef(X_i)$. Suppose that we want to prove an asymptotic equicontinuity result of the form $M_n(f_n) = o_P(n^{-1/2})$ for a certain sequence $(f_n)_{n\geq 1}$, while only assuming that $\sigma(f_n) = o_P(1)$ and not making any assumptions on $(\|f_n\|_\infty)_{n\geq 1}$. Then, as $n \to \infty$, we can bound $M_n(f_n)$ with high probability by the supremum of $M_n(f)$ over subsets of $\mathcal{F}$ with σ radius arbitrarily close to zero. This allows us to bound $M_n(f_n)$ with high probability by the right-hand side above with the upper bound r of the entropy integral arbitrarily close to zero. Unfortunately, letting the upper bound of the entropy integral converge to zero isn't enough to make the expression converge to zero faster than $n^{-1/2}$. Indeed, since the term r^- needs to be at least as large as $M/\sqrt{n}$, with M an upper bound on the $\|\cdot\|_\infty$ radius of the class over which we take the supremum, we need to be able to let M get arbitrarily close to zero with high probability to make the RHS of (7.10) go to zero faster than $n^{-1/2}$. This explains why, given our maximal inequality (7.10), we need to control $(\|f_n\|_\infty)_{n\geq 1}$.

That being said, one might still wonder why in our maximal inequality the lower bound r^- of the entropy integral needs to be larger than $M/\sqrt{n}$, thus making us pay an approximation error price $r^- \geq M/\sqrt{n}$. The reason is that we obtain our result by applying a chaining device to the following deviation bound Merlevède et al. [2009] for fixed f. Under exponential α-mixing (condition 7.22) and finiteness of the sum of ρ-mixing coefficients, if $\|f\|_\infty \leq M$, their result gives that

$$P\left[\sqrt{n} M_n(f) \geq x\right] \lesssim \exp\left(- \frac{x^2}{\sigma(f)^2 + \frac{M^2}{n} + \frac{Mx(\log n)^2}{\sqrt{n}}} \right), \tag{7.11}$$

Compare this with the usual Bernstein inequality i.i.d. random variables:

$$P\left[\sqrt{n} M_n(f) \geq x\right] \lesssim \exp\left(- \frac{x^2}{\|f\|_{2,P}^2 + \frac{Mx}{\sqrt{n}}} \right)$$

While the i.i.d. Bernstein inequality implies that $\sqrt{n} M_n(f)$ scales as the L_2 norm $\|f\|_{P,2}$ (as long as the ratio $\|f\|_{2,P}/\|f\|_\infty \gtrsim 1/\sqrt{n}$), the concentration bound for mixing sequences implies that it scales as $\sigma(f) + \|f\|_\infty/\sqrt{n}$. In the chaining argument, we consider ϵ_j-bracketings in $\|\cdot\|_{2,P}$ norm in the i.i.d. case, and in σ norm in the weakly dependent case, with $\epsilon_j = r2^{-j}$, for increasingly large j, where j has the interpretation of the depth of the chains.

Let us first discuss chaining in the i.i.d. case, and in the case where we want to obtain a bound on $E_P\sqrt{n}\sup_{f\in\mathcal{F}} M_n(f)$ (as opposed to obtaining a high probability bound on $\sup_{f\in\mathcal{F}} M_n(f)$, which is slightly more technical). Denote $\{[\lambda_{j,k}, \upsilon_{j,k}] : k \in [N_j]\}$ the ϵ_j-bracketing of $\mathcal{F}$ used in the chaining device. The contribution of depth j to the final bound on $E_P\sqrt{n}\sup_{f\in\mathcal{F}} M_n(f)$ is a supremum over links $\lambda_{j,k} - \lambda_{j-1,k}$ between depths j and $j-1$, which can essentially be bounded, using Bernstein's inequality, by

$$\sup_{k\in[N_j],k\in[N_{j-1}]} \|\lambda_{j,k} - \lambda_{j-1,k}\|_{P,2}\sqrt{\log N_{[]}(\epsilon_j, \mathcal{F}, \|\cdot\|_{P,2})} \lesssim \epsilon_j\sqrt{\log N_{[]}(\epsilon_j, \mathcal{F}, \|\cdot\|_{P,2})}.$$

(To be rigorous, the bound obtained from Bernstein's inequality has another term, but in adaptive chaining, we choose the maximal depth of the chains so that this term is no larger than the first one above). By contrast, in the weakly dependent case, the concentration bound (7.11) gives that the corresponding contribution is bounded by

$$\begin{aligned}&\left(\sup_{k\in[N_j],k\in[N_{j-1}]} \sigma(\lambda_{j,k} - \lambda_{j-1,k}) + \frac{M}{\sqrt{n}}\right)\sqrt{\log N_{[]}(\epsilon_j, \mathcal{F}, \sigma)}\\ \lesssim &\left(\epsilon_j + \frac{M}{\sqrt{n}}\right)\sqrt{\log N_{[]}(\epsilon_j, \mathcal{F}, \sigma)}.\end{aligned}$$

A consequence of this is that when the depth j of the chain is such that $\epsilon_j \leq M/\sqrt{n}$, then the term $M/\sqrt{n}$ becomes the main scaling factor. It can be checked that as result of this, the sum of these bounds diverges as $j \to \infty$. This is why in our chaining decomposition, we impose that our chains must have depth no larger that J such that $\epsilon_J \geq M/\sqrt{n}$. This gives us a bound involving an entropy integral with lower bound ϵ_J and an approximation error $\sqrt{n}\epsilon_J$.

7.C.3 Exponential deviation bound for empirical risk minimizers

Let ℓ be a functional defined on $\mathbb{R}^{\mathcal{X}}$, the space of functions $\mathcal{X} \to \mathbb{R}$, such that, for every $\mathbb{R}^{\mathcal{X}}$, $\ell(f)$ is a function $\mathcal{X} \to \mathbb{R}$. We call ℓ a loss function. For every $f : \mathcal{X} \to \mathbb{R}$, we define the population risk and the empirical risk as

$$R_n(f) := \frac{1}{n}\sum_{i=1}^n E[\ell(f)(X_i)] \qquad \text{and} \qquad \widehat{R}_n(f) := \frac{1}{n}\sum_{i=1}^n \ell(f)(X_i).$$

Let f_n be an empirical risk minimizer over $\mathcal{F}$, that is an element of $\mathcal{F}$ such that

$$\widehat{R}_n(f_n) = \inf_{f\in\mathcal{F}} \widehat{R}_n(f).$$

and $f^* \in \mathcal{F}$ be a minimizer of the population risk, that is a function such that

$$R_n(f^*) = \inf_{f\in\mathcal{F}} R_n(f).$$

In this section, we give exponential bounds on the excess population risk $R_n(f_n) - R_n(f^*)$, and on the norm $\sigma(f - f^*)$. We rely on the following assumptions.

Assumption 7.26 (Variance bound). *For every $f \in \mathcal{F}$, $\sigma^2(\ell(f) - \ell(f^*)) \lesssim (R_n(f) - R_n(f^*))$.*

Assumption 7.26 can be checked in common settings, such as in non-parametric regression settings when ℓ is the square loss and the dependent variable has bounded range.

Assumption 7.27 (A power of σ dominates $\|\cdot\|_\infty$ over $\ell(\mathcal{F})$). *There exists $\alpha \in (0,1)$ such that, for all $f \in \mathcal{F}$, $\|\ell(f) - \ell(f^*)\|_\infty \lesssim (\sigma(\ell(f) - \ell(f^*)))^\alpha$.*

Assumption 7.28 (Excess risk dominates norm of difference). *Suppose that $\sigma^2(f - f^*) \lesssim R_n(f) - R_n(f^*)$ for every $f \in \mathcal{F}$.*

Assumption 7.27 holds for instance if all functions in $\ell(\mathcal{F})$ are all L-Lipschitz for the same L, as formally presented in lemma 7.4.

Assumption 7.29 (Entropy). *There exists $p \in (0,2)$ such that*

$$\log N_{[]}(\epsilon, \ell(\mathcal{F}), \sigma) \lesssim \epsilon^{-p}.$$

Theorem 7.13 (Exponential deviation bound for ERM). *Suppose that $\mathcal{F}$ is a convex set, and that ℓ is convex on $\mathcal{F}$. Suppose that assumptions 7.22, 7.23, 7.26, 7.27, 7.28 and 7.29 hold. Let*

$$\phi_n : r \mapsto \frac{r^\alpha}{\sqrt{n}} + \frac{\log n}{\sqrt{n}} r^{1-p/2} + \frac{(\log n)^2}{n} r^{\alpha - p},$$

and let $r_n > 0$ such that $r_n^2/3 = \phi_n(r_n)$ (there exists such an r_n from lemma 7.7 applied to $3\phi_n$). Let $r > 0$ such that

$$r \geq \max\left\{ n^{-\frac{1}{2(1-\alpha)}}, r_n, \sqrt{3}\log n \sqrt{\frac{x}{n}}, (\log n)^{\frac{2}{2-\alpha}} \left(\frac{3x}{n}\right)^{\frac{1}{2-\alpha}} \right\}.$$

Then, with probability at least $1 - 2e^{-x}$, $R_n(f_n) - R_n(f^*) \lesssim r^2$ and $\sigma(f_n - f^*) \lesssim r$.

