University of California
Santa Barbara

Noise-adding and beyond: A study in data-adaptive methods for differential privacy

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Computer Science

by

Rachel Redberg

Committee in charge:

Professor Yu-Xiang Wang, Chair
Professor Xifeng Yan
Professor Alexander Franks

December 2023

The Dissertation of Rachel Redberg is approved.

Professor Xifeng Yan

Professor Alexander Franks

Professor Yu-Xiang Wang, Committee Chair

December 2023

Noise-adding and beyond: A study in data-adaptive methods for differential privacy

by

Rachel Redberg

Acknowledgements

Finishing a PhD has been a massive undertaking and I am enormously grateful to have had such strong support from many different corners. First and foremost I would like to thank my advisor Yu-Xiang Wang, whose belief in me has meant so much.

I am also incredibly fortunate to have benefited from the friendship and mentorship of Sourav Medya and Arlei Silva, who helped me steer out of some very tight spots and who got me up to speed on many of the unspoken rules in academia. The last person I'll mention in my (not exclusive) list of notable mentors is Amr El Abbadi, whose passion for teaching greatly inspired me first as his student and then as his TA for distributed systems.

During my time at UCSB, I was able to learn from amazing faculty both in and out of the classroom. I'd like to thank Daniel Lokshtanov for ski lessons and bike-riding pointers, and Yekaterina Kharitonova for two very memorable quarters as her TA. Thanks also to Elizabeth Belding, who offered me support at a pivotal crossroads. Last but not least, my heartfelt thanks to Amr El Abbadi and Divy Agrawal for letting me unabashedly use their lab space after campus re-opened. My time working in the distributed systems lab was arguably the most productive era of my PhD!

The computer science department at UCSB also has incredible staff. I am deeply grateful to Karen Van Gool, who in one of our first encounters commandeered the department chair's office so that we could have a private conversation. Tim Robinson was there from the start of my PhD and hugely helpful in managing all things IGERT. I'd also like to thank Maritza Fuljencio and Bella Cardoso for coming to my defense even though I know they are insanely busy! And lastly, I'd like to thank Greta Carl-Halle because without her, the department would fall apart.

I am fortunate to have collaborated with Yuqing Zhu and Antti Koskela on research

projects that have been very meaningful to me. I'd also like to thank my other collaborators: Yingyu Lin, Yian Ma, Fuheng Zhao and Dan Qiao. Many thanks also to Xifeng Yan and Alexander Franks for serving on my committee and for helping me understand the broader impact of my research.

Lastly, I can't imagine how I would have gotten through a PhD without the support of friends, family and labmates. I am using "labmates" as a generous term that encompasses labs I used to belong to (Dynamo Lab); labs that I currently belong to (S2ML lab); and labs that I never actually belonged to (Distributed Systems Lab and Theory Lab). I am deeply grateful to everyone who came along on this journey with me.

Curriculum Vitæ

Rachel Redberg

Education

2023	Ph.D. in Computer Science, University of California, Santa Barbara.
2022	M.S. in Computer Science, University of California, Santa Barbara.
2015	B.A. in Applied Mathematics, University of California, Berkeley.

Publications

Lin, Y., Ma, Y., Wang, Y. X., & Redberg, R. (2023). Tractable MCMC for Private Learning with Pure and Gaussian Differential Privacy. arXiv preprint arXiv:2310.14661.

Redberg, R. E., Koskela, A., & Wang, Y. X. (2023, November). Improving the privacy and practicality of objective perturbation for differentially private linear learners. In Thirty-seventh Conference on Neural Information Processing Systems.

Redberg, R., Zhu, Y., & Wang, Y. X. (2023, April). Generalized PTR: User-Friendly Recipes for Data-Adaptive Algorithms with Differential Privacy. In International Conference on Artificial Intelligence and Statistics (pp. 3977-4005). PMLR.

Zhao, F., Qiao, D., Redberg, R., Agrawal, D., El Abbadi, A., & Wang, Y. X. (2022). Differentially private linear sketches: Efficient implementations and applications. Advances in Neural Information Processing Systems, 35, 12691-12704.

Redberg, R., & Wang, Y. X. (2021). Privately publishable per-instance privacy. Advances in Neural Information Processing Systems, 34, 17335-17346.

Ye, W., Wang, Z., Redberg, R., & Singh, A. (2019). Tree++: Truncated tree based graph kernels. IEEE Transactions on Knowledge and Data Engineering, 33(4), 1778-1789.

Abstract

Noise-adding and beyond: A study in data-adaptive methods for differential privacy

by

Rachel Redberg

As technology has evolved, so too have our privacy needs. AI is now a household name and machine learning (ML) applications a part of daily life. But an ML model is only as good as the data on which it's trained — and what happens when that data needs to be protected?

Differential privacy (DP) is a rigorous mathematical definition which can be used to provably bound the privacy leakage (or loss) of running a machine learning algorithm. This relatively young field of study has started to gain considerable traction in the ML research community. There is, however, a narrowing but still precipitous gap between theory and practice which has prevented DP from seeing widespread deployment in the real world. This dissertation proposes several tools and algorithms to bridge this gap.

In Chapter 2, we parameterize the privacy loss as a function of the data and investigate how to privately publish these data-dependent DP losses for the objective perturbation mechanism. These data-dependent DP losses might be significantly smaller than the worst-case DP bound, thus serving as justification for using a looser privacy guarantee — hence achieving better utility — in practice. Chapter 3 then demonstrates how data-dependent DP losses can be used in order to develop DP algorithms which can adapt to favorable properties of the data, in order to achieve a better privacy-utility trade-off. Chapter 4 returns to objective perturbation and provides this time-honored DP mechanism with new tools and privacy analyses that allow it to compete with more modern algorithms.

