Title
Incorporating Intent, Impact, and Context for Beneficial Machine Learning

Author
Rolf, Esther

Publication Date
2022

Peer reviewed|Thesis/dissertation

Incorporating Intent, Impact, and Context for Beneficial Machine Learning

by

Esther Grace Rolf

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Benjamin Recht, Co-chair
Professor Michael I. Jordan, Co-chair
Professor Solomon Hsiang

Spring 2022

Incorporating Intent, Impact, and Context for Beneficial Machine Learning

Abstract

Incorporating Intent, Impact, and Context for Beneficial Machine Learning

by

Esther Grace Rolf

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Benjamin Recht, Co-chair

Professor Michael I. Jordan, Co-chair

Advancements in machine learning hold unprecedented potential to help humans understand and shape our world, from deriving data-driven policies with global prediction systems to combating long-standing prejudice in social decision-making. However, in practice, machine learning is not achieving its potential. Algorithmic systems that are theoretically well motivated fail to live up to anticipated performance in the real world, and all too often, they exacerbate inequality rather than relieve it. In order to translate potential into real benefit, machine learning systems need to address the *context* of the domain they're applied in to better interface with system *intent* and system *impact*.

In this thesis, we present research on the context-aware design of machine learning systems that interface with individuals, our environment, and our societies. We emphasize the core themes of intent, impact, and context through three interweaving threads: (i) contrasting algorithmic impact with desired intent, (ii) contextualizing learning algorithms through structures in input data, and (iii) advancing machine learning with remotely sensed data as precise applications of intent- and context-aware design in practice. We conclude with overarching lessons learned that carry into constructive opportunities for future research.

To my mother Trude and my sister Helen,
the women who inspire me to find the best in myself
and ceaselessly support me in the path of pursuing it.

Contents

Acknowledgments

Before presenting the research that constitutes this thesis, I would like to acknowledge all of the people who had a hand in helping shape this work and my doctoral studies.

I have an inexpressible amount of gratitude for my PhD advisors, Ben Recht and Mike Jordan, who welcomed me into their research groups many years ago, and have supported me ever since. In addition to being stellar researchers and teachers, Ben and Mike have been phenomenal advisors. Mike has always encouraged me to think about a bigger picture, introducing me to new fields of research and thought to round out my perspectives on a problem. At the same time, Ben has always encouraged me to find problems that matter to me, questions I can't get out of my head. Together, their guidance led me to develop the varied yet cohesive research agenda outlined in this thesis. I hope that alongside the research and technical skills I grew under their guidance, I also take with me a small a flavor of their abilities for mentoring and teaching.

I also wish to thank Sol Hsiang, who in addition to serving on my dissertation committee has been a dedicated research mentor to me since the beginning of my PhD. As an interdisciplinary researcher himself, Sol has inspired me to push the boundaries of how I conceive of my research interfacing with the broader world and has encouraged me to tackle head-first the most pressing problems I can pose. A special thanks also to my informal mentor Moritz Hardt, who chaired my qualifying exam committee. Moritz has greatly influenced my perspectives on the social contexts that surround computing technologies, with a critical eye toward both how key phenomena can be codified and explained through mathematical frameworks, and when to look beyond mathematical frameworks from the start.

I have been blessed to have many wonderful academic mentors during my dissertation work. One of these mentors was Joshua Blumenstock, who has helped me find a balance between meaningful practical problems and overarching academic insights in my research. Another was Tamma Carleton, who on top of being an incredibly supportive friend and collaborator, has helped me navigate where I want to make a difference with my research, and how to get there. Thank you also to Nebojsa Jojic for hosting me during an internship at Microsoft Research, and to Fernando Diaz, Ben Packer, and Alex Beutel for hosting me during an internship at Google Research. Remote internships posed an interesting challenge to which these mentors rose, innovating on ways to help communicate, share, and build upon each other's research ideas and interests.

Thank you to my many wonderful collaborators who fueled my passion to take on and chip away at hard research problems – to the collaborators who had a hand in the research presented in this thesis (in addition to those listed above): Lydia Liu, Sarah Dean, Max Simchowitz, Dan Björkegren, Claire Tomlin, David Fridovich-Keil, Teddi Worledge, Jon Proctor, Ian Bolliger, Miyabi Ishihara, and Vaishaal Shankar – and to the other collaborators with whom I've worked in the course of my doctoral studies, notably Emily Aiken, Kolya Malkin, Caleb Robinson, and Alexandros Graikos. These were the people who shared in the ideation stages, the grind and re-grind of realizing those ideas, and late nights of paper writing, memories I look back on fondly today.

I would also like to acknowledge the many staff and administrators of the Computer Science Department and my lab groups who made my research possible. In particular thank you to Shirley Salanio, Audrey Sillers, Jean Nguyen, Tiffany Reardon, Pat Hernan, Kattt Atchley, Naomi Yamasaki, Ria Briggs, Taylor Kee, Jon Kuroda, and Kostadin Ilov.

I have been lucky to have phenomenal lab mates in the research groups to which I belong. My research and life perspectives have benefited greatly from all the students and postdocs of SAIL, Modest Yachts, and the Global Policy Lab who have overlapped with my time at UC Berkeley. I am thankful to Orianna DeMasi, Ashia Wilson, Becca Roelofs, Max Simchowitz, Vaishaal Shankar, Stephen Tu, Chelsea Zhang, Ludwig Schmidt, Eric Jonas, Horia Mania, Sarah Dean, Lydia Liu, Ross Boczar, Ahmed El Alaoui, Sara Fridovich-Keil, Mihaela Curmei, Karl Krauth, Romain Lopez, Tijana Zrnic, Wenshuo Guo, Paula Gradu, Deb Raji, Juanky Perdomo, John Miller, Eric Mazumdar, and Nilesh Tripuraneni for being amazing blend of friends, office-mates, colleagues, support system, and sounding boards. I am especially thankful to Shivaram Venkataraman, my desk-mate my first year of my PhD. I think Shivaram could always tell when I was having a rough day; I always knew he was always there ready to brew a pot of tea, which always made it better. And of course, I am grateful for Vaishaal Shankar and Ludwig Schmidt for always being game to run up the fire trails after work, often waiting for me to catch up at the bench before heading back down.

I was incredibly fortunate at Berkeley to find my friends Carolyn Matl, Anusha Nagabandi, Andrea Bajcsy, Hani Gomez, and Alyssa Morrow who were always there to turn bad days into good, hunger into delicious meals, and almost-tears into belly laughs. I've also had incredible housemates, including Caroline Lemieux, Rachel Chen, and of course Melvin Walls, whose gourmet cooking fed me through most of the pandemic. I've shared unforgettable travels inside and far outside of California with Kelly Fernandez, Lucy Stephenson, Robert Nishihara, Jeff Mahler, Alyssa Morrow, Ahmed El Alaoui, Arya Reais-Parsi, and Ryan Kaveh. To Ryan especially I am thankful for sharing so many laughs, adventures, and day-to-day confidences, and for being a constant source of joy and support during the latter years of my PhD.

Long before even dreaming of starting a PhD, I had an incredible amount support from my family – my aunts, uncles, grandparents and cousins who every winter would want to know how school was going, and encourage my passions for whatever subject I loved most at the time. Sometimes what I've needed most was that unconditional support, something embodied in the love of Katie Rourke, who has known me for as long as I can remember, and has helped foster many of the qualities I appreciate most about myself today. My sister Helen has always been a model of hard work and perseverance to look up to, and a soft and understanding voice to fall back on. And of course, my mother Trude has been the most patient, kind, and understanding parent I could imagine. I will forever be grateful for the lengths she went through to give me opportunities in life, and the encouragement and support to see me through them.

In summary, thank you all.
– Esther

Chapter 1

Introduction

Recent advancements in how machine learning models are designed and trained have spurred immense interest in using these models for the benefit of society. Increasingly large-scale datasets contain key high-dimensional patterns that sophisticated predictive algorithms can uncover. As a consequence, machine learning holds unprecedented potential to help people understand and shape our world, from deriving data-driven policies with global prediction systems to combating long-standing prejudice in social decision-making. However, in practice, machine learning is not achieving its potential. Algorithmic systems that are theoretically well motivated fail to live up to anticipated performance in the real world, and all too often, they exacerbate inequality rather than relieve it. This thesis aims in the direction of achieving this illusive potential for benefit with data-driven algorithms by addressing key concepts of intent, impact, and context, and presenting methods by which to integrate them in statistical machine learning frameworks and models.

The aspiration of global benefit is founded: machine learning can translate vast amounts of data into inferences that help people make more informed decisions. As an example, the United Nations released discussion of how data science and analytics can contribute to their 17 sustainable development goals (SDGs) [34], ranging from eliminating poverty and hunger to taking climate action. Examples already in evidence include generating predicted measures of poverty with satellite imagery and computer vision [145], extracting information from call data records and machine learning to target humanitarian assistance to those who need it most [6], and monitoring ecological change like deforestation with satellite imagery and computer vision [46, 85, 215]. The potential of achieving societal benefit with machine learning extends to applications in the domains of healthcare [235], climate change [248], and wildlife conservation [277].

At the same time, there is increasing evidence of deployed machine learning systems exacerbating existing social problems, and potentially generating new ones. For example, learned classification systems have been shown to amplify gender discrimination in resume reading tools [83]. Data-driven risk assessment tools used in bail decisions have resulted in higher likelihood of falsely predicting black defendants would re-offend, compared to white defendants [8]. In a potentially more subtle failure mode, systems that work well in one

context or simulated environment can fail to generalize to new application settings. Audits have exposed commercial facial recognition software exhibiting accuracy disparities across intersectional subgroupings of race and gender [49], presumably in part due to training data that was unrepresentative of the full diversity of human faces. Despite best efforts, a majority of machine learning models put forth in academic literature to diagnose or prognosticate COVID-19 cases have been assessed to have little, if any, clinical value [240, 295].

In each of these examples, some part of the machine learning prediction system (Figure 1.1) – from dataset design and collection, to model architecture and training, to performance objective definition, evaluation, and monitoring – missed the mark. One way to explain this is that these systems failed to address the full *context* of the domain that they were deployed in, resulting in a discrepancy between a desired *intent* (presumably, benefit) and their realized *impact*. While context is clearly integral to producing functioning (efficient and fair) learning systems, integrating context in general machine learning frameworks remains difficult, precisely because context entails something different in each application domain (Section 1.1). Much work remains to make existing machine learning algorithms and frameworks amenable to modeling diverse notions of context. In this thesis, we emphasize data as a mathematical formalizer by which to clarify and unify domain-specific considerations, in the service of reaching precisely delineated, but possibly multi-faceted learning objectives. Alongside this study of context in statistical learning frameworks, we anchor these findings with key applications combining remote sensing and machine learning.

Aligning progress in machine learning with the potential for real-world benefit will require deep collaboration across disciplinary boundaries. The research featured in the chapters of this thesis integrates theory, practice, and human values in collaboration with development economists (Chapter 3), roboticists (Chapter 7) and policy researchers (Chapter 8). These chapters thread a balance between application-driven instances of domain context and general notions of intent, impact, and context that apply across learning paradigms. Parts I and II situate recurring themes from these interdisciplinary viewpoints within more overarching statistical machine learning frameworks. Looking forward, the research presented in this thesis is inspired by, and provides evidence to the idea that multi-disciplinary and context-centric perspectives can drive novel machine learning insights, and are essential factor for designing and implementing algorithms with positive social benefit.

1.1 Intent, Impact, and Context

As discussed above, the line of research presented in thesis integrates the concepts of intent, impact, and context with statistical machine learning frameworks (Figure 1.1), toward understanding failure points and designing explicitly for specific intents.

In considering the consequences of data-driven algorithms, it is natural to think about the **impact** of machine learning systems, namely the ways in which these algorithms interact with society, its individuals, and our environment. This realized impact is distinct from the desired **intent** of the system. While intent can and should be defined during the conceptu-

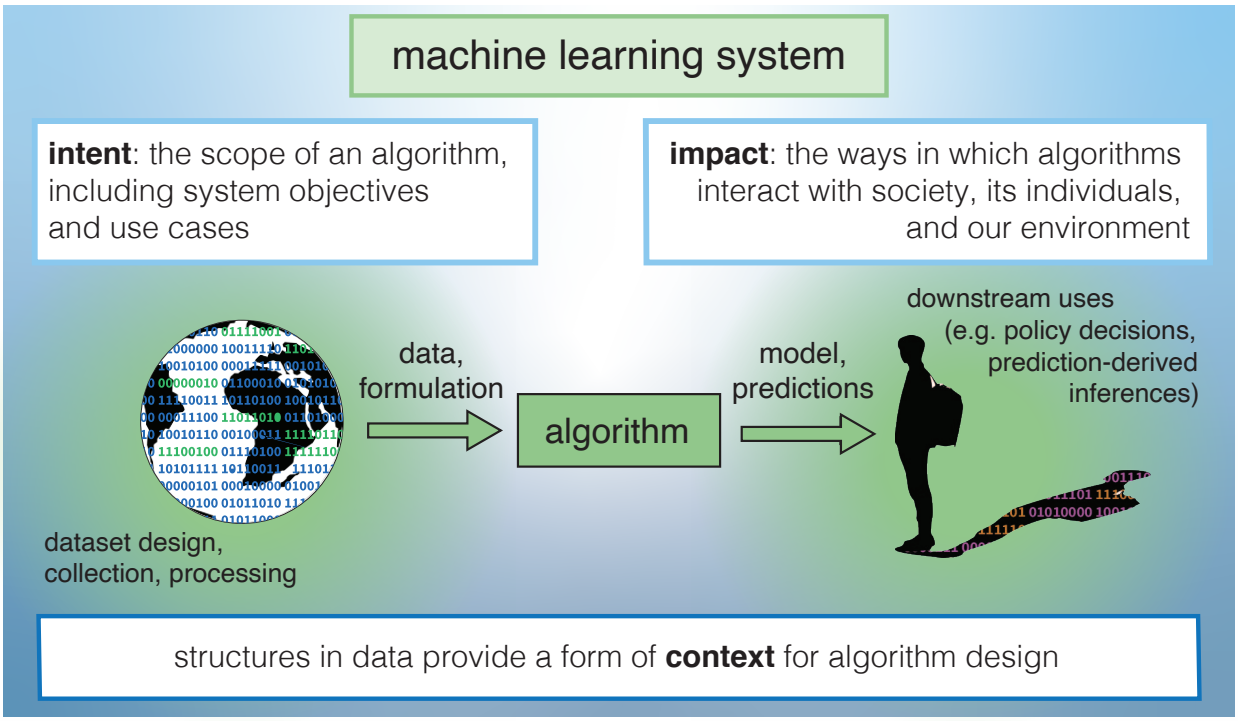


Figure 1.1: Depiction of the algorithmic system. The machine learning *algorithm* is surrounded within the larger machine learning *system* – the frameworks we develop and study in this thesis will generally incorporate algorithmic innovations within specific constraints of a larger system, bookended by design, collection, and processing of the input data and eventual downstream use cases of the model or predictions output by the algorithm. The themes of intent and impact are part of a larger context surrounding the algorithmic system. In this thesis, we consider structures in input data as one way to formalize key aspects of this broader context.

alization and design of a learning system, impact is measured or audited once a system is in use. Intent can be delineated by specifying the scope of an algorithmic system, including for example system objectives, constraints, requirements, and valid use cases.

Harm can arise from algorithmic systems when intent and impact of machine learning systems are not aligned.[1] While the intent of learning models designed for COVID-19 diagnosis and prognosis was to aid doctors toward better clinical decision-making, the overall impact of the models was minimal if not detrimental, as as some over-estimated their effectiveness for clinical decision-making [240, 295]. In a different example, concern has been raised that complex machine learning models designed to help combat climate change actually require a substantial amount of energy to train and deploy [78]. To aid intentional design and use

[1]Harm can also arise from algorithmic systems when the intent of those systems is not clearly specified, as we discuss further in Chapter 2.

of machine learning models, Mitchell et al. [201] develop model cards, a framework that systematizes transparent documentation of key aspects of what this thesis terms intent and impact, including benchmarking performance across salient subgroups and providing context of intended use.

Aligning intent and impact requires understanding the broader scientific and societal **context** in which algorithms are situated. Blumenstock [37] warns that in applying machine learning to development domains, these "next-generation solutions must be designed and produced by people who understand the problems and context — not just by those who understand the algorithms." This sentiment is reinforced by Bondi et al. [40] and Kshirsagar et al. [170] in discussions of the importance of participation, inclusion, and communication across diverse stakeholders and contributors in "machine learning for social good" endeavors. Chen et al. [62] discuss unique concerns and opportunities in utilizing machine learning methodologies in medical contexts. The importance of context extends far beyond algorithm design; Abebe et al. [2] discuss the importance of understanding social and political context for designing and enforcing data sharing practices.

As illustrated by the examples and references above, the ways intent, impact and context manifest with machine learning systems is inherently domain-dependent. Recognizing this, we do not aim here to define the terms intent, impact and context in their relationship to machine learning in any absolute sense. Rather, we instantiate these themes through targeted mathematical and statistical characterizations, and present methods by which to integrate these characterizations in both general machine learning frameworks and applied to specific problem settings.

Intent, impact, and context enter machine learning frameworks at many levels of abstraction. Figure 1.1 differentiates between a machine learning *algorithm* and a machine learning *system*, a broader unit of inquiry including all procedural steps of instantiating that model. Here we consider a machine learning system to include steps of dataset design, collection, and processing, as well as steps taken after predictions are made, including downstream decisions, tasks, or inferences based on algorithmic outputs.

To illustrate this distinction, consider a machine learning system used to target humanitarian aid based on predicted poverty measures from satellite imagery. Key aspects of this broader algorithmic system would include questions of dataset formulation: how poverty is measured, at what resolution and in what regions labeled data is acquired for training and evaluation, and which sources of satellite imagery are matched with these labels. The learning system also includes the procedures by which predictions of poverty are used in the downstream allocation of aid: whether aid is allocated to individuals or groups, how the amount of aid given to each individual unit is determined as a result of predicted poverty measures, etc.

An algorithm, on the other hand, is generally scoped to either abstract away these broader questions or to incorporate them as fixed constraints on the structure of the problem at hand. In the aid targeting example, the algorithmic component comprises the definition and training of a machine learning model that outputs predicted poverty measures from satellite imagery outputs. Depending on the specification of the larger learning system, the

algorithm might be designed to output predictions at pixel-resolution or image-resolution; the training objective might prioritize average accuracy over the entire training region, or it might optimize the worst-case accuracy of any given region.

Such abstractions are necessary to formulating a precise problem statement amenable to standard machine learning techniques; for example delineating an objective function which can be optimized with respect to parameters of a predictive model. At the same time, focusing only on solving the abstracted problem neglects the broader context of the learning system [256]. Thus, it is important to develop techniques and analysis that center *between* these levels. Throughout this thesis we both center algorithm-level innovations within larger system constraints and characterize system-level frameworks in terms of key properties of the underlying learning algorithms. These perspectives complement each other to shed light on how context intersects across algorithm and system levels of abstraction, described in Section 1.3 and synthesized in the conclusion of this document in Part IV.

1.2 Remote Sensing and Machine Learning

Applications combining machine learning with remote sensing will anchor the somewhat abstract themes of intent and context discussed in the previous section. We focus on applications to remote sensing for several reasons. First, there is a clear potential for benefit, ranging from monitoring and addressing climate change, to informing and evaluating data-driven policies across the globe. Second, the scale and high-dimensional nature in remotely sensed data are particularly amenable to machine learning analysis, whereas they could quickly become a bottleneck for manual analysis. At the same time, structures in data acquisition patterns and sensor measurements provide an opportunity to tailor algorithmic techniques, evidencing the value of incorporating domain context through data structures and specific intents; in this sense these application areas can anchor the themes of intent, impact, and context in specific examples of machine learning systems.

Remotely sensed data, for example from satellite imagery or robots enabled with mobile sensors, can shed light on remote or difficult to measure regions of the world where sending humans to collect data could be prohibitively costly or dangerous. Combining machine learning and remote sensing can thus help researchers monitor environmental and social change and policymakers develop more comprehensive data-driven decisions. Examples include robotic localization during safety-critical search and rescue missions [133], weather forecasting with sensor data [140], and ample opportunities in combining satellite imagery and machine learning to address the United Nation's Sustainable Development Goals [51].

However, remotely sensed data presents challenges for traditional learning paradigms, such as immense scale of image data [186, 245] and sparsity, quality or resolution issues of labeled data [38, 188]. With these challenges come opportunities to leverage structures specific to remotely sensed data. Delineating precise knowledge of sensor structures enables adaptation of established statistical learning frameworks to new real-world settings, as exemplified by our approach to the problem of radioactive source localization in Chapter 7. In a

second example, invariances in scale, rotation, and translation in satellite imagery prediction tasks can guide algorithmic simplifications, as discussed in Chapter 8.

It is important to address unique issues of fairness, responsibility, and ethics of the emerging field intersecting machine learning, remote sensing, and policy. Addressing unique issues of social context including potential harms that could arise from combining remote sensing and machine learning will be critical to ensuring that progress in emerging field aligns with intendended benefits; we return to this point in Section 9.2.

1.3 Document Overview

The body of this thesis is organized in three parts: (I) contrasting impact and intent of algorithmic classification systems, (II) utilizing structures in training data as mathematical form of context in statistical machine learning frameworks, and (III) anchoring these lessons to applications combining remote sensing and machine learning. Each part considers a different depth of generality in the design and use of learning systems, starting from in some sense the most general.

In Part I, we consider algorithmic decision rules at large – classification decisions that are often based on learned models, but need not be. In Chapter 2, we discuss several examples where deployed machine learning systems exhibit social harm or bias. Often these harms are unintentional; in Chapter 2 we go on to define and discuss notions of fairness in machine learning that aim to mitigate potential harms resulting from algorithmic systems. In Chapter 3, we present an alternative framework for delineating welfare goals explicitly via multi-objective optimization, drawing connections with the approaches described in Chapter 2. Contrasting the intents and impacts of algorithmic decisions leads us to study how to achieve diverse objectives when decisions are in fact based on learned models.

In Part II, we ground our study of context in data-driven learning by focusing on two key elements of structure in training data: *how* data instances are collected across different subpopulations and data sources and *what* information different features of a dataset convey. In Chapter 4, we incorporate dataset collection as a design step in the learning procedure. This enables us to assess and leverage the importance of different data allocations toward reaching high group and population accuracies. In Chapter 5, we consider structures in data from another angle: utilizing the full information associated with each data instance. We show that "side information" like time or location metadata generally encode different structures than traditional "feature" representations, yet can still amplify predictive signal when utilized correctly. We propose a post-processing procedure as a natural and effective way to utilize this structure, and show it is robust to different tasks, predictors, and sampling patterns.

Together, Chapters 4 and 5 build an understanding of fundamental dataset characteristics and qualities that apply across machine learning paradigms. We underscore that data is a useful, but notably incomplete notion of context in Chapter 6, where we highlight key

limitations of viewing data as context and expose opportunities for expanding this line of study.

Part III anchors the ideas from Parts I and II with two applications combining remote sensing and machine learning. In Chapter 7, we develop statistically-efficient trajectory planning methods for sensor-equipped robots to localize sources of environmental radiation as quickly as possible. In Chapter 8, we detail a large scale project that combines machine learning and satellite imagery, where structures in the imagery data and downstream use cases guide algorithmic innovations that increase accessibility and computational efficiency of this technology. While each chapter in Part III incorporates both the themes of delineating intent during design and viewing data as a form of context, the structure within Part III largely echoes that of the first two parts. In line with the themes of Part I, our approach in Chapter 7 frames the precise intent of the source identification task in the decision-making problem. This allows us to leverage a multi-armed bandit inspired sampling approach, which we modify to apply to problem settings with physical sensing constraints. In Chapter 8, we tailor computer vision techniques to the unique structures of satellite imagery as a data source, in line with the overall themes of Part II.

Chapter 9 in Part IV concludes this thesis with connections, conclusions, and overarching perspectives. Organized into three distinct parts based on their focuses, Chapters 2 through 8 all involve the themes of intent and impact, data as context, and application-driven design to some degree. As these themes interweave, so to do threads of interdisciplinary perspectives which both inspire novel learning frameworks and targeted algorithmic innovations. The result is a broad set of results surrounding incorporating intent, impact, and context in statistical machine learning, with evidenced potential for societal benefit.

Part I

Contrasting Intent & Impact of Algorithmic Selection Rules

And because the upsides are so obvious, it's particularly important to step back and ask ourselves, what are the possible downsides? ... How do we get the benefits of this while mitigating the risk?

— Emily M. Bender, quoted in Hao, "We read the paper that forced Timnit Gebru out of Google. Here's what it says," 2020 [124]

In Part I of this thesis, we contrast the social impacts of machine learning systems with their intended benefit. In Chapter 2, we discuss several examples where deployed machine learning systems exhibit social harm or bias. Often these harms are unintentional; in Chapter 2 we go on to define and discuss notions of fairness in machine learning that aim to mitigate potential harms resulting from algorithmic systems. In Chapter 3, we present an alternative approach of explicitly delineating and designing for social welfare as a desired intent. This alternative framing encompasses several existing fairness notions (Section 3.5) and handles more general settings applicable when individual notions of welfare can be measured, approximated, or predicted.

The themes of intent and impact of algorithmic systems that are the focus of Part I interweave throughout the latter parts of this thesis. We further discuss connections and implications of these themes to Parts II and III in Part IV.

Chapter 2

Background: Un-Intended Impacts of Learned Algorithmic Policies.

As discussed in Chapter 1, harm can arise from algorithmic systems when the intent of a system is not clearly specified, or when it does not incorporate the full relevant context in which the algorithm will be used.[1] This chapter provides a short introduction to how and when unintended harms can arise in classification systems.

In Section 2.1, we organize several examples of unintended bias and harm resulting from use of machine learning algorithms. The variety of examples highlight different manifestations of un-intended impacts across different contexts. There is also a line of work developing more general frames of thought surrounding how and when issues of fairness, bias, or harm arise in algorithmic systems [25, 214, 269]. Suresh and Guttag [269] discuss the ways in which bias can enter specific processes within the machine learning system life cycle. Barocas et al. [25] and Crawford [80] distinguish between allocative and representational harms. Allocative harms include unfair dispersion of resources or benefits. Representational harms, on the other hand, surround the establishment, codification, or reinforcement of negative associations with identities and group categories.

In Section 2.2, we present two key notions of group or demographic fairness that address largely allocative harms. In Section 2.3, we discuss limitations of demographic fairness constraints as a mitigation strategy, setting the stage for Chapter 3. We briefly discuss alternative approaches, noting that achieving fairness or benefit is never a "one size fits all" solution – strategies to mitigate harms must also incorporate context. Selbst et al. [256], for example, argue that overemphasis on abstraction and modular design – foundations of modern computing and machine learning – can create a mismatch between the intent of "fair" algorithms and interventions and their actual suitability to real contexts. The authors suggest avenues by which to broaden boundaries of abstraction to include relevant social context.

[1] Harm can also arise when system intent is specified *and* realized, but that specified intent causes harm. While this of real concern, such systems fall outside the scope of this chapter.

2.1 Examples of Bias in Machine Learning Systems

Examining critical examples of bias in learning systems sheds light on the complex interplay between the different sources and forms of harm that one might hope to characterize. For example, take the distinction between allocative and representative harms discussed above. When Sweeney [270] identified that online search engine ads suggested arrest records at higher rates for queries of black identifying first names than white identifying first names, she exposed a representational harm: something in the depths of the algorithmic system caused substantially more negative content to be delivered when the queries were names associated with certain race groups. This representational harm results in the witnessed allocative harm. For example, when a potential employer searches for a candidates name and the search engine returns suggestions of prior arrest, this can decrease the chance of that candidate receiving the job opportunity. Another study found that women were less likely to be shown ads for high paying jobs than men [84], yet another example of allocative harm.

Biased representations and outcomes are key concerns across diverse fields of machine learning, from computer vision to natural language processing. Gender bias has been addressed in models across computer vision tasks from image captioning to classification [131, 290]. Audits of facial recognition models have exposed significant accuracy and efficacy gaps across intersectional and non-binary demographic identities [49, 253]. In the field of natural language processing, numerous works have exposed bias in word embeddings across demographic and social groups [39, 53, 271, 302]. While much scholarship aims to debias word embeddings, there is concern these proposed corrections may mask, rather than remedy, the deep-seated mechanisms by which such biases arise [110]. Bender et al. [32] warn that these concerns are likely to be amplified as language models become larger, increasingly relying on big data at the expense of carefully curated and documented data.

It can be tempting to blame such harms on the data on which algorithmic systems are trained. After all, any collected data encodes a context of its time and measurement, and is susceptible to containing historical biases or personal prejudices as a result. For example, algorithmically generated credit scores estimate "creditworthiness" as a function of an individuals previous repayment of loans, which are inseparable from the historical systems through which loan terms and decisions were made [27]. Obermeyer et al. [212] show that using available medical expenditure data as a proxy for medical need results in underestimation of medical need for black patients, who historically have lower medical costs on average than white patients due to a variety of factors. Similar concerns over uses of observational data have been raised in recidivism prediction [8] and child welfare analytics [67]. These domains exhibit the added challenge of what the machine learning community has termed a *selective labels* problem [10, 163, 171]: the subset of all possible observations that makes up the observational dataset is a function of prior (usually human) judgement. Lum and Isaac [185] discuss the dangers of hidden feedback loops that can occur when predictive policing algorithms over-predict crime in historically over-surveilled areas.

While quality and quantity of training data is integral to building effective machine learning systems (as is the focus of Chapter 4), they are by no means sufficient to achieving

benefit, or even avoiding harm. As Wang et al. [290] show, representational biases can persist despite best efforts to balance datasets across groups. Biases can arise throughout the learning process [269], and it is imperative to understand how the large-scale use of algorithmic predictions in consequential settings can serve to exacerbate social biases or reinforce existing structures of power [33, 211]. The field of fairness in machine learning is one that aims to identify, characterize, and mitigate potential harms of learning systems.

2.2 Group Fairness

At a high level, group fairness techniques address scenarios where unmitigated classification decisions would result in unequal in treatment or impact across groups, or sub-populations of a larger target population. The mitigation strategies are often encoded as a constraint on the classification or selection rule, enforcing equality of a carefully defined statistic, or *demographic fairness criteria*[2], across groups. While the landscape of group fairness criteria and methodologies to satisfy them are broad [26], we focus in this section on two commonly studied fairness criteria. This will serve to give a flavor of demographic fairness criteria and to establish relevant background and definitions for Section 2.3 and Chapter 3.

The group fairness settings we discuss here assume that the population is made up of disjoint groups $\mathcal{G}$, where each individual in the population belongs to one of $g \in \mathcal{G}$. Each individual instance has associated with them a vector of features $x \in \mathcal{X}$ used as input to the classification policy $\pi_g : \mathcal{X} \to [0, 1]$ which outputs probabilities of selection. Each individual might also have a label y denoting the desired prediction or true value for each individual. In this chapter we take the labels to be binary variables, i.e. $y \in \{0, 1\}$. Let $\{\mathcal{D}_g^{\text{emp}}\}_{g\in\mathcal{G}}$ denote the empirical distribution of features and labels for each group, so that drawing $(X, Y) \sim \mathcal{D}_g^{\text{emp}}$ puts equal probability mass on the (feature, label) pairs corresponding to each individual in group g.

A *demographic parity* (also referred to as statistical parity) constraint [96] requires that the same fraction of each group is selected (classified as a positive instance). In terms of the setting and notation developed above:

Definition 2.1 (Demographic parity)**.** The (exact) demographic parity constraint requires that *selection rates* are equal across groups:

$$\mathbb{E}_{(X,Y)\sim\mathcal{D}_g^{\text{emp}}}[\pi_g(X)] = \mathbb{E}_{(X,Y)\sim\mathcal{D}_{g'}^{\text{emp}}}[\pi_{g'}(X)] \quad \forall g, g' \in \mathcal{G} \ .$$

While intuitive, demographic parity only requires that the same fraction of each group is selected; it does not put constraints on which individuals in each group should be selected. In other words, it ignores the outcomes or true labels y. A concern is that the selection policy for one group might correctly classify or select β fraction of individuals from group

[2]We use the terms group fairness constraints and demographic fairness criteria interchangeably, noting that in the past group fairness has been used to refer to demographic parity (Definition 2.1).

g, but choose completely at random β fraction of individuals from group g'. An alternative constraint, *equality of opportunity* [125], requires parity of true positive rates across groups.

Definition 2.2 (Equality of opportunity)**.** The equal opportunity (equalized odds) constraint requires that *true positive rates* are equal across groups:

$$\mathbb{E}_{(X,Y)\sim\mathcal{D}_g^{\text{emp}}}[\pi_g(X)|Y=1] = \mathbb{E}_{(X,Y)\sim\mathcal{D}_{g'}^{\text{emp}}}[\pi_{g'}(X)|Y=1] \quad \forall g, g' \in \mathcal{G} \ .$$

The equality of opportunity constraint is a statement on conditional expectations – of the individuals in each group whose true labels are 1, the selection rates must be equal between groups. A related constraint, *equalized odds* [125], additionally requires selection rates be equal among individuals in each group whose true labels are 0.

Definitions 2.1 and 2.2 state the fairness criteria in terms of probabilistic selection or classification policies $\pi_g(x)$. In some cases it might be more convenient to consider the output of a learning algorithm to be a predicted *score* $\hat{y} \in \mathbb{R}$ rather than probability of selection directly. Liu et al. [179] discuss a correspondence between these settings: informally, when higher scores $\hat{y}$ correspond to higher expected values of both the main metric (e.g. higher probability that $y = 1$) and outcomes for individuals as a result of selection, optimal policies subject to demographic parity and equal opportunity constraints correspond to selecting individuals with scores $\hat{y}$ above some group-dependent thresholds.

2.3 Limitations and Alternatives to Group Fairness

While group fairness metrics and criteria can encode rich and meaningful notions, scholarship has also exposed key limitations of group-based fairness metrics. Alternative frameworks for describing fairness expand the the set of outcomes that can be identified and achieved with algorithmic systems. However, some issues with existing fairness notions may be inherent to classification or prediction systems at their root. In such cases, it is important to critically consider the value of deploying learned algorithmic systems alongside their potential for harm.

Limitations of Group Fairness

Even within the problem framing described in Section 2.2, there are limitations to demographic fairness notions. Formative works prove the impossibility of satisfying multiple group fairness criteria simultaneously in general problem settings [66, 164]. It has also been shown that applying fairness notions can lead to harmful "delayed impacts," as over- or under-selecting of individuals from certain groups to satisfy fairness conditions can leave those groups worse off than had no fairness constraint been applied [179].

The focus on fixed, categorical group labels has also received scrutiny, as demographic identities such as race or gender, are more complex than a reduction to a fixed number of

categories can capture [143, 253]. A number of alternative frameworks do not rely strictly on such groupings, as we describe next.

Alternative Framings of Fairness

One way to avoid assigning explicit demographic groups for fairness interventions is to ask that such fairness hold for *all possible groupings* of instances, usually up to some computational limits on group size or identifiability [128, 162]. In a different vein, *individual fairness* posits that instances that are similar in factors relevant to prediction should have similar predicted values [96, 301]. Several works have exposed tensions between jointly achieving individual and group fairness [96] or achieving individual fairness and accuracy [306]. Binns [35] discusses that understanding the normative or empirical assumptions underlying how disparities might arise can contextualize whether individual or group fairness would be more suitable.

Returning to Intent and Impact

Fairness has become a sub-field in machine learning largely in the service of mitigating unintented harmful impacts of learned algorithmic systems. To understand the impacts *of* imposing fairness criteria on classification systems, Liu et al. [179] posit a model of how individuals will fair as a result of classification. If such a model of individual welfare is available, could we conceivably encode welfare directly as an intent? The next chapter in this thesis directly addresses that question, encoding welfare as one objective in a multi-objective optimization problem.

Before diving in to another solution-oriented view in Chapter 3, it is worth taking a moment to reflect on the larger nature of aligning intent and impact of algorithmic systems. Here is a natural point to consider whether machine learning or computing-based solutions can and will confer the impact the designer intends.[3] An unavoidable answer is that some machine learning systems will pose more threat of harm than potential benefit (and thus should not be deployed or require major modification) [24]. Understanding how and when bias and harm might arise is key to assessing algorithmic systems within their context of intended and eventual use. While machine learning and computing more broadly do have roles to play toward increasing societal well-being [3, 248], realizing the potential for these technologies may mean reimagining their scopes and modes of deployment.

[3] A starting point is to delineate and publicize the intent of machine learning systems, an idea central to the process of "model cards for model reporting" [201].

Chapter 3

Delineating Intent with Multiple Objectives: Welfare-Aware Optimization

This chapter is based on the paper "Balancing competing objectives with noisy data: Score-based classifiers for welfare-aware machine learning" [246], written in collaboration with Max Simchowitz, Sarah Dean, Lydia T. Liu, Daniel Björkegren, Moritz Hardt, and Joshua Blumenstock.

While real-world decisions involve many competing objectives, algorithmic decisions are often evaluated with a single objective function. In this chapter, we study algorithmic policies which explicitly trade off between a private objective (such as profit) and a public objective (such as social welfare). We analyze a natural class of policies which trace an empirical Pareto frontier based on learned scores, and focus on how such decisions can be made in noisy or data-limited regimes. Our theoretical results characterize the optimal strategies in this class, bound the Pareto errors due to inaccuracies in the scores, and show an equivalence between optimal strategies and a rich class of fairness-constrained profit-maximizing policies. We then present empirical results in two different contexts — online content recommendation and sustainable abalone fisheries — to underscore the applicability of our approach to a wide range of practical decisions. Taken together, these results shed light on inherent trade-offs in using machine learning for decisions that impact social welfare.

3.1 Background

From medical diagnosis and criminal justice to financial loans and humanitarian aid, consequential decisions increasingly rely on data-driven algorithms. Machine learning algorithms used in these contexts are mostly trained to optimize a single metric of performance. As a result, the decisions made by such algorithms can have unintended adverse side effects: profit-maximizing loans can have detrimental effects on borrowers [264] and fake news can undermine democratic institutions [223].

As discussed in Chapter 2, the field of fair machine learning proposes algorithmic approaches that mitigate the adverse effects of single objective maximization. Thus far it has predominantly done so by defining various fairness criteria that an algorithm ought to satisfy (see e.g., [26], and references therein). However, a growing literature highlights the inability of any one fairness definition to solve more general concerns of social equity [75]. The impossibility of satisfying all desirable criteria [164] and the unintended consequences of enforcing parity constraints based on sensitive attributes [158] indicate that existing fairness solutions are not a panacea for these adverse effects. Recent work [136, 179] contend that while social welfare is of primary concern in many applications, common fairness constraints may be at odds with the relevant notion of welfare.

In this chapter, we consider *welfare-aware machine learning* as an inherently multi-objective problem that requires explicitly balancing multiple objectives and outcomes. A central challenge is that certain objectives, like welfare, may be harder to measure than others. Building on the traditional notion of Pareto optimality, which provides a characterization of optimal policies under complete information, we develop methods to balance multiple objectives when those objectives are measured or predicted with error.

We study a natural class of selection policies that balance multiple objectives (e.g., private profit and public welfare) when each individual has predicted *scores* for each objective (e.g., their predicted contribution to total welfare and profit). We show that this class of score-based policies has a natural connection to statistical parity constrained classifiers and their ϵ-fair analogs. In the likely case where scores are imperfect predictors, we bound the sub-optimality of the multi-objective utility as a function of the estimator errors. Simulation experiments highlight characteristics of problem settings (e.g. correlation of the true scores) that affect the extent to which we can jointly maximize multiple objectives.

We apply the multi-objective framework to data from two diverse decision-making settings. We first consider an ecological setting of sustainable fishing, where we study score degradation to mimic certain dimensions being costly or impossible to measure. In our second empirical study, we use existing data on the popularity and "social health" of roughly 40,000 videos promoted by YouTube's recommendation algorithm. We show that multi-objective optimization could produce substantial increases in average video quality for almost negligible reductions in user engagement.

This chapter provides a characterization, theoretical analysis, and empirical study of a score-based multi-objective optimization framework for learning welfare-aware policies. We hope that our framework may help decouple the complex problem of defining and measuring welfare, which has been studied at length in the social sciences, e.g. [87], from a machine toolkit geared towards optimizing it.

Fair and Welfare-Aware Machine Learning

The growing subfield of *fairness in machine learning* has investigated the implementation and implications of machine learning algorithms that satisfy definitions of fairness [26, 27, 96]. Machine learning systems in general cannot satisfy multiple definitions of group fairness

[66, 164], and there are inherent limitations to using observational criteria [160]. Alternative notions of fairness more directly encode specific trade-offs between separate objectives, such as per-group accuracies [162] and overall accuracy versus a continuous fairness score [314]. These fairness strategies represent trade-offs with domain specific implications, for example in tax policy [104] or targeted poverty prediction [210].

An emerging line of work is concerned with the long-term impact of algorithmic decisions on societal welfare and fairness [100, 137, 181, 206]. Liu et al. [179] investigated the potentially harmful delayed impact that a fairness-satisfying decision policy has on the well-being of different subpopulations. In a similar spirit, Hu and Chen [136] showed that always preferring "more fair" classifiers does not abide by the Pareto Principle (the principle that a policy must be preferable for at least one of multiple groups) in terms of welfare. Motivated by these findings, our framework acknowledges that algorithmic policies affect individuals and institutions in many dimensions, and explicitly encodes these dimensions in policy optimization.

We will show that fairness constrained policies that result in per-group score thresholds and their ϵ-fair equivalent soft-constrained analogs [99] can be cast as specific instances of the Pareto framework that we study. Analyzing the limitations of this optimization regime with imperfect scores therefore connects to a recent literature on achieving group fairness with noisy or missing group class labels [14, 172], including using proxies of group status [63]. The explicit welfare effects of selection in our model also complement the notion of utilization in fair allocation problems [92, 99].

Multi-Objective Machine Learning

We consider two simultaneous goals of a learned classifier: achieving high profit value of the classification policy, while improving a measure of social welfare. This relates to an existing literature on multi-objective optimization in machine learning [148, 149], where many algorithms exist for finding or approximating global optima under different problem formulations [89, 91, 165].

This chapter studies the Pareto solutions that arise from learned score functions, a problem space related to, but distinct from a large literature on learning Pareto frontiers directly. Evolutionary strategies are a popular class of approaches to estimating a Pareto frontier from empirical data, as they refine a class of several policies at once [89, 161]. Many of these strategies use surrogate convex loss functions to afford better convergence to solutions. Surrogate functions can be defined over each dimension independently [165], or as a single function over both objective dimensions [183]. While surrogate loss functions play an important role in a direct optimization of non-convex utility functions, our framework provides an alternative approach, so long as scores functions can be reliably estimated.

Another class of methods explicitly incorporates models of uncertainty in dual-objective optimization [219, 221]. For sequential decision-making, there has been recent work on finding Pareto-optimal policies for reinforcement learning settings [178, 242, 283]. To promote applicability of our results to a variety of real-world domains where noise sources are diverse

and the effects of single policy enactments complex, we first develop a methodology under a noise-free setting, then extend to reasonable forms of error in provided estimates.

Measures of Social Welfare

The definition and measurement of welfare is a complex problem that has received considerable attention in the social science literature (cf. [87, 88, 267]). There, a standard approach is to sum up individual measures of welfare, to obtain an aggregate measure of societal welfare. The separability assumption (independent individual scores) is a standard simplifying assumption (e.g. [105]) that appears in the foundational work of Pigou [224], as well as many others [9, 50, 252, 257]. Future work may explore alternative social welfare function (cf. [69]). Our focus is on bringing machine learning to the most common notion of welfare.

3.2 Problem Setting: Pareto-Optimal Policies

In this chapter we consider a setting in which a centralized policymaker has two simultaneous objectives: to maximize some private return (such as revenue or user engagement), which we generically refer to as *profit*; and to improve a public objective (such as social welfare or user health), which we refer to as *welfare*. The policymaker makes decisions about *individuals*, who are specified by feature vectors $x \in \mathbb{R}^d$. Decision policies are functions that output a randomized decision $\pi(x) \in [0, 1]$ corresponding to the *probability* that an individual with features x is selected. To each individual we associate a value p representing the expected profit to be garnered from approving this individual and w encoding the change in welfare. The profit and welfare objectives are thus expectations over the joint distribution of (w, p, x):

$$\mathcal{U}_{\mathsf{W}}(\pi) = \mathbb{E}[w \cdot \pi(x)] \quad \text{and} \quad \mathcal{U}_{\mathsf{P}}(\pi) = \mathbb{E}[p \cdot \pi(x)] \,. \tag{3.1}$$

This aggregate measure of societal welfare is defined as a sum of individual measures of welfare; this is a standard approach in the social science literature (see Section 3.1). While this induces limitations on the form of the welfare function, it affords flexibility when focusing instead on the resulting binary decision, a point we expand on in Section 3.5.

Given two objectives, one can no longer define a unique optimal policy π. Instead, we focus on policies π which are *Pareto-optimal* [218], in the sense that they are not strictly dominated by any alternative policy, i.e. there is no π' such that both $\mathcal{U}_{\mathsf{P}}$ and $\mathcal{U}_{\mathsf{W}}$ are strictly larger under π'.

For a general set of policy classes (defined in Proposition 3.5)[1], it is equivalent to consider policies that maximize a weighted combination of both objectives. We can thus parametrize the Pareto-optimal policies by $\eta \in [0, 1]$:

[1]For clarity of exposition, some formal results are defered to Section 3.A.

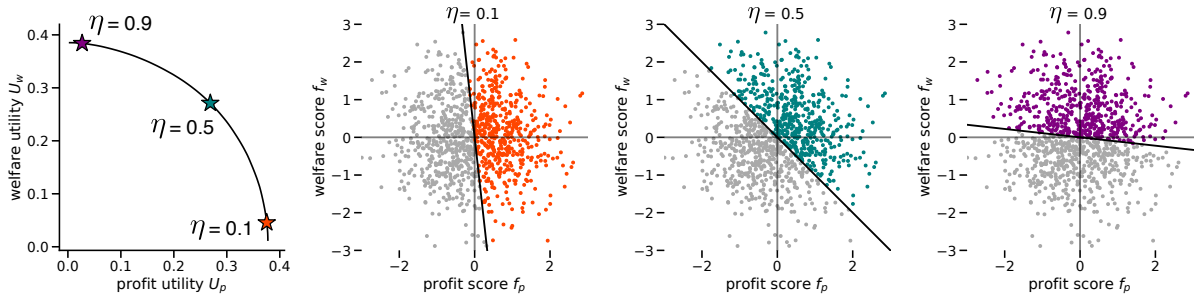


Figure 3.1: Illustration of a Pareto curve (left) and the decision boundaries induced by three different trade-off parameters η. Colored (darker in gray scale) points indicate selected individuals.

Definition 3.1 (Pareto-optimal policies)**.** For policy class Π, an η-Pareto-optimal policy (for $\eta \in [0,1]$) satisfies:

$$\pi^\star_\eta \in \operatorname*{argmax}_{\pi\in\Pi} \mathcal{U}_\eta(\pi),$$
$$\mathcal{U}_\eta(\pi) := (1-\eta)\mathcal{U}_\mathsf{P}(\pi) + \eta\mathcal{U}_\mathsf{W}(\pi).$$

The *Pareto frontier (curve)* of a policy class is the set of all profit-welfare utilities (p, w) obtained by Pareto-optimal policies:

$$\mathcal{P}(\Pi) := \{(p,w) = (\mathcal{U}_\mathsf{P}(\pi), \mathcal{U}_\mathsf{W}(\pi)) \in \mathbb{R}^2 : \ \pi \text{ is } \eta\text{-Pareto-optimal for } \Pi \text{ and some } \eta \in [0,1]\}.$$

In instantiating Definition 3.1 above, a natural choice might be to take Π to be class of randomized policies $\pi(x) \to [0,1]$. We next show that when features x can exactly encode scores, the optimal policy is a threshold function of the scores.

Optimal Policies with Exact Scores

We briefly consider an idealized setting where the welfare and profit contributions w and p can be directly determined from the features x via exact *score functions*, $f_\mathsf{W}(x) = w$, $f_\mathsf{P}(x) = p$. These exact score functions can be thought of as sufficient statistics for the decision: the expected weighted contribution from accepted individuals is described by $((1-\eta)p + \eta w)$. Therefore, one can show (Proposition 3.6) that the optimal policy is given by thresholding this composite:

$$\pi^\star_\eta(p,w) = \mathbb{I}((1-\eta)p + \eta w \geq 0). \tag{3.2}$$

Though they are all Pareto-optimal, the policies $\pi^\star_\eta$ induce different trade-offs between the two objectives. The parameter η determines this trade-off, tracing the Pareto frontier:

$$\mathcal{P}_{\text{exact}} := \{(\mathcal{U}_\mathsf{P}(\pi^\star_\eta), \mathcal{U}_\mathsf{W}(\pi^\star_\eta)) : \eta \in [0,1]\}.$$

Figure 3.1 plots an example of this curve (bottom-left panel) and the corresponding decision rules for three points along it. We note the concave shape of this curve, a manifestation of *diminishing marginal returns*: as a decision policy forgoes profit to increase total welfare, less welfare is gained for the same amount of profit forgone. The notion of diminishing returns is formalized in Theorem 3.3.

3.3 Pareto Frontiers with Inexact Scores

In many settings, we typically do not know the profit score p or welfare score w — or the score functions f_{P} and f_{W} — for all individuals a priori. Instead, we might estimate score functions $\widehat{f_{\mathsf{P}}}(x)$ and $\widehat{f_{\mathsf{W}}}(x)$ from data in the hope that these models can provide good predictions on future examples. We study the class of *score-based policies* that act on the predicted scores:

Definition 3.2 (Score-based policy class)**.**

$$\Pi_{\text{emp}} := \{\, \pi : (\widehat{f_{\mathsf{P}}}(X), \widehat{f_{\mathsf{W}}}(X)) \mapsto [0,1] \,\}$$

Focusing on this class of policies allows us to characterize optimal policies within this class, derive diagnosable bounds the utility of suboptimal policies, and relate our results to common fairness criteria. We summarize additional benefits as well as potential limitations of restricting our study to this policy class in Section 3.6.

Pareto-Optimality for Learned Scores

To characterize Pareto-optimal policies over Π_{emp}, we define the following conditional expectations over the distribution $\mathcal{D}$ of (x, p, w):

$$\overline{\mu}_{\mathsf{P}}(\widehat{f_{\mathsf{P}}}(x), \widehat{f_{\mathsf{W}}}(x)) := \mathbb{E}_{\mathcal{D}}[p \mid \widehat{f_{\mathsf{P}}}(x), \widehat{f_{\mathsf{W}}}(x)] \,,$$
$$\overline{\mu}_{\mathsf{W}}(\widehat{f_{\mathsf{P}}}(x), \widehat{f_{\mathsf{W}}}(x)) := \mathbb{E}_{\mathcal{D}}[w \mid \widehat{f_{\mathsf{P}}}(x), \widehat{f_{\mathsf{W}}}(x)] \,.$$

Intuitively, these values represent our best guesses of p and w, given the predicted scores. We define π_η^{opt} as the threshold policy on the composite of these predictions:

$$\pi_\eta^{\text{opt}} := \mathbb{I}((1-\eta) \cdot \overline{\mu}_{\mathsf{P}} + \eta \cdot \overline{\mu}_{\mathsf{W}} \geq 0).$$

Theorem 3.1 (Pareto frontier in inexact knowledge case)**.** *Given any population distribution $\mathcal{D}$ over (x, p, w) and empirical score functions $\widehat{f_{\mathsf{W}}}$ and $\widehat{f_{\mathsf{P}}}$,*

(i) The policies π_η^{opt} are Pareto optimal over the class Π_{emp}, with $\pi_\eta^{\text{opt}} \in \operatorname{argmax}_{\pi \in \Pi_{\text{emp}}} \mathcal{U}_\eta(\pi)$.

(ii) The Pareto frontier $\mathcal{P}(\Pi_{\text{emp}})$ is given by $\{(\mathcal{U}_{\mathsf{P}}(\pi_\eta^{\text{opt}}), \mathcal{U}_{\mathsf{W}}(\pi_\eta^{\text{opt}}) : \eta \in [0,1]\}$. The associated function mapping $\sup_{\pi \in \Pi_{\text{emp}}} \{\mathcal{U}_{\mathsf{W}}(\pi) : \mathcal{U}_{\mathsf{P}}(\pi) = p\}$ is concave and non-increasing in p.

(iii) The empirical frontier $\mathcal{P}_{\text{emp}}$ is dominated by the exact frontier $\mathcal{P}_{\text{exact}}$. That is, if $(p, w_{\text{exact}}) \in \mathcal{P}_{\text{exact}}$ and $(p, w_{\text{emp}}) \in \mathcal{P}_{\text{emp}}$, then $w_{\text{emp}} \leq w_{\text{exact}}$.

Theorem 3.1 says that an optimal empirical-score based policy can also be realized as a threshold policy (this time of the conditional expectations $\overline{\mu}_{\mathsf{P}}$ and $\overline{\mu}_{\mathsf{W}}$), and it obeys the same diminishing-returns phenomenon as in the exact score case. One example of score predictors that achieves this optimality is the *Bayes optimal estimators* i.e., $\widehat{f}_{\mathsf{P}}(x) = \mathbb{E}[p \mid x]$ and $\widehat{f}_{\mathsf{W}}(x) = \mathbb{E}[w \mid x]$.

Proof of Theorem 3.1. For the theorem, we assume that both policies based on empirical scores and those based on exact scores are *well-behaved* in the sense defined in Assumption 3.1. We first establish part (i), namely that

$$\pi_\eta^{\text{opt}} \in \operatorname*{argmax}_{\pi \in \Pi_{\text{emp}}} \mathbb{E}\left[\mathcal{U}_\eta(\pi(\widehat{f}_{\mathsf{W}}, \widehat{f}_{\mathsf{P}}))\right].$$

Recall that $\pi_\eta^{\text{opt}} := (1-\eta)\cdot\overline{\mu}_{\mathsf{P}} + \eta\cdot\overline{\mu}_{\mathsf{W}}$. We have that

$$\begin{aligned}
\mathcal{U}_\eta(\pi) &= \mathbb{E}[((1-\eta)p + \eta w)\,\pi(\widehat{f}_{\mathsf{P}}, \widehat{f}_{\mathsf{W}})] \\
&= \mathbb{E}[\Big((1-\eta)\mathbb{E}[p \mid \widehat{f}_{\mathsf{P}}, \widehat{f}_{\mathsf{W}}] + \eta\mathbb{E}[w \mid \widehat{f}_{\mathsf{P}}, \widehat{f}_{\mathsf{W}}]\Big)\cdot\pi(\widehat{f}_{\mathsf{P}}, \widehat{f}_{\mathsf{W}})] \\
&:= \mathbb{E}[\Big((1-\eta)\overline{\mu}_{\mathsf{P}}(\widehat{f}_{\mathsf{P}}, \widehat{f}_{\mathsf{W}}) + \eta\overline{\mu}_{\mathsf{W}}(\widehat{f}_{\mathsf{P}}, \widehat{f}_{\mathsf{W}})\Big)\cdot\pi(\widehat{f}_{\mathsf{P}}, \widehat{f}_{\mathsf{W}})] \\
&\leq \mathbb{E}\left[\max\left\{(1-\eta)\overline{\mu}_{\mathsf{P}}(\widehat{f}_{\mathsf{P}}, \widehat{f}_{\mathsf{W}}) + \eta\overline{\mu}_{\mathsf{W}}(\widehat{f}_{\mathsf{P}}, \widehat{f}_{\mathsf{W}}), 0\right\}\right] \;=\; \mathcal{U}_\eta(\pi_\eta^{\text{opt}}).
\end{aligned}$$

We obtain the Pareto optimality of π_η^{opt} by Proposition 3.5 (proved in Section 3.A), which establishes the correspondence between Pareto optimal policies optimizers of a composite objective given in Definition 3.1.

Part (ii) is a direct consequence of Lemma 3.2 (see Section 3.A) and the assumption that our policy class is well behaved. For part (iii), empirical policies are dominated by those induced by the true score functions because, as established, the Pareto optimal policies based on the true score functions are in fact Pareto optimal over all policies that are induced by a function of the features x. □

Plug-In Policies

In general, we may have access to score predictions or the ability to learn them from data, but not a guarantee that the predictions are Bayes optimal. In the hopes that the predicted scores will suffice, a natural selection rule is based on η-defined *plug-in threshold policies*:

Definition 3.3 (Plug-in policy)**.** For $\eta \in [0,1]$ and score predictions $\widehat{f}_{\mathsf{P}}(x), \widehat{f}_{\mathsf{W}}(x)$, the η-plug-in policy is:

$$\pi_\eta^{\text{plug}}(x) = \mathbb{I}\left((1-\eta)\widehat{f}_{\mathsf{P}}(x) + \eta\widehat{f}_{\mathsf{W}}(x) \geq 0\right). \tag{3.3}$$

Since π_η^{opt} requires computing conditional expectations over the distribution $\mathcal{D}$, it will in general will differ from the plug-in policy. The following corollary of Theorem 3.1 gives a condition in which π_η^{opt} and π_η^{plug} coincide.

Corollary 3.1. *The plug-in policies π_η^{plug} are optimal in the class Π_{emp} as long as the predicted score functions are well-calibrated, in the sense that $\mathbb{E}[p \mid \widehat{f}_{\mathsf{P}}(x), \widehat{f}_{\mathsf{W}}(x)] = \widehat{f}_{\mathsf{P}}(x)$ and $\mathbb{E}[w \mid \widehat{f}_{\mathsf{P}}(x), \widehat{f}_{\mathsf{W}}(x)] = \widehat{f}_{\mathsf{W}}(x)$.*

Proof of Corollary 3.1. In this case, $\bar{\mu}_p = \widehat{f}_{\mathsf{P}}(x)$ and $\bar{\mu}_w = \widehat{f}_{\mathsf{W}}(x)$, so we may invoke Theorem 3.1. □

Under typical conditions [180], this form of calibration can be achieved by empirical risk minimization.

In Proposition 3.1, we bound the error in the plug-in policies by the error by the individual errors in each score. In Section 3.4, simulation experiments detail the use of the plug in policy under controlled degradations of learned score accuracy and real-data experiments provide further insight into using the plug-in policy for welfare-aware optimization in practice.

Bounding Pareto Inefficiencies

Even when plug-in policies are not optimal, the sub-optimality of the resulting classifier in terms of the utility function $\mathcal{U}_\eta$ is bounded by the η-weighted sum of ℓ_1 errors in the profit and welfare scores.

Proposition 3.1 (Sub-optimality bound)**.** *For any score prediction functions $\widehat{f}_{\mathsf{P}}(x), \widehat{f}_{\mathsf{W}}(x)$ and $\eta \in [0,1]$, the gap in η-utility from applying the plug-in policy* (3.3) *with $\widehat{f}_{\mathsf{P}}(x), \widehat{f}_{\mathsf{W}}(x)$ versus applying the optimal policy* (3.2) *with true scores $f_{\mathsf{P}}, f_{\mathsf{W}}$, is bounded above by*

$$\mathcal{U}_\eta(\pi_\eta^\star) - \mathcal{U}_\eta(\pi_\eta^{\text{plug}}) \le (1-\eta)\mathbb{E}[|\widehat{f}_{\mathsf{P}}(x) - f_{\mathsf{P}}(x)|] + \eta\mathbb{E}[|\widehat{f}_{\mathsf{W}}(x) - f_{\mathsf{W}}(x)|]. \tag{3.4}$$

Proof of Proposition 3.1. We compute

$$\mathcal{U}_\eta(\pi_\eta^{\text{plug}}) - \mathcal{U}_\eta(\pi_\eta) = \mathbb{E}[((1-\eta)p + \eta w)\left(\pi_\eta^{\text{plug}} - \pi_\eta\right)] \ .$$

Define the functions $Y(x) = (1-\eta)f_{\mathsf{P}}(x) + \eta f_{\mathsf{W}}(x)$, and let $E(x) = (1-\eta)(\widehat{f}_{\mathsf{P}}(x) - f_{\mathsf{P}}(x)) + \eta(\widehat{f}_{\mathsf{W}}(x) - f_{\mathsf{W}}(x))$. Then, $\pi_\eta^{\text{plug}}(x) - \pi_\eta^\star = \mathbb{I}(Y(x) + E(x) \ge 0) - \mathbb{I}(Y(x) \ge 0)$. We see that this difference is at most 1 in magnitude, and is 0 unless possibly if $|Y(x)| \le |E(x)|$. Hence,

$$|Y(x)| \cdot |\pi_\eta^{\text{plug}}(x) - \pi_\eta^\star(x)| \le |E(x)| \ .$$

Therefore

$$\begin{aligned}
|\mathcal{U}_\eta(\pi_\eta^{\text{plug}}) - \mathcal{U}_\eta(\pi_\eta^\star)| &= |\mathbb{E}[Y(x)(\pi_\eta^{\text{plug}}(x) - \pi_\eta^\star(x))]| \\
&\le \mathbb{E}[|Y(x)| \cdot |\pi_\eta^{\text{plug}}(x) - \pi_\eta^\star(x)|] \\
&\le \mathbb{E}[|E(x)|] = \mathbb{E}[|(1-\eta)(\widehat{f}_{\mathsf{P}}(x) - f_{\mathsf{P}}(x)) + \eta(\widehat{f}_{\mathsf{W}}(x) - f_{\mathsf{W}}(x))|] \\
&\le (1-\eta)\mathbb{E}[|\widehat{f}_{\mathsf{P}}(x) - f_{\mathsf{P}}(x)|] + \eta\mathbb{E}[|\widehat{f}_{\mathsf{W}}(x) - f_{\mathsf{W}}(x))|].
\end{aligned}$$

□

Note that by definition of $\pi^\star_\eta$, $\mathcal{U}_\eta(\pi^\star_\eta) - \mathcal{U}_\eta(\pi^{\text{plug}}_\eta) \geq 0$. Proposition 3.1 provides a general bound on the η-performance of the plug-in policy which holds for any distribution on scores and estimator errors.[2] To provide further insight, we consider a specific distributional setting:

Example 3.1. Suppose that individuals' true scores are distributed as:

$$(w_i, p_i) \sim_{i.i.d.} \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma_w^2 & \rho\sigma_w\sigma_p\\ \rho\sigma_w\sigma_p & \sigma_p^2\end{bmatrix}\right) \tag{3.5}$$

Let the prediction errors $\varepsilon_{p_i} := \hat{p}_i - p_i$ and $\varepsilon_{w_i} := \hat{w}_i - w_i$ be independent of the true scores p_i, w_i, zero-mean, and sub-Gaussian with parameters σ_{ε_p} and σ_{ε_w}, respectively. ♢

This example elucidates how correlation between profit and welfare scores affects the empirical Pareto frontier.

Proposition 3.2. *In the setting of Example 3.1 with* $-1 \leq \rho \leq 1$, $\mathbb{E}[\mathcal{U}_\eta(\pi^\star_\eta)] = \frac{\sigma_y}{\sqrt{2\pi}}$ *and the expected* η*-utility of the plug in policy is at least:*[3]

$$\mathbb{E}[\mathcal{U}_\eta(\pi^{\text{plug}}_\eta)] \geq \mathbb{E}[\mathcal{U}_\eta(\pi^\star_\eta)]\left(1 - \frac{2\cdot\tilde{\sigma}^2}{\tilde{\sigma}^2 + \sigma_y^2}\right) \tag{3.6}$$

where $\sigma_y^2 = \eta^2\sigma_w^2 + (1-\eta)^2\sigma_p^2 + 2\rho\eta(1-\eta)\sigma_w\sigma_p$ *and* $\tilde{\sigma}^2 = 4(\eta^2\sigma^2_{\varepsilon_w} + (1-\eta)^2\sigma^2_{\varepsilon_p})$.

The proof of Proposition 3.2 is given in Section 3.B. This lower bound is in terms of both the optimal η-utility and a discount factor. Because σ_y^2 is increasing in ρ for any $\eta \in (0,1)$, both of these terms are increasing in ρ. Thus, the expected η-utility of the plug in policy is higher for correlated scores, not only because the optimal η-utility is higher, but also because the discount factor is closer to 1.

Figure 3.2 shows the lower bound on expected η-utility with noisy scores as a function of possible score correlations ρ and trade-off parameters η, for a fixed setting of predictor noise in Example 3.1. For comparatively small error in profit scores and moderate welfare error, the lower bound on the η-utility increases as the correlation (ρ) between the scores increases. This captures how the low-noise profit score indirectly improves decisions about the high-noise welfare. The lower bound is decreasing in η for positive ρ, which reflects the higher variance introduced by placing more weight on the noisier welfare score.

[2] We remark that in general, optimizing arbitrary loss functions for function value states (e.g. estimating η-utilities for all η directly from features) requires a prohibitively large sample [20]. The structures of the combined learning problems and η-utility in our setting allow us to circumvent this lower bound.

[3] The constant on $\tilde{\sigma}^2$ can be reduced to 1 when prediction errors are independent, $\varepsilon_w \perp\!\!\!\perp \varepsilon_p$.

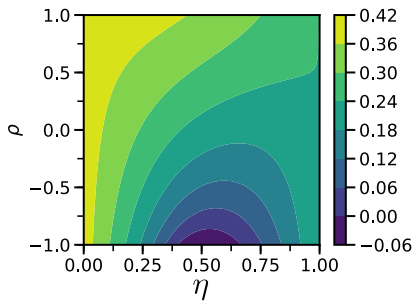


Figure 3.2: Lower bound on expected η-utility as a function of η and correlation in the true scores, from Proposition 3.2, with $\sigma_w = \sigma_p = 1$; $\sigma_{\varepsilon_w} = 0.5$; $\sigma_{\varepsilon_p} = 0.1$.

3.4 Experiments

This section presents three sets of empirical results. We first corroborate our theoretical results under different simulated distributions on scores and prediction errors. Our second experiment studies empirical Pareto frontiers from learned scores with realistic degradation of training data, in the context of sustainable abalone collection. Our third experiment shows how our methods facilitate trading off between user engagement with predicted quality of content in a corpus of YouTube videos, using pre-learned scores.

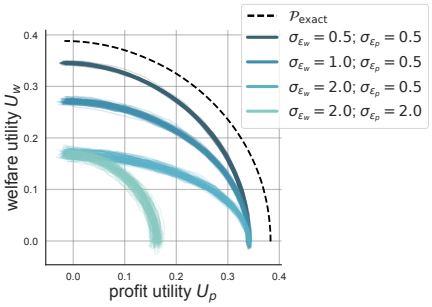
Simulation Experiments

Our first set of simulations shows the performance of the plug-in policy when scores are perturbed by additive noise of varying degrees in each dimension (Figure 3.3a). We instantiate true scores w_i and p_i as in Eq. (3.5) with $\rho = 0$ and $\sigma_w^2 = \sigma_p^2 = 1$, and instantiate predicted scores as:

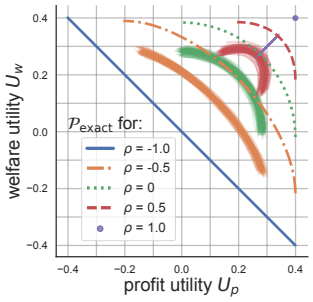
$$\begin{aligned} \widehat{f}_{\mathsf{W}}(x_i) &= w_i + \varepsilon_{w_i} \quad \varepsilon_{w_i} \sim \mathcal{N}(0, \sigma^2_{\varepsilon_w}), \\ \widehat{f}_{\mathsf{P}}(x_i) &= p_i + \varepsilon_{p_i} \quad \varepsilon_{p_i} \sim \mathcal{N}(0, \sigma^2_{\varepsilon_p}). \end{aligned} \tag{3.7}$$

These score predictions satisfy the *well-calibrated* condition of Corollary 3.1. The results for different pairs $(\sigma^2_{\varepsilon_w}, \sigma^2_{\varepsilon_p})$ are shown in Figure 3.3a. As the noise in scores increases, the empirical Pareto frontiers recede from the exact frontier $\mathcal{P}_{\text{exact}}$. Additionally, higher noise in the predicted scores imposes a wider distribution of empirical Pareto frontiers.

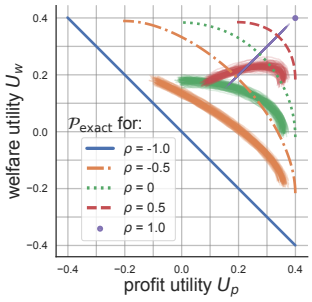
Next, we study the effect of noise in predictions when scores are correlated (Figure 3.3b). We draw w_i and p_i according to Eq. (3.5) with $\sigma_w = \sigma_p = 1$ and correlation parameter ρ. We then add random noise as in Eq. (3.7) with parameters $\sigma_{\varepsilon_w} = \sigma_{\varepsilon_p} = 1.0$. Note that in this setting, scores are in general not calibrated due to the correlation between w_i and p_i. For positive values of ρ, the exact and empirical utilities are greatest at $\eta = 0.5$, since the correlation in the scores allows us to overcome some of the noise in each individual parameter, as explained by Proposition 3.2.



(a) No correlation, varying noise.



(b) Varying correlation, fixed and equal noise $\sigma_{\varepsilon_p} = \sigma_{\varepsilon_w} = 1$.



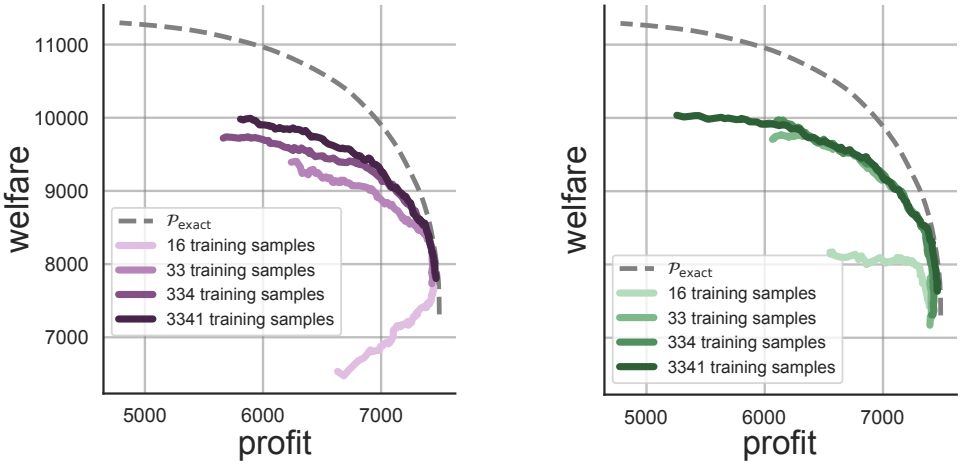
(c) Varying correlation, fixed and unequal noise $\sigma_{\varepsilon_p} = 0.5$, $\sigma_{\varepsilon_w} = 2.0$.