The proof of 7.13 is a relatively straightforward adaptation of the proof of lemma 13 in Bartlett et al. [2006]. It relies on the following two intermediate lemmas

Lemma 7.7. *Let $\phi : (0, \infty) \to \mathbb{R}_+$ such that $r \mapsto \phi(r)/r$ is strictly decreasing on $(0,\infty)$ and $\lim_{r \to 0^+} \phi(r)/r > 1$. Then, there exists a unique $r_* \in (0, \infty)$ such that $r_*^2 = \phi(r_*)$, and for any $r \in (0, \infty)$, $r^2 \geq \phi(r)$ if and only if $r \geq r^*$.*

Lemma 7.8. *Suppose that the assumptions of theorem 7.13 hold and let r_n and r be as defined in theorem 7.13. Then, there exists a constant $C > 0$ such that, for any $x > 0$,*

$$P\left[\sup\left\{M_n(\ell(f) - \ell(f^*)) : f \in \mathcal{F}, R_n(f) - R_n(f^*) \leq r^2\right\} \geq Cr^2\right] \leq 2e^{-x}.$$

Proof of lemma 7.7. The claim follows directly from the fact that, since both $r \mapsto \phi(r)/r$ and $r \mapsto 1/r$ are strictly decreasing on $(0, \infty)$, $r \mapsto \phi(r)/r^2$ is also strictly decreasing on $(0, \infty)$. □

Proof of lemma 7.8. From assumption 7.26,

$$\begin{aligned}&P\left[\sup\left\{M_n(\ell(f)-\ell(f^*)) : f\in\mathcal{F}, R_n(f)-R_n(f^*)\leq r^2\right\}\gtrsim r^2\right]\\ \leq& P\left[\sup\left\{M_n(\ell(f)-\ell(f^*)) : f\in\mathcal{F}, \sigma(\ell(f)-\ell(f^*))\lesssim r\right\}\gtrsim r^2\right].\end{aligned} \tag{7.12}$$

Under assumption 7.27, $\|M_n(\ell(f)-\ell(f^*))\|_\infty \lesssim r^\alpha$ for any $f\in\mathcal{F}$ such that $\sigma(\ell(f)-\ell(f^*))\lesssim r$. Therefore, since $r > n^{-1/(2(1-\alpha))}$, implies $r > r^\alpha n^{-1/2}$, applying theorem 7.11 with $r^- = r^\alpha/\sqrt{n}$, we have that

$$P\left[\sup\left\{M_n(\ell(f)-\ell(f^*)) : f\in\mathcal{F}, \sigma(\ell(f)-\ell(f^*))\lesssim r\right\}\gtrsim \psi_n(r,x)\right]\leq 2e^{-x},$$

with

$$\begin{aligned}\psi_n(r,x) :=& \frac{r^\alpha}{\sqrt{n}} + \frac{\log n}{\sqrt{n}}\int_{\frac{r^\alpha}{\sqrt{n}}}^{r} u^{-p/2}du + \frac{(\log n)^2}{n}r^{\alpha-p}\\ &+ r\log n\sqrt{\frac{x}{n}} + r^\alpha(\log n)^\alpha\frac{x}{n}.\end{aligned}$$

Observe that

$$\psi_n(r,x)\leq\phi_n(r) + r\log n\sqrt{\frac{x}{n}} + r^\alpha(\log n)^\alpha\frac{x}{n}. \tag{7.13}$$

From the definition of r in the statement of theorem 7.11, we have $r\geq r_n$, which from lemma 7.7 implies that $r^2/3\geq\phi_n(r)$. We also have $r^2/3\geq r\log n\sqrt{x/n}$, and $r^2/3\geq r^\alpha(\log n)^2x/n$. Therefore, $r^2\geq\psi_n(r,x)$. Therefore, from (7.12) and (7.13), we have

$$\begin{aligned}&P\left[\sup\left\{M_n(\ell(f)-\ell(f^*)) : f\in\mathcal{F}, \sigma(\ell(f)-\ell(f^*))\lesssim r\right\}\gtrsim r^2\right]\\ \leq& P\left[\sup\left\{M_n(\ell(f)-\ell(f^*)) : f\in\mathcal{F}, \sigma(\ell(f)-\ell(f^*))\lesssim r\right\}\gtrsim \psi_n(r,x)\right]\\ \leq& 2e^{-x}.\end{aligned}$$

□

Proof of theorem 7.13. From the convexity of $\mathcal{F}$ and the convexity of ℓ on $\mathcal{F}$, the following assertion holds for every $r\geq 0$:

$$\exists f\in\mathcal{F},\ \widehat{R}_n(f)-\widehat{R}_n(f^*)\leq 0 \text{ and } R_n(f)-R_n(f^*)\geq r^2$$

implies that

$$\exists f\in\mathcal{F},\ \widehat{R}_n(f)-\widehat{R}_n(f^*)\leq 0 \text{ and } R_n(f)-R_n(f^*)= r^2.$$

Using this fact, and the fact that by definition of f_n, $\widehat{R}_n(f_n)-\widehat{R}_n(f^*)\leq 0$, we have

$$P\left[\widehat{R}_n(f_n)-\widehat{R}_n(f^*)\geq r^2\right]$$

$$
\begin{aligned}
\leq & P\left[\exists f \in \mathcal{F}, \widehat{R}_n(f) - \widehat{R}_n(f^*) \leq 0 \text{ and } R_n(f) - R_n(f^*) \geq r^2\right] \\
\leq & P\left[\exists f \in \mathcal{F}, \widehat{R}_n(f) - \widehat{R}_n(f^*) \leq 0 \text{ and } R_n(f) - R_n(f^*) = r^2\right] \\
\leq & P\left[\sup\left\{M_n(\ell(f) - \ell(f^*)) : f \in \mathcal{F},\ \sigma(\ell(f) - \ell(f^*)) \lesssim r\right\} \geq r^2\right] \\
\leq & 2e^{-x}
\end{aligned}
$$

from lemma 7.8 □

7.D Proofs for the analysis of the TMLE

7.D.1 Proof of lemma 7.3 on the stabilization of the variance of the EIF

Proof of lemma 7.3. It is immediate to check that $\bar{D}_{T,N}(q_0)$ and $\bar{D}_{0,\infty,N}$ are centered, and therefore

$$
\begin{aligned}
\operatorname{Var}\left(\bar{D}_{T,N}(q_0)(L(t,i), C(t,i))\right) =& \|\bar{D}_{T,N}(q_0)\|^2_{2,q_0,h_{0,t,i}}, \\
\operatorname{Var}\left(\bar{D}_{0,\infty,N}(L_\infty, C_\infty)\right) =& \|\bar{D}_{0,\infty,N}\|^2_{2,q_0,h_{0,\infty,N}}.
\end{aligned}
$$

We have that

$$
\begin{aligned}
& \|\bar{D}_{T,N}(q_0)\|_{2,q_0,h_{0,t,i}} - \|\bar{D}_{0,\infty,N}\|_{2,q_0,h_{0,\infty,N}} \\
=& \|\bar{D}_{T,N}(q_0)\|_{2,q_0,h_{0,t,i}} - \|\bar{D}_{0,\infty,N}\|_{2,q_0,h_{0,t,i}} \\
& + \|\bar{D}_{0,\infty,N}\|_{2,q_0,h_{0,t,i}} - \|\bar{D}_{0,\infty,N}\|_{2,q_0,h_{0,\infty,N}}.
\end{aligned}
$$

We first start with the first term. We have that

$$
\begin{aligned}
& \left| \|\bar{D}_{T,N}(q_0)\|_{2,q_0,h_{0,t,i}} - \|\bar{D}_{0,\infty,N}\|_{2,q_0,h_{0,t,i}} \right| \\
\leq & \|\bar{D}_{T,N}(q_0) - \bar{D}_{0,\infty,N}\|_{2,q_0,h_{0,t,i}} \\
\leq & \sum_{s=1}^{\tau} \sum_{j=1}^{N} \left\| \frac{1}{\bar{h}_{0,T,N}} - \frac{1}{h_{0,\infty,N}} \right\|_{2,h_{0,t,i}} \|h^*_{0,s,j} \widetilde{D}_{s,j,N}(q_0)\|_\infty \\
\leq & 2B\boldsymbol{\varphi}\tau \left\| \frac{1}{\bar{h}_{0,T,N}} - \frac{1}{h_{0,\infty,N}} \right\|_{2,h_{0,t,i}} \\
=& o(1).
\end{aligned}
$$

The third line above follows from the triangle inequality. The fourth line above is a consequence of lemma 7.1. The fifth line follows from assumption 7.11.

We now turn to the second term. We have that

$$
\begin{aligned}
& \|\bar{D}_{0,\infty,N}\|^2_{2,q_0,h_{0,t,i}} - \|\bar{D}_{0,\infty,N}\|^2_{2,q_0,h_{0,\infty,N}} \\
=& E\left[\bar{D}^2_{0,\infty,N}(L(t,i), C(t,i))\right] - E\left[\bar{D}^2_{0,\infty,N}(L_\infty, C_\infty)\right] \\
& \to \infty 0,
\end{aligned}
$$

since, from assumption 7.12, $((L(t,i), C(t,i)) \to (L_\infty, C_\infty))$, and $\bar{D}_{0,\infty,N}$ is a bounded continuous function. □

7.D.2 Proof of the weak convergence of the martingale term

Proof of 7.6. Order the couples (t, i) as $(1, 1), \ldots, (1, N), \ldots, (T, 1), \ldots, (T, N)$, and let $(t(k), i(k))$ be the k-th couple in this ordering. Let

$$Z_{k,TN} := \frac{\bar{D}_{TN}(q_0)(X(t(k), i(k))}{\sigma_{0,\infty,N}\sqrt{TN}}.$$

Observe that under assumptions 7.9, 7.10, from lemma 7.2, $\|\bar{D}_{TN}(q_0)\|_\infty \leq C$, for some $C < \infty$ that does not depend on N.

Therefore, assumption (7.5) in theorem 7.5 is trivially checked. Let us now turn to assumption (7.6). First, observe that, as $E[Z^2_{k,TN}] = \mathrm{Var}_{q_0,h_{0,i(k),t(k)}}(\bar{D}_{TN}(q_0)(X(t(k), i(k))/(TN\sigma^2_{0,\infty,N})$, we have that $NTE[Z^2_{k,TN}] \to 1$ as $k \to \infty$, and therefore, by Cesaro's lemma for deterministic sequences of real valued numbers,

$$\sum_{k=1}^{\lfloor xTN \rfloor} E[Z^2_{k,TN}] \to x.$$

We now need to show that $V_{TN}(x) := \sum_{k=1}^{\lfloor xTN \rfloor} Z^2_{k,TN}$ converges in probability to its mean as $T \to \infty$. We proceed by taking the variance of $V_{TN}(x)$ and showing that it converges to zero, which, by Chebyshev's inequality will give us the wished result. We have

$$\begin{aligned}
\mathrm{Var}(V_{TN}(x)) = &\sum_{k=1}^{\lfloor xNT \rfloor} \mathrm{Var}(Z^2_{k,TN}) + \sum_{1 \leq k_1 < k_2 \leq \lfloor xNT \rfloor} \mathrm{Cov}(Z^2_{k_1,TN}, Z^2_{k_2,TN}) \\
\leq &xNT\|Z_{k,TN}\|^4_\infty + \sum_{k_1=1}^{\lfloor xNT \rfloor} \sum_{k_2=k_1+1}^{\lfloor xNT \rfloor} \|Z_{k_1,TN}\|^4_\infty \alpha\left(X(t(k_1), i(k_1)), X(t(k_2), i(k_2)\right) \\
\leq &xNT \frac{C^4}{(TN)^2\sigma^4_{0,\infty,N}}(1 + o(NT)) \\
= &o(1),
\end{aligned}$$

where the third line follows from assumption 5. By Chebyshev's inequality, we thus obtain that

$$\sum_{k=1}^{\lfloor xNT \rfloor} Z^2_{k,TN} - E\left[Z^2_{k,TN}\right] \to 0,$$

which implies that (7.6). This concludes the proof. □

7.E Proof of the exponential deviation bound for nuisance estimators

Proof of theorem 7.9. Most of the work in this proof is to check the conditions of our generic theorem 7.13 for empirical risk minimizers, in particular the variance bound (assumption 7.26, the assumption connecting the $\|\cdot\|_\infty$ norm to the norm σ (assumption 7.27), and the entropy bound. In doing so, we follow closely the techniques presented in section 3.4.1, chapter 3.4 of van der Vaart and Wellner [1996] for the analysis of maximum likelihood estimators.

Notation. We introduce the alternative loss

$$\widetilde{\ell}_n(q)(c,o) := -\log\left(\frac{q+q_n}{2q_n}(o \mid c)\right),$$

the corresponding empirical and population risks

$$\widehat{\widetilde{R}}_n(q) := \frac{1}{n}\sum_{i=1}^n \widetilde{\ell}_n(q)(\widetilde{C}(k),\widetilde{O}(k)) \qquad \text{and} \qquad \widetilde{R}_{0,n} := \frac{1}{n}\sum_{i=1}^n E_{q_0,\widetilde{h}_k}\left[\widetilde{\ell}_n(q)(\widetilde{C}(k),\widetilde{O}(k))\right].$$

For any c and any two conditional densities $q_1(\cdot \mid c)$ and $q_1(\cdot \mid c)$, we introduce the conditional Hellinger distance:

$$H(q_1,q_2 \mid c) := \left(\int \left(\sqrt{q_1(o,c)} - \sqrt{q_2(o,c)}\right)^2 do\right)^{1/2}.$$

For any marginal density $h : \mathcal{C} \to \mathbb{R}$, and any two q_1 and q_2, we define the conditional Hellinger distance integrated against h:

$$H_h(q_1,q_2) := \left(\int H^2(q_1,q_2 \mid c)h(c)dc\right)^{1/2}.$$

We further define

$$H_n(q_1,q_2 \mid c) = H(q_1+q_n, q_2+q_n \mid c) \qquad \text{and} \qquad H_{h,n}(q_1,q_2) = H_h(q_1+q_n, q_2+q_n).$$

For a conditional density $q(\cdot \mid c)$, a positive number number $p \geq 1$, let, for any $f : \mathcal{O} \times \mathcal{C} \to \mathbb{R}$,

$$\|f(\cdot,c)\|_{q(\cdot|c),p} := \left(\int |f(c,o)|^p\, q(o \mid c)do\right)^{1/p},$$

$$\text{and } \|f(\cdot \mid c)\|_{q(\cdot|c),B} := \left(\sum_{p\geq 2} \frac{\|f(\cdot,c)\|^p_{q(\cdot|c),p}}{p!}\right)^{1/2},$$

be the L_p norm, and the so-called Bernstein "norm" [1] with respect to $q(\cdot \mid c)$.

[1] It is not actually a norm, but this doesn't matter for what follows.

Checking the entropy condition. We have that

$$\begin{aligned}
&\left\| \widetilde{\ell}_n(q_1) - \widetilde{\ell}_n(q_2)(\cdot, c) \right\|^2_{q_0(\cdot|c),2} \\
\leq &\left\| \widetilde{\ell}_n(q_1) - \widetilde{\ell}_n(q_2)(\cdot, c) \right\|^2_{q_0(\cdot|c),B} \\
\lesssim &H^2_n(q_1, q_2 \mid c) \\
= &\int \left(\frac{q_1 - q_2}{\sqrt{q_1 + q_n} + \sqrt{q_2 + q_n}} \right)^2 (o \mid c) do \\
\lesssim &\int (q_1 - q_2)^2 (o \mid c) do.
\end{aligned}$$

The inequality in the third line above is proven to hold under assumption 7.19 in section 3.4.1 of van der Vaart and Wellner [1996]. The last line follows assumption 7.18.

Let $\mu_{\mathcal{O}}$ be the Lebesgue measure on $\mathcal{O}$. Integrating the previous inequality against $\widetilde{h}_i$, and recalling the definition of σ, we have that

$$\begin{aligned}
(1 + 2\boldsymbol{\rho})^{-1/2} \sigma \left(\widetilde{\ell}_n(q_1) - \widetilde{\ell}_n(q_2) \right) := &\sup_{i \geq 1} \left\| \widetilde{\ell}_n(q_1) - \widetilde{\ell}_n(q_2) \right\|_{q_0, \widetilde{h}_i, 2} \\
\lesssim &\sup_{i \geq 1} \| q_1 - q_2 \|_{\mu_{\mathcal{O}}, \widetilde{h}_i, 2} \\
\lesssim &\| q_1 - q_2 \|_{\mu, 2},
\end{aligned}$$

where the last inequality follows from assumption 7.20. Therefore, denoting $\widetilde{\ell}_n(\mathcal{Q}) := \{ \widetilde{\ell}_n(q) : q \in \mathcal{Q} \}$, we have that

$$\begin{aligned}
\log N_{[]}(\epsilon, \widetilde{\ell}_n(\mathcal{Q}), \sigma) \lesssim &\log N_{[]}(\epsilon, \mathcal{Q}, L_2(\mu)) \\
\lesssim &\epsilon^{-1} \log(1/\epsilon))^{2(d-1)},
\end{aligned}$$

where the last inequality is the claim of proposition 7.1.