Figure 3.3: Simulated experiments corresponding to the setting in Example 3.1 (fixing $\sigma_w = \sigma_p = 1$). Empirical frontiers $\mathcal{P}(\Pi_{\text{emp}})$ for 100 random trials with $n = 5,000$ each are shown as overlaid translucent curves. Exact frontiers $\mathcal{P}_{\text{exact}}$ are shown as dashed curves.

Lastly, we study the space of empirical and exact frontiers with degraded noise when scores are correlated and prediction error is higher in the welfare the score, with $\sigma_{\varepsilon_p} = 0.5$ whereas $\sigma_{\varepsilon_w} = 2.0$ (Figure 3.3c). While the optimal Pareto frontiers are the same as in Figure 3.3b, we see a stark change in the empirical Pareto frontiers. Compared to the case of no correlation, the empirical Pareto frontier is expanded when $\rho > 0$ and when $\rho < 0$ the frontier recedes. Additionally, we see evidence that due to the correlation, π_η^{plug} is no longer guaranteed to be optimal, as welfare utility decreases for large enough η when $\rho = 0.5$.

Learned Scores with Imperfect Data: Abalone

Our next example is motivated by the domain of ecologically sustainable selection, where the goal is to select profitable mollusks to catch and keep, while having minimal impact on the natural development of the mollusks' ecosystem. We learn scores for the age and profitability of each abalone from data, and perform experiments to test the degradation of the empirical Pareto frontiers under realistic degradations of the data. While our characterization of the problem is highly simplified, the main focus of this experiment is to demonstrate the instantiation of Pareto curves for different predictor function classes and different regimes of data availability.



(a) Ridge regression model.

(b) Random forest model.

Figure 3.4: Abalone empirical frontiers as training set size increases.

The welfare measure we use is an increasing function of age,[4] encoding that it is more sustainable to harvest older abalones. We define the profit score of each abalone as a linear function of meat weight and shell area. We use the features (sex, total weight, height, width, and diameter) to train score predictors. We derive these measures from physical data collected by Nash et al. [208] (accessed via the UCI data repository [94]). The correlation of the profit and welfare scores is 0.56.

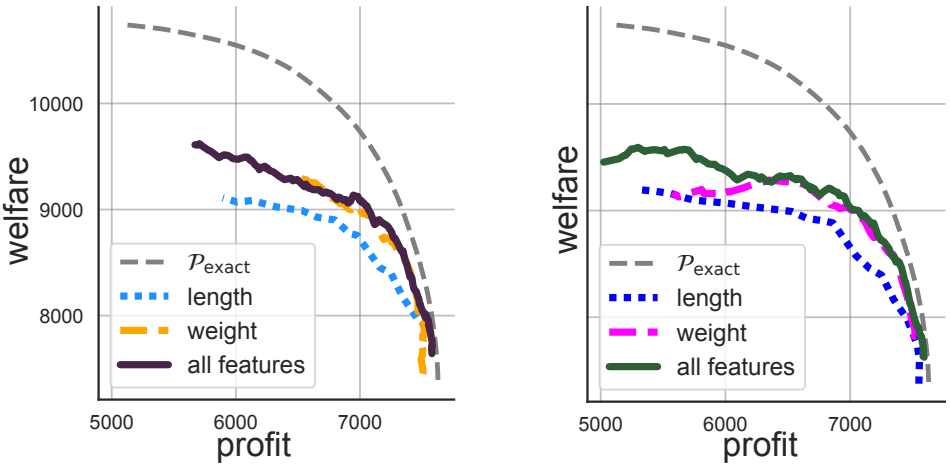
In this setting, we study the effectiveness of two models — ridge regression and random forests — to learn scores with which to instantiate the plug-in policy. To assess how the empirical Pareto frontiers degrade under realistic notions of imperfect data, we subsample training instances to reflect a hypothetical regime were data is sparse and we subsample features to reflect a hypothetical regime where entire measurements were not recorded in the original dataset.

Figure 3.4 shows the empirical Pareto frontiers reached as we change the size of the training data set from which learn the profit and welfare scores. Even with 33 training samples (1% of the original training set), the set of plug-in policies traces a meaningful trade-off over η. For severely degraded scores (16 training samples - just 0.5% of the original

[4]Specifically, we instantiate scores as:

$$p := \texttt{meat\_price\_per\_gram} \cdot (200 \cdot \texttt{shucked\_weight}) + \texttt{shell\_price\_by\_cm}^2 \cdot (20 \cdot \texttt{length}) \cdot (20 \cdot \texttt{diameter})$$
$$w := c \cdot \log((\texttt{rings} + 1.5)/10)$$

where $\texttt{meat\_price\_per\_gram} = 0.25$ and $\texttt{shell\_price\_per\_cm}^2 = 0.32$, and the constant factors of 20 and 200 match units of the original data with units of these prices. We add 1.5 to the ring count to get age, and divide by 10 before taking the logarithm to encode that harvesting abalone less than 10 years of age has negative welfare. We scale the welfare weights by constant c so that the distribution of welfare and profit have the same standard deviation.



(a) Ridge regression model.

(b) Random forest model.

Figure 3.5: Abalone empirical frontiers for different feature sets.

training sets), the error on the welfare score predictions is so high that instantiating a plug-in policy with $\eta > 0$ actually decreases welfare overall.

Figure 3.5 shows the empirical Pareto frontiers reached as we change the features used to train the model, using just length, just weight, or all seven features as in Figure 3.4. The mean average error of welfare scores is substantially greater than the mean average error of profit scores for most prediction settings, thus the empirical frontiers are farther from $\mathcal{P}_{\text{exact}}$ in the welfare dimension than the profit dimension.

Altogether, the empirical Pareto frontiers are relatively robust to small data regimes, as well as to missing predictors. However, when predictions have very high error (diagnosable by cross-validation or holdout set error), empirical Pareto frontiers degrade quickly.

Balancing User Engagement and Health

We now illustrate how the multi-objective framework can be used to balance the desire to promote high quality content with the need for profit. We work with a dataset that contains measures of content quality and content engagement for 39,817 YouTube videos, which was constructed as part of an independent effort to automatically ascertain the quality and truthfulness of YouTube videos [103].

The measure of quality $\widehat{f}_{\mathsf{W}}$ we use is a function of the "conspiracy score" developed by Faddoul, Chaslot, and Farid [103], which estimates the probability that the video promotes a debunked conspiracy theory. From this score $s_{\text{conspiracy}} \in [0, 1]$ we derive a predicted "quality score" as $(0.95 - s_{\text{conspiracy}})$.

We instantiate the profit score $f_{\mathsf{P}}[i]$ for video i as $\log((1 + \#\text{ views}[i])/100,000)$. Dividing by a large constant represents that videos with low view counts may not be profitable due to

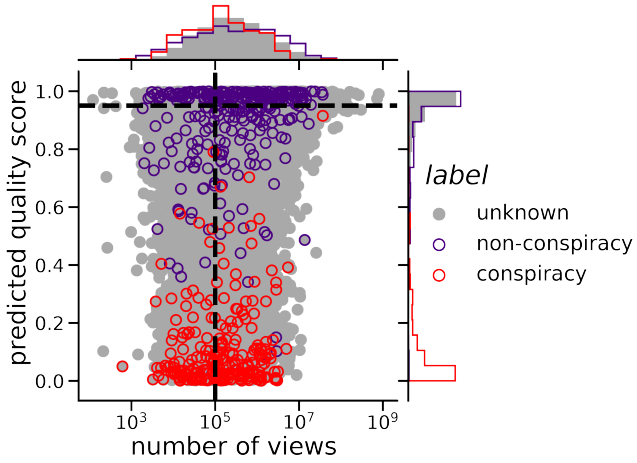
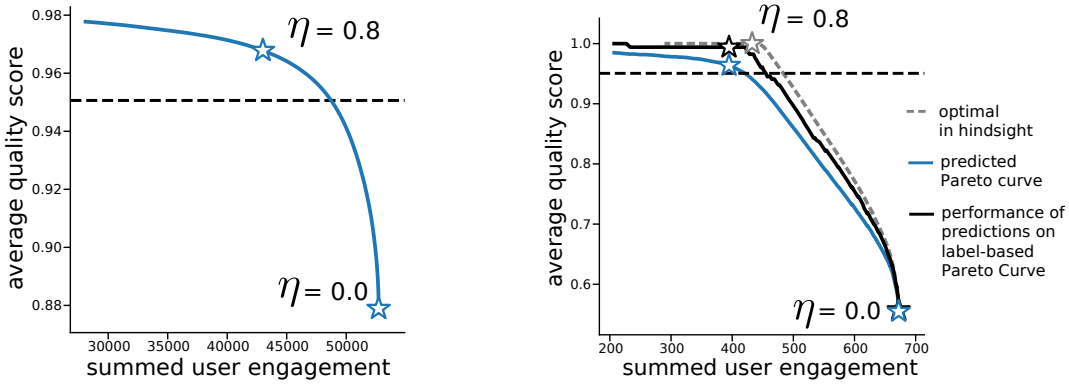


Figure 3.6: Distribution of YouTube data predicted quality scores unlabeled videos (gray), and hand labeled conspiracy (red) and non-conspiracy (purple) videos.



(a) Estimated Pareto curve using pre-computed predictions. Stars indicate specific η trade-offs.

(b) Estimated Pareto curve on labeled data subset (blue). Optimal-in-hindsight curve (dashed gray) and performance of predictions on label set (black).

Figure 3.7: Balancing user engagement and health of hosted YouTube videos.

storage and hosting costs. The resulting distribution over f_{P} and $\widehat{f}_{\mathsf{W}}$ is shown in Figure 3.6 (gray dots), where dotted lines denote 0-utility thresholds in each score.

Using these scores and predictions, we estimate a Pareto frontier using the optimal policies π_η^{plug} for learned scores from Eq. (3.3). The resulting estimated Pareto curve is shown in Figure 3.7a. The curve is concave, demonstrating the phenomenon of diminishing returns in the trade-off between total user engagement and average video quality. While there is always some quality to gain by sacrificing some total engagement, these relative gains are

greatest when the starting point is close to an engagement-maximizing policy. Specifically, at the maximum-engagement end of the spectrum (lower right star), we can gain a 1.1% increase in average video quality for a 0.1% loss in total engagement. However, for a policy with trade-off rate $\eta = 0.8$ (upper left star), to obtain an increase of 0.3% in welfare, a larger loss of 5.2% in user engagement is required.

Next, we assess the validity of this estimated Pareto curve using the small set of 541 hand-labeled training set instances from which $s_{\text{conspiracy}}$ was learned. This assessment is likely optimistic due to the fact that the score predictor functions were trained on this same data; nonetheless, this is an important check to perform on the estimated Pareto frontier.

In Figure 3.7b we plot the optimal-in-hindsight Pareto frontier (dashed gray line) had we known the labels a priori and applied thresholds according to Eq. (3.2). We also plot the performance of our estimated policy π_η^{plug} on the labeled instances (black line). The stars on each curve correspond to decision thresholds with $\eta = 0$ and $\eta = 0.8$, and illustrate the alignment of the curves.

Relating back to Theorem 3.1, we see that performance of the learned scores (black line) is dominated by that of the optimal classifier, as is the predicted Pareto curve (thick blue line). Here the predicted Pareto curve under-predicts the actual performance; in general it is possible for the opposite to be true. Encouragingly, we observe that the curves representing the predicted and actual performance show similar qualitative trade-offs.

3.5 Connections to Fairness Constraints

Having shown our main results on learning Pareto-optimal policies with limited data, we now illustrate connections between our framework and approaches based on fair machine learning that constrain classification decisions to satisfy certain criteria (see Chapter 2). For example, in the setting of hiring or admissions, one might require that the same proportion of male and female candidates are admitted, i.e. demographic parity. We demonstrate that profit maximization with group fairness constraints corresponds to multi-objective optimization over profit and welfare for an induced definition of welfare. This connection illustrates that even though we consider a welfare function defined from individual welfare scores, our framework can encode more collective conceptions of welfare, like those arising from group fairness constraints.

Consider a population partitioned into subgroups $g \in \mathcal{G}$ and a classifier which has access to the profit score p of each individual. In this case, we decompose policies over groups such that $\pi = (\pi_g)_{g \in \mathcal{G}}$ and the fairness-constrained profit maximization is given as

$$\pi^\epsilon_{\text{fair}} \in \underset{\pi,\beta}{\text{argmax}} \ \mathcal{U}_{\mathsf{P}}(\pi) \ \text{ s.t. } \ \mathbb{E}[\pi_g(p) \mid \text{in group } g \cap \mathcal{C}] = \beta_g, \ \ |\beta_{g'} - \beta_g| \leq \epsilon \text{ for all } g, g' \in \mathcal{G} \tag{3.8}$$

where the choice of $\mathcal{C}$ encodes particular fairness criteria. For a large class of fairness criteria including demographic parity (Definition 2.1) and equal opportunity (Definition 2.2), we can

restrict our attention to threshold policies $\pi_g(p) = \mathbb{I}(p \geq t_g)$ where t_g are group-dependent thresholds [179]. Notice that due to the definition of profit utility (3.1), the unconstrained solution would simply be $\pi^{\texttt{MaxUtil}}(p) = \mathbb{I}(p \geq 0)$ for all groups. For this reason, we refer to groups with $t_g < 0$ as comparatively *disadvantaged* (since their threshold increases in the absence of fairness constraints) and $t_g > 0$ as *advantaged*.

In this setting, there exist fixed welfare scores w which achieve the same solution policy for any population.

Proposition 3.3. *Any fairness-constrained threshold policy giving rise to thresholds $\{t_g^\star\}_{g\in\mathcal{G}}$ is equivalent to a set of η-Pareto policies for $\eta \in (0,1)$ in (3.2) with welfare scores fixed within each group and defined as*

$$w_g = -\frac{1-\eta}{\eta} t_g^\star \,.$$

In particular, w_g and $t_g^\star$ have opposite signs for all settings of $\eta \in (0,1)$, and any relative scale between them achieved by some choice of η.

Proof of Proposition 3.3. The equivalence follows by comparing the policies

$$\pi_\eta(w,p) = \mathbb{I}(\eta w + (1-\eta)p \geq 0) \quad \text{and} \quad \pi^\epsilon_{\text{fair},g}(p) = \mathbb{I}(p \geq t_g^\star) \,.$$

Restricting the choice to a fixed score within each group yields the expression

$$w_g = -\frac{1-\eta}{\eta} t_g^\star =: -c t_g^\star \,.$$

Thus we have that $w_g \propto -t_g^*$ for all $g \in \mathcal{G}$. Further, notice that for any $c > 0$ there exists some $\eta \in (0,1)$ achieving that c with $\eta = \frac{1}{1+c}$. □

Trade-Offs between Profit and Fairness.

While the result presented above is valid for even inexact fairness constraints, it does not shed light on the trade-off between profit and fairness as the parameter ϵ varies. We now show how this trade-off in the fairness setting is reflected in an induced problem setting in the multi-objective framework. For simplicity, we restrict our attention to the setting of two groups and criteria of demographic parity. We note that with additional mild assumptions, our arguments extend naturally to other criteria, including equal opportunity (analogously to Section 6.2 of [179]).

Define the two groups as A and B. Assume that the distribution of the profit score p has continuous support within these populations. The following proposition shows that the solution to the constrained profit maximization problem in Eq. (3.8) changes monotonically with the fairness parameter ϵ.

Proposition 3.4. *Suppose that the unconstrained selection rate in group* A *is less than or equal to the unconstrained selection rate in group* B*. Then the policies* $\pi^\epsilon_\mathsf{A}, \pi^\epsilon_\mathsf{B}$ *that optimize Eq.* (3.8) *with the demographic parity constraint are equivalent to randomized group-dependent threshold policies with thresholds* t^ϵ_A *and* t^ϵ_B *satisfying the following:*

- $t^\epsilon_\mathsf{A} \leq 0$ *for all* $\epsilon \geq 0$ *and* t^ϵ_A *is increasing in* ϵ*,*
- $t^\epsilon_\mathsf{B} \geq 0$ *for all* $\epsilon \geq 0$ *and* t^ϵ_B *is decreasing in* ϵ*.*

Notice that the unconstrained selection rate in group A being less than the unconstrained selection rate in group B is equivalent to A being disadvantaged compared with B. Thus we see that as ϵ increases, the group-dependent optimal thresholds shrink toward the unconstrained profit maximizing solution, where $t_\mathsf{A} = t_\mathsf{B} = 0$. For clarity of exposition, we present the proof of this result in Section 3.C.

We define the map $\epsilon_\mathsf{A}(p) := \epsilon$ s.t. $t^\epsilon_\mathsf{A} = p$ for $p \in [t^0_\mathsf{A}, 0]$. By Proposition 3.4, $\epsilon_\mathsf{A}(p)$ in increasing in p. Using this ingredients, we define a policy based on welfare scores which is equivalent to a fair policy.

Theorem 3.2. *Under the conditions of Proposition 3.4, the family of policies* $\pi^\epsilon_{\text{fair}}$ *parametrized by* ϵ *corresponds to a family of* η*-Pareto policies solutions for a fixed choice of group-dependent welfare weightings. In particular, denoting the associated thresholds as* t^ϵ_A *and* t^ϵ_B *and defining for each individual in* A *with profit score* p*,*

$$w_\mathsf{A} = \begin{cases} -\frac{p}{t^{\epsilon_\mathsf{A}(p)}_\mathsf{B}} & t^0_\mathsf{A} \leq p \leq 0 \\ 0 & \textit{otherwise} \end{cases},$$

and for all individuals in B*,*

$$w_\mathsf{B} = \begin{cases} -1 & 0 \leq p \leq t^0_\mathsf{B} \\ 0 & \textit{otherwise} \end{cases},$$

then for each $\pi^\epsilon_{\text{fair}}$ *there exists an equivalent* η^ϵ*-Pareto policy* π_{η^ϵ} *where the trade-off parameter* η^ϵ *decreases in* ϵ*.*

Proof of Theorem 3.2. By Proposition 3.4, the policy π^ϵ is equivalent to a threshold policy with group dependent thresholds denoted t^ϵ_A and t^ϵ_B. The group dependent threshold policy $\mathbb{I}(p \geq t^\epsilon_g)$ is equivalent to an η-Pareto optimal policy (for some definition of welfare score w) if and only if for all values of p:

$$\mathbb{I}\left(p \geq t^\epsilon_g\right) = \mathbb{I}\left(\eta^\epsilon w + (1-\eta^\epsilon)p \geq 0\right).$$

It is sufficient to restrict our attention to welfare scores w that depend on profit score and group membership, which we denote as w^p_g. Starting with group B, we have that for $0 \leq p \leq t^0_\mathsf{B}$, $w^p_\mathsf{B} = -1$, so

$$\pi_{\eta^\epsilon} = \mathbb{I}\left(-\eta^\epsilon + (1-\eta^\epsilon)\, p \geq 0\right) = \mathbb{I}\left(p \geq \frac{\eta^\epsilon}{1-\eta^\epsilon}\right).$$

Thus, equivalence is achieved for this case if $\frac{\eta^\epsilon}{1-\eta^\epsilon} = t^\epsilon_{\mathsf{B}}$, or equivalently,

$$\eta^\epsilon = \frac{t^\epsilon_{\mathsf{B}}}{1 + t^\epsilon_{\mathsf{B}}} \,. \tag{3.9}$$

We will use this definition for η^ϵ moving forward, and verify that the proposed welfare score definitions work.

We now turn to group A in the case that $t^0_{\mathsf{A}} \leq p \leq 0$. We have $w^p_{\mathsf{A}} = \frac{p}{t^{\epsilon_{\mathsf{A}}(p)}_{\mathsf{B}}}$, so

$$\pi_{\eta^\epsilon} = \mathbb{I}\left(-\frac{t^\epsilon_{\mathsf{B}}}{t^{\epsilon_{\mathsf{A}}(p)}_{\mathsf{B}}} \frac{p}{1 + t^\epsilon_{\mathsf{B}}} + \frac{p}{1 + t^\epsilon_{\mathsf{B}}} \geq 0 \right) .$$

Because $1 + t^\epsilon_{\mathsf{B}} \geq 0$ and $p \leq 0$, the indicator will be one if and only if $t^\epsilon_{\mathsf{B}} \geq t^{\epsilon_{\mathsf{A}}(p)}_{\mathsf{B}}$. By Proposition 3.4, this is true if and only if $\epsilon \leq \epsilon_{\mathsf{A}}(p)$, which is true if any only if $t^\epsilon_{\mathsf{A}} \leq t^{\epsilon_{\mathsf{A}}(p)}_{\mathsf{A}} = p$. This is exactly the condition for $\pi^\epsilon_{\text{fair},\mathsf{A}}$, as desired.

Then finally we consider the remaining cases. In the case that $p \leq t^0_{\mathsf{A}}$ in A or $p \leq 0$ in B, we have that $\pi^\epsilon_{\text{fair},g} = 0$ for all ϵ by Proposition 3.4. Then as desired, $0 + (1 - \eta^\epsilon)p \leq 0$ in this case. In the case that $p \geq 0$ in A or $p \geq t^0_{\mathsf{B}}$ in B, we have that $\pi^\epsilon_{\text{fair},g} = 1$ for all ϵ. Then as desired, $0 + (1 - \eta^\epsilon)p \geq 0$ in this case.

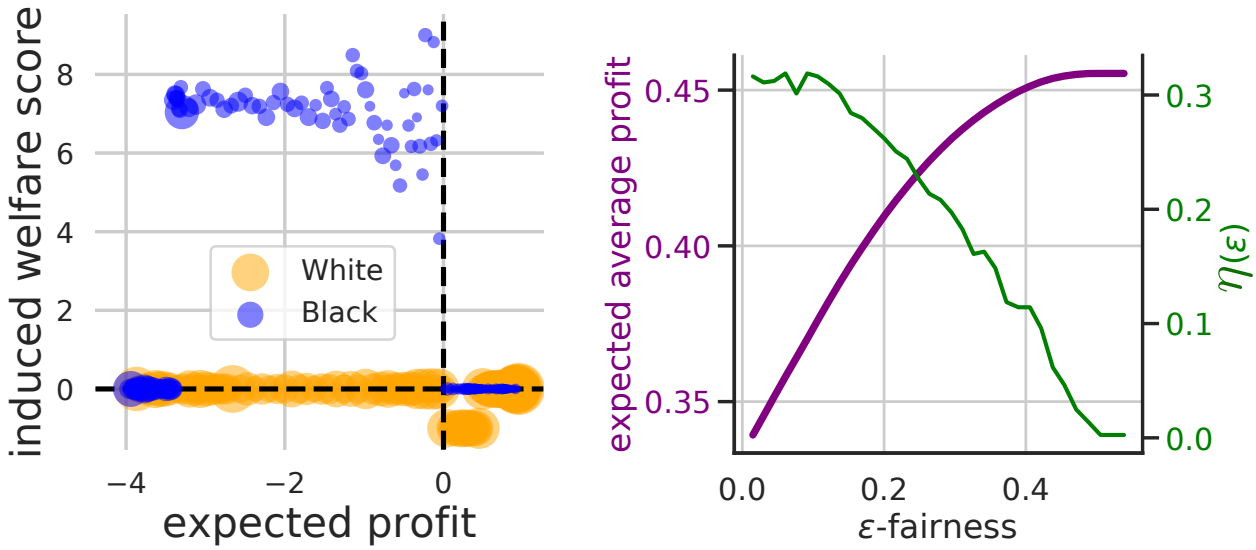
Finally, we remark on the form of η^ϵ. By Proposition 3.4, $t^\epsilon_{\mathsf{B}} \geq 0$ and is decreasing in ϵ, so η^ϵ is decreasing in ϵ. □

This correspondence between any ϵ-fair setting and its corresponding welfare-aware setting is synthesized in the statement of Corollary 3.2, which follows from Theorem 3.2.

Corollary 3.2. *It is possible to define fixed welfare scores such that the family of inexact fair policies parameterized by any $\epsilon \geq 0$ in Eq. (3.8) corresponds to a family of Pareto-optimal policies parameterized by $\eta(\epsilon)$. The group-dependent welfare scores are such that $w \geq 0$ for all individuals in the disadvantaged group and $w \leq 0$ in the advantaged group. Furthermore, the induced trade-off parameter $\eta(\epsilon)$ increases as ϵ decreases.*

Fairness constraints can be seen as encoding implicit group-dependent welfare scores for individuals, where members of disadvantaged groups are assigned positive welfare weights and members of advantaged groups are assigned negative weights. Figure 3.8 illustrates this result applied to data from a credit lending scenario from Barocas, Hardt, and Narayanan [26][5], where welfare scores are induced for individuals depending on their race and likelihood of repayment. This correspondence is related to the analysis of welfare weights by Hu

[5]More specifically, we estimate a distribution of profit scores using repayment information from a sample of 301,536 TransUnion TransRisk scores from 2003 published by US Federal Reserve [282], preprocessed by Hardt, Price, and Srebro [125], and accessible at `https://fairmlbook.org/` [26]. In this demonstration, we define the profit score as the expected gain from lending to an individual, $p = u_+ \cdot \rho + u_- \cdot (1 - \rho)$, where ρ is the individual's probability of repayment. For results in Figure 3.8, we set $u_+ = 1$ and $u_- = -4$, indicating that a default is more costly than a repayment.



(a) Distribution of profit and welfare scores. Marker size indicates population sizes.

(b) The fairness parameter ϵ determines the profit trade-off and corresponds to the welfare weight η.

Figure 3.8: Trade-offs between profit and fairness in lending can be equivalently encoded by a multi-objective framework.

and Chen [137], however, our perspective focuses on trade-offs between welfare and profit objectives, in contrast to pure welfare maximization.

In the case that group membership is believed to correspond to the welfare impact of selection, Corollary 3.2 connects our results in Section 3.3 with a body of work on achieving fairness when group labels are approximate or estimated [153]. While some applications may directly call for statistical parity as a criterion, Corollary 3.2 emphasizes the inevitability of fairness constraints as trade-offs between multiple objectives, and frames these trade-offs explicitly in terms of welfare measures.

3.6 Conclusions

In this chapter we presented a methodology for developing welfare-aware policies that jointly optimize a private return (such as profit) with a public objective (such as social welfare). Taking care to consider data-limited regimes, we develop theory around the optimality of using learned predictors to make decisions. Experiments corroborate our theoretical results, showing that thresholding on predicted scores can approach a Pareto-optimal policy.

This score-based approach to balancing competing objectives with noisy data is attractive for several reasons:

- Score-based policies can trade off multiple objectives with scalar predictions, with error bounded by a weighted sum of the errors in the learned scores.
- The plug-in policy is a learned decision rule that is easily explained and diagnosed — in line with the desire for transparent classification rules in practice.
- It provides a crisp and interpretable connection to fair-constrained profit maximization, but reframes the problem as one of multi-objective optimization (see Section 3.5).

While separating the problem of instantiating learned policies from the problem of learning scores has desirable benefits, we note the limitations of this approach as well. First, the plug-in policy is not guaranteed to be the optimal policy learned from data. Thus, when further assumptions on the problem structure are appropriate, it may be worthwhile to consider more general policy classes learned from data. Second, the score-based approach shifts much of the difficulty of welfare-aware machine learning toward defining and predicting welfare, which is an area of active academic and policy debate [114, 152].

When welfare utilities are estimable, the ability to trade off context-sensitive measures with general policies can improve upon the status quo of applying machine learning policies in welfare-sensitive domains. Further, a multi-objective framework could allow communities to understand the trade-offs between competing definitions of welfare or fairness in data constrained situations.

Taken together, these results help illustrate how machine learning can be used to design policies that prioritize the social impact of an algorithmic decision from the outset, rather than as an afterthought. By elucidating the possible trade-offs between competing objectives, and by illustrating the importance of measurement and prediction error in multi-objective optimization, we hope this work encourages new ways of thinking about welfare-aware machine learning.

3.A Proofs for Characterization of Pareto Curves

Pareto Policies Optimize Weighted Combination of Utilities

Proposition 3.5 (Pareto optimal policies optimize a composite objective)**.** *Let* Π *be a set of policies which is convex, and compact in a topology in which* $\pi \mapsto \mathcal{U}_{\mathsf{P}}(\pi)$ *and* $\mathcal{U}_{\mathsf{W}}(\pi)$ *are continuous.*[6] *Then, a policy* $\pi^\star \in \Pi$ *is Pareto optimal if and only if there exists an* $\eta \in [0, 1]$ *for which*

$$\pi^\star \in \operatorname*{argmax}_{\pi \in \Pi} \ \mathcal{U}_\eta(\pi)$$
$$\mathcal{U}_\eta(\pi) := (1-\eta)\mathcal{U}_{\mathsf{P}}(\pi) + \eta \mathcal{U}_{\mathsf{W}}(\pi).$$

[6]The convexity of Π means that Π is closed under the randomized combination of policies. In the simplest case, compactness is achieved when the space of features is finite (e.g. features x can only take a values in a discrete, finite subset of $\mathbb{R}^d$).

Proof of Proposition 3.5. First, we prove that if $\pi^\star \in \operatorname{argmax}_\pi \ \mathcal{U}_\eta(\pi) := (1-\eta)\mathcal{U}_\mathsf{P}(\pi) + \eta\mathcal{U}_\mathsf{W}(\pi)$, then $\pi^\star$ is Pareto optimal. Suppose that there exists an η for which the policy $\pi^\star \in \operatorname{argmax}_\pi \ \mathcal{U}_\eta(\pi)$. If $\eta \in \{0,1\}$, then $\pi^\star$ maximizes either $\mathcal{U}_\mathsf{W}(\cdot)$ or $\mathcal{U}_\mathsf{P}(\cdot)$, and is therefore Pareto optimal by definition. Otherwise, if $\eta \in (0,1)$, suppose for the sake of contradiction that $\pi^\star$ is not Pareto optimal. Then there exists a policy π for which $\mathcal{U}_\mathsf{W}(\pi^\star) \le \mathcal{U}_\mathsf{W}(\pi)$ and $\mathcal{U}_\mathsf{P}(\pi^\star) \le \mathcal{U}_\mathsf{P}(\pi)$, where one of these inequalities is strict. We can then check that $\mathcal{U}_\eta(\pi^\star) < \mathcal{U}_\eta(\pi)$, contradicting the assumption that $\pi^\star \in \operatorname{argmax}_\pi \ \mathcal{U}_\eta(\pi)$.

To show the other direction, suppose that $\pi^\star$ is Pareto optimal. If $\pi^\star$ maximizes either profit or welfare, then $\pi^\star \in \operatorname{argmax}_\pi \ \mathcal{U}_\eta(\pi)$ for either $\eta = 1$ or $\eta = 0$. Otherwise, let $W = \mathcal{U}_\mathsf{W}(\pi^\star)$. Then, by Pareto optimality,

$$\begin{aligned}
\pi^\star &\in \operatorname{argmax}\{\mathcal{U}_\mathsf{P}(\pi) : \mathcal{U}_\mathsf{W}(\pi) \ge \mathcal{U}_\mathsf{W}(\pi^\star), \pi \in \Pi\} \\
&= \operatorname*{argmax}_{\pi\in\Pi} \left(\mathcal{U}_\mathsf{P}(\pi) + \min_{t\ge 0} t(\mathcal{U}_\mathsf{W}(\pi) - \mathcal{U}_\mathsf{W}(\pi^\star)) \right) \\
&= \operatorname*{argmax}_{\pi\in\Pi} \min_{t\ge 0} \left(\mathcal{U}_\mathsf{P}(\pi) + t\left(\mathcal{U}_\mathsf{W}(\pi) - \mathcal{U}_\mathsf{W}(\pi^\star)\right)\right).
\end{aligned}$$

The map $\mathcal{U}_\mathsf{P}(\pi)$ and $\mathcal{U}_\mathsf{W}(\pi)$ are both linear functions in π. Hence, if Π is a a convex, and compact in a topology in which $\pi \mapsto \mathcal{U}_\mathsf{P}(\pi)$ and $\mathcal{U}_\mathsf{W}(\pi)$ are continuous, Sion's minimax theorem [166] ensures that strong duality holds, which means that we can switch order of the minimization over t and maximization over π. Thus, for some $t \ge 0$,

$$\begin{aligned}
\pi^\star \in \operatorname*{argmax}_\pi \left(\mathcal{U}_\mathsf{P}(\pi) + t\cdot\mathcal{U}_\mathsf{W}(\pi) - t\cdot\mathcal{U}_\mathsf{W}(\pi^\star)\right) &= \operatorname*{argmax}_\pi \left(\mathcal{U}_\mathsf{P}(\pi) + t\cdot\mathcal{U}_\mathsf{W}(\pi)\right) \\
= \operatorname*{argmax}_\pi \left(\frac{1}{1+t}\mathcal{U}_\mathsf{P}(\pi) + \frac{t}{1+t}\mathcal{U}_\mathsf{W}(\pi)\right) &= \operatorname*{argmax}_\pi \left(\mathcal{U}_{t/(1+t)}(\pi)\right),
\end{aligned}$$

as needed. □

Optimal Policies under Exact Information

Here we verify the optimality of threshold policies under exact information:

Proposition 3.6 (Pareto optimal policies on exact scores are threshold policies)**.** *For the weighted utility $\mathcal{U}_\eta$, the optimal policy $\pi^\star_\eta$ over the unrestricted class $\Pi_\star$ is a* threshold *on a weighted combination of w and p, namely*[7]*,*

$$\pi^\star_\eta(p,w) = \mathbb{I}(\eta w + (1-\eta)p \ge 0). \tag{3.10}$$

Proof of Proposition 3.6. Consider the reward of an arbitrary policy π. Recall that $\pi(x) \in [0,1]$ denotes the probability that π classifies an individual with features x as a 1. Since $\pi(x)\cdot z \le z\cdot\mathbb{I}(z\ge 0)$ for $\pi(x)\in[0,1]$ and $z\in\mathbb{R}$, we can bound $\mathcal{U}_\eta(\pi) = \mathbb{E}_x[\pi(x)\cdot(\eta w(x) + (1-\eta)p(x))] \le \mathbb{E}_x[\max\{0, \eta w(x) + (1-\eta)p(x)\}]$. We can directly check that the threshold policy $\pi^\star_\eta(p,w) = \mathbb{I}(\eta w + (1-\eta)p \ge 0)$ saturates this bound. □

[7] When the distribution over (w,p) is sufficiently smooth, we can ignore the case where $\eta w + (1-\eta)p = 0$.

Well-Behaved Pareto Curves

In this section, we establish that under mild regularity conditions, the Pareto frontier takes of the form of a continuous curve. The following assumption stipulates these conditions:

Assumption 3.1 (Well-behaved policy class)**.** *Let* Π *be a policy class, and let* $\mathsf{P}_{\max} := \sup_{\pi\in\Pi}\mathcal{U}_{\mathsf{P}}(\pi)$*,* $\mathsf{W}_{\max} := \sup_{\pi}\mathcal{U}_{\mathsf{W}}(\pi)$*, let* $\mathsf{P}_{\min} := \sup_{\pi\in\Pi}\{\mathcal{U}_{\mathsf{P}}(\pi) : \mathcal{U}_{\mathsf{W}}(\pi) = \mathsf{W}_{\max}\}$*. A policy class* Π *is said to be well-behaved if:*

(a) $\{p : (p, w) \in \mathcal{P}(\Pi)$ *for some* $w \in \mathbb{R}\} = [\mathsf{P}_{\min}, \mathsf{P}_{\max}]$

(b) For any $p \in [\mathsf{P}_{\min}, \mathsf{P}_{\max}]$*,* $\operatorname{argmax}_{\pi\in\Pi}\{\mathcal{U}_{\mathsf{W}}(\pi) : \mathcal{U}_{\mathsf{P}}(\pi) = p\}$ *is achieved.*

The following lemma shows that the above assumptions are reasonable, in that we shouldn't expect Pareto optimal policies with $p \notin [\mathsf{P}_{\max}, \mathsf{P}_{\min}]$.

Lemma 3.1. *Suppose that* Π *is* any *policy class such that there exists* π_{W} *attaining* $\mathcal{U}_{\mathsf{W}}(\pi_{\mathsf{W}}) = \mathsf{W}_{\max}$*. Then, for any* $(p, w) \in \mathcal{P}(\Pi)$ *which is a Pareto-optimal pair,* $p \in [\mathsf{P}_{\min}, \mathsf{P}_{\max}]$*.*

Proof. Clearly $p \le \mathsf{P}_{\max}$, since $\mathsf{P}_{\max}$ is the maximal attainable profit. Now, suppose that $(p, w) \in \mathcal{P}(\Pi)$ is a Pareto optimal pair, and assume for the sake of contradiction that $p < \mathsf{P}_{\min}$. Since there exists a $\mathcal{U}_{\mathsf{W}}(\pi_{\mathsf{W}}) = \mathsf{W}_{\max}$, there exists, for any $\epsilon > 0$, some policy π_ϵ such that $\mathcal{U}_{\mathsf{W}}(\pi_\epsilon) = \mathcal{U}_{\mathsf{W}}(\pi_{\mathsf{W}})$, and $\mathcal{U}_{\mathsf{P}}(\pi_\epsilon) \ge \mathsf{P}_{\min} - \epsilon$. By making ϵ sufficiently small, we can ensure that $\mathcal{U}_{\mathsf{P}}(\pi_\epsilon) > p$. On the other hand, $\mathcal{U}_{\mathsf{W}}(\pi_\epsilon) = \mathcal{U}_{\mathsf{W}}(\pi_{\mathsf{W}}) = \mathsf{W}_{\max}$, so that in particular $\mathcal{U}_{\mathsf{W}}(\pi_\epsilon) \ge w$. Hence, π_ϵ dominates the policy with utilities (p, w) in a Pareto sense, so that $(p, w) \notin \mathcal{P}(\Pi)$. □

We now establish the existence of Pareto curves for for well-behaved policy classes.

Lemma 3.2 (Properties of the Pareto curve)**.** *For well-behaved function classes (Assumption 3.1), there exists a unique, non-increasing function* $\mathbf{g}_\Pi$ *such that*

$$\mathcal{P}(\Pi) = \{(p, \mathbf{g}_\Pi(p)) : p \in [\mathsf{P}_{\min}, \mathsf{P}_{\max}]\}\,, \tag{3.11}$$

where $\mathbf{g}_\Pi(p) := \sup_{\pi\in\Pi}\{\mathcal{U}_{\mathsf{W}}(\pi) : \mathcal{U}_{\mathsf{P}}(\pi) = p\}$*, and where we recall* $\mathsf{P}_{\min}, \mathsf{P}_{\min}$ *from Assumption 3.1. If in addition* Π *is* convex*, then* $\mathbf{g}_\Pi(p)$ *is concave.*

Proof of Lemma 3.2. Suppose Assumption 3.1 holds.

First, we show $\mathcal{P}(\Pi) \subseteq \{(p, \mathbf{g}_\Pi(p)) : p \in [\mathsf{P}_{\min}, \mathsf{P}_{\max}]\}$. Given $(p, w) \in \mathcal{P}(\Pi)$ corresponding to a policy π, we must have that $p \in [\mathsf{P}_{\min}, \mathsf{P}_{\max}]$ by Lemma 3.1. Moreover, by Pareto optimality, $w = \mathbf{g}_\Pi(p) := \sup_{\pi\in\Pi}\{\mathcal{U}_{\mathsf{W}}(\pi) : \mathcal{U}_{\mathsf{P}}(\pi) = p\}$ since this optimal is attained by Assumption 3.1.

For the reverse inclusion, we know that if $p \in [\mathsf{P}_{\min}, \mathsf{P}_{\max}]$, then by Assumption 3.1(a), there exists a policy π such that $\mathcal{U}_{\mathsf{P}}(\pi) = p$. Then, by Assumption 3.1(b), there exists a policy π which maximizes $\mathcal{U}_{\mathsf{W}}(\pi) : \mathcal{U}_{\mathsf{P}}(\pi) = p$. By definition, $\mathcal{U}_{\mathsf{W}}(\pi) = \mathbf{g}_\Pi(p)$, and π is Pareto optimal by definition. Hence $(p, w) \in \mathcal{P}(\Pi)$.

We now show that that convexity of Π implies concavity of $\mathbf{g}_\Pi$. It suffices to show that, for any points $(p_1, w_1), (p_2, w_2) \in \mathcal{P}(\Pi)$, and any $\lambda \in [0,1]$, $\lambda w_1 + (1-\lambda)w_2 \leq \mathbf{g}_\Pi(\lambda p_1 + (1-\lambda)p_2)$. Indeed, by definition of the Pareto curve, there exist policies π_1 and π_2 such that $(\mathcal{U}_\mathsf{P}(\pi_i), \mathcal{U}_\mathsf{W}(\pi_i)) = (p_i, w_i)$ for $i \in \{1,2\}$. By convexity of Π, the policy $\pi := \lambda\pi_1 + (1-\lambda)\pi_2 \in \Pi$. Moreover, $\mathcal{U}_\mathsf{P}(\pi) = \lambda p_1 + (1-\lambda)p_2$ and $\mathcal{U}_\mathsf{W}(\pi) = \lambda w_1 + (1-\lambda)w_2$. Finally, by definition of $\mathbf{g}_\Pi$, $\mathcal{U}_\mathsf{W}(\pi) \leq \mathbf{g}_\Pi(\mathcal{U}_\mathsf{P}(\pi))$, which concludes the proof. □

Lemma 3.2 confirms that the Pareto curve, as we might intuitively imagine it, actually exists. With this in hand, we now show that given the Pareto-optimal policies with *exact* scores, the parameterizing function $\mathbf{g}_\Pi$ is concave. Pictorially, $\mathbf{g}_\Pi(p)$ is the Pareto frontier interpreted as a function of allowable profit p which returns the maximum amount of welfare w that can be achieved at this profit level (e.g. the black curve in Figure 3.1, left).

Theorem 3.3 (Pareto frontier under exact knowledge)**.** *Suppose that the unconstrained policies are a well-behaved class. Consider the setting where the welfare and profit are specified exactly by scores w and p. Then, given any population distribution over p, w, the Pareto optimal policies $\pi^\star_\eta$ are given by Eq.* (3.2) *and the Pareto frontier $\mathcal{P}(\Pi_\star)$ is given by*

$$\mathcal{P}_{\text{exact}} := \{(\mathcal{U}_\mathsf{P}(\pi^\star_\eta), \mathcal{U}_\mathsf{W}(\pi^\star_\eta)) : \eta \in [0,1]\}.$$

Moreover, the associated function $\mathbf{g}_{\text{exact}}(p)$ is non-increasing and concave.

Proof of Theorem 3.3. The first statement follows from Proposition 3.6, and that $\mathbf{g}_{\mathcal{P}_{\text{exact}}}(p)$ is non-increasing follows from Lemma 3.2.

Concavity of $\mathbf{g}_{\mathcal{P}_{\text{exact}}}(p)$ follows from Lemma 3.2, and the fact that $\Pi_\star$ is convex. □

3.B Proof of Proposition 3.2

Proof of Proposition 3.2. By definition of $\pi^\star_\eta$,

$$\mathcal{U}_\eta(\pi^\star_\eta) - \mathcal{U}_\eta(\pi^{\text{plug}}_\eta) = \mathbb{E}[|\eta w + (1-\eta)p| \cdot \mathbb{I}(\pi^\star_\eta \neq \pi^{\text{plug}}_\eta)] \ . \tag{3.12}$$

Now consider the event $\pi^\star_\eta \neq \pi^{\text{plug}}_\eta$. This happens only when the predicted scores incur an opposite classification by the η threshold policy, that is, $(\eta w_i + (1-\eta)p_i) \cdot (\eta \hat{w}_i + (1-\eta)\hat{p}_i) < 0$. Define the quantities

$$\begin{aligned} y_i &:= \eta w_i + (1-\eta)p_i \\ z_i &:= \eta(\hat{w}_i - w_i) + (1-\eta)(\hat{p}_i - p_i) \ , \end{aligned}$$

so that

$$\mathcal{U}_\eta(\pi^\star_\eta) - \mathcal{U}_\eta(\pi^{\text{plug}}_\eta) = \mathbb{E}[|y| \cdot \mathbb{I}(y(y+z) < 0)] \ ,$$

where $y \sim \mathcal{N}(0, \eta^2\sigma_w^2 + (1-\eta)^2\sigma_p^2 + 2\rho\eta(1-\eta)\sigma_w\sigma_p)$ and z is sub-Gaussian with squared parameter $\tilde{\sigma}^2 = 4(\eta^2\sigma_{\varepsilon_w}^2 + (1-\eta)^2\sigma_{\varepsilon_p}^2)$ [287].[8] By assumption, the errors are independent of the scores, so that

$$\begin{aligned}\mathbb{E}[|y| \cdot \mathbb{I}(y(y+z) < 0)] &= \mathbb{E}\left[|y| \cdot \mathbb{E}\left[\mathbb{I}(y(y+z) < 0)|y\right]\right] \\ &= \mathbb{E}\left[|y| \cdot \mathcal{P}\left(y(y+z) < 0\right)\right] \ .\end{aligned}$$

Now by sub-Gaussianity of z, we bound $\mathcal{P}\left(y(y+z) < 0\right)$ for any fixed y:

$$\begin{aligned}\mathcal{P}\left(y(y+z) < 0\right) &\leq \begin{cases}\mathcal{P}\left(z < -y\right) & y > 0 \\ \mathcal{P}\left(z > -y\right) & y < 0\end{cases} \\ &\leq e^{-\frac{y^2}{2\tilde{\sigma}^2}} \ .\end{aligned}$$

By symmetry of the distribution of y, the expectation can be bounded as

$$\begin{aligned}\mathbb{E}[|y| \cdot \mathbb{I}(y(y+z) < 0)] &\leq 2\int_0^\infty y(e^{-y^2/(2\tilde{\sigma}^2)}) \cdot \frac{1}{\sigma_y\sqrt{2\pi}} e^{-y^2/(2\sigma_y^2)} dy \\ &= \frac{\sqrt{2}}{\sigma_y\sqrt{\pi}} \int_0^\infty y e^{-\frac{y^2}{2}\left(\frac{1}{\tilde{\sigma}^2} + \frac{1}{\sigma_y^2}\right)} dy \\ &= \frac{1}{\sigma_y\left(\frac{1}{\tilde{\sigma}^2} + \frac{1}{\sigma_y^2}\right)^{1/2}} \cdot \frac{\sqrt{2}\left(\frac{1}{\tilde{\sigma}^2} + \frac{1}{\sigma_y^2}\right)^{1/2}}{\sqrt{\pi}} \int_0^\infty y e^{-\frac{y^2}{2}\left(\frac{1}{\tilde{\sigma}^2} + \frac{1}{\sigma_y^2}\right)} dy \ .\end{aligned}$$

This is a scaled mean of a half-normal distribution with scale parameter $\left(\frac{1}{\tilde{\sigma}^2} + \frac{1}{\sigma_y^2}\right)^{-1/2}$. Thus, difference in η-utility is bounded as

$$\begin{aligned}\mathcal{U}_\eta(\pi_\eta^\star) - \mathcal{U}_\eta(\pi_\eta^{\text{plug}}) &\leq \frac{\sqrt{2}}{\sigma_y\sqrt{\pi}} \left(\frac{1}{\tilde{\sigma}^2} + \frac{1}{\sigma_y^2}\right)^{-1} \\ &= \frac{\sigma_y \cdot \sqrt{2}}{\sqrt{\pi}} \left(\frac{\tilde{\sigma}^2}{\tilde{\sigma}^2 + \sigma_y^2}\right).\end{aligned}$$

The expected utility of the optimal classifier is $\mathbb{E}[y\mathbb{I}(y \geq 0) \geq 0]$ for $y \sim \mathcal{N}(0, \sigma_y^2)$ as defined above. As one half the expectation of a half-normal distribution with scale parameter σ_y,

$$\mathbb{E}[U_\eta(\pi^\star)] = \frac{\sigma_y}{\sqrt{2\pi}} = \frac{1}{\sqrt{2\pi}}\sqrt{\eta^2\sigma_w^2 + (1-\eta)^2\sigma_p^2 + 2\rho\eta(1-\eta)\sigma_w\sigma_p} \ ,$$

[8] When ϵ_w and ϵ_p are assumed to be independent, $\tilde{\sigma}^2 = (\eta^2\sigma_{\varepsilon_w}^2 + (1-\eta)^2\sigma_{\varepsilon_p}^2)$ [287].

which is non-decreasing in ρ for $\eta \in [0,1]$, and increasing in ρ for $\eta \in (0,1)$. Then, the expected Pareto utility of the plug in policy can be lower bounded as

$$\begin{aligned}\mathbb{E}[\mathcal{U}_\eta(\pi_\eta^{\text{plug}})] &= \mathbb{E}[\mathcal{U}_\eta(\pi_\eta^\star) - \big(\mathcal{U}_\eta(\pi_\eta^\star) - \mathcal{U}_\eta(\pi_\eta^{\text{plug}})\big)] \\ &\geq \frac{\sigma_y}{\sqrt{2\pi}}\left(1 - c\frac{\tilde{\sigma}^2}{\tilde{\sigma}^2 + \sigma_y}\right) \\ &= \mathbb{E}[\mathcal{U}_\eta(\pi_\eta^\star)]\left(1 - c\frac{\tilde{\sigma}^2}{\tilde{\sigma}^2 + \sigma_y^2}\right),\end{aligned}$$

where $c = 2$. Since σ_y is increasing in ρ for $\eta \in (0,1)$, the lower bound on $\mathbb{E}[\mathcal{U}_\eta(\pi_\eta^{\text{plug}})]$ is increasing as well. □

3.C Proof of Proposition 3.4

Before presenting the supporting lemmas that will lead into the proof of of Proposition 3.4, we define important quantities. Recall from Chapter 2 that demographic parity constrains the selection rates of policies. Define rate function for each group as

$$r_g(\pi) = \mathbb{E}[\pi_g(p) \mid \text{in group } g] \,.$$

Because we focus on threshold policies, we will equivalently write $r_g(t) := r_g(\mathbb{I}(p \geq t))$. This function is monotonic in the threshold t, and therefore its inverse maps acceptance rates to thresholds which achieve that rate, i.e. $r_g^{-1}(\beta) = t_g$. Define $f_{\mathsf{A}}(\beta)$ and $f_{\mathsf{B}}(\beta)$ to be components of the objective function in (3.8) due to each group, that is,

$$f_g(\beta) = \mathbb{P}\{\text{in } g\}\mathbb{E}[p \cdot \mathbb{I}\{p > r_g^{-1}(\beta)\} \mid \text{in } g] \,.$$

By Proposition 5.3 in [179], the functions f_g are concave. Therefore, the combined objective function is concave in each argument,

$$\mathcal{U}_{\mathsf{P}}(\beta_{\mathsf{A}}, \beta_{\mathsf{B}}) := \sum_{g \in \{\mathsf{A},\mathsf{B}\}} \mathbb{P}\{\text{in } g\}\mathbb{E}[p \cdot \mathbb{I}\{p > r_g^{-1}(\beta_g)\} \mid \text{in } g] = \sum_{g \in \{\mathsf{A},\mathsf{B}\}} f_g(\beta_g) \,. \tag{3.13}$$

We restrict our attention to the case that p has continuous support. In this case, the functions f_g are differentiable.

Lemma 3.3. *If the distribution of exact profit scores for groups* A *and* B *are such that that maximum profit selection rates* $\beta_{\mathsf{A}}^{\texttt{MaxUtil}} = r_{\mathsf{A}}(0)$ *and* $\beta_{\mathsf{B}}^{\texttt{MaxUtil}} = r_{\mathsf{B}}(0)$ *and* $r_{\mathsf{A}}(0) \leq r_{\mathsf{A}}(0)$*, then for any* $\epsilon \geq 0$*, the selection rates maximizing the optimization problem* (3.8) *under demographic parity satisfy the following:*

$$\beta_{\mathsf{A}}^{\texttt{MaxUtil}} \leq \beta_{\mathsf{A}}^{\epsilon} \leq \beta_{\mathsf{B}}^{\epsilon} \leq \beta_{\mathsf{B}}^{\texttt{MaxUtil}} \,.$$

Proof of Lemma 3.3. First, note we must have that $\beta_{\mathsf{A}}^{\epsilon} \leq \beta_{\mathsf{B}}^{\texttt{MaxUtil}}$. If it were that $\beta_{\mathsf{A}}^{\epsilon} > \beta_{\mathsf{B}}^{\texttt{MaxUtil}}$, then the alternate solution $\beta_{\mathsf{A}} = \beta_{\mathsf{B}}^{\texttt{MaxUtil}}$ and $\beta_{\mathsf{B}} = \beta_{\mathsf{B}}^{\texttt{MaxUtil}}$ would be feasible for (3.8) and achieve a higher objective value by the concavity of Eq. (3.13).

Then we show that $\beta_{\mathsf{B}}^{\epsilon} \leq \beta_{\mathsf{B}}^{\texttt{MaxUtil}}$. Assume for the sake of contradiction that $\beta_{\mathsf{B}}^{\epsilon} > \beta_{\mathsf{B}}^{\texttt{MaxUtil}}$. Then since $\beta_{\mathsf{A}}^{\epsilon} \leq \beta_{\mathsf{B}}^{\texttt{MaxUtil}}$, setting $\beta_{\mathsf{B}} = \beta_{\mathsf{B}}^{\texttt{MaxUtil}}$ achieves higher objective value without increasing $|\beta_{\mathsf{B}} - \beta_{\mathsf{A}}^{\epsilon}|$, and thus would be feasible for (3.8). A similar argument shows that $\beta_{\mathsf{A}}^{\epsilon} \geq \beta_{\mathsf{A}}^{\texttt{MaxUtil}}$.

Then, we show that for any optimal selection rates, $\beta_{\mathsf{A}}^{\epsilon} \leq \beta_{\mathsf{B}}^{\epsilon}$ for all $\epsilon \geq 0$. Suppose for the sake of contradiction that $\beta_{\mathsf{A}}^{\epsilon} > \beta_{\mathsf{B}}^{\epsilon}$. In this case, we can equivalently write that

$$\beta_{\mathsf{B}}^{\texttt{MaxUtil}} - \beta_{\mathsf{B}}^{\epsilon} > \beta_{\mathsf{B}}^{\texttt{MaxUtil}} - \beta_{\mathsf{A}}^{\epsilon} \;\text{ and/or }\; \beta_{\mathsf{A}}^{\epsilon} - \beta_{\mathsf{A}}^{\texttt{MaxUtil}} > \beta_{\mathsf{B}}^{\epsilon} - \beta_{\mathsf{A}}^{\texttt{MaxUtil}} .$$

In either case, setting $\beta_{\mathsf{A}}^{\epsilon} = \beta_{\mathsf{B}}^{\epsilon}$ would be a feasible solution which would achieve a higher objective function value, by the concavity of Eq. (3.13). This contradicts the assumption that $\beta_{\mathsf{A}}^{\epsilon} > \beta_{\mathsf{B}}^{\epsilon}$, and thus it must be that $\beta_{\mathsf{A}}^{\epsilon} \leq \beta_{\mathsf{B}}^{\epsilon}$. □

Lemma 3.4. *Under the conditions of Lemma 3.3, the maximizer $(\beta_{\mathsf{A}}^{\epsilon}, \beta_{\mathsf{B}}^{\epsilon})$ of the ϵ-demographic parity constrained problem in* (3.8) *is either satisfied with the maximum profit selection rates $(\beta_{\mathsf{A}}^{\texttt{MaxUtil}}, \beta_{\mathsf{B}}^{\texttt{MaxUtil}})$, or $\beta_{\mathsf{B}}^{\epsilon} - \beta_{\mathsf{A}}^{\epsilon} = \epsilon$ (or the two conditions coincide).*

Proof of Lemma 3.4. If it were that $|\beta_{\mathsf{A}}^{\epsilon} - \beta_{\mathsf{B}}^{\epsilon}| = \gamma < \epsilon$ then we could construct an alternative solution using the remaining $\epsilon - \gamma$ slack in the constraint which would achieve a higher objective function value, since the functions f_g are concave. Furthermore, by Lemma 3.3, we have that $|\beta_{\mathsf{A}}^{\epsilon} - \beta_{\mathsf{B}}^{\epsilon}| = \beta_{\mathsf{B}}^{\epsilon} - \beta_{\mathsf{A}}^{\epsilon}$. □

This result implies that the complexity of the maximization in (3.8) can be reduced to a single variable search:

$$\beta^{\star} = \operatorname*{argmax}_{\beta} f_{\mathsf{A}}(\beta) + f_{\mathsf{B}}(\beta + \epsilon), \quad \pi_{\text{fair}}^{\epsilon} = (\mathbb{I}\{p \geq r_g^{-1}(\beta^{\star})\}, \mathbb{I}(p \geq r_g^{-1}(\beta^{\star} + \epsilon)\}) \tag{3.14}$$

This expression holds when $|\beta_{\mathsf{A}}^{\texttt{MaxUtil}} - \beta_{\mathsf{B}}^{\texttt{MaxUtil}}| > \epsilon$, and otherwise the solution is given by $(\beta_{\mathsf{A}}^{\texttt{MaxUtil}}, \beta_{\mathsf{B}}^{\texttt{MaxUtil}})$.

Lemma 3.5. *Under the conditions of Lemma 3.3, as $\epsilon \geq 0$ decreases, the group-dependent selection rates $\beta_{\mathsf{A}}^{\epsilon}$ and $\beta_{\mathsf{B}}^{\epsilon}$ become closer to the profit maximizing selection rates for each group. That is, the functions $|\beta_{\mathsf{A}}^{\epsilon} - \beta_{\mathsf{A}}^{\texttt{MaxUtil}}|$ and $|\beta_{\mathsf{B}}^{\epsilon} - \beta_{\mathsf{B}}^{\texttt{MaxUtil}}|$ are both increasing in ϵ.*

Proof of Lemma 3.5. We show that for any $\epsilon' \geq \epsilon \geq 0$, it must be that $|\beta_g^{\epsilon} - \beta_g^{\texttt{MaxUtil}}| \leq |\beta_g^{\epsilon'} - \beta_g^{\texttt{MaxUtil}}|$. First, we remark that if $|\beta_{\mathsf{A}}^{\texttt{MaxUtil}} - \beta_{\mathsf{B}}^{\texttt{MaxUtil}}| \leq \epsilon$ or if $\epsilon \leq |\beta_{\mathsf{A}}^{\texttt{MaxUtil}} - \beta_{\mathsf{B}}^{\texttt{MaxUtil}}| \leq \epsilon'$, the claim holds by application of Lemma 3.3.

Otherwise, let the ϵ-demographic parity constrained solution be optimized by $(\beta, \beta + \epsilon)$ and the ϵ'-demographic parity constrained solution be optimized by $(\beta', \beta' + \epsilon')$. This is valid

by Lemma 3.4. Equivalently, $\beta \in \text{argmax}\{f_{\mathsf{A}}(\beta) + f_{\mathsf{B}}(\beta+\epsilon)\}$ and $\beta' \in \text{argmax}\{f_{\mathsf{A}}(\beta') + f_{\mathsf{B}}(\beta'+\epsilon')\}$. Since f_{A} and f_{B} are concave and differentiable,

$$f'_{\mathsf{A}}(\beta) + f'_{\mathsf{B}}(\beta+\epsilon) = 0 \quad \text{and} \quad f'_{\mathsf{A}}(\beta') + f'_{\mathsf{B}}(\beta'+\epsilon') = 0 \ .$$

Assume for sake of contradiction that $\beta < \beta'$ and recall that by Lemma 3.3 we further have $\beta' > \beta \geq \beta_{\mathsf{A}}^{\texttt{MaxUtil}}$, so by the concavity of f_{A},

$$f_{\mathsf{A}}(\beta) \geq f_{\mathsf{A}}(\beta') \quad \text{and} \quad f'_{\mathsf{A}}(\beta') \leq f'_{\mathsf{A}}(\beta) \ .$$

Analogously, we must have that $\beta_{\mathsf{B}}^{\texttt{MaxUtil}} \geq \beta' + \epsilon' > \beta + \epsilon$, so that

$$f_{\mathsf{B}}(\beta'+\epsilon') \geq f_{\mathsf{B}}(\beta+\epsilon) \quad \text{and} \quad f'_{\mathsf{B}}(\beta'+\epsilon') \geq f'_{\mathsf{B}}(\beta+\epsilon) \ .$$

Using the equations above, we have that

$$f'_{\mathsf{B}}(\beta+\epsilon) = -f'_{\mathsf{A}}(\beta) \leq -f'_{\mathsf{A}}(\beta') = f'_{\mathsf{B}}(\beta'+\epsilon') \ .$$

Since f_{B} is concave and thus its derivative is decreasing, this statement implies that $\beta+\epsilon \geq \beta'+\epsilon'$, which is a contradiction. Thus, it must be that $\beta \geq \beta'$, i.e. $\beta_{\mathsf{A}}^{\epsilon} \geq \beta_{\mathsf{A}}^{\epsilon'}$. With an analogous proof by contradiction, one can show that $\beta_{\mathsf{B}}^{\epsilon'} \geq \beta_{\mathsf{B}}^{\epsilon}$.

Combining these two inequalities in Lemma 3.3 completes the proof of Lemma 3.5. □

Proof of Proposition 3.4. The proof makes use of Lemma 3.3 and Lemma 3.5.

First, we show that $t_{\mathsf{A}}^{\epsilon} \leq 0$ for all $\epsilon \geq 0$. This is a consequence of Lemma 3.3, which shows that $\beta_{\mathsf{A}}^{\epsilon} \geq \beta_{\mathsf{A}}^{\texttt{MaxUtil}}$. Since r_{A} is a decreasing function (and thus, r_{A}^{-1} is also a decreasing function), this implies that

$$t_{\mathsf{A}}^{\epsilon} = r_{\mathsf{A}}^{-1}(\beta_{\mathsf{A}}^{\epsilon}) \leq r_{\mathsf{A}}^{-1}(\beta_{\mathsf{A}}^{\texttt{MaxUtil}}) = 0 \ .$$

A similar argument holds to show that $t_{\mathsf{B}}^{\epsilon} \geq 0$ for all $\epsilon \geq 0$.

Now we show that $t_{\mathsf{A}}^{\epsilon}$ is increasing in ϵ and $t_{\mathsf{B}}^{\epsilon}$ is decreasing in ϵ to show that both are shrinking toward 0 as ϵ increases. Since $t_g = r_g^{-1}(\beta)$ is decreasing in β, Lemma 3.5 implies that the functions $|t_{\mathsf{A}}^{\epsilon}| = |t_{\mathsf{A}}^{\epsilon} - t_{\mathsf{A}}^{\texttt{MaxUtil}}|$ and $|t_{\mathsf{B}}^{\epsilon}| = |t_{B}^{\epsilon} - t_{\mathsf{B}}^{\texttt{MaxUtil}}|$ are also decreasing in ϵ toward the max profit thresholds of $t_{\mathsf{A}}^{\texttt{MaxUtil}} = t_{\mathsf{B}}^{\texttt{MaxUtil}} = 0$. Since $t_{\mathsf{A}}^{\epsilon} \leq 0$ and $t_{\mathsf{B}}^{\epsilon} \geq 0$ for all $\epsilon \geq 0$, this concludes the proof of Proposition 3.4. □

Part II

Training Data as a Form of Context: Two Views for Supervised Learning

Model building and model testing, however, are necessarily constrained by the properties of the data with which we are dealing: one's philosophy of measurement influences the hypotheses that one can formulate and the ways in which those hypotheses may be examined.

— David J. Hand, "Statistics and the Theory of Measurement," 1996 [122]

Alongside understanding and designing for the intent of learning systems as discussed in Part I, it is crucial to understand how we can reach these diverse objectives when learning statistical models from data. In Part II of this thesis, we ground our study of context in learning systems by focusing on two key elements of structure in training data: *how* data instances are collected across different sub-populations and data sources and *what* information different features of a dataset convey.

In Chapter 4, we incorporate dataset collection as a design step in the learning procedure. This enables us to assess and leverage the importance of different data allocations toward reaching high group and population accuracies. The results expose that in many settings, representation across all diverse groups in training data is aligned with reaching population-level accuracy goals. In service of obtaining accurate predictors, no components of data should be overlooked. In Chapter 5, we study how to make the most of information that is available to us in data, focusing on careful use of auxilliary "side information" like time or location metadata. Such data are natural to many settings, but should be incorporated with care. We propose that post-processing is a natural and effective way to utilize this structure, and show it is robust to different tasks, predictors, and sampling patterns.

Together, Chapters 4 and 5 provide two complementary and encouraging takes on understanding and utilizing data as a form of context for delineating system requirements and improving model performance. These two views illustrate how data can be used to formalize certain notions of context in learning systems. However, data are not the whole of context; in Chapter 6, we highlight key limitations of viewing data as context, emphasizing opportunities for expanding this line of research.

Chapter 4

The Importance of Numerical Representation of Sub-Groups in Training Data

This chapter is based on the paper "Representation matters: Assessing the importance of subgroup allocations in training data" [247], written in collaboration with Theodora Worledge, Benjamin Recht, and Michael I. Jordan.

Collecting more diverse and representative training data is often touted as a remedy for the disparate performance of machine learning predictors across subpopulations. However, a precise framework for understanding how dataset properties like diversity affect learning outcomes is largely lacking. By casting data collection as part of the learning process, we demonstrate that diverse representation in training data is key not only to increasing subgroup performances, but also to achieving population-level objectives. Analysis and experiments presented in this chapter describe how dataset compositions influence performance and provide constructive results for using trends in existing data, alongside domain knowledge, to help guide intentional, objective-aware dataset design.

4.1 Background

Datasets play a critical role in shaping the perception of performance and progress in machine learning —the way we collect, process, and analyze data affects the way we benchmark success and form new research agendas [93, 220]. A growing appreciation of this determinative role of datasets has sparked a concomitant concern that standard datasets used for training and evaluating machine learning models lack diversity along significant dimensions, for example, geography, gender, and skin type [49, 258].

Lack of diversity in evaluation data can obfuscate disparate performance when evaluating based on aggregate accuracy [49]. Lack of diversity in training data can limit the extent

to which learned models can adequately apply to all portions of a population, a concern highlighted in recent work in the medical domain [119, 134].

This chapter aims to develop a general unifying perspective on the way that dataset composition affects outcomes of machine learning systems. We focus on *dataset allocations*: the number of datapoints from predefined subsets of the population. While we acknowledge that numerical inclusion of groups is an imperfect proxy of representation, we believe that allocations provide a useful initial mathematical abstraction for formulating relationships among diversity, data collection, and statistical risk. We discuss broader implications of our formulation in Section 4.5 and Chapter 6.

With the implicit assumption that the learning task is well specified and performance evaluation from data is meaningful for all groups, we first ask:

1. *Are group allocations in training data pivotal to performance? To what extent can methods that up-weight underrepresented groups help, and when might upweighting actually hurt performance?*

Taking a point of view that data collection is a critical component of the overall machine learning process, we study the effect that dataset composition has on group and population accuracies. This complements work showing that simply gathering more data can mitigate some sources of bias or unfairness in learned outputs [61], a phenomenon which has been observed in practice as well. Indeed, in response to the Gender Shades study [49], which exposed disparate intersectional accuracies of commercially available facial recognition software, companies selectively collected additional data to decrease the exposed inaccuracies of their facial recognition models for certain groups, often raising aggregate accuracy in the process [234]. Given the potential for targeted data collection efforts to repair unintended outcomes of machine learning systems, we next ask:

2. *How might we describe "optimal" dataset allocations for different learning objectives? Does the often-witnessed lack of diversity in large-scale datasets align with maximizing population accuracy?*

We show that purposeful data collection efforts can proactively support intentional objectives of a learning system, and that diversity and population objectives are often aligned. Many datasets have recently been designed or amended to exhibit diversity of the underlying population [250, 275, 297]. Such endeavors are significant undertakings, as data gathering and annotation must consider consent, privacy, and power concerns in addition to inclusivity, transparency and reusability [106, 150, 292]. In light of the importance of more representative and diverse datasets, and the effort required to create them, our final question asks:

3. *When and how can we leverage existing datasets to help inform better allocations, towards achieving diverse objectives in a subsequent dataset collection effort?*

Representation bias, or systematic underrepresentation of meaningful sub-population in a dataset, is one of many causes of unintended consequences of machine learning [269]. It is our intention that the results of this work can provide a resource toward comprehensive and contextual scoping of machine learning pipelines and their limitations so that they may be applied conscientiously and successfully. In this chapter:

1. We analyze the complementary roles of dataset allocations and algorithmic interventions for achieving per-group and total-population performance (Section 4.2). Our experiments show that while algorithmically up-weighting underrepresented groups can help, dataset composition is the most consistent determinant of performance (Section 4.4).

2. We propose a scaling model that describes the impact of dataset allocations on group accuracies (Section 4.3). Under this model, when parameters governing the relative values of within-group data are equal for all groups, the allocation that minimizes *population* risk *overrepresents* minority groups relative to their population proportions.

3. We demonstrate that our proposed scaling model captures major trends of the relationship between dataset allocations and performance (Section 4.4). Further experiments evidence that a small initial sample can be used to inform subsequent data collection efforts to, for example, maximize the minimum accuracy over groups.

Sections 4.2 and 4.3 formalize data collection as part of the learning problem and derive results under illustrative settings. Experiments in Section 4.4 support these results and expose nuances inherent to real-data contexts. Section 4.5 synthesizes results and delineates possible generalizations and extensions of our findings.

Additional Related Work

Targeted data collection in machine learning. Recent research evidences that targeted data collection can be an effective way to reduce disparate performance of machine learning models evaluated across sub-populations [234]. Chen, Johansson, and Sontag [61] present a formal argument that the addition of training data can lessen discriminatory outcomes while improving accuracy of learned models, and adaptively collecting data from the lowest-performing sub-population has been proposed as a method by which to increase the minimum accuracy over groups [4, 259].

At the same time, there are many complexities of gathering data as a solution to disparate performance across groups. Targeted data collection from specific groups can present undue burdens of surveillance or skirt consent [220]. When learning systems fail portions of their intended population due to issues of measurement and construct validity, more thorough data collection is unlikely to solve the issue without further intervention [143].

With these complexities in mind, we study the importance of numerical representation in training datasets in achieving diverse objectives. Optimal allocation of subpopulations

in statistical survey designs dates back to at least Neyman in 1934 [209], including stratified sampling methods to ensure coverage across sub-populations [182]. For more complex prediction systems, the field of optimal experimental design [227] studies what inputs are most valuable for reaching a given objective, often focusing on linear prediction functions. We consider a constrained sampling structure and directly model the impact of group allocations on subgroup performance for general model classes, more similar to the setting of Hashimoto [127].

Valuing data. In economics, allocations indicate a division of goods to various entities [74]. While we focus on the influence of data allocations on model accuracies across groups, there are many approaches to valuing data. Methods centering on a theory of Shapley valuations [107, 298] complement studies of the influence of individual data points on model performance to aid subsampling data [286].

Methods for handling group-imbalanced data. Importance sampling and importance weighting are standard approaches to addressing class imbalance or small groups sizes [48, 120], though the effects of importance weighting for deep learning may vary with regularization [52]. Other methods specifically address differential performance between groups. Maximizing minimum performance across groups can reduce accuracy disparities [251] and promote fairer sequential outcomes [128]. For broader classes of group-aware objectives, techniques exist to mitigate unfairness or disparate performance of black box prediction functions [97, 162]. It might not be clear a priori which subsets need attention; Sohoni et al. [265] propose a method to identify and account for hidden strata, while other methods are defined for any subsets [128, 162].

In addition to modifying the optimization objective or learning algorithm, one can also modify the input data itself to match the desired population by downsampling or by upsampling with data augmentation techniques [60, 142].

Notation

Δ^k denotes the k-dimensional simplex. $\mathbb{Z}^+$ denotes non-negative integers and $\mathbb{R}^+$ non-negative reals.

4.2 Training Set Allocations and Alternatives

We study settings in which each data instance is associated with a group g_i, so that the training set can be expressed as $\mathcal{S} = \{x_i, y_i, g_i\}_{i=1}^n$ where x_i, y_i denote the features and labels of each instance. We index the discrete **groups** by integers $\mathcal{G} = \{1, .., |\mathcal{G}|\}$, or when we specifically consider just two groups, we write $\mathcal{G} = \{A, B\}$. We assume that groups are disjoint and cover the entire population, with $\gamma_g = P_{(X,Y,G)\sim\mathcal{D}}[G = g]$ denoting the **population prevalence** of group g, so that $\vec{\gamma} \in \Delta^{|\mathcal{G}|}$. Groups could represent inclusion in one of many binned demographic categories, discrete sources of data, or simply a general association with latent characteristics that are relevant to prediction.

For a given population with distribution $\mathcal{D}$ over features, labels, and groups, we are interested in the population level risk, $\mathcal{R}(\hat{f}(\mathcal{S}); \mathcal{D}) := \mathbb{E}_{(X,Y,G)\sim\mathcal{D}}[\ell(\hat{f}(X), Y)]$, of a predictor $\hat{f}$ trained on dataset $\mathcal{S}$, as well as group specific risks. Denoting the **group distributions** by $\mathcal{D}_g$, defined as conditional distributions, via

$$P_{(X,Y)\sim\mathcal{D}_g}[X = x, Y = y] = P_{(X,Y,G)\sim\mathcal{D}}[X = x, Y = y, G = g]/\gamma_g,$$

the population risk decomposes as a weighted average over group risks:

$$\mathcal{R}(\hat{f}(\mathcal{S}); \mathcal{D}) = \sum_{g\in\mathcal{G}} \gamma_g \cdot \mathcal{R}(\hat{f}(\mathcal{S}); \mathcal{D}_g) \ . \tag{4.1}$$

In the remainder of this section we will assume that the loss $\ell(\hat{y}, y)$ is a separable function over data instances. While this holds for many common loss functions, some objectives do not decouple in this sense (e.g., group losses and associated classes of fairness-constrained objectives; see [97]). We revisit this point in Sections 4.4 and 4.5.

Training Set Allocations

In light of the decomposition of the population-level risk as a weighted average over group risks in Eq. (4.1), we now consider the composition of fixed-size training sets, in terms of how many samples come from each group.

Definition 4.1 (Allocations)**.** Given a dataset of n triplets, $\{x_i, y_i, g_i\}_{i=1}^n$, the **allocation** $\vec{\alpha} \in \Delta^{|\mathcal{G}|}$ describes the relative proportions of each group in the dataset:

$$\alpha_g := \tfrac{1}{n} \sum_{i=1}^{n} \mathbb{I}[g_i = g], \quad g \in \mathcal{G}. \tag{4.2}$$

It will be illuminating to consider $\vec{\alpha}$ not only as a property of an existing dataset, but as a parameter governing dataset construction, as captured in the following definition.

Definition 4.2 (Sampling from allocation $\vec{\alpha}$)**.** Given the sample size n, group distributions $\{\mathcal{D}_g\}_{g\in\mathcal{G}}$, and allocation $\vec{\alpha} \in \Delta^{|\mathcal{G}|}$, such that $n_g := \alpha_g n \in \mathbb{Z}^+, \forall g \in \mathcal{G}$, to **sample from allocation** $\vec{\alpha}$ is procedurally equivalent to independent sampling of $|\mathcal{G}|$ disjoint datasets $\mathcal{S}_g$ and concatenating:

$$\begin{aligned} \mathcal{S}(\vec{\alpha}, n) &= \bigcup_{g\in\mathcal{G}} \mathcal{S}_g \\ \mathcal{S}_g &= \{x_i, y_i, g\}_{i=1}^{n_g}, \quad (x_i, y_i) \sim_{i.i.d.} \mathcal{D}_g \ . \end{aligned} \tag{4.3}$$

For $\vec{\alpha}$ not satisfying the requirement that $\alpha_g n$ is integral, we could randomize the fractional allocations, or take $n_g = \lfloor \alpha_g n \rfloor$, reducing the total number of samples to $\sum_g \lfloor \alpha_g n \rfloor$. In the following sections we will generally allow allocations with $n_g \notin \mathbb{Z}$, assuming that the effect of up to $|\mathcal{G}|$ fractionally assigned instances is negligible for large n.

The procedure given in Definition 4.2 suggests formalizing data collection as a component of the learning process in the following way: in addition to choosing a loss function and method for minimizing the risk, choose the relative proportions at which to sample the groups in the training set:

$$\vec{\alpha}^* = \underset{\vec{\alpha}\in\Delta^{|\mathcal{G}|}}{\operatorname{argmin}} \min_{\hat{f}\in\mathcal{F}} \mathcal{R}\left(\hat{f}\left(\mathcal{S}(\vec{\alpha}, n)\right); \mathcal{D}\right).$$

In Section 4.3, we show that when a dataset curator can design dataset allocations in the sense of Definition 4.2, they have the opportunity to improve accuracy of the trained model. Of course, one does not always have the opportunity to collect new data or modify the composition of an existing dataset. We next consider methods for using fixed datasets that have groups with small training set allocation α_g, relative to γ_g, or high risk for some groups relative to the population.

Accounting for Small Group Allocations in Fixed Datasets

In classical **empirical risk minimization** (ERM), one learns a function from class $\mathcal{F}$ that minimizes average prediction loss over the training instances $(x_i, y_i, g_i) \in \mathcal{S}$ (we also abuse notation and write $i \in \mathcal{S}$) with optional regularization R:

$$\hat{f}(\mathcal{S}) = \underset{f\in\mathcal{F}}{\operatorname{argmin}} \sum_{i\in\mathcal{S}} \ell(f(x_i), y_i) + R(f, \mathcal{S}).$$

There are many methods for addressing small group allocations in data (see Section 4.1). Of particular relevance to our work are objective functions that minimize group or population risks. In particular, one approach is to use **importance weighting** (IW) to re-weight training samples with respect to a target distribution defined by $\vec{\gamma}$:

$$\hat{f}^{\text{IW}}(\mathcal{S}) = \underset{f\in\mathcal{F}}{\operatorname{argmin}} \sum_{g\in\mathcal{G}} \frac{\gamma_g}{\alpha_g} \left(\sum_{i\in\mathcal{S}_g} \ell(f(x_i), y_i) \right) + R(f, \mathcal{S}).$$

This empirical risk with instances weighted by $\gamma_g/\alpha_g = \gamma_g n/n_g$ is an unbiased estimate of the population risk, up to regularization. While unbiasedness is often desirable, importance weighting can induce high variance of the estimator when γ_g/α_g is large for some group [76], which happens when group g is severely underrepresented in the training data relative to their population prevalence.

Alternatively, **group distributionally robust optimization** (GDRO) [138, 251] minimizes the maximum empirical risk over all groups:

$$\hat{f}^{\text{GDRO}}(\mathcal{S}) = \underset{f\in\mathcal{F}}{\operatorname{argmin}} \max_{g\in\mathcal{G}} \left(\tfrac{1}{n_g} \sum_{i\in\mathcal{S}_g} \ell(f(x_i), y_i) + R(f, \mathcal{S}_g) \right) .$$

For losses ℓ which are continuous and convex in the parameters of f, the optimal GDRO solution corresponds to the minimizer of a group-weighted objective: $\frac{1}{n}\sum_{i=1}^{n} w(g_i)\cdot\ell(f(x_i),y_i)$, though this is not in general true for nonconvex losses (see Proposition 1 of [251] and the remark immediately thereafter).

Given the correspondence of GDRO (for convex loss functions) and IW to the optimization of group-weighted ERM objectives, we now investigate the joint roles of sample allocation and group re-weighting for estimating group-weighted risks. For prediction function f, loss function ℓ, and group weights $w : \mathcal{G} \to \mathbb{R}^+$, let $\hat{L}(w,\alpha,n;f,\ell)$ be the random variable defined by:

$$\hat{L}(w,\alpha,n;f,\ell) := \tfrac{1}{n}\sum_{i\in\mathcal{S}(\vec{\alpha},n)} w(g_i)\cdot\ell(f(x_i),y_i)\ ,$$

where the randomness in $\hat{L}$ comes from the draws of x_i, y_i from $\mathcal{D}_{g_i}$ according to procedure $\mathcal{S}(\vec{\alpha},n)$ (Definition 4.2), as well as any randomness in f.

The following proposition shows that group weights and allocations play complementary roles in risk function estimation. In particular, if $w(g)$ depends on the sampling allocations α_g, then there are alternative group weights w^* and allocation $\vec{\alpha}^*$ such that the alternative estimator has the same expected value but lower variance.

Proposition 4.1 (Weights and allocations)**.** *For any loss ℓ, prediction function f and group distributions $\mathcal{D}_g$, there exist weights with $w^*(g) \propto \left(\mathrm{Var}_{(x,y)\sim\mathcal{D}_g}[\ell(f(x),y)]\right)^{-1/2}$ such that for any triplet $(\vec{\alpha},w,n)$ with $\sum_g \alpha_g w(g) > 0$, if $w \not\propto w^*$,[1] there exists an alternative allocation $\vec{\alpha}^*$ such that*

$$\mathbb{E}[\hat{L}(w^*,\vec{\alpha}^*,n;f,\ell)] = \mathbb{E}[\hat{L}(w,\vec{\alpha},n;f,\ell)]$$
$$\mathrm{Var}[\hat{L}(w^*,\vec{\alpha}^*,n;f,\ell)] < \mathrm{Var}[\hat{L}(w,\vec{\alpha},n;f,\ell)]\ .$$

If $w(g) > w^*(g)$, $\alpha_g^* > \alpha_g$ and if $w(g) < w^*(g)$, $\alpha_g^* < \alpha_g$.

Proof of Proposition 4.1. For any total training set size n, any $(w,\vec{\alpha})$ pair induces a vector $\vec{\gamma}'$ with entries $\gamma'_g(w,\vec{\alpha}) := \frac{w(g)\alpha_g}{\sum_{g\in\mathcal{G}} w(g)\alpha_g}$, where

$$\begin{aligned}\mathbb{E}[\hat{L}(w,\vec{\alpha},n;f,\ell)] &= \frac{1}{n}\sum_{(x_i,y_i,g_i)\in\mathcal{S}(\vec{\alpha},n)} w(g_i)\mathbb{E}[\ell(f(x_i),y_i)]\\ &= \sum_{g\in\mathcal{G}}\frac{n_g}{n}w(g)\mathbb{E}_{(x,y)\sim\mathcal{D}_g}[\ell(f(x),y)]\\ &= c\cdot\mathbb{E}_{g\sim\text{Multinomial}(\vec{\gamma}')}\left[\mathbb{E}_{(x,y)\sim\mathcal{D}_g}\left[\ell(f(x),y)\right]\right]\end{aligned}$$

[1] We use the symbol $\not\propto$ to denote "not approximately proportional to." The approximately part of this relation stems from finite and integer sample concerns; for example, the proposition holds if we consider $w \not\propto w^*$ to mean $\exists g\in\mathcal{G} : |1-\frac{w(g)}{w^*(g)}| > \frac{|\mathcal{G}|}{\alpha_g n}$.

for constant $c = \sum_g \alpha_g w(g)$. The vector $\vec{\gamma}'$ in this sense describes an implicit "target distribution" induced by applying weights w after sampling with allocation $\vec{\alpha}$. Note that unless $w_g = 0$ for all g with $\alpha_g > 0$, $\vec{\gamma}'$ has at least one nonzero entry. The constant c re-scales the weighted objective function with original weights w so as to match the expected loss with respect to the group proportions $\vec{\gamma}'$. Stated another way, for any alternative allocation $\vec{\alpha}'$, we could pick weights $w'(g) = c\gamma'_g/\alpha'_g$ (letting $w'(g) = 0$ if $\alpha'_g = 0$), and satisfy

$$\mathbb{E}[\hat{L}(w', \vec{\alpha}', n; f, \ell)] = \mathbb{E}[\hat{L}(w, \vec{\alpha}, n; f, \ell)] \ .$$

Given this correspondence, we now find the pair $(\vec{\alpha}^*, w^*)$ which minimizes $\mathrm{Var}[\hat{L}(cw', \vec{\alpha}', n; f, \ell)]$, subject to $w'(g)\alpha'_g = c\gamma'_g$. Since the original pair $(\vec{\alpha}, w)$ satisfies this constraint (by construction), we must have

$$\min_{\vec{\alpha}', w' : w'(g)\alpha'_g = c\gamma'_g} \mathrm{Var}[\hat{L}(w', \vec{\alpha}', n; f, \ell)] \leq \mathrm{Var}[\hat{L}(w, \vec{\alpha}, n; f, \ell)] \ .$$

We first compute $\mathrm{Var}[\hat{L}(w', \vec{\alpha}', n; f, \ell)]$. By Definition 4.2, samples (x_i, y_i) are assumed to be independent draws from distributions $\mathcal{D}_{g_i}$, so that the variance of the estimator can be written as (for convenience we assume here that $n\alpha'_g \in \mathbb{Z}$, see the discussion below):

$$\begin{aligned}
\mathrm{Var}\left[\hat{L}(w', \vec{\alpha}', n; f, \ell)\right] &= \frac{1}{n^2} \sum_{g \in \mathcal{G}} w'(g)^2 \sum_{i=1}^{n\alpha'_g} \mathbb{E}_{(x_i, y_i) \sim \mathcal{D}_g} \left[\left(\ell(f(x_i), y_i) - \mathbb{E}_{(x,y)\sim\mathcal{D}_g}[\ell(f(x), y)] \right)^2 \right] \\
&= \frac{1}{n} \sum_{g \in \mathcal{G}} \alpha'_g w'(g)^2 \mathrm{Var}\left[\ell_g^{(i)}\right] \ ,
\end{aligned}$$

where $\mathrm{Var}\left[\ell_g^{(i)}\right]$ denotes $\mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_g} \left[\left(\ell(f(x_i), y_i) - \mathbb{E}_{(x,y)\sim\mathcal{D}_g}[\ell(f(x), y)] \right)^2 \right]$. Now, to respect the constraint $w'(g)\alpha'_g = c\gamma'_g$ means that for any g with $\gamma'_g > 0$, $w'(g)$ is a deterministic function of α'_g, since c and $\vec{\gamma}'$ are determined by the initial pair $(\vec{\alpha}, w)$. Then it is sufficient to compute

$$\operatorname*{argmin}_{\alpha' \in \Delta^{|\mathcal{G}|}} \frac{1}{n} \sum_{g \in \mathcal{G} : \gamma'_g > 0} \alpha'_g \left(\frac{c\gamma'_g}{\alpha'_g} \right)^2 \mathrm{Var}\left[\ell_g^{(i)}\right] = \operatorname*{argmin}_{\alpha' \in \Delta^{|\mathcal{G}|}} \frac{c^2}{n} \sum_{g \in \mathcal{G} : \gamma'_g > 0} \frac{(\gamma'_g)^2}{\alpha'_g} \mathrm{Var}\left[\ell_g^{(i)}\right] \ .$$

The minimizer $\vec{\alpha}^*$ has entries $\alpha^*_g = \gamma'_g \sqrt{\mathrm{Var}[\ell_g^{(i)}]} / \left(\sum_{g \in \mathcal{G}} \gamma'_g \sqrt{\mathrm{Var}[\ell_g^{(i)}]} \right)$. Because $\vec{\alpha}^*$ is unique and determines w^*, $\mathrm{Var}[\hat{L}(w^*, \vec{\alpha}^*, n; f, \ell)] \leq \mathrm{Var}[\hat{L}(w', \vec{\alpha}', n; f, \ell)]$ with strict inequality unless $(w', \vec{\alpha}') = (w^*, \vec{\alpha}^*)$. The optimal weights are

$$w^*(g) = c\gamma'_g/\alpha^*_g = c \left(\sum_{g \in \mathcal{G}} \gamma'_g \sqrt{\mathrm{Var}[\ell_g^{(i)}]} \right) / \sqrt{\mathrm{Var}[\ell_g^{(i)}]} \ .$$

Note that for any pair of groups (g, g'), the relative weights satisfy $\frac{w^*(g)}{w^*(g')} = \sqrt{\frac{\text{Var}[\ell_{g'}^{(i)}]}{\text{Var}[\ell_g^{(i)}]}}$, and thus do not depend on $\vec{\gamma}'$.

If we consider finite sample concerns, the minimizer α^* must satisfy integer values $n\alpha_g^* \in \mathbb{Z}^+ \, \forall g \in \mathcal{G}$. In this case, efficient algorithms exist for finding the integral solution to allocating $\{n_g\}_{g\in\mathcal{G}}$ [294]. However, the non-integer restricted solution α^* has a closed form solution, and we will use the fact that for any group g, α_g^* as defined above and its variant with the additional constraint that $n\alpha_g^* \in \mathbb{Z}^+$ can differ by at most $\frac{|\mathcal{G}|}{n}$. This means that any $\vec{\alpha}$ with $|\alpha_g - \alpha_g^*| > \frac{|\mathcal{G}|}{n}$ cannot be a minimizer of the objective function, even constrained to $n\alpha_g^* \in \mathbb{Z}$. Since $w(g)\alpha_g = c\gamma_g'$, an equivalent statement in terms of w is $|1 - \frac{w(g)}{w^*(g)}| > \frac{w(g)|\mathcal{G}|}{nc\gamma'(g)} = \frac{|\mathcal{G}|}{n\alpha_g} = \frac{|\mathcal{G}|}{n_g}$.

Finally, we show that if $w^*(g) < w(g)$, $\alpha_g^* > \alpha_g$. This follows from our definition of $\vec{\gamma}'$ such that $w(g) = c\gamma_g'/\alpha_g$, and our constraint, such that $w^*(g) = c\gamma_g'/\alpha_g^*$. From these, we must have that $w^*(g)\alpha_g^* = w(g)\alpha_g$, from which the claim follows. □

Since the estimation of risk functions is a key component of learning, Proposition 4.1 illuminates an interplay between the roles of sampling allocations and group-weighting schemes like IW and GDRO. When allocations and weights are jointly maximized, the optimal allocation accounts for an implicit target distribution $\vec{\gamma}'$ (defined above), which may vary by objective function. The optimal weights account for per-group variabilities $\text{Var}_{(x,y)\sim\mathcal{D}_g}[\ell(f(x), y))]$. In Section 4.4 we find that it can be advantageous to use IW and GDRO when some groups have small α_g/γ_g; though the boost in accuracy is less than having an optimally allocated training set to begin with, and diminishes when all groups are appropriately represented in the training set allocation.

4.3 Allocating Samples to Minimize Population-Level Risk

Having motivated the importance of group allocations, we now investigate the direct effects of training set allocations on group and population risks. Using a model of per-group performance as a function of allocations, we study the optimal allocations under a variety of settings.

A Per-Group Power-Law Scaling Model

We model the impact of allocations on performance with scaling laws that describe per-group risks as a function of the number of data points from their respective group, as well as the total number of training instances.

Assumption 4.1 (Group risk scaling with allocation)**.** *The group risks denoted as* $\mathcal{R}(\hat{f}; \mathcal{D}_g) := \mathbb{E}_{(x,y)\sim\mathcal{D}_g}[\ell(\hat{f}(x), y)]$ *scale approximately as the sum of inverse power functions on the number of samples from group* g *and the total number of samples. That is,* $\exists\ M_g > 0$, $\sigma_g, \tau_g, \delta_g \geq 0$,

and $p, q > 0$ such that for a learning procedure which returns predictor $\hat{f}(\mathcal{S})$, and training set $\mathcal{S}$ with group sizes $n_g \geq M_g$:

$$\mathcal{R}\left(\hat{f}\big(\mathcal{S}(\vec{\alpha}, n)\big); \mathcal{D}_g\right) := \mathbb{E}_{(x,y)\sim\mathcal{D}_g}\left[\ell(\hat{f}(\mathcal{S})(x), y)\right] \approx r(\alpha_g n, n; \sigma_g, \tau_g, \delta_g, p, q) \quad \forall\, g \in \mathcal{G}$$

$$r(n_g, n; \sigma_g, \tau_g, \delta_g, p, q) := \sigma_g^2 n_g^{-p} + \tau_g^2 n^{-q} + \delta_g \ . \quad (4.4)$$

Assumption 4.1 is similar to the scaling law in Chen, Johansson, and Sontag [61], but includes a $\tau_g^2 n^{-q}$ term to allow for data from other groups to influence the risk evaluated on group g. It additionally requires that the same exponents p, q apply to each group, an assumption that underpins our theoretical results in Section 4.3. We examine the extent to which Assumption 4.1 holds empirically in Section 4.4, and will modify Eq. (4.4) to include group-dependent terms p_g, q_g when appropriate. The following examples give intuition into the form of Eq. (4.4).

Example 4.1 (Split classifiers per group)**.** When separate models are trained for each group, using training data only from that group, we expect Eq. (4.4) to apply with $\tau_g = 0\ \forall g \in \mathcal{G}$. The parameter p could derived through generalization bounds [42], or through modeling assumptions (Example 4.3). ♢

It is often advantageous to pool training data to learn a single classifier. In this case, model performance evaluated on group g will depend on both n_g and n, as the next examples show.

Example 4.2 (Groups irrelevant to prediction)**.** When groups are irrelevant for prediction and the model class $\mathcal{F}$ correctly accounts for this, we expect Eq. (4.4) to apply with $\sigma_g = 0$ $\forall g \in \mathcal{G}$. ♢

Example 4.3 (Shared linear model with group-dependent intercepts)**.** Consider a $(d+1)$-dimensional linear model, where two groups, $\{A, B\}$, share a weight vector β and features $x \sim \mathcal{N}(0, \Sigma_x)$, but the intercept varies by group:

$$y_i = \beta^\top x_i + c_A \mathbb{I}[g_i = A] + c_B \mathbb{I}[g_i = B] + \mathcal{N}(0, \sigma^2).$$

The ordinary least squares predictor has group risks[2]

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_g}[(x^\top \hat{\beta} + \hat{c}_g - y)^2] = \sigma^2\left(1 + 1/n_g + O(d/n)\right),$$

where the $1/n_g$ arises because we need samples from group g to estimate the intercept c_g, whereas samples from both groups help us estimate β. ♢

[2]The derivation of the group risks is given in Section 4.A.

Example 4.3 suggests that in some settings, we can relate σ_g and τ_g to *group specific* and *group agnostic* model components that affect performance for group g. In general, the relationship between group sizes and group risks can be more nuanced. Data from different groups may be correlated, so that samples from groups similar to or different from g have greater effect on $\mathcal{R}(\hat{f}; \mathcal{D}_g)$ (see e.g. Figure 4.5). Equation 4.4 is meant to capture the dominant effects of training set allocations on group risks and serves as our main structural assumption in the next section, where we study the allocation that minimizes the approximate *population risk*.

Optimal (w.r.t. Population Risk) Allocations

We now study properties of the allocation that minimizes the approximated population risk:

$$\hat{\mathcal{R}}(\vec{\alpha}, n) := \sum_{g \in \mathcal{G}} \gamma_g r(\alpha_g n, n; \sigma_g, \tau_g, \delta_g, p, q) \approx \sum_{g \in \mathcal{G}} \gamma_g \mathcal{R}\left(\hat{f}(\mathcal{S}); \mathcal{D}_g\right) = \mathcal{R}\left(\hat{f}(\mathcal{S}); \mathcal{D}\right) \ . \tag{4.5}$$

The following proposition lays the foundation for two corollaries which show that:

(1) when only the population prevalences $\vec{\gamma}$ vary between groups, the allocation that minimizes the approximate *population risk* up-represents groups with small γ_g;

(2) for two groups with different scaling parameters σ_g, the optimal allocation of the group with $\gamma_g < \frac{1}{2}$ is bounded by functions of σ_A, σ_B, and $\vec{\gamma}$.

Proposition 4.2. *Given a population made up of disjoint groups $g \in \mathcal{G}$ with population prevalences γ_g, under the conditions of Assumption 4.1, the allocation $\vec{\alpha}^* \in \Delta^{|\mathcal{G}|}$ that minimizes the approximated population risk $\hat{\mathcal{R}}$ in Eq. (4.5) has elements:*

$$\alpha_g^* = \frac{\left(\gamma_g \sigma_g^2\right)^{1/(p+1)}}{\sum_{g \in \mathcal{G}} \left(\gamma_g \sigma_g^2\right)^{1/(p+1)}} \ . \tag{4.6}$$

If $\sigma_g = 0 \ \forall g \in \mathcal{G}$, then any allocation in $\Delta^{|\mathcal{G}|}$ minimizes $\hat{\mathcal{R}}$.

Note that $\vec{\alpha}^*$ does not depend on n, $\{\tau_g\}_{g \in \mathcal{G}}$, or q; this will in general not hold if powers p_g differ by group.

Proof of Proposition 4.2. Recall the decomposition of the estimated population risk:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(\hat{f}_{\mathcal{S}}(x), y)\right] = \sum_{g \in \mathcal{G}} \gamma_g \cdot \mathbb{E}_{(x,y)\sim\mathcal{D}_g}\left[\ell(\hat{f}(x), y)\right] \approx \sum_{g \in \mathcal{G}} \gamma_g \left(\sigma_g^2 (n\alpha_g)^{-p} + \tau_g^2 n^{-q} + \delta_g\right) \ .$$

Now we find

$$\begin{aligned}\vec{\alpha}^* &= \underset{\vec{\alpha}\in\Delta^{|\mathcal{G}|}}{\operatorname{argmin}} \sum_{g\in\mathcal{G}} \gamma_g \left(\sigma_g^2 (n\alpha_g)^{-p} + \tau_g^2 n^{-q} + \delta_g\right) \\ &= \underset{\vec{\alpha}\in\Delta^{|\mathcal{G}|}}{\operatorname{argmin}} (n^{-p}) \sum_{g\in\mathcal{G}} \gamma_g \left(\sigma_g^2 (\alpha_g)^{-p}\right) \\ &= \underset{\vec{\alpha}\in\Delta^{|\mathcal{G}|}}{\operatorname{argmin}} \sum_{g\in\mathcal{G}} \gamma_g \sigma_g^2 \alpha_g^{-p} \ .\end{aligned}$$

If $\sigma_g = 0\ \forall g \in \mathcal{G}$, then any allocation $\vec{\alpha}^* \in \Delta^{|\mathcal{G}|}$ minimizes the approximated population loss. Otherwise, α_g^* will be 0 for any group with $\sigma_g = 0$; what follows describes the solution α_g^* for g with $\sigma_g > 0$. If any $\alpha_g = 0$, then the objective is unbounded above, so we can restrict our constraints to $\vec{\alpha} \in (0,1)^{|\mathcal{G}|}$. As the sum of convex functions, the objective function is convex in $\vec{\alpha}$. It is continuously differentiable when $\alpha_g > 0, \forall g \in \mathcal{G}$. The Karush-Kuhn-Tucker (KKT) conditions of this constrained optimization problem are satisfied when

$$\begin{aligned} p\gamma_g \sigma_g^2 \alpha_g^{-(p+1)} &= \lambda \quad \forall g \\ \sum_g \alpha_g &= 1 \ . \end{aligned}$$

Solving this system of equations yields that the KKT conditions are satisfied when $\alpha_g^* = \left(\gamma_g\sigma_g^2\right)^{1/(p+1)} / \sum_g \left(\gamma_g\sigma_g^2\right)^{1/(p+1)}$. Since this is the only solution to the KKT conditions, it is the unique minimizer. □

We now study the form of $\vec{\alpha}^*$ under illustrative settings. Corollary 4.1 shows that when the group scaling parameters σ_g in Eq. (4.4) are equal across groups, the allocation that minimizes the approximate *population risk* allocates samples to minority groups at higher than their *population prevalences*.

Corollary 4.1 (Many groups with equal σ_g)**.** *When $\sigma_g = \sigma > 0,\ \forall g \in \mathcal{G}$, the allocation that minimizes $\hat{\mathcal{R}}$ in Eq.* (4.5) *satisfies $\alpha_g^* \geq \gamma_g$ for any group with $\gamma_g \leq \frac{1}{|\mathcal{G}|}$.*

Proof of Corollary 4.1. Let $m = |\mathcal{G}|$ denote the number of groups. When $\sigma_g = \sigma\ \forall g \in \mathcal{G}$,

$$\alpha_g^* = \frac{\gamma_g^{1/(p+1)}}{\sum_g \gamma_g^{1/(p+1)}} = \gamma_g \cdot \frac{1}{\gamma_g + \gamma_g^{p/(p+1)} \sum_{g'\neq g} \gamma_{g'}^{1/(p+1)}} \ .$$

Since $p > 0$ by Assumption 4.1, we can use the fact that in general, $\sum_{i=1}^n {\gamma_i}^{1/(p+1)}$ with γ_i subject to (a) $\sum_{i=1}^n \gamma_i = s$ for some constant s, and (b) $\gamma_i > 0$, is maximized when all γ_i are

equal. In our setting, since $\sum_{g' \neq g} \gamma'_g = 1 - \gamma_g$,

$$\gamma_g \cdot \frac{1}{\gamma_g + \gamma_g^{p/(p+1)} \sum_{g' \neq g} \gamma_{g'}^{1/(p+1)}} \geq \gamma_g \cdot \frac{1}{\gamma_g + \gamma_g^{p/(p+1)} \sum_{g' \neq g} \left(\frac{1-\gamma_g}{m-1}\right)^{1/(p+1)}}$$
$$= \gamma_g \cdot \frac{1}{\gamma_g + ((m-1)\gamma_g)^{p/(p+1)} (1-\gamma_g)^{1/(p+1)}} .$$

When $\gamma_g \leq 1/m$, $\gamma_g/(1-\gamma_g) \leq \frac{1}{m-1}$, so that

$$\alpha_g^* \geq \gamma_g \cdot \frac{1}{\gamma_g + (1-\gamma_g)(m-1)^{p/(p+1)} (\gamma_g/(1-\gamma_g))^{p/(p+1)}} \geq \gamma_g \cdot \frac{1}{\gamma_g + (1-\gamma_g)} = \gamma_g . \quad \square$$

This shows that the allocation that minimizes population risk can differ from the actual population prevalences $\vec{\gamma}$. In fact, Corollary 4.1 asserts that near the allocation $\vec{\alpha} = \vec{\gamma}$, the marginal returns to additional data from group g are largest for groups with small α_g, enough so as to offset the small weight γ_g in Eq. (4.1). This result provides evidence *against* the idea that small training set allocation to minority groups might comply with minimizing population risk as a result of a small relative contribution to the population risk.

Remark. A counterexample shows that $\alpha_g^* \leq \gamma_g$ does not hold for all g with $\gamma_g > 1/|\mathcal{G}|$. Take $\vec{\gamma} = [.68, .30, .01, .01]$ and $p = 1$; Eq. (4.6) gives $\alpha_2^* > 0.3 = \gamma_2 > 1/4$. In general, whether group g with $\gamma_g \geq 1/|\mathcal{G}|$ gets up- or down-sampled depends on the distribution of $\vec{\gamma}$ across all groups.

Complementing the investigation of the role of the population proportions $\vec{\gamma}$ in Corollary 4.1, the next corollary shows that the optimal allocation $\vec{\alpha}^*$ generally depends on the relative values of σ_g between groups. Inspecting Eq. (4.4) shows that σ_g defines a limit of performance: if σ_g^2 is large, the only way to make the approximate risk for group g small is to make n_g large. For two groups, we can bound the optimal allocations $\vec{\alpha}^*$ in terms of $\{\sigma_g\}_{g \in \mathcal{G}}$ and the population proportions $\vec{\gamma}$. We let A be the smaller of the two groups without loss of generality.

From Eq. (4.6), we know that for two groups, α_A^* is increasing in $\frac{\sigma_A}{\sigma_B}$; Corollary 4.2 gives upper and lower bounds on α_A^* in terms of σ_A and σ_B.

Corollary 4.2 (Unequal per-group constants)**.** *For two groups $\{A, B\} = \mathcal{G}$ with $\gamma_A < \gamma_B$, and parameters $\sigma_A, \sigma_B > 0$ in Eq.* (4.4)*, the allocation of the smaller group α_A^* that minimizes $\hat{\mathcal{R}}$ in Eq.* (4.5) *is upper and lower bounded as*

$$\frac{\gamma_A(\sigma_A^2)^{1/(p+1)}}{\gamma_A(\sigma_A^2)^{1/(p+1)} + \gamma_B(\sigma_B^2)^{1/(p+1)}} < \alpha_A^* < \frac{(\sigma_A^2)^{1/(p+1)}}{(\sigma_A^2)^{1/(p+1)} + (\sigma_B^2)^{1/(p+1)}} .$$

When $\sigma_A \geq \sigma_B$, $\alpha_A^* > \gamma_A$, and when $\sigma_A \leq \sigma_B$, $\alpha_A^* < 1/2$.

Proof of Corollary 4.2. For the two group setting, we can express the optimal allocations as:

$$\alpha_A^* = \frac{(\gamma_A \sigma_A^2)^{1/(p+1)}}{(\gamma_A \sigma_A^2)^{1/(p+1)} + ((1-\gamma_A)\sigma_B^2)^{1/(p+1)}}, \quad \alpha_B^* = 1 - \alpha_A^*$$

Rearranging,

$$\alpha_A^* = \gamma_A \frac{1}{\gamma_A + (\sigma_B^2/\sigma_A^2)^{1/(p+1)}(1-\gamma_A)^{1/(p+1)}(\gamma_A)^{p/(p+1)}} \ .$$

For $p > 0$, it holds that $0 < \frac{1}{p+1} < 1$. Therefore, for any $p > 0$ and $\gamma < 0.5$,

$$\gamma < (\gamma)^{\frac{1}{p+1}}(1-\gamma)^{\frac{p}{p+1}} < (1-\gamma) \ .$$

From this, we derive the upper bound

$$\alpha_A^* < \gamma_A \frac{1}{\gamma_A + (\sigma_B^2/\sigma_A^2)^{1/(p+1)}(\gamma_A)} = \frac{(\sigma_A^2)^{1/(p+1)}}{(\sigma_A^2)^{1/(p+1)} + (\sigma_B^2)^{1/(p+1)}},$$

and the lower bound

$$\alpha_A^* > \gamma_A \frac{1}{\gamma_A + (\sigma_B^2/\sigma_A^2)^{1/(p+1)}(1-\gamma_A)} = \gamma_A \frac{(\sigma_A^2)^{1/(p+1)}}{\gamma_A(\sigma_A^2)^{1/(p+1)} + (\sigma_B^2)^{1/(p+1)}(1-\gamma_A)} \ .$$

When $\sigma_A \geq \sigma_B$,

$$\alpha_A^* > \gamma_A \frac{(\sigma_A^2)^{1/(p+1)}}{\gamma_A(\sigma_A^2)^{1/(p+1)} + (\sigma_B^2)^{1/(p+1)}(1-\gamma_A)} > \gamma_A \frac{(\sigma_A^2)^{1/(p+1)}}{(\gamma_A + 1 - \gamma_A)(\sigma_A^2)^{1/(p+1)}} = \gamma_A \ ,$$

and when $\sigma_A \leq \sigma_B$,

$$\alpha_A^* < \frac{(\sigma_A^2)^{1/(p+1)}}{(\sigma_A^2)^{1/(p+1)} + (\sigma_B^2)^{1/(p+1)}} < \frac{(\sigma_A^2)^{1/(p+1)}}{(\sigma_A^2)^{1/(p+1)} + (\sigma_A^2)^{1/(p+1)}} = 1/2 \ .$$

□

Altogether, these results highlight key properties of training set allocations that minimize population risk. Experiments in Section 4.4 give further insight into the values of weights and allocations for minimizing group and population risks and apply the scaling law model in real data settings.

4.4 Experiments

Having shown the importance of training set allocations from a theoretical perspective, we now undertake an empirical investigation of this phenomenon. Throughout our experiments, we use a diverse collection of datasets to give as full a picture of the empirical phenomena as possible (Table 4.1).

Following a description of the datasets used, the first set of experiments in this section investigates group and population accuracies as a function of the training set allocation $\vec{\alpha}$ by sub-sampling different training set allocations for a fixed training set size. We also study the amount to which importance weighting and group distributionally robust optimization can increase group accuracies, complementing the results of Proposition 4.1 with an empirical perspective. The second set of experiments uses a similar subsetting procedure to examine the fit of the scaling model proposed in Section 4.3.

The third set of experiments investigates using the estimated scaling law fits to inform future sampling practices. We simulate using a small pilot training set to inform targeted dataset augmentation through collecting additional samples. In the final experiment of this chapter, we probe a setting where we might expect the scaling law to be too simplistic, exposing the need for more nuanced modelling in such settings.

In contrast to Section 4.2, here losses are defined over sets of data; note in particular that AUROC is not separable over groups, and thus Eq. (4.1) does not apply for this metric.

Datasets

Here we give a brief description of each dataset we use, with summary in Table 4.1.[3].

Modified CIFAR-4. To create a dataset where we can ensure class balance (equal number of positive and negative labels) across groups, we modify the CIFAR-10 dataset [169] by subsetting to the bird, car, horse, and plane classes. We predict whether the image subject moves primarily by air (plane/bird) or land (car/horse) and group by whether the image contains an animal (bird/horse) or vehicle (car/plane); see Figure 4.1. We set $\gamma = 0.9$.

ISIC. The International Skin Imaging Collaboration (ISIC) dataset is a set of labeled images of skin lesions designed to aid development and standardization of imaging procedures for automatic detection of melanomas [73]. For our main analysis, we follow similar preprocessing steps to [265], removing any images with patches. We predict whether a lesion is benign or malignant, and group instances by the approximate age of the patient of whom the image was taken.

Goodreads. Given publicly available book reviews compiled from the Goodreads database [288], we predict the rating (1-5) corresponding to each review using tf-idf embeddings to featurize reviews. We use data from two genres: "fantasy & paranormal" and "history & biography," and group instances by genre.

[3]See [247] for further details on each experimental setup.

Figure 4.1: Modified CIFAR-4 dataset setup.

Mooc. The HarvardX-MITx Person-Course Dataset contains student demographic and activity data from 2013 offerings of online courses [126]. We predict whether a student earned a certification in the course and we group instances by highest completed level of education.

Adult. The Adult dataset, downloaded from the UCI Machine Learning Repository [94] and originally taken from the 1994 Census database, has a prediction task of whether an individual's annual income is over $\$50,000$. We group instances by sex, codified as binary male/female, and exclude features that directly determine group status.

Table 4.1: Summary descriptions of datasets used throughout Section 4.4.

dataset	groups $\{A, B\}$	γ_A	$\min_g n_g$	n_{test}	target label	loss metric	main model used
CIFAR-4	{animal, vehicle}	0.1	10,000	4,000	air/land	0/1 loss	resnet-18
ISIC	{age ≥ 55, age < 55}	0.43	4,092	2,390	benign/malignant	1 - AUROC	resnet-18
Goodreads	{history, fantasy}	0.38	50,000	25,000	book rating (1-5)	ℓ_1 loss	logistic regression
Mooc	{edu $\leq 2^\circ$, edu $> 2^\circ$}	0.16	3,897	6,032	certified	1 - AUROC	random forest
Adult	{female, male}	0.5	10,771	16,281	income > \$50K	0/1 loss	random forest

Allocation-Aware Objectives vs. Ideal Allocations

We first investigate (a) the change in group and population performance at different training set allocations, and (b) the extent to which optimizing the three objective functions defined in Section 4.2 decreases average and group errors.

For each dataset, we vary the training set allocations $\vec{\alpha}$ between $(0, 1)$ and $(1, 0)$ while fixing the training set size as $n = \min_g n_g$ (see Table 4.1) and evaluate the per-group and population losses on subsets of the heldout test sets. For the image classification tasks, we compare group-agnostic empirical risk minimization (ERM) to importance weighting (implemented via importance sampling (IS) batches following the findings of Buda, Maki, and Mazurowski [48]) and group distributionally robust optimization (GDRO) with group-dependent regularization as described in [251]. For the non-image datasets, we implement

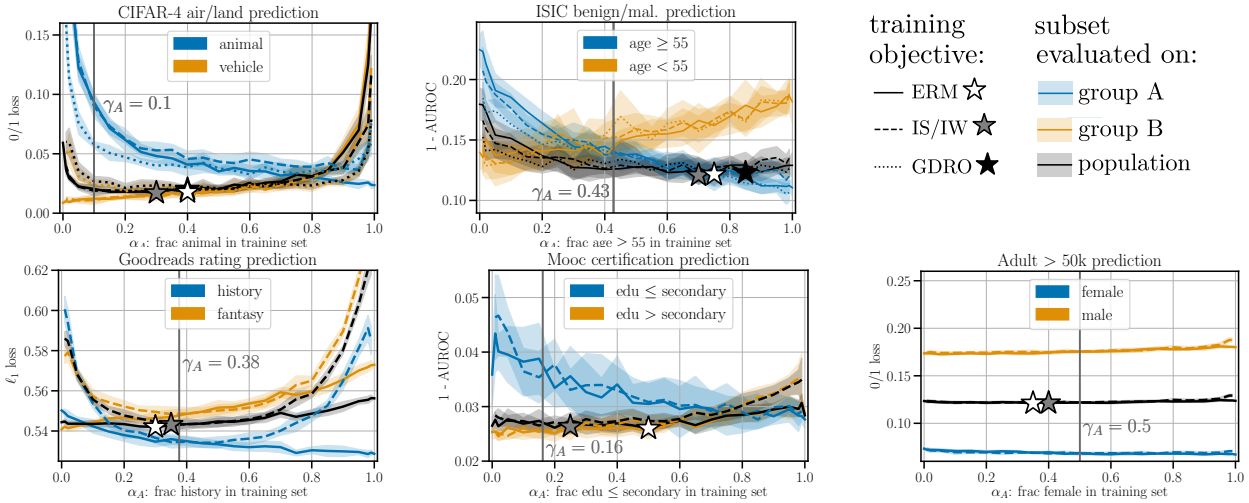


Figure 4.2: Performance across $\vec{\alpha}$ for the different datasets studied. Shaded regions denote one standard deviation above and below the mean over 10 trials. Stars indicate population minima for each objective (ERM: white, IS/IW: grey, GDRO: black).

importance weighting (IW) by weighting instances in the loss function during training, and do not compare to GDRO, as the gradient-based algorithm of Sagawa et al. [251] is not easily adaptable to the predictors we use for these datasets. We pick hyperparameters for each method based on cross-validation results over a coarse grid of $\vec{\alpha}$; for IS, IW, and GDRO, we allow the hyperparameters to vary with $\vec{\alpha}$; for classical ERM we choose a single hyperparameter setting for all $\vec{\alpha}$ values.

Figure 4.2 highlights the importance of at least a minimal representation of each group in order to achieve low population loss (black curves) for all objectives. For CIFAR-4, the population loss increases sharply for $\alpha_A < 0.1$ and $\alpha_A > 0.8$, and for ISIC, when $\alpha_A < 0.2$. While not as crucial for achieving low population losses for the remaining datasets, the *optimal* allocations $\vec{\alpha}^*$ (stars) do require a minimal representation of each group. The $\vec{\alpha}^*$ are largely consistent across the training objectives (different star colors). The population losses (black curves) are largely consistent across mid-range values of α_A for all training objectives. The relatively shallow slopes of the black curves for α_A near α_A^* (stars) stand in contrast to the per-group losses (blue and orange curves), which can vary considerably as $\vec{\alpha}$ changes. From the perspective of model evaluation, this reinforces a well-documented need for more comprehensive reporting of performance. From the view of dataset design, this exposes an opportunity to choose allocations which optimize diverse evaluation objectives while maintaining low population loss. Experiments below investigate this further.

Across the CIFAR-4 and ISIC tasks, GDRO (dotted curves) is more effective than IS (dashed curves) at reducing per-group losses. This is expected, as minimizing the largest loss of any group is the explicit objective of GDRO. Figure 4.2 shows that GDRO can also

improve the *population loss* (see $\alpha_A > 0.7$ for CIFAR-4 and $\alpha_A < 0.2$ for ISIC) as a result of minimizing the worst group loss. Importance weighting (dashed curves) has little effect on performance for Mooc and Adult (random forest models), and actually increases the loss for Goodreads (multiclass logistic regression model).

For all the datasets we study, the advantages of using IS or GDRO are greatest when one group has very small training set allocation (α_A near 0 or 1). When allocations are optimized (stars in Figure 4.2), the boost that these methods give over ERM diminishes. In light of Proposition 4.1, these results suggest that in practice, part of the value of such methods is in compensating for sub-optimal allocations. We find, however, that explicitly optimizing the maximum per-group loss with GDRO can reduce population loss more effectively than directly accounting for allocations with IS.

Assessing the Scaling Law Fits

For each dataset, we combine the results in Figure 4.2 with extra subsetting runs where we vary both n_g and n. From the combined results, we use nonlinear least squares to estimate the parameters of modified scaling laws, where exponents can differ by group:

$$\text{loss}_g \approx \sigma_g^2 n_g^{-p_g} + \tau_g^2 n^{-q_g} + \delta_g \ . \tag{4.7}$$

The estimated parameters of Eq. (4.7) given in Table 4.2 capture different high-level phenomena across the five datasets. For CIFAR-4, $\hat{\tau}_g \approx 0$ for both groups, indicating that most of the group performance is explained by n_g, the number of training instances from that group, whereas the total number of data points n has less influence. For Goodreads, both n_g and n have influence in the fitted model, though $\hat{\tau}_g$ and $\hat{q}_g$ are larger than $\hat{\sigma}_g$ and $\hat{p}_g$, respectively. For ISIC, $\hat{\tau}_A \approx 0$ but $\hat{\tau}_B \not\approx 0$, suggesting other-group data has little effect on the first group, but is beneficial to the latter. For the non-image datasets (Goodreads, Mooc, and Adult), $0 < \hat{\sigma}_g < \hat{\tau}_g$ and $\hat{p}_g < \hat{q}_g$ for all groups.

These results shed light on the applicability of the assumptions made in Section 4.3. Figure 4.3 shows that the fitted curves capture the overall trends of per-group losses as a function of n and n_g. However, the assumptions of Proposition 4.2 and Corollaries 4.1 and 4.2 (e.g., equal p_g for all $g \in \mathcal{G}$) are not always reflected in the empirical fits. Results in Section 4.3 use Eq. (4.4) to describe optimal allocations under different hypothetical settings; we find that allowing the scaling parameters vary by group as in Eq. (4.7) is more realistic in empirical settings.

The estimated models describe the overall trends (Figure 4.3), but the parameter estimates are variable (Table 4.2), indicating that a range of parameters can fit the data, a well-known phenomenon in fitting power laws to data [70]. While we caution against absolute or prescriptive interpretations based on the estimates given in Table 4.2, if such interpretations are desired [61], we suggest evaluating variation due to subsetting patterns and comparing to alternative models such as log-normal and exponential fits (cf. Clauset, Shalizi, and Newman [70]).

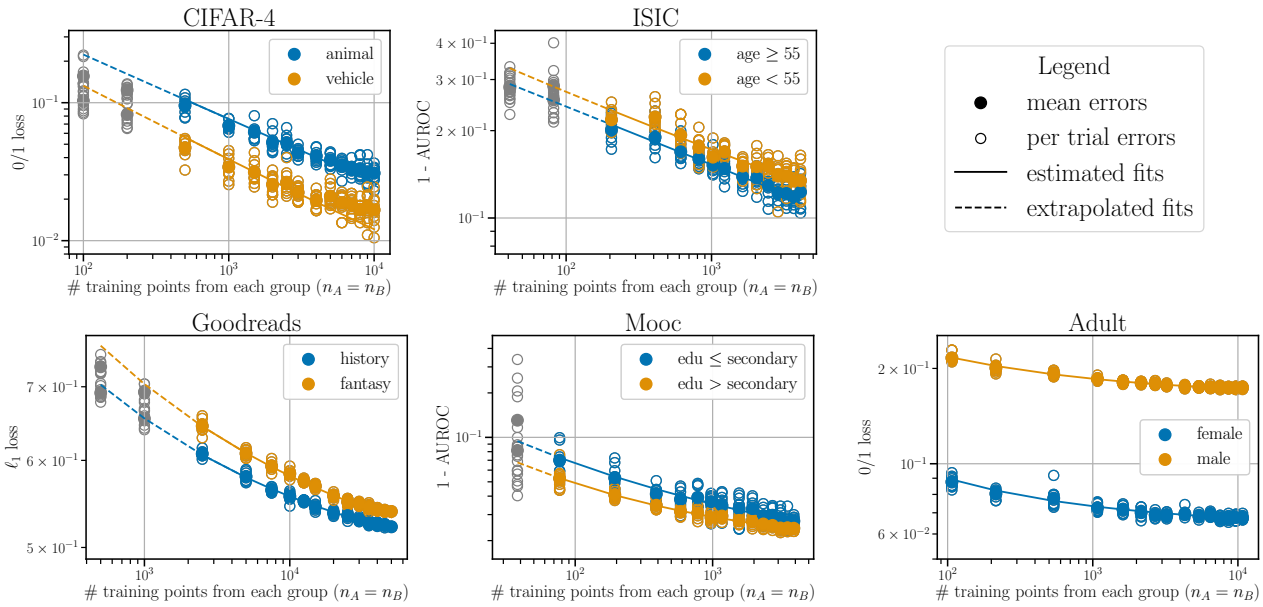


Figure 4.3: Estimated scaling law fits describe observed trends of group errors as a function of (n_g, n). Grey points are not included in the scaling law fit, as $n_g < M_g$ (see Table 4.2).

Table 4.2: Estimated scaling parameters for Eq. (4.7). Parentheses denote standard deviations estimated by the nonlinear least squares fit. Parameters are constrained so that $\hat{\tau}_g, \hat{\sigma}_g, \hat{\delta}_g \geq 0$ and $\hat{p}_g, \hat{q}_g \in [0, 2]$.

dataset	M_g	group g	$\hat{\sigma}_g$	$\hat{p}_g$	$\hat{\tau}_g$	$\hat{q}_g$	$\hat{\delta}_g$
CIFAR-4	500	animal	1.9 (0.12)	0.47 (9.8e-04)	4.5e-09 (1.8e+06)	2.0 (0.0e+00)	1.1e-03 (8.9e-06)
		vehicle	1.6 (0.19)	0.54 (2.0e-03)	3.2e-12 (1.1e+06)	2.0 (0.0e+00)	1.4e-03 (2.8e-06)
ISIC	200	age $\geq$ 55	0.61 (1.7e-03)	0.20 (1.1e-03)	1.7e-09 (1.9e+04)	1.9 (0.0e+00)	1.4e-15 (6.1e-04)
		age $<$ 55	0.26 (9.3e-04)	0.13 (0.012)	0.61 (0.044)	0.3 (7.5e-03)	7.5e-11 (7.2e-03)
Goodreads	2500	history	0.16 (1.2e-03)	0.074 (2.5e-03)	2.5 (0.058)	0.37 (2.0e-04)	0.41 (3.0e-03)
		fantasy	0.62 (0.69)	0.020 (1.2e-03)	3.1 (0.093)	0.39 (1.9e-04)	7.2e-21 (0.72)
Mooc	50	edu $\leq 2°$	0.08 (2.6e-05)	0.14 (6.0e-03)	0.73 (0.059)	0.63 (4.8e-03)	1.3e-15 (2.6e-04)
		edu $> 2°$	0.038 (6.2e-04)	0.068 (6.3e-03)	0.54 (6.5e-03)	0.61 (9.8e-04)	2.8e-12 (8.0e-04)
Adult	50	female	0.078 (0.051)	0.018 (3.6e-03)	0.43 (8.3e-03)	0.59 (1.6e-03)	8.0e-16 (0.052)
		male	0.066 (2.6e-05)	0.21 (1.2e-03)	0.47 (6.5e-03)	0.50 (1.1e-03)	0.16 (5.4e-06)

Targeted Data Collection with Fitted Scaling Laws

We now study the use of scaling models fitted on a small pilot dataset to inform a subsequent data collection effort. Given the result summarized in Figure 4.2, we aim to collect a training set that minimizes the maximum loss on any group. This procedure goes beyond the descriptive use of the estimated scaling models in the previous experiments; important considerations for operationalizing these findings are discussed below.

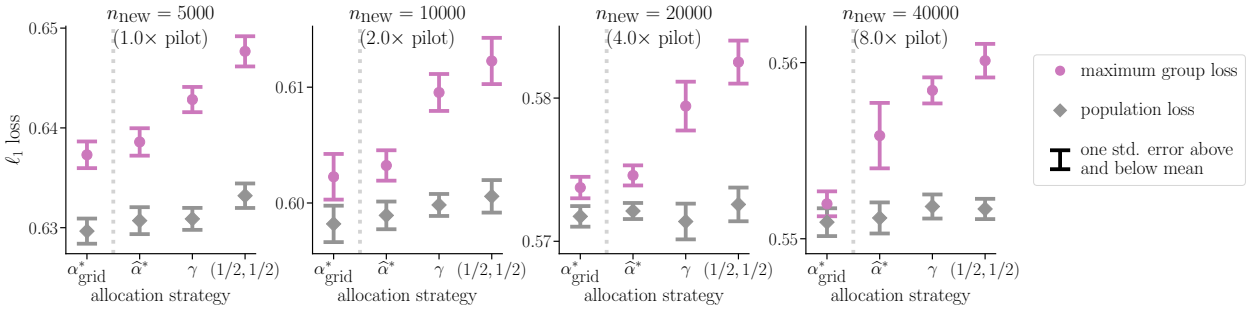


Figure 4.4: Pilot sample experiment. Panels show the result of the three allocations $\vec{\alpha} \in \{\hat{\alpha}^*_{\text{minmax}}, \vec{\gamma}, (1/2, 1/2)\}$ for different sizes of the new training sets compared with an α^*_{grid} baseline that minimizes the maximum group loss over a grid of resolution 0.01, averaged over 10 random trials. Purple circles indicate average maximum error over groups and grey diamonds indicate average population error. Ranges denote standard errors.

We perform this experiment with the Goodreads dataset, the largest of the five we study. The pilot sample contains 2,500 instances from each group, drawn at random from the full training set. We estimate the parameters of Eq. (4.7) using a procedure similar to that described in the experiment above. For a new training set of size n_{new}, we suggest an allocation to minimize the maximum forecasted loss of any group:

$$\hat{\alpha}^*_{\text{minmax}} = \operatorname*{argmin}_{\vec{\alpha} \in \Delta^2} \max_{g \in \mathcal{G}} \left(\hat{\sigma}_g^2 (\alpha_g n_{\text{new}})^{-\hat{p}_g} + \hat{\tau}_g^2 n_{\text{new}}^{-\hat{q}_g} + \hat{\delta}_g \right).$$

For $n_{\text{new}} \in \{1\times, 2\times, 4\times, 8\times\}$ the pilot sample size, we simulate collecting a new training set by drawing n_{new} fresh samples from the training set with allocation $\vec{\alpha} = \hat{\alpha}^*_{\text{minmax}}(n_{\text{new}})$. We train a model on this sample (ERM objective) and evaluate on the test set. For comparison, we also sample at $\vec{\alpha} = \vec{\gamma}$ (population proportions) and $\vec{\alpha} = (0.5, 0.5)$ (equal allocation to both groups). We repeat the experiment, starting with the random instantiation of the pilot dataset, for ten trials. As a point of comparison, we also compute the results for all α in a grid of resolution 0.01, and denote the allocation value in this grid that minimizes the average maximum group loss over the ten trials as α^*_{grid}.

Among the three allocation strategies we compare, $\vec{\alpha} = \hat{\alpha}^*_{\text{minmax}}$ minimizes the average maximum loss over groups, across n_{new} (Figure 4.4). Since per-group losses generally decrease with the increased allocations to that group, we expect the best minmax loss over groups to be achieved when the purple and grey bars meet in Figure 4.4. The allocation strategy $\vec{\alpha} = \hat{\alpha}^*_{\text{minmax}}$ does not quite achieve this; however, it does not increase the population loss over that of the other allocation strategies. This reinforces the finding of previous experiments that different per-group losses can be reached for similar population losses and provides evidence that we can navigate these possible outcomes by leveraging information from a small initial sample.

While the results in Figure 4.4 are promising, error bars highlight the variation across trials. The variability in performance across trials for allocation baseline α^*_{grid} (which is kept constant across the ten trials) is largely consistent with that of the other allocation sampling strategies examined (standard errors in Figure 4.4). However, the estimation of $\hat{\alpha}^*$ in each trial does introduce additional variation: across the ten draws of the pilot data, the range of $\hat{\alpha}^*$ values for subsequent dataset size $n_{\text{new}} = 5000$ is [2e-04,0.04], for $n_{\text{new}} = 10000$ it is [1e-04,0.05], for $n_{\text{new}} = 20000$ it is [5e-05,0.14], and for $n_{\text{new}} = 40000$ it is [2e-05,0.82].

Therefore, the estimated $\hat{\alpha}^*$ should be leveraged with caution, especially if the subsequent sample will be much larger than the pilot sample. Further caution should be taken if there may be distribution shifts between the pilot and subsequent samples. We suggest to interpret estimated $\hat{\alpha}^*$ values as one signal among many that can inform a dataset design in conjunction with current and emerging practices for ethical data collection (see Chapter 6).

Interactions Between Groups

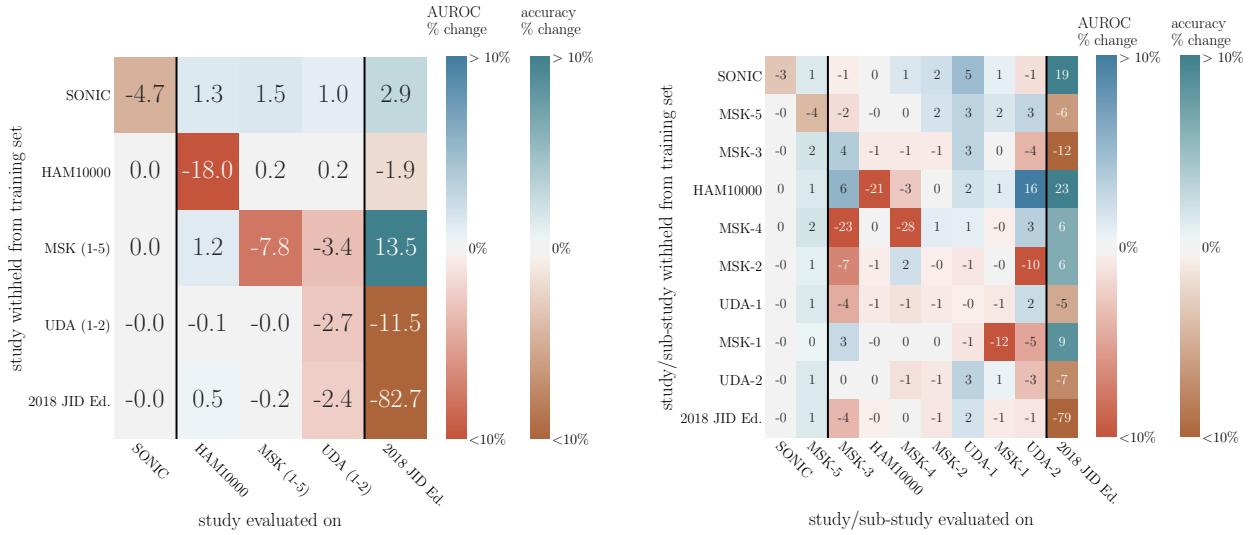
We now shift the focus of our analysis to explore potential between- and within-group interactions that are more nuanced than the scaling law in Eq. (4.4) provides for. The results highlight the need for and encourage future work extending our analysis to more complex notions of groups (e.g., intersectional, continuous, or proxy groups).

As discussed in Section 4.3, data from groups similar to or different from group g may have greater effect on $\mathcal{R}(\hat{f}(\mathcal{S}); \mathcal{D}_g)$ compared to data drawn at random from the entire distribution. We examine this possibility on the ISIC dataset, which is aggregated from different studies [73]). We measure baseline performance of the model trained on data from all of the studies. We then remove one study at a time from the training set, retrain the model, and evaluate the change in performance for all studies in the test set.

Figure 4.5a shows the percent changes in performance due to leaving out studies from the training set. The MSK and UDA studies are comprised of 5 and 2 sub-studies, respectively; Figure 4.5b shows the results of leaving out each sub-study. Rows correspond to the study withheld from the training set and columns correspond to the study used for evaluation. Rows and columns are ordered by % malignancy. For Figure 4.5a this is the same as ordering by dataset size, SONIC being the largest study.

Consistent with our modelling assumptions and results so far, accuracies evaluated on group g decrease as a result of removing group g from the training set (diagonal entries of Figure 4.5). However, additional patterns show more nuanced relationships between groups. Positive values in the upper right regions of Figures 4.5a and 4.5b show that excluding studies with low malignancy rates can raise performance evaluated on studies with high malignancy rates.

This could be partially due to differences in label distributions when removing certain studies from the training data. Importantly, this provides a counterpoint to consequence of Assumption 4.1, that group risks decrease in the total training set size n, regardless of the groups these n instances belong to. To study more nuanced interactions between pairs $g' \neq g$, future work could modify Eq. (4.4) by reparameterizing $r(\cdot)$ to directly account for $n_{g'}$.



(a) Grouped by study.

(b) Grouped by sub-study.

Figure 4.5: Percent change in performance (AUROC / accuracy) due to withholding a study from the training set. Studies are ordered by % malignancy of the data in the evaluation set. SONIC and MSK-5 contain all benign instances and the 2018 JID Editorial Images dataset has all malignant instances; for these we report % change in binary accuracy. For the remaining groups, we report % change in AUROC. Note that the random training/test splits differ between (a) and (b), accounting for the differences in values for corresponding cells between the two figures.

Grouping by substudies within the UDA and MSK studies reveals that even within well defined groups, interactions between subgroups can arise. Negative off-diagonal entries in Figure 4.5b suggest strong interactions between different groups, underscoring the importance of evaluating results across hierarchies and intersections of groups when feasible.

Of the 16,965 images in the full training set, 7,401 are from the SONIC study. When evaluating on all non-SONIC instances (like the evaluation set from the rest of the section), withholding the SONIC study from the training set (akin to the training set from previous experiments) leads to higher AUROC (.905) than training on all studies (0.890). This demonstrates that more data is not always better, especially if the distributional differences between the additional data and the target populations are not well accounted for.

4.5 Conclusions

This chapter studies the ways in which group and population performances are affected by the numerical allocations of discrete groups in training sets. We demonstrate that representation in data is fundamental to training machine learning models that work for the entire

population of interest. By casting dataset design as part of the learning procedure, we can formalize the characteristics that training data must satisfy in order to reach the objectives and specifications of the overall machine learning system. Empirical results bolster our theoretical findings and explore the nuances of real-data phenomena that call for domain dependent analyses in order to operationalize our general results in specific contexts.

While focusing on discrete groups allows us to derive meaningful results, understanding similar phenomena for intersectional groups and continuous notions of inclusion is an important next step. Addressing the more nuanced relationships between the allocations of different data sources (as in the experiment summarized by Figure 4.5) is a first step in this direction. Extending and applying this framework to other objectives and loss functions (e.g., robustness for out-of-distribution prediction and fairness objectives) will also be an important area of future work.

We find that underrepresentation of groups in training data can limit group and population accuracies. However, assuming we can easily remedy these effects by collecting more data about any group is often naive. There may be unintended consequences of upweighting certain groups in an objective function. Naive targeted data collection attempts can present undue burdens of surveillance or skirt consent [220]. When ML systems fail to represent subpopulations due to measurement or construct validity issues, more comprehensive interventions are needed [143]. Chapter 6 discusses these concerns in further detail.

4.A Derivation of Example 4.3

In Example 4.3, we consider the model $y_i = x_i^\top \beta + \alpha_1 \mathbb{I}[g_i = A] + \alpha_2 \mathbb{I}[g_i = B] + \mathcal{N}(0, \sigma^2)$ where $x_i \sim \mathcal{N}(0, \Sigma_x)$ and denote $\theta = [\alpha_1, \alpha_2, \beta^\top]$ (note: here α_i denote the model coefficients, not allocations). We want to compute

$$\begin{aligned}
&\mathbb{E}_{(x,y,g)\sim\mathcal{D}}\left[(\hat{f}(x) - y)^2 | g = A\right] \\
&\quad = \sigma^2 + \mathbb{E}\left[\|x^\top(\hat{\beta} - \beta) + (\hat{\alpha}_1 - \alpha_1)\|^2\right] \\
&\quad = \sigma^2 + \mathbb{E}\left[x^\top(\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top x\right] + 2\mathbb{E}\left[(\hat{\alpha}_1 - \alpha_1)(\hat{\beta} - \beta)^\top x\right] + \mathbb{E}\left[(\hat{\alpha}_1 - \alpha_1)^2\right].
\end{aligned}$$

Since the draw $(x, y, g) \sim \mathcal{D}$ is independent of the data from which the ordinary least squares solution $\hat{\theta}$ is predicted, we can write out each of these terms in terms of the dependence on n, the total number of data points, as well as n_A and n_B (where $n_A + n_B = n$), the total number of datapoints for each group, from which $\hat{\theta}$ is estimated. To do this, we'll solve for

the entries of the covariance matrix:

$$\begin{aligned}
&\mathbb{E}[(\hat{\theta}-\theta)(\hat{\theta}-\theta)^\top] \\
&= \begin{bmatrix} \mathbb{E}\left[(\hat{\alpha_1}-\alpha_1)^2\right] & \mathbb{E}\left[(\hat{\alpha_1}-\alpha_1)(\hat{\alpha_2}-\alpha_2)\right] & \mathbb{E}\left[(\hat{\beta}-\beta)(\hat{\alpha_1}-\alpha_1)\right]^\top \\ \mathbb{E}\left[(\hat{\alpha_1}-\alpha_1)(\hat{\alpha_2}-\alpha_2)\right] & \mathbb{E}\left[(\hat{\alpha_2}-\alpha_2)^2\right] & \mathbb{E}\left[(\hat{\beta}-\beta)(\hat{\alpha_2}-\alpha_2)\right]^\top \\ \mathbb{E}\left[(\hat{\beta}-\beta)(\hat{\alpha_1}-\alpha_1)\right] & \mathbb{E}\left[(\hat{\beta}-\beta)(\hat{\alpha_2}-\alpha_2)\right] & \mathbb{E}\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^\top\right] \end{bmatrix} \\
&= \sigma^2(Z^\top Z)^{-1}
\end{aligned}$$

where Z is the $n \times (d+2)$ design matrix with rows $\{(\mathbb{I}[g_i = A], \mathbb{I}[g_i = B], x_i^\top)\}_{i=1}^n$. Next, we find the block entries of the matrix $(Z^\top Z)^{-1}$. We first interrogate the term within the inverse:

$$Z^\top Z = \begin{bmatrix} n_A & 0 & n_A\bar{x}_A^\top \\ 0 & n_B & n_B\bar{x}_B^\top \\ n_A\bar{x}_A & n_B\bar{x}_B & X^\top X \end{bmatrix}$$

where $\bar{x}_A = \frac{1}{n_A}\sum_{i=1}^n x_i \cdot \mathbb{I}[g_i = A]$, and similarly for $\bar{x}_B$. We'll now use the Schur complement to compute the desired blocks of Σ. The Schur complement is

$$S = X^\top X - \begin{bmatrix} n_A\bar{x}_A & n_B\bar{x}_B \end{bmatrix} \begin{bmatrix} n_A^{-1} & 0 \\ 0 & n_B^{-1} \end{bmatrix} \begin{bmatrix} n_A\bar{x}_A^\top \\ n_B\bar{x}_B^\top \end{bmatrix} = X^\top X - n\bar{x}\bar{x}^\top \ ,$$

which we simplify to $S = X^\top X$ by assuming that we zero-mean the sample feature matrix X before calculating the least squares solution. Using the Schur complement, the covariance matrix in block form is

$$(Z^\top Z)^{-1} = \begin{bmatrix} \frac{1}{n_A} + \bar{x}_A^\top S^{-1}\bar{x}_A & \bar{x}_A^\top S^{-1}\bar{x}_B & -\bar{x}_A^\top S^{-1} \\ \bar{x}_B^\top S^{-1}\bar{x}_A & \frac{1}{n_B} + \bar{x}_B^\top S^{-1}\bar{x}_B & -\bar{x}_B^\top S^{-1} \\ -S^{-1}\bar{x}_A & -S^{-1}\bar{x}_B & S^{-1} \end{bmatrix} \ .$$

Plugging in the appropriate blocks to our original equation, we get:

$$\begin{aligned}
&\mathbb{E}_{(x,y,g)\sim\mathcal{D}}\left[(\hat{f}(x)-y)^2 | g = A\right] \\
&= \sigma^2 + \mathbb{E}_{(x,y)\sim\mathcal{D}_A}\left[x^\top(\hat{\beta}-\beta)(\hat{\beta}-\beta)^\top x + 2(\hat{\alpha}_1-\alpha_1)(\hat{\beta}-\beta)^\top x + (\hat{\alpha}_1-\alpha_1)^2\right] \\
&= \sigma^2(1 + \frac{1}{n_A} + \mathbb{E}_{g=A}\left[x^\top S^{-1}x + \bar{x}_A^\top S^{-1}\bar{x}_A\right]),
\end{aligned}$$

where the middle term can be removed since $\mathbb{E}[x] = 0$. Note that S is the scaled sample covariance matrix. The vectors x_i are drawn i.i.d. from $\mathcal{N}(0, \Sigma_x)$ so that S^{-1} follows an inverse Wishart distribution with parameters n, d, Σ_x. For a freshly drawn independently sampled x,

$$\mathbb{E}[x^\top S^{-1}x] = \text{Trace}(\mathbb{E}[S^{-1}]\mathbb{E}[xx^\top]) = \frac{1}{n-d-1}Tr(\Sigma_x^{-1}\Sigma_x) = \frac{d}{n-d-1} \ .$$

For the $\bar{x}_A^\top S^{-1}\bar{x}_A$ term we invoke the matrix inversion lemma. For a single row x_i of X, let X_{-i} denote the $(n-1)\times d$ matrix comprised of all rows of X except X_i. Then

$$\begin{aligned} x_i^\top (X^\top X)^{-1} x_i &= x_i^\top (X_{-i}^\top X_{-i} + x_i x_i^\top)^{-1} x_i \\ &= x_i^\top \left((X_{-i}^\top X_{-i})^{-1} - (X_{-i}^\top X_{-i})^{-1} x_i (1 + x_i^\top (X_{-i}^\top X_{-i})^{-1} x_i)^{-1} x_i^\top (X_{-i}^\top X_{-i})^{-1} \right) x_i \ . \end{aligned}$$

Letting $a_i = x_i^\top (X_{-i}^\top X_{-i})^{-1} x_i \geq 0$, we rewrite the above as

$$x_i^\top (X^\top X)^{-1} x_i = a_i - \frac{a_i^2}{1+a_i} = \frac{a_i}{1+a_i} \leq a_i \ .$$

Since the x_i are independent and zero mean, $\mathbb{E}\left[x_i^\top (X_{-i}^\top X_{-i})^{-1} x_j\right] = \mathbb{E}[x_i^\top]\mathbb{E}[(X_{-i}^\top X_{-i})^{-1} x_j] = 0 \, \forall i \neq j$. From a similar argument to that given above, we derive that $\mathbb{E}[a_i] = d/(n-d-2)$, so that

$$\begin{aligned} \mathbb{E}\left[\bar{x}_A^\top S^{-1} \bar{x}_A\right] = \mathbb{E}\left[\left(\frac{1}{n_A} \sum_i^{n_A} x_i \right)^\top S^{-1} \left(\frac{1}{n_A} \sum_i^{n_A} x_i \right) \right] &= \frac{1}{n_A^2} \mathbb{E}\left[\sum_i^{n_A} x_i^\top S^{-1} x_i \right] \\ &\leq \frac{1}{n_A} \cdot \frac{d}{n-d-2} \ . \end{aligned}$$

Putting this all together, we conclude that for $n \gg d$,

$$\begin{aligned} \mathbb{E}_{(x,y,g)\sim\mathcal{D}}\left[(\hat{f}(x) - y)^2 | g = A \right] &= \sigma^2 \left(1 + \frac{1}{n_A} + \frac{d}{n-d-1} + \mathbb{E}[\bar{x}_A^\top S^{-1} \bar{x}_A] \right) \\ &= \sigma^2 \left(1 + \frac{1}{n_A} + O\left(\frac{d}{n}\right) \right) \ . \end{aligned}$$

Chapter 5

Post-Estimation Smoothing for Learning with Structural Side Information

This chapter is based on the paper "Post-estimation smoothing: A simple baseline for learning with side information" [244], written in collaboration with Benjamin Recht and Michael I. Jordan.