Checking the variance bound condition. The first claim of theorem 3.4.4 in van der Vaart and Wellner [1996] asserts that

$$H^2(q, q_n \mid c) \lesssim \int \left(\widetilde{\ell}_n(q) - \widetilde{\ell}_n(q_n) \right) (c, o) q_0(o \mid c) do.$$

In section 3.4.1 of van der Vaart and Wellner [1996], the authors also show the following claim, which we transpose to our notation:

$$\left\| \left(\widetilde{\ell}_n(q) - \widetilde{\ell}_n(q_n) \right) (c, \cdot) \right\|_{q_0(\cdot|c), B} \lesssim H^2(q, q_n \mid c).$$

Therefore, putting the previous two inequalities together, integrating w.r.t. $\bar{\widetilde{h}}_n$ and recalling the definition of $\widetilde{R}_{0,n}$, we have that

$$\left\| \widetilde{\ell}_n(q) - \widetilde{\ell}_n(q_n) \right\|_{q_0, \bar{\widetilde{h}}_n, 2} \lesssim \widetilde{R}_{0,n}(q) - \widetilde{R}_{0,n}(q_n).$$

Using assumption 7.21 then yields that

$$\sigma^2(q, q_n) \lesssim \widetilde{R}_{0,n}(q) - \widetilde{R}_{0,n}(q_n),$$

which is the wished variance bound condition.

Checking assumption 7.27. Lemma 7.4 gives us that assumption 7.27 holds for $\alpha = 1/(d+1)$.

Upper bounding the rate of convergence. We calculate r_n defined in theorem 7.13. Observing that, from lemma 7.1, it holds for any $\nu > 0$ that $\log N_{[]}(\epsilon, \mathcal{Q}, \sigma) \lesssim \epsilon^{-(1+\nu)}$, we have

$$\phi_n(r) = \frac{r^{\frac{1}{d+1}}}{\sqrt{n}} + \log n \frac{r^{\frac{1-\nu}{2}}}{\sqrt{n}} + \frac{(\log n)^2}{n} r^{\frac{1}{d+1} - 1 - \nu}.$$

For $d \geq 2$, ν small enough, and n large enough, it is straightforward to observe

$$n^{-\frac{2}{4-2\alpha}} \gtrsim \phi_n(n^{-\frac{1}{4-2\alpha}}),$$

which, from lemma 7.7 implies that $r_n \lesssim n^{-\frac{1}{4-2\alpha}}$. We have thus checked the assumptions of theorem 7.13 and shown that r_n is upper bounded by the wished rate, which implies the claim. □

7.F Proof of the type-I error guarantee for the adaptive stopping rule (theorem 7.10)

Proof of theorem 7.10. Under H_0, we have that $\Psi(P_0^{T,N}) = 0$. That the procedure rejects is therefore equivalent to the following event:

$$\begin{aligned}
&\left\{ \exists T \in [t_0 T_{\max}, T_{\max}],\ \sqrt{\frac{T}{T_{\max}}} \sqrt{TN} \frac{\left(\widehat{\Psi}_{T,N} - \Psi(P_0^{T,N})\right)}{\sigma_{0,\infty,N}} \in [\pm a_\alpha(T/T_{\max})] \right\} \\
= &\left\{ \exists T \in [t_0 T_{\max}, T_{\max}],\ \frac{T}{T_{\max}} \sqrt{T_{\max} N} \frac{\left(\widehat{\Psi}_{T,N} - \Psi(P_0^{T,N})\right)}{\sigma_{0,\infty,N}} \in [\pm a_\alpha(T/T_{\max})] \right\} \\
= &\left\{ \sup_{t \in [t_0, 1]} t \sqrt{T_{\max} N} \frac{\left| \widehat{\Psi}_{t T_{\max}, N} - \Psi(P_0^{t T_{\max}, N}) \right|}{\sigma_{0,\infty,N} a_\alpha(t)} \leq 1 \right\}
\end{aligned}$$

Let ϕ be the function defined on the set $\mathbb{D}([0,1])$ of real-valued cadlag functions on $[0,1]$ by

$$\phi : f \mapsto \sup_{t \in [t_0, 1]} \frac{|f(t)|}{a_\alpha(t)}.$$

For $([\pm a_\alpha(t) : t \in [0,1])$ to be an $(1-\alpha)$ joint confidence band for $(W(t) : t \in [0,1])$, each $[\pm a_\alpha(t)]$ must contain $W(t)$ with probability at least $1-\alpha$, and therefore, for every $t \geq t_0$, we must have $a_\alpha(t) \geq \sqrt{t_0} q_{1-\alpha/2}$, where $q_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal. Therefore, the denominator in the definition of ϕ remains uniformly in t bounded away from 0, thus ensuring that ϕ is continous w.r.t $\|\cdot\|_\infty$, and bounded.

Therefore, from the fact that,

$$\left\{t\sqrt{T_{\max}N}\left(\widehat{\Psi}_{N,T} - \Psi(P_0^{T,N})\right) : t \in [t_0, 1]\right\} \xrightarrow{d} W,$$

by definition of weak convergence, and continuity and boundedness of ϕ as a mapping $(\mathbb{D}([0,1]), \|\cdot\|_\infty) \to (\mathbb{R}, |\cdot|)$, we have that

$$\begin{aligned}
&\lim_{T_{\max}\to\infty} P_0\left[\sup_{t\in[t_0,1]} t\sqrt{T_{\max}N}\frac{\left|\widehat{\Psi}_{tT_{\max},N} - \Psi(P_0^{tT_{\max},N})\right|}{\sigma_{0,\infty,N}a_\alpha(t)} \leq 1\right]\\
=&P_0\left[\sup_{t\in[t_0,1]}\frac{|W(t)|}{a_\alpha(t)} \leq 1\right]\\
\geq&1-\alpha,
\end{aligned}$$

where the latter inequality follows by definition of $a_\alpha(t)$ □

Chapter 8

Sufficient and insufficient conditions for the stochastic convergence of Cesaro means

Aurélien Bibaut, Alexander Luedtke, Mark van der Laan

Cesaro means and sequential causal inference. Cesaro means often arise in estimators of statistical parameters in sequential decision problems. Suppose for instance that we are in the stochastic contextual bandit setting, as presented in subsection 1.2.1, and that we have collected a sequence of triples context-action-rewards $(O(t))_{t \in [n]} = (X(t), A(t), Y(t))_{t \in [T]}$, where action $A(t)$ is drawn from policy g_t. Suppose we care about estimating and making inference for the value of a counterfactual policy g^*. One valid estimator for this task if the stabilized one-step estimator of Luedtke and van der Laan [2016]:

$$\widehat{\Psi}_T := \left(\frac{1}{T} \sum_{t=1}^{T} \widehat{\sigma_t}^{-1} \right)^{-1} \frac{1}{T} \sum_{t=1}^{T} \widehat{\sigma_t}^{-1} D_t(O(t)),$$

where

$$D_t(o) = \frac{g^*(a \mid x)}{g_t(a \mid x)} (y - \widehat{\bar{Q}}_{t-1}(a, x)) + \sum_{a'=1}^{K} g^*(a' \mid x) \widehat{\bar{Q}}_{t-1}(a', x),$$

with $\widehat{\bar{Q}}_{t-1}$ an $O(1), \ldots, O(t-1)$-measurable estimator of the outcome regression function $\bar{Q}(a, x)$ $:= E[Y(1) \mid A(1) = a, X(1) = x]$, and $\widehat{\sigma}_t$ an $\widehat{\bar{Q}}_{t-1}$ an $O(1), \ldots, O(t-1)$ of

$$\sigma_t := \sqrt{\mathrm{Var}(D_t(O(t) \mid O(1), \ldots, O(t-1)))}.$$

A condition in the analysis of $\widehat{\Psi}_T$ is the convergence of the Cesaro mean of estimators $T^{-1} \sum_{t=1}^{T} \widehat{\sigma}_t^{-1}$.

Cesaro means also appear in other estimation problems where estimators are computed in an online fashion, as we illustrate in section 8.2 further down.

Our contribution. We study the stochastic convergence of the Cesàro mean of a sequence of random variables. We show that establishing a rate of convergence in probability for a sequence is not sufficient in general to establish a rate in probability for its Cesàro mean. We also present several sets of conditions on the sequence of random variables that are sufficient to guarantee a rate of convergence for its Cesàro mean. We identify common settings in which these sets of conditions hold.