Observational data are often accompanied by natural structural indices, such as time stamps or geographic locations, which are meaningful to prediction tasks but are often discarded. In this chapter we aim to leverage semantically meaningful indexing data while ensuring robustness to potentially uninformative or misleading indices. We propose a post-estimation smoothing operator as a fast and effective method for incorporating structural index data into prediction. Because the smoothing step is separate from the original predictor, it applies to a broad class of machine learning tasks, with no need to retrain models. Our theoretical analysis details simple conditions under which post-estimation smoothing will improve accuracy over that of the original predictor. Our experiments on large scale spatial and temporal datasets highlight the speed and accuracy of post-estimation smoothing in practice. Together, these results illuminate a novel way to consider and incorporate the natural structure of index variables in machine learning.

5.1 Background

The canonical machine learning setup models pairs of features and labels as originating from some underlying distribution, $\{x_i, y_i\} \sim \mathcal{D}(x, y)$; the problem is to learn a predictor $\widehat{y}(x)$ which describes y as faithfully as possible. However, a recent narrative in machine learning is that well-annotated, large-scale datasets are rare, whereas less curated data are abundant; this has led to a taxonomy of supervision including distant-, weak-, and semi-supervision. Whether labels are noisy by nature (distant) [200], programmatically generated (weak) [236], or missing altogether (semi) [309], it stands that characteristics of some data

necessitate making use of additional sources of constraints.

Semi-supervised methods in particular aim to leverage unlabeled data to elicit an underlying structure which can aid prediction [263]. In practice, however, semi-supervised methods can be computationally expensive, and are sensitive to distribution shifts [213]. In this chapter we propose to use readily-available data that is inherently structural, and apply a robust post-processing method which is independent of the original predictor to incorporate this structure.

We consider scenarios where each datum (x, y) has an associated index t with some linking or semantic meaning. We thus represent observations as triplets:

$$\{x_i, y_i, t_i\} \quad i = 1, ..., n \ .$$

Examples of such triplets include {image, annotation, frame number} in video prediction, {house attributes, price, address} in house price prediction, and {document, sentiment, keywords} in sentiment analysis. While intuition suggests that index variables t may be *correlated* with the label values y and thus are highly informative to the prediction task, in many cases they are not well suited as *predictors* of y without major modification. For example, in object detection in videos, we may expect objects to move smoothly across frames, but the frame number itself does not carry predictive power from one video to another.

We aim to leverage the structural information encoded in t without over-relying on it. This motivates a main question of our work: *how can we utilize the dependence of x and y on t even for predictors that might ignore or underestimate such dependence?* We propose a *post-estimation smoothing* (P-ES) operator $S(t)$ that only depends on t to obtain smoothed predictions:

$$\widetilde{y} = S(t)\widehat{y}(x).$$

Decoupling smoothing $S(t)$ from the initial feature-based prediction step $\widehat{y}(x)$ allows us to efficiently smooth any off-the-shelf model. P-ES applies to any precomputed predictions made over time or space, regardless of the original predictive model. The ease of applying P-ES facilitates robust and reproducible incorporation of index variable structure in predictions.

Throughout this work we consider the setting in which we have a dataset indexed by $t_i \in \mathbb{R}^l$, as well as predictions $\widehat{y}_i \in \mathbb{R}$ associated with each index. It is natural to consider that there is also a set of features $x_i \in \mathbb{R}^d$ and model $f : \mathbb{R}^d \to \mathbb{R}$ from which predictions $\widehat{y} = f(x)$ were generated; we take this as given and work with directly with $\widehat{y}$.

We study the post-prediction application of a P-ES matrix operator $S(t) : \mathbb{R}^n \to \mathbb{R}^n$ to form smoothed predictions, $\widetilde{y} := S(t)\widehat{y}$, such that the $\widetilde{y}$ are closer to the true labels y than the original unsmoothed predictions $\widehat{y}$ are. In our theoretical analysis in Section 5.2, this is measured using the expected mean-squared error: $\mathbb{E}\left[\frac{1}{n}\|\widetilde{y} - y\|_2^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n(\widetilde{y}_i - y_i)^2\right]$, while experiments in Section 5.3 consider different accuracy metrics suitable to different contexts.

In this chapter:

- We formulate a structural-index-based post-process smoothing procedure, P-ES, which is applicable to any predictor.
- We derive theoretical results proving that under mild conditions, P-ES will improve accuracy relative to the original predictor (Theorem 5.1), and characterizing when a linear smoothing operation can greatly increase predictive accuracy (Lemma 5.1).
- We perform experiments on large-scale datasets for human pose estimation and house price prediction demonstrating that P-ES improves accuracy of state-of-the-art predictors at minimal extra cost.

These contributions are made possible by incorporating the general index variables separately from the feature based predictor. This results in a fast, accurate, and robust method for local variance reduction, with the potential to change how we consider and leverage structural variables in machine learning predictions.

More broadly, we demonstrate the effectiveness of a simple method that extends and generalizes previous scholarship in locality-based semi-supervised learning and nonparametric regression, applied in a modern context of abundant but weakly predictive data. Given recent exposition of the systematic underreporting of simple baselines [82, 190] in machine learning and especially in semi-supervised learning [213], it is worth considering P-ES as a theoretically motivated and easily implementable baseline for semi-supervised learning and smoothing in large scale, real-data contexts.

Related Work

Semi-supervised learning (SSL) methods leverage large amounts of unlabeled data along with some labeled data under a local consistency assumption: instances which are near to each other should have similar label values. Distance is commonly determined with respect to an underlying manifold or graph defined by the features [29, 310].

To encourage local consistency of predictions, Belkin, Niyogi, and Sindhwani [30] add a Laplacian regularization term to least squares and support vector machines, and Jean, Xie, and Ermon [147] add a spatial regularization to the loss function of deep neural nets. There is also a considerable amount of work incorporating additive consistency regularization and similar notions in neural nets [18, 112, 273]. As noted by Oliver et al. [213], however, such methods can be sensitive to distribution shifts and require large validation sets and heavy computation to tune parameters; thus they are often poorly suited for "real-world" applications.

Unfortunately, adding a local consistency regularization term multiplies the number of parameter configurations in the optimization problem, and only works for predictors with an explicit objective function. For example, it not straightforward to add a spatial consistency term to random forests. An alternative method, Gaussian harmonic energy minimization (HEM) [310], augments an underlying graph with noisy predictions and solves for spatially

consistent predictions on this larger graph. The local and global consistency (LCG) algorithm [307] solves a similar optimization problem iteratively.

Singh, Nowak, and Zhu [263] show that unlabeled data is useful in SSL precisely when it illuminates the underlying structure of the data beyond what was discernible by the labeled data alone. However, if modeling assumptions incorrectly summarize the true structure of the data, unlabeled data can be misleading and even degrade performance [79, 213]. One approach to mitigate this is to fortify semi-supervised learning methods to be robust to this mismatch [174]; we obtain robustness by decoupling feature-based prediction from a nonparametric incorporation of structural indices.

The literature on nonparametric regression methods is extensive [117, 276]; we focus on two prominent approaches. Gaussian process regression (GPR) places a Gaussian prior on label covariances, specified by feature variables [293]. GPR has been widely adopted and extended in the geospatial statistics community, under the names of "kriging" and "inverse distance interpolation" [16, 184]. As pointed out by Banerjee et al. [23], when applied to large datasets, GPR has large computation and memory requirements, or necessitates approximations [293].

Kernel smoothing [276] is another type of nonparametric regression in which predictions are locally weighted averages of observations. Of particular note is the Nadaraya-Watson estimator [207, 291] in which weights are determined by a kernel relation on all instances. GPR, Laplacian regularized least squares [30], HEM [310], and exact LCG [307] can all be cast as instances of linear smoothing operators for which computing the smoothing matrix involves inverting a matrix of size $n \times n$.

Lastly, recent work on the statistical optimality of data interpolation in machine learning [31] highlights that averaging methods are quite powerful for prediction. We apply a locally-weighted average to *predictions* themselves, which are the output of some prior model. Like semi-supervised learning methods, P-ES encodes spatial consistency properties, but P-ES takes the perspective of refining given predictions with minimal restrictions on the underlying structure. In light of the previous work, P-ES can be seen as a fast and robust way to leverage structure, and to interpolate. Application-specific references are provided in Section 5.3.

5.2 Analysis

Here we answer the questions *(i) how should we form a useful post-estimation smoothing matrix while maintaining robustness to possible distributional misspecification?* and *(ii) for what data distributions and predictors is linear smoothing beneficial?* Throughout the analysis, we model true values y, predictions $\widehat{y}$, and error residuals ε as stochastic processes indexed by t:

$$\widehat{y}(t) = y(t) + \varepsilon(t) \ . \tag{5.1}$$

Accuracy Increases with General Smoothing Matrices

While we may have strong intuition that there is *some* locality-based structure in certain domains, the choice of distributional priors governing this structure will most often be inexact. We use a matrix $W(t) \in \mathbb{R}^{n \times n}$, where weights W_{ij} denote how much the j^{th} prediction should contribute to a smoothed estimate for the i^{th} instance, depending on the values of t_i and t_j.

Theorem 5.1 below shows that using a reasonable weight matrix $W(t)$ which captures correlation in the underlying data can improve performance. A key insight is that shrinking W towards the identity matrix tempers potential misspecification gracefully. Therefore, we form our smoothing matrix as the convex combination:

$$S_c(t) = c \cdot W(t) + (1 - c) \cdot I \ , \tag{5.2}$$

where in practice $c \in [0, 1]$ can be chosen through cross-validation along with any parameters of W.

For any weight matrix, define the following quantities: $\gamma(\varepsilon, W)$, which describes the amount by which W acts as a zero operator on the errors, and $\beta(\varepsilon, W; y)$, which describes the amount by which W acts as the identity operator on the true labels, both scaled by $\mathbb{E}[\|\varepsilon\|_2^2]$:

$$\gamma(\varepsilon, W) := \mathbb{E}[\varepsilon^\top W \varepsilon] / \mathbb{E}[\|\varepsilon\|_2^2],$$
$$\beta(\varepsilon, W; y) := \mathbb{E}[\varepsilon^\top (W - I) y] / \mathbb{E}\left[\|\varepsilon\|_2^2\right] \ .$$

Intuitively, we want to use a weight matrix W such that both γ and β are small, so that W averages out erroneous error signals while decreasing correlation between y and ε. Theorem 5.1 shows that an imperfect W will suffice, so long as the sum $\gamma + \beta$ is controlled.

Theorem 5.1. *Given any predictor $\widehat{y}$ of y with error residuals satisfying $\mathbb{E}\left[\|\varepsilon\|_2^2\right] \neq 0$, and any weight matrix W satisfying $\gamma(\varepsilon, W) + \beta(\varepsilon, W; y) < 1$, there exists a constant $c \in (0, 1]$ such that the smoothing matrix $S_c = c \cdot W + (1 - c) \cdot I$ strictly reduces expected mean squared error (MSE) of the resulting predictions:*

$$\mathbb{E}\left[\tfrac{1}{n}\|S_c \widehat{y} - y\|_2^2\right] < \mathbb{E}\left[\tfrac{1}{n}\|\widehat{y} - y\|_2^2\right] \ .$$

Proof of Theorem 5.1. Let $\mu := \mathbb{E}[\varepsilon] = \mathbb{E}[\hat{y} - y]$. The squared error of using smoothing matrix $S_c = cW + (1 - c)I$ decomposes as:

$$\begin{aligned}
\|S_c \hat{y} - y\|_2^2 &= \|c(W\hat{y} - y) + (1 - c)(\hat{y} - y)\|_2^2 \\
&= \|c(W\hat{y} - y) + (1 - c)(\varepsilon)\|_2^2 \\
&= c^2 \|W\hat{y} - y\|_2^2 + (1 - c)^2 \|\varepsilon\|_2^2 + 2c(1 - c)(\varepsilon^\top W \varepsilon + \varepsilon^\top (W - I) y)
\end{aligned}$$

so that the expected reduction in MSE is given by

$$\begin{aligned}
\mathbb{E}\left[\|S_c \hat{y} - y\|_2^2\right] - \mathbb{E}\left[\|\hat{y} - y\|_2^2\right] &= c^2 \mathbb{E}\left[\|W\hat{y} - y\|_2^2\right] + \left(1 + (c^2 - 2c) + 2(c - c^2)\gamma\right) \mathbb{E}\left[\|\varepsilon\|_2^2\right] \\
&\quad + 2c(1 - c)\mathbb{E}\left[\varepsilon^\top (W - I) y)\right] - \mathbb{E}\left[\|\varepsilon\|_2^2\right] \\
&= c^2 \mathbb{E}\left[\|W\hat{y} - y\|_2^2\right] + \left((c^2 - 2c) + 2(c - c^2)(\gamma + \beta)\right) \mathbb{E}\left[\|\varepsilon\|_2^2\right] \ .
\end{aligned}$$

This is a quadratic in c:

$$\mathbb{E}\left[\|S_c\hat{y} - y\|_2^2\right] - \mathbb{E}\left[\|\hat{y} - y\|_2^2\right] = c^2\left(\mathbb{E}\left[\|W\hat{y} - y\|_2^2\right] + (1 - 2(\gamma+\beta))\mathbb{E}\left[\|\varepsilon\|_2^2\right]\right) + 2c\left((\gamma+\beta-1)\mathbb{E}\left[\|\varepsilon\|_2^2\right]\right) \ .$$

We first show that under the assumptions above, this expression is convex. Afterwards, we will show that the nonzero root is strictly greater than zero, and therefore conclude that there must be a value $c \in (0, 1]$ for which the objective is negative. We first get a handle on the coefficient of the quadratic term:

$$\begin{aligned}\mathbb{E}\left[\|W\hat{y} - y\|_2^2 + (1 - 2(\gamma+\beta))\|\varepsilon\|_2^2\right] &= \mathbb{E}\left[\|W\hat{y} - y\|_2^2 + \|\varepsilon\|_2^2 - 2\varepsilon^\top W\varepsilon + 2\varepsilon^\top(I - W)y\right] \\ &= \mathbb{E}\left[\|W\hat{y}\|_2^2 - 2y^\top W\widehat{y} + \|\widehat{y}\|_2^2 - 2\varepsilon^\top W\widehat{y}\right] \\ &= \mathbb{E}\left[\|(W - I)\hat{y}\|_2^2\right] \\ &\geq 0 \ .\end{aligned}$$

The coefficient on the quadratic term is nonnegative, so that the expression is convex in c. Now we show that under the conditions outlined in the theorem statement, the coefficient on the linear term is negative. Recall the condition that the matrix W acts close to the identity on y but close to the zero matrix on ε, with respect to the errors: $\gamma(\varepsilon, W) + \beta(\varepsilon, W; y) < 1$. When this conditions holds, we have

$$2\left((\gamma+\beta-1)\mathbb{E}\left[\|\varepsilon\|_2^2\right]\right) < 0 \ .$$

Thus, the optimal c value is given as

$$c^* = \frac{(1 - (\gamma+\beta))\,\mathbb{E}\left[\|\varepsilon\|_2^2\right]}{\mathbb{E}\left[\|W\hat{y} - y\|_2^2\right] + (1 - 2\gamma - 2\beta)\mathbb{E}\left[\|\varepsilon\|_2^2\right]} \ .$$

Since c^* is always positive, by convexity and continuity of the objective function, the optimal value for c within the range $(0, 1]$ is $\min(c^*, 1)$. If $c^* > 1$, this implies that

$$\begin{aligned}(1 - (\gamma+\beta))\,\mathbb{E}\left[\|\varepsilon\|_2^2\right] &> \mathbb{E}\left[\|W\hat{y} - y\|_2^2\right] + (1 - 2\gamma - 2\beta)\mathbb{E}\left[\|\varepsilon\|_2^2\right] \\ (\gamma+\beta))\mathbb{E}\left[\|\varepsilon\|_2^2\right] &> \mathbb{E}\left[\|W\hat{y} - y\|_2^2\right] \ .\end{aligned}$$

If this is the case, then clipping the chosen c to be $c = 1$ (denote the resulting smoothing matrix S_1) will result in expected MSE decrease

$$\begin{aligned}\mathbb{E}\left[\tfrac{1}{n}\|S_1\hat{y} - y\|_2^2\right] - \mathbb{E}\left[\tfrac{1}{n}\|\varepsilon\|_2^2\right] &= \mathbb{E}\left[\tfrac{1}{n}\|W\hat{y} - y\|_2^2\right] - \mathbb{E}\left[\tfrac{1}{n}\|\varepsilon\|_2^2\right] \\ &< -(1 - \gamma - \beta)\mathbb{E}\left[\tfrac{1}{n}\|\varepsilon\|_2^2\right] \ .\end{aligned}$$

Otherwise (if $c^* \leq 1$), the resulting expected MSE decrease is upper bounded as

$$\begin{aligned}&\mathbb{E}\left[\tfrac{1}{n}\|S_{c^*}\hat{y} - y\|_2^2\right] - \mathbb{E}\left[\tfrac{1}{n}\|\varepsilon\|_2^2\right] \\ &\quad\leq -\frac{(1 - \gamma - \beta)^2\,\mathbb{E}\left[\|\varepsilon\|_2^2\right]^2}{n\left(\mathbb{E}\left[\|W\hat{y} - y\|_2^2\right] + (1 - 2\gamma - 2\beta)\mathbb{E}\left[\|\varepsilon\|_2^2\right]\right)} \\ &\quad= -(1 - \gamma - \beta)\mathbb{E}\left[\tfrac{1}{n}\|\varepsilon\|_2^2\right] \cdot \frac{(1 - \gamma - \beta)\,\mathbb{E}\left[\|\varepsilon\|_2^2\right]}{\left(\mathbb{E}\left[\|W\hat{y} - y\|_2^2\right] + (1 - 2\gamma - 2\beta)\mathbb{E}\left[\|\varepsilon\|_2^2\right]\right)} \ .\end{aligned}$$

The optimal resulting MSE reduction from using S_c where $c = \min\{c^*, 1\}$ is then bounded as

$$\begin{aligned}
&\mathbb{E}\left[\tfrac{1}{n}\|S_c\hat{y} - y\|_2^2\right] - \mathbb{E}\left[\tfrac{1}{n}\|\hat{y} - y\|_2^2\right] \\
&\qquad \leq -(1-\gamma-\beta)\mathbb{E}\left[\tfrac{1}{n}\|\varepsilon\|_2^2\right] \cdot \min\left\{1, \frac{(1-\gamma-\beta)\mathbb{E}\left[\|\varepsilon\|_2^2\right]}{\left(\mathbb{E}\left[\|W\hat{y} - y\|_2^2\right] + (1-2\gamma-2\beta)\mathbb{E}\left[\|\varepsilon\|_2^2\right]\right)}\right\} \\
&\qquad < 0\ . \qquad \square
\end{aligned}$$

Theorem 5.1 inverts the standard statistical smoothing analysis [261, 276]—which, provided generative processes as in Eq. (5.1), calculates the bias and variance of the resulting smoothed estimator—and instead characterizes properties of the underlying signal $y(t)$, the prediction errors $\varepsilon(t)$, and the weight matrix W that make smoothing beneficial. As in standard kernel smoothing regression [289], in P-ES we are willing to tolerate an increase in the bias of our predictor, so long as the variance decreases. Formulating the conditions in terms of γ and β allows us to assess this trade-off in terms of the conditions on the prediction errors directly.

We can guarantee that $\gamma(\varepsilon, W) \leq 1$ by ensuring $\lambda_{\max}(W) \leq 1$, for example by taking any right-stochastic matrix (here $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue). Controlling $\beta(\varepsilon, W; y)$ depends on both the true values and the errors in predictions. A sufficient condition to achieving $\beta \leq c$ is to pick W which satisfies $\mathbb{E}\left[\|(W-I)y\|_2^2\right] \leq c^2\mathbb{E}\left[\|\varepsilon\|_2^2\right]$, encoding that we tolerate a deviation in labels due to W limited by the magnitude of errors that can potentially be reduced. The next two paragraphs show that under a reasonable assumption on the predictions, $\beta + \gamma < 1$ can always be satisfied and detail practical considerations in picking W and checking the conditions of the theorem.

On ensuring $\beta+\gamma < 1$: Manipulation of the definitions of γ and β shows that the condition $\beta + \gamma < 1$ is equivalent to the condition $\mathbb{E}[\varepsilon^\top(W-I)\hat{y}] < 0$, which is always satisfiable with some W, so long as $\mathbb{E}[\varepsilon\hat{y}^\top]$ is not the all zeros matrix. Further, when $|\mathbb{E}[\varepsilon^\top y]| < \mathbb{E}[\varepsilon^\top\varepsilon]$, $W = t \cdot I$ for any $t < 1$ will suffice so that $\beta + \gamma < 1$. The wide range of possible t is because the matrix W is combined in a convex combination with the identity matrix to form $S_c(t)$ in Eq. (5.2).

Lemma 5.1 and Example 5.1 below show that an optimal smoothing matrix averages out errors in the predictions, depending on the structure in y and ε. We'd like our empirical choice of W to be close to this optimal matrix. For practical applications, we could (a) use empirical covariance matrices from training/validation data to inform our choice of W, and/or (b) for a pre-specified W we could estimate γ and β by using the training/validation data to estimate ε and y. We suspect that estimating γ and β in this way may not be practically necessary, for the following reason. If the chosen matrix W does not reduce the mean squared error for any choice of $c \in (0, 1]$, then cross validation over parameter c will result in $c = 0$, such that no smoothing occurs. Since cross-validating over c amounts to only vector (not matrix) operations, it is practical to sweep over a large number of possible c's. Thus, it could be just as fast to check if smoothing with matrix S_c (for any of the c's)

reduces the MSE as to check the condition $\gamma + \beta < 1$.

In general, the condition $\beta + \gamma < 1$ represents a trade-off in choosing a weight matrix that acts approximately as a zero matrix with respect to the errors (small γ), while acting close to an identity matrix with respect to the true values (small β). In order to keep the sum small, W needs to incorporate knowledge in the structure of the domain-specific labels, y, as well as the distribution of prediction errors, ε, for a given predictor. For all experiments (Section 5.3), we use the Nadaraya-Watson smoothing matrix with a Gaussian kernel (Eq. (5.4)). This matrix is right-stochastic, and the Gaussian kernel encodes the constraint that nearby data points (measured with respect to structural index variable t) should have similar label values y.

The best-case reduction in MSE attainable by P-ES is bounded in the final line of the proof of Theorem 5.1. The MSE reduction depends on covariances between W, y, and the error residuals in $\hat{y}$. This motivates us to study the form of the optimal smoothing operator and the resulting expected error reduction, when these covariances are known.

Optimal P-ES for Known Distributions

Having proposed a P-ES matrix S_c in the previous section, we now study the form of an optimal linear smoothing matrix S^* when the distributions governing the labels and error residuals are known. This reinforces the high-level structures we wish to capture in S_c, and provides a baseline for simulation experiments in Section 5.3. Denote the cross-correlation matrices K element-wise as $K_{xy}[t, s] = \mathbb{E}[x(t)y(s)]$.

Lemma 5.1. *For a predictor $\widehat{y}$ of y with error residuals distributed as $\varepsilon(t) = \widehat{y}(t) - y(t)$, when $K_{\widehat{y}\widehat{y}} \succ 0$, the optimal linear smoothing matrix has the form*

$$\begin{aligned} S^* &= \operatorname*{argmin}_{S \in \mathbb{R}^{n \times n}} \mathbb{E}\left[\tfrac{1}{n}\|S\widehat{y} - y\|_2^2\right] \\ &= I - (K_{\varepsilon\varepsilon} + K_{y\varepsilon})^\top (K_{yy} + K_{y\varepsilon} + K_{\varepsilon y} + K_{\varepsilon\varepsilon})^{-1} \ . \end{aligned}$$

The expected MSE reduction of applying S^* versus using the original predictions $\widehat{y}$ is always non-negative, and is given by

$$\tfrac{1}{n}\mathbb{E}\left[\|\widehat{y} - y\|_2^2\right] - \|S^*\widehat{y} - y\|_2^2\right] = \tfrac{1}{n}\mathrm{tr}\left(K_{\widehat{y}y}^\top (K_{\widehat{y}\widehat{y}})^{-1} K_{\widehat{y}y}\right) \ .$$

Proof of Lemma 5.1. Setting the matrix differential of the following convex objective to zero, any solution S^* to

$$S^* = \operatorname*{argmin}_{S \in \mathbb{R}^{n \times n}} \mathbb{E}\left[\tfrac{1}{n}\|S\widehat{y} - y\|_2^2\right]$$

satisfies

$$\frac{\partial}{\partial S}\mathbb{E}\left[(S\widehat{y} - y)^\top (S\widehat{y} - y)\right] = 2(SK_{\widehat{y}\widehat{y}} - K_{y\widehat{y}}) = 0 \ .$$

If $K_{\widehat{y}\widehat{y}}$ is positive definite (and thus invertible), the objective is strictly convex and the unique optimal solution is

$$\begin{aligned} S^* &= K_{y\widehat{y}}(K_{\widehat{y}\widehat{y}})^{-1} \\ &= I - K_{\varepsilon\widehat{y}}(K_{\widehat{y}\widehat{y}})^{-1} \\ &= I - (K_{\varepsilon\varepsilon} + K_{\varepsilon y})(K_{yy} + K_{y\varepsilon} + K_{\varepsilon y} + K_{\varepsilon\varepsilon})^{-1} \ . \end{aligned}$$

Since the identity matrix I is within the set of possible estimators ($\mathbb{R}^{n\times n}$), we know that the resulting objective satisfies $\mathbb{E}\left[\|S^*\widehat{y} - y\|_2^2\right] \leq \mathbb{E}\left[\|\widehat{y} - y\|_2^2\right]$. In fact, applying properties of the trace operator (cyclic property, invariance to transposes) gives the following expression for the reduction in expected squared error:

$$\begin{aligned} \mathbb{E}\left[\|\widehat{y} - y\|_2^2 - \|S^*\widehat{y} - y\|_2^2\right] &= \mathrm{tr}\left(K_{yy} + K_{\widehat{y}\widehat{y}} - 2K_{y\widehat{y}} - K_{yy} + K_{y\widehat{y}}\left(K_{\widehat{y}\widehat{y}}\right)^{-1}K_{\widehat{y}y}\right) \\ &= \mathrm{tr}\left((K_{\widehat{y}\widehat{y}} - K_{y\widehat{y}})(K_{\widehat{y}\widehat{y}})^{-1}(K_{\widehat{y}\widehat{y}} - K_{\widehat{y}y})\right) \\ &= \mathrm{tr}\left((K_{\varepsilon\varepsilon} + K_{y\varepsilon})^\top(K_{yy} + K_{\varepsilon\varepsilon} + K_{\varepsilon y} + K_{y\varepsilon})^{-1}(K_{\varepsilon\varepsilon} + K_{y\varepsilon})\right) \ . \end{aligned}$$

Applying a matrix trace inequality for positive definite matrix A and positive semi-definite matrix B: $\mathrm{tr}\left(A^{-1}B\right) \geq \lambda_{\min}(A^{-1})\mathrm{tr}\left(B\right) = \mathrm{tr}\left(B\right)/\lambda_{\max}(A) \geq \mathrm{tr}\left(B\right)/\mathrm{tr}\left(A\right)$ gives an upper bound on the reduction:

$$\mathbb{E}\left[\|\widehat{y} - y\|_2^2 - \|S^*\widehat{y} - y\|_2^2\right] \geq \frac{\mathrm{tr}\left(\left(K_{\varepsilon\varepsilon} + K_{y\varepsilon}\right)\left(K_{\varepsilon\varepsilon} + K_{y\varepsilon}\right)^\top\right)}{\mathrm{tr}\left(K_{yy} + K_{\varepsilon\varepsilon} + K_{\varepsilon y} + K_{y\varepsilon}\right)} \ .$$

Note that $(K_{\varepsilon\varepsilon} + K_{y\varepsilon})(K_{\varepsilon\varepsilon} + K_{y\varepsilon})^\top$ and $K_{yy} + K_{\varepsilon\varepsilon} + K_{\varepsilon y} + K_{y\varepsilon} = K_{\hat{y}\hat{y}}$ are positive semi-definite by construction and positive definite by assumption, respectively. □

Lemma 5.1 shows that smoothing can reduce prediction error associated with $K_{\varepsilon\varepsilon}$ and $K_{y\varepsilon}$, but that the extent to which errors can be smoothed out depends on the forms of K_{yy} and $K_{y\varepsilon}$. The following example underscores this point for an illustrative data generating model and sets the stage for simulation experiments in Section 5.3. The main details of the example are given here, with more extensive exposition in [244].

Example 5.1. Consider zero-mean stochastic processes $x(t)$ and $y(t)$ which are dependent on a third zero-mean hidden process $z(t)$, but with independent additive Gaussian noise. In particular:

$$\begin{aligned} z &\sim \mathcal{N}(0, K_{zz}) & & \\ x(t) &= z(t) + \omega(t), & \omega(t) &\sim_{i.i.d.} \mathcal{N}(0, \sigma_x^2) \\ y(t) &= a \cdot z(t) + \mu(t), & \mu(t) &\sim_{i.i.d.} \mathcal{N}(0, \sigma_y^2) \ . \end{aligned} \tag{5.3}$$

The autocorrelation matrices show that there is shared variation due to the "hidden" process z:

$$K_{xx} = K_{zz} + K_{\omega\omega}, \qquad K_{yy} = a^2K_{zz} + K_{\mu\mu}, \qquad K_{xy} = aK_{zz} \ .$$

Without any specific knowledge of the covariance structure in z, this could be modeled with an "errors in variables" model, for which total least squares (TLS) gives a statistically consistent estimator of a. In the setting of this example, the asymptotic distribution of the TLS estimator is normal, with mean a, and variance approaching 0 as $n \to \infty$ [139, 254], from which it follows that the expected MSE of the TLS predictions approaches

$$\mathbb{E}\left[\tfrac{1}{n}\|\widehat{y}_{TLS} - y\|_2^2\right] \approx \sigma_y^2 + a^2\sigma_x^2$$

as n grows large. Invoking Lemma 5.1, the expected smoothed performance using S^* approaches

$$\mathbb{E}\left[\tfrac{1}{n}\|S^*\widehat{y}_{TLS} - y\|_2^2\right] \approx \sigma_y^2 + a^2\sigma_x^2\left(1 - \tfrac{1}{n}\mathrm{tr}\left((\sigma_x^{-2}K_{zz} + I)^{-1}\right)\right) \geq \sigma_y^2 \ .$$

The expected MSE reduction is approximately $\frac{a^2\sigma_x^2}{n}\mathrm{tr}\left((\sigma_x^{-2}K_{zz} + I)^{-1}\right)$ which is strictly positive for $a > 0$ and $\sigma_x^2 > 0$, and increasing with σ_x^2. $\Diamond$

5.3 Experiments

We first use simulated experiments to study situations in which smoothing is beneficial and to demonstrate that a simple instantiation of Eq. (5.2) achieves close to optimal accuracy in these settings. We then apply P-ES to predictions on real-world datasets with temporal and spatial structure: human-pose prediction in video and house-price prediction over space. P-ES improves performance of all predictors we consider, including some that already incorporate locality. P-ES compares favorably to statistical smoothing and SSL methods, both in predictive accuracy and computation time. As a simple local-averaging weight matrix, all experiments use as W the Nadaraya-Watson smoothing matrix with squared exponential kernel on t, where $D_{ij}(t;\sigma) = e^{-\frac{1}{2\sigma^2}\|t_i - t_j\|_2^2}$:

$$S_c(t;\sigma) = c \cdot \mathrm{diag}^{-1}\left(D\vec{1}\right)D + (1-c)\cdot I \ . \tag{5.4}$$

Simulations

We return to the distribution defined in Eq. (5.3) in Example 5.1, where the processes $x(t)$ and $y(t)$ are influenced by a third hidden process $z(t)$. We now make a specific assumption for the covariance of z:

$$\vec{z} = \mathcal{N}\left(\vec{0}, \Sigma(t)\right), \quad \Sigma_{ij}(t) = e^{-\frac{1}{2\sigma_z^2}(t_i - t_j)^2} \ .$$

We take $t = (0, 1/n, 2/n, \ldots, (n-2)/n, (n-1)/n)$ with $n = 2000$, and $\sigma_z = 0.2$. Half of the points are chosen at random to form a training set from which we learn the total least squares (TLS) estimator $\widehat{y}(x)$. The remaining 1000 points are used to evaluate performance with and without P-ES. To show the expressiveness of the matrix S_c, in simulations we

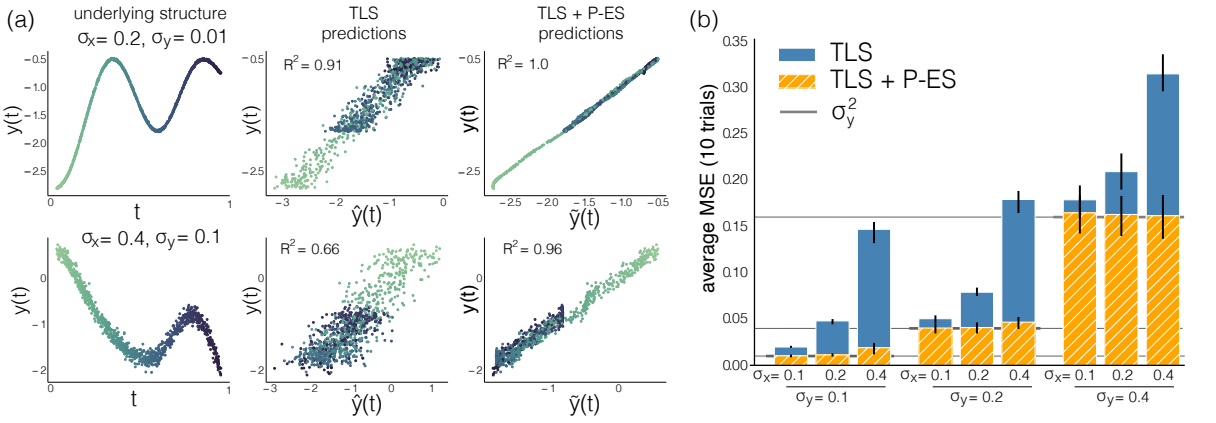


Figure 5.1: Simulation results. (a) Two examples of structure in z (left column), where a total least squares (TLS) estimator recovers structure (middle column), but is improved upon using P-ES (right column). (b) Aggregate performance over different noise parameters for unsmoothed and P-ES estimates, compared to a lower bound of σ_y^2 for any linear smoother. Vertical black lines show minima and maxima over 10 trials.

pick the parameters c, σ of S_c so as to maximize performance on the evaluation set. In experiments with real data below, we pick parameters on a validation set before applying to a holdout set.

Figure 5.1(a) shows the process of P-ES as local variance reduction. Each row shows a different setting of σ_x, σ_y. The leftmost column shows observed labels $y(t)$ as a function of indices t. The middle and right columns show the TLS predictions, without and with P-ES, respectively. Errors in $\widehat{y}$ that are made in the horizontal axis are reducible by smoothing, as $S_c(t)$ gives more weight to pairs closer in t (similar hue in Figure 5.1(a)). The smoothed predictions $\widetilde{y}$ exhibit a similar structure to the original predictions, with significantly reduced horizontal error bands. The difference in the performance of the TLS estimator with and without P-ES (Figure 5.1(b)) indicates that P-ES reduces prediction errors that are uncorrelated with the index variable t.

Human Pose Prediction in Video

Recent work has shown that improvements in human pose estimation [81, 156, 303] and object detection and classification in videos [226, 300, 311] can be obtained by encoding temporal consistency as part of a larger predictive pipeline. The intuition is that exploiting continuity of motion over video frames can reduce the noise in per-frame predictions. For example, a recent state-of-the-art method for pose estimation [156] learns both a temporal encoder and temporal human dynamics as part of the predictive pipeline. In the following experiment, we apply P-ES to predictions from this model as well as to predictions from a per-frame baseline model [155].

Table 5.1: Holdout set performance for human pose estimation in video. Arrows indicate the direction of desired performance; bold numbers indicate the best performance for each metric. For all metrics, P-ES predictions (italicized methods) have the best performance of all methods considered. Other methods are attributed as "temporal, temporal + dynamics:" Kanazawa et al. [156], "per-frame:" Kanazawa et al. [155, 156].

	3DPW				Penn Action	
method	PCK ↑	MPJPE ↓	PA-MPJPE ↓	Acc. Err. ↓	PCK ↑	Accel ↓
per-frame	84.06	129.95	76.68	37.41	73.17	79.91
per-frame, with P-ES	84.46	128.44	**75.84**	20.46	73.74	48.22
temporal	82.59	139.19	78.35	15.15	71.16	29.30
temporal + dynamics	86.37	127.08	80.05	16.42	77.88	29.66
temporal + dynamics with P-ES	**86.57**	**126.14**	79.73	**8.14**	**78.07**	**4.96**

We use the same validation and holdout splits for the 3D Poses in the Wild (3DPW) [191] and the Penn Action datasets [305] as in Kanazawa et al. [156]. Before testing results on the holdout set, smoothing parameters were chosen from $\sigma \in [0.5, 1, 2, 3, 4]$ frames, and $c \in [0.0, 0.2, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$ to maximize average validation accuracy measured by the key-point accuracy (PCK) metric.

Holdout test set performance is given in Table 5.1. We evaluate performance using the same metrics as reported in [156], and the same code to calculate these metrics. All metrics are defined per video, and averaged over all videos. A description of each metric is given here; See [17, 156] for further explanation:

- Percentage key points (PCK): percentage of 2D key points that fall within $\alpha \cdot \max\{h, w\}$ of the labeled key point, where h and w parameters of a per-frame tight bounding box around the entire person; here $\alpha = 0.05$.
- Mean per joint position error (MPJPE): Mean euclidean distance of predicted to ground truth joint, averaged over joints in the human pose model (calculated after aligning root joints), measured in millimeters.
- Mean per joint position error after Procrustes alignment (PA-MPJPE): MPJPE after alignment to the ground truth by Procrustes alignment method, measured in millimeters.
- Acceleration Error (Accel Err): defined in [156] as "the average difference between ground truth 3D acceleration and predicted 3D acceleration of each joint in mm/s^2."
- Acceleration (Accel) For 2D datasets, measures "acceleration in mm/s^2" [156]. Note that this metric is only useful in conjunction with other metrics, as a baseline constant predictor would achieve 0 acceleration. However, for predictions that also do well on PCK, lower acceleration is more meaningful.

Table 5.2: Optimal hyperparameter values on validation set per predictor for 3D Poses in the Wild (3DPW) and the Penn Action datasets.

	per-frame		temporal + dynamics	
	3DPW	PA	3DPW	Penn Action
σ	2	3	2	2
c	0.5	0.4	0.8	1.0

While we optimized P-ES parameters according to the PCK metric, P-ES improves performance in both models, across all metrics. Smoothing confers greater gains in the time-agnostic per-frame model than the temporal dynamics model. Interestingly, the "temporal model" without human dynamics does worse in almost all metrics than the "per-frame" model. This underscores our motivation that temporal information must be encoded with care, as well as our claim that P-ES is a suitable baseline for such tasks.

The optimal hyperparameter pairs chosen are given in Table 5.2. For both models, the optimal σ was around 2 frames, but the optimal c for the per-frame predictions was much smaller for the per-frame model (avg. 0.45) than for the model that already incorporated temporal structure (avg. 0.9). This may be because predictions for the temporal model are smoother, so that we do not alter the signal in predictions as much with P-ES.

In summary, P-ES confers performance gains to both the per-frame model and the more accurate temporal and dynamics model, showing that P-ES can improve performance even when the base estimator is a complex model incorporating locality.

Predicting House Price from Attributes

The usefulness of applying semi-parametric techniques merging feature-based prediction and spatial regularization in predicting house prices has been documented from many perspectives (see, e.g. [54, 55, 68, 95]). This motivates house price prediction as a domain in which to compare the performance of P-ES and other methods exploiting spatial consistency. Using data on house sales from the Zillow Transaction and Assessment Database (ZTRAX) [312], we first demonstrate the effectiveness of P-ES on various machine learning regression methods (Figure 5.2), and then in comparison to standard semi-supervised learning techniques (Table 5.3).

In this experiment we predict sale prices y of single family homes, given features x about the homes (e.g., number of bedrooms, year of home sale, etc.).[1] Location t is the latitude and

[1]Features included were: year built (from 2010), number of stories, number of rooms, number of bedrooms, number of baths, number of partial baths, size (sqft), whether there is heating, whether there is air conditioning, the contract year, the contract month, and whether the home was new; location was encoded as the latitude and longitude of the home, and target label is the most recent sale price of the home. We drop any instances missing values for any of these features. The results and opinions are those of the authors

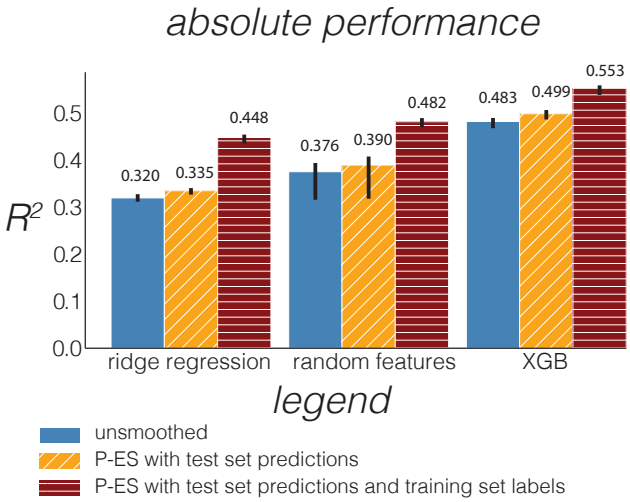


Figure 5.2: P-ES performance for various base predictors. Bars denote average performance; vertical black lines show minima and maxima over 10 trials. All six average relative differences (unsmoothed minues smoothed) are positive with p-value $< 1e^{-4}$, those that include the training set have p-value $< 1e^{-7}$.

longitude of the homes. After preprocessing to restrict to single family homes and only most recent sales, the dataset contains roughly 600,000 home sales at unique locations. We test three diverse machine learning models: ridge regression, random feature regression [232], and gradient boosted decision trees (XGB) [64], with and without post-estimation smoothing. For all three resulting models, we chose parameters jointly over the model parameters and P-ES parameters to maximize validation set accuracy, measured in R^2, the percent of label variation explained by the predictions.

Figure 5.2 shows the holdout test set performance of smoothed and unsmoothed models for the three machine learning algorithms. Training, validation, and test sets are of size $n = 20,000$ each. We performed this experiment over 10 random data draws. Smoothing is performed with respect to just the predictions, as well as with respect to the concatenated training set labels and test set predictions (validated on training set labels and validation set predictions).

For all three methods, applying P-ES with the test set predictions improves accuracy over the original predictions. Smoothing with the training points, in the spirit of semi-supervised learning, boosts accuracy further, as we might expect since there is no estimation error for the training labels.

Table 5.3 shows a comparison to alternative methods for reducing variance or inducing spatial consistency: kernel smoothing based only on the training set labels (without predictions), Gaussian process regression (GPR), Laplacian regularized least squares (LapRLS),

of this work and do not reflect the position of Zillow Group.

Table 5.3: Comparison with nonparametric and semi-supervised methods, for 10 random trials with train, validation, and holdout sets of size $n = 10,000$. Values in parentheses denote standard deviations.

method	model function (f) or post-processing (pp) form	average holdout accuracy mean (std) in r^2	average runtime mean (std) in secs
kernel smoothing	$f(t, y)$	0.277 (0.185)	20.6 (0.2)
GPR (Kriging)	$f(t, y)$	0.386 (0.011)	1336.1 (4.6)
LapRLS [30]	$f(x, t, y)$	0.452 (0.012)	1683.6 (12.4)
XGB	$f(x, y)$	0.458 (0.014)	7.8 (0.1)
XGB + shrinkage	$pp(t, \widehat{y}_{\text{XGB}})$	0.457 (0.014)	+ 0.0 (0.0)
XGB + P-ES	$pp(t, y, \widehat{y}_{\text{XGB}})$	0.526 (0.015)	+ 27.0 (0.2)
HEM [310]	$f(t, y, \widehat{y}_{\text{XGB}})$	0.544 (0.015)	+ 898.6 (4.8)
HEM [310] + P-ES	$pp\left(t, y, f(t, y, \widehat{y}_{\text{XGB}})\right)$	0.546 (0.015)	+ 287.8 (8.0)

a variance reducing shrinkage estimator (taking $S = \delta \cdot (\frac{1}{n}\vec{1}\vec{1}^\top) + (1 - \delta)I$), and Gaussian harmonic energy minimization (HEM) (see Section 5.1 for more details on these methods).

Timing results underscore that P-ES is a fast way to incorporate spatial structure (it incurs an $\mathcal{O}(n^2)$ additional runtime as opposed to $\mathcal{O}(n^3)$ for GPR, LapRLS, and HEM). The high variance in performance of kernel smoothing alone may be explained by inherent difficulties in choosing hyperparameters in semi-supervised settings, as discussed by [213]. Accuracy of post-processing with P-ES is within 1.2 standard deviations of the HEM method which takes roughly 30× as long to run in this instance over the chosen set of hyperparameters (grid of 9 σ values and 11 c values).

Runtimes for post-processing procedures are reported as the additional time compared to not running the post-processing procedure (on average, computing P-ES predictions takes 27 seconds on top of the 7.8 seconds to run XGB over multiple hyperparameter configurations). The runtime numbers reported in Table 5.3 are for solving the exact HEM and LapRLS problems, using the inverse and with as much shared computation as possible. We omit experimental comparison to LGC [307, 308] as the adaption from multi-class classification and ranking problems to regression problems is nontrivial, but we note that it is an iterative method where each iteration is $\mathcal{O}(n^2)$. The first iteration of the LGC algorithm is very similar to P-ES, so that similarity of accuracy of HEM and P-ES with the smoothing matrix defined in Eq. (5.4) suggests that a one-iteration approximation of these algorithms can be sufficient in some cases.

The last line in table Table 5.3 confirms that we get a very small increase in accuracy by smoothing the best SSL method; this is consistent with our understanding that this

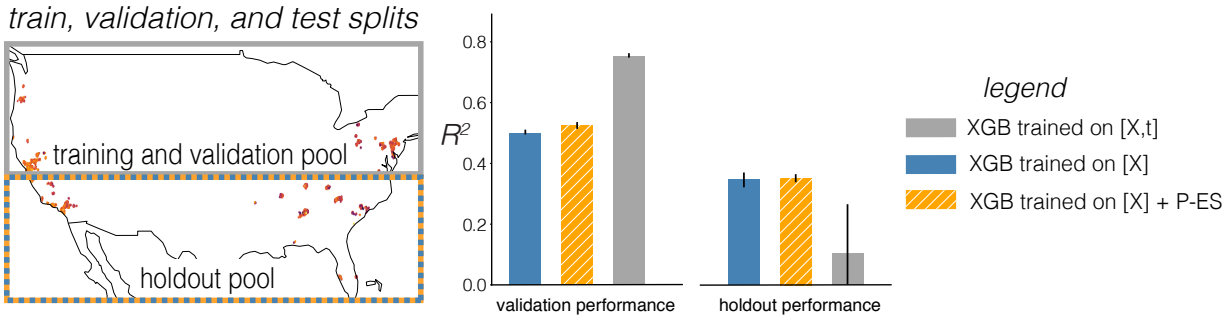


Figure 5.3: In-sample (unsmoothed) validation performance and out-of-sample holdout performance for different methods of incorporating spatial index variables for housing price prediction. Vertical lines on the barplots show minima and maxima over 10 trials.

smoothing operator acts similarly to the graph-defined HEM operator. The reduction in computation to apply P-ES significantly reduces the barrier to comparing to this family of algorithms as a baseline. Additionally, the computational speed of P-ES makes it much easier to explore different choices of the weight matrix $W(t;\sigma)$ from which to form the smoothing matrix; we consider this future work since such a matrix will likely be domain-specific.

We conclude our experiments with a final example of the danger of incorporating index variables with unique characteristics as predictive features. We compare the out-of-sample generalizability of a model $\widehat{y}(x,t)$ trained on the concatenated set of home attributes and geographic locations as features with that of the model from previous experiments $\widehat{y}(x)$ trained only on attributes with smoothing applied.

Figure 5.3 shows the results of these two approaches for a spatial extrapolation experiment, where a training set and validation set of size $20,000$ each are sampled from the northern U.S. and a holdout set of size $20,000$ is sampled from a disjoint southern segment.

We might expect that incorporating latitude and longitude as features could cause the first model to overfit, whereas distribution shift in neighborhood structure might negatively impact the learning of parameters for P-ES. Indeed, while the XGB predictor trained with locations and attributes (XGB trained on $[X,t]$) has better validation performance than the location-agnostic predictor (XGB trained on $[X]$), it performs much worse on the holdout set (0.10 vs 0.35 average R^2). However, applying P-ES does not degrade performance (0.002 average R^2 increase). In this scenario, the distribution shift with respect to t is so severe that it is possible to overfit by using t as a feature. Incorporating structure in t with P-ES, on the other hand, is much more robust to this distribution shift.

5.4 Conclusions

In this chapter we introduce post-estimation smoothing as a method for incorporating structural indices like time or location as valuable information sources in machine learning pre-

dictions. Theory and experiments underscore that P-ES is an effective and robust way to incorporate structured index variables in prediction, at much less cost than traditional semi-supervised methods.

The performance of P-ES depends on the accuracy of the original predictions. If predictions $\hat{y}$ are very far from y, smoothing is unlikely to remedy this. While decoupling smoothing from the original prediction may be limiting, we have shown that it can be advantageous when viewing P-ES as a diagnostic method or a baseline with which to compare more complex methods.

The results detailed in this chapter open a door for extensions to applications where index variables satisfy less physical notions of distance (e.g., word embeddings), and to analysis characterizing when decoupling local consistency and prediction can be close to the optimal integrated approach. Future work could also consider multivariate labels y with correlation among their elements. Lastly, it will likely be worthwhile to investigate more structured weight matrices W (other than the Nadaraya-Watson estimator) which form the basis of the smoothing matrix S. Due to the decoupling of smoothing from the original prediction, practitioners can try domain-specific weight matrices with marginal extra cost.

In summary, when the goal is to obtain accurate predictors, no data should be overlooked. However, index variables such as time and space should be incorporated with care. We propose that post-processing is a natural and effective way to utilize this structure, and show it is robust to different tasks, predictors, and sampling patterns.

Chapter 6

Limitations of "Data as Context"

While Chapters 4 and 5 provide encouraging results toward understanding and utilizing data as a form of context for statistical learning systems, it is important to recognize aspects of context that a data-centric view is not currently poised to address. Studying numerical allocations alone will not address issues of misuse of prediction systems, nor will it address issues of mismeasurement, when the data available does not fully or faithfully represent the salient factors of the phenomena studied. This chapter delineates potential limitations of identifying context within a data-centric framework. Understanding these limitations can help to clearly characterize the failure modes and opportunities to improve or augment data-centered analysis, essential steps toward developing even more nuanced and holistic methods of incorporating context in learning systems and frameworks.

6.1 Limitations of Numerical Allocation as Representation

Representation is a broad and often ambiguous concept [59], and numerical allocation is an imperfect proxy of representation or inclusion. One concern is that the data annotation process may systematically misrepresent individuals from certain groups or sup-populations. For example, Abbasi et al. [1] expose and quantify representational harms that can arise from stereotyping individuals in groups. If this is the case, solely optimizing allocations as in Chapter 4 to maximize accuracy with respect to those labels would not reflect an intended goal of high efficacy across all groups.

True representation thus requires that each data instance captures, or measures, the intended variables and their complexities in addition to having sufficient numerical allocations. The field of measurement theory provides a language by which to discuss these concepts of representation and suitability of measurements and their use in statistical analysis [22, 121, 122]. Specifically addressing learning contexts, Jacobs and Wallach [143] delineate key concepts of construct reliability and construct validity – commonly used in the quantitative social sciences – as they pertain to concerns of unfairness in algorithmic systems. For ex-

ample, in the group fairness setting introduced in Chapter 2, the definition of groups and assignment of instances to them requires measuring those demographic groups, which are often imperfectly described by a group structure at all. This is further complicated by the fact that "many demographic factors, such as race or gender, are themselves essentially contested constructs, with theoretical understandings that vary across cultures and over time" [143].

Focusing primarily on the numerical allocations of data from certain groups is thus only suitable to the extent that those data accurately measure the relevant phenomena for individuals in each group, *and* to the extent that the definition and measurement of those groups are valid. Understanding and accounting for key concepts and tools in measurement theory offers an opportunity to frame, extend, and possibly contest specific uses of data as context.

6.2 Whose Context do Data Reflect?

While data can encode certain types of context about a prediction problem, it is important to understand the ways in which data *have* social context: be that political, historical, or otherwise. As Boyd and Crawford [45] explain, "data are not generic. There is value to analyzing data abstractions, yet retaining context remains critical, particularly for certain lines of inquiry."

For example, relying on neighborhood effects for predicting house prices may increase accuracy on an existing evaluation set, but could also exacerbate deflation of home prices in historically undervalued neighborhoods, similarly to examples seen in Chapter 2. Historical context salient at the time of data creation or publication may get lost over time as datasets are reused or repurposed. Radin [229] details that as the Pima Indian Diabetes Dataset (PIDD) became widely used as a benchmark dataset in machine learning, its use became largely detached from the subjects and community from whom the data was collected. Couldry and Mejias [77] use the term "data colonialism" to illustrate the political and social disconnect between ways in which data is collected, processed, analysed, and shared, and the people whom that data is meant to represent.

Understanding and accounting for the context of data goes hand in hand with utilizing the context within it to the extent possible in statistical models. For example, results in Chapter 5 show that structural index variables can mislead predictors to rely on trends that do not generalize if used naively in the predictive model. Similarly, different data types and sources may differ in quality across regions, as we discuss for satellite imagery in Chapter 8. Understanding and delineating these characteristics of data with statistical tools can aid reliable reporting of machine learning model performance. More importantly, we must realize and account for the fact that the application of these statistical tools to real-world settings occurs within a broader social context beyond what can be encoded in numerical data.

6.3 Additional Data Collection is Not a Panacea

The results in this part of this thesis indicate the importance of having suitable representation in both the number of data points from all salient data sources or groups (Chapter 4) and utilizing the information carried in the feature representations of each instance (Chapter 5). At first glance, this might seem to indicate that a simple remedy to poor performance of a learning system would be to collect more data, either by collecting more samples (increasing n), or collecting more information about each data point (increasing d). But this is not necessarily the case. As Borgman [41] puts it quite simply, "having the right data is usually better than having more data."

While having more high quality data is generally preferable, assuming that just collecting more data will easily or necessarily remedy undesirable effects of machine learning systems can be naive. When machine learning systems fail to represent subpopulations due to measurement or construct validity issues discussed above, more comprehensive interventions are needed [1, 143]. Moreover, naive targeted data collection attempts can present undue burdens of surveillance or skirt consent [220]. Modeling characteristics of data from statistical perspectives as in Chapters 4 and 5 is thus complementary to current and emerging practices for ethical, contextualized data collection and curation [2, 90, 106, 150]. Integrating these perspectives together will be an important step toward operationalizing our findings.

Part III

Context-Driven Applications in Remote Sensing and Machine Learning

Data is the lifeblood of decision-making and the raw material for accountability... New sources of data - such as satellite data -, new technologies, and new analytical approaches, if applied responsibly, can enable more agile, efficient and evidence-based decision-making and can better measure progress on the Sustainable Development Goals (SDGs) in a way that is both inclusive and fair.

— United Nations Global Issues, "Big Data for Sustainable Development" [34]

Whereas the first two parts of this thesis center around the themes of intent and impact (Part I) and context (Part II) in general machine learning systems, in Part III we detail two applications of these themes to combining machine learning and remote sensing. These specific applications are indicative of a broader class of problems in using machine learning to aid environmental monitoring, a compelling area for jointly achieving social benefit and leveraging structures in remotely sensed data to anchor context-aware prediction in practice.

In Chapter 7, we study the problem of robotic source localization. We develop statistically efficient trajectory planning methods for robots to find sources of environmental radiation. Our solution makes use of precise measurement structures of the on-board radiation sensors to optimize a sampling procedure for the robot, despite physical travel constraints. In Chapter 8, we detail a large scale project that combines machine learning and satellite imagery, where structures in the imagery data and downstream use cases guide algorithmic innovations that increase accessibility and computational efficiency of this technology.

Applied perspectives like these shed light on how the themes of intent, impact, and context can manifest in real-world machine learning systems. Together, Chapters 7 and 8 evidence that structural domain knowledge about our application space—including properties of our data sensing process and intended downstream decisions or use cases—can be leveraged to delineate and optimize for our intent with targeted algorithmic innovations.

Chapter 7

A Successive-Elimination Approach to Adaptive Robotic Source Seeking

This chapter is based on the paper "A successive-elimination approach to adaptive robotic sensing" [243], written in collaboration with David Fridovich-Keil, Max Simchowitz, Benjamin Recht, and Claire Tomlin.[1]

In this chapter, we study an adaptive source seeking problem, in which a mobile robot must identify the strongest emitter(s) of a signal in an environment with background emissions. Background signals may be highly heterogeneous and can mislead algorithms that are based on receding horizon control. We propose AdaSearch, a general algorithm for adaptive source seeking in the face of heterogeneous background noise. AdaSearch combines global trajectory planning with principled confidence intervals in order to concentrate measurements in promising regions while guaranteeing sufficient coverage of the entire area. Theoretical analysis shows that AdaSearch confers gains over a uniform sampling strategy when the distribution of background signals is highly variable. Simulation experiments demonstrate that when applied to the problem of radioactive source-seeking, AdaSearch outperforms both uniform sampling and a receding time horizon information maximization approach based on previous literature. We also demonstrate AdaSearch in hardware, providing further evidence of its potential for real-time implementation.

7.1 Background

Robotic source seeking is a problem domain in which a mobile robot must traverse an environment to locate the maximal emitters of a signal of interest, usually in the presence of background noise. Adaptive source seeking involves adaptive sensing and active infor-

[1]Full citation: Esther Rolf et al. "A successive-elimination approach to adaptive robotic source seeking". In: *IEEE Transactions on Robotics* 37.1 (2021), pp. 34–47. DOI: 10.1109/TRO.2020.3005537

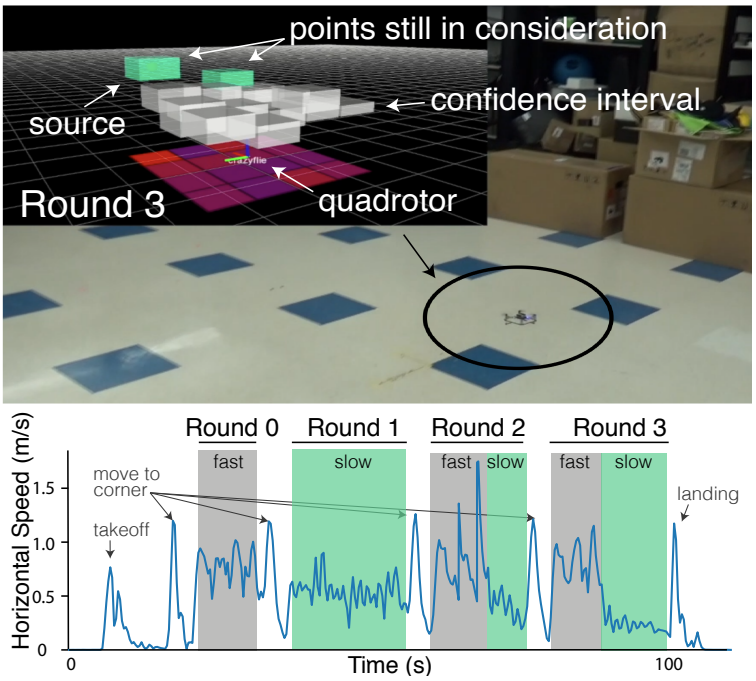


Figure 7.1: A Crazyflie 2.0 quadrotor in a motion capture room during a hardware demo of `AdaSearch` (round 3 of the algorithm). Top: confidence intervals for estimated radioactivity (simulated). Bottom: horizontal speed over time, indicating slow and fast sections in which `AdaSearch` allocates more and fewer sensor measurements, respectively. In later rounds, more time is spent measuring points that are still in consideration (in teal). ©2021 IEEE

mation gathering, and encompasses several well-studied problems in robotics, including the rapid identification of accidental contamination leaks and radioactive sources [196, 285], and finding individuals in search and rescue missions [133]. We consider a specific motivating application of radioactive source seeking (RSS), in which an unmanned aerial vehicle (UAV) (Figure 7.1) must identify the k-largest radioactive emitters in a planar environment, where k is a user-defined parameter. RSS is a particularly interesting instance of source seeking due to the challenges posed by the highly heterogeneous background noise [216].

A well-adopted methodology for approaching source seeking problems is information maximization, in which measurements are collected in the most promising locations following a receding planning horizon. Information maximization favors measuring regions that are likely to contain the highest emitters and avoids wasting time elsewhere. However, when operating in real-time, computational constraints necessitate approximations such as limits on planning horizon and trajectory parameterization. These limitations scale with size of the search region and complexity of the sensor model and may cause the algorithm to be excessively greedy, spending extra travel time tracking down false leads.

To overcome these limitations, we introduce `AdaSearch`, a successive-elimination frame-

work for general source seeking problems with multiple sources, and demonstrate it within the context of RSS. `AdaSearch` explicitly maintains confidence intervals over the emissions rate at each point in the environment. Using these confidence intervals, the algorithm identifies a set of candidate points likely to be among the top-k emitters, and eliminates points that are not. Rather than iteratively planning for short, receding time horizons, `AdaSearch` repeats a *fixed, globally-planned path*, adjusting the robot's speed in real-time to focus measurements on promising regions. This approach offers coverage of the full search space while affording an adaptive measurement allocation in the spirit of information maximization. By maintaining a single fixed, global path, `AdaSearch` reduces the online computational overhead, yielding an algorithm easily amenable to real-time implementation.

In this chapter:

- We present `AdaSearch`, a general framework for designing efficient sensing trajectories for robotic source seeking problems.
- We derive theoretical runtime analysis of `AdaSearch` and compare to that of a naive, uniform sampling baseline which follows the same fixed global path but moves at constant speed.
- We detail simulation experiments for RSS evaluating `AdaSearch` in comparison with a uniform baseline and information maximization.

Our theoretical analysis sharply quantifies `AdaSearch`'s improvement over its uniform sampling analog. Experiments validate this finding in practice, and also show that `AdaSearch` outperforms a custom implementation of information maximization tailored to the RSS problem. Together, these results suggest that the accuracy and efficient runtime of `AdaSearch` are robust to heterogeneous background noise, which stands in contrast to existing alternative methods. This robustness is particularly valuable in real-world applications where the exact distribution of background signals in the environment is likely unknown.

This chapter is organized as follows. The remainder of this section presents a brief survey of related literature. In Section 7.2, we provide a formal statement of the source seeking problem and introduces our solution, `AdaSearch`. In Section 7.3, we consider a radioactive source seeking (RSS) case study and develop two appropriate sensing models which allow us to apply `AdaSearch` to RSS. In Section 7.4, we analyze the theoretical runtime complexity of `AdaSearch` and its uniform sampling analog for the RSS problem. In Section 7.5, we present simulation experiments which corroborate these theoretical results. A hardware demonstration provides further evidence of `AdaSearch`'s potential for real-time application. In Section 7.6, we suggests a number of extensions and generalizations to `AdaSearch`, and in Section 7.7, we conclude with a summary of our results.

Related Work

There is a breadth of existing work related to source seeking. Much of this literature, particularly when tailored to robotic applications, leverages some form of information maximization,

often using a Gaussian process prior. However, our own work is inspired by approaches from the pure exploration multi-armed bandit literature, even though bandits are not typically used to model physical sensing problems with realistic motion constraints. We survey the most relevant work in both information maximization and multi-armed bandits below.

Information Maximization Methods

A popular approach to active sensing and source seeking in robotics, e.g. in active mapping [44] and target localization [199], is to choose trajectories that maximize a measure of information gain [19, 44, 58, 173, 187]. In the specific case of linear Gaussian measurements, Atanasov et al. [11] formulate the informative path planning problem as an optimal control problem that affords an offline solution. Similarly, Lim, Hsu, and Lee [176] propose a recursive divide and conquer approach to active information gathering for discrete hypotheses, which is near-optimal in the noiseless case.

Planning for information maximization-based methods typically proceeds with a receding horizon [19, 116, 192, 193, 195]. For example, Ristic, Morelande, and Gunatilaka [239] formulate information gathering as a partially observable Markov decision process and approximate a solution using a receding horizon. Marchant, Roman and Ramos, Fabio [193] combine upper confidence bounds (UCBs) at potential source locations with a penalization term for travel distance to define a greedy acquisition function for Bayesian optimization. Their subsequent work [192] reasons at the path level to find longer, more informative trajectories. Noting the limitations of a greedy receding horizon approach, Hitz et al. [132] incentivize exploration by using a look-ahead step in planning. Though similar in spirit to these information seeking approaches, a key benefit of `AdaSearch` is that it is not greedy, but rather iterates over a global path.

Information maximization methods typically require a prior distribution on the underlying signals. Many active sensing approaches model this prior as being drawn from a Gaussian process (GP) over an underlying space of possible functions [19, 193, 199], tacitly enforcing the assumption that the sensed signal is smooth [193]. In certain applications, this is well motivated by physical laws, e.g. diffusion [132]. However, GP priors may not reflect the sparse, heterogeneous emissions encountered in radiation detection and similar problem settings.

Multi-Armed Bandit Methods

`AdaSearch` draws heavily on confidence-bound based algorithms from the pure exploration bandit literature [13, 102, 144]. In contrast to these works, our method explicitly incorporates a physical sensor model and allows for efficient measurement allocation despite the physical movement constraints inherent to mobile robotic sensing. Other works have studied spatial constraints in the online, "adversarial" reward setting [47, 168]. Baykal et al. [28] consider spatial constraints in a persistent surveillance problem, in which the objective is to observe as many events of interest as possible despite unknown, time-varying event statistics.

Recently, Ma, Garnett, and Schneider [187] encode a notion of spatial hierarchy in designing informative trajectories, based on a multi-armed bandit formulation. While [187] and `AdaSearch` are similarly motivated, hierarchical planning can be inefficient for many sensing models, e.g. for short-range sensors, or signals that decay quickly with distance from the source.

Bandit algorithms are also studied from a Bayesian perspective, where a prior is placed over underlying rewards. For example, Srinivas et al. [266] provide an interpretation of the GP upper confidence bound (GP-UCB) algorithm in terms of information maximization. `AdaSearch` does not use such a prior, and is more similar to the lower and upper confidence bound (LUCB) algorithm [154], but opts for successive elimination over the more aggressive LUCB sampling strategy for measurement allocation.

A multi-armed bandit approach to active exploration in Markov decision processes (MDPs) with transition costs is studied in [272], which details trade-offs between policy mixing and learning environment parameters. This work highlights the potential difficulties of applying a multi-armed bandit approach while simultaneously learning robot policies. In contrast, we show that decoupling the use of active learning during the sampling decisions from a fixed global movement path confers efficiency gains under reasonable environmental models.

Other Source Seeking Methods

Other notable extremum seeking methods include those that emulate gradient ascent in the physical domain [36, 197, 225], take into account specific environment signal characteristics [159], or are specialized for particular vehical dynamics [198]. Modeling emissions as a continuous field, gradient-based approaches estimate and follow the gradient of the measured signal toward local maxima [36, 197, 225]. One of the key drawbacks of gradient-based methods is their susceptibility to finding local, rather than global, extrema. Moreover, the error margin on the noise of gradient estimators for large-gain sensors measuring noisy signals can be prohibitively large [284], as is the case in RSS. Khodayi-mehr, Aquino, and Zavlanos [159] handle noisy measurements by combining domain, model, and parameter reduction methods to actively identify sources in steady state advection-diffusion transport system problems such as chemical plume tracing. Their approach combines optimizing an information theoretic quantity based on these approximations with path planning in a feedback loop, specifically incorporating the physics of advection-diffusion problems. In comparison, we consider planning under specific sensor models, and plan motion path and optimal measurement allocation separately.

7.2 AdaSearch Planning Strategy

Problem Statement

We consider signals (e.g. radiation) which emanate from a finite set of environment points $\mathcal{S}$. Each point $x \in \mathcal{S}$ emits signals $\{\mathbf{X}_t(x)\}$ indexed by time t with means $\mu(x)$, independent and identically distributed over time. Our aim is to correctly and exactly discern the set of the k points in the environment that emit the maximal signals:

$$\mathcal{S}^*(k) = \underset{S' \subseteq \mathcal{S}, |S'|=k}{\operatorname{argmax}} \sum_{x \in S'} \mu(x) \tag{7.1}$$

for a pre-specified integer $1 \leq k \leq |\mathcal{S}|$. Throughout, we assume that the set of maximal emitters $\mathcal{S}^*(k)$ is unique.

In order to decide which points are maximal emitters, the robot takes sensor measurements along a fixed path $\mathcal{Z} = (z_1, \ldots, z_n)$ in the robot's configuration space. Measurements are determined by a known *sensor model* $h(x, z)$ that describes the contribution of environment point $x \in \mathcal{S}$ to a sensor measurement collected from sensing configuration $z \in \mathcal{Z}$. We consider a linear sensing model in which the total observed measurement at time t, $\mathbf{Y}_t(z)$, taken from sensing configuration z, is the weighted sum of the contributions $\{\mathbf{X}_t(x)\}$ from all environment points:

$$\mathbf{Y}_t(z) = \sum_{x \in \mathcal{S}} h(x, z)\mathbf{X}_t(x). \tag{7.2}$$

Note that while $h(x, z)$ is known, the $\{\mathbf{X}_t\}$ are unknown and must be estimated via the observations $\{\mathbf{Y}_t\}$.

The path of sensing configurations, $\mathcal{Z}$, should be as short as possible, while providing sufficient information about the entire environment. This may be expressed as a condition on the minimum aggregate sensitivity α to any given environment point x over the sensing path $\mathcal{Z}$:

$$\sum_{z \in \mathcal{Z}} h(x, z) \geq \alpha > 0 \quad \forall x \in \mathcal{S}. \tag{7.3}$$

Moreover, we need to disambiguate between contributions from different environment points $x, x' \in \mathcal{S}$. We define the matrix $H \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{Z}|}$ that encodes the sensitivity of each sensing configuration $z_j \in \mathcal{Z}$ to each point $x_i \in \mathcal{S}$, so that $H_{ij} := h(x_i, z_j)$. Disambiguation then translates to a rank constraint $\operatorname{rank}(H) \geq |\mathcal{S}|$, enforcing invertibility of HH^T. In Section 7.3, we define two specific sensitivity functions that we consider in the context of the RSS problem. In Section 7.6, we discuss sensitivity functions that may arise in other application domains.

Algorithm 1: `AdaSearch`

Input Candidate points of interest $\mathcal{S}$; sensing path of configurations $\mathcal{Z}$; number of points of interest k; minimum measurement duration τ_0; procedure for constructing $[\mathrm{LCB}_i(x), \mathrm{UCB}_i(x)]$ (e.g., as in Section 7.3); confidence parameter δ_{tot}.
Initialize $\mathcal{S}_0^{\mathrm{top}} = \emptyset, \mathcal{S}_0 = \mathcal{S}$.
For rounds $i = 0, 1, 2, \ldots$
  If $\mathcal{S}_i = \emptyset$, **Return** $\mathcal{S}_i^{\mathrm{top}}$.
  Choose configuration subset $\mathcal{Z}_i \subseteq \mathcal{Z}$ that is informative about environment points $x \in \mathcal{S}_i$.
  Execute a trajectory along path $\mathcal{Z}$ that spends time $\tau_i = \tau_0 \cdot 2^i$ at each $z \in \mathcal{Z}_i$ and time τ_0 at each $z \in \mathcal{Z} \setminus \mathcal{Z}_i$. Meanwhile, observe signal measurements according to (7.2).
  Update $[\mathrm{LCB}_i(x), \mathrm{UCB}_i(x)]$ for all $x \in \mathcal{S}$.
  Update Augment $\mathcal{S}_i^{\mathrm{top}}$ according to (7.4), and prune $\mathcal{S}_i$ according to (7.5).

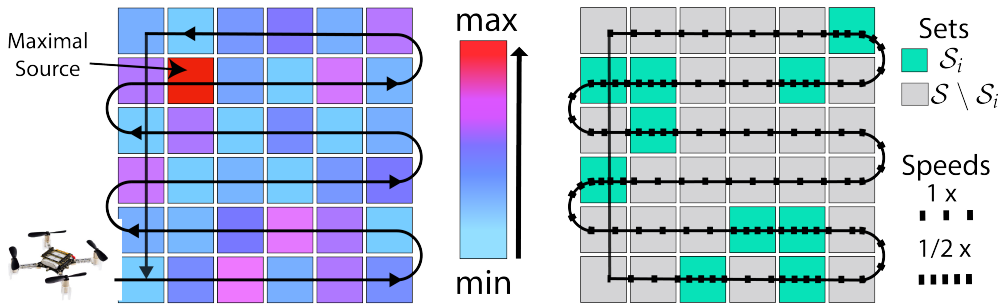


Figure 7.2: (left) Raster path $\mathcal{Z}$ over an example grid environment of size 6×6. The path ensures that each point is sufficiently measured during each round. (right) Illustrative trajectory for round $i = 1$. Dots indicate measurements. `AdaSearch` slows to take twice as many measurements over points $x \in \mathcal{S}_i$. ©2021 IEEE

The `AdaSearch` Algorithm

`AdaSearch` (Algorithm 1) concentrates measurements in regions of uncertainty until we are confident about which points belong to $S^*(k)$. At each round i, we maintain a set of environment points $\mathcal{S}_i^{\mathrm{top}}$ that we are confident are among the top-k, and a set of candidate points $\mathcal{S}_i$ about which we are still uncertain. As the robot traverses the environment, new sensor measurements allow us to update the lower and upper confidence bounds

$$[\mathrm{LCB}_i(x), \mathrm{UCB}_i(x)] \text{ for each } x \in \mathcal{S}_i$$

and prune the uncertainty set $\mathcal{S}_i$. The procedure for constructing these intervals from observations should ensure that for every $x \in \mathcal{S}_i$, $\mathrm{LCB}_i(x) \leq \mu(x) \leq \mathrm{UCB}_i(x)$ with high

probability. In Section 7.3, we detail the definition of these confidence intervals under different sensing models.

Using the updated confidence intervals, we expand the set $\mathcal{S}_{i+1}^{\text{top}}$ and prune the set $\mathcal{S}_{i+1}$. We add to the top-set all points $x \in \mathcal{S}_i$ whose lower confidence bounds exceed the upper confidence bounds of all but $(k - |\mathcal{S}_i^{\text{top}}|)$ points in $\mathcal{S}_i$; formally,

$$\mathcal{S}_{i+1}^{\text{top}} \leftarrow \mathcal{S}_i^{\text{top}} \cup \{x \in \mathcal{S}_i \mid \text{LCB}_i(x) > (k - |\mathcal{S}_i^{\text{top}}| + 1)\text{-th largest UCB}_i(x'), x' \in \mathcal{S}_i\}. \quad (7.4)$$

In (7.4), we need not re-evaluate confidence intervals for points x already in $\mathcal{S}_i^{\text{top}}$ when producing the set $\mathcal{S}_{i+1}^{\text{top}}$, and can only consider new points. This is explained in the proof of correctness (Lemma 7.1) and in Theorem 7.1, where we show that with high probability, points $x \notin \mathcal{S}^*(k)$ are never incorrectly added to the estimate of the top set $\mathcal{S}_i^{\text{top}}$.

Next, the points added to $\mathcal{S}_{i+1}^{\text{top}}$ are removed from $\mathcal{S}_{i+1}$, since we are now certain about them. Additionally, we remove all points in $\mathcal{S}_i$ whose upper confidence bound is lower than the lower confidence bounds of at least $k - |\mathcal{S}_{i+1}^{\text{top}}|$ points in $\mathcal{S}_i$. The set $\mathcal{S}_{i+1}$ is defined constructively as:

$$\mathcal{S}_{i+1} \leftarrow \{x \in \mathcal{S}_i \mid x \notin \mathcal{S}_{i+1}^{\text{top}} \text{ and } \text{UCB}_i(x) \geq (k - |\mathcal{S}_{i+1}^{\text{top}}|)\text{-th largest LCB}_i(x'), x' \in \mathcal{S}_i\}. \quad (7.5)$$

Trajectory Planning for `AdaSearch`

The update rules (7.4) and (7.5) only depend on confidence intervals for points $x \in \mathcal{S}_i$. At each round, `AdaSearch` chooses a subset of the sensing configurations $\mathcal{Z}_i \subseteq \mathcal{Z}$ which are informative to disambiguating the points remaining in $\mathcal{S}_i$.

`AdaSearch` defines a trajectory by following the fixed path $\mathcal{Z}$ over all configurations, slowing down to spend time $2^i\tau_0$ at informative configurations in $\mathcal{Z}_i$, and spending minimal time τ_0 at all other configurations in $\mathcal{Z} \setminus \mathcal{Z}_i$. Doubling the time spent at each $z \in \mathcal{Z}_i$ in each round amortizes the time spent traversing the entire path $\mathcal{Z}$. For omnidirectional sensors, a simple raster pattern (Figure 7.2) suffices for $\mathcal{Z}$ and choosing $\mathcal{Z}_i$ is relatively straightforward. Finally, we remark that one can set the time per configuration $z \in \mathcal{Z}_i$ as $\tau_i = \tau_0 c^i$ for any constant $c > 1$; this yields similar theoretical guarantees, and constants other than $c = 2$ may confer slight benefits in practice.

We could also design a trajectory that visits the $z \in \mathcal{Z}_i$ and minimizes total travel distance each round, e.g. by approximating a traveling salesman solution. In practice, this would improve upon the runtime of the fixed raster path suggested above. In this chapter, we use a raster pattern to emphasize the gains due to our main algorithmic components: global coverage and adaptive measurement allocation.

Correctness

Lemma 7.1 establishes that the two update rules above guarantee the overall correctness of `AdaSearch`, whenever the confidence intervals $[\text{LCB}_j(x), \text{UCB}_j(x)]$ actually contain the correct mean $\mu(x)$:

Lemma 7.1 (Sufficient condition for correctness)**.** *For each round $i \geq 0$, $\mathcal{S}_i^{\text{top}} \cap \mathcal{S}_i = \emptyset$. Moreover, whenever the confidence intervals satisfy the coverage property:*

$$\text{for all rounds } j \leq i \quad \text{and all } x \in \mathcal{S}_j, \quad \mu(x) \in [\text{LCB}_j(x), \text{UCB}_j(x)], \tag{7.6}$$

then $\mathcal{S}_{i+1}^{\text{top}} \subseteq \mathcal{S}^*(k) \subseteq \{\mathcal{S}_{i+1}^{\text{top}} \cup \mathcal{S}_{i+1}\}$. If (7.6) holds for all rounds i, then `AdaSearch` *terminates and correctly returns $\mathcal{S}^*(k)$.*

Proof of Lemma 7.1. First, we verify that for each round $i \geq 0$, $\mathcal{S}_i^{\text{top}} \cap \mathcal{S}_i = \emptyset$. At round $i = 0$, $\mathcal{S}_i^{\text{top}} = \emptyset$, so the bound holds immediately. Suppose by an inductive hypothesis that $\mathcal{S}_i^{\text{top}} \cap \mathcal{S}_i = \emptyset$ for some $i \geq 0$. Then, for any $x \in \mathcal{S}_{i+1}^{\text{top}}$, we have two cases:

(a) $x \in \mathcal{S}_i^{\text{top}}$. Then, $x \notin \mathcal{S}_i$ by the inductive hypothesis, and $\mathcal{S}_i \supset \mathcal{S}_{i+1}$ by (7.5).

(b) x is added to $\mathcal{S}_{i+1}^{\text{top}}$ via (7.4). Then $x \notin \mathcal{S}_{i+1}$ by (7.5).

Next, we verify that if the confidence intervals are correct in all rounds leading up to round i, i.e.

$$\text{for all rounds } j \leq i \text{ and all } x \in \mathcal{S}_j, \quad \mu(x) \in [\text{LCB}_j(x), \text{UCB}_j(x)], \tag{7.7}$$

then $\mathcal{S}_{i+1}^{\text{top}} \subset \mathcal{S}^*(k)$, and $\mathcal{S}^*(k) \subset \mathcal{S}_{i+1}^{\text{top}} \cup \mathcal{S}_{i+1}$. We again use induction. Initially, we have $\mathcal{S}_0^{\text{top}} = \emptyset \subset \mathcal{S}^*(k) \subset \mathcal{S} = \mathcal{S}_0$. Now, suppose that at round i, one has that $\mathcal{S}_i^{\text{top}} \subset \mathcal{S}^*(k)$, and $\mathcal{S}^*(k) \subset \mathcal{S}_i^{\text{top}} \cup \mathcal{S}_i$.

To show that $\mathcal{S}_{i+1}^{\text{top}} \subset \mathcal{S}^*(k)$, it suffices to show that if x is added to $\mathcal{S}_{i+1}^{\text{top}}$, then $x \in \mathcal{S}^*(k)$. By the inductive hypothesis there exists $k - |\mathcal{S}_i^{\text{top}}|$ elements of $\mathcal{S}^*(k)$ in $\mathcal{S}_i$. Hence, if x is added to $\mathcal{S}_{i+1}^{\text{top}}$, and if (7.7) holds, then

$$\begin{aligned}
\mu(x) &\geq \text{LCB}_i(x) \\
&> (k - |\mathcal{S}_i^{\text{top}}| + 1)\text{-th largest value of } \text{UCB}_i(x),\ x \in \mathcal{S}_i && \text{by (7.4)} \\
&\geq (k - |\mathcal{S}_i^{\text{top}}| + 1)\text{-th largest value of } \mu(x),\ x \in \mathcal{S}_i && \text{by (7.7)} \\
&\geq (k+1)\text{-th largest value of } \mu(x), x \in \mathcal{S}_i \cup \mathcal{S}_i^{\text{top}}.
\end{aligned}$$

Hence, $\mu(x)$ is among the k largest values of $\mu(x)$ for $x \in \mathcal{S}_i \cup \mathcal{S}_i^{\text{top}}$. Since $\mathcal{S}^*(k) \subset \mathcal{S}_i \cup \mathcal{S}_i^{\text{top}}$, we therefore have that $x \in \mathcal{S}^*(k)$.

Similarly, to show $\mathcal{S}^*(k) \subset \mathcal{S}_{i+1} \cup \mathcal{S}_{i+1}^{\text{top}}$, it suffices to show that if $x \in \mathcal{S}_i \setminus \mathcal{S}_{i+1}$, and $x \notin \mathcal{S}_{i+1}^{\text{top}}$, then $x \notin \mathcal{S}^*(k)$. For x such that $x \in \mathcal{S}_i \setminus \mathcal{S}_{i+1}$, and $x \notin \mathcal{S}_{i+1}^{\text{top}}$, it follows that

$$\begin{aligned}
\mu(x) &\leq \text{UCB}_i(x) \\
&< (k - |\mathcal{S}_{i+1}^{\text{top}}|)\text{-th largest value of } \text{LCB}_i(x),\ x \in \mathcal{S}_i && \text{by (7.5)} \\
&\leq (k - |\mathcal{S}_{i+1}^{\text{top}}|)\text{-th largest value of } \mu(x),\ x \in \mathcal{S}_i && \text{by (7.7)} \\
&\leq k\text{-th largest value of } \mu(x), x \in \mathcal{S}_i \cup \mathcal{S}_{i+1}^{\text{top}} \\
&\leq k\text{-th largest value of } \mu(x), x \in \mathcal{S},
\end{aligned}$$

hence $\mu(x) \notin \mathcal{S}^*(k)$.

Finally, we verify that if (7.7) holds at each round, then at the termination round i_{fin}, $\mathcal{S}_{i_{\text{fin}}} = \emptyset$, so that $\mathcal{S}^{\text{top}}_{i_{\text{fin}}} \subset \mathcal{S}^*(k) \subset \mathcal{S}^{\text{top}}_{i_{\text{fin}}} \cup \mathcal{S}_{i_{\text{fin}}} = \mathcal{S}^{\text{top}}_{i_{\text{fin}}}$, and thus $\mathcal{S}^*(k) = \mathcal{S}^{\text{top}}_{i_{\text{fin}}}$. □

The proof of Lemma 7.1 holds *for any* instantiation of Algorithm 1, regardless of the sensing model or the planning strategy. Lemma 7.1 provides a backbone upon which we construct a probabilistic correctness guarantee in Section 7.4. If the event (7.6) holds over all rounds with some probability $1 - \delta_{\text{tot}}$, then `AdaSearch` returns the correct set $\mathcal{S}^*(k)$ with the same probability $1 - \delta_{\text{tot}}$.

7.3 Radioactive Source Seeking with Poisson Emissions

While `AdaSearch` applies to a range of adaptive sensing problems, for concreteness we now refine our focus to the problem of radioactive source seeking (RSS) with an omnidirectional sensor. The environment is defined by potential emitter locations which lie on the ground plane, i.e. $x \in \mathcal{S} \subset \mathbb{R}^2 \times \{0\}$, and sensing configurations encode spatial position, i.e. $z \in \mathcal{Z} \subset \mathbb{R}^3$. Environment points emit gamma rays according to a Poisson process, i.e. $\mathbf{X}_t(x) \sim \text{Poisson}\,(\mu(x))$. Here, $\mu(x)$ corresponds to rate or intensity of emissions from point x.

Thus, the number of gamma rays observed over a time interval of length τ from configuration z has distribution

$$\mathbf{Y}_t(z) \sim \text{Poisson}\Big(\tau \cdot \sum_{x \in \mathcal{S}} h(x, z)\mu(x)\Big) \tag{7.8}$$

where $h(x, z)$ is specified by the sensing model. In the following sections, we introduce two sensing models: a pointwise sensing model amenable to theoretical analysis, and a more physically realistic sensing model for experiments.

In both settings, we develop appropriate confidence intervals for use in the `AdaSearch` algorithm. We introduce the specific path used for global trajectory planning, and conclude Section 7.3 with two benchmark algorithms to which we compare `AdaSearch`.

Pointwise Sensing Model

First, we consider a simplified sensing model, where the set of sensing locations $\mathcal{Z}$ coincides with the set $\mathcal{S}$ of all emitters, i.e. each $z \in \mathcal{Z}$ corresponds to exactly one $x \in \mathcal{S}$ and vice versa. The sensitivity function is defined as $h(x, z) := \mathbb{I}(x = z) = [1 \text{ if } x = z, 0 \text{ if } x \neq z]$.

Now we derive confidence intervals for Poisson counts observed according to this sensing model. Define $\mathbf{N}(x)$ to be the total number of gamma rays observed during the time interval of length τ spent at x. The maximum likelihood estimator (MLE) of the emission rate for point x is $\widehat{\mu}(x) = \frac{\mathbf{N}(x)}{\tau}$. In Section 7.A, we introduce the *bounding functions* $U_-(\cdot, \cdot)$ and

$U_+(\cdot,\cdot)$:

$$U_+(\mathbf{N},\delta) := 2\log(1/\delta) + \mathbf{N} + \sqrt{2\mathbf{N}\log(1/\delta)} \text{ and}$$
$$U_-(\mathbf{N},\delta) := \max\left\{0, \mathbf{N} - \sqrt{2\mathbf{N}\log(1/\delta)}\right\}.$$

Then for any $\lambda \geq 0$, $\mathbf{N} \sim \text{Poisson}(\lambda)$, and $\delta \in (0,1)$,

$$\mathbb{P}[U_-(\mathbf{N},\delta) \leq \lambda \leq U_+(\mathbf{N},\delta)] \geq 1 - 2\delta.$$

Let $\mathbf{N}_i(x)$ denote the number of gammas rays observed from emitter x during round i, so that $\mathbf{N}(x) \sim \text{Poisson}(\tau_i\mu(x))$. For any point $x \in \mathcal{S}_i$, the corresponding duration of measurement would be τ_i. The bounding functions above provide the desired confidence intervals for signals $\mu(x), \forall x \in \mathcal{S}_i$:

$$\text{LCB}_i(x) := \tfrac{1}{\tau_i}U_-\left(\mathbf{N}_i(x),\delta_i\right), \quad \text{UCB}_i(x) := \tfrac{1}{\tau_i}U_+\left(\mathbf{N}_i(x),\delta_i\right) . \tag{7.9}$$

This bound implies that the inequality $\tau_i\text{LCB}_i(x) \leq \tau_i\mu(x) \leq \tau_i\text{UCB}_i(x)$ holds with probability $1-2\delta_i$. Dividing by τ_i, we see that $\text{LCB}_i(x)$ and $\text{UCB}_i(x)$ are valid confidence bounds for $\mu(x)$.

The term δ_i can be thought of as an "effective confidence" for each interval that we construct during round i. In order to achieve the correctness in Lemma 7.1 with overall probability $1-\delta_{\text{tot}}$, we set the effective confidence δ_i at each round to be $\delta_i = \delta_{\text{tot}}/(4|\mathcal{S}|i^2)$ (see Lemma 7.4 in Section 7.B).

Physical Sensing Model

A more physically accurate sensing model for RSS reflects that the gamma ray counts at each location are sensitivity-weighted combinations of emissions from each environment point. Conservation of energy in free space allows us to approximate the sensitivity with an inverse-square law $h(x,z) := c/\|x-z\|_2^2$, with c a known, sensor-dependent constant. More sophisticated approximations are also possible [239].

Because multiple environment points x contribute to the counts observed from any sensor position z, the MLE $\hat{\mu}$ for the emission rates at all $x \in \mathcal{S}$ is difficult to compute efficiently. However, we can approximate it in the limit: $\frac{1}{\sqrt{\tau}}\text{Poisson}(\tau\mu) \xrightarrow{d} \mathcal{N}(\mu,\mu)$ as $\tau \to \infty$. Thus, we may compute $\hat{\mu}$ as the least squares solution:

$$\hat{\mu} = \arg\min_{\vec{\mu}} \|\tilde{H}^T\vec{\mu} - \vec{\mathbf{Y}}\|_2^2 , \tag{7.10}$$

where $\vec{\mu} \in \mathbb{R}^{|\mathcal{S}|}$ is a vector representing the mean emissions from each $x \in \mathcal{S}$, $\vec{\mathbf{Y}} \in \mathbb{R}^m$ is a vector representing the observed number of counts at each of m consecutive time intervals, and $\tilde{H} \in \mathbb{R}^{|\mathcal{S}|\times m}$ is a rescaled sensitivity matrix such that $\tilde{H}_{ij}$ gives the measurement-adjusted

sensitivity of the i^{th} environment point to the sensor at the j^{th} sensing position.[2] The resulting confidence bounds are given by the standard Gaussian confidence bounds:

$$[\mathrm{LCB}_i(x_k), \mathrm{UCB}_i(x_k)] := \hat{\mu}(x_k) \pm \alpha(\delta_i) \cdot \Sigma_{kk}^{1/2} \quad \text{where } \Sigma := (\tilde{H}\tilde{H}^T)^{-1} \tag{7.11}$$

$\alpha(\delta_i)$ controls the round-wise effective confidence widths in Eq. (7.11) as a function of the desired threshold probability of overall error, δ_{tot}. We use a Kalman filter to solve the least squares problem (7.10) and compute the confidence intervals (7.11).

Design and Planning for `AdaSearch` : Choosing $\mathcal{Z}$ and $\mathcal{Z}_i$

Pointwise sensing model. In the pointwise sensing model, $\mathcal{Z} = \mathcal{S}$ and the most informative sensing locations $\mathcal{Z}_i$ at round i are precisely $\mathcal{S}_i$. We therefore choose the path $\mathcal{Z}$ to be a simple space filling curve over a raster grid, which provides coverage of all of $\mathcal{S}$. We adopt a simple dynamical model of the quadrotor in which it can fly at up to a pre-specified top speed, and where acceleration and deceleration times are negligible. This model is suitable for large outdoor environments where travel times are dominated by movement at maximum speed, denoted as τ_0. Figure 7.2 shows an example environment with raster path $\mathcal{Z}$ overlaid (left) and trajectory during round $i = 1$ with $\mathcal{Z}_1$ shown in teal (right).

Physical sensing model. Because the physical sensitivity follows an inverse-square law, the most informative measurements about $\mu(x)$ are those taken at locations near to x. We take measurements at points $z \in \mathbb{R}^3$ two meters above points $x \in \mathcal{S}$ on the ground plane. Flying at relatively low height improves the conditioning of the sensitivity matrix H. We use the same design and planning strategy as in the pointwise model, following the raster pattern depicted in Figure 7.2. More generally, one should chose configurations z_i from $\mathcal{Z}_i$ so that environment point x_i on the ground below each is still in the set $\mathcal{S}_i$ of environment points we are unsure about.

Baselines

We compare `AdaSearch` to two baselines: a uniform-sampling based algorithm `NaiveSearch`, and a spatially-greedy information maximization algorithm `InfoMax`.

NaiveSearch algorithm. As a non-adaptive baseline, we consider a uniform sampling scheme that follows the raster pattern in Figure 7.2 at constant speed. This global `NaiveSearch` trajectory results in measurements uniformly spread over the grid, and avoids redundant movements between sensing locations. The only difference between `NaiveSearch` and `AdaSearch` is that `NaiveSearch` flies at a constant speed, while `AdaSearch` varies its speed. Comparing to `NaiveSearch` thus separates the advantages of `AdaSearch`'s adaptive

[2] Specifically, we define $\tilde{H}_{ij} = h(x_i, z_j)/(\mathbf{Y}_j + b)$. The rescaling term $\mathbf{Y}_j + b$ is a plug-in estimator for the variance of $\mathbf{Y}_j$ (with small bias b introduced for numerical stability), which down-weights higher variance measurements.

measurement allocation from the effects of its global trajectory heuristic. Theoretical analysis in Section 7.4 considers a slight variant in which the sampling time is doubled at each round. This doubling has theoretical benefits, but for all experiments we implement the more practical fixed-speed baseline.

InfoMax algorithm. As discussed in Section 7.1, one of the most successful methods for active search in robotics is receding horizon informative path planning, e.g. [192, 195]. We implement `InfoMax`, a version of this approach based on [192] and specifically adapted for RSS. Each planning invocation solves an information maximization problem over the space of trajectories $\xi : [t, t + T_{\text{plan}}] \longrightarrow \mathcal{B}$ mapping from time in the next T_{plan} seconds to a box $\mathcal{B} \subset \mathbb{R}^3$.

We measure the information content of a candidate trajectory ξ by accumulating the sensitivity-weighted variance at each grid point $x \in \mathcal{S}$ at N evenly-spaced times along ξ, i.e.

$$\xi_t^* = \arg\max_{\xi} \sum_{i=1}^{N} \sum_{j=1}^{|\mathcal{S}|} \Sigma_{jj} \cdot h\left(x_j, \xi(t + T_{\text{plan}} i/N)\right) \quad . \tag{7.12}$$

This objective favors taking measurements sensitive to regions with high uncertainty. As a consequence of the Poisson emissions model, these regions will also generally have high expected intensity μ; therefore we expect this algorithm to perform well for the RSS task. We parameterize trajectories ξ as Bezier curves in $\mathbb{R}^3$, and use Bayesian optimization (see [194]) to solve Eq. (7.12). Empirically, we found that Bayesian optimization outperformed both naive random search and a finite difference gradient method. We set T_{plan} to 30 s and use second-order Bezier curves.

Stopping criteria and metrics. All three algorithms use the same stopping criterion, which is satisfied when the k^{th} highest LCB exceeds the $(k+1)^{th}$ highest UCB. For $k = 1$ emitter, this corresponds to the first round i in which $\text{LCB}_i(x) > \text{UCB}_i(x'), \forall x' \in \mathcal{S} \setminus \{x\}$ for some environment point x. For sufficiently small probability of error δ_{tot}, this ensures that the top-k sources are almost always correctly identified by all algorithms.

7.4 Theoretical Runtime and Sampling Analysis

Separation of sample-based planning and a repeated global trajectory make `AdaSearch` particularly amenable to runtime and sample complexity analysis. We analyze `AdaSearch` and `NaiveSearch` under the pointwise sensing model and trajectory planning strategy outlined in Section 7.3. Runtime and sample guarantees are given in Theorem 7.1, with further analysis for a single source in Corollary 7.1 to complement experiments. For ease of exposition, we defer some proofs and complimentary lower bounds to Sections 7.A to 7.C. Simulations (Section 7.5) show that our theoretical results are indeed predictive of the relative performance of `AdaSearch` and `NaiveSearch`.

We analyze `AdaSearch` with the trajectory and planning strategy outlined in Section 7.3. For `NaiveSearch`, the robot spends time τ_i at each point in each round i until termina-

tion, which is determined by the same confidence intervals and termination criterion for `AdaSearch`.

We will be concerned with the *total runtime*. Recall that τ_0 is the time spent over any point when the robot is moving at maximum speed; τ_i is the time spent sampling candidate points at the slower speed of round i. The total runtime is then

$$T^{\mathrm{run}} = \begin{cases} \sum_{i=0}^{i_{\mathrm{fin}}} \left(\tau_i |\mathcal{S}_i| + \tau_0 |\mathcal{S} \setminus \mathcal{S}_i|\right) & (\texttt{AdaSearch}) \\ \sum_{i=0}^{i_{\mathrm{fin}}} \tau_i |\mathcal{S}| & (\texttt{NaiveSearch}) \end{cases}, \tag{7.13}$$

where i_{fin} is the round at which the algorithm terminates. We state bounds in terms of divergences between emission rates $\mu_2 \geq \mu_1 > 0$:

$$\mathrm{d}(\mu_1, \mu_2) = (\mu_2 - \mu_1)^2 / \mu_2 \ .$$

These divergences approximate the KL-divergence between distributions Poisson(μ_1) and Poisson(μ_2) (see Lemma 7.3), and hence the sample complexity of distinguishing between points emitting photons at rates μ_1, μ_2. Analogous divergences are available for any exponential family, for example Gaussian distributions where the divergences are symmetric.

To achieve the termination criterion (when $\mathcal{S}^*(k)$ is determined with confidence $1 - \delta_{\mathrm{tot}}$), all points with emission rate below the lowest in $\mathcal{S}^*(k)$ must be distinguished from $\mu^{(k)}$, the lowest emission rate of points in $\mathcal{S}^*(k)$. Therefore, for points $x \notin \mathcal{S}^*(k)$, we consider divergences $\mathrm{d}(\mu(x), \mu^{(k)})$. Similarly, all points in $\mathcal{S}^*(k)$ must be distinguished from the highest background emitter corresponding to the divergences $\mathrm{d}(\mu^{(k+1)}, \mu(x))$, describing how close $\mu(x)$ is to the mean rate of the highest background emitter.

Theorem 7.1. *(Sample and runtime guarantees). Define the general adaptive and uniform sample complexity terms $\mathcal{C}^{(k)}_{\texttt{adapt}}$ and $\mathcal{C}^{(k)}_{\texttt{unif}}$:*

$$\begin{aligned} \mathcal{C}^{(k)}_{\texttt{adapt}} &:= |S|\tau_0 + \sum_{x \in \mathcal{S}^*(k)} \tfrac{1}{\mathrm{d}(\mu^{(k+1)}, \mu(x))} + \sum_{x \in \mathcal{S} \setminus \mathcal{S}^*(k)} \tfrac{1}{\mathrm{d}(\mu(x), \mu^{(k)})} \\ \mathcal{C}^{(k)}_{\texttt{unif}} &:= |\mathcal{S}|\tau_0 + |\mathcal{S}| \frac{1}{\mathrm{d}(\mu^{(k+1)}, \mu^{(k)})} \ . \end{aligned} \tag{7.14}$$

$\mathcal{C}^{(k)}_{\texttt{adapt}} \geq \mathcal{C}^{(k)}_{\texttt{unif}}$ for any integer number of sources $k \geq 1$ and any distribution of emitters. For any $\delta_{\mathrm{tot}} \in (0, 1)$, the following hold each with probability at least $1 - \delta_{\mathrm{tot}}$: [3]
(i) `AdaSearch` *correctly returns $\mathcal{S}^*(k)$, with runtime at most*

$$T^{\mathrm{run}}(\texttt{AdaSearch}) \leq \ \mathcal{C}^{(k)}_{\texttt{adapt}} \cdot \widetilde{\mathcal{O}}\left(\log(|\mathcal{S}|/\delta_{\mathrm{tot}})\right) + \widetilde{\mathcal{O}}\left(|\mathcal{S}| \log_+\left(\mathcal{C}^{(k)}_{\texttt{unif}}/|\mathcal{S}|\right)\right) \ .$$

(ii) `NaiveSearch` *correctly returns $\mathcal{S}^*(k)$ with runtime bounded by*

$$T^{\mathrm{run}}(\texttt{NaiveSearch}) \leq \mathcal{C}^{(k)}_{\texttt{unif}} \cdot \widetilde{\mathcal{O}}\left(\log(|\mathcal{S}|/\delta_{\mathrm{tot}})\right) \ .$$

[3] $\widetilde{\mathcal{O}}(\cdot)$ notation suppresses doubly-logarithmic factors.

Theorem 7.1 is a consequence of results in Sections 7.A to 7.C. We sketch the main ideas of the proof here:

Proof of Theorem 7.1 (sketch). The runtimes (7.13) of each algorithm depend on how quickly we can reduce the set $\mathcal{S}_i$ in each round. For each point x, let $i_{\text{fin}}(x)$ denote the round at which `AdaSearch` removes x from $\mathcal{S}_i$; at this point we are confident as to whether or not x is in $\mathcal{S}^*(k)$, so we do not sample it on successive rounds. At round i, we spend time $\tau_i = 2^i$ sampling each point still in $\mathcal{S}_i$, so that we spend $\sum_{x\in\mathcal{S}} \sum_{i=0}^{i_{\text{fin}}(x)} \tau_i \leq \sum_{x\in\mathcal{S}} 2^{1+i_{\text{fin}}(x)}$ time sampling x throughout the run of the algorithm. For `NaiveSearch`, we sample all points in all rounds, so we spend time $\leq |\mathcal{S}| \max_{x\in\mathcal{S}} 2^{1+i_{\text{fin}}(x)}$ sampling.

Now we bound $i_{\text{fin}}(x)$ for each algorithm. These quantities depend on the estimated means $\hat{\mu}_i(x)$. Using the concentration bounds that informed the bounding functions in Section 7.3, we can form deterministic bounds $[\overline{\text{LCB}}_i, \overline{\text{UCB}}_i]$ that depend only on the true means $\mu(x)$. We choose these to encompass the algorithm confidence intervals, so that: $\overline{\text{LCB}}_i(x) \leq \text{LCB}_i(x) \leq \mu(x) \leq \text{UCB}_i(x) \leq \overline{\text{UCB}}_i(x)$ with high probability. If each of these inequalities holds with probability $\delta_{\text{tot}}/(4|\mathcal{S}|i^2)$, then a union bound gives that the probability of failure of any inequality over all rounds is at most δ_{tot}. By Lemma 7.1, this ensures correctness with probability at least $1 - \delta_{\text{tot}}$.

Because $\overline{\text{LCB}}_i(x)$ and $\overline{\text{UCB}}_i(x)$ are deterministic given $\mu(x)$ and are contracting to $\mu(x)$ nearly geometrically in i, we can bound $i_{\text{fin}}(x)$ by inverting the intervals to find the smallest integer i such that $\overline{\text{LCB}}_i(x^*) > \overline{\text{UCB}}_i(x)$ for all $x^* \in \mathcal{S}^*(k)$ and $x \in \mathcal{S} \setminus \mathcal{S}^*(k)$. This requires an inversion lemma from the best arm identification literature (Eq. (110) in [260]). The specific forms of $\overline{\text{LCB}}_i$ and $\overline{\text{UCB}}_i$ yield the bounds on $i_{\text{fin}}(x)$ in terms of approximate KL divergences, which are added across all environment points to obtain the sample complexity terms for each algorithm in (7.14).

The form of $\mathcal{C}^{(k)}_{\texttt{unif}}$ results from noting that the function $(a, b) \mapsto \frac{a}{(a-b)^2}$ is decreasing in a and increasing in b for $a > b$, therefore $\max\left\{\max_{x\in\mathcal{S}^*(k)} \frac{1}{\text{d}(\mu^{(k+1)},\mu(x))}, \max_{x\in\mathcal{S}\setminus\mathcal{S}^*(k)} \frac{1}{\text{d}(\mu(x),\mu^{(k)})}\right\}$ $= 1/\text{d}(\mu^{(k+1)}, \mu^{(k)})$. □

The $\widetilde{\mathcal{O}}\left(\log(|\mathcal{S}|/\delta_{\text{tot}})\right)$ term in the `AdaSearch` runtime bounds accounts for travel times of transitioning between measurement configurations. The second term $|\mathcal{S}| \log(\mathcal{C}_{\texttt{unif}}/|\mathcal{S}|)$ accounts for the travel time of traversing the uninformative points in the global path $\mathcal{Z}$ at a high speed. This term is never larger than $T^{\text{run}}(\texttt{NaiveSearch})$ and is typically dominated by $\mathcal{C}_{\texttt{adapt}} \cdot \widetilde{\mathcal{O}}\left(\log(|\mathcal{S}|/\delta_{\text{tot}})\right)$. With a uniform strategy, runtime scales with the *largest value* of $1/\text{d}(\mu(x_1), \mu(x_2))$ over $x_1 \notin \mathcal{S}^*(k), x_2 \in \mathcal{S}^*(k)$ because that quantity alone determines the number of rounds required. In contrast, `AdaSearch` scales with the *average* of $1/\text{d}(\mu(x_1), \mu(x_2))$ because it dynamically chooses which regions to sample more precisely.

In many scenarios, the number k may estimate the number of hotspots, or there may be more than k sources with similarly high emissions. The extreme case is when emissions $\mu^{(k)}$ and $\mu^{(k+1)}$ are *equal*, the divergences $\text{d}(\mu^{(k+1)}, \mu^{(k)})$ zero. Here the resultant sample complexities $\mathcal{C}_{\texttt{adapt}}$ and $\mathcal{C}_{\texttt{unif}}$ are infinite – because no statistical test can distinguish between

$\mu^{(k)}$ and $\mu^{(k+1)}$, the algorithm continues to collect additional samples without terminating. Later in this section, we resolve this issue by proposing a simple modification of the stopping condition which returns a set $\widehat{\mathcal{S}}$ of possibly greater than k sources. This modification enjoys similar guarantees to our default termination rule (Theorem 7.2).

Sample Complexity for Heterogeneous Sources

Our sample complexity results qualitatively match standard bounds for active top-k identification with sub-Gaussian rewards in the general multi-armed bandit setting (e.g. [154]). The following corollary suggests that when the values of $\mathrm{d}(\mu(x), \mu^*)$ are heterogeneous, `AdaSearch` yields significant speedups over `NaiveSearch`.

Corollary 7.1 (Performance under heterogeneous background noise)**.** *For a large environment with a single source x^* with emission rate μ^* and background signals distributed as $\mu(x) \sim \mathrm{Unif}[0, \overline{\mu}]$ for $x \neq x^*$, the ratio of the upper bounds on sample complexities of* `AdaSearch` *to* `NaiveSearch` *scales with the ratio of $\overline{\mu}$ to μ^* as $1 - \overline{\mu}/\mu^*$.*

Proof of Corollary 7.1. To control the complexity of `NaiveSearch`, note that

$$\mathcal{C}_{\texttt{unif}} = \widetilde{\mathcal{O}}(\max_{x \neq x^*} 1/\mathrm{d}(\mu(x), \mu^*)) = \widetilde{\mathcal{O}}(\mu^*/(\mu^* - \max_{x \neq x^*} \mu(x))^2).$$

It is well known that that the maximum of N uniform random variables on $[0, 1]$ is approximately $1 - \Theta(\frac{1}{N})$ with probability $1 - \Theta(\frac{1}{N})$, implying that $\max_{x \neq x^*} \mu(x) \approx (1 - \frac{1}{|\mathcal{S}|})\overline{\mu} \approx \overline{\mu}$ with probability at least $1 - \Theta(1/|\mathcal{S}|)$. Hence, the sample complexity of `NaiveSearch` scales as $\widetilde{\mathcal{O}}\left(|\mathcal{S}|\mu_*/(\overline{\mu} - \mu_*)^2\right)$. On the other hand, the sample complexity of `AdaSearch` grows as

$$\mathcal{C}_{\texttt{adapt}} = \widetilde{\mathcal{O}}(\sum_{x \neq x^*} 1/\mathrm{d}(\mu(x), \mu^*)) = \widetilde{\mathcal{O}}(\sum_{x \neq x^*} \mu^*(\mu^* - \mu(x))^{-2}).$$

When $\mu(x) \sim \mathrm{Unif}[0, \overline{\mu}]$ are random and $|\mathcal{S}|$ is large, the law of large numbers implies that this tends to $\widetilde{\mathcal{O}}\left(\mu^*|\mathcal{S}| \cdot \mathbb{E}_{\mu(x) \sim \mathrm{Unif}[0,\overline{\mu}]}(\mu^* - \mu(x))^{-2}\right) = \widetilde{\mathcal{O}}\left(|\mathcal{S}|(\mu^* - \overline{\mu})^{-1}\right)$. Therefore, the ratio of sample bounds of `AdaSearch` to `NaiveSearch` is $\left(|\mathcal{S}|(\mu^* - \overline{\mu})^{-1}\right) / \left(|\mathcal{S}|\mu^*(\mu^* - \overline{\mu})^{-2}\right) = 1 - \overline{\mu}/\mu^*$. □

Extension: Unknown Number of High-Emission Sources

In many scenarios, there may be considerably more sources of radiation than anticipated. For example, suppose that `AdaSearch` is specified with one target source ($k = 1$), but in fact there exist two sources $x_1, x_2 \in \mathcal{S}$ with $\mu(x_1) = \mu(x_2) = \mu^*$. Then, `AdaSearch` will *not* terminate with high probability, because it cannot differentiate between these two sources.

To remedy this, we can introduce a slightly more aggressive stopping criterion which will terminate even if multiple sources have similar emissions. The stopping criterion can be stated to terminate if the following holds for an error parameter $\epsilon > 0$:

Definition 7.1 (ϵ-approximate termination rule)**.** Under the ϵ-approximate termination rule, `AdaSearch` either (a) terminates when $\mathcal{S}_i = \emptyset$ and returns $\widehat{\mathcal{S}} \leftarrow \mathcal{S}_i^{\text{top}}$, or (b) terminates when

$$\min\{\text{LCB}_i(x) : x \in \mathcal{S}_i\} \geq \max\{\text{UCB}_i(x) : x \in \mathcal{S}_i\} - \epsilon \tag{7.15}$$

and returns the *union* of the sets $\widehat{\mathcal{S}} = \mathcal{S}_i \cup \mathcal{S}_i^{\text{top}}$.

This criterion will ensure bounded runtime even when there are multiple sources whose mean emissions are close to that of $\mu^{(k)}$. For another possible approach to addressing unknown k, we direct the reader to Section 7.6. Under the modification of Definition 7.1, we can modify Theorem 7.1 as follows. For $\mu_2 \geq \mu_1 > 0$, introduce

$$\text{d}_\epsilon(\mu_1, \mu_2) := \frac{\max\{\mu_2 - \mu_1, \epsilon\}^2}{\mu_2}.$$

Observe that $\text{d}_\epsilon(\mu_1, \mu_2) \geq \text{d}(\mu_1, \mu_2)$, and is always at least ϵ^2/μ_2. We define complexity terms analogous to Eq. (7.14):

$$\begin{aligned} \mathcal{C}_{\texttt{adapt}}^{(k)}(\epsilon) &:= |S|\tau_0 + \sum_{x \in \mathcal{S}^*(k)} \tfrac{1}{\text{d}_\epsilon(\mu^{(k+1)}, \mu(x))} + \sum_{x \in \mathcal{S} \setminus \mathcal{S}^*(k)} \tfrac{1}{\text{d}_\epsilon(\mu(x), \mu^{(k)})} \\ \mathcal{C}_{\texttt{unif}}^{(k)}(\epsilon) &:= |\mathcal{S}|\tau_0 + |\mathcal{S}| \frac{1}{\text{d}_\epsilon(\mu^{(k+1)}, \mu^{(k)})} \ . \end{aligned} \tag{7.16}$$

The above describe the adaptive and uniform complexities analogous to those applied in Theorem 7.1. Essentially, these complexities prevent the runtime from suffering if there are many sources whose emissions are close to that of $\mu^{(k)}$.

Lastly, we introduce a relaxed definition of correctness, which requires that an estimate set $\widehat{\mathcal{S}}$ of the top emitters contains all top emitters $\mathcal{S}^*(k)$, and all remaining sources in the set have emissions close to $\mu^{(k)}$.

Definition 7.2. A set $\widehat{\mathcal{S}} \subset \mathcal{S}$ is said to be (k, ϵ)-correct if $\widehat{\mathcal{S}} \supseteq \mathcal{S}^*(k)$, and for any $x \in \widehat{\mathcal{S}}$, $\mu(x) \geq \mu^{(k)} - \epsilon$.

For the stopping rule in Definition 7.1 and approximate notion of correctness in Definition 7.2, Theorem 7.1 generalizes as follows:

Theorem 7.2. *(Sample and runtime guarantees). For any $\delta_{\text{tot}} \in (0, 1)$ and $\epsilon \leq \mu^{(k)}$, the following hold each with probability at least $1 - \delta_{\text{tot}}$:*
(i) `AdaSearch` *with the termination rule in Definition 7.1 returns a (k, ϵ)-correct $\widehat{\mathcal{S}}$ with runtime at most*

$$T^{\text{run}}(\texttt{AdaSearch}) \leq \ \mathcal{C}_{\texttt{adapt}}^{(k)}(\epsilon) \cdot \widetilde{\mathcal{O}}\left(\log(|\mathcal{S}|/\delta_{\text{tot}})\right) + \widetilde{\mathcal{O}}\left(|\mathcal{S}| \log_+\left(\mathcal{C}_{\texttt{unif}}^{(k)}(\epsilon)/|\mathcal{S}|\right)\right) \ .$$

(ii) `NaiveSearch` *with the termination rule in Definition 7.1 returns a (k, ϵ)-correct $\widehat{\mathcal{S}}$ with runtime bounded by*

$$T^{\text{run}}(\texttt{NaiveSearch}) \leq \mathcal{C}_{\texttt{unif}}^{(k)}(\epsilon) \cdot \widetilde{\mathcal{O}}\left(\log(|\mathcal{S}|/\delta_{\text{tot}})\right) \ .$$

Proof. We first prove correctness. To see that the modified termination rule returns a (k, ϵ)-correct set, we consider the two possible termination criteria. If the default stopping rule $\mathcal{S}_i = \emptyset$, is triggered, then the correctness follows from Theorem 7.1.

Suppose instead that the stopping rule in Eq. (7.15) is triggered, so that $\widehat{\mathcal{S}} = \mathcal{S}_i \cup \mathcal{S}_i^{\text{top}}$. By the original correctness analysis, with high probability, the top emitters are never eliminated from $\mathcal{S}_i \cup \mathcal{S}_i^{\text{top}}$. Thus, $\widehat{\mathcal{S}}$ contains $\mathcal{S}^*(k)$. To prove (k, ϵ)-correctness, let us now show that for any $x \in \widehat{\mathcal{S}}$, $\mu(x) \geq \mu^{(k)} - \epsilon$. With high probability, $\mathcal{S}_i^{\text{top}} \subset \mathcal{S}^*(k)$ for all rounds i, so we will verify this while restricting to $x \in \mathcal{S}_i$. We note that on the high probability event that top emitters are never eliminated from $\mathcal{S}_i \cup \mathcal{S}_i^{\text{top}}$, and if $|\mathcal{S}_i^{\text{top}}| < k$, there exists some $x_0 \in \mathcal{S}^*(k)$ such that $x_0 \in \mathcal{S}_i$; for this source, $\mu(x_0) \geq \mu^{(k)}$. By definition of the stopping rule, we then have

$$\forall x \in \mathcal{S}_i, \quad \text{LCB}_i(x) \geq \text{UCB}_i(x_0) - \epsilon. \tag{7.17}$$

Under the high-probability event that the confidence intervals are correct (see Section 7.B), this means that

$$\forall x \in \mathcal{S}_i, \quad \mu(x) \geq \max\{\mu(x) : x \in \mathcal{S}_i\} - \epsilon. \tag{7.18}$$

Hence, Eq. (7.18) implies that $\forall x \in \mathcal{S}_i, \mu(x) \geq \mu^{(k)} - \epsilon$, as desired.

Let us now account for the improved sample complexity. In the proof of Theorem 7.1, we considered an upper bound $i_{\text{fin}}(x)$ on the last round i at which $x \in \mathcal{S}$ remains in $\mathcal{S}_i$. The total number of samples were then $2^{i_{\text{fin}}(x)}$, and up to logarithmic factors, this upper bound scaled as $\frac{1}{\text{d}(\mu^{(k+1)}, \mu(x))}$ for $x \in \mathcal{S}^*(k)$, and with $\frac{1}{\text{d}(\mu(x), \mu^{(k)})}$ for $x \notin \mathcal{S}^*(k)$.

For the new stopping rule, we have two cases: if $\mu(x) \notin [\mu^{(k)} - \epsilon/4, \mu^{(k)} + \epsilon/4]$, then the desired sample complexity analysis carries through, as is. This is is because, if $\mu(x) \geq \mu^{(k)} + \epsilon/4$ then $\text{d}_\epsilon(\mu^{(k+1)}, \mu(x)) = \Omega(\text{d}(\mu^{(k+1)}, \mu(x)))$ (and analogously when $\mu(x) \leq \mu^{(k)} - \epsilon/4$).

Now, consider i_0 to be sufficiently large so that $i_0 \geq \max\{\inf(x) : \mu(x) \notin [\mu^{(k)} - \epsilon/4, \mu^{(k)} + \epsilon/4]\}$. Then, for $i \geq i_0$, all that remain are means $\mu(x) \in [\mu^{(k)} - \epsilon/4, \mu^{(k)} + \epsilon/4]$. But once the confidence intervals are at most $\epsilon/8$ in width, all LCB's and UCB's will be within, say $\epsilon/4$ of one another, triggering the new stopping condition Eq. (7.15). Hence, the sample complexity for these means is governed by how many samples are required to shrink these intervals to a width of $\Omega(\epsilon)$, which is bounded by $\frac{\mu^{(k)} + \epsilon/4}{\epsilon^2}$. Under the assumption that $\mu^{(k)} \leq \epsilon$, this is at most $\frac{\mu^{(k)}}{\epsilon^2}$, which is bounded by $1/\text{d}_\epsilon(\mu^{(k)}, \mu(x))$ for $x \notin \mathcal{S}^*(k)$, and by $1/\text{d}_\epsilon(\mu(x), \mu^{(k+1)})$ for $x \in \mathcal{S}^*(k)$. □

7.5 Experiments

We compare the performance of `AdaSearch` with the baselines defined in Section 7.3 in simulation for the RSS problem with physical sensing model (also defined in Section 7.3), and validate `AdaSearch` in a hardware demonstration.

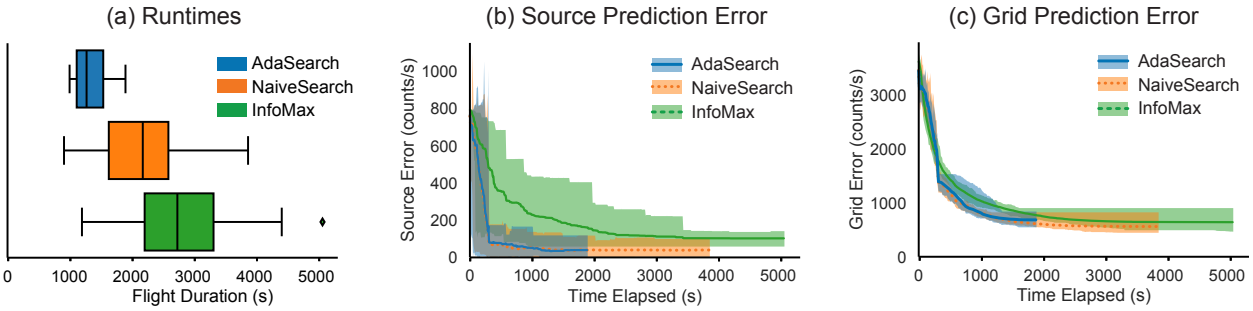


Figure 7.3: Simulation results of `AdaSearch`, `NaiveSearch` and `InfoMax` for 25 randomly instantiated environments with $\overline{\mu} = 400$ and $\mu^* = 800$ $\frac{\text{counts}}{\text{s}}$. Lighter shaded areas denote the range of values at each time over 25 runs of each algorithm; dark lines show the mean. We include final errors for runs that have already finished in max, min, and mean computations. Each algorithm was given the same 25 randomized grids. (a) Runtimes of `AdaSearch`, `NaiveSearch`, and `InfoMax` over 25 random trials. (b) Absolute source error $|\hat{\mu}(x^*)(t) - \mu(x^*)|$ over time. (c) Total grid error $\sqrt{\sum_{x \in \mathcal{S}} \|\hat{\mu}(x)(t) - \mu(x)\|_2^2}$ over time. ©2021 IEEE

Simulation Methodology

We evaluate `AdaSearch`, `InfoMax`, and `NaiveSearch` in simulation using the Robot Operating System (ROS) framework [228]. Environment points $\mathcal{S}$ lie in a 16×16 planar grid, spread evenly over a total area 64×64 m^2. Radioactive emissions are detected by a simulated sensor following the physical sensing model given in Section 7.3 and constrained to fly above a minimum height of 2m at all times (see inset of Figure 7.1 for simulation setup with 4×4 planar grid environment). For all experiments, we set confidence parameter $\alpha = 0.0001$.

For the first set of experiments (Figures 7.3 and 7.5), we set $k = 1$, so that the set of sources $\mathcal{S}^*(k) = \{x^*\}$ is a single point in the environment. We set $\mu^* = \mu(x^*) = 800$ photons/s. In this setting, we investigate algorithm performance in the face of heterogeneous background signals by varying a maximum environment emission rate parameter $\overline{\mu} \in \{300, 400, 500, 600\}$. For each setting of $\overline{\mu}$, we test all three algorithms on 25 grids randomly generated with background emission rates drawn uniformly at random from the interval $[0, \overline{\mu}]$.

We also examine the relative performance of all three algorithms as the number of sources increases (Figure 7.7). For all experiments with $k > 1$, we randomly assign k unique environment points from the grid as the point sources, with emissions rates set to span evenly the range $[800, 1000]$ photons/s. The signals of the remaining background emitters are drawn randomly as before, with $\overline{\mu} = 400$.

Finally, we examine the relative performance of each algorithm for different environment sizes: square grids with widths $2\times$ and $4\times$ that of the previous experiments (Table 7.2). To keep the number of sources that must be disambiguated consistent, we instantiate environment points $\mathcal{S}$ in a 16×16 grid, so that the size of each cell changes with the environment

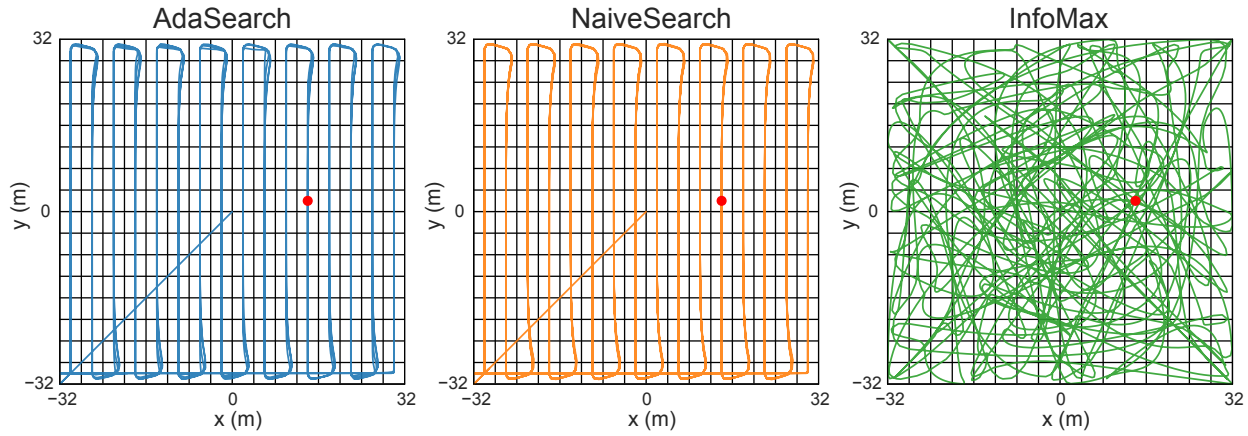


Figure 7.4: Indicative flight paths for each algorithm. The highest emitting source is denoted by the red dot. `AdaSearch` and `NaiveSearch` follow a fixed pattern over several rounds, whereas `InfoMax` does not. ©2021 IEEE

scale factor. Here we set $k = 1$, $\overline{\mu} = 400$, and $\mu^* = 800$. For the `InfoMax` baseline, we adjust the planning horizon to scale with the width of the environment, setting $T_{\text{plan}} = 60$ for the doubled grid and $T_{\text{plan}} = 120$ for the largest grid.

Results

Figure 7.3 shows performance across the three algorithms with respect to the following metrics: (a) total runtime (time from takeoff until x^* is located with confidence), (b) absolute difference between the predicted and actual emission rate of x^*, and (c) aggregate difference between predicted and actual emission rates for all environment points $x \in \mathcal{S}$, measured in Euclidean norm. The uniform baseline `NaiveSearch` terminates significantly earlier than `InfoMax`, and `AdaSearch` terminates even earlier, on average. Of these 25 runs, `AdaSearch` finished faster than `NaiveSearch` in 21 runs, and finished faster than `InfoMax` in 24. The flight patterns for the first trial of each algorithm are shown in Figure 7.4.

To examine the variation in runtimes due to factors other than the environment instantiation, we also conducted 25 runs of the same exact environment grid. Due to delays in timing and message passing in simulation (just like there would be in a physical system), measurements of the simulated emissions can still be thought of as random though the environment is fixed. Indeed, the variance in runtimes was comparable to the variance in runtimes in Figure 7.3; over the 25 trials of a fixed grid, the variance in runtimes were $265s$ (`AdaSearch`), $537s$ (`NaiveSearch`), and $1028s$ (`InfoMax`). Of these 25 runs, `AdaSearch` finished faster than `NaiveSearch` in 18 runs, and finished faster than `InfoMax` in all 25. Figure 7.3(b) plots the absolute difference in the estimated emission rate $\hat{\mu}(x^*)$ and the true emission rate $\mu(x^*)$ at the one source. `AdaSearch` and `NaiveSearch` perform comparably over time, and `AdaSearch` terminates significantly earlier. Figure 7.3(c) plots the Euclidean

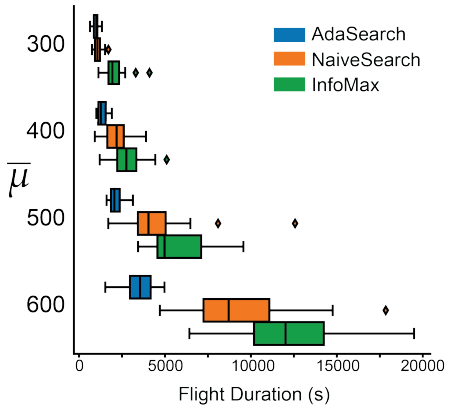


Figure 7.5: Performance of all three algorithms for grids with maximum background varying, and set as $\overline{\mu} \in \{300, 400, 500, 600\}$, $\mu^* = 800$ counts/s. For each value of $\overline{\mu}$, each algorithm was given the same 25 randomized grids. ©2021 IEEE

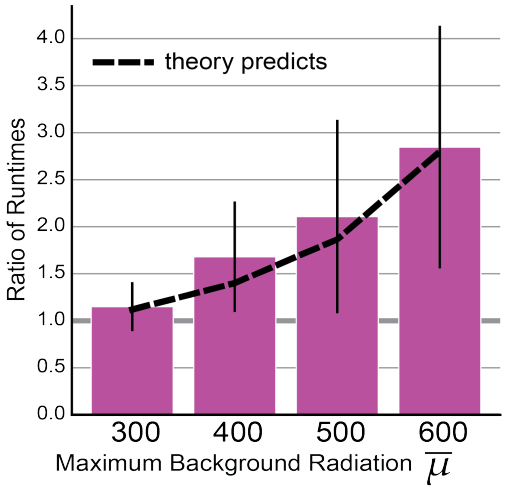


Figure 7.6: Relative performance of `AdaSearch` and `NaiveSearch` per random grid, measured as $T^{\text{run}}(\texttt{NaiveSearch})/T^{\text{run}}(\texttt{AdaSearch})$; magenta bars indicate mean, and horizontal black lines denote one standard deviation from the mean in either direction. Dashed line shows an approximate fit according to Corollary 7.1. Data is obtained from the same randomized grids as in Figure 7.5. ©2021 IEEE

error between the estimated and the ground truth grids; in this metric the gaps in error between all three algorithms are smaller. `AdaSearch` is fast at locating the highest-mean sources without sacrificing performance in total environment mapping.

Figure 7.5 shows performance of all three algorithms across different maximum background radiation thresholds $\overline{\mu}$. As $\overline{\mu}$ increases, all algorithms take longer to terminate

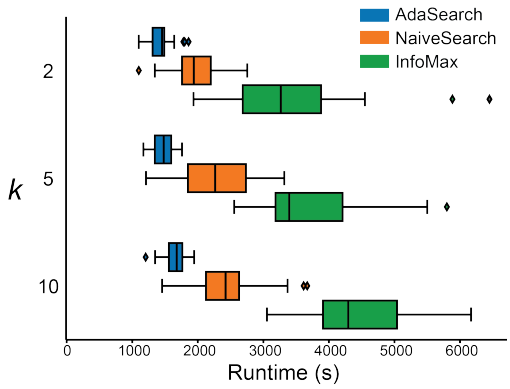


Figure 7.7: Performance across number of sources, k for $k \in [2, 5, 10]$. $\mu(x) \in [800, 1000]$ for all $x \in \mathcal{S}^*(k)$. For each value of k, Each algorithm was given the same 25 randomized grids.

because the source is harder to distinguish from increasing heterogeneous background signals (left). For high background radiation values (e.g. $\overline{\mu} = 600$), the difference in runtimes between all three algorithms is larger; the runtime of `AdaSearch` increases gradually under high background signals, whereas `NaiveSearch` and `InfoMax` are greatly affected. Figure 7.6 shows that as $\overline{\mu}$ approaches μ^*, the relative speedup of using adaptivity, $T^{\text{run}}(\texttt{NaiveSearch})/T^{\text{run}}(\texttt{AdaSearch})$, increases. This is consistent with the theoretical analysis in Section 7.4; the dashed line plots a fit curve with rule $0.7 \cdot \mu^*/(\mu^* - \bar{\mu})$.

Figure 7.7 compares algorithm runtimes across different numbers of sources, k. As suggested from Theorem 7.1, both absolute and relative performance is consistent across k for all three algorithms.

The runtimes and number of rounds executed by `AdaSearch` and `NaiveSearch` for these experiments are summarized in Table 7.1. In every trial, `AdaSearch` takes no more rounds than `NaiveSearch` to reach the termination criterion, suggesting that slowing down over informative points saves the algorithm from having to do more entire passes over the environment. Note that each successive round of the `AdaSearch` algorithm takes longer than a round of `NaiveSearch`, since the robot slows down over informative regions, whereas `NaiveSearch` does not.

`AdaSearch` is inherently a probabilistic algorithm, returning the true sources with probability $1 - \delta_{\text{tot}}$, as a function of the number of rounds and the confidence with parameter, α. Of the 175 trials run throughout these experiments, `AdaSearch` locates the correct source in 174 of them (99.4%). We set $\alpha = 0.0001$ in our experiments to facilitate fair comparison of algorithms while maintaining reasonable runtime of the slower methods (`NaiveSearch`, `InfoMax`). Given the speed with which `AdaSearch` returns a source, in practice it would be feasible to reduce α to reduce the probability of a mistake, δ_{tot}. As evidenced Figure 7.3(c), even in the low-probability case that an incorrect source is returned, `AdaSearch` still provides valuable information about the environment.

Table 7.1: Round number at time of termination and runtime for `AdaSearch` and `NaiveSearch`. Averages and standard deviations taken over 25 trials. ©2021 IEEE

		round # at term.: avg (std)		runtime in seconds: avg (std)	
k	$\overline{\mu}$	`AdaSearch`	`NaiveSearch`	`AdaSearch`	`NaiveSearch`
1	300	3.0 (0.4)	3.8 (0.8)	981 (1778)	1103 (237)
1	400	3.8 (0.6)	7.4 (2.4)	1352 (255)	2208 (694)
1	500	5.0 (0.7)	14.4 (6.8)	2136 (391)	4436 (2145)
1	600	6.6 (0.6)	30.5 (9.5)	3558 (767)	9483 (3004)
2	400	4.0 (0.4)	6.6 (1.2)	1442 (200)	1955 (386)
5	400	3.8 (0.4)	7.8 (2.0)	1465 (167)	2295 (569)
10	400	4.2 (0.5)	8.2 (1.7)	1667 (168)	2502 (532)

Table 7.2: Runtime for each algorithm with different environment sizes, as well as the round number at termination for `AdaSearch` and `NaiveSearch`. Each grid consists of 16×16 cells with $k = 1$ and $\overline{\mu} = 400$. Averages and standard deviations taken over 10 trials. ©2021 IEEE

	round # at termination: avg (std)		runtime in seconds: avg (std)		
grid size	`AdaSearch`	`NaiveSearch`	`AdaSearch`	`NaiveSearch`	`InfoMax`
64×64	3.8 (0.4)	7.6 (2.4)	1345 (229)	2222 (692)	2460 (701)
128×128	2.0 (0.0)	2.4 (0.5)	1356 (170)	1322 (193)	3301 (837)
256×256	1.9 (0.3)	1.9 (0.3)	1916 (548)	2025 (424)	6792 (1949)

The performance of each algorithm for environments at larger scale factors is given in Table 7.2. Doubling the environment scale factor has two effects on the difficulty of the problem. First, it essentially doubles the flight time to fulfill a "snaking" path across the environment (see Figure 7.2). Second, doubling the environment grid width distributes environment points farther from each other in space, so that contributions from individual environment points are easier to disambiguate. The results in Table 7.2 show that for larger grid sizes, both `AdaSearch` and `NaiveSearch` outperform `InfoMax` in terms of runtime. Additionally, the difference in average runtime between `AdaSearch` and `NaiveSearch` is small for the larger grid sizes (128×128 and 256×256 m^2), a consequence of the easier sampling problem (due to dispersed environment points), indicated by a reduced number of rounds needed for the larger grid environments, compared to the 64×64 m^2 grid. In all runs summarized in Table 7.2, the algorithms locate the correct source.

Discussion

While all three methods eventually locate the correct source x^* the vast majority of the time, the two algorithms with global planning heuristics, `AdaSearch` and `NaiveSearch`, terminate considerably earlier than `InfoMax`, which uses a greedy, receding horizon approach (Figure 7.3). Moreover, the adaptive algorithm `AdaSearch` consistently terminates before its non-adaptive counterpart, `NaiveSearch`. These trends hold over differing background noise threshold $\overline{\mu}$ and number of sources, k (Figures 7.5 and 7.7).

The `AdaSearch` algorithm excels when it can quickly rule out points in early rounds. From (7.14) we recall that the `AdaSearch` sample complexity scales with the average value of $\mu(x)/(\mu^* - \mu(x))^2$ (rather than the maximum, for `NaiveSearch`). Hence, `AdaSearch` will outperform `NaiveSearch` when there are varying levels of background radiation.

As $\overline{\mu}$ approaches μ^* and the gaps $\mu^* - \mu(x)$ become more variable, adaptivity confers even greater advantages over uniform sampling. From Corollary 7.1, we expect the ratio of `NaiveSearch` runtime to `AdaSearch` runtime to scale as $\mu^*/(\mu^* - \overline{\mu})$, which is corroborated by the fit of the dashed line to the average runtime ratios in Figure 7.6. The stability of `AdaSearch` in spite of increasing background noise is striking, especially in comparison to the two alternatives presented here; this suggests that in settings where background noise could be misleading to discerning the true signal, a confidence-bound based sampling scheme is likely preferable.

The performance differences between `AdaSearch`, `InfoMax`, and `NaiveSearch` hold as the number of sources increases, indicating that `AdaSearch` is preferable for a range of different environments and source seeking instances.

`InfoMax`'s strength lies in quickly reducing global uncertainty across the entire emissions landscape. However, `InfoMax` takes considerably longer to identify x^* (Figure 7.3(a)) and, surprisingly, `AdaSearch` and `NaiveSearch` perform similarly to `InfoMax` in mapping the entire emissions landscape on longer time scales (Figure 7.3(c)). We attribute this to the effects of greedy, receding horizon planning. Initially, `InfoMax` has many locally-promising points to explore and reduces the Euclidean error quickly. Later on, it becomes harder to find informative trajectories that route the quadrotor near the few under-explored regions. The results in Table 7.2 evidence that this problem remains for larger environments as well. These results suggest that when a path $\mathcal{Z}$ such as the raster path used here is available, it is well worth considering.

High variation in all experiments is expected due to the noisy Poisson emissions signals. While this noise effects the runtime of all algorithms, the range of runtimes for `AdaSearch` is consistently tight compared to the other two methods, suggesting that carefully allocated measurements are indeed increasing robustness under heterogeneous background signals.

Hardware Demonstration

The previous results are based on a simulation of two key physical processes: radiation sensing and vehicle dynamics. We also test `AdaSearch` on a Crazyflie 2.0 palm-sized quadro-

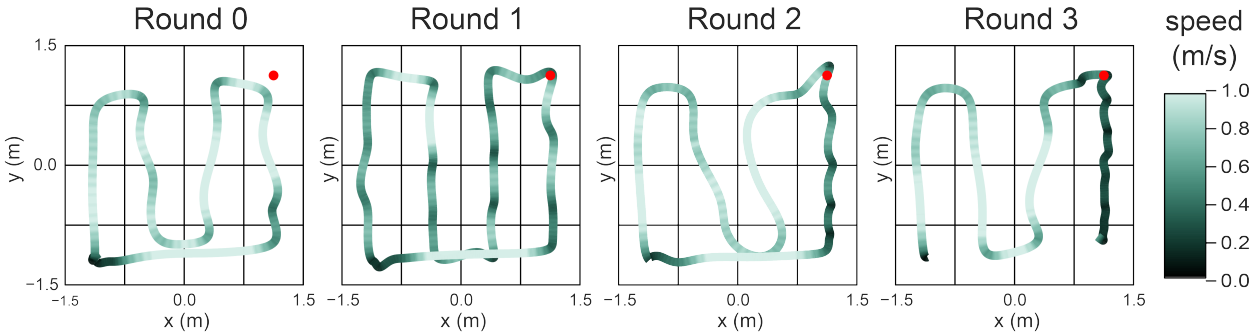


Figure 7.8: Hardware experiment trajectories for each round of `AdaSearch`, color coded by speed (m/s). The highest emitting source is denoted by the red dot. ©2021 IEEE

tor in a motion capture room with simulated radiation readings. The motion capture data (position and orientation) is acquired at roughly 235 Hz and processed in real time using precisely the same implementation of `AdaSearch` used in our software simulations. Figure 7.1 visualizes the confidence intervals and the absolute source point estimation error, as well as the horizontal speed during a representative flight over a small 4×4 grid, roughly 3m / side.