8.1 Introduction

The following fact is well known [Cauchy, 1821, Cesàro, 1888] for deterministic real-valued sequences $(x_n)_{n \geq 1}$:

$$n^\beta x_n \to 0 \text{ for some } \beta \geq 0 \quad \Longrightarrow \quad n^\beta \bar{x}_n := n^\beta \frac{1}{n} \sum_{i=1}^{n} x_n \to 0. \tag{8.1}$$

In this note, we investigate the extent to which this kind of result carries over to a sequence $(X_n)_{n \geq 1}$ of random variables defined on a complete probability space $(\mathcal{X}, \mathcal{A}, P)$. Specifically, we aim to answer the following questions:

Question 8.1. *Is $n^\beta X_n \xrightarrow{p} 0$ sufficient to ensure that $n^\beta \bar{X}_n := n^\beta \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{p} 0$?*

Question 8.2. *Do reasonable conditions on $(X_n)_{n\geq 1}$ imply the convergence of $n^\beta \bar{X}_n$ in probability? almost surely? in mean?*

Question 8.3. *Does knowing that $n^\beta X_n$ satisfies an exponential tail bound imply a similar bound for $n^\beta \bar{X}_n$?*

Generalizing the deterministic result (8.1) to the stochastic case is important in many statistical problems that have an online or sequential component [e.g., Luedtke and van der Laan, 2016]. In these settings, X_n is often a function of an estimator computed on data available at time n. For example, X_n may be equal to $\widehat{\theta}_n - \theta_0$, where θ_0 is a scalar statistical parameter and $\widehat{\theta}_n$ is an estimator of θ_0 based on the first n observations. Alternatively, X_n may be an excess risk $R(\widehat{\theta}_n) - \inf_{\theta\in\Theta} R(\theta)$, where R a risk function and Θ is an indexing set.

Guarantees for estimators are generally stated in terms of some form of stochastic convergence, where the type of convergence established varies depending on the setting. For example, results for empirical risk minimizers (also called minimum contrast estimators or M-estimators) have been given in terms of rates in probability [e.g., van der Vaart and Wellner, 1996] and exponential tail bounds on excess risks [e.g., Bartlett et al., 2005, 2006]. Convergence rates for kernel density and kernel regression estimators are often given in probability [see e.g. Hansen, 2008], in mean squared error [see e.g. Tsybakov, 2008], or almost surely [see e.g. Hansen, 2008].

Section 8.3 answers 8.1 in the negative via a counterexample. Section 8.4 answers 8.2 in the affirmative for convergence in mean, and Section 8.5 similarly answers this question for almost sure convergence. Since convergence in mean or convergence almost surely imply convergence in probability, these sections also yield reasonable conditions for the convergence in probability of $n^\beta \bar{X}_n$. Section 8.6 answers 8.3 in the affirmative, and also evaluates the implications of this finding for empirical risk minimizers.

Whenever we do not make it explicit in the notation, we use the convention that probabilistic notions are with respect to the measure P. This convention is applied to expectations E, almost sure convergence, convergence in mean, and $L^r(P)$ norms $\|\cdot\|_r$. Here we recall that $\|f\|_r := \{\int |f(\omega)|^r dP(\omega)\}^{1/r}$ when $r \in (1,\infty)$ and that $\|f\|_\infty$ denotes the P-essential supremum. We call the sequence (X_n) uniformly bounded if $(\|X_n\|_\infty)_{n\geq 1}$ is a bounded sequence.

8.2 Motivating examples

8.2.1 Online estimator of the Bayes risk in binary classification.

Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are n i.i.d. copies of a couple of random variables (X, Y), with X a vector of predictors and $Y \in \{-1, 1\}$ a binary label. Denote $\eta(x) := \Pr(Y = 1 \mid X = x)$, and let $f_\eta(x) := \operatorname{sign}\{2\eta(x) - 1\}$ be the Bayes classifier. For any classifier f, consider $\ell(f)(x, y) := \mathbf{1}[y \neq \operatorname{sign}\{f(x)\}]$ the 0-1 classification loss of f, and let $R(f) := E\{\ell(f)(X, Y)\}$, the corresponding classification risk. Say we want to estimate the Bayes risk $R^* := R(f_\eta)$. Suppose that $(\widehat{f}_i)_{i\geq 1}$ is an $(\mathcal{H}_i)_{i\geq 1}$-adapted sequence of estimators of the Bayes classifiers f_η, where

$\mathcal{H}_i := \sigma\{(X_1, Y_1), \ldots, (X_i, Y_i)\}$ is the filtration induced by the first i observations. Consider the online estimator $\widehat{R}_n := n^{-1} \sum_{i=1}^n \ell(\widehat{f}_{i-1})(X_i, Y_i)$ of R^*. It can be checked that the following decomposition holds:

$$\widehat{R}_n - R^* = \frac{1}{n} \sum_{i=1}^n \ell(f_{i-1})(X_i, Y_i) - E\left\{\ell(f_{i-1})(X_i, Y_i) \mid \mathcal{H}_{i-1}\right\} + \frac{1}{n} \sum_{i=1}^n \{R(\widehat{f}_{i-1}) - R^*\}.$$

The first average can be easily checked to be $O(n^{-1/2})$ with high probability via Azuma-Hoeffding. We would then like to show that the second average is $o(n^{-1/2})$ in some stochastic sense. It is known that, under some well-studied assumptions, the individual terms $R(\widehat{f}_{i-1}) - R^*$ can be shown to converge faster than $i^{-1/2}$ [see, e.g., Audibert and Tsybakov, 2007]. We would like to be able to prove the same for their average.

8.2.2 Online estimator of the mean outcome under missingness at random.

Consider $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$ to be a random couple, with, for instance, X having the interpretation of an individual's demographics and Y representing a person's vote intention. Suppose that R is a third random variable, representing whether a person's outcome is measured. We observe i.i.d. copies $Z_1 := (X_1, R_1, R_1 Y_1), \ldots, Z_n := (X_n, R_n, X_n Y_n)$ of $Z := (X, R, RY)$.

The objective is to estimate $\Psi(P) := E_P\{E_P(Y \mid R = 1, X)\}$, which under some assumptions (missingness at random of the outcome measurement, and non-zero probability of the conditioning event), equals the mean outcome Y across the entire population. Let $(\widehat{Q}_i)_{i \geq 1}$, and $(\widehat{g}_i)_{i \geq 1}$ be sequences of $(\mathcal{H}_i)_{i \geq 1}$-adapted estimators of the conditional missingness probability $g : (r, x) \mapsto \Pr_P(R = r \mid X = x)$ and outcome regression function $\bar{Q} : (r, x) \mapsto E_P(Y \mid R = r, X = x)$. Then, denoting $D(P)(x, r, y) := \{g(r, x)\}^{-1} r\{y - \bar{Q}(y, x)\} + Q(1, x) - \Psi(P)$, the online estimator $\widehat{\Psi}_n := n^{-1} \sum_{i=1}^n \Psi(\widehat{P}_{i-1}) + D(\widehat{P}_{i-1})(Z_i)$ admits the following decomposition:

$$\widehat{\Psi}_n - \Psi(P) = \frac{1}{n} \sum_{i=1}^n D(\widehat{P}_{i-1})(Z_i) - E_P\{D(\widehat{P}_{i-1})(Z_i) \mid \mathcal{H}_{i-1}\} + \frac{1}{n} \sum_{i=1}^n \mathrm{Rem}(\widehat{P}_{i-1}, P),$$

where $\mathrm{Rem}(\widehat{P}_{i-1}, P) := E_P[\{g(R, X)\}^{-1}\{\widehat{g}_{i-1}(R, X) - g(R, X)\}\{\widehat{Q}(R, X) - \bar{Q}(R, X)\}]$ is a remainder term. If g is uniformly lower bounded over its domain by some $\delta > 0$, then the Cauchy-Schwarz inequality implies that $\mathrm{Rem}(\widehat{P}_{i-1}, P) \leq \delta^{-1} \|\widehat{g} - g\|_2 \|\widehat{Q} - \bar{Q}\|_2$. Convergence guarantees on $\mathrm{Rem}(\widehat{P}_{i-1}, P)$ can therefore be obtained from convergence guarantees on $\widehat{g}$ and $\widehat{Q}$. Analyzing the online estimator $\widehat{\Psi}_n$ requires characterizing the stochastic convergence of the average of the remainder terms.

8.3 An example where $n^\beta X_n$ converges to zero in probability, yet $n^\beta \bar{X}_n$ does not

The following counterexample shows that $n^\beta X_n \xrightarrow{p} 0$ does not generally even imply that $(n^\beta \bar{X}_n)_{n \geq 1}$ is uniformly tight, even if the further condition is imposed that the random variables

$(X_n)_{n\geq 1}$ are uniformly bounded. Therefore, it is certainly not the case that $n^\beta X_n \xrightarrow{p} 0$ implies that $n^\beta \bar{X}_n \xrightarrow{p} 0$.