Figure 7.8 shows the flight paths for each round, color coded by speed (darker is slower). Despite imperfections in following the snake path and velocity changes, the robot's trajectory successfully represents the algorithm. After two rounds, `AdaSearch` identifies the two highest emitting points as the highlighted pixels in the top inset, and the absolute error in estimating $\mu(x^*)$ is very small. `AdaSearch` spends most of its remaining runtime sensing these two points and avoids taking redundant measurements elsewhere. The plot of horizontal speed over time (lower inset of Figure 7.1) shows this reallocation of sensor measurements; in the final two rounds, the quadrotor moves quickly at first, then slows down over the two candidate points. This hardware demonstration gives preliminary validation that `AdaSearch` is indeed safe and reasonable to use onboard a physical system.

7.6 Generalizations and Extensions

Before concluding, we briefly discuss several extensions and generalizations of `AdaSearch`.

Unknown Number of Sources

At the end of Section 7.4, we presented a modified termination criterion which accommodates the possibility of multiple high-emission sources (Definition 7.1). This criterion is particularly suitable if there are multiple sources whose emissions are near that of the k-th largest source $\mu^{(k)}$. The modified algorithm correctly recovers all top-k sources, as well as possibly returning *some* additional sources for which emissions $\mu(x)$ are within ϵ of $\mu^{(k)}$.

In other scenarios, it may be more suitable to return *all* sources for which emissions $\mu(x)$ are within ϵ of $\mu^{(k)}$ and *no* sources whose emission are with a factor of $\epsilon' > \epsilon$. With more sophisticated termination criterion and candidate sets $\mathcal{S}_i^{\text{top}}$ and $\mathcal{S}_i$, `AdaSearch` can be modified to accommodate this alternate guarantee. More broadly, the `AdaSearch` design principle—combining confidence-interval based elimination with simple raster movement planning—is amenable to other approximate-search criteria which may arise in given application domains. We also note that, if one runs `AdaSearch` with a small k, the algorithm will still collect measurements from other high-emission locations that can re-used if the practitioner wishes to consider a greater number of $k' > k$ on a subsequent run.

Oriented Sensor

A natural extension of the radioactive source-seeking example is to consider a sensing model with a sensitivity function which depends upon orientation. The additional challenge lies in identifying informative sensing configuration sets $\mathcal{Z}_i$ and a reasonably efficient equivalent fixed global path $\mathcal{Z}$. More broadly, the sensing configurations $z \in \mathcal{Z}$ could be taken to represent generalized configurations of the robot and sensor, e.g., they could encode the position and angular orientation of a directional sensor or joint angles of a manipulator arm.

Pointwise Sensing Model

We motivated the pointwise sensing model where sensitivity function is $h(x, z) = \mathbb{I}\{x = z\}$ as a model conducive to theoretical analysis. Though it is only a coarse (yet still predictive) approximation of the physical process of radiation sensing, this sensitivity model is a more precise descriptor of other sensing processes. For example, the pointwise model is appropriate for survey design. As a concrete example, suppose a government agency with enough funding to set up k medical clinics seeks to identify which k towns had the highest rates of disease. It is reasonable to think that the data collected about town i is mostly informative about only the rate of disease in that town, so that the pointwise sensing model may be quite appropriate.

Surveying

Although we demonstrate `AdaSearch` operating onboard a UAV in the context of RSS, the core algorithm applies more broadly, even to non-robotic embodied sensing problems. Consider the above problem of a government planning k clinic locations, each in a different town. Because surveys are conducted in person, the government may be resource limited, both in terms of the time it takes to survey a single person or clinic within a town, and in terms of travel time between towns. A survey planner could use `AdaSearch` to guide the decisions of how long to spend in each town counting new cases of the disease before moving on to the next, and to trade-off the travel time of returning to collect more data from a certain town with spending extra time at the town in the first place.

While `AdaSearch` provides a good starting point for solving such problems, the high cost of transportation would likely make it worthwhile to further optimize the surveying trajectory at each round, e.g. by (approximately) solving a traveling salesman problem.

7.7 Conclusions

In this chapter, we show that statistical methods from pure exploration active learning offer a promising, under-explored toolkit for robotic source seeking. Specifically, we show that motion constraints need not impede active learning strategies.

Our main contribution, `AdaSearch`, outperforms a greedy information-maximization baseline in a radioactive source seeking task. Its success can be understood as a consequence of two structural phenomena: planning horizon and implicit design objective. The information-maximization baseline operates on a receding horizon and seeks to reduce global uncertainty, which means that even if its planned trajectories are individually highly informative, they may lead to sub-optimal performance over a long time scale. In contrast, `AdaSearch` uses an application-dependent global path that efficiently covers the entire search space and allocates measurements using principled, statistical confidence intervals.

While our results for the problem of RSS are encouraging, it is likely that in many applications, performance could be limited by the range, field of view, or orientation of the sensors. In some cases (e.g. oriented sensors), such limitations could be addressed by the extensions suggested in Section 7.6, and in others, might necessitate new innovations. We are hopeful that the abstraction of sensing models, statistical measurement, and path planning as separate but integrated components of source seeking can guide such future innovations.

`AdaSearch` excels in situations with a heterogeneous distribution of the signal of interest; it would be interesting to make a direct comparison with Gaussian process (GP)-based methods in a domain where the smooth GP priors are more appropriate. Furthermore, as `AdaSearch` is explicitly designed for general embodied sensing problems, it would be exciting to test it in a wider variety of application domains.

7.A Theoretical Results for Pointwise Sensing

In this section, we present formal statements of the measurement complexities provided in Section 7.4.

Notation. Throughout, we use the notation $f \lesssim g$ to denote that there exists a universal constant C, independent of problem parameters, for which $f \leq C \cdot g$. We also define $\log_+(x) := \max\{1, \log x\}$.

Formal setup. Throughout, we consider a rectangular grid $\mathcal{S}$ of $|\mathcal{S}|$ points, and let $\mu(x)$ denote the mean emission rate of each point $x \in \mathcal{S}$ in counts/second. We let $\mu^{(k)}$ denote the k-th largest mean $\mu(x)$. In the case that $k = 1$, we denote $\mu^* := \mu^{(1)}$, and let

$x^* := \arg\max_{x\in\mathcal{S}} \mu(x)$ denote the highest-mean point, with emission rate $\mu^* := \mu(x^*)$. For identifiability, we assume $\mu^{(k)} > \mu^{(k+1)}$.

Measurements. As described in Section 7.3, we assume a point-wise sensing model in which `AdaSearch` and `NaiveSearch` can measure each point directly. Recall that at each round i, `AdaSearch` takes $\tau_i := 2^i\tau_0$ measurements at each point $x \in \mathcal{S}_i$, and `NaiveSearch` takes τ_i measurements at each $x \in \mathcal{S}$. We let $\mathbf{N}_i(x)$ denote the total number of counts collected at position x at round i. We further assume that $\mu(x)$ are standardized according to the same time units as τ_0, so that measuring a source of mean $\mu(x)$ for time interval of length τ_i yields counts distributed according to $\mathbf{N}_i \sim \text{Poisson}(2^i \cdot \mu(x))$. Finally, we shall let i_{fin} denote the (random) round at which a given algorithm - either `AdaSearch` or `NaiveSearch` - terminates.

Confidence intervals. At the core of our analysis are rigorous $1-\delta$ upper and lower confidence intervals for Poisson random variables. It is well known that the upper Poisson tail satisfies Bennet's inequality, and its lower tail is sub-Gaussian, yielding the following exponential tail bounds (see, e.g. [43]):

Lemma 7.2. *Let* $\mathbf{N} \sim \text{Poisson}(\mu)$. *Then,*

$$\mathbb{P}[\mathbf{N} \geq \mu + x] \leq \exp\left(-\frac{x^2}{2(\mu + x/3)}\right) \text{ and } \mathbb{P}[\mathbf{N} \leq \mu - x] \leq \exp\left(-\frac{x^2}{2\mu}\right) . \tag{7.19}$$

.

At each round i and $x \in \mathcal{S}_i$ (`AdaSearch`) or $x \in \mathcal{S}$ (`NaiveSearch`), recall that we use upper and lower confidence intervals with definitions recalled in Eq. (7.20) below:

$$\text{LCB}_i(x) := \frac{1}{\tau_i} U_-\left(\mathbf{N}_i(x), \delta_i\right), \quad \text{UCB}_i(x) := \frac{1}{\tau_i} U_+\left(\mathbf{N}_i(x), \delta_i\right), \quad \text{where } \delta_i := \delta/(4|\mathcal{S}|i^2).$$

The following proposition bounds the probability that the true mean emission rate is above or below these quantities.

Proposition 7.1. *Fix any* $\mu \geq 0$ *and let* $\mathbf{N} \sim \text{Poisson}(\mu)$. *Define*

$$\begin{aligned} U_+\left(\mathbf{N}, \delta\right) &:= 2\log(1/\delta) + \mathbf{N} + \sqrt{2\mathbf{N}\log\left(1/\delta\right)} \quad \text{and} \\ U_-\left(\mathbf{N}, \delta\right) &:= \max\left\{0, \mathbf{N} - \sqrt{2\mathbf{N}\log(1/\delta)}\right\} \end{aligned} \tag{7.20}$$

Then, it holds that $\mathbb{P}[\mu > U_+(\mathbf{N}, \delta)] \leq \delta$ *and* $\mathbb{P}[\mu < U_-(\mathbf{N}, \delta)] \leq \delta$.

Proof of Proposition 7.1. We first prove that $\mathbb{P}[\mu \leq U_+(\mathbf{N}, \delta)] \geq 1 - \delta$. Recall the definition

$$U_+\left(\mathbf{N}, \delta\right) := 2\log(1/\delta) + \mathbf{N} + \sqrt{2\mathbf{N}\log\left(1/\delta\right)}.$$

We begin by bounding the lower tail of $\mathbf{N}$, which corresponds to the upper confidence U_+ bound on μ. Let $\mathcal{E}_+(\delta)$ denote the event $\{\mathbf{N} \geq \mu - \sqrt{2\mu\log(1/\delta)}\}$. By Lemma 7.2, we have

that $\mathbb{P}[\mathcal{E}_+(\delta)^c] \leq \delta$; hence, it suffices to show that $\mathcal{E}_+(\delta)$ implies $\{\mu \leq U_+(\mathbf{N}, \delta)\}$. If $\mathcal{E}_+(\delta)$ holds, the quadratic equation implies

$$\mu^{1/2} \leq \frac{\sqrt{2\log(1/\delta)} + \sqrt{2\log(1/\delta) + 4\mathbf{N}}}{2}. \tag{7.21}$$

Hence, we have

$$\begin{aligned}\mu &\leq \frac{2\log(1/\delta) + 2\log(1/\delta) + 4\mathbf{N} + 2\sqrt{2\log(1/\delta)}\sqrt{4\mathbf{N} + 2\log(1/\delta)}}{4} \\ &\leq 2\log(1/\delta) + \mathbf{N} + \sqrt{2\mathbf{N}\log(1/\delta)} = U_+(\mathbf{N}, \delta), \text{ as needed.}\end{aligned}$$

Next we prove that $\mathbb{P}[\mu \geq U_-(\mathbf{N}, \delta)] \geq 1 - \delta$. Analogous to the above, let $\mathcal{E}_-(\delta) := \{\mathbf{N} \leq \mu + \sqrt{2\mu\log(1/\delta)} + \frac{2}{3}\log(1/\delta)\}$. Since $\mathbb{P}[\mathbf{N} \geq \mu + x] \leq \exp(-\frac{x^2}{2(\mu + x/3)})$, we have that $\mathbb{P}[\mathcal{E}_-(\delta)] \geq 1 - \delta$. Thus, again it suffices to show that $\mathcal{E}_-(\delta)$ implies $\{\mu \geq U_-(\mathbf{N}, \delta)\}$. We have two cases:

(a) $\mathbf{N} \leq \frac{2}{3}\log(1/\delta)$. Then, $U_-(\mathbf{N}, \delta) = 0$, so $\mu \geq U_-(\mathbf{N}, \delta)$ trivially.

(b) Otherwise, by solving the quadratic in the definition $\mathcal{E}_-(\delta)$, we find that on $\mathcal{E}_-(\delta)$,

$$\begin{aligned}\mu^{1/2} &\geq \frac{-\sqrt{2\log(1/\delta)} \pm \sqrt{2\log(1/\delta) - \frac{8}{3}\log(1/\delta) + 4\mathbf{N}}}{2} \\ &= \frac{-\sqrt{2\log(1/\delta)} \pm \sqrt{4\mathbf{N} - 2/3\log(1/\delta)}}{2},\end{aligned}$$

where we note that the discriminant is positive since $\mathbf{N} \geq \frac{2}{3}\log(1/\delta)$. Squaring,

$$\begin{aligned}\mu &\geq \frac{2\log(1/\delta) + 4\mathbf{N} - 2\log(1/\delta)/3 - 2\sqrt{2\log\delta^{-1}}\sqrt{4\mathbf{N} - 2/3\log(1/\delta)}}{4} \\ &\geq \mathbf{N} + (1 - 1/6 - \sqrt{1/3})\log(1/\delta) - \sqrt{2\mathbf{N}\log(1/\delta)} \geq \mathbf{N} - \sqrt{2\mathbf{N}\log(1/\delta)}.\end{aligned}$$

Since we also have $\mu \geq 0$, we see that on $\mathcal{E}_0(\delta)$, we have that

$$\mu \geq \max\{\mathbf{N} - \sqrt{2\mathbf{N}\log(1/\delta)}, 0\} = U_-(\mathbf{N}, \delta), \text{ as needed.}$$

□

Trajectory for `AdaSearch`. `AdaSearch` follows a trajectory where, at each round i, `AdaSearch` spends time $\tau_i = 2^i \cdot \tau_0$ measuring each $x \in \mathcal{S}_i$, and spends τ_0 travel time traveling over each $x \notin \mathcal{S}_i$. For the radioactive sensing problem, this is achieved by following the "snaking pattern" depicted in Figure 7.2, in which the quadrotor speeds up or slows

down over each point to match the specified measurement times. We define the *total sample complexity* and *total run time* respectively as

$$T^{\text{sample}} := \sum_{i=0}^{i_{\text{fin}}} \tau_i |\mathcal{S}_i| \quad \text{and} \quad T^{\text{run}} := T^{\text{sample}} + \sum_{i=0}^{i_{\text{fin}}} \tau_0 |\mathcal{S} \setminus \mathcal{S}_i|.$$

The first quantity above captures the total number of measurements taken at points we still wish to measure, and the second captures the total flight time of the algorithm. For simplicity, we will normalize our units of time so that $\tau_0 = 1$.

Trajectory for `NaiveSearch`. Whereas we implement `NaiveSearch` to travel at a constant speed at each point for each round, our analysis considers a variant where `NaiveSearch` halves its speed each round - that is, takes 2^i measurements at each point for each round; this doubling yields slightly better bounds on sample complexity, and makes `NaiveSearch` compare even more favorably compared to `AdaSearch` in theory.[4] This results in a total of $2^i \tau_0 |\mathcal{S}|$ measurements per round. For `NaiveSearch`, the total sample complexity and total run time are equal, and given by

$$T^{\text{run}} = T^{\text{sample}} = |\mathcal{S}| \sum_{i=0}^{i_{\text{fin}}} \tau_i.$$

Termination criterion for `NaiveSearch`. For an arbitrary number of k emitters, `NaiveSearch` terminates at the first round i in which the k-th largest lower confidence bound of all points $x \in \mathcal{S}$ is higher than the $(k+1)$-th largest upper confidence bound of all points $x \in \mathcal{S}$.

Main Results for $k = 1$ Emitters

We are now ready to state our main theorems for $k = 1$ emitters. Recall the divergence terms

$$\mathrm{d}(\mu_1, \mu_2) := \frac{(\mu_2 - \mu_1)^2}{\mu_2}, \ (\mu_1 < \mu_2)$$

and, in particular,

$$\mathrm{d}(\mu(x), \mu^*) := \frac{(\mu^* - \mu(x))^2}{\mu^*}, \ (\mu(x) < \mu^*)$$

from Section 7.4. When term $\mathrm{d}(\mu(x), \mu^*)$ is small, it is difficult to distinguish between x^* and x. The following lemma shows that $\mathrm{d}(\mu(x), \mu^*)$ approximates the the KL-divergence between the distribution Poisson($\mu(x)$) and Poisson(μ^*):

[4] In practice, we keep the speed constant between trials because, for uniform sampling, this is more efficient; that is, in both theory and practical evaluations, we choose the variant of `NaiveSearch` performs the best.

Lemma 7.3. *There exists universal constants c_1 and c_2 such that, for any $\mu_2 \geq \mu_1 > 0$,*

$$c_1 \cdot \mathrm{d}(\mu_1, \mu_2) \leq \mathrm{KL}(\mathrm{Poisson}(\mu_1), \mathrm{Poisson}(\mu_2)) \leq c_2 \cdot \mathrm{d}(\mu_1, \mu_2), \tag{7.22}$$

where $\mathrm{d}(\mu_1, \mu_2) = (\mu_2 - \mu_1)^2/\mu_2$.

Proof of Lemma 7.3. We begin by stating a standard computation of the KL-divergence between two Poisson distributions.

Fact 7.1. $\mathrm{KL}(\mathrm{Poisson}(\mu_1), \mathrm{Poisson}(\mu_2)) = \mu_1 \log(\mu_1/\mu_2) + (\mu_2 - \mu_1)$.

To prove Lemma 7.3, recall that we assume that $\mu_2 \geq \mu_1$. We may therefore reparameterize $\mu \leftarrow \mu_2$, and $\mu_1 \leftarrow (1-\alpha)\mu_2$ for $\alpha \in (0,1)$. One then has

$$\mathrm{KL}(\mathrm{Poisson}(\mu_1), \mathrm{Poisson}(\mu_2)) = \mu\{\alpha + (1-\alpha)\log(1-\alpha)\} .$$

Since $\mu_2 - \mu_1 = \mu(1-(1-\alpha)) = \alpha\mu$, it suffices to show that there exists constants c_1 and c_2 such that

$$c_1 \mu \alpha^2 \leq \mathrm{KL}(\mathrm{Poisson}(\mu_1), \mathrm{Poisson}(\mu_2)) \leq c_2 \mu \alpha^2 . \tag{7.23}$$

To this end, it suffices to show that there exists a universal constant $\alpha_0 > 0$, such for all $\alpha \leq \alpha_0$, one has

$$\mathrm{KL}(\mathrm{Poisson}(\mu_1), \mathrm{Poisson}(\mu_2)) \in \left[\frac{1}{4}, \frac{3}{4}\right] \mu\alpha^2 . \tag{7.24}$$

Indeed, for any $\alpha \geq \alpha_0$, we have that

$$0 < \mu(\alpha_0 + (1-\alpha_0)\log(1-\alpha_0)) \leq \mathrm{KL}(\mathrm{Poisson}(\mu_1), \mathrm{Poisson}(\mu_2)) \leq \mu ,$$

which implies that, for $\alpha \in [\alpha_0, 1]$,

$$\begin{aligned} 0 &< \left(\alpha^2\mu\right) \cdot (\alpha_0 + (1-\alpha_0)\log(1-\alpha_0)) \\ &\leq \alpha^2\mu \cdot \frac{\alpha_0 + (1-\alpha_0)\log(1-\alpha_0)}{\alpha^2} \\ &\leq \mathrm{KL}(\mathrm{Poisson}(\mu_1), \mathrm{Poisson}(\mu_2)) \\ &\leq \alpha^2\mu \cdot \frac{1}{\alpha^2} \leq \left(\mu\alpha^2\right) \cdot \frac{1}{\alpha_0^2} . \end{aligned}$$

Hence taking $c_1 = (\alpha_0 + (1-\alpha_0)\log(1-\alpha_0))$ and $c_2 = \frac{1}{\alpha_0^2}$, we see that (7.23) holds for $\alpha \in [\alpha_0, 1]$. We now turn to prove (7.24). Note that $\log'(1-x) = -1/(1-x)$, $\log''(1-x) = 1/(1-x)^2$, and $\log'''(x) = -2/(1-x)^3$. Hence, by Taylor's theorem, there exists an $\alpha' \in [0, \alpha]$ such that

$$\log(1-\alpha) = \alpha\log'(1) + \frac{\alpha^2}{2}\log''(1) + \frac{\alpha^3}{6}\log'''(1-\alpha') = -\alpha - \frac{\alpha^2}{2} - \frac{2\alpha^3}{(1-\alpha')^3} .$$

Hence,

$$\begin{aligned}\mathrm{KL}(\mathrm{Poisson}(\mu_1), \mathrm{Poisson}(\mu_2)) &= \mu\left\{\alpha - (1-\alpha)\left(\alpha + \frac{\alpha^2}{2} + \frac{2\alpha^3}{(1-\alpha')^3}\right)\right\} \\ &= \mu\left\{\frac{\alpha^2}{2} + \alpha^3\left(\frac{-2}{(1-\alpha')^3} + \frac{1}{2} + \frac{2\alpha}{(1-\alpha')^3}\right)\right\} .\end{aligned}$$

In particular, there exists a universal constant α_0 such that, for all $\alpha \le \alpha_0$,

$$\mathrm{KL}(\mathrm{Poisson}(\mu_1), \mathrm{Poisson}(\mu_2)) \in \left[\frac{1}{4}, \frac{3}{4}\right] \mu\alpha^2 .$$

□

Up to log factors, the sample complexities for `AdaSearch` and `NaiveSearch` in the $k = 1$ case are given by $\mathcal{C}_{\texttt{adapt}}$ and $\mathcal{C}_{\texttt{unif}}$, respectively, below:

$$\mathcal{C}_{\texttt{adapt}} := \sum_{x \in \mathcal{S}\setminus\{x^*\}} \left(\tau_0 + \frac{1}{\mathrm{d}(\mu(x), \mu^*)}\right) \quad \text{and} \quad \mathcal{C}_{\texttt{unif}} = |\mathcal{S}| \left(\tau_0 + \max_{x \in S} \frac{1}{\mathrm{d}(\mu(x), \mu^*)}\right) \tag{7.25}$$

Similarly to the definitions in Theorem 7.1, $\mathcal{C}_{\texttt{adapt}}$ and $\mathcal{C}_{\texttt{unif}}$ differ in that $\mathcal{C}_{\texttt{adapt}}$ considers the sum over all these point-wise complexities, whereas $\mathcal{C}_{\texttt{unif}}$ replaces this sum with the number of points multiplied by the worst per-point complexity. $\mathcal{C}_{\texttt{adapt}}$ can be thought of as the complexity of sampling each point $x \neq x^*$ the exact number of times to distinguish it from x^*, whereas $\mathcal{C}_{\texttt{unif}}$ is the complexity of sampling each point the exact number of times to distinguish the best point from *every other* point. Note that we always have that $\mathcal{C}_{\texttt{adapt}} \le \mathcal{C}_{\texttt{unif}}$, and in fact $\mathcal{C}_{\texttt{unif}}$ can be as large as $\mathcal{C}_{\texttt{adapt}} \cdot |\mathcal{S}|$.

Our next theorem bounds the *sample complexity* of `AdaSearch` for the $k = 1$ version of the case presented in Section 7.4. We recall that the sample complexity is the total time spent at all $x \in \mathcal{S}_i$ until termination.

Theorem 7.3. *For any $\delta \in (0, 1)$, the following holds with probability at least $1 - \delta$:* `AdaSearch` *correctly returns x^*, the total sample complexity is bounded by bounded above by*

$$T^{\text{sample}} \lesssim |\mathcal{S}| + \sum_{x \neq x^*} \frac{\log_+\left(|\mathcal{S}| \log_+\left(\frac{1}{\mathrm{d}(\mu(x), \mu^*)}\right)/\delta\right)}{\mathrm{d}(\mu(x), \mu^*)} = \widetilde{O}\left(|\mathcal{S}| + \sum_{x \neq x^*} \frac{\log(|\mathcal{S}|/\delta)}{\mathrm{d}(\mu(x), \mu^*)}\right),$$

and the runtime is bounded above by

$$T^{\text{run}} \lesssim T^{\text{sample}} + |\mathcal{S}| \log_+\left(\log_+ \tfrac{|\mathcal{S}|}{\delta} \cdot \max_{x \neq x^*} \tfrac{1}{\mathrm{d}(\mu(x), \mu^*)}\right) = \widetilde{\mathcal{O}}\left(\mathcal{C}_{\texttt{adapt}} \log(|\mathcal{S}|/\delta) + \tau_0 |\mathcal{S}| \log\left(\tfrac{\mathcal{C}_{\texttt{unif}}}{|\mathcal{S}|}\right)\right),$$

where $\widetilde{O}(\cdot)$ hides the doubly logarithmic factors in $1/\mathrm{d}(\mu(x), \mu^*)$.

Theorem 7.3 is a direct consequence of our more general bound for $k \geq 1$ emitters, given by Theorem 7.5, which is proved in Section 7.B. The next theorem, proved in Section 7.C, controls the sample complexity of `NaiveSearch`.

Theorem 7.4. *For any $\delta \in (0,1)$, the following holds with probability at least $1-\delta$:* `NaiveSearch` *correctly returns x^*, and the total runtime is bounded by bounded above by*

$$T^{\mathrm{run}} \lesssim |\mathcal{S}| \cdot \left(\max_{x \neq x^*} \frac{\log_+ \left(|\mathcal{S}| \log_+ \left(\frac{1}{\mathrm{d}(\mu(x), \mu^*)} \right) / \delta \right)}{\mathrm{d}(\mu(x), \mu^*)} \right) \quad = \quad \widetilde{O}\left(\mathcal{C}_{\texttt{unif}} \log(|\mathcal{S}|/\delta) \right) \ .$$

Lastly, we show that our adaptive and uniform sample complexities are near optimal. The following proposition lower bounds the number of samples any adaptive algorithm must take.

Proposition 7.2. *There exists a universal constant c such that, for any $\delta \in (0, 1/4)$, any adaptive sampling allocation which correctly identifies the top emitting point x^* with probability at least $1-\delta$ must collect at least*

$$c \log(1/\delta) \cdot \mathcal{C}_{\texttt{adapt}}$$

samples in expectation. Moreover, any uniform sampling allocation which identifies the top emitting point x^* with probability at least $1-\delta$ must take at least

$$c \log(1/\delta) \cdot \mathcal{C}_{\texttt{unif}}$$

samples in expectation.

Proof of Proposition 7.2. The basic proof strategy follows along the lines of the information-theoretic lower bounds in [157]. Consider a grid $\mathcal{S}$ of $|\mathcal{S}|$ points, with means $\mu(x), x \in \mathcal{S}$. We fix a given sampling algorithm, adaptive or otherwise, and let $T(x)$ denote the expected number of measurements from point x given that the means are given by $\mu(x)$. Suppose $x^* := \arg\max_{x \in \mathcal{S}} \mu(x)$ is unique. We will argue that for a universal constant c_1 and any $x \neq x^*$,

$$T(x) \geq \frac{c_1 \log(1/\delta)}{\mathrm{KL}(\mu(x), \mu(x^*))} \ . \tag{7.26}$$

By the KL approximation in Lemma 7.3, this implies that for some universal constant c_2,

$$T(x) \geq \frac{c_2 \log(1/\delta)}{\mathrm{d}(\mu(x), \mu^*)} \ . \tag{7.27}$$

For adaptive sampling, the expected number of samples is at least $\sum_{x \neq x^*} T(x)$, which by (7.27) is at least

$$c_2 \log(1/\delta) \cdot \sum_{x \neq x^*} \frac{\mu(x^*)}{\Delta_x^2} \ .$$

This completes the proof for adaptive sampling. For non-adaptive sampling, $T(x) = T(x')$ for all $x, x' \in \mathcal{S}$. Hence, the expected number of samples is at least

$$\sum_{x\in\mathcal{S}} T(x) = |\mathcal{S}| \max_{x\neq x^*} T(x) \overset{(7.27)}{\geq} |\mathcal{S}| \cdot \max_{x\neq x^*} c_2 \log(1/\delta)\frac{\mu(x^*)}{\Delta_x^2} \ .$$

We now verify Equation (7.26). To do so, consider an alternative grid $\mathcal{S}$ of $|\mathcal{S}|$ pixels, with means $\mu'(x)$. Suppose moreover that $x^{*'} := \arg\max_{x\in\mathcal{S}} \mu'(x)$ is unique, and that $x^* \neq x$. The key insight from Kaufmann, Cappé, and Garivier [157] is that any algorithm which identifies x^* with probability $1-\delta$ must be able to distinguish between the means $\mu'(x)$ and the with means $\mu(x)$. Kaufmann, Cappé, and Garivier [157] show that this requires that the expected number of samples $T(x)$ satisfy

$$\sum_{x\in\mathcal{S}} T(x)\mathrm{KL}(\mu(x), \mu'(x)) \geq c_1 \log(1/\delta) \ . \tag{7.28}$$

Now let's fix a particular $x_0 \neq x^*$ and an $\epsilon > 0$. We can define the means $\mu'(x)$ to be

$$\mu'(x) := \begin{cases} \mu(x^*) + \epsilon & x = x_0 \\ \mu(x) & \text{otherwise} \end{cases} \ .$$

Note then that $\arg\max_x \mu'(x) = x_0$, and $\mu(x_0) = \mu(x^*) + \epsilon$. Hence, Eq. (7.28) holds for the means $\mu'(\cdot)$. Moreover, $\mu(x) = \mu'(x)$ for all $x \neq x_0$, so that $\mathrm{KL}(\mu(x), \mu'(x)) = 0$ for $x \neq x_0$ and Eq. (7.28) simplifies to

$$T(x_0)\mathrm{KL}(\mu'(x_0), \mu'(x^*) + \epsilon) = T(x_0)\mathrm{KL}(\mu(x_0), \mu'(x_0)) \geq c_1 \log(1/\delta) \ .$$

Since $\mathrm{KL}(\mu'(x_0), \mu'(x^*) + \epsilon)$ is continuous in ϵ (see Fact 7.1), taking $\epsilon \to 0$ yields

$$T(x_0)\mathrm{KL}(\mu'(x_0), \mu'(x^*)) \geq c_1 \log(1/\delta) \ \text{ as needed.}$$

□

Analysis for Top-k Poisson Emitters

We now continue our analysis of `AdaSearch`, addressing the full problem of identifying the k Poisson emitters with the highest emission rates. Our goal is to identify the unique set

$$\mathcal{S}^*(k) := \{x \in \mathcal{S} : \mu(x) \geq \mu^{(k)}\} \ . \tag{7.29}$$

To ensure the top-k emitters are unique, we assume that $\mu^{(k)} > \mu^{(k+1)}$ (recall that $\mu^{(k)}$ denotes the k-th largest value of $\mu(x)$ among all $x \in \mathcal{S}$). The complexity of identifying the top-k emitters can then described in terms of the gaps of the divergence terms

$$\mathrm{d}(\mu^{(k+1)}, \mu(x)) = \frac{(\mu(x) - \mu^{(k+1)})^2}{\mu(x)}, \ (\mu(x) > \mu^{(k+1)}) \quad \text{and}$$
$$\mathrm{d}(\mu(x), \mu^{(k)}) = \frac{(\mu^{(k)} - \mu(x))^2}{\mu^{(k)}}, \quad (\mu(x) < \mu^{(k)}).$$

For $x \in \mathcal{S}^*(k)$, $\mathrm{d}(\mu^{(k+1)}, \mu(x))$ describes how close the emission rate $\mu(x)$ is to the "best" alternative in $\mathcal{S} \setminus \mathcal{S}^*(k)$. For $x \in \mathcal{S} \setminus \mathcal{S}^*(k)$, $\mathrm{d}(\mu(x), \mu^{(k)})$ describes how close $\mu(x)$ is to the mean $\mu^{(k)}$ of the emitter in $\mathcal{S}^*(k)$ from which it is hardest to distinguish. The analogues of $\mathcal{C}_{\texttt{adapt}}$ and $\mathcal{C}_{\texttt{unif}}$ are then

$$\mathcal{C}^{(k)}_{\texttt{adapt}} := \sum_{x \in \mathcal{S}^*(k)} \frac{1}{\mathrm{d}(\mu^{(k+1)}, \mu(x))} + \sum_{x \in \mathcal{S} \setminus \mathcal{S}^*(k)} \frac{1}{\mathrm{d}(\mu(x), \mu^{(k)})} \quad \text{and} \tag{7.30}$$

$$\begin{aligned} \mathcal{C}^{(k)}_{\texttt{unif}} &:= |\mathcal{S}| \cdot \max\left\{ \max_{x \in \mathcal{S}^*(k)} \frac{1}{\mathrm{d}(\mu^{(k+1)}, \mu(x))}, \max_{x \in \mathcal{S} \setminus \mathcal{S}^*(k)} \frac{1}{\mathrm{d}(\mu(x), \mu^{(k)})} \right\} \\ &= |\mathcal{S}| \mathrm{d}(\mu^{(k+1)}, \mu^{(k)}) \end{aligned} \tag{7.31}$$

where the equality follows by noting that the function $(x, a) \mapsto \frac{x}{(x-a)^2}$ is decreasing in x and increasing in a for $x > a$. The following theorem, proved in Section 7.B, provides an upper bound on the sample complexity for top-k identification.

Theorem 7.5. *For any $\delta \in (0, 1)$, the following holds with probability at least $1 - \delta$:* `AdaSearch` *correctly returns $\mathcal{S}^*(k)$, the total sample complexity is bounded above by*

$$\begin{aligned} T^{\mathrm{sample}} &\lesssim |\mathcal{S}| \sum_{x \in \mathcal{S}^*(k)} \frac{\log_+\left(|\mathcal{S}| \log_+\left(\frac{1}{\mathrm{d}(\mu^{(k+1)}, \mu(x))} \right) / \delta \right)}{\mathrm{d}(\mu^{(k+1)}, \mu(x))} \\ &\qquad + \sum_{x \in \mathcal{S} \setminus \mathcal{S}^*(k)} \frac{\mu^{(k)} \log_+\left(|\mathcal{S}| \log_+\left(\frac{1}{\mathrm{d}(\mu(x), \mu^{(k)})} \right) / \delta \right)}{\mathrm{d}(\mu(x), \mu^{(k)})} \\ &= \widetilde{\mathcal{O}}\left(\mathcal{C}^{(k)}_{\texttt{adapt}} \cdot \log(|\mathcal{S}|/\delta) \right), \end{aligned}$$

and the total runtime is bounded by

$$T^{\mathrm{run}} \lesssim T^{\mathrm{sample}} + |\mathcal{S}| \log_+\left(\frac{\log_+ \frac{|\mathcal{S}|}{\delta}}{\mathrm{d}(\mu^{(k+1)}, \mu^{(k)})} \right) = \widetilde{\mathcal{O}}\left(\mathcal{C}^{(k)}_{\texttt{adapt}} \cdot \log \tfrac{|\mathcal{S}|}{\delta} + |\mathcal{S}| \log_+\left(\tfrac{\mathcal{C}^{(k)}_{\texttt{unif}}}{|\mathcal{S}|} \right) \right) .$$

We remark that our sample complexity qualitatively matches standard bounds for active top-k identification with sub-Gaussian rewards in the non-embodied setting (see, e.g. [154]). Our results differ by considering the appropriate modifications for Poisson emissions, as well as accounting for total travel time. Lastly, we have the bound for uniform sampling for the case of general k.

Theorem 7.6. *For any $\delta \in (0, 1)$, the following holds with probability at least $1 - \delta$:* `NaiveSearch` *correctly returns $\mathcal{S}^*(k)$, and the total runtime is bounded by bounded above by*

$$T^{\mathrm{run}} \lesssim \widetilde{\mathcal{O}}\left(\mathcal{C}_{\texttt{unif}} \right) \cdot \log(|\mathcal{S}|/\delta) .$$

7.B Analyzing AdaSearch: Proof of Theorem 7.5

Analysis Roadmap

To simplify the analysis, we assume that at round i, we take a fresh $\tau_i = 2^i$ samples (recall we have normalized $\tau_0 = 1$) from each remaining $x \in \mathcal{S}_i$.[5] For $x \in \mathcal{S}_i$, $\mathbf{N}_i(x)$ denotes the number of counts observed from point x over the interval of length τ_i, and $\widehat{\mu}_i(x)$ denotes the empirical average emissions; that is, $\widehat{\mu}_i(x) = \mathbf{N}_i(x)/\tau_i$. With this notation, our confidence intervals take the following form:

$$\mathrm{LCB}_i(x) := \frac{1}{2^i} U_- \left(\mathbf{N}_i(x), \frac{\delta}{4|\mathcal{S}|i^2} \right) \text{ and } \mathrm{UCB}_i(x) := \frac{1}{2^i} U_+ \left(\mathbf{N}_i(x), \frac{\delta}{4|\mathcal{S}|i^2} \right) . \tag{7.32}$$

We first argue that there exists a good event, $\mathcal{E}_{\mathrm{good}}(\delta)$, occurring with probability at least $1-\delta$, on which the true mean $\mu(x)$ of each pixel x lies between $\mathrm{LCB}_i(x)$ and $\mathrm{UCB}_i(x)$ for all i. Moreover, $\mathrm{LCB}_i(x)$ and $\mathrm{UCB}_i(x)$ are contained within the interval defined by $[\overline{\mathrm{UCB}}_i(x), \overline{\mathrm{LCB}}_i(x)]$, which depends explicitly upon $\mu(x)$, but *not* on $\widehat{\mu}_i(x)$. To derive $\overline{\mathrm{UCB}}_i(x)$ and $\overline{\mathrm{LCB}}_i(x)$, we begin by deriving high probability upper and lower bounds $\overline{U}_+(\mu, \delta)$ and $\overline{U}_-(\mu, \delta)$ for the functions $U_+(\mathbf{N}, \delta)$ and $U_-(\mathbf{N}, \delta)$ that hold for Poisson random variables. Formally, we have the following

Proposition 7.3. *Let $\mu \geq 0$ and let and $\mathbf{N} \sim \mathrm{Poisson}(\mu)$. Define*

$$\begin{aligned} \overline{U}_+(\mu, \delta) &:= \mu + \frac{14}{3} \log(1/\delta) + 2\sqrt{2\mu \log(1/\delta)} \text{ and} \\ \overline{U}_-(\mu, \delta) &:= \max\left\{ 0, \mu - 2\sqrt{2\mu \log(1/\delta)} \right\} . \end{aligned}$$

Then, it holds that

$$\mathbb{P}\left[\overline{U}_-(\mu, \delta) \leq U_-(\mathbf{N}, \delta) \leq \mu \leq U_+(\mathbf{N}, \delta) \leq \overline{U}_+(\mu, \delta) \right] \geq 1 - 2\delta . \tag{7.33}$$

Proof of Proposition 7.3. From the proof of Proposition 7.1, recall the events

$$\begin{aligned} \mathcal{E}_+(\delta) &:= \{ \mathbf{N} \geq \mu - \sqrt{2\mu \log(1/\delta)} \} \quad \text{and} \\ \mathcal{E}_-(\delta) &:= \{ \mathbf{N} \leq \mu + \sqrt{2\mu \log(1/\delta)} + \frac{2}{3} \log(1/\delta) \} . \end{aligned}$$

Further, recall that on $\mathcal{E}_+(\delta)$, we have $\mu \leq U_+(\mathbf{N}, \delta)$ and $\mu \geq U_-(\mathbf{N}, \delta)$. We now show that on $\mathcal{E}_-(\delta)$, we also have $U_+(\mathbf{N}, \delta) \leq \overline{\mathrm{UCB}}(\mu, \delta)$, and on $\mathcal{E}_+(\delta)$, we have $U_-(\mathbf{N}, \delta) \geq \overline{\mathrm{LCB}}(\mu, \delta)$.

[5] The analysis is nearly the same as if we used the total $2^{i+1} - 1$ samples collected throughout.

To bound $U_-(\mathbf{N},\delta) \geq \overline{U}_-(\mu,\delta)$, observe that $\mathbf{N} \mapsto U_-(\mathbf{N},\delta)$ is increasing in $\mathbf{N}$, and on $\mathcal{E}_+(\delta)$, one has $\{\mathbf{N} \geq \mu - \sqrt{2\mu\log(1/\delta)}\}$. Thus

$$\begin{aligned} U_-(\mathbf{N},\delta) &:= \mathbf{N} - \sqrt{2\mathbf{N}\log(1/\delta)} \\ &\geq \mu - \sqrt{2\mu\log(1/\delta)} - \sqrt{2\log(1/\delta)\mu(\mu - \sqrt{2\mu\log(1/\delta)})} \\ &\geq \mu - \sqrt{2\mu\log(1/\delta)} - \sqrt{2\mu\log(1/\delta)} \\ &\geq \mu - 2\sqrt{2\mu\log(1/\delta)} := \overline{U}_-(\mu,\delta) \ . \end{aligned}$$

Next, we prove the bound bound $U_+(\mathbf{N},\delta) \leq \overline{U}_+(\mu,\delta)$. On $\mathcal{E}_-(\delta)$, we have

$$\begin{aligned} \mathbf{N} &\leq \mu + \frac{2}{3}\log(1/\delta) + \sqrt{2\mu\log(1/\delta)} \\ &= (\mu^{1/2} + \sqrt{2\log(1/\delta)})^2 - 2\log(1/\delta) + \frac{2}{3}\log(1/\delta) \\ &\leq (\mu^{1/2} + \sqrt{2\log(1/\delta)})^2 \ . \end{aligned}$$

Hence, when the above occurs, we have

$$\begin{aligned} U_+(\mathbf{N},\delta) &= 2\log(1/\delta) + \mathbf{N} + \sqrt{2\mathbf{N}\log(1/\delta)} \\ &= \frac{8}{3}\log(1/\delta) + \mu + \sqrt{2\mu\log(1/\delta)} + \sqrt{2((\mu^{1/2} + \sqrt{2\log(1/\delta)})^2)\log(1/\delta)} \\ &\leq \frac{8}{3}\log(1/\delta) + \mu + \sqrt{2\mu\log(1/\delta)} + \sqrt{2\mu\log(1/\delta)} + 2\log(1/\delta) \\ &= \frac{14}{3}\log(1/\delta) + \mu + 2\sqrt{2\mu\log(1/\delta)} \ . \end{aligned}$$

□

As a consequence of Proposition 7.3, we can show that $\overline{\mathrm{LCB}}_i(x)$ and $\overline{\mathrm{UCB}}_i(x)$ are probabilistic lower and upper bounds on $\mathrm{LCB}_i(x)$ and $\mathrm{UCB}_i(x)$:

Lemma 7.4. *Introduce the confidence intervals*

$$\overline{\mathrm{LCB}}_i(x) := \frac{1}{\tau_i}\overline{U}_-\left(\tau_i\mu(x), \frac{\delta}{4|\mathcal{S}|i^2}\right) \text{ and } \overline{\mathrm{UCB}}_i(x) := \frac{|S|}{\tau_i}\overline{U}_+\left(\tau_i\mu(x), \frac{\delta}{4|\mathcal{S}|i^2}\right) \ .$$

Then, there exists an event $\mathcal{E}_{\mathrm{good}}$ for which $\mathbb{P}[\mathcal{E}_{\mathrm{good}}] \geq 1-\delta$, and

$$\forall i \geq 1, x \in \mathcal{S}_i : \overline{\mathrm{LCB}}_i(x) \leq \mathrm{LCB}_i(x) \leq \mu(x) \leq \mathrm{UCB}_i(x) \leq \overline{\mathrm{UCB}}_i(x). \tag{7.34}$$

Proof of Lemma 7.4. Lemma 7.4 is a simple consequence of Propositions 7.1 and 7.3, and a union bound:

$$\begin{aligned} &\mathbb{P}\left[\exists x \in \mathcal{S}_0, i \geq 1 : \left\{\overline{\mathrm{LCB}}_i(x) \leq \mathrm{LCB}_i(x) \leq \mu(x) \leq \mathrm{UCB}_i(x) \leq \overline{\mathrm{UCB}}_i(x)\right\} \text{ fails}\right] \\ &\overset{\text{union bound}}{\leq} \sum_{x\in\mathcal{S}_0, i\geq 1} \mathbb{P}\left[\left\{\overline{\mathrm{LCB}}_i(x) \leq \mathrm{LCB}_i(x) \leq \mu(x) \leq \mathrm{UCB}_i(x) \leq \overline{\mathrm{UCB}}_i(x)\right\} \text{ fails}\right] \\ &\overset{\text{Prop. 7.1\&7.3}}{\leq} \sum_{x\in\mathcal{S}_0, i\geq 1} \frac{\delta}{2|\mathcal{S}|i^2} = |\mathcal{S}| \sum_{i\geq 1} \frac{\delta}{2|\mathcal{S}|i^2} = \frac{\delta}{2}\sum_i i^{-2} \leq \delta \ . \end{aligned}$$

□

Note that on $\mathcal{E}_{\text{good}}$, one has that $\mu(x) \in [\text{LCB}_i(x), \text{UCB}_i(x)]$ for all rounds i and all $x \in \mathcal{S}_i$; hence, as a result of Lemma 7.1,

Lemma 7.5. *If $\mathcal{E}_{\text{good}}$ holds, then for all rounds i, $\mathcal{S}_i^{\text{top}} \subset \mathcal{S}^*(k) \subset \mathcal{S}_i^{\text{top}} \cup \mathcal{S}_i$; in particular, if* `AdaSearch` *terminates at round i_{fin}, then it correctly returns $\mathcal{S}^*(k)$.*

Finally, the next lemma gives a *deterministic* condition under which a point $x \in \mathcal{S}$ can be removed from $\mathcal{S}_i$, in terms of the deterministic confidence bounds $\overline{\text{LCB}}_i(x)$ and $\overline{\text{UCB}}_i(x)$.

Lemma 7.6. *Suppose $\mathcal{E}_{\text{good}}$ holds. Let $x^{(k)}$ and $x^{(k+1)}$ denote arbitrary points in $\mathcal{S}$ with $\mu(x^{(k)}) = \mu^{(k)}$ and $\mu(x^{(k+1)}) = \mu^{(k+1)}$. Define the function*

$$i_{\text{fin}}(x) := \begin{cases} \inf\{i : \overline{\text{LCB}}_i(x) > \overline{\text{UCB}}_i(x^{(k+1)})\} & x \in \mathcal{S}^*(k) \\ \inf\{i : \overline{\text{UCB}}_i(x) < \overline{\text{LCB}}_i(x^{(k)})\} & x \in \mathcal{S} \setminus \mathcal{S}^*(k) \end{cases}.$$

Then, on $\mathcal{E}_{\text{good}}$, $x \notin \mathcal{S}_i$ for all $i > i_{\text{fin}}(x)$.

Proof of Lemma 7.6. Assume $\mathcal{E}_{\text{good}}$ holds, and let $x \in \mathcal{S}$, and set $i = i_{\text{fin}}(x)$. Then

(a) If $x \in \mathcal{S}^*(k)$, then $\overline{\text{LCB}}_i(x) > \overline{\text{UCB}}_i(x^{(k+1)})$. In this case, we shall show that x will be added to $\mathcal{S}_i^{\text{top}}$ via (7.4).

(b) If $x \in \mathcal{S} \setminus \mathcal{S}^*(k)$, then $\overline{\text{UCB}}_i(x) < \overline{\text{LCB}}_i(x^{(k)})$. In this case, we shall show that x will be removed from $\mathcal{S}_i$ via (7.5).

Case (a): $\overline{\text{LCB}}_i(x) > \overline{\text{UCB}}_i(x^{(k+1)})$. By (7.4), x is added to $\mathcal{S}_{i+1}^{\text{top}}$ if $\text{LCB}_i(x)$ is larger than all but $k - |\mathcal{S}_i^{\text{top}}|$ values of $\text{UCB}_i(x')$, $x' \in \mathcal{S}_i$. Since $\text{LCB}_i(x) \geq \overline{\text{LCB}}_i(x)$ and $\text{UCB}_i(x') \leq \overline{\text{UCB}}_i(x')$ on $\mathcal{E}_{\text{good}}$, it is enough that

$$\overline{\text{LCB}}_i(x) > (k - |\mathcal{S}_i^{\text{top}}| + 1)\text{-st largest value of } \overline{\text{UCB}}_i(x'),\ x' \in \mathcal{S}_i.$$

We now observe that $\overline{\text{UCB}}_i(x')$ is monotonic in $\mu(x')$. Hence, it is enough that

$$\overline{\text{LCB}}_i(x) > \overline{\text{UCB}}_i(x_+),\ \text{where } \mu(x_+) = (k - |\mathcal{S}_i^{\text{top}}| + 1)\text{-st largest value of } \mu(x'),\ x' \in \mathcal{S}_i.$$

But since there are exactly $k - |\mathcal{S}^{\text{top}}|$ elements of $\mathcal{S}^*(k)$ in $\mathcal{S}_i$ by Lemma 7.5, x_+ is not among the top k, and thus $\mu(x_+) \leq \mu(x^{(k+1)})$. Hence, $\overline{\text{UCB}}_i(x_+) \leq \overline{\text{UCB}}_i(\mu(x^{(k+1)}))$, so it is enough that $\overline{\text{LCB}}_i(x) > \overline{\text{UCB}}_i(\mu(x^{(k+1)}))$.

Case (b): $\overline{\text{UCB}}_i(x) < \overline{\text{LCB}}_i(x^{(k)})$. Following the reasoning of case (a) applied to (7.5), we can see that it is enough that

$$\overline{\text{UCB}}_i(x) < \overline{\text{LCB}}_i(x_-),\ \text{where } \mu(x_-) = (k - |\mathcal{S}_{i+1}^{\text{top}}|)\text{-st largest value of } \mu(x),\ x' \in \mathcal{S}_i \setminus \mathcal{S}_{i+1}^{\text{top}}.$$

Lemma 7.5 ensures that there are $(k - |\mathcal{S}_{i+1}^{\text{top}}|)$ members of $\mathcal{S}^*(k)$ in $\mathcal{S}_i \setminus \mathcal{S}_{i+1}^{\text{top}}$, so $\mu(x_-) \geq \mu(x^{(k)})$. Hence, it is enough that $\overline{\text{UCB}}_i(x) < \overline{\text{LCB}}_i(x^{(k)})$. □

In view of Lemma 7.6, we can bound

$$\begin{aligned} T^{\text{sample}} := \sum_{i=0}^{i_{\text{fin}}} \tau_i |\mathcal{S}_i| \;&=\; \sum_{x\in\mathcal{S}} \sum_{i\geq 0} \tau_i \mathbb{I}(x \in \mathcal{S}_i) \;\leq\; \sum_{x\in\mathcal{S}} \sum_{i=0}^{i_{\text{fin}}(x)} \tau_i \quad \text{(by Lemma 7.6)} \\ &\leq \sum_{x\in\mathcal{S}} 2^{i_{\text{fin}}(x)+1} \quad \text{(since } \tau_i = 2^i\text{)} \end{aligned} \tag{7.35}$$

and further, bound

$$\begin{aligned} T^{\text{run}} := T^{\text{sample}} + \sum_{i=0}^{i_{\text{fin}}} \tau_0 |\mathcal{S} \setminus \mathcal{S}_i| \;&\leq\; T^{\text{sample}} + |\mathcal{S}| i_{\text{fin}} \quad \text{(recall } \tau_0 = 1\text{)} \\ &\leq T^{\text{sample}} + |\mathcal{S}| \max_{x\in\mathcal{S}} i_{\text{fin}}(x). \end{aligned} \tag{7.36}$$

where the last line uses Lemma 7.6. Lastly, we prove an upper bound on $i_{\text{fin}}(x)$ for all $x \in \mathcal{S}$.

Proposition 7.4. *There exists a universal constant $C > 1$ such that, for $x \in \mathcal{S}^*(k)$,*

$$2^{i_{\text{fin}}(x)} \leq C \cdot \left\{ 1 + \frac{\log_+\left(|\mathcal{S}| \log_+\left(\frac{1}{\mathrm{d}(\mu^{(k+1)}, \mu(x))} \right) / \delta \right)}{\mathrm{d}(\mu^{(k+1)}, \mu(x))} \right\}$$

whereas for $x \in \mathcal{S} \setminus \mathcal{S}^*(k)$,

$$2^{i_{\text{fin}}(x)} \leq C \cdot \left\{ 1 + \frac{\log_+\left(|\mathcal{S}| \log_+\left(\frac{1}{\mathrm{d}(\mu^{(k+1)}, \mu(x))} \right) / \delta \right)}{\mathrm{d}(\mu^{(k+1)}, \mu(x))} \right\}$$

The proof of Proposition 7.4 will invoke an inversion lemma from the best arm identification literature (see, e.g. Equation (110) in [260]):

Lemma 7.7. *For any $\delta, u > 0$, let $\mathcal{T}(u, \delta) := 1 + \log_+(\delta^{-1} \log_+(u))/u$. There exists a universal constant C_0 such that, for all $n \geq C_0 \mathcal{T}(u, \delta)$, we have $\log(\delta^{-1} \log n)/n < u$.*

Proof of Proposition 7.4. Let $n = 2^i$, let $\delta_0 = \delta/(4|\mathcal{S}| \log_2 e)$, and let $\Delta = \mu(x_1) - \mu(x_2)$. Then $\overline{\mathrm{UCB}}_i(x_2) < \overline{\mathrm{LCB}}_i(x_1)$ is equivalent to

$$\begin{aligned} \mu(x_2) + \frac{14}{3n} \log\left(\delta_0^{-1} \log n\right) + 2\sqrt{2\mu(x_2) \log\left(\delta_0^{-1} \log n\right)/n} \\ < \mu(x_1) - 2\sqrt{2\mu(x_1) \log\left(\delta_0^{-1} \log n\right)/n} \\ \text{implied by} \quad \frac{14}{3n} \log\left(\delta_0^{-1} \log n\right) + 4\sqrt{2\mu(x_1) \log\left(\delta_0^{-1} \log n\right)/n} < \Delta \;, \end{aligned}$$

where the second line uses $\mu(x_2) \leq \mu(x_1)$. For the second line to hold, it is enough that

$$\frac{1}{n}\log\left(\delta_0^{-1}\log n\right) < \frac{3\Delta}{28} \text{ and } \log\left(\delta_0^{-1}\log n\right)/n < \left(\frac{\Delta}{8\sqrt{2}}\right)^2 . \tag{7.37}$$

Lemma 7.7 and (7.37) imply that it is sufficent that

$$n \geq C_0 \mathcal{T}\left(\min\left\{\frac{3\Delta}{28}, \frac{1}{\mu(x_1)}\left(\frac{\Delta}{8\sqrt{2}}\right)^2\right\}, \delta_0\right) ,$$

from which it follows that

$$\inf\left\{2^i : \overline{\mathrm{UCB}}_i(x_2) < \overline{\mathrm{LCB}}_i(x_1)\right\} \leq 2C_0\left\{\mathcal{T}\left(\min\left\{\frac{3\Delta}{28}, \frac{1}{\mu(x_1)}\left(\frac{\Delta}{8\sqrt{2}}\right)^2\right\}, \delta_0\right)\right\} .$$

Absorbing constants and plugging in $\delta_0 = \delta/(4|\mathcal{S}|\log_2 e)$, algebraic manipulation finally implies that there exists a universal constant C such that, for any x_1, x_2 with $\mu(x_1) < \mu(x_2)$,

$$\begin{aligned}
\inf\left\{2^i : \overline{\mathrm{UCB}}_i(x_1) < \overline{\mathrm{LCB}}_i(x_2)\right\} &\leq C\left\{1 + \mathcal{T}\left(\min\left\{\Delta, \frac{\Delta^2}{\mu(x_1)}\right\}, \delta/M\right)\right\} \\
&= C\left\{1 + \mathcal{T}\left(\frac{\Delta^2}{\mu(x_1)}, \delta/|\mathcal{S}|\right)\right\} \\
&= C\left\{1 + \mathcal{T}\left(\mathrm{d}(\mu(x_2), \mu(x_1)), \delta/|\mathcal{S}|\right)\right\} .
\end{aligned}$$

To conclude, we select $x_1 = x^{(k+1)}$ and $x_2 = x$ when $x \in \mathcal{S}^*(k)$, and $x_1 = x$ and $x_2 = x^{(k)}$ for $x \in \mathcal{S} \setminus \mathcal{S}^*(k)$. □

Theorem 7.5 now follows by plugging in Proposition 7.4 into Eq.'s (7.35) and (7.36).

7.C Analyzing `NaiveSearch`: Proof of Theorem 7.4

In this section, we present a brief proof of Theorem 7.4 and Theorem 7.6. The arguments are quite similar to those in the analysis of `AdaSearch`, and we point out modifications as we go allow.

Let $\mathcal{E}_{\mathrm{good}}$ denote the event of Lemma 7.4, modified to hold for all $x \in \mathcal{S}$ at each round i (rather than all $x \in \mathcal{S}_i$, as in the case of `AdaSearch`). The proof of Lemma 7.4 extends to this case as well, yielding that

$$\mathbb{P}[\mathcal{E}_{\mathrm{good}}] \geq 1 - \delta .$$

It suffices to show that on $\mathcal{E}_{\mathrm{good}}$, `NaiveSearch` correctly returns $\mathcal{S}^*(k)$, and satisfies the desired runtime guarantees.

Correctness: On $\mathcal{E}_{\text{good}}$, we have that $x \in \mathcal{S}^*(k)$, $\mu(x) \leq \text{UCB}_i(x)$, and for $x' \in \mathcal{S} \setminus \mathcal{S}^*(k)$, $\mu(x') \geq \text{LCB}_i(x')$. Hence, for any $x' \in \mathcal{S} \setminus \mathcal{S}^*(k)$ and for all $x \in \mathcal{S}^*(k)$, $\text{LCB}_i(x') \leq \mu(x') < \mu(x) \leq \text{UCB}_i(x)$. Thus, the termination criterion can only fulfilled when $\text{LCB}_i(x)$, each $x \in \mathcal{S}^*(k)$, are greater than the remained $|\mathcal{S}| - k$ values of $\text{UCB}_i(x)$. This yields correctness.

Runtime: Recall that for `NaiveSearch` with the standardization τ_0, we have

$$T^{\text{run}} = |\mathcal{S}| \sum_{i=0}^{i_{\text{fin}}} \tau_i = |\mathcal{S}| \sum_{i=0}^{i_{\text{fin}}} 2^i \leq 2|\mathcal{S}| \cdot 2^{i_{\text{fin}}} \ .$$

Arguing as in the analysis for `AdaSearch`, it suffices to show that, on $\mathcal{E}_{\text{good}}$,

$$i_{\text{fin}} \leq \inf\{i : \overline{\text{LCB}}_i(x) > \overline{\text{UCB}}_i(x') \quad \forall x \in \mathcal{S}^*(k), x' \in \mathcal{S} \setminus \mathcal{S}^*(k)\} := \overline{i_{\text{fin}}}, \tag{7.38}$$

for which we bound

$$\begin{aligned} T^{\text{run}} &\lesssim |\mathcal{S}| 2^{\overline{i_{\text{fin}}}} \\ &\overset{(i)}{\lesssim} |\mathcal{S}| \cdot \max_{x \in \mathcal{S}^*(k), x' \in \mathcal{S} - \mathcal{S}^*(k)} \frac{\log_+\left(\frac{|\mathcal{S}|}{\delta} \log_+(\frac{1}{\text{d}(\mu(x'), \mu(x))})\right)}{\text{d}(\mu(x'), \mu(x))} \\ &= |\mathcal{S}| \cdot \frac{\log_+\left(\frac{|\mathcal{S}|}{\delta} \log_+\left(\frac{1}{\text{d}(\mu^{(k+1)}, \mu^{(k)})}\right)\right)}{\text{d}(\mu^{(k+1)}, \mu^{(k)})} \\ &= \widetilde{O}(\mathcal{C}_{\texttt{unif}} \log(|\mathcal{S}|/\delta)), \end{aligned}$$

where (i) follows from the same argument as in the proof of Proposition 7.4. To verify (7.38), suppose that $\mathcal{E}_{\text{good}}$ holds, and that `NaiveSearch` has not terminated before round $\overline{i_{\text{fin}}}$. Then, by definition of $\overline{i_{\text{fin}}}$,

$$\overline{\text{LCB}}_i(x) > \overline{\text{UCB}}_i(x'), \ \forall x \in \mathcal{S}^*(k), x' \in \mathcal{S} \setminus \mathcal{S}^*(k) \ .$$

Moreover, on $\mathcal{E}_{\text{good}}$, $\text{LCB}_i(x) \geq \overline{\text{LCB}}_i(x)$ and $\text{UCB}_i(x') \leq \overline{\text{UCB}}_i(x')$ for all $x \in \mathcal{S}^*(k)$ and all $x' \in \mathcal{S} \setminus \mathcal{S}^*(k)$. Thus,

$$\text{LCB}_i(x) > \text{UCB}_i(x'), \ \forall x \in \mathcal{S}^*(k), x' \in \mathcal{S} \setminus \mathcal{S}^*(k),$$

which directly implies the termination criterion for `NaiveSearch`.

Chapter 8

MOSAIKS: A Generalizable and Accessible Approach to Machine Learning with Global Satellite Imagery

This chapter is based on the paper "A generalizable and accessible approach to machine learning with global satellite imagery" [245][1], written in collaboration with Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang.

Combining satellite imagery with machine learning (SIML) has the potential to address global challenges by remotely estimating socioeconomic and environmental conditions in data-poor regions, yet the resource requirements of SIML limit its accessibility and use. We show that a single encoding of satellite imagery can generalize across diverse prediction tasks (e.g. forest cover, house price, road length). Our method achieves accuracy competitive with deep neural networks at orders of magnitude lower computational cost, scales globally, delivers label super-resolution predictions, and facilitates characterizations of uncertainty. Since image encodings are shared across tasks, they can be centrally computed and distributed to unlimited researchers, who need only fit a linear regression to their own ground truth data in order to achieve state-of-the-art SIML performance.

8.1 Background

Addressing complex global challenges—such as managing global climate changes, population movements, ecosystem transformations, or economic development—requires that many different researchers and decision-makers (hereafter, *users*) have access to reliable, large-scale observations of many variables simultaneously. Planet-scale ground-based monitoring systems are generally prohibitively costly for this purpose, but satellite imagery presents a

[1]Reproduced with permission from Springer Nature

viable alternative for gathering globally comprehensive data, with over 700 earth observation satellites currently in orbit [280]. Application of machine learning is proving to be an effective approach for transforming these vast quantities of unstructured imagery data into structured estimates of ground conditions. For example, combining satellite imagery and machine learning (SIML) has enabled better characterization of forest cover [123], land use [141], poverty rates [145] and population densities [241], thereby supporting research and decision-making. We refer to such prediction of an individual variable as a *task*. Demand for SIML-based estimates is growing, as indicated by the large number of private service-providers specializing in predicting one or a small number of these tasks.

The resource requirements for deploying SIML technologies, however, limit their accessibility and usage. Satellite-based measurements are particularly under-utilized in low-income contexts, where the technical capacity to implement SIML may be low, but where such measurements would likely convey the greatest benefit [118, 299]. For example, government agencies in low-income settings might want to understand local waterway pollution, illegal land uses, or mass migrations. SIML, however, remains largely out of reach to these and other potential users because current approaches require a major resource-intensive enterprise, involving a combination of task-specific domain knowledge, remote sensing and engineering expertise, access to imagery, customization and tuning of sophisticated machine learning architectures, and large computational resources [21].

To remove many of these barriers, we develop a new approach to SIML that enables non-experts to obtain state-of-the-art performance without manipulating imagery, using specialized computational resources, or developing a complex prediction procedure. We design a one-time, task-agnostic encoding that transforms each satellite image into a vector of variables (hereafter, features). We then show that these features (x) perform well at predicting ground conditions (y) across diverse tasks, using only a linear regression implemented on a personal computer. Prior work has similarly sought an unsupervised encoding of satellite imagery [65, 146, 222, 249]; however, to the best of our our knowledge, we are the first to demonstrate that a single set of features both achieves performance competitive with deep-learning methods across a variety of tasks and scales globally.

We focus here on the problem of predicting properties of small regions (e.g. average house price) at a single time period, using high-resolution daytime satellite imagery as the only input. We develop a simple yet high-performing system that is tailored to address the challenges and opportunities specific to SIML applications. We achieve large computational efficiency gains in model training and testing, relative to deep neural networks, through algorithmic simplifications that take advantage of the fact that satellite images are collected from a fixed distance and viewing angle and capture repeating patterns and objects. This contrasts with deep-learning approaches to SIML that use techniques originally developed for natural images (e.g. photos taken from handheld cameras), where inconsistency in many key factors, such as subject or camera perspective, require complex solutions that our results suggest may be mostly unnecessary for SIML applications.

Here we show that a single set of general purpose features can encode rich information in satellite images. We utilize an unsupervised featurization process, which separates feature

construction from model-fitting. This approach dramatically increases computational speed for any given researcher and delivers large computational gains at the research-system level by reorganizing how imagery is processed and distributed. Traditionally, hundreds or thousands of researchers use the same images to solve different and unrelated tasks. Our approach allows common sources of imagery to be converted into centralized sets of features that can be accessed by many researchers, each solving different tasks. This isolates future users from the costly steps of obtaining, storing, manipulating, and processing imagery themselves. The magnitude of the resulting benefits grow with the size of the expanding SIML user community and the scale of global imagery data, which currently increases by more than 80TB/day [177].

8.2 Multi-Task Observation using Satellite Imagery & Kitchen Sinks

Our objective is to enable any user with basic resources to predict ground conditions using only satellite imagery and a limited sample of task-specific ground truth data which they possess. Our SIML system, "Multi-task Observation using Satellite Imagery and Kitchen Sinks" (MOSAIKS), separates the prediction procedure into two independent steps: a fixed "featurization step" which translates satellite imagery into succinct vector representations ($images \rightarrow x$), and a "regression step" which learns task-specific coefficients that map these features to outcomes for a given task ($x \rightarrow y$). For each image, the unsupervised featurizaton step can be centrally executed once, producing one set of outputs that are used to solve many different tasks through repeated application of the regression step by multiple independent users (Figure 8.1). Because the regression step requires relatively little computation, MOSAIKS scales efficiently across unlimited users and tasks.

The *accessibility* of our approach stems from the simplicity and computational efficiency of the regression step for potential users, given features which are already computed once and stored centrally. To generate SIML predictions, a user of MOSAIKS (i) queries these tabular data for a vector of K features for each of their N locations of interest; (ii) merges these features x with label data y, i.e. the user's independently collected ground truth data; (iii) implements a linear regression of y on x to obtain coefficients β; (iv) uses coefficients β and and features x to predict labels $\hat{y}$ in new locations where imagery and features are available but ground truth data are not (Figure 8.1).

The *generalizability* of our approach means that a single mathematical summary of satellite imagery (x) performs well across many prediction tasks ($y_1, y_2, ...$) without any task-specific modification to the procedure. The success of this generalizability relies on how images are encoded as features, which we describe below.

In contrast to many recent alternative approaches to SIML, MOSAIKS does not require training or using the output of a deep neural network and encoding images into unsupervised features requires no labels. Nonetheless, MOSAIKS achieves competitive performance at a

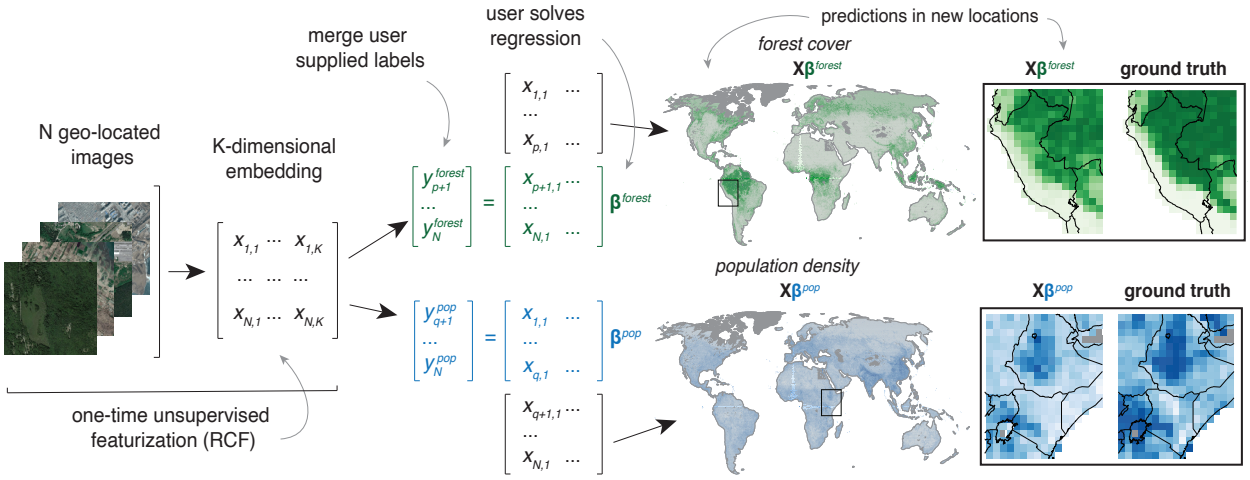


Figure 8.1: Schematic of the MOSAIKS process. N images are transformed using random convolutional features (RCF) into K-dimensional feature vectors before labels are known. Once the feature matrix $\mathbf{X} \in \mathbb{R}^{N \times K}$ is computed, it can be used for unlimited tasks without recomputation. Users interested in a new task s (e.g. forest cover or population density) merge their own labels y^s to features for training.

large computational advantage that grows linearly with the number of SIML users and tasks, due to shared computation and storage. In principle, any unsupervised featurization would enable these computational gains. However, to date, a single set of unsupervised features has neither achieved accuracy competitive with supervised CNN-based approaches across many SIML tasks, nor at the scale that we study. In Sections 8.3 to 8.5, we show that MOSAIKS achieves a practical level of generalization and effectiveness in real world contexts.

Random Convolutional Features

We design a featurization function by building on the theoretically grounded machine learning concept of random kitchen sinks [231], which we apply to satellite imagery by constructing random convolutional features (RCFs). RCFs are suitable for the structure of satellite imagery and have established performance encoding genetic sequences [204], classifying photographs [72], and predicting solar flares [151] . RCFs capture a flexible measure of similarity between every sub-image across every pair of images without using contextual or task-specific information. The regression step in MOSAIKS then treats these features x as an overcomplete basis for predicting any y, which may be a nonlinear function of image elements.

Notation. In our notation, the input variable z is a set of satellite images $\mathbf{I}$, each corresponding to a physical location, ℓ. We use brackets to denote indexing into images, with colons denoting sub-regions of images (e.g. $\mathbf{I}_\ell[i, j]$ is the $(i, j)^{th}$ pixel of image $\mathbf{I}_\ell$,

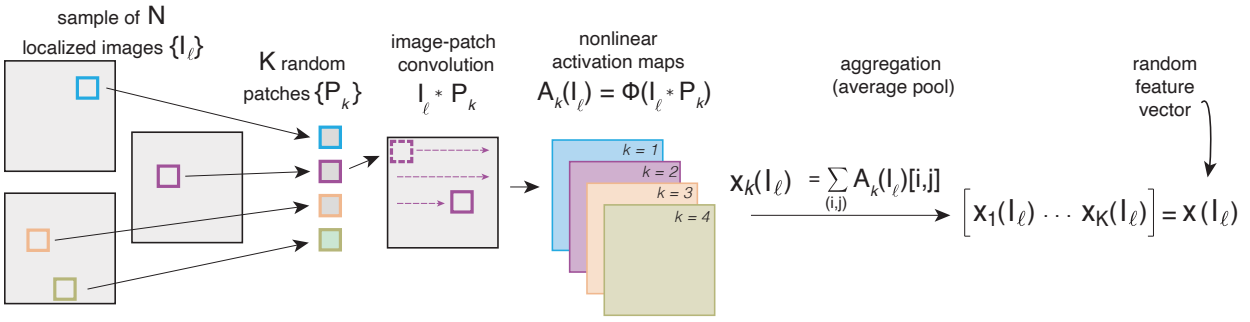


Figure 8.2: MOSAIKS featurization process (random convolutional features). Given a large sample of N satellite images, a random sample of K patches are drawn. These K random patches $\mathbf{P}_k$ are convolved over each image $\mathbf{I}_\ell$ and passed through a nonlinear function $\Phi(\cdot) = \text{ReLU}(\cdot)$ to generate K activation maps. Pixel-specific activations are pooled across each image to generate K-dimensional features that can be stored and distributed to users.

$\mathbf{I}_\ell[i : i + M, j : j + M]$ is the square sub-image of size $M \times M$ starting at pixel (i, j).) Because images have a third dimension (spectral bands), a colon $\mathbf{I}_\ell[i, j, :]$ denotes all bands at pixel (i, j). Indexing into non-image objects is denoted with subscripts (e.g. the k^{th} element of vector $\mathbf{x}$ is denoted as $\mathbf{x}_k$ and the k^{th} patch in a set of patches $\mathbf{P}$ is denoted as $\mathbf{P}_k$). We denote inner products with angular brackets $\langle \cdot, \cdot \rangle$ and the convolution operator with $*$.

Figure 8.2 depicts our featurization process. We begin with a set of images $\{\mathbf{I}_\ell\}_{\ell=1}^N$, each of which is centered at locations indexed by $\ell = \{1, \ldots, N\}$. MOSAIKS generates task-agnostic feature vectors $\mathbf{x}(\mathbf{I}_\ell)$ for each satellite image $\mathbf{I}_\ell$ by convolving K "patches", $\mathbf{P}_k \in \mathbb{R}^{M \times M \times S}$, across the entire image. Each patch is a randomly sampled sub-image from the set of training images $\{\mathbf{I}_\ell\}_{\ell=1}^N$.[2] M is the width and height of each patch in units of pixels and S is number of spectral bands; in our analysis, we use patches of width and height $M = 3$ and $S = 3$ bands (red, green, and blue). The dimension of the resulting feature space is equal to K, the number of patches used; in our main analyses we employ $K = 8,192$.[3]

Both images and patches are whitened according to zero components analysis (ZCA), a common pre-processing routine in image processing [169].[4] We then convolve each patch $\mathbf{P}_k$ over each of the N images and apply a pixel-wise nonlinearity operator Φ to each resulting matrix to obtain K *nonlinear* activation maps $\mathbf{A}_k(\mathbf{I}_\ell) = \Phi(\mathbf{P}_k * \mathbf{I}_\ell + \mathbf{b}_k)$ for each image $\mathbf{I}_\ell$.

[2]This patch selection process is almost identical to the filter selection methods described in [5, 71, 237].

[3]To improve efficiency of the featurization process, our implementation calculates the inner product of patch and image only for $K/2 = 4096$ unique patches. We then create an additional $K/2$ values equal to the negative of each of the original inner products.

[4]ZCA whitening pre-multiplies each patch by a transformation such that the resulting empirical covariance matrix of the whitened patches is the identity matrix. In practice, we apply the whitening operator as a right multiplication to the original 8192×27 whitened patch matrix in order to reduce computation.

The $(i, j)^{th}$ pixel of the k^{th} activation map is defined as

$$\mathbf{A}_k(\mathbf{I}_\ell)[i, j] = \Phi(\langle \mathbf{I}_\ell[i : i + M, j : j + M, :], \mathbf{P}_k \rangle + b_k), \tag{8.1}$$

where b_k is a bias term from the constant bias matrix $\mathbf{b}_k$, in which every element is equal to $b_k = 1$. We use $\Phi(\mathbf{I}_\ell; \mathbf{P}_k, \mathbf{b}_k) = \text{ReLU}(\mathbf{P}_k * \mathbf{I}_\ell + \mathbf{b}_k) := \max\{\mathbf{P}_k * \mathbf{I}_\ell + \mathbf{b}_k, 0\}$ as the nonlinear operator. We then aggregate across the image by taking the average of the nonlinear activation maps. The combination of the nonlinear operator $\Phi(\cdot)$ and average pooling results in a scalar value for each patch k and image ℓ pair. With images of shape 256×256 pixels and patches of width and height $M = 3$, the features are expressed as:

$$\mathbf{x}_k(\mathbf{I}_\ell) = \frac{1}{254^2} \sum_{i=1}^{254} \sum_{j=1}^{254} \mathbf{A}_k(\mathbf{I}_\ell)[i, j] \ . \tag{8.2}$$

Stacking these scalars across all K patches provides the resulting K-dimensional feature vector $\mathbf{x}(\mathbf{I}_\ell) := \begin{bmatrix} \mathbf{x}_1(\mathbf{I}_\ell) & \mathbf{x}_2(\mathbf{I}_\ell) & ... & \mathbf{x}_K(\mathbf{I}_\ell) \end{bmatrix} \in \mathbb{R}^K$. This featurization thus embeds the original image $\mathbf{I}_\ell$ into a K-dimensional feature space, which can then be mapped to many different outcomes as illustrated in Figure 8.1. The linear link between labels and features in Figure 8.1 may express a relationship between labels and image pixels that is highly nonlinear because the features themselves are nonlinear with respect to the images.

Connections to the Random Kitchen Sinks Framework

Random kitchen sinks approximate arbitrary functions by creating a finite series of features generated by passing the input variables z through a set of K nonlinear functions $g(z; \Theta_k)$, each paramaterized by draws of a random vector Θ. The realized vectors Θ_k are drawn independently from a pre-specified distributions for each of $k = 1...K$ features. Given an expressive enough function g and infinite K, such a featurization would be a universal function approximator [230]. In our case, such a function g would encode interactions between all subsets of pixels in an image. For an image of size $256 \times 256 \times 3$, there are $2^{256 \times 256 \times 3}$ such subsets. Therefore, the fully-expressive approach is inefficient in generating predictive skill with reasonably concise K because each feature encodes more pixel interactions than are empirically useful.

To adapt random kitchen sinks for satellite imagery, we use random convolutional features (described above), making the simplifying assumption that most information contained within satellite imagery is represented in *local image structure*. Connecting our implementation and notation to the framework of random kitchen sinks, the random variables Θ_k are instantiated as the values of a random patch $\mathbf{P}_k$ and the bias b_k. The input variable z is an image $\mathbf{I}_\ell$, and $g(z; \Theta_k)$ represents the convolution of the patch over the image, followed by addition of the bias b_k and application of a element-wise ReLU function and an average pool. Applied to satellite images, random convolutional features reduce the number of effective parameters in the function by considering only local spatial relationships between pixels. This results in a highly expressive, yet computationally tractable model for prediction.

Relevant Structures of Satellite Imagery and SIML Tasks

Three particular properties provide the the motivation for our choice of a convolution and average-pool mapping to define g.

First, we hypothesize that convolutions of small patches will be sufficient to capture nearly all of the relevant spatial information encoded in images because objects of interest (e.g. a car or a tree) tend to be contained in a small sub-region of the image. This is particularly true in satellite imagery, which has a much lower spatial resolution than most natural imagery.

Second, we expect a single layer of convolutions to perform well because satellite images are taken from a constant perspective (from above the region of interest) at a constant distance and are often orthorectified to remove the effects of image perspective. Together, these characteristics mean that a given object will tend to appear the same size when captured in different images. This allows for MOSAIKS's relatively simple, translation invariant featurization scheme to achieve high performance, and avoids the need for more complex architectures designed to provide robustness to variation in object size and orientation.

Third, we average-pool the convolution outputs because most labels in the types of problems we study can be approximately decomposed into a sum of sub-image characteristics. For example, forest cover is measured by the percent of total image area covered in forest, which can equivalently be measured by averaging the percent forest cover across sub-regions of the image. Labels that are strictly averages, totals, or counts of sub-image values (such as forest cover, road length, population density, elevation, and night lights) will all exhibit this decomposition. While this is not strictly true of all SIML tasks, for example income and average housing price, we demonstrate that MOSAIKS still recovers strong predictive skill on these tasks. This suggests that some components of the observed variance in these labels may still be decomposable in this way, possibly because they are well-approximated by functions of sums of observable objects.

8.3 Assessing Generalization Across Tasks

We design a set of experiments to test whether and under what settings MOSAIKS can provide access to high-performing, computationally-efficient, global-scale SIML predictions. In the next several sections, we demonstrate generalization across tasks, and compare MOSAIKS's performance and cost to existing state-of-the-art SIML models (Section 8.3), assess its performance when data are limited and when predicting far from observed labels (Section 8.4), scale the analysis to make global predictions and try recreating the results of a national survey (Section 8.5), and detail an algorithmic extension that enables us to make predictions at finer resolution than the provided labels (Section 8.6).

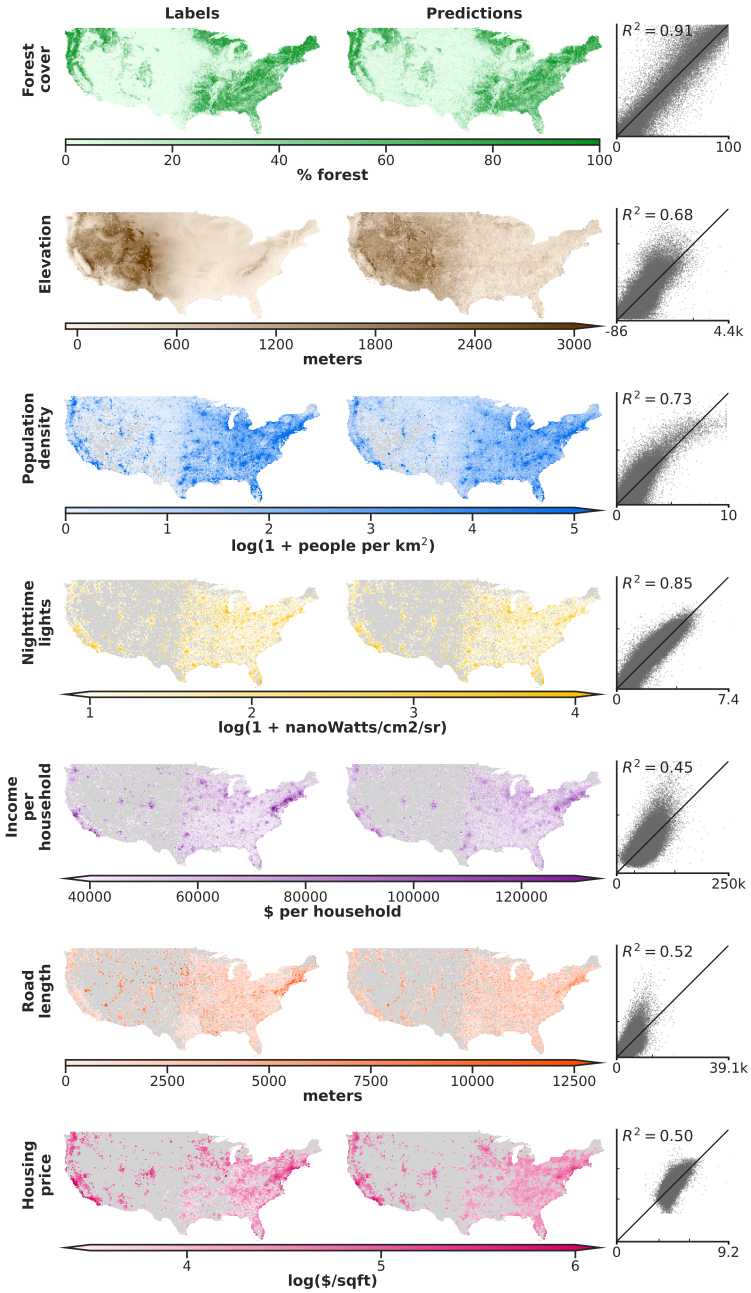


Figure 8.3: 1km × 1km resolution prediction of many tasks across the continental US. Left maps: observations. Right maps: predictions (concatenated validation set estimates from 5-fold cross-validation for the same 80,000 grid cells as the observations). Scatters: spatially disagreggated ground-truth observations (vertical axis) vs. predictions (horizontal axis); each point is a ∼1km×1km grid cell. Values in maps are aggregated up to 20km×20km cells for display. Grey areas were not sampled in the experiment.

Multi-Task Performance of MOSAIKS in the US

We first test whether MOSAIKS achieves a practical level of generalization by applying it to a diverse set of pre-selected tasks in the continental United States (US). While many applications of interest for SIML are in remote and/or data-limited environments where ground-truth may be unavailable or inaccurate, systematic evaluation and validation of SIML methods are most reliable in well-observed and data-rich environments [37].

We sample daytime images using the Google Static Maps API [111] from across the continental US, each covering ∼1km×1km (640 × 640, downsampled to 256 × 256 pixels) and aligned to an equal-area grid. For our primary experiment we subsample sets of 100,000 observations, roughly 1.25% of the grid cells in the continental US, using two distinct sampling strategies.[5] First, we sample uniformly-at-random (UAR) from all grid cells within the continental US. This sampling strategy is most appropriate for tasks like forest cover, where there is meaningful variation in most regions of the country. Second, we implement a population-weighted (POP) sampling strategy, where each grid cell is sampled with probability weighted by population density estimates for the year 2015 from [57]. This weighted sampling strategy is most applicable to tasks like housing price, where the most meaningful variation lies in more populated regions of the US. We use the UAR sample for the forest cover, elevation, and population density tasks. We use the POP sample for nighttime lights, income, road length, and housing price. We model labels whose distribution is approximately log-normal (population density, nighttime lights, and housing price) using a log transformation.

We first implement the featurization step, passing these images through the MOSAIKS feature extraction algorithm (Section 8.2) to produce $K = 8,192$ features per image. Using the resulting matrix of features ($\mathbf{X}$), we then repeatedly implement the regression step by solving a 5-fold cross-validated ridge regression for each task and predict forest cover ($R^2 = 0.91$), elevation ($R^2 = 0.68$), population density ($R^2 = 0.72$), nighttime lights ($R^2 = 0.85$), average income ($R^2 = 0.45$), total road length ($R^2 = 0.53$), and average house price ($R^2 = 0.52$)[6] in a holdout test sample made up of 20% of the sampled data (Figure 8.3).[7] Computing the feature matrix $\mathbf{X}$ from imagery for a 100,000 image sample took less than 2 hours on a cloud computing node (Amazon EC2 p3.2xlarge instance, Tesla V100 GPU). Subsequently, solving a cross-validated ridge regression for each task took 6.8 minutes to compute on a local workstation with ten cores (Intel Xeon CPU E5-2630).

These results indicate that MOSAIKS is skillful for a wide range of possible applications without changing the procedure or features and without task-specific expertise. Note that due to the absence of metadata describing the exact time of observation in the Google imagery, as well as task-specific data availability constraints, these performance measures

[5]We discard marine grid cells, but do not discard grid cells that are composed only of lakes or smaller inland bodies of water.

[6]Performance observed for housing using our published data will be higher ($R^2 = 0.60$) because privacy concerns mandate the withholding of a subset of this data (see Section 8.A).

[7]Test set and validation set performance are essentially identical (Table 8.1); validation set values are used in Figure 8.3 for display purposes as there are more observations.

Task	MOSAIKS R^2	Fine-tuned ResNet-18 R^2	Pre-trained ResNet-152 R^2
Forest cover	0.91	0.94	0.66
Elevation	0.68	0.80	0.32
Population density	0.72	0.80	0.29
Nighttime lights	0.85	0.89	0.48
Income	0.45	0.47	0.07
Road length	0.53	0.58	0.16
Housing price	0.52	0.50	0.01

Table 8.1: Task-specific MOSAIKS test-set performance in contrast to an 18-layer variant of the ResNet Architecture (ResNet-18) trained end-to-end for each task and an unsupervised featurization using the last hidden layer of a 152-layer ResNet variant pre-trained on natural imagery and applied using ridge regression.

are conditional on a certain degree of unknown temporal mismatch between imagery and task labels (Section 8.A).

Benchmarking Performance with Alternative SIML Approaches

We contextualize this performance by comparing MOSAIKS to existing deep-learning based SIML approaches (Table 8.1). First, we retrain end-to-end a commonly-used deep convolutional neural network (CNN) architecture [115, 129, 175] (ResNet-18) using identical imagery and labels for the seven tasks above. This training took 7.9 hours per task on a cloud computing node (Amazon EC2 p3.xlarge instance, Tesla V100 GPU). We find that MOSAIKS exhibits predictive accuracy competitive with the CNN for all seven tasks (mean $R^2_{\text{CNN}} - R^2_{\text{MOSAIKS}} = 0.04$; smallest $R^2_{\text{CNN}} - R^2_{\text{MOSAIKS}} = -0.02$ for housing; largest $R^2_{\text{CNN}} - R^2_{\text{MOSAIKS}} = 0.12$ for elevation).

Second, we apply transfer learning [217] using a ResNet-152 CNN pre-trained on natural images to featurize the same satellite images [115, 175]. We then apply ridge regression to the CNN-derived features. The speed of this approach is similar to MOSAIKS, but its performance is dramatically lower on all seven tasks (Table 8.1).

Third, we compare MOSAIKS to an approach from prior studies [130, 145, 296] where a deep CNN (VGG16 [262] pretrained on the ImageNet dataset) is trained end-to-end on nighttime lights and then each task is solved via transfer learning. We apply MOSAIKS to the imagery from Rwanda, Haiti, and Nepal used in [130] to solve all eleven development-oriented tasks they analyze. We find MOSAIKS matches prior performance across tasks in Rwanda and Haiti, and has slightly lower performance (average $\Delta R^2 = 0.08$) on tasks in Nepal (Figure 8.4). The regression step of this transfer learning approach and MOSAIKS are similarly fast, but the transfer learning approach requires country-specific retraining of the CNN, limiting its accessibility and reducing its generalizability.

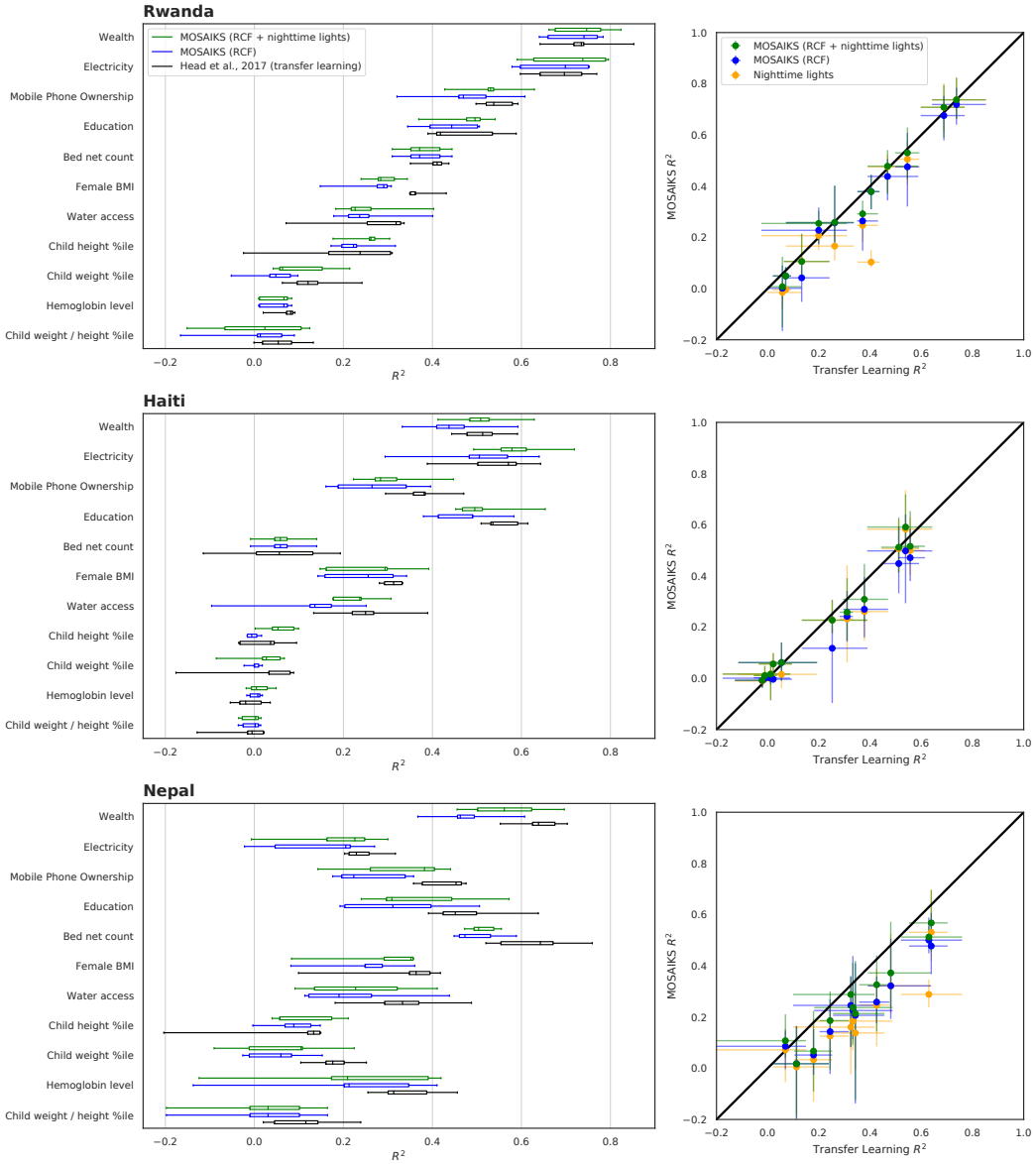


Figure 8.4: Comparison of accuracy between MOSAIKS and a transfer learning model. Box plots (left) show task-specific performance of MOSAIKS models (random convolutional features, RCF, in blue and RCF + nighttime lights appended as an additional column in the feature matrix in green) in contrast to a transfer learning model (black). Box and whiskers show the performance over the 5 cross-validation folds. Scatter plots (right) show the performance of MOSAIKS models and a nighttime lights-only model (orange) versus the transfer learning model performance. Each point in the scatter is the average R^2 over the 5 cross-validation folds, while whiskers indicate the full range of performance across folds.

Together, these three experiments illustrate that with a single set of task-independent features, MOSAIKS predicts outcomes across a diverse set of tasks, with performance and speed that favorably compare to existing SIML approaches. However, throughout this set of experiments, we find that some sources of variation in labels are not recovered by MOSAIKS. For example, extremely high elevations (>3,000m) are not reliably distinguished from high elevations (2,400-3,000m) that appear visually similar (Figure 8.3). Additionally, roughly half the variation in incomes and housing prices is unresolved (Figure 8.3), presumably because they depend on factors not observable from orbit, such as tax policies or school districts.

Comparing Costs

In practice, high computational costs can limit the use of SIML methods – especially when resources are scarce, such as in government agencies of low-income countries [118] or research teams and NGOs with limited budgets. Specifically designed to address this challenge, MOSAIKS scales across many research tasks by decoupling featurization from task selection, model-fitting, and prediction. The computationally costly step of featurization is done centrally on a fast computer with a graphics processing unit (GPU). Implementation of this one-time unsupervised featurization results in a roughly 6 to 1 compression of stored and transmitted imagery data with $K = 8,912$ features.[8] Because features are created and stored by a central entity, the research community could make use of a cached set of computations, potentially reducing the overall computational burden of widespread SIML and any external social costs generated by these computations [268]. Additionally, decoupling task-agnostic computations from task-specific computations enables practitioners to run more diagnostic analyses on their tasks, such as those described in Section 8.4.

From the perspective of a user who can access pre-computed MOSAIKS features to train and validate a new task, we find that MOSAIKS is $\sim 250\times$ to $10,000\times$ faster than a standard deep neural net architecture (ResNet-18), depending on the computational resources available to a MOSAIKS user (Table 8.2). Moreover, MOSAIKS performance is competitive with the ResNet on all tasks we have studied (Table 8.1). From the perspective of the entire computational ecosystem, which bears the cost of image featurization in addition to model training and testing, we find that MOSAIKS is $5.3\times$ faster than the ResNet-18 when solving a single task. The relative efficiency of MOSAIKS grows with the number of tasks studied because MOSAIKS features can be reused across tasks.

For the ResNet-18 model, the times in Table 8.2 reflect our wall-clock time on a single Amazon EC2 instance for a single task, so that the time costs are similar to that of introducing a single new domain *ex post*. For MOSAIKS, we report wall-clock times on three different computational platforms, as users may have access to different resources. We show times using the same GPU instance as we use for the ResNet comparisons, times on a local

[8]This is calculated as: $(256 * 256 * 3)/(8192 * 4) = 6\times$ compression, where 256 * 256 * 3 integer values per image are compressed into 8192 float32 features, each of which takes $4\times$ the storage of an integer. Using 100 features gives a $500\times$ compression.

	ResNet	**MOSAIKS**
Component	Time (GPU)	Time
Training set featurization ($N = 80$k)	**$\sim$ 7.9 hours**	$\sim$ 1.2 hours (GPU)
Model training		$\sim$ 2.8 seconds (GPU)
		$\sim$ 50 seconds (10 cores)
		$\sim$ 1.8 minutes (laptop)
Holdout set featurization ($N = 20$k)	**$\sim$ 40 seconds**	$\sim$18 minutes (GPU)
Holdout set prediction		$<$ 0.01 seconds (GPU)
		$\sim$ 0.1 seconds (10 cores)
		$\sim$ 0.7 seconds (laptop)
Total cost to ecosystem	**$\sim$ 7.9 hours**	$\sim$ 1.5 hours (GPU)
Total cost to user	**$\sim$ 7.9 hours**	$\sim$ 2.8 seconds (GPU)
		$\sim$ 50.1 seconds (10 cores)
		$\sim$ 1.8 minutes (laptop)

Table 8.2: Wall-clock times of components of MOSAIKS compared with a fine-tuned CNN. Bold times are those that a practitioner using each method would incur (assuming MOSAIKS users have access to a standard laptop only). Model training time is reported for a single hyperparameter configuration and a single task for both ResNet and MOSAIKS (using K=8,192 features). ResNet operations were run on an Amazon EC2 p3.2xlarge instance with a Tesla V100 GPU and 60GB of onboard RAM, which cost roughly \$3/hr at the time we ran our experiments. MOSAIKS operations are shown for this same instance, a local workstation with ten cores (Intel Xeon CPU E5-263), and a laptop (2018 MacBook Pro).

workstation with ten cores, and times on a laptop. For both ResNet and MOSAIKS, we report model training time for a single set of hyperparameters. The ecosystem-wide costs of featurization per task shown in Table 8.2 would decline as MOSAIKS becomes more widely adopted, because features can be cached centrally and distributed without modification to multiple users who are training and/or testing SIML in common locations.

For cost comparisons, we considered only one CNN architecture, which we chose because of its use in previous remote sensing applications [145]. We did not attempt to innovate in deep neural net architectural design or algorithms to decrease computational constraints. Our results suggest that such innovations may be both possible and worthwhile, as increasing training efficiency can in turn increase accessibility and enable context-sensitive analysis.

8.4 Evaluating Model Sensitivity

There is growing recognition that understanding the accuracy, precision, and limits of SIML predictions is important, since consequential decisions increasingly depend on these outputs, such as which households should receive financial assistance [12, 37]. However, historically,

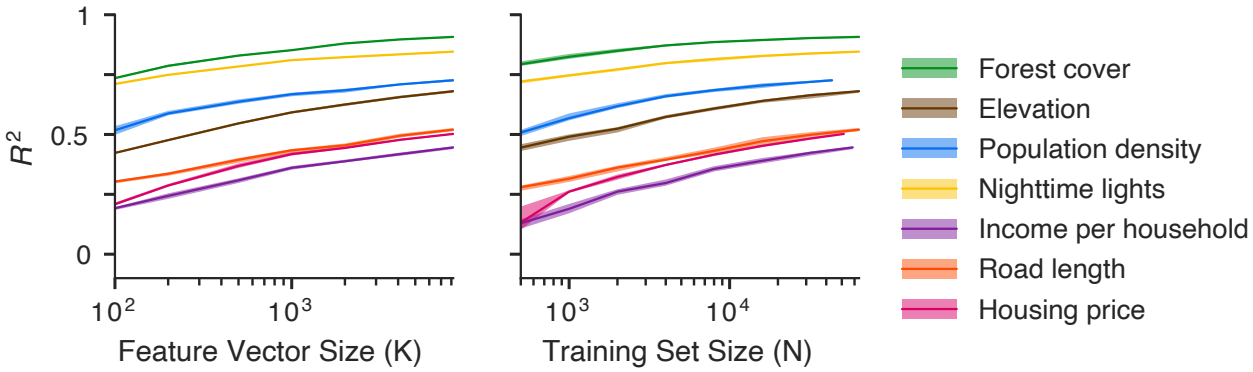


Figure 8.5: Prediction accuracy using smaller K and N. Validation set R^2 performance for all seven tasks while varying the number of random convolutional features K and holding $N = 64,000$ (left) and while varying N and holding $K = 8,192$ (right). Lines indicate average accuracy across folds; shaded bands indicate accuracy range across folds.

the high costs of training deep-learning models have generally prevented the stress-testing and bench-marking that would ensure accuracy and constrain uncertainty. To characterize the performance of MOSAIKS, we test its sensitivity to the number of features (K) and training observations (N), as well as the extent of spatial extrapolation.

Changes to Training Data

Due to the sampling of random patches in the MOSAIKS featurization step, the computational complexity of the regression step can be manipulated by simply including more or fewer features. Repeatedly re-solving the linear regression step in MOSAIKS with a varied number of features indicates that increasing K above 1,000 features provides minor predictive gains (Figure 8.5). A majority of the observable signal in the baseline experiment using $K = 8,192$ is recovered using $K = 200$ (min 55% for income, max 89% for nighttime lights), reducing each 65,536-pixel tri-band image to just 200 features ($\sim 250\times$ data compression). Similarly, re-solving MOSAIKS predictions with a different number of training observations demonstrates that models trained with fewer samples may still exhibit high accuracy. A majority of the available signal is recovered for many outcomes using only $N = 500$ (55% for road length to 87% for forest cover), with the exception of income (28%) and housing price (26%) tasks, which require larger samples. Together, these experiments suggest that users with computational, data acquisition, or data storage constraints can easily tailor MOSAIKS to match available resources and can estimate the performance impact of these alterations.

To systematically evaluate the ability of MOSAIKS to make accurate predictions in large contiguous areas where labels are not available, we conduct a spatial cross-validation experiment by partitioning the US into a checkerboard pattern (Figure 8.6), training on the black squares and testing on the white squares. Increasing the width of squares (δ) in the

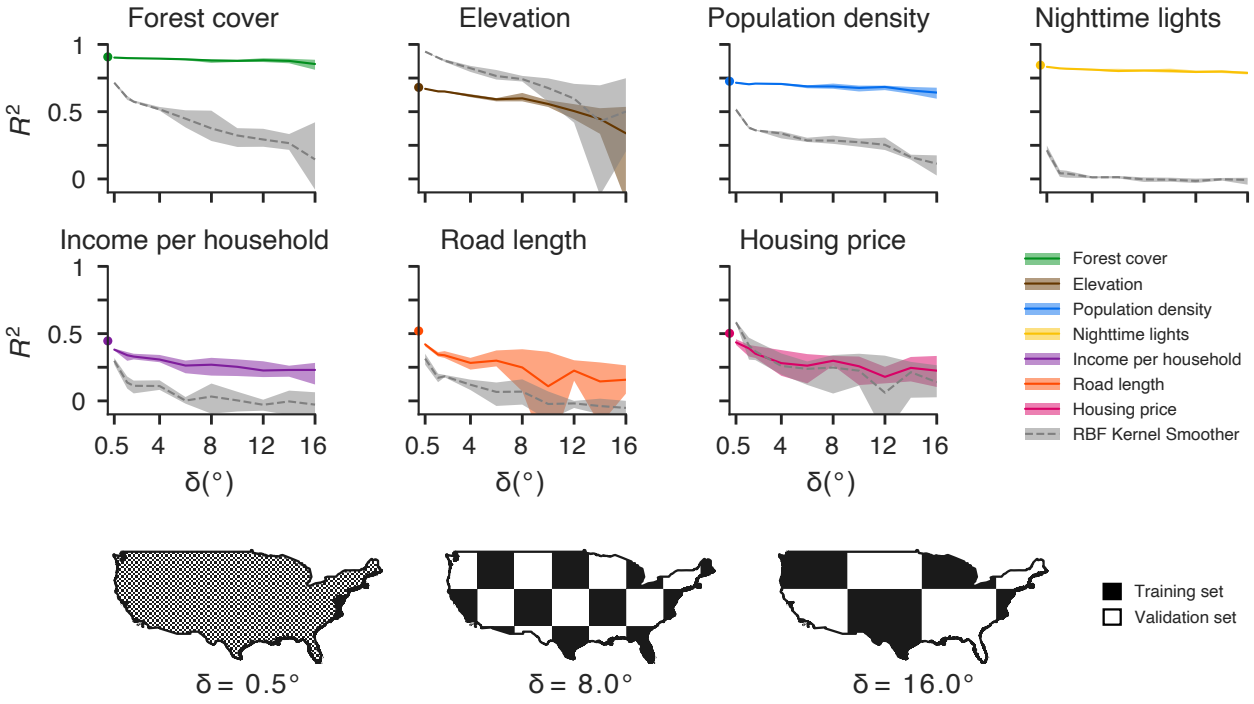


Figure 8.6: Evaluation of performance over regions of increasing extents that that are excluded from training sample. Data are split using a checkerboard partition, where the width and height of each square is δ (measured in degrees). Example partitions with $\delta = 0.5°$, 8°, 16° are shown in maps at the bottom of the figure. For a given δ, training occurs using data sampled from black squares and performance is evaluated in white squares. Plots show colored lines representing average performance of MOSAIKS in the US across δ values for each task. Benchmark performance numbers from Figure 8.3 are indicated as circles at $\delta = 0$. Grey dashed lines indicate corresponding performance using only spatial interpolation with an optimized radial basis function (RBF) kernel instead of MOSAIKS. For each δ, colored or grey bands denote ranges of performance across four instantiations of the checkerboard using different center vertices to define checkerboard cells.

checkerboard increases the average distances between train and test observations, simulating increasingly large spatial extrapolations. We find that for three of seven tasks (forest cover, population density, and nighttime lights), performance declines minimally regardless of distance (maximum R^2 decline of 10% at $\delta = 16°$ for population density). For income, road length, and housing price, performance falls moderately at small degrees of spatial extrapolation (19%, 33%, and 35% decline at $\delta = 4°$, respectively), but largely stabilizes thereafter. Finally, elevation exhibits steady decline with increasing distances between training and testing data (49% decline at $\delta = 16°$).

To contextualize this performance, we compare MOSAIKS to kernel-based spatial in-

terpolation using a Gaussian radial basis function (RBF) kernel, a simple and widely used approach to fill in regions of missing data (see Chapter 5). In this approach, the value for a point in the validation set at location $\ell_v \in \mathbb{R}^2$ is predicted to be a weighted sum of the values of all the points in the training set ℓ_t, as follows:

$$\hat{y}_v^s = \frac{\sum_{\ell_t \in [\text{Train}]} y_t^s w(\ell_t, \ell_v)}{\sum_{\ell_t \in [\text{Train}]} w(\ell_t, \ell_v)}; \qquad w(\ell_t, \ell_v) = e^{-\frac{1}{2\sigma^2}\|\ell_t - \ell_v\|^2} .$$

Here, w is the weight assigned to each observation in the training set based on distance, such that w decreases as the distance between the point being predicted and the point in the training set increases. We select σ, the parameter that determines the rate at which w degrades with distance, to maximize average performance on the validation set across four spatially-offset runs of the checkboard partitions shown in Figure 8.6, and tune the regularization parameter in the MOSAIKS ridge regression in the same fashion.

Using the same samples, MOSAIKS substantially outperforms spatial interpolation (grey dashed lines in Figure 8.6) across all tasks except for elevation, where interpolation performs almost perfectly over small ranges ($\delta = 0.5^\circ : R^2 = 0.95$), and housing price, where interpolation slightly outperforms MOSAIKS at small ranges. For both, interpolation performance converges to that of MOSAIKS over larger distances. Thus, in addition to generalizing across tasks, MOSAIKS generalizes out-of-sample across space, outperforming spatial interpolation of ground-truth in 5 of 7 tasks. This suggests that MOSAIKS, and SIML generally, exploits the spectral and structural content of information within an image to generate predictions at national scale that extend beyond what can be captured by geographic location alone.

The above sensitivity tests are enabled by the speed and simplicity of training MOSAIKS. These computational gains also enable quantification of uncertainty in model performance within each diagnostic test. As demonstrated by the shaded bands in Figures 8.5 and 8.6, uncertainty in MOSAIKS performance due to variation in splits of training-validation data remains modest under most conditions.

8.5 Testing Performance at Scale

Having evaluated MOSAIKS systematically in the data-rich US, we test its performance at planetary scale and its ability to recreate results from a national survey.

Global Observation

We test the ability of MOSAIKS to scale globally using the four tasks for which global labels are readily available. Using a random sub-sample of global land locations (training and validation: N = 338,781, test: N = 84,692), we construct planet-scale, multi-task

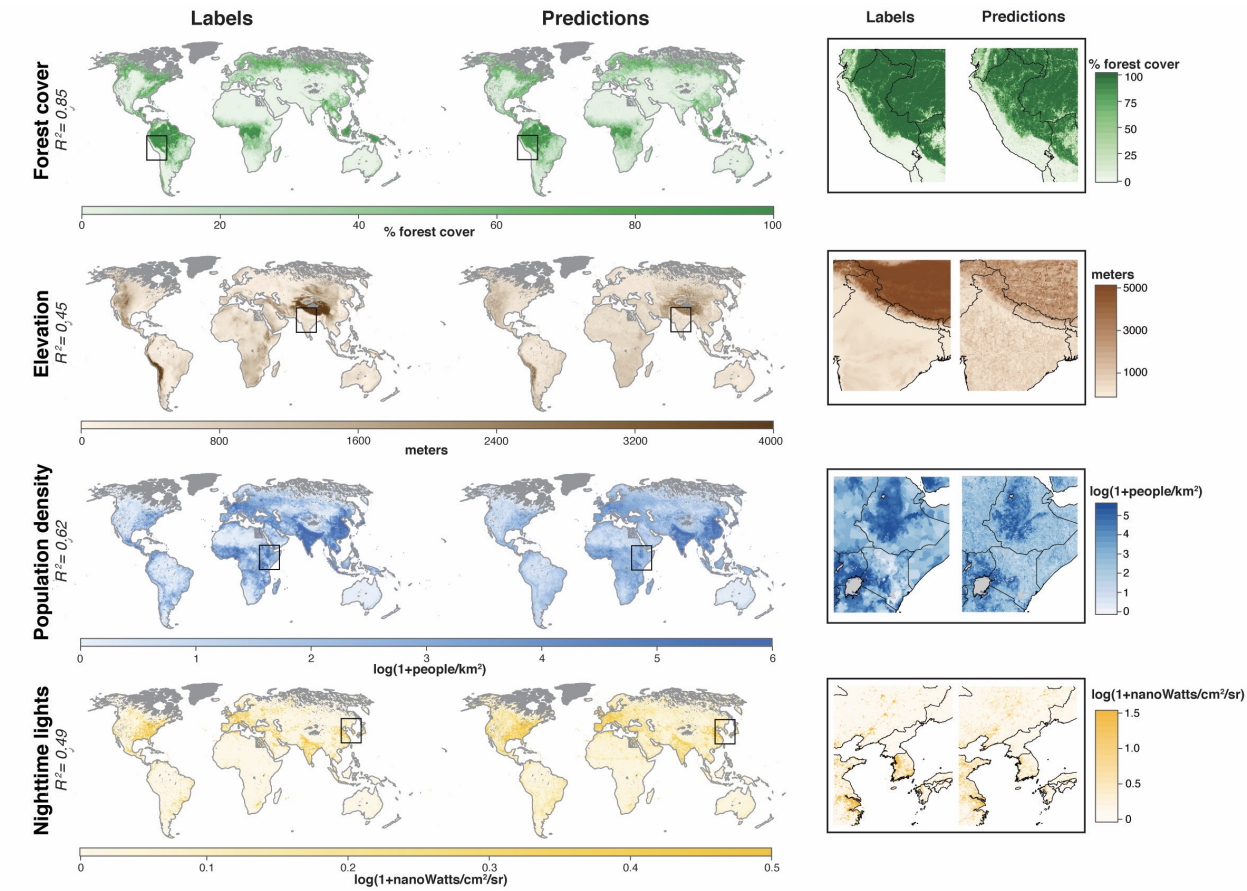


Figure 8.7: Global training data (left maps) and estimates using a single featurization of daytime imagery (right maps). Insets (far right) marked by black squares in global maps. Out-of-sample predictions are constructed using 5-fold cross-validation. For display purposes, maps depict ∼50km × 50km average values (ground truth and predictions at ∼1km × 1km).

estimates using a single set of label-independent features ($K = 2048$, Figure 8.7)[9], predicting the distribution of forest cover ($R^2 = 0.85$), elevation ($R^2 = 0.45$), population density ($R^2 = 0.62$), and nighttime lights ($R^2 = 0.49$).

For our global analysis, we draw samples from a global grid, composed of roughly 420 million cells just over 1km^2 in size, using an identical structure to that we use for the US. To obtain observations for our global analysis, we sub-sample 1,000,000 cells from this grid, sampling UAR from non-marine grid cells.

One of the difficulties in sub-sampling from the global grid is that there are many grid cells where we do not have images available (there are negligibly few missing images in the US

[9]When generating features ($K = 2,048$) for our global model, we use patches drawn randomly from the *global* sample of images, not just from within the US.

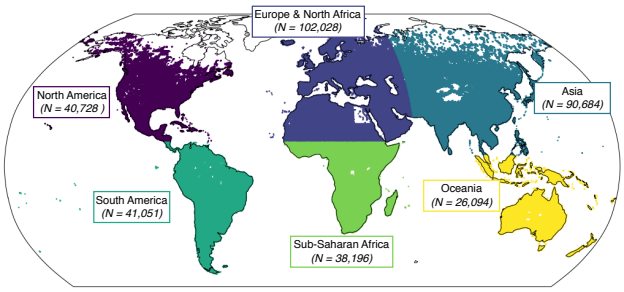


Figure 8.8: MOSAIKS predictions at global scale (Figure 8.7) are generated from six separate cross-validated ridge regressions using random convolutional features. Each continent model is trained on 80% of the sample size shown (N).

grid). After discarding grid cells with missing imagery from our original sample of 1,000,000 observations, we are left with $N = 498,063$ valid observations. After removing observations for which labeled data are missing for any of the tasks we analyze at global scale (forest cover, elevation, population density, and nighttime lights), we are left with $N = 423,476$ observations, which we use to train and evaluate the model.

When training the global model, we follow the approach outlined in Section 8.2, solving for grid cell labels as a linear function of the random convolutional features using ridge regression and cross-validation to tune the regularization parameter λ. However, recovered regression weights are likely to differ across regions of the globe due to heterogeneity in image quality and in visual signal of task labels or their derivatives. We divide our global sample into six continental regions before solving each task. The continents (and sample sizes used for training and testing each continent-specific model) are shown in Figure 8.8.

Modeling heterogeneity using the continents shown in this way leads to meaningful gains in performance over ignoring continent effects, resulting in R^2 values of 0.85, 0.45, 0.62, and 0.49, for forest cover, elevation, population density, and nighttime lights, respectively (Figure 8.7). In contrast, a global model that pools all observations across the globe and solves for a single linear function of random convolutional features generates R^2 values of 0.80, 0.26, 0.48, and 0.41, for the same tasks.

Scaling Across Tasks

It has been widely suggested that SIML could be used by resource-constrained governments to reduce the cost of surveying their citizens [86, 130, 145, 203, 238]. To demonstrate MOSAIKS's performance in this theoretical use-case, we simulate a field test with the goal of recreating results from an existing nationally representative survey. Using the pre-computed features from the first US experiment above, we generate predictions for 12 pre-selected questions in the 2015 American Community Survey (ACS) conducted by the US Census Bureau [278]. We obtain R^2 values ranging from 0.06 (% household income spent on rent)

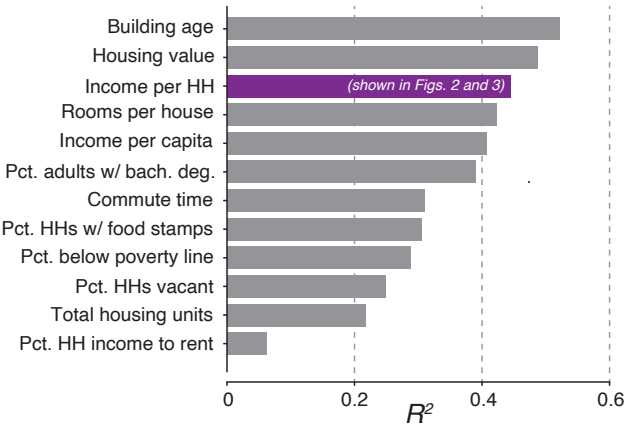


Figure 8.9: Test-set performance in the US shown for 12 variables from the 2015 American Community Survey (ACS) conducted by the US Census Bureau [278]. Income per household (HH) (in purple) is also shown in Figure 8.3 and Table 8.1.

to 0.52 (building age), with an average R^2 of 0.34 across 12 tasks (Figure 8.9).

Compared to a baseline of no ground survey, or a costly survey extension, these results suggest that MOSAIKS predictions could provide useful information to a decision-maker for almost all tasks at low cost; noting that, in contrast, the ACS costs > \$200 million to deploy annually [279]. However, some variables (e.g. percent household income spent on rent) may continue to be retrievable only via ground survey.

8.6 Extension: Label Super-Resolution

Many use cases would benefit from SIML predictions at finer resolution than is available in training data [188, 274]. Here we show that MOSAIKS can estimate the relative contribution of sub-regions within an image to overall image-level labels, even though only aggregated image-level labels are used in training (Figure 8.10). Such "label super-resolution" prediction follows from the functional form of the featurization and linear regression steps in MOSAIKS, allowing it to be analytically derived for labels that represent nearly linear combinations of ground-level conditions.

The featurization method in MOSAIKS exploits the fact that many image-level outcomes of interest are linearly decomposable across sub-image regions. This is done by creating image-level features that are averages of statistics from all sub-image regions. Because these features are ultimately used in linear regression, a natural property of this approach is that weights estimated in this linear regression can be used not only to generate predictions of outcome variables at the image-scale, but also at the scale of any sub-image region, as illustrated in Figure 8.11.

Given an image-label pair $\{\mathbf{I}_\ell, y_\ell^s\}$, the goal of label super-resolution is to resolve which

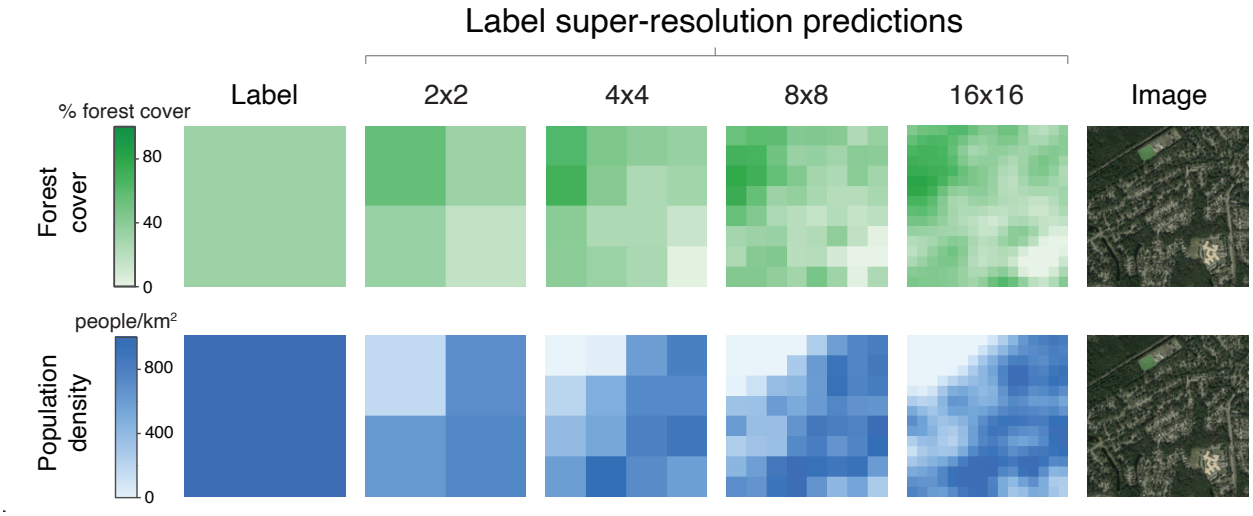


Figure 8.10: Label super-resolution predictions for an example image. Shown here are the image-level labels (column 1), predicted outcomes from MOSAIKS at increasing levels of label super-resolution (columns 2-5), and the image from Google Static Maps (column 6).

sub-regions of the image $\mathbf{I}_\ell$ contribute to high or low values of y_ℓ^s. Recall that for image $\mathbf{I}_\ell$, feature vector $\mathbf{x}(\mathbf{I}_\ell)$ is a K dimensional vector, where each scalar element $\mathbf{x}_k(\mathbf{I}_\ell)$ of $\mathbf{x}(\mathbf{I}_\ell)$ is an average across the pixels of the image of the values obtained by convolving sub-regions of the image with patch $\mathbf{P}_k$. As in Section 8.2, denote by $\mathbf{X}$ the full random feature matrix in $\mathbb{R}^{N\times K}$, so that $\mathbf{X}_{\ell k}$ denotes the k^{th} element of the feature vector describing image $\mathbf{I}_\ell$. When we perform a linear regression for task s, the resulting regression weights are a vector $\hat{\beta}^s \in \mathbb{R}^K$ such that the scalar $\hat{\beta}_k^s$ describes the relative weight of feature k in the image-scale predictions. The prediction of outcome s using image $\mathbf{I}_\ell$ thus decomposes as:

$$\hat{y}_\ell^s = \mathbf{X}_\ell \hat{\beta}^s \tag{8.3}$$

$$= \sum_{k=1}^{K} \mathbf{X}_{\ell k} \cdot \hat{\beta}_k^s \tag{8.4}$$

$$= \sum_{k=1}^{K} \left(\frac{1}{254^2} \sum_{i=1}^{254} \sum_{j=1}^{254} \mathbf{A}_k(\mathbf{I}_\ell)[i,j] \right) \cdot \hat{\beta}_k^s \tag{8.5}$$

$$= \frac{1}{254^2} \sum_{i=1}^{254} \sum_{j=1}^{254} \underbrace{\left(\sum_{k=1}^{K} \hat{\beta}_k^s \cdot (\mathbf{A}_k(\mathbf{I}_\ell)[i,j]) \right)}_{\text{super-resolution prediction}} \tag{8.6}$$

where the third line follows from substituting $\mathbf{X}_{\ell k}$ according to Eq. (8.2). Therefore, we can

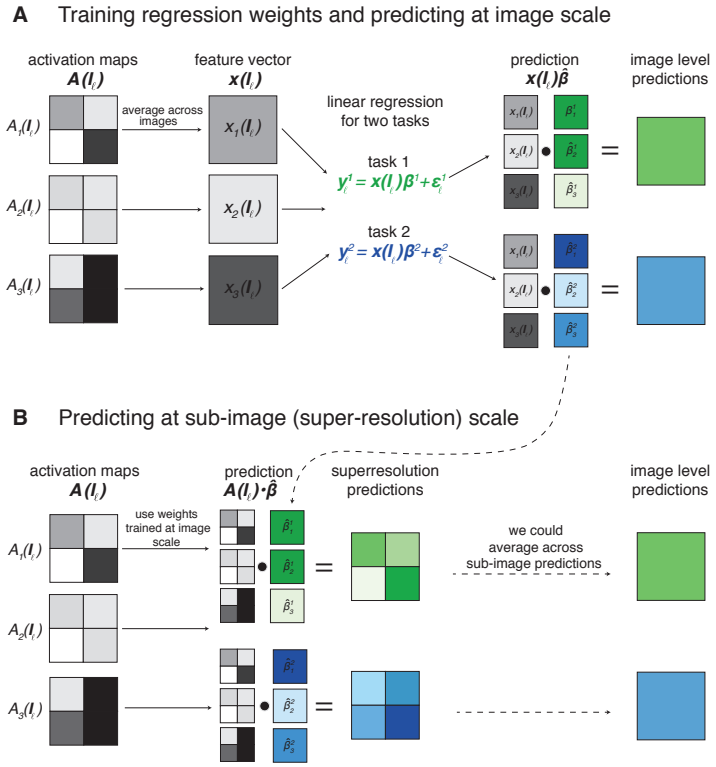


Figure 8.11: Illustration of the procedure to construct predictions at image resolution and label super-resolution. Panel A illustrates the standard MOSAIKS prediction pipeline. Panel B illustrates how the weights trained using labels and features at image-scale in panel A can be used to generate predictions at resolutions higher than the images and labeled data, achieving predictions at label super-resolution. The last column of panel B illustrates the fact that label super-resolution predictions, when averaged across an image, are identical to predictions generated from the standard process in panel A.

associate with each pixel indexed by (i, j) a predicted super-resolution value:

$$\hat{y}^s_{\ell,(i,j)} = \sum_{k=1}^{K} \hat{\beta}^s_k \cdot (\mathbf{A}_k(\mathbf{I}_\ell)[i, j]) \tag{8.7}$$

which is that pixel's predicted label value, and thus its contribution to the overall predicted image-level label value $\hat{y}_\ell$ for $\mathbf{I}_\ell$. We use a Gaussian filter to smooth these per-pixel predictions to enforce spatial consistency and reduce variance of the high-resolution predictions, using a kernel bandwidth of $\sigma = 16$ pixels. These smoothed pixel-level predictions can be average-pooled to larger sub-image scales as shown in Figure 8.10. The procedure to construct label super-resolution predictions, and a comparison to the procedure to construct image-level predictions, is illustrated in Figure 8.11.

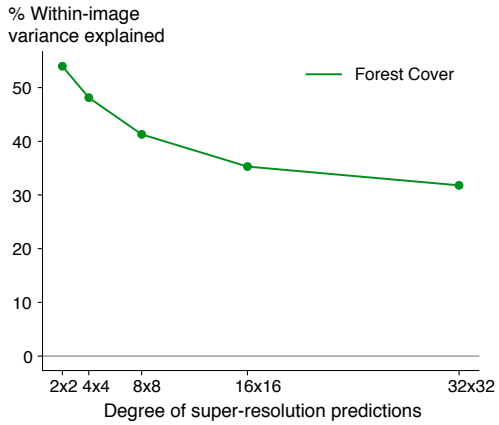


Figure 8.12: Evaluation of within-image R^2 recovered in the forest cover task.

We numerically assess label super-resolution predictions of MOSAIKS for the forest cover task, since raw label data are available at much finer resolution than our image labels. Provided only a single label per image, MOSAIKS recovers substantial within-image signal when predicting forest cover in 4 to 1,024 sub-labels per label[10] (within-image $R^2 = 0.54$-0.32, Figure 8.12) across a subsample of $N = 16,000$ images.

To assess the performance of label super-resolution at a variety of scales, we calculate the percent of the variance of the raw within-image forest cover labels that can be explained by the super-resolution label predictions at each scale. For example, to assess the performance of 2×2 label super-resolution predictions, we average predictions from the 254×254 label super-resolution predictions by quadrants, resulting in four predicted values (twice the original resolution).[11] We perform the same per-quadrant average for the raw fine-resolution forest cover labels. We demean both the within-image predictions and labels to eliminate across-image variation, thereby focusing this test on the ability of the predictions to explain residual within-image variation. We then concatenate these within-image predictions and labels across the $N = 16,000$ sampled images, so that the resulting R^2 value reported is the percent of super-resolution label variance explained by label super-resolution predictions, across $64,000 = 16,000 \cdot 2^2$ label-prediction pairs.

8.7 Conclusions

In this chapter, we develop a new approach to SIML that achieves practical generalization across tasks while exhibiting performance that is competitive with deep learning models op-

[10] We test up to $w = 32$ because the native width of the forest cover labels ($\sim$30m) is just under 1/32 the width of the original image ($\sim$1km).

[11] For the analysis, we clip the images and predictions to 224 x 224 pixels so they are evenly divisible by a 32x super-resolution factor.

timized for a single task. Crucial to context-driven planet-scale analyses, MOSAIKS requires orders of magnitude less computation time to solve a new task than CNN-based approaches. We hope these computational gains, paired with the relative simplicity of using MOSAIKS, will democratize access to global-scale SIML technology and accelerate its application to solving pressing global challenges.

While we have shown that in many cases MOSAIKS is a faster and simpler alternative to existing deep learning methods, there remain contexts in which custom-designed SIML pipelines will continue to play a key role in research and decision-making. Existing ground-based surveys will also remain important. We expect MOSAIKS can complement these efforts, especially in resource constrained settings. For example, MOSAIKS can provide fast assessments to guide slower SIML systems or extend the range and resolution of traditional surveys.

As real-world policy actions increasingly depend on SIML predictions, it is crucial to understand the accuracy, precision and sensitivity of these measurements. The low cost and high speed of re-training MOSAIKS enables thorough stress tests that can support robust SIML-based decision systems. Here, we tested the sensitivity of MOSAIKS to model parameters, number of training points, and degree of spatial extrapolation, and expect that many more tests can be developed and implemented to analyze model performance and prediction accuracies in context. The high performance of RCF, a relatively simple featurization, suggests that developing and benchmarking other unsupervised SIML methods across tasks at scale may be a rich area for future research.

By distilling SIML to a pipeline with simple and mathematically interpretable components, MOSAIKS facilitates development of methodologies for additional SIML use cases and enhanced performance. For example, the ability of MOSAIKS to achieve label super-resolution is easily derived analytically (Section 8.6). Furthermore, while we have focused here on tri-band daytime imagery, we showed that MOSAIKS can seamlessly integrate data from multiple sensors through simple concatenation, extracting useful information from each source to maximize performance. We conjecture that integrating new diverse data, from both satellite and non-satellite sources, may substantially increase the predictive accuracy of MOSAIKS for tasks not entirely resolved by daytime imagery alone; such integration using deep learning models is an active area of research [135].

We hope that MOSAIKS lays the foundation for the future development of an accessible and democratized system of global information sharing, where, over time, imagery from all available global sensors is continuously encoded as features and appended to a single table of data, which is distributed and used planet-wide. Such a unified global system may enhance our collective ability to observe and understand the world, a necessary condition for tackling pressing global challenges.

8.A Data Details

This section first describes the datasets we use to construct our ground truth labels across all seven of our tasks: forest cover, elevation, population density, nighttime lights, income, road length, and housing price. We then describe the imagery used in the analysis, and lastly how label data is assigned to the spatial extent of each sample image.

In evaluating the ability of MOSAIKS to generalize, we are interested in its ability to recover different types of variables, including: (i) variables that are averages of sub-image properties, (ii) variables that are not directly observable through daytime imagery but are a function of visible objects in the image, such as nighttime lights, and (iii) variables that are an underlying factor that determines what material appears in the image, such as elevation. Labels may also be a combinations of (i)-(iii), such as housing price or household income.

For each task, we obtain an up-to-date and geographically complete publicly available datasource to match with the images. Most of these data are based on measurements from 2010 - 2015, though our data on population density draws from sources that date back as far as 2005 in order to achieve global coverage. Our imagery data, from the Google Static Maps API [111], was mostly acquired in 2018, though in some cases images may be a few years older.

Labels

Tasks were chosen to represent outcomes of classes (i)-(iii) above, subject to the condition that high resolution and up-to-date label data are available across the US. Below we describe these data sources.

Forest cover. To measure forest cover, we use globally comprehensive raster data from [123], which is designed to accurately measure forest cover in 2010. This dataset is commonly used to measure forest cover when ground-based measurements are not available [7, 56]. Forest in these data is defined as vegetation greater than 5m in height, and measurements of forest cover are given at a raw resolution of roughly 30m $\times$ 30m. These estimates of annual maximum forest cover are derived from a model based on Landsat imagery captured during the growing season and were derived using different spectral bands than we observe in our imagery, and using information about how surface reflectance changes over the growing season, which we did not observe. This gives us confidence that we are indeed learning to map visual, static, high-resolution imagery to forest cover, rather than simply recovering the model used in [123].[12]

[12]These data were originally accessed at: `https://landcover.usgs.gov/glc/TreeCoverDescriptionAndDownloads.php`. They can now be found from the University of Maryland, Department of Geographical Sciences and USGS at `https://glad.umd.edu/dataset/global-2010-tree-cover-30-m`.

Elevation. We use data on elevation provided by Mapzen, and accessed via the Amazon Web Services (AWS) Terrain Tile service. These Mapzen terrain tiles provide global elevation coverage in raster format. The underlying data behind the Mapzen tiles comes from the Shuttle Radar Topography Mission (SRTM) at NASA's Jet Propulsion Laboratory (JPL), in addition to other open data projects.

These data can be accessed through AWS at different zoom levels, which range from 1 to 14 and, along with latitude, determine the resolution of the resulting raster. To align with the resolution of our satellite imagery, we use zoom level 8, which leads to a raw resolution of 611.5 meters at the equator.[13]

Population density. We use data on population density from the Gridded Population of the World (GPW) dataset [57]. The GPW data estimates population on a global 30 arc-second (roughly 1 km at the equator) grid using population census tables and geographic boundaries. It compiles, grids, and temporally extrapolates population data from 13.5 million administrative units. It draws primarily from the 2010 Population and Housing Censuses, which collected data between 2005 and 2014. GPW data in the US comes from the 2010 census.[14]

Nighttime lights. We use luminosity data generated from nighttime satellite imagery, which is provided by the Earth Observations Group at the National Oceanic and Atmospheric Administration (NOAA) and the National Geogphysical Data Center (NGDC). The values we use are Version 1.3 annual composites representing the average radiance captured from satellite images taken at night by the Visible Infrared Imaging Radiometer Suite (VIIRS). We use values from 2015, the most recent annual composite available [98].

This composite is created after the Day/Night VIIRS band is filtered to remove the effects of stray light, lightening, lunar illumination, lights from aurora, fires, boats, and background light. Cloud cover is removed using the VIIRS Cloud Mask product. These values are provided across the globe from a latitude of 75N to 65S at a resolution of 15 arc-seconds. The radiance units are nW cm^{-2} sr^{-1} (nanowatts per square centimeter per steradian).

Like forest cover, these labels are themselves derived from satellite imagery. However, because they capture luminosity at night, while our satellite imagery is taken during the day, the labels for luminosity and the imagery used to predict luminosity represent independent data sources. Our ability to predict nighttime lights depends on how well objects visible during the day are indicative of light emissions at night.[15]

Income. We use the American Community Survey (ACS) 5-year estimates of median annual household income in 2015. These data are publicly available at the census block group

[13]We accessed these data via the R function `get_aws_terrain` from the `elevatr` package. Code and documentation can be found here: `https://www.github.com/jhollist/elevatr`.

[14]These data can be accessed at `http://sedac.ciesin.columbia.edu/data/collection/gpw-v4` and are licensed for use under a Creative Commons Attribution 4.0 International License.

[15]These data can be accessed at `https://eogdata.mines.edu/products/vnl/`.

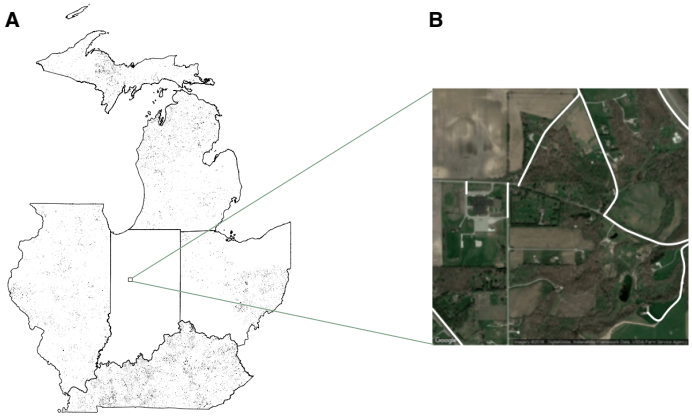


Figure 8.13: Quality of ground truth road data varies by region. (A) Private roads in the northern Midwest recorded in the USGS National Transportation Dataset. There is a noticeable lack of recorded private roads in Indiana and sections of Ohio. (B) Overlaying recorded roads of all types (shown in white) over a single satellite image (from Google Static Maps) in Indiana demonstrates that some roads that are easily visible from satellite imagery are missing in the data that we use to construct labels.

level, of which there are 211,267 in the US, including Puerto Rico. On average, block groups are around 38 km^2, though block groups are smaller in more densely populated areas.[16]

Note that all estimates in this chapter that are based off of the ACS use the Census Bureau Data API but are not endorsed or certified by the Census Bureau.

Road length. We use road network data from the United States Geological Survey (USGS) National Transportation Dataset, which is based on TIGER/Line data provided by US Census Bureau in 2016. Shapefiles for each state provide road locations and types, including highways, local neighborhood roads, rural roads, city streets, unpaved dirt trails, ramps, service drives, and private roads. The variable we predict is road length (in meters), which is computed as the total length of all types of roads that are recorded in a given grid cell.

The Census Bureau database is created and corrected via a combination of partner supplied data, aerial images, and fieldwork. The spatial accuracy of linear features of roads and coordinates vary by source materials used. The accuracy also differs by region, causing cases in which some regions lack recordings of certain road types, the most common one being private roads and dirt trails. For example, private roads are rarely recorded in Indiana and some regions in Ohio despite satellite images that suggest they are present (Figure 8.13).[17]

[16]These data are accessible using the `acs` package in R [109], table number B19013.

[17]The data can be accessed at: `https://prd-tnm.s3.amazonaws.com/index.html?prefix=StagedProducts/Tran/Shape/`.

Housing price. We estimate housing price per square foot using sale price and assessed square footage values for residential buildings. Data are provided by Zillow through the Zillow Transaction and Assessment Dataset (ZTRAX). This dataset aggregates transaction and assessment data across the United States, combining reported values from states and counties with widely varying regulations and standards. Thus, significant data cleaning is required. Furthermore, because some states do not require mandatory disclosure of the sale price, we have limited data for the following states: Idaho, Indiana, Kansas, Mississippi, Missouri, Montana, New Mexico, North Dakota, South Dakota, Texas, Utah, and Wyoming. To address data quality issues, we develop a quality assurance and quality control (QA/QC) approach that is based on approaches employed in previous work [108, 205, 281] but adapted for our case.

ZTRAX contains data on the majority of buildings in the United States, initially comprising 374 million detailed records of transactions across more than 2,750 counties. The data is organized into two components - *transaction data* and *assessment data*. These two datasets are linked, allowing us to merge the latest sale price of a property to the latest assessment data. To minimize the effect of nation-wide trends in housing price that would be unobservable from our cross-sectional satellite imagery, we limit our dataset to sales occurring in 2010 or later. Further, we restrict our analysis to buildings coded as "residential" or "residential income - multi-family" and drop any sale that was coded as an intra-family transfer. To obtain a square footage value, we follow the example in Zillow Research's GitHub repository [313] and take the maximum reported square footage for a given improvement, and then sum over all improvements on a given property.

To reduce the number of potentially miscoded outliers at the bottom end of the distribution of sale price and property size, we drop any remaining sales that fall under $10,000 USD, any properties that fall under 100 sq. ft., and any $/sq. ft. values under $10. To address outliers on the high end of the distribution, we take this restricted sample and further cut our dataset at the 99th percentile of $/sq. ft. by state. Afterwards, we select the most recent recorded sale price for each property (divided by the most recent assessed square footage). We then average across all of the remaining units within each grid cell to comprise our final dataset of housing price per square foot.

To protect potentially identifiable information, our public data release contains housing price labels only for grid cells that contain 30 or more sales meeting the aforementioned criteria. This reduces the size of the dataset from $N = 80,420$ to $N = 52,355$ and makes the model performance better than that stated in the main text. For example, the public dataset will yield a test set R^2 of 0.60, rather than 0.52. This could be due to the fact that the average housing price label we train on is noisier when estimated in a grid cell with few valid sales prices. It could also be because the average housing price of areas with few recent sales may be inherently harder to predict via satellite imagery than that of areas with a greater number of recent sales.

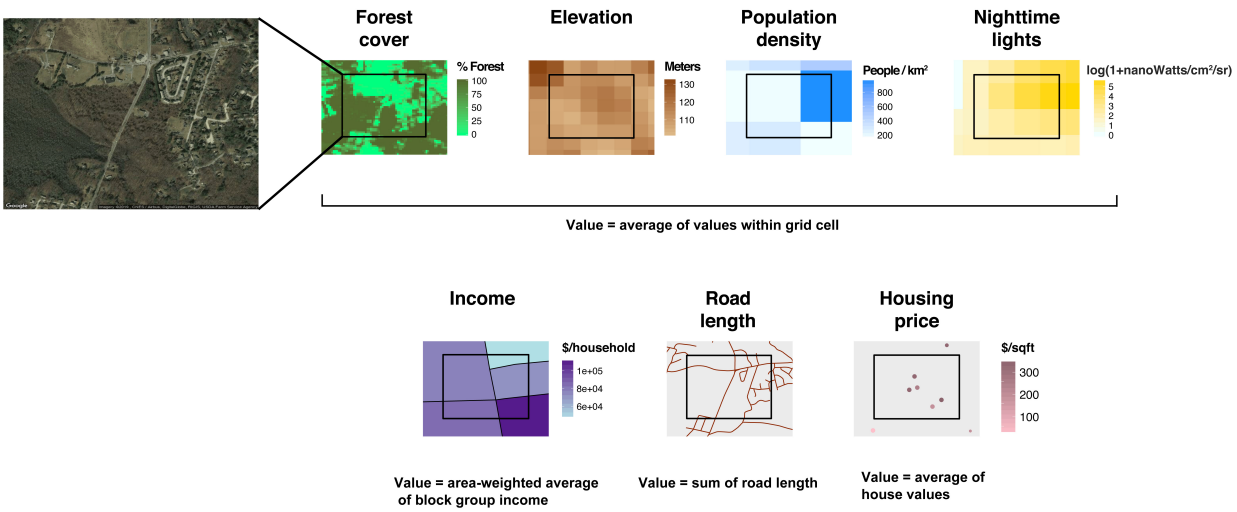


Figure 8.14: Assignment of grid cell labels from raw data. We calculate labels by spatially overlaying our grid cells and raw labeled data. We calculate labels as the average of raw label values that fall within the grid cell, except for roads where we calculate the label as the sum of road length within the grid cell. Example image from Google Static Maps.