Proposition 8.1. *For any $\beta \in (0,1)$ and $b > 0$, there exists a sequence of random variables $(X_n)_{n\geq 1}$ such that (1) $n^\beta \bar{X}_n = o_p(1)$ and (2) $|X_n| \leq b$ a.s. for all n, and such that $(n^\beta \bar{X}_n)_{n\geq 1}$ is not uniformly tight.*

Proof. Without loss of generality, suppose that $b = 1$. Fix $\beta \in (0,1)$ and $\alpha \in (0,\beta)$. For all $n \geq 1$, let $p_{n,\alpha} := (2^{\lfloor \log_2 n \rfloor})^{-\alpha}$. Consider a sequence of independent random variables $(X_n)_{n\geq 1}$ such that, for all $n \geq 1$, $X_n \sim$ Bernoulli$(p_{n,\alpha})$. The definition of $(p_{n,\alpha})_{n\geq 1}$ ensures that for every $k \geq 1$, $X_{2^{k-1}}, \ldots, X_{2^k - 1}$ is a block of 2^{k-1} i.i.d. observations with marginal distribution Bernoulli$(p_{2^{k-1},\alpha})$.

Observe that, for any $M > 0$, $\Pr(X_n \geq Mn^{-\beta}) = p_{n,\alpha} \to 0$, that is, $X_n = o_p(n^{-\beta})$ holds. We will show that $\bar{X}_n$ is not uniformly tight, which implies in particular that it is not true that $\bar{X}_n = O_p(n^{-\beta})$. In what follows, we will denote $\bar{X}_{n_1:n_2} := (n_2 - n_1 + 1)^{-1} \sum_{i=n_1}^{n_2} X_i$.

Fix $M > 0$ and $k \geq 1$. For $n = 2^k$, we have that

$$\Pr\left(\bar{X}_{n-1} \geq n^{-\beta} M\right) = \Pr\left(\frac{1}{n}\sum_{i=n/2}^{n-1} X_i \geq n^{-\beta} M\right) = \Pr\left(\bar{X}_{n/2:n-1} \geq 2n^{-\beta} M\right)$$

$$=\Pr\left[\left\{\frac{n}{p_{n/2,\alpha}(1-p_{n/2,\alpha})}\right\}^{1/2} \left(\bar{X}_{n/2:n-1} - p_{n/2,\alpha}\right) \right.$$
$$\left. \geq \left\{\frac{n}{p_{n/2,\alpha}(1-p_{n/2,\alpha})}\right\}^{1/2} \left(2n^{-\beta}M - p_{n/2,\alpha}\right)\right]. \tag{8.2}$$

We now use the Berry–Esseen theorem to lower bound the last line in the above display. We have that, for every $i \in \{n/2, \ldots, n-1\}$, $E(X_i) = p_{n/2,\alpha}$, $E\{(X_i - p_{n/2,\alpha})^2\} = p_{n/2,\alpha}(1-p_{n/2,\alpha})$, and

$$\begin{aligned} E\{|X_i - E(X_i)|^3\} &= p_{n/2,\alpha}(1-p_{n/2,\alpha})^3 + (1-p_{n/2,\alpha})p_{n/2,\alpha}^3 \\ &= p_{n/2,\alpha}(1-p_{n/2,\alpha})\left\{(1-p_{n/2,\alpha})^2 + p_{n/2,\alpha}^2\right\} \\ &\leq p_{n/2,\alpha}(1-p_{n/2,\alpha}). \end{aligned}$$

Hence, $E\{|X_i - E(X_i)|^3\}/\text{var}(X_i) \leq 1$ for every $i = n/2, \ldots, n-1$. Using the Berry–Esseen bound, and letting Φ denote the cumulative distribution function of the standard normal distribution, we see that

$$\Pr\left[\left\{\frac{n}{p_{n/2,\alpha}(1-p_{n/2,\alpha})}\right\}^{1/2} \left(\bar{X}_{n/2:n-1} - p_{n/2,\alpha}\right)\right.$$
$$\left. \geq \left\{\frac{n}{p_{n/2,\alpha}(1-p_{n/2,\alpha})}\right\}^{1/2} \left(2n^{-\beta}M - p_{n/2,\alpha}\right)\right]$$
$$\geq 1 - \Phi\left[\left\{\frac{n}{p_{n/2,\alpha}(1-p_{n/2,\alpha})}\right\}^{1/2} \left(2n^{-\beta}M - p_{n/2,\alpha}\right)\right] - \frac{C}{\sqrt{n}}, \tag{8.3}$$

where C is a universal positive constant. Noting that $p_{n/2,\alpha} = (n/2)^{-\alpha}$, we see that, for all $n = 2^k$ large enough, $p_{n/2,\alpha} \leq 1/2$, and so, for such n,

$$\left\{ \frac{n}{p_{n/2,\alpha}(1 - p_{n/2,\alpha})} \right\}^{1/2} (2Mn^{-\beta} - p_{n/2,\alpha}) \geq 2^{(1-\alpha)/2} n^{(1+\alpha)/2} (2Mn^{-\beta} - [n/2]^{-\alpha}),$$

and the right-hand side diverges to $-\infty$ as $n \to \infty$ since $0 < \alpha < \beta < 1$. Hence, the right-hand side of (8.3) converges to 1 as $n \to \infty$. Combining this with (8.2) and recalling that (8.2) assumed that $n = 2^k$ shows that $\Pr(\bar{X}_{2^k-1} \geq 2^{-k\beta} M) \to 1$, and so there exists an infinite subsequence (n_k) of the natural numbers such that $\Pr(\bar{X}_{n_k - 1} \geq n_k^{-\beta} M) \to 1$. As $M > 0$ was arbitrary, $n^\beta \bar{X}_n$ is not uniformly tight. □

8.4 Convergence in mean

The following proposition shows that convergence in mean of $n^\beta X_n$ implies convergence in mean of $n^\beta \bar{X}_n$.

Proposition 8.2. *Suppose that $E(|X_n|) = o(n^{-\beta})$. Then $E(|\bar{X}_n|) = o(n^{-\beta})$.*

Proof. From the triangle inequality, $n^\beta E(|\bar{X}_n|) \leq n^\beta \times n^{-1} \sum_{i=1}^n E(|X_i|)$. From (8.1) applied to the deterministic sequence $\{E(|X_i|)\}_{n \geq 1}$, we have that $n^\beta \times n^{-1} \sum_{i=1}^n E(|X_i|) \to 0$, which establishes the claim. □

The above proposition can be restated by recalling that, if $X_n \xrightarrow{p} 0$, then the convergence in mean of X_n to zero is equivalent to the asymptotic uniform integrability of $(X_n)_{n \geq 1}$ [Theorem 2.20 in Van der Vaart, 2000]. Therefore, the above proposition immediately yields the following corollary.

Corollary 8.1. *Suppose that $n^\beta X_n \xrightarrow{p} 0$ and also that $(n^\beta X_n)_{n \geq 1}$ is asymptotically uniformly integrable, in the sense that*

$$\lim_{x \to \infty} \limsup_{n \to \infty} n^\beta E\{|X_n| \mathbf{1}(n^\beta |X_n| > x)\} = 0. \tag{8.4}$$

Then, $E(|\bar{X}_n|) = o(n^{-\beta})$.

The above can be used to prove the following corollary.

Corollary 8.2. *Fix $r \in (1, \infty]$ and let q denote the Hölder conjugate of r. Suppose that $n^\beta X_n \xrightarrow{p} 0$ in probability and that r is such that $(\|X_n\|_r)_{n \geq 1}$ is a bounded sequence. If*

$$\lim_{x \to \infty} \limsup_{n \to \infty} n^{\beta q} \Pr(n^\beta |X_n| > x) = 0, \tag{8.5}$$

then $E(|\bar{X}_n|) = o(n^{-\beta})$.

Proof. For any $\beta \geq 0$ and $n \geq 1$, Hölder's inequality shows that $n^\beta E\{|X_n|\mathbf{1}(n^\beta|X_n| > x)\} \leq n^\beta \|X_n\|_r \mathrm{Pr}(n^\beta|X_n| > x)^{1/q}$. Since $(\|X_n\|_r)_{n\geq 1}$ is bounded and $z \mapsto z^q$ is continuous at zero, (8.5) implies (8.4), and so the result follows by Corollary 8.1. □

In the special case where $\beta = 0$ and $r = \infty$ (and, therefore, $q = 1$), (8.5) automatically follows from the condition that $X_n \xrightarrow{p} 0$. Put another way, if $X_n \xrightarrow{p} 0$ and $(X_n)_{n\geq 1}$ is uniformly bounded, then $E(|\bar{X}_n|) = o(1)$.

Observe that, in the context of the counterexample from the proof of Proposition 8.1, Corollary 8.1 (applied with $r = \infty$) shows that for any $\beta < \alpha$, $\bar{X}_n = O_p(n^{-\beta})$.

8.5 Almost sure convergence

Proposition 8.3. *If $n^\beta X_n \to 0$ almost surely, then $n^\beta \bar{X}_n \to 0$ almost surely.*

Proof. Let $\mathcal{E}$ be the event $\{n^\beta X_n \to 0\}$. That $n^\beta X_n \to 0$ almost surely means that $\mathrm{Pr}(\mathcal{E}) = 1$. Suppose that $\mathcal{E}$ holds. Then, by (8.1) applied to the realization of the sequence $(X_n)_{n\geq 1}$, we have that $n^\beta \bar{X}_n \to 0$. Therefore, $\mathrm{Pr}(n^\beta \bar{X}_n \to 0) \geq \mathrm{Pr}(\mathcal{E}) \geq 1$, hence the claim. □

Example 8.1 (Uniform almost sure convergence of kernel estimators)**.** *Consider $(X_1, Y_1), \ldots, (X_n, Y_n)$ a stationary sequence of observations with $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$. For all x, let $m(x) := E(Y_1 \mid X_1 = x)$ and consider the Nadaraya-Watson estimator $\widehat{m}_n(x) := \sum_{i=1}^n Y_i K\{(X_i - x)/h_n\}/\sum_{i=1}^n K\{(X_i - x)/h_n\}$, where $K : \mathbb{R}^d \to \mathbb{R}$ is a symmetric multivariate kernel and h_n is the bandwidth, which converges to zero. Hansen [2008] gives conditions for uniform almost sure convergence of $\widehat{m}_n - m$. In particular, for a compact set $\mathcal{C} \subset \mathbb{R}^d$, and under the conditions of [Theorem 9 in Hansen, 2008], it holds that $\sup_{x\in\mathcal{C}} |\widehat{m}_n(x) - m(x)| \leq O((\log n/n)^{2/(d+4)})$ almost surely.*

We now present two corollaries of Proposition 8.3 that provide sufficient conditions for $n^\beta X_n \to 0$ almost surely, and therefore for $n^\beta \bar{X}_n \to 0$ almost surely. Like Corollary 8.2, the first imposes a bound on the tail of $n^\beta X_n$.

Corollary 8.3. *Suppose that, for any $x > 0$, there exists $\alpha(x) > 0$ such that $\mathrm{Pr}(n^\beta X_n > x) = O(n^{-1-\alpha(x)})$. Then, $n^\beta \bar{X}_n \to 0$ almost surely.*

Proof. For $x > 0$ and $n \geq 1$, define the event $\mathcal{E}(n, x) := \{n^\beta X_n > x\}$. Because $\mathrm{Pr}\{\mathcal{E}(n, x)\} = O(n^{-1-\alpha(x)})$, there exists a constant $C < \infty$ such that $\sum_{n=1}^\infty \mathrm{Pr}\{\mathcal{E}(n, x)\} = C \sum_{n=1}^\infty n^{-1-\alpha(x)} < \infty$. Hence, by the Borel-Cantelli lemma, $\mathrm{Pr}\{\limsup_n \mathcal{E}(n, x)\} = 0$. As $x > 0$ was arbitrary, $\mathrm{Pr}\{\cap_{k=1}^\infty \limsup_n \mathcal{E}(n, 1/k)\} = 0$, which implies that $n^\beta X_n \to 0$ almost surely. The result follows by Proposition 8.3. □

The second corollary works in the setting where $(X_n)_{n\geq 1}$ is an adapted process. The corollary imposes a condition that is considerably weaker than the requirement that $n^\beta |X_n|$ almost surely converge to zero, but, in the case where $\beta > 0$, is stronger than the condition that $|X_n|$ is a supermartingale. In the case where $\beta = 0$, the imposed condition is equivalent to requiring that $|X_n|$ is a supermartingale.

Corollary 8.4. *Suppose that $(X_n)_{n\geq 1}$ is a sequence of random variables that is adapted to the filtration $(\mathcal{H}_n)_{n=1}^{\infty}$, that $n^{\beta}X_n \xrightarrow{p} 0$, and that*

$$(1+1/n)^{\beta}E(|X_{n+1}| \mid \mathcal{H}_n) \leq |X_n| \text{ for all } n \geq 1. \tag{8.6}$$

Under these conditions, $n^{\beta}\bar{X}_n \to 0$ almost surely.

Proof. Let $Y_n := n^{\beta}|X_n|$. Eq. 8.6 imposes that $(Y_n)_{n=1}^{\infty}$ is a supermartingale adapted to the filtration $(\mathcal{H}_n)_{n=1}^{\infty}$. Since $|X_n|$ is nonnegative, $E[Y_n^-] = 0 < \infty$. Hence, by Doob's martingale convergence theorem, Y_n converges almost surely to a random variable Y_{∞}. Moreover, since $Y_n \xrightarrow{p} 0$, it must be the case that $Y_{\infty} = 0$. Hence, $Y_n \to 0$ almost surely. Proposition 2 then gives the result. □

Since $1 + 1/n \leq \exp(1/n)$ for all n, the above corollary remains true if (8.6) is replaced by the condition that $\exp(\beta/n)E(|X_{n+1}| \mid \mathcal{H}_n) \leq |X_n|$ for all n.

8.6 Exponential deviation bounds

The following result shows that if X_n satisfies an exponential deviation bound, then $\bar{X}_n$ also satisfies such a bound.

Proposition 8.4. *Suppose that $(X_n)_{n\geq 1}$ is a sequence of random variables, for which there exists $C_0 \geq 0$, $C_1, C_2 > 0$, $\beta \in (0,1)$, and $\gamma \in (0, 1/\beta)$, such that, for any $x > 0$ and any $n \geq 1$,*

$$\operatorname{pr}\left(X_n \geq C_0 n^{-\beta} + x\right) \leq C_1 \exp(-C_2 n x^{\gamma}).$$

Let $\delta \in (\beta, \min(\gamma^{-1}, 1))$. Then, there exists a constant $C_4 > 0$ depending only on the constants of the problem ($C_0, C_1, C_2, \beta, \gamma$, and δ), such that, for any $y \geq 1$, it holds that

$$\operatorname{pr}\left(\bar{X}_n \geq \frac{C_0}{1-\beta}n^{-\beta} + \frac{3}{1-\delta}n^{-\delta}y\right) \leq C_4 n^{\alpha} \exp\left\{-C_2 n^{\alpha(1-\gamma\delta)} y^{\gamma}\right\},$$

with $\alpha := \gamma(1-\delta)/\{\gamma(1-\delta) + (1-\gamma\delta)\}$.

We defer the proof of the above result to the end of the current section.

Empirical risk mininizers are a common type of estimators for which the excess risk satisfies an exponential tail bound, as the following example shows. This example is a weakened version of Theorem 17 in Bartlett et al. [2006].

Example 8.2. *Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. copies of a a couple of random variables (X, Y) taking values in $\mathcal{X} \times \mathcal{Y}$. Consider a class of functions $\mathcal{F}$ defined as $\mathcal{F} := B\mathrm{absconv}(\mathcal{G})$, for some constant $B > 0$, and function class $\mathcal{G} \subseteq \{\pm 1\}^{\mathcal{X}}$, where* absconv *denotes the absolute convex hull (or symmetric convex hull). Let ℓ be a loss on $\mathcal{F}$, that is, a mapping defined on $\mathcal{F}$, such that for all $f \in \mathcal{F}$, $\ell(f)$ is a mapping $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. For any f, let $R(f) := E\{\ell(f)(X,Y)\}$. Let $\widehat{f}$ be an empirical risk minimizer over $\mathcal{F}$, that is, $f \in \arg\min_{f\in\mathcal{F}} \sum_{i=1}^{n} \ell(f)(X_i, Y_i)$. Let $f^* \in \arg\min_{f\in\mathcal{F}} R(f)$, a minimizer of the population risk over $\mathcal{F}$. Suppose that the following conditions are met.*

Condition 8.1. *There exists $L > 0$ such that, for any $x, y \in \mathcal{X} \times \mathcal{Y}$, and any $f_1, f_2 \in \mathcal{F}$, $|\ell(f_1)(x,y) - \ell(f_2)(x,y)| \leq |f_1(x) - f_2(x)|$.*

Condition 8.2. *There exists $c > 0$ such that, for any $f \in \mathcal{F}$, $E[\{\ell(f)(X,Y) - \ell(f^*)(X,Y)\}^2] \leq c\{R(f) - R(f^*)\}$.*

Condition 8.3. *It holds that $d_{VC}(\mathcal{F}) \leq d$, for some $d \geq 1$, where d_{VC} is the Vapnik-Chervonenkis dimension.*

Then, it holds that, for any $x > 0$,

$$\Pr\left\{R(\widehat{f}) - R(f^*) \geq C_0 n^{-(d+2)/(2d+2)} + x\right\} \leq \exp(-C_1 n x),$$

for some $C_0, C_1 > 0$ depending on the B, L, and c.

Remark 8.1. *Observe that the bound from Example 8.2 above is of the form* $\Pr(X_n \geq C_0 n^{-\beta} + x) \leq C_1 \exp(-C_2 n x^\gamma)$ *with* $\beta\gamma < 1$.