Imagery

We use satellite imagery from Google Static Maps API [111], zoom level 16. This gives roughly 1km × 1km images which are 640 × 640 pixels across and 3 dimensions deep (red, green, and blue spectral bands). We coarsen these images to 256 × 256 × 3 prior to featurizing, meaning that our models are trained on images with roughly 4m resolution. These images can be composites of several satellite images – sources include the Landsat, Sentinel, SPOT, Pleiades, WorldView and QuickBird satellites.[18] Prior to downloading, images were geo-rectified and pre-processed to remove cloud occlusions.[19]

Assigning Label Data to Sampled Imagery

To assign labels to each grid cell, we spatially overlay our raw labeled data and our custom grid. The native format and spatial resolution of the labeled data vary across the tasks studied, necessitating different aggregation or disaggregation procedures for each task. Here, we describe the approach taken in each task (Figure 8.14). In these cases, we aggregate labeled data up to the grid cell level.

The raw forest cover, elevation, population density and nighttime lights data are provided natively as rasters with higher spatial resolution than our custom grid. For these tasks, we

[18]In some cases aerial photography is also integrated into images.

[19]More information is available at: `https://developers.google.com/maps/documentation/maps-static/dev-guide`.

perform aggregation by calculating the mean of all labeled pixels with centroids that fall within the imagery grid cell. The resulting labels indicate mean forest cover, mean elevation, mean population density, and mean nighttime lights across the image grid cell.

Our road length data are provided as high-resolution spatial line segments. To aggregate these data to the image grid cell, we calculate the sum of road length segments within each image grid cell.

Our housing price data are available as individual house sales. We aggregate these geocoded prices to the image grid cell by taking the average housing price per square foot across all sale prices that fall within the extent of the image. The resulting labels indicate the average housing price per square foot across all observed houses within a grid cell.

Our income data are provided at the block-group level. In some parts of the U.S., these block-groups are larger in total area than our image grid cells. However, in other regions, block-groups are smaller than our image grid cells. To treat both cases consistently, we aggregate incomes to the grid cell level by taking the weighted average of block-group incomes, where the weights are the area of intersection between the image grid cell and the block-group polygons. These weights are normalized to unity for each grid cell. The resulting labels indicate the area-weighted average median income across the grid cell.

Future users of a production-scale version of MOSAIKS would employ label data of arbitrary format and resolution. The above approaches provide guidelines for how to match various forms of label data to the pre-computed image feature grid, but other methods may be used. In the simplest case, for example, sparse point data could be directly matched to the nearest grid cell centroid.

Part IV

Connections, Conclusions, and Perspectives

Chapter 9

Discussion

We began this thesis by contrasting the potential for statistical machine learning systems to help humans improve our world with the evidence that in practice, such systems often fail to live up to this potential. Motivated to move toward this potential for benefit with machine learning systems, in Chapter 1 we introduced the themes of intent, impact, and context that interweave throughout this thesis.

In Part I, we saw that algorithmic systems that are theoretically well motivated can fail to live up to anticipated performance in the real world, and can exacerbate inequality rather than relieve it. Contrasting the *intents* and *impacts* of machine learning systems led us to develop a framework in which an intended benefit — for example, social welfare — can be expressly delineated and optimized alongside multiple objectives.

In Part II, we studied context in general statistical machine learning settings by focusing on specific structures in input data. We leveraged group structures in datasets to study the importance of numerical representation to achieving high group- and population-level accuracies. In a different problem setting, we designed a post-processing method to utilize structural side information like spatio-temporal correlations that can amplify predictive signal, but can hurt generalization performance if treated like traditional features. Alongside these examples of considering data as a *useful* form of context, in Chapter 6 we discussed limitations of using data to encode context, underscoring that data should never be the *only* form of context in real-world machine learning systems.

In Part III, we anchored the themes of intent and context with applications in machine learning with remotely sensed data. Making use of precise structures in sensor data, we designed simple but effective learning algorithms in two different settings. In an adaptive data acquisition problem, we developed a statistically-aware trajectory planning method to speed up localization of environmental radioactivity with quadcopter robots. In light of the potential for combining machine learning and remotely sensed data to confer social benefit discussed in Chapter 1, our second example focused on simplifying machine learning with satellite imagery, with the intent to make such technologies accessible to a much broader set of researchers.

We now conclude by synthesizing the three major parts of this thesis and the connections

that interweave throughout them. In Section 9.1, we draw connections between chapters spanning Parts I to III, emphasizing potential for future work at their unions. We conclude with overarching perspectives in Section 9.2.

9.1 Connections and Conclusions

Interconnections between Intent and Context

The first two parts of this thesis largely separated the focuses of intent and impact (Part I) from using data as a form of context (Part II). Separating these focuses allowed us to isolate and frame our scope of inquiry in each chapter. Of course, producing effective data-driven learning systems will require integrating these themes together. We need both the capability to delineate and balance specific system objectives and the understanding of how different input data influence our ability to reach diverse and multifaceted intents.

For example, consider a scenario in which a prediction system is used to recommend individuals for selection for targeted humanitarian aid [6]. If the program designers aim to disburse aid so as to improve welfare along multiple dimensions (e.g. education and health), Chapter 3 details how policies acting on multiple learned scores can be used to strike optimal balances between these objectives. As we saw in Section 3.3, the range of possible utilities in both dimensions expands if the predicted scores are more accurate. If the program designers are willing to spend some of their total budget to collect more data to update and improve the predicted scores, they will be able to target aid more precisely and efficiently with the remaining budget, possibly reaching higher utility solutions. The question is then how much budget to spend on improving the model versus targeting aid with an imperfect model. A comprehensive answer requires the designers to understand the costs of acquiring different forms of data and the effects on system performance (broadly construed) that additional data will confer. The latter of those questions was a fundamental point of study in Part II of this thesis.

The interconnections between intent and impact span a broad range of exciting research directions. Like the example in the previous paragraph, future research could integrate constraints on downstream uses cases that can help guide system design, including data acquisition and algorithmic improvements. Integrating the intent-driven perspective of Chapter 3 with the focus on representation in Chapter 4 can guide end-to-end design of machine learning systems that incorporate diverse system objectives, including targeted benefit, privacy, or participation. Moreover, perhaps such a synthesis can also help describe why existing systems might not reach these diverse goals. Extending our study of the different characteristics data might take on, as in Chapter 5, with careful considerations of the limitations of using data as a way to encode context (Chapter 6) and potential harmful impacts of well intentioned systems (Chapter 2) can guide future work toward nuanced, effective, and explainable notions of how data and representation reflect one another.

Intent, Impact, and Context in Remote Sensing and Machine Learning

Part III of this thesis added application-driven perspectives to situating learning systems within their broader context, complementing the more generalized problem settings considered in Parts I and II. In Chapter 7, we studied a physical sensing problem with a clear and singular intent: use the sensor-equipped quadcopter to locate the environment point emitting the largest amount of radiation, as quickly as possible. Modeling the specific structure of the onboard omni-directional radiation sensor and its measurements of the environment guided us to cast the source identification problem as a multi-armed bandit problem with travel constraints, from which we derived a sample-efficient trajectory planning method. A key insight was to consider the value of sensing configurations across a global set of possible configurations. As a result, our algorithm adaptively allocates data acquisition in a principled manor that does not fall prey to the same myopic phenomena of more local or greedy trajectory planning approaches.

In Chapter 8, we presented a very different instantiation of intent- and context-aware design to combining machine learning and satellite imagery. Like in the previous example, we made use of specific structures of the remotely sensed data to design suitable machine learning methods within our predictive system. Levering scale and rotation invariances common to many satellite imagery prediction tasks, we used a simple computer vision algorithm to obtain an effective unsupervised embedding of imagery. The desire for such simplifications was motivated by the intent of having this single embedding of the satellite imagery encode relevant statistical properties useful to *multiple* possible downstream prediction tasks, where the prediction step can be achieved with a simple model (in the case of MOSAIKS, linear regression). The practical effect of these design choices was to share computational cost in the primary embedding step, and to make global prediction with satellite imagery accessible for researchers with a broader set of possible backgrounds and computational resources. Understanding and designing for this potential research-ecoystem level impact required integrating a broader type of context: the ways in which such a prediction system would be likely to be used in practice, and by whom.

The application-centric perspectives in Chapters 7 and 8 not only evidence intent- and context-aware design in practice, but they also show how real-world problems can expose opportunities to expand general learning frameworks. Our application of the multi-armed bandits framework in Chapter 7 showed that a statistical framework typically abstracted from physical problems can confer advantages when adapted for embodied settings. Furthermore, we incorporated characteristics of the physical environment into performance and regret bounds from which we can ascertain for which physical environments our algorithm will be most appropriate. In Chapter 8, we delineated the potentially transformative impacts of a single embedding of satellite imagery to generalize across prediction tasks. This insight highlights the importance of a growing research focus on unsupervised and self-supervised methods for satellite imagery (e.g. [15, 146, 189]). The competitiveness of random convolutional features – a relatively simple computer vision algorithm – to fine-tuning with full

supervision suggests a baseline for such improvements, as well as an algorithmic building block from which to instantiate them.

Looking forward, it is important to address unique issues of fairness, responsibility, and ethics of the emerging field intersecting machine learning, remote sensing, and policy. For example, sparse or clustered label data may elicit concerns over data representation in learning settings [38]. Disparately inaccurate or uncertain predictions across geographies can affect the validity and fairness of downstream policy decisions [167, 304].

Addressing these issues requires a blend of algorithmic innovations and interdisciplinary viewpoints, drawing on and expanding beyond the themes of intent, impact and context developed in Parts I to III. Future work bridging these themes could focus on the key notions of representation, data quality, and fairness with remotely sensed data. Relating to and expanding beyond the instantiation of these themes in Chapters 2 to 8, an important area of future work is to address issues of privacy, surveillance, power, and politics that arise in using machine learning and remotely sensed data for consequential predictions.

9.2 Perspectives

Through the work presented in this thesis, we make a start toward understanding how we might characterize and design for context, intent, and impact in machine learning systems. In many ways, it is just a start. In Section 9.1, we outlined opportunities for future research bridging the core themes of Parts I to III. While such work will be integral to framing statistical learning systems so that they might confer benefit, actually achieving that benefit will often necessitate taking a broader perspective. We illustrate this point with two challenges that exemplify the need for interconnecting diverse perspectives if we are to address real world problems with machine learning.

Our first exemplary challenge is that of *reconciling the potential of machine learning for the environment with the environmental and social impacts of training and deploying large predictive models.* As we saw in Chapters 1, 7 and 8, combining machine learning with remote sensing can help shed light on hard-to-reach regions of the world, helping humans make more informed decisions to monitor and address climate and ecological change. But alongside examples applying machine learning to monitor deforestation [85, 215], biodiversity changes [277], and climate change [248] come numerous concerns of the environmental toll of training and deploying large machine learning models [32, 255]. The tension between the potential for machine learning to promote or impede sustainability goals has been noted by previous researchers [78], notably in the practice of "green AI" [255].

Navigating the range of possible positive and negative impacts machine learning solutions might have on our environment will be an interdisciplinary endeavor. It will for example involve estimating and monitoring policy impacts of the proposed technologies in light of the environmental costs associated with training and maintaining machine learning models. Toward such an effort, Schwartz et al. [255] encourage creators of machine learning models to publish metrics of energy and computation requirements for training and prediction

alongside the models themselves. They suggest focusing on efficiency of models as a complementary effort to advancing benchmark accuracy metrics, a goal exemplified by our work in Chapter 8. Simpler and more efficient models have the additional advantage of increasing accessibility to a wider range of researchers who can use, study, or audit their use in practice – a current downside of large models [268]. Equity in access to predictive models is especially pivotal as machine learning systems are monitored and maintained by human oversight and intervention.

This brings us to a second exemplary challenge: *ensuring that data driven strategies enhance rather than divert resources from addressing underlying social problems.* A common thread underlying Chapters 2, 3 and 6 is that data-driven algorithms encode social context in intricate and imperfect ways. In the best case, predictive models can still help guide decisions by providing valuable, if incomplete information. Predictions of poverty levels using remote sensing and digital trace data can help target aid to the poorest people in a region, especially valuable in places where traditional censuses are sparse [6, 145]. In using such machine learning systems, one must contend with the knowledge that "automated decision-making... reframes shared social decisions about who we are and who we want to be as systems engineering problems," as scholar Virginia Eubanks writes in her book *Automating inequality: How high-tech tools profile, police, and punish the poor* [101].

The dangers of focusing on efficiency above efficacy and short-term metrics above long-term solutions are many. As we argued in Section 6.2, data exist within social contexts and structures of power. Prediction models and their outputs are no different, and thus can reflect or reinforce harmful or unjust social systems [33, 101, 211]. A perhaps more subtle concern is that assigning resources to deriving machine learning-based solutions can divert those resources from directly addressing social problems. As Moore [202] puts it, "many of these problems, like poverty, recidivism, and the distribution of resources, are ones of institutional failure. Technology-based approaches, when not aimed at the root of problems, divert attention from the proper recourse: structural change." Even within a well-motivated machine learning solution, over-reliance on imperfect proxy predictions has the potential to disincentivize valuable data collection efforts that would result in more reliable information in the future.

Virginia Eubanks goes on to explain that "political contests are more than informational; they are about values, group membership, and balancing conflicting interests" [101]. Situating technical solutions within political systems thus requires critical thought as to what information is helpful and how that information can be used [3, 113]. Moreover, understanding the myriad risks and benefits of machine learning systems is not a one-time affair, but something that can be continuously re-assessed from multiple perspectives.

Addressing these types of broader challenges requires re-contextualizing what aspects of a machine learning *system* (Figure 1.1) to include in the design space. Current lines of inquiry within machine learning research already incorporate certain aspects of context, for example by addressing human incentives, computational efficiency, and deep integration with a variety of application spaces. Still, a dominant focus in machine learning research aims toward improving model performance on benchmark datasets as a measure of progress [233,

255]. The body of work presented in this thesis suggests that a different sort of progress can be made by expanding the purview of the machine learning system to incorporate key forms of context by design.

As echoed throughout this thesis, context is layered. Incorporating these different layers of context in machine learning systems offers several opportunities for individual researchers and the field at large. First is an opportunity to leverage data as a formalizable encoding of context, toward answering what exactly data-driven algorithms and analysis *can and cannot* provide. This could take shape using tools in statistics, mathematics, and computing to delineate capabilities and guide-rails of learning systems. A second layer of context wraps tightly around a particular machine learning system and its scope of application. Here researchers in machine learning have the opportunity to leverage cross-disciplinary collaborations and partnerships to understand the structural, methodological and ethical elements of the problem before co-designing a solution. A broader layer of social, environmental, and political context encompasses more than any one or two disciplines. This layer evolves at a quicker pulse, as our societies change rapidly in tandem with the technologies that interface with us. It confers the opportunities to listen, learn, and adapt capabilities of machine systems to the needs of the world around us.

Bibliography

[1] Mohsen Abbasi, Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. "Fairness in representation: Quantifying stereotyping as a representational harm". In: *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM. 2019, pp. 801–809.

[2] Rediet Abebe, Kehinde Aruleba, Abeba Birhane, Sara Kingsley, George Obaido, Sekou L Remy, and Swathi Sadagopan. "Narratives and counternarratives on data sharing in Africa". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 329–341.

[3] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. "Roles for computing in social change". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 252–260.

[4] Jacob Abernethy, Pranjal Awasthi, Matthäus Kleindessner, Jamie Morgenstern, and Jie Zhang. "Adaptive sampling to reduce disparate performance". In: *arXiv preprint arXiv:2006.06879* (2020).

[5] Alekh Agarwal, Sham M Kakade, Nikos Karampatziakis, Le Song, and Gregory Valiant. "Least squares revisited: Scalable approaches for multi-class prediction". In: *International Conference on Machine Learning*. 2014, pp. 541–549.

[6] Emily Aiken, Suzanne Bellue, Dean Karlan, Chris Udry, and Joshua E Blumenstock. "Machine learning and phone data can improve targeting of humanitarian aid". In: *Nature* (2022), pp. 1–7.

[7] Ramdane Alkama and Alessandro Cescatti. "Biophysical climate impacts of recent changes in global forest cover". In: *Science* 351.6273 (Feb. 2016), pp. 600–604.

[8] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine bias". In: *Ethics of Data and Analytics*. Auerbach Publications, 2016, pp. 254–264.

[9] Kenneth Joseph Arrow. *Social choice and individual values*. 12. Yale University Press, 1963.

[10] Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. "Learning under selective labels in the presence of expert consistency". In: *2018 Workshop on Fairness, Accountability, and Transparency in Machine Learning* (2018). URL: `https://arxiv.org/abs/1807.00905`.

[11] Nikolay Atanasov, Jerome Le Ny, Kostas Daniilidis, and George J Pappas. "Information acquisition with sensing robots: Algorithms and error bounds". In: *International Conference on Robotics and Automation*. IEEE. 2014, pp. 6447–6454. URL: http://ieeexplore.ieee.org/document/6907811/?section=abstract.

[12] Susan Athey. "Beyond prediction: Using big data for policy problems". In: *Science* 355.6324 (Feb. 2017), pp. 483–485. DOI: 10.1126/science.aal4321. URL: http://science.sciencemag.org/content/355/6324/483.abstract.

[13] Jean-Yves Audibert and Sébastien Bubeck. "Best arm identification in multi-armed bandits". In: *Conference on Learning Theory*. 2010, 13–p. URL: %7Bhttps://hal.inria.fr/hal-00654404/%7D.

[14] Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. "Equalized odds postprocessing under imperfect group information". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 1770–1780.

[15] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. "Geography-aware self-supervised learning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10181–10190.

[16] Olena Babak and Clayton V Deutsch. "Statistical approach to inverse distance interpolation". In: *Stochastic Environmental Research and Risk Assessment* 23.5 (2009), pp. 543–553.

[17] Sudharshan Chandra Babu. *Human-pose-estimation-101*. 2019. URL: https://github.com/cbsudux/Human-Pose-Estimation-101.

[18] Philip Bachman, Ouais Alsharif, and Doina Precup. "Learning with pseudo-ensembles". In: *Advances in Neural Information Processing Systems*. 2014, pp. 3365–3373.

[19] Shi Bai, Jinkun Wang, Fanfei Chen, and Brendan Englot. "Information-theoretic exploration with Bayesian optimization". In: *International Conference on Intelligent Robots and Systems*. IEEE. 2016, pp. 1816–1822. URL: http://personal.stevens.edu/~benglot/Bai_Wang_Chen_Englot_IROS2016_AcceptedVersion.pdf.

[20] Eric Balkanski and Yaron Singer. "The sample complexity of optimizing a convex function". In: *Conference on Learning Theory*. 2017, pp. 275–301.

[21] John E. Ball, Derek T. Anderson, and Chee Seng Chan. "A comprehensive survey of deep learning in remote sensing: Theories, tools and challenges for the community". In: *Journal of Applied Remote Sensing* 11.4 (2017).

[22] Deborah L Bandalos. *Measurement theory and applications for the social sciences*. New York, NY: Guilford Publications, Mar. 2018.

[23] Sudipto Banerjee, Alan E Gelfand, Andrew O Finley, and Huiyan Sang. "Gaussian predictive process models for large spatial data sets". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.4 (2008), pp. 825–848.

[24] Solon Barocas, Asia J. Biega, Benjamin Fish, Jundefineddrzej Niklas, and Luke Stark. "When not to design, build, or deploy". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, p. 695. ISBN: 9781450369367. DOI: 10.1145/3351095.3375691. URL: https://doi.org/10.1145/3351095.3375691.

[25] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. "The problem with bias: Allocative versus representational harms in machine learning". In: *9th Annual Conference of the Special Interest Group for Computing, Information and Society*. 2017.

[26] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. http://www.fairmlbook.org. fairmlbook.org, 2019.

[27] Solon Barocas and Andrew D Selbst. "Big data's disparate impact". In: *California Law Review* 104 (2016), p. 671.

[28] Cenk Baykal, Guy Rosman, Sebastian Claici, and Daniela Rus. "Persistent surveillance of events with unknown, time-varying statistics". In: *2017 IEEE International Conference on Robotics and Automation*. 2017, pp. 2682–2689. DOI: 10.1109/ICRA.2017.7989313.

[29] Mikhail Belkin and Partha Niyogi. "Semi-supervised learning on Riemannian manifolds". In: *Machine Learning* 56.1-3 (2004), pp. 209–239.

[30] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples". In: *Journal of Machine Learning Research* 7.Nov (2006), pp. 2399–2434. URL: http://www.jmlr.org/papers/volume7/belkin06a/belkin06a.pdf.

[31] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. "Does data interpolation contradict statistical optimality?" In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 1611–1619.

[32] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the dangers of stochastic parrots: Can language models be too big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 610–623.

[33] Ruha Benjamin. *Race after technology: Abolitionist tools for the new jim code*. Oxford, England: Polity Press, 2019.

[34] *Big data for sustainable development*. URL: https://www.un.org/en/global-issues/big-data-for-sustainable-development.

[35] Reuben Binns. "On the apparent conflict between individual and group fairness". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 514–524.

[36] Emrah Bıyık and Murat Arcak. "Gradient climbing in formation via extremum seeking and passivity-based coordination rules". In: *Asian Journal of Control* 10.2 (2008), pp. 201–211. URL: `http://onlinelibrary.wiley.com/doi/10.1002/asjc.19/full`.

[37] Joshua Blumenstock. *Don't forget people in the use of big data for development.* 2018.

[38] Ian Bolliger, Tamma Carleton, Solomon Hsiang, Jonathan Kadish, Jonathan Proctor, Benjamin Recht, Esther Rolf, and Vaishaal Shankar. "Ground control to Major Tom: The importance of field surveys in remotely sensed data analysis". In: *arXiv preprint arXiv:1710.09342* (2017).

[39] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings". In: *Advances in Neural Information Processing Systems* 29 (2016).

[40] Elizabeth Bondi, Lily Xu, Diana Acosta-Navas, and Jackson A Killian. "Envisioning communities: A participatory approach towards AI for social good". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* 2021, pp. 425–436.

[41] Christine L Borgman. *Big data, little data, no data: Scholarship in the networked world.* MIT press, 2017.

[42] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. "Theory of classification: A survey of some recent advances". In: *ESAIM: Probability and Statistics* 9 (2005), pp. 323–375.

[43] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A Nonasymptotic theory of independence.* Oxford university press, 2013.

[44] Frederic Bourgault, Alexei A Makarenko, Stefan B Williams, Ben Grocholsky, and Hugh F Durrant-Whyte. "Information based adaptive robotic exploration". In: *International Conference on Intelligent Robots and Systems.* Vol. 1. IEEE. 2002, pp. 540–545. URL: `https://ieeexplore.ieee.org/abstract/document/1041446/`.

[45] Danah Boyd and Kate Crawford. "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon". In: *Information, Communication & Society* 15.5 (2012), pp. 662–679.

[46] Maria Antonia Brovelli, Yaru Sun, and Vasil Yordanov. "Monitoring forest change in the amazon using multi-temporal remote sensing data and machine learning classification on Google Earth Engine". In: *ISPRS International Journal of Geo-Information* 9.10 (2020), p. 580.

[47] Sébastien Bubeck, Michael B Cohen, Yin Tat Lee, James R Lee, and Aleksander Mądry. "K-server via multiscale entropic regularization". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing.* 2018, pp. 3–16.

[48] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. "A systematic study of the class imbalance problem in convolutional neural networks". In: *Neural Networks* 106 (2018), pp. 249–259.

[49] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Conference on Fairness, Accountability and Transparency*. 2018, pp. 77–91.

[50] Abram Burk. "A reformulation of certain aspects of welfare economics". In: *The Quarterly Journal of Economics* 52.2 (1938). Publisher: MIT Press, pp. 310–334.

[51] Marshall Burke, Anne Driscoll, David B Lobell, and Stefano Ermon. "Using satellite imagery to understand and promote sustainable development". In: *Science* (2021).

[52] Jonathon Byrd and Zachary Chase Lipton. "What is the effect of importance weighting in deep learning?" In: *Proceedings of the 36th International Conference on Machine Learning,* vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 872–881. URL: http://proceedings.mlr.press/v97/byrd19a.html.

[53] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334 (2017), pp. 183–186.

[54] Ayse Can. "Specification and estimation of hedonic housing price models". In: *Regional Science and Urban Economics* 22.3 (1992), pp. 453–474.

[55] Andrew Caplin, Sumit Chopra, John Leahy, Yann LeCun, and Trivikraman Thampy. "Machine learning and the spatial structure of house prices and housing returns". In: *Available at SSRN 1316046* (2008).

[56] Kimberly M Carlson, Robert Heilmayr, Holly K Gibbs, Praveen Noojipady, David N Burns, Douglas C Morton, Nathalie F Walker, Gary D Paoli, and Claire Kremen. "Effect of oil palm sustainability certification on deforestation and fire in Indonesia." In: *Proceedings of the National Academy of Sciences of the United States of America* 115.1 (Jan. 2018), pp. 121–126.

[57] Center for International Earth Science Information Network (CIESIN). *Gridded Population of the World, Version 4*. 2016. DOI: http://dx.doi.org/10.7927/H4NP22DQ.

[58] Benjamin Charrow, Sikang Liu, Vijay Kumar, and Nathan Michael. "Information-theoretic mapping using Cauchy-Schwarz quadratic mutual information". In: *International Conference on Robotics and Automation*. IEEE. 2015, pp. 4791–4798. URL: http://mrsl.grasp.upenn.edu/bcharrow/ICRA2015tech.pdf.

[59] Kyla Chasalow and Karen Levy. "Representativeness in statistics, politics, and machine learning". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 77–89. URL: https://doi.org/10.1145/3442188.3445872.

[60] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. "SMOTE: Synthetic minority over-sampling technique". In: *Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357.

[61] Irene Y. Chen, Fredrik D. Johansson, and David A. Sontag. "Why is my classifier discriminatory?" In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems*. 2018, pp. 3543–3554.

[62] Irene Y. Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. "Ethical machine learning in healthcare". In: *Annual Review of Biomedical Data Science* 4.1 (2021), pp. 123–144. DOI: 10.1146/annurev-biodatasci-092820-114757.

[63] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. "Fairness under unawareness: Assessing disparity when protected class is unobserved". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019, pp. 339–348.

[64] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2016, pp. 785–794.

[65] Anil M. Cheriyadat. "Unsupervised feature learning for aerial scene classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 52.1 (2014), pp. 439–451.

[66] Alexandra Chouldechova. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *Big Data* 5.2 (2017), pp. 153–163.

[67] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions". In: *Conference on Fairness, Accountability and Transparency*. PMLR. 2018, pp. 134–148.

[68] John M Clapp, Hyon–Jung Kim, and Alan E Gelfand. "Predicting spatial patterns of house prices using LPR and Bayesian smoothing". In: *Real Estate Economics* 30.4 (2002), pp. 505–532.

[69] Andrew Clark and Andrew Oswald. "Satisfaction and comparison income". In: *Journal of Public Economics* 61.3 (1996). Publisher: Elsevier, pp. 359–381. ISSN: 0047-2727. URL: https://econpapers.repec.org/article/eeepubeco/v_3a61_3ay_3a1996_3ai_3a3_3ap_3a359-381.htm.

[70] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. "Power-law distributions in empirical data". In: *SIAM Review* 51.4 (2009), pp. 661–703.

[71] Adam Coates, Ann Arbor, and Andrew Y Ng. "An Analysis of single-layer networks in unsupervised feature learning". In: *International Conference on Articial Intelligence and Statistics* (2011), pp. 215–223. DOI: 10.1109/ICDAR.2011.95.

[72] Adam Coates and Andrew Y Ng. "Learning feature representations with K-means". In: *Neural Networks: Tricks of the Trade*. Springer, Berlin, Heidelberg, 2012. DOI: 10.1007/978-3-642-35289-8{_}30.

[73] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)". In: *arXiv preprint arXiv:1902.03368* (2019).

[74] Richard Cole, Vasilis Gkatzelis, and Gagan Goel. "Mechanism design for fair division: Allocating divisible items without payments". In: *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*. EC '13. Philadelphia, Pennsylvania, USA: Association for Computing Machinery, 2013, pp. 251–268.

[75] Sam Corbett-Davies and Sharad Goel. "The measure and mismeasure of fairness: A critical review of fair machine learning". In: *arXiv preprint arXiv:1808.00023* (2018).

[76] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. "Learning bounds for importance weighting". In: *24th Annual Conference on Neural Information Processing Systems 2010*. 2010. URL: https://proceedings.neurips.cc/paper/2010/hash/59c33016884a62116be975a9bb8257e3-Abstract.html.

[77] Nick Couldry and Ulises A Mejias. "Data colonialism: Rethinking big data's relation to the contemporary subject". In: *Television & New Media* 20.4 (2019), pp. 336–349.

[78] Josh Cowls, Andreas Tsamados, Mariarosaria Taddeo, and Luciano Floridi. "The AI gambit: Leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations". In: *AI & Society* (2021), pp. 1–25.

[79] Fabio Gagliardi Cozman and Ira Cohen. "Unlabeled data can degrade classification performance of generative classifiers." In: *Flairs Conference*. 2002, pp. 327–331.

[80] Kate Crawford. "The trouble with bias". NIPS Keynote. 2017. URL: https://www.youtube.com/watch?v=fMym_BKWQzk.

[81] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. "Learning 3d human pose from structure and motion". In: *Proceedings of the European Conference on Computer Vision*. 2018, pp. 668–683.

[82] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. "Are we really making much progress? A worrying analysis of recent neural recommendation approaches". In: *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM. 2019, pp. 101–109.

[83] Jeffrey Dastin. "Amazon scraps secret AI recruiting tool that showed bias against women". In: *Ethics of Data and Analytics*. Auerbach Publications, 2018, pp. 296–299.

[84] Amit Datta, Michael Carl Tschantz, and Anupam Datta. "Automated experiments on ad privacy settings". In: *Proceedings on Privacy Enhancing Technologies* (2015), pp. 92–112.

[85] Pablo Pozzobon De Bem, Osmar Abílio de Carvalho Junior, Renato Fontes Guimarães, and Roberto Arnaldo Trancoso Gomes. "Change detection of deforestation in the Brazilian Amazon using landsat data and convolutional neural networks". In: *Remote Sensing* 12.6 (2020), p. 901.

[86] A. M. De Sherbinin, G. Yetman, K. MacManus, and S. Vinay. "Improved mapping of human population and settlements through integration of remote sensing and socioeconomic data". In: *AGUFM* (2017).

[87] Angus Deaton. "Measuring and understanding behavior, welfare, and poverty". In: *American Economic Review* 106.6 (2016), pp. 1221–43.

[88] Angus Deaton. *The measurement of welfare: Theory and practical guidelines*. English. World Bank, Development Research Center, 1980.

[89] Kalyanmoy Deb and Deb Kalyanmoy. *Multi-objective optimization using evolutionary algorithms*. New York, NY, USA: John Wiley & Sons, Inc., 2001. ISBN: 047187339X.

[90] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. "Bringing the people back in: Contesting benchmark machine learning datasets". In: *arXiv preprint arXiv:2007.07399* (2020).

[91] Jean-Antoine Désidéri. "Multiple-gradient descent algorithm (MGDA) for multiobjective optimization". In: *Comptes Rendus Mathematique* 350.5-6 (2012), pp. 313–318.

[92] Kate Donahue and Jon Kleinberg. "Fairness and utilization in allocating resources with uncertain demand". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 658–668.

[93] Ravit Dotan and Smitha Milli. "Value-laden disciplinary shifts in machine learning". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 294–294.

[94] Dheeru Dua and Casey Graff. *UCI machine learning repository*. 2017. URL: `http://archive.ics.uci.edu/ml`.

[95] Robin A Dubin. "Predicting house prices using multiple listings data". In: *The Journal of Real Estate Finance and Economics* 17.1 (1998), pp. 35–59.

[96] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness through awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. Cambridge, Massachusetts: ACM, 2012, pp. 214–226. ISBN: 978-1-4503-1115-1. DOI: `10.1145/2090236.2090255`.

[97] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. "Decoupled classifiers for group-fair and efficient machine learning". In: *Conference on Fairness, Accountability and Transparency*. 2018, pp. 119–133.

[98] Christopher D Elvidge, Kimberly Baugh, Mikhail Zhizhin, Feng Chi Hsu, and Tilottama Ghosh. "VIIRS night-time lights". In: *International Journal of Remote Sensing* 38.21 (2017), pp. 5860–5879.

[99] Hadi Elzayn, Shahin Jabbari, Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, and Zachary Schutzman. "Fair algorithms for learning in allocation problems". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019, pp. 170–179.

[100] Danielle Ensign, Sorelle A Friedler, Scott Nevlle, Carlos Scheidegger, and Suresh Venkatasubramanian. "Runaway feedback loops in predictive policing". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Vol. 81. 2018.

[101] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[102] Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. "Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems". In: *Journal of Machine Learning Research* 7.6 (2006).

[103] Marc Faddoul, Guillaume Chaslot, and Hany Farid. "A longitudinal analysis of YouTube's promotion of conspiracy videos". In: *arXiv preprint arXiv:2003.03318* (2020).

[104] Marc Fleurbaey and Francois Maniquet. "Optimal income taxation theory and principles of fairness". In: *Journal of Economic Literature* 56.3 (2018), pp. 1029–79.

[105] Massimo Florio. *Applied welfare economics: Cost-benefit analysis of projects and policies*. Routledge, 2014.

[106] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. "Datasheets for datasets". In: *Communications of the ACM* 64.12 (2021), pp. 86–92.

[107] Amirata Ghorbani and James Y. Zou. "Data Shapley: Equitable valuation of data for machine learning". In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. 2019, pp. 2242–2251. URL: `http://proceedings.mlr.press/v97/ghorbani19c.html`.

[108] Marina Gindelsky, Jeremy Moulton, and Scott A Wentland. "Valuing housing services in the era of big data: A user cost approach leveraging Zillow microdata". In: *Big Data for 21st Century Economic Statistics*. University of Chicago Press, 2019.

[109] Ezra Haber Glenn. *acs: Download, manipulate, and present American Community Survey and decennial data from the US Census*. 2019.

[110] Hila Gonen and Yoav Goldberg. "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)",* 2019.

[111] Google. *Google Static Maps API*. Accessed: 2018-03-18. 2018.

[112] Yves Grandvalet and Yoshua Bengio. "Semi-supervised learning by entropy minimization". In: *Advances in Neural Information Processing Systems*. 2005, pp. 529–536.

[113] Ben Green and Salomé Viljoen. "Algorithmic realism: Expanding the boundaries of algorithmic thought". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 19–31.

[114] James Griffin. *Well-being: Its meaning, measurement and moral importance*. Clarendon Press, 1986.

[115] Yating Gu, Yantian Wang, and Yansheng Li. "A survey on deep learning-driven remote sensing image scene understanding: Scene classification, scene retrieval and scene-guided object detection". In: *Applied Sciences* 9.10 (2019).

[116] Carlos Guestrin, Andreas Krause, and Ajit Paul Singh. "Near-optimal sensor placements in Gaussian processes". In: *International Conference on Machine Learning*. ACM. 2005, pp. 265–272. URL: `https://las.inf.ethz.ch/files/guestrin05near.pdf`.

[117] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[118] Barry Haack and Robert Ryerson. "Improving remote sensing research and education in developing countries: Approaches and recommendations". In: *International Journal of Applied Earth Observation and Geoinformation* 45 (Mar. 2016), pp. 77–83. URL: `https://www.sciencedirect.com/science/article/pii/S0303243415300477`.

[119] Ahsan Habib, Chandan Karmakar, and John Yearwood. "Impact of ECG dataset diversity on generalization of CNN model for detecting QRS complex". In: *IEEE Access* 7 (2019), pp. 93275–93285.

[120] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. "Learning from class-imbalanced data: Review of methods and applications". In: *Expert Systems with Applications* 73 (2017), pp. 220–239.

[121] David J Hand. *Measurement theory and practice*. Hoboken, NJ: Wiley-Blackwell, July 2004.

[122] David J Hand. "Statistics and the theory of measurement". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 159.3 (1996), pp. 445–473.

[123] M C Hansen, P V Potapov, R Moore, M Hancher, S A Turubanova, A Tyukavina, D Thau, S V Stehman, S J Goetz, T R Loveland, A Kommareddy, A Egorov, L Chini, C O Justice, and J R G Townshend. "High-resolution global maps of 21st-century forest cover change." In: *Science (New York, N.Y.)* 342.6160 (Nov. 2013), pp. 850–3. URL: http://www.ncbi.nlm.nih.gov/pubmed/24233722.

[124] Karen Hao. "We read the paper that forced Timnit Gebru out of Google. Here's what it says". In: *MIT Technology Review* (2020). URL: https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/.

[125] Moritz Hardt, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning". In: *Advances in Neural Information Processing Systems*. 2016, pp. 3315–3323.

[126] HarvardX. *HarvardX person-course academic year 2013 de-identified dataset, version 3.0*. Version V11. 2014. DOI: 10.7910/DVN/26147. URL: https://doi.org/10.7910/DVN/26147.

[127] Tatsunori Hashimoto. "Model performance scaling with multiple data sources". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 4107–4116.

[128] Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. "Fairness without demographics in repeated loss minimization". In: *Proceedings of the 35th International Conference on Machine Learning ICML*. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1934–1943. URL: http://proceedings.mlr.press/v80/hashimoto18a.html.

[129] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[130] Andrew Head, Mélanie Manguin, Nhat Tran, and Joshua E. Blumenstock. "Can human development be measured with satellite imagery?" In: ICTD '17. New York, NY, USA: Association for Computing Machinery, 2017. URL: https://doi.org/10.1145/3136560.3136576.

[131] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. "Women also snowboard: Overcoming bias in captioning models". In: *Proceedings of the European Conference on Computer Vision*. 2018, pp. 771–787.

[132] Gregory Hitz, Alkis Gotovos, Marie-Éve Garneau, Cédric Pradalier, Andreas Krause, Roland Y Siegwart, et al. "Fully autonomous focused exploration for robotic environmental monitoring". In: *International Conference on Robotics and Automation*. IEEE. 2014, pp. 2658–2664.

[133] Gabriel M Hoffmann and Claire J Tomlin. "Mobile sensor network control using mutual information methods and particle filters". In: *IEEE Transactions on Automatic Control* 55.1 (2009), pp. 32–47.

[134] Johannes Hofmanninger, Forian Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch, and Georg Langs. "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem". In: *European Radiology Experimental* 4.1 (2020), pp. 1–13.

[135] Danfeng Hong, Lianru Gao, Naoto Yokoya, Jing Yao, Jocelyn Chanussot, Qian Du, and Bing Zhang. "More diverse means better: Multimodal deep learning meets remote sensing imagery classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 0196.2892 (2020), pp. 1–15. ISSN: 23318422. DOI: 10.1109/tgrs.2020.3016820.

[136] Lily Hu and Yiling Chen. "Fair classification and social welfare". In: *ACM FAT** (2020).

[137] Lily Hu and Yiling Chen. "Welfare and distributional impacts of fair classification". In: *Fairness, Accountability, and Transparency in Machine Learning. Stockholm, Sweden.* (July 2018). arXiv: 1807.01134. URL: http://arxiv.org/abs/1807.01134.

[138] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. "Does distributionally robust supervised learning give robust classifiers?" In: *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2018, pp. 2034–2042. URL: http://proceedings.mlr.press/v80/hu18a.html.

[139] Sabine Van Huffel. *The total least squares problem: Computational aspects and analysis*. Society for Industrial and Applied Mathematics, 1991. ISBN: 0898712750.

[140] Jessica Hwang, Paulo Orenstein, Judah Cohen, Karl Pfeiffer, and Lester Mackey. "Improving subseasonal forecasting in the western US with machine learning". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 2325–2335.

[141] Jordi Inglada, Arthur Vincent, Marcela Arias, Benjamin Tardy, David Morin, and Isabel Rodes. "Operational high resolution land cover map production at the country scale using satellite image time series". In: *Remote Sensing* 9.1 (2017), p. 95.

[142] Vasileios Iosifidis and Eirini Ntoutsi. "Dealing with bias via data augmentation in supervised learning scenarios". In: *Proceedings of the Workshop on Bias in Information, Algorithms*. 2018, pp. 24–29.

[143] Abigail Z. Jacobs and Hanna Wallach. "Measurement and fairness". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 375–385. DOI: 10.1145/3442188.3445901.

[144] Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. "lil'ucb: An optimal exploration algorithm for multi-armed bandits". In: *Conference on Learning Theory*. PMLR. 2014, pp. 423–439.

[145] Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. "Combining satellite imagery and machine learning to predict poverty". In: *Science* 353.6301 (2016), pp. 790–794.

[146] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. "Tile2Vec: Unsupervised representation learning for spatially distributed data". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 3967–3974. DOI: 10.1609/aaai.v33i01.33013967.

[147] Neal Jean, Sang Michael Xie, and Stefano Ermon. "Semi-supervised deep kernel learning: Regression with unlabeled data by minimizing predictive variance". In: *Advances in Neural Information Processing Systems*. 2018, pp. 5322–5333.

[148] Yaochu Jin. *Multi-objective machine learning*. Vol. 16. Springer Science & Business Media, 2006.

[149] Yaochu Jin and Bernhard Sendhoff. "Pareto-based multiobjective machine learning: An overview and case studies". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38.3 (2008), pp. 397–415.

[150] Eun Seo Jo and Timnit Gebru. "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2020, pp. 306–316. DOI: 10.1145/3351095.3372829.

[151] Eric Jonas, Monica Bobra, Vaishaal Shankar, J. Todd Hoeksema, and Benjamin Recht. "Flare prediction using photospheric and coronal image data". In: *Solar Physics* 293.3 (2018), pp. 1–22. URL: http://dx.doi.org/10.1007/s11207-018-1258-9.

[152] Daniel Kahneman and Alan B Krueger. "Developments in the measurement of subjective well-being". In: *Journal of Economic perspectives* 20.1 (2006), pp. 3–24.

[153] Nathan Kallus, Xiaojie Mao, and Angela Zhou. "Assessing algorithmic fairness with unobserved protected class using data combination". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 110–110.

[154] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. "PAC subset selection in stochastic multi-armed bandits." In: *International Conference on Machine Learning*. Vol. 12. 2012, pp. 655–662.

[155] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. "End-to-end recovery of human shape and pose". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7122–7131.

[156] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. "Learning 3D human dynamics from video". In: *Computer Vision and Pattern Recognition*. 2019.

[157] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. "On the complexity of best-arm identification in multi-armed bandit models". In: *Journal of Machine Learning Research* 17.1 (2016), pp. 1–42. URL: http://www.jmlr.org/papers/volume17/kaufman16a/kaufman16a.pdf.

[158] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness". In: *International Conference on Machine Learning*. 2018, pp. 2564–2572.

[159] Reza Khodayi-mehr, Wilkins Aquino, and Michael M Zavlanos. "Model-based active source identification in complex environments". In: *IEEE Transactions on Robotics* (2019).

[160] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. "Avoiding discrimination through causal reasoning". In: *Advances in Neural Information Processing Systems* 30 (2017).

[161] Il Yong Kim and Oliver L de Weck. "Adaptive weighted-sum method for bi-objective optimization: Pareto front generation". In: *Structural and multidisciplinary optimization* 29.2 (2005), pp. 149–158.

[162] Michael P Kim, Amirata Ghorbani, and James Zou. "Multiaccuracy: Black-box post-processing for fairness in classification". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 247–254.

[163] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. "Human decisions and machine predictions". In: *The quarterly Journal of Economics* 133.1 (2018), pp. 237–293.

[164] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores". In: *Proceedings of the 8th Conference on Innovations in Theoretical Computer Science (ITCS)* (2017).

[165] Joshua Knowles. "ParEGO: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems". In: *IEEE Transactions on Evolutionary Computation* 10.1 (2006), pp. 50–66.

[166] Hidetoshi Komiya. "Elementary proof for Sion's minimax theorem". In: *Kodai Mathematical Journal* 11.1 (1988), pp. 5–7.

[167] Lukas Kondmann and Xiao Xiang Zhu. "Under the radar–Auditing fairness in ML for humanitarian mapping". In: *arXiv preprint arXiv:2108.02137* (2021).

[168] Tomer Koren, Roi Livni, and Yishay Mansour. "Multi-armed bandits with metric movement costs". In: *Conference on Neural Information Processing Systems*. 2017, pp. 4122–4131. URL: https://papers.nips.cc/paper/7000-multi-armed-bandits-with-metric-movement-costs.pdf.

[169] Alex Krizhevsky. *Learning multiple layers of features from tiny images*. Tech. rep. University of Toronto, 2009.

[170] Meghana Kshirsagar, Caleb Robinson, Siyu Yang, Shahrzad Gholami, Ivan Klyuzhin, Sumit Mukherjee, Md Nasir, Anthony Ortiz, Felipe Oviedo, Darren Tanner, et al. "Becoming good at AI for good". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 664–673.

[171] Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. "The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017, pp. 275–284.

[172] Alex Lamy, Ziyuan Zhong, Aditya K Menon, and Nakul Verma. "Noise-tolerant fair classification". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 294–305.

[173] Daniel Levine, Brandon Luders, and Jonathan How. "Information-rich path planning with general constraints using rapidly-exploring random trees". In: *AIAA Infotech@ Aerospace 2010*. 2010, p. 3360. URL: `http://acl.mit.edu/papers/Levine10_InfoTech.pdf`.

[174] Yu-Feng Li and Zhi-Hua Zhou. "Towards making unlabeled data never hurt". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.1 (2014), pp. 175–188.

[175] Ying Li, Haokui Zhang, Xizhe Xue, Yenan Jiang, and Qiang Shen. "Deep learning for remote sensing image classification: A survey". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.6 (2018), pp. 1–17.

[176] Zhan Wei Lim, David Hsu, and Wee Sun Lee. "Adaptive informative path planning in metric spaces". In: *The International Journal of Robotics Research* 35.5 (2016), pp. 585–598.

[177] Jay Littlepage. *DigitalGlobe moves to the cloud with AWS Snowmobile*. URL: `http://blog.digitalglobe.com/industry/digitalglobe-moves-to-the-cloud-with-aws-snowmobile/`.

[178] Chunming Liu, Xin Xu, and Dewen Hu. "Multiobjective reinforcement learning: A comprehensive overview". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45.3 (2014), pp. 385–398.

[179] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. "Delayed impact of fair machine learning". In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. Stockholm, Sweden, 2018, pp. 3156–3164.

[180] Lydia T. Liu, Max Simchowitz, and Moritz Hardt. "The implicit fairness criterion of unconstrained learning". In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, 2019, pp. 4051–4060.

[181] Lydia T. Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. "The disparate equilibria of algorithmic decision making when individuals invest rationally". In: *ACM FAT** (2020).

[182] Sharon L. Lohr. *Sampling: Design and Analysis*. Chapman and Hall/CRC, 2021. DOI: 10.1201/9780429298899.

[183] Ilya Loshchilov, Marc Schoenauer, and Michèle Sebag. "A mono surrogate for multiobjective optimization". In: *Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation*. ACM. 2010, pp. 471–478.

[184] George Y Lu and David W Wong. "An adaptive inverse-distance weighting spatial interpolation technique". In: *Computers & Geosciences* 34.9 (2008), pp. 1044–1055.

[185] Kristian Lum and William Isaac. "To predict and serve?" In: *Significance* 13.5 (2016), pp. 14–19.

[186] Dalton Lunga, Jonathan Gerrand, Lexie Yang, Christopher Layton, and Robert Stewart. "Apache Spark accelerated deep learning inference for large scale satellite image analytics". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13 (2020), pp. 271–283. DOI: 10.1109/JSTARS.2019.2959707.

[187] Yifei Ma, Roman Garnett, and Jeff Schneider. "Active search for sparse signals with region sensing". In: *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[188] Kolya Malkin, Caleb Robinson, Le Hou, Rachel Soobitsky, Jacob Czawlytko, Dimitris Samaras, Joel Saltz, Lucas Joppa, and Nebojsa Jojic. "Label super-resolution networks". In: *International Conference on Learning Representations*. 2018.

[189] Oscar Mañas, Alexandre Lacoste, Xavier Giro-i-Nieto, David Vazquez, and Pau Rodriguez. "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 9414–9423.

[190] Horia Mania, Aurelia Guy, and Benjamin Recht. "Simple random search of static linear policies is competitive for reinforcement learning". In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018. URL: https://proceedings.neurips.cc/paper/2018/file/7634ea65a4e6d9041cfd3f7de18e334a-Paper.pdf.

[191] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. "Recovering accurate 3D human pose in the wild using IMUs and a moving camera". In: *European Conference on Computer Vision*. 2018.

[192] Roman Marchant and Fabio Ramos. "Bayesian optimisation for informative continuous path planning". In: *International Conference on Robotics and Automation*. IEEE. 2014, pp. 6136–6143. URL: ieeexplore.ieee.org/document/6907763/.

[193] Marchant, Roman and Ramos, Fabio. "Bayesian Optimisation for Intelligent Environmental Monitoring". In: *International Conference on Intelligent Robots and Systems*. IEEE. 2012, pp. 2242–2249. URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6385653.

[194] Ruben Martinez-Cantin. "BayesOpt: a Bayesian optimization library for nonlinear optimization, experimental design and bandits." In: *Journal of Machine Learning Research* 15.1 (2014), pp. 3735–3739.

[195] Ruben Martinez-Cantin, Nando de Freitas, Eric Brochu, José Castellanos, and Arnaud Doucet. "A Bayesian exploration-exploitation approach for optimal online sensing and planning with a visually guided mobile robot". In: *Autonomous Robots* 27.2 (2009), pp. 93–103. URL: https://link.springer.com/article/10.1007/s10514-009-9130-2.

[196] Frank Mascarich, Taylor Wilson, Christos Papachristos, and Kostas Alexis. "Radiation source localization in GPS-denied environments using aerial robots". In: *International Conference on Robotics and Automation*. IEEE. 2018. URL: https://www.autonomousrobotslab.com/robotics-for-nuclear-sites.html.

[197] Alexey S Matveev, Michael C Hoy, and Andrey V Savkin. "Extremum seeking navigation without derivative estimation of a mobile robot in a dynamic environmental field". In: *Transactions on Control Systems Technology* 24.3 (2016), pp. 1084–1091. URL: http://ieeexplore.ieee.org/document/7243315/.

[198] Chiara Mellucci, Prathyush P Menon, Christopher Edwards, and Peter Challenor. "Source seeking using a single autonomous vehicle". In: *2016 American Control Conference*. IEEE. 2016, pp. 6441–6446.

[199] Lauren M Miller, Yonatan Silverman, Malcolm A MacIver, and Todd D Murphey. "Ergodic exploration of distributed information". In: *IEEE Transactions on Robotics* 32.1 (2016), pp. 36–52.

[200] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. "Distant supervision for relation extraction without labeled data". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pp. 1003–1011.

[201] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "Model cards for model reporting". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019, pp. 220–229.

[202] Jared Moore. "AI for not bad". In: *Frontiers in Big Data* (2019), p. 32.

[203] *More than a pretty picture: Using poverty maps to design better policies and interventions*. Washington, DC: The World Bank, 2007. ISBN: 0-8213-6932-6. DOI: 10.1596/978-0-8213-6931-9.

[204] Alyssa Morrow, Vaishaal Shankar, Devin Petersohn, Anthony Joseph, Benjamin Recht, and Nir Yosef. "Convolutional kitchen sinks for transcription factor binding site prediction". In: *arXiv preprint* (2017). URL: `http://arxiv.org/abs/1706.00125`.

[205] Jeremy Moulton and Scott Wentland. "Monetary policy and the housing market". In: *Annual Meeting of the American Economic Association*. Philadelphia, PA, 2018. URL: `https://www.aeaweb.org/conference/2018/preliminary/paper/HTnsAQrn`.

[206] Hussein Mouzannar, Mesrob I. Ohannessian, and Nathan Srebro. "From fair decision making to social equality". In: *ACM FAT** (2019).

[207] Elizbar A Nadaraya. "On estimating regression". In: *Theory of Probability & Its Applications* 9.1 (1964), pp. 141–142.

[208] Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn, and Wes B Ford. "The population biology of abalone (haliotis species) in Tasmania. I. Blacklip Abalone (h. rubra) from the north coast and islands of Bass Strait". In: *Sea Fisheries Division, Technical Report* 48 (1994), p411.

[209] Jerzy Neyman. "On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection". In: *Journal of the Royal Statistical Society* 97.4 (1934), pp. 558–625.

[210] Alejandro Noriega, Bernardo Garcia-Bulle, Luis Tejerina, and Alex Pentland. "Algorithmic fairness and efficiency in targeting social welfare programs at scale". In: *Bloomberg Data for Good Exchange Conference* (2018).

[211] Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[212] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019), pp. 447–453.

[213] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. "Realistic evaluation of deep semi-supervised learning algorithms". In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.

[214] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. "Social data: Biases, methodological pitfalls, and ethical boundaries". In: *Frontiers in Big Data* 2 (2019), p. 13.

[215] Mabel Ortega Adarme, Raul Queiroz Feitosa, Patrick Nigri Happ, Claudio Aparecido De Almeida, and Alessandra Rodrigues Gomes. "Evaluation of deep learning techniques for deforestation detection in the Brazilian Amazon and cerrado biomes from remote sensing imagery". In: *Remote Sensing* 12.6 (2020), p. 910.

[216] Chetan D Pahlajani, Jianxin Sun, Ioannis Poulakakis, and Herbert G Tanner. "Error probability bounds for nuclear detection: Improving accuracy through controlled mobility". In: *Automatica* 50.10 (2014), pp. 2470–2481.

[217] Sinno Jialin Pan and Qiang Yang. "A survey on transfer learning". In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), pp. 1345–1359. DOI: 10.1109/TKDE.2009.191.

[218] Vilfredo Pareto. *Manuale di economia politica*. Vol. 13. Societa Editrice, 1906.

[219] Biswajit Paria, Kirthevasan Kandasamy, and Barnabás Póczos. "A flexible multi-objective Bayesian optimization approach using random scalarizations". In: *Uncertainty in Artificial Intelligence* (2019).

[220] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. "Data and its (dis) contents: A survey of dataset development and use in machine learning research". In: *Patterns* 2.11 (2021).

[221] Sebastian Peitz and Michael Dellnitz. "Gradient-based multiobjective optimization with uncertainties". In: *NEO 2016*. Springer, 2018, pp. 159–182.

[222] Otávio A B Penatti, Keiller Nogueira, Jefersson A Dos Santos, and Jefersson A Dos Santos. "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2015), pp. 44–51.

[223] Nathaniel Persily. "The 2016 US election: Can democracy survive the internet?" In: *Journal of Ddemocracy* 28.2 (2017), pp. 63–76.

[224] A. C. Pigou. *The economics of welfare*. English. Macmillan London, 1920.

[225] B. Porat and A. Nehorai. "Localizing vapor-emitting sources by moving sensors". In: *IEEE Transactions on Signal Processing* 44.4 (1996), pp. 1018–1021. DOI: 10.1109/78.492560.

[226] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. "Learning object class detectors from weakly annotated video". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2012, pp. 3282–3289.

[227] Friedrich Pukelsheim. *Optimal design of experiments*. Classic ed. Classics in applied mathematics; 50. Society for Industrial and Applied Mathematics, 2006.

[228] Morgan Quigley, Ken Conley, Brian P. Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y. Ng. "ROS: an Open-source robot operating system". In: *ICRA Workshop on Open Source Software*. 2009. URL: http://www.willowgarage.com/sites/default/files/icraoss09-ROS.pdf.

[229] Joanna Radin. ""Digital natives": How medical and indigenous histories matter for big data". In: *Osiris* 32.1 (2017), pp. 43–64.

[230] Ali Rahimi and Benjamin Recht. "Uniform approximation of functions with random bases". In: *46th Annual Allerton Conference on Communication, Control, and Computing*. IEEE, Sept. 2008, pp. 555–561. DOI: 10.1109/ALLERTON.2008.4797607. URL: http://ieeexplore.ieee.org/document/4797607/.

[231] Ali Rahimi and Benjamin Recht. "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning". In: *Neural Information Processing Systems*. Vol. 1. 1. Vancouver, B.C., 2008, pp. 1313–1320.

[232] Ali Rahimi and Benjamin Recht. "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning". In: *Advances in Neural Information Processing Systems*. 2009, pp. 1313–1320.

[233] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. "AI and the everything in the whole wide world benchmark". In: *35th Conference on Neural Information Processing Systems Track on Datasets and Benchmarks*. 2021.

[234] Inioluwa Deborah Raji and Joy Buolamwini. "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 429–435.

[235] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J. Topol. "AI in health and medicine". In: *Nature Medicine* 28.1 (Jan. 2022), pp. 31–38. DOI: 10.1038/s41591-021-01614-0. URL: https://doi.org/10.1038/s41591-021-01614-0.

[236] Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. "Data programming: Creating large training sets, quickly". In: *Advances in Neural Information Processing Systems*. 2016, pp. 3567–3575.

[237] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. "Do ImageNet classifiers generalize to ImageNet?" In: *International Conference on Machine Learning*. 2019, pp. 5389–5400.

[238] Fennis Reed, Andrea Gaughan, Forrest Stevens, Greg Yetman, Alessandro Sorichetta, and Andrew Tatem. "Gridded population maps informed by different built settlement products". In: *Data* 3.3 (Sept. 2018), p. 33. DOI: 10.3390/data3030033. URL: http://www.mdpi.com/2306-5729/3/3/33.

[239] Branko Ristic, Mark Morelande, and Ajith Gunatilaka. "Information driven search for point sources of gamma radiation". In: *Signal Processing* 90.4 (2010), pp. 1225–1239.

[240] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, et al. "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans". In: *Nature Machine Intelligence* 3.3 (2021), pp. 199–217.

[241] Caleb Robinson, Fred Hohman, and Bistra Dilkina. "A deep learning approach for population estimation from satellite imagery". In: *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities - GeoHumanities 2017*. New York, New York, USA: ACM Press, 2017, pp. 47–54. URL: http://dl.acm.org/citation.cfm?doid=3149858.3149863.

[242] Diederik M Roijers and Shimon Whiteson. "Multi-objective decision making". In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 11.1 (2017), pp. 1–129.

[243] Esther Rolf, David Fridovich-Keil, Max Simchowitz, Benjamin Recht, and Claire Tomlin. "A successive-elimination approach to adaptive robotic source seeking". In: *IEEE Transactions on Robotics* 37.1 (2021), pp. 34–47. DOI: 10.1109/TRO.2020.3005537.

[244] Esther Rolf, Michael I. Jordan, and Benjamin Recht. "Post-estimation smoothing: A simple baseline for learning with side information". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 1759–1769. URL: https://proceedings.mlr.press/v108/rolf20a.html.

[245] Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. "A generalizable and accessible approach to machine learning with global satellite imagery". In: *Nature Communications* 12.1 (2021), pp. 1–11.

[246] Esther Rolf, Max Simchowitz, Sarah Dean, Lydia T Liu, Daniel Bjorkegren, Moritz Hardt, and Joshua Blumenstock. "Balancing competing objectives with noisy data: Score-based classifiers for welfare-aware machine learning". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 8158–8168.

[247] Esther Rolf, Theodora T Worledge, Benjamin Recht, and Michael Jordan. "Representation matters: Assessing the importance of subgroup allocations in training data". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 9040–9051.

[248] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Sasha Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. "Tackling climate change with machine learning". In: *ACM Comput. Surv.* 55.2 (2022). ISSN: 0360-0300. URL: https://doi.org/10.1145/3485128.

[249] Adriana Romero, Carlo Gatta, and Gustau Camps-Valls. "Unsupervised deep feature extraction for remote sensing image classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 54.3 (2016), pp. 1349–1362.

[250] Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. "Inclusivefacenet: Improving face attribute detection with race and gender diversity". In: *arXiv:1712.00193* (2017).

[251] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. "Distributionally robust neural networks". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL: https://openreview.net/forum?id=ryxGuJrFvS.

[252] Paul A. Samuelson. *Foundations of economic analysis*. Harvard University Press, 1947.

[253] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. "How computers see gender: An evaluation of gender classification in commercial facial analysis services". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–33.

[254] H Schneeweiss. "Consistent estimation of a regression with errors in the variables". In: *Metrika* 23.1 (1976), pp. 101–115.

[255] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. "Green AI". In: *Communications of the ACM* 63.12 (2020), pp. 54–63.

[256] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. "Fairness and abstraction in sociotechnical systems". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019, pp. 59–68.

[257] Amartya Sen. "Behaviour and the concept of preference". In: *Economica* 40.159 (1973), pp. 241–259.

[258] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. "No classification without representation: Assessing geodiversity issues in open data sets for the developing world". In: *arXiv preprint arXiv:1711.08536* (2017).

[259] Shubhanshu Shekhar, Greg Fields, Mohammad Ghavamzadeh, and Tara Javidi. "Adaptive sampling for minimax fair classification". In: *Advances in Neural Information Processing Systems* 34 (2021).

[260] Max Simchowitz, Kevin Jamieson, and Benjamin Recht. "Best-of-K bandits". In: *Conference on Learning Theory*. 2016, pp. 1440–1489.

[261] Jeffrey S. Simonoff. *Smoothing methods in statistics (Springer Series in Statistics)*. Springer, 1998. ISBN: 0387947167.

[262] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *International Conference on Learning Representations*. 2015. URL: `http://arxiv.org/abs/1409.1556`.

[263] Aarti Singh, Robert Nowak, and Xiaojin Zhu. "Unlabeled data: Now it helps, now it doesn't". In: *Advances in Neural Information Processing Systems*. 2009, pp. 1513–1520.

[264] Paige Marta Skiba and Jeremy Tobacman. "Do payday loans cause bankruptcy?" In: *The Journal of Law and Economics* 62.3 (2019), pp. 485–519.

[265] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. "No subclass left behind: Fine-grained robustness in coarse-grained classification problems". In: *Advances in Neural Information Processing Systems* 33 (2020).

[266] Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. "Information-theoretic regret bounds for Gaussian process optimization in the bandit setting". In: *IEEE Transactions on Information Theory* 58.5 (2012), pp. 3250–3265. DOI: 10.1109/TIT.2011.2182033.

[267] Joseph Stiglitz, Amartya Sen, and Jean-Paul Fitoussi. "The measurement of economic performance and social progress revisited". In: *Reflections and Overview. Commission on the Measurement of Economic Performance and Social Progress, Paris* (2009).

[268] Emma Strubell, Ananya Ganesh, and Andrew McCallum. "Energy and policy considerations for deep learning in NLP". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 3645–3650.

[269] Harini Suresh and John Guttag. "A framework for understanding sources of harm throughout the machine learning life cycle". In: *Equity and Access in Algorithms, Mechanisms, and Optimization*. 2021, pp. 1–9.

[270] Latanya Sweeney. "Discrimination in online ad delivery". In: *Communications of the ACM* 56.5 (2013), pp. 44–54.

[271] Yi Chern Tan and L Elisa Celis. "Assessing social and intersectional biases in contextualized word representations". In: *Advances in Neural Information Processing Systems* 32 (2019).

[272] Jean Tarbouriech and Alessandro Lazaric. "Active exploration in Markov decision processes". In: *International Conference on Artificial Intelligence and Statistics*. 2019.

[273] Antti Tarvainen and Harri Valpola. "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results". In: *Advances in Neural Information Processing Systems*. 2017, pp. 1195–1204.

[274] Grigorios Tsagkatakis, Anastasia Aidini, Konstantina Fotiadou, Michalis Giannopoulos, Anastasia Pentari, and Panagiotis Tsakalides. "Survey of deep-learning approaches for remote sensing observation enhancement". In: *Sensors (Switzerland)* 19.18 (2019), pp. 1–39. DOI: 10.3390/s19183929.

[275] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions". In: *Scientific Data* 5.1 (2018). URL: https://doi.org/10.1038/sdata.2018.161.

[276] Alexandre B. Tsybakov. *Introduction to nonparametric estimation (Springer Series in Statistics)*. Springer, 2010. ISBN: 9781441927095.

[277] Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R. Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W. Mathis, Frank van Langevelde, Tilo Burghardt, Roland Kays, Holger Klinck, Martin Wikelski, Iain D. Couzin, Grant van Horn, Margaret C. Crofoot, Charles V. Stewart, and Tanya Berger-Wolf. "Perspectives in machine learning for wildlife conservation". In: *Nature Communications* 13.1 (Feb. 2022). DOI: 10.1038/s41467-022-27980-y. URL: https://doi.org/10.1038/s41467-022-27980-y.

[278] U.S. Census Bureau. *2015 American Community Survey 5-Year estimates, table B19013*. URL: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_15_5YR_B19013&prodType=table.

[279] U.S. Census Bureau. *Budget estimates, fiscal year 2021*. 2021. URL: https://www2.census.gov/about/budget/census-fiscal-year-21-presidents-budget.pdf.

[280] Union of Concerned Scientists. *UCS satellite database*. Jan. 2019. URL: https://www.ucsusa.org/nuclear-weapons/space-weapons/satellite-database.

[281] Union of Concerned Scientists. *Underwater: Rising seas, chronic floods, and the implications for US coastal real estate*. Tech. rep. Union of Concerned Scientists, 2018. URL: http://www.zillow.com/ztrax..

[282] US Federal Reserve. *Report to the congress on credit scoring and its effects on the availability and affordability of credit*. 2007.

[283] Kristof Van Moffaert and Ann Nowé. "Multi-objective reinforcement learning using sets of pareto dominating policies". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 3483–3512.

[284] Luma K. Vasiljevic and Hassan K. Khalil. "Error bounds in differentiation of noisy signals by high-gain observers". In: *Systems & Control Letters* 57.10 (2008), pp. 856–862.

[285] Kai Vetter, Ross Barnowksi, Andrew Haefner, Tenzing HY Joshi, Ryan Pavlovsky, and Brian J. Quiter. "Gamma-ray imaging for nuclear security and safety: Towards 3-D gamma-ray vision". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 878 (2018), pp. 159–168. URL: https://www.sciencedirect.com/science/article/pii/S0168900217309269.

[286] Kailas Vodrahalli, Ke Li, and Jitendra Malik. "Are all training examples created equal? An empirical study". In: *arXiv preprint arXiv:1811.12569* (2018).

[287] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2019.

[288] Mengting Wan and Julian J. McAuley. "Item recommendation on monotonic behavior chains". In: *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*. ACM, 2018, pp. 86–94. DOI: 10.1145/3240323.3240369. URL: https://doi.org/10.1145/3240323.3240369.

[289] Matt P Wand and M Chris Jones. *Kernel smoothing*. Chapman and Hall/CRC, 1994.

[290] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 5310–5319.

[291] Geoffrey S Watson. "Smooth regression analysis". In: *Sankhyā: The Indian Journal of Statistics, Series A* (1964), pp. 359–372.

[292] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1 (2016), pp. 1–9.

[293] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*. Vol. 2. 3. MIT Press Cambridge, MA, 2006.

[294] Tommy Wright. "A general exact optimal sample allocation algorithm: With bounded cost and bounded sample sizes". In: *Statistics & Probability Letters* (2020), p. 108829.

[295] Laure Wynants, Ben Van Calster, Gary S Collins, Richard D Riley, Georg Heinze, Ewoud Schuit, Marc MJ Bonten, Darren L Dahly, Johanna A Damen, Thomas PA Debray, et al. "Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal". In: *BMJ* 369 (2020).

[296] Michael Xie, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. "Transfer learning from deep features for remote sensing and poverty mapping". In: *Thirtieth AAAI Conference on Artificial Intelligence*. 2016.

[297] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. "Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 547–558.

[298] Gal Yona, Amirata Ghorbani, and James Zou. "Who's responsible? Jointly quantifying the contribution of the learning algorithm and data". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 1034–1041.

[299] Le Yu, Lu Liang, Jie Wang, Yuanyuan Zhao, Qu Cheng, Luanyun Hu, Shuang Liu, Liang Yu, Xiaoyi Wang, Peng Zhu, Xueyan Li, Yue Xu, Congcong Li, Wei Fu, Xuecao Li, Wenyu Li, Caixia Liu, Na Cong, Han Zhang, Fangdi Sun, Xinfang Bi, Qinchuan Xin, Dandan Li, Donghui Yan, Zhiliang Zhu, Michael F. Goodchild, and Peng Gong. "Meta-discoveries from a synthesis of satellite-based land-cover mapping research". In: *International Journal of Remote Sensing* 35.13 (July 2014), pp. 4573–4588.

[300] Kaan Yucer, Oliver Wang, Alexander Sorkine-Hornung, and Olga Sorkine-Hornung. "Reconstruction of articulated objects from a moving camera". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2015, pp. 28–36.

[301] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. "Learning fair representations". In: *International Conference on Machine Learning*. PMLR. 2013, pp. 325–333.

[302] Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. "Hurtful words: Quantifying biases in clinical contextual word embeddings". In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. 2020, pp. 110–120.

[303] Jason Y. Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. "Predicting 3D human dynamics from video". In: *International Conference on Computer Vision*. 2019.

[304] Miao Zhang, Harvineet Singh, Lazarus Chok, and Rumi Chunara. "Segmenting across places: The need for fair transfer learning with satellite imagery". In: *arXiv preprint arXiv:2204.04358* (2022).

[305] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. "From actemes to action: A strongly-supervised representation for detailed action understanding". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 2248–2255.

[306] Han Zhao and Geoff Gordon. "Inherent tradeoffs in learning fair representations". In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.

[307] Dengyong Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. "Learning with local and global consistency". In: *Advances in Neural Information Processing Systems*. 2004, pp. 321–328.

[308] Dengyong Zhou, Jason Weston, Arthur Gretton, Olivier Bousquet, and Bernhard Schölkopf. "Ranking on data manifolds". In: *Advances in Neural Information Processing Systems*. 2004, pp. 169–176.

[309] Xiaojin Zhu. *Semi-supervised learning literature survey*. Tech. rep. 1530. Computer Sciences, University of Wisconsin-Madison, 2005.

[310] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. "Semi-supervised learning using gaussian fields and harmonic functions". In: *Proceedings of the 20th International Conference on Machine Learning*. 2003, pp. 912–919.

[311] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. "Deep feature flow for video recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 2349–2358.

[312] Zillow. *ZTRAX: Zillow Transaction and Assessor Dataset*. 2018. URL: `http://www.zillow.com/ztrax/`.

[313] Zillow Research. *zillow-research/ztrax*. URL: `https://github.com/zillow-research/ztrax`.

[314] Indre Zliobaite. "On the relation between accuracy and fairness in binary classification". In: *arXiv preprint arXiv:1505.05723* (2015).