The proof of proposition 4 relies on the following lemma.

Lemma 8.1. *Suppose that $(X_n)_{n\geq 1}$ is a sequence of random variables, for which there exists $C_0 \geq 0$, $C_1, C_2 > 0$, $\beta \in (0,1)$, and $\gamma \in (0, 1/\beta)$, such that, for any $x > 0$ and any $n \geq 1$,*

$$\mathrm{pr}\left(X_n \geq C_0 n^{-\beta} + x\right) \leq C_1 \exp(-C_2 n x^\gamma).$$

Consider $\delta \in (0, \min(\gamma^{-1}, 1))$. There exists a constant C_1' that depends only on the constants of the problem ($C_0, C_1, C_2, \beta, \gamma, \delta$) such that, for any integer $m \geq 1$, and any real number $y \geq 1$,

$$\Pr\left(\exists k \geq m+1 : X_k \geq C_0 k^{-\beta} + k^{-\delta} y\right) \leq C_1' m \exp\left(-C_2 m^{1-\gamma\delta} y^\gamma\right).$$

Proof. Let $y \geq 1$. We have that

$$\begin{aligned}\Pr\left(\exists k \geq m+1 : X_k \geq C_0 k^{-\beta} + k^{-\delta} y\right) \leq & C_1 \sum_{k \geq m+1} \exp\left(-C_2 k^{1-\gamma\delta} y^\gamma\right) \\ \leq & C_1 \int_m^\infty \exp\left(-C_2 k^{1-\gamma\delta} y^\gamma\right) dk\end{aligned}$$

Making the change of variable $u = k^{1-\gamma\delta} y^\delta$, we obtain

$$\begin{aligned}\Pr\left(\exists k \geq m+1 : X_k \geq C_0 k^{-\beta} + k^{-\delta} y\right) \leq & C_1 y^{-\gamma/(1-\gamma\delta)} \int_{m^{1-\gamma\delta} y^\gamma}^\infty \exp(-C_2 u) u^{1/(1-\gamma\delta)-1} du \\ \leq & C_1 y^{-\gamma\kappa} \int_{m^{1-\gamma\delta} y^\gamma}^\infty \exp(-C_2 u) u^{\lceil \kappa - 1 \rceil} du,\end{aligned}$$

where we denote $\kappa := 1/(1-\gamma\delta)$. Observe that $\kappa > 1$.

We now prove a general identity for the type of integral that appears in the last line of the above display. Denote, for any integer $q \geq 1$, and real numbers $a \geq 1$, and $c > 0$, $I_q(a,c) := \int_a^\infty \exp(-cu)u^q du$. By integration by parts, we have that

$$I_q(a,c) = \exp(-ca)\frac{a^q}{c} + \frac{q}{c}I_{q-1}(a,c).$$

Reasoning by induction, we obtain that

$$\begin{aligned} I_q(a,c) =& \exp(-ca)\left(\frac{a^q}{c} + \frac{qa^{q-1}}{c^2} + \dots \frac{q!}{c^{q+1}}\right) \\ \leq& q \times q! \max\{c^{-1}, c^{-(q+1)}\} a^q \exp(-ca), \end{aligned}$$

where we have used in the last line that $a \geq 1$.

Therefore, denoting $C_3 := C_3(C_2, \kappa) := \lceil \kappa - 1 \rceil \times \lceil \kappa - 1 \rceil ! \max\{c^{-1}, c^{-(q+1)}\}$, we have that

$$\begin{aligned} \Pr\left(\exists k \geq m+1 : X_k \geq C_0 k^{-\beta} + k^{-\delta} y\right) \leq& C_1 C_3 y^{-\gamma\kappa} m^{(1-\gamma\kappa)\lceil\kappa-1\rceil} y^{\gamma\lceil\kappa-1\rceil} \exp\left(-C_2 m^{1-\gamma\delta} y^\gamma\right) \\ \leq& C_1 C_3 m \exp\left(-C_2 m^{1-\gamma\delta} y^\delta\right), \end{aligned}$$

where we have used in the last line that $m \geq 1$ and $y \geq 1$. □

We now prove proposition 8.4.

of proposition 8.4. Let $y_1 \geq 1$ and $y \geq 1$, and let $1 \leq m < n$. From lemma 8.1, we have that, with probability at least $1 - C_1 C_3 \exp(-C_2 y_1^\gamma)$,

$$\frac{1}{n}\sum_{k=2}^{m} X_k - C_0 k^{-\beta} \leq \frac{1}{1-\delta}\frac{m^{1-\delta}}{n} y_1, \tag{8.7}$$

and, with probability at least $1 - C_1 C_3 \exp(-C_2 m^{1-\gamma\delta} y^\gamma)$,

$$\frac{1}{n}\sum_{k=m+1}^{n} X_k - C_0 k^{-\beta} \leq \frac{1}{1-\delta} n^{-\delta} y. \tag{8.8}$$

□

Set $y_1 = m^{(1-\gamma\delta)/\gamma} y$ and $m = n^\alpha$, with $\alpha := \gamma(1-\delta)/\{\gamma(1-\delta) + (1-\gamma\delta)\}$. Observe that these choices are consistent with the conditions $y_1 \geq 1$ and $1 \leq m \leq n$. We then have that $y_1^\gamma = m^{1-\gamma\delta} y^\gamma$ and $m^{1-\delta}/n y_1 = n^{-\delta} y$, which renders equal the right-hand sides in (8.7) and (8.8) and the corresponding exponential probability bounds. From a union bound, we then have that, with probability at least $1 - C_1 C_2 (n^\alpha + 1) \exp\{-C_2 n^{\alpha(1-\gamma\delta)} y^\gamma\}$,

$$\frac{1}{n}\sum_{k=2}^{n} X_k - C_0 k^{-\beta} \leq \frac{2}{1-\delta} n^{-\delta} y.$$

We now turn to the first term of $\bar{X}_n$. We have that

$$P\left(\frac{1}{n}(X_1 - C_0) \geq \frac{1}{1-\delta} n^{-\delta} y\right) \leq C_1 \exp\left(-C_2(1-\delta)^{-\gamma} n^{1+\gamma(1-\delta)} y^\gamma\right).$$

Observe that $1 + \gamma(1-\delta) > \alpha(1-\gamma\delta)$. Therefore, there exists C_1' that depends only on the constants of the problem $(C_0, C_1, C_2, \beta, \gamma, \delta)$, such that, for any $y \geq 1$,

$$P\left(\frac{1}{n}(X_1 - C_0) \geq \frac{1}{1-\delta} n^{-\delta} y\right) \leq C_1' \exp(-C_2 n^{\alpha(1-\gamma\delta)} y^\gamma).$$

Therefore, gathering the previous bounds via a union bound yields that there exists a constant C_4 that depends only on the constants of the problem such that, for any $y \geq 1$, with probability at least $1 - C_4 n^\alpha \exp\{-C_2 n^{\alpha(1-\gamma\delta)} y^\gamma\}$, $\bar{X}_n \leq C_0/(1-\beta)n^{-\beta} + 3/(1-\delta)n^{-\delta} y$.

8.7 Conclusion

In this chapter we studied conditions under which the Cesaro means of random variables converge stochastically. We have shown in particular that almost sure convergence, L_1 convergence, and high probability convergence yield the same rate of convergence for the Cesaro means. We also gave a counterexample of a situation where convergence in probability at a certain rate does not imply convergence in probability for the same rate of the Cesaro means.

As we pointed out in the introduction, we were motivated to work on the question of stochastic convergence of Cesaro means are these arise in particular in the study of semiparametric estimators built from a sequence of nuisance estimators. Fortunately, there exists high probability, almost sure, or in L_1 norm convergence guarantees for a wide range of nonparametric estimators we might want to use for nuisance estimation, as we discussed in our examples.

Acknowledgements

The authors are grateful to Iosif Pinelis, who answered a question by AFB on MathOverflow [Pinelis, 2019] that partly inspired our counterexample in Section 8.3. AL was supported by the National Institutes of Health (NIH) under award number DP2- LM013340. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Bibliography

J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35(2): 608–633, 04 2007.

P. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Ann. Statist.*, 33 (4):1497–1537, 08 2005.

A-L. Cauchy. *Cours d'Analyse de l'École royale polytechnique*, pages 48–52. 1821.

E. Cesàro. Sur la convergence des séries. *Nouvelles annales de mathématiques*, 7:49–59, 1888.

B. E. Hansen. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24(3):726–748, 2008.

A. R. Luedtke and M. J. van der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.*, 44(2):713–742, 04 2016.

I. Pinelis. Convergence in probability of cesaro means. MathOverflow, 2019. URL `https://mathoverflow.net/q/347079`. URL:https://mathoverflow.net/q/347079 (version: 2019-11-27).

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.