Contents

Chapter 1

Introduction

> "That the individual shall have full protection in person and in property is a principle as old as the common law; but it has been found necessary from time to time to define anew the exact nature and extent of such protection."
>
> — Samuel D. Warren II and Louis Brandeis, *The Right to Privacy, 1890*

In 2006 Netflix staged an open competition to improve its recommendation algorithm for films. Thousands of teams from over 100 countries competed for the $1 million prize (Thompson, 2008). Each competing team had access to a dataset containing movie ratings created by almost 500,000 Netflix subscribers between December 1999 and December 2005 (Lohr, 2009). The dataset had been anonymized by removing all customers' identifying information; only ratings and dates remained.

Netflix announced a winner in 2009, and plans for a second contest were already underway. This second contest never saw the light of day (Singel, 2010). The culprit was privacy concerns: In a 2008 paper, researchers at University of Texas were able to identify users from the "anonymous" Netflix prize dataset by linking it with an auxiliary IMDB dataset (Narayanan and Shmatikov, 2008).

Stories like these demonstrate that by the first years of the millenium, there was a clear need for rigorous and provable privacy protection for data. But what properties should such a framework have?

1.1 Differential Privacy

Suppose that we can query a dataset D of manatee preferences. We ask, "How many users in D like manatees?" and then we ask, "How many users in D like manatees, who are not Rachel?" If an adversary knows that I do, in fact, like manatees, then a differencing attack like this could reveal whether or not my data is in dataset D.

From this example, we might desire that our privacy definition is robust against auxiliary information (such as the fact that I like manatees) and also that our privacy definition is agnostic to the differences between nearly identical datasets (such as datasets which differ only in the presence or absence of a Rachel).

Our goal is to learn something meaningful about the data *in aggregate* without leaking information about any particular individual in the dataset. One way to view this is that the outcome of analyzing a dataset should be roughly the same regardless of whether an individual contributed her data or not. Now, let's say that our study of dataset D has taught us that swimming increases the odds of liking manatees. Should swimmers be wary of having contributed their data to D? Maybe, but there's nothing we can (or rather, should) do about it. A privacy definition that prevents us from drawing meaningful conclusions from data is not particularly useful!

These guiding principles motivated the development of *differential privacy* (Dwork et al., 2006), now the gold standard for data privacy protection. Differential privacy is defined as a bound over *neighboring datasets* D and D' which are identical up to a single datapoint, e.g. one could obtain D' by adding or removing a datapoint from D.

Definition 1.1.1 (Differential privacy)**.** A mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ satisfies (ϵ, δ)-differential privacy if for all neighboring datasets $D, D' \in \mathcal{D}$ and output sets $S \subseteq \mathcal{R}$,

$$\Pr\left[\mathcal{M}(D) \in S\right] \leq e^{\epsilon} \Pr\left[\mathcal{M}(D') \in S\right] + \delta.$$

Suppose that we apply a randomized algorithm $\mathcal{M}$ to a dataset D and to one of its neighbors D', which differs from D by only one datapoint. Differential privacy guarantees that an adversary will have difficulty distinguishing whether an output was produced from running $\mathcal{M}(D)$ or running $\mathcal{M}(D')$ — in other words, an adversary can't tell from the output whether or not an individual datapoint was present in the dataset. The degree of difficulty is determined by the parameter ϵ, which effectively bounds the log probability density ratio between the output distributions of $\mathcal{M}(D)$ and $\mathcal{M}(D')$. The parameter δ allows for some slack in this bound.

1.2 Dissertation Overview

This dissertation is organized around three papers which share common goals and themes. The first of these is developing data-adaptive tools and algorithms for differential privacy (Chapters 2 and 3), which can provide a more meaningful privacy-utility trade-off for "typical" data compared to the worst-case data hypothesized by DP. Another main theme (which we will see in Chapters 2 and 4) is moving beyond the common DP paradigm of calibrating noise to the *sensitivity* of a function, i.e. the maximum change in a function between any two neighboring datasets.

1.2.1 Privately Publishable Per-instance Privacy

In Chapter 2 we consider how to privately share the personalized privacy losses incurred by the objective perturbation mechanism, using per-instance differential privacy (pDP). Differential privacy gives us a worst-case bound that might be orders of magnitude larger than the privacy loss to a particular individual relative to a fixed dataset. The pDP framework provides a more fine-grained analysis of the privacy guarantee to a target individual, but the per-instance privacy loss itself might be a function of sensitive data. In

this chapter, we will analyze the per-instance privacy loss of releasing a private empirical risk minimizer learned via objective perturbation, and propose a group of methods to privately and accurately publish the pDP losses at little to no additional privacy cost.

1.2.2 Generalized Propose-Test-Release

The "Propose-Test-Release" (PTR) framework [Dwork and Lei, 2009] is a classic recipe for designing differentially private (DP) algorithms that are data-adaptive, i.e. those that add less noise when the input dataset is "nice". In Chapter 3 we extend PTR to a more general setting by privately testing data-dependent privacy losses rather than local sensitivity, hence making it applicable beyond the standard noise-adding mechanisms, e.g. to queries with unbounded or undefined sensitivity. We demonstrate the versatility of generalized PTR using private linear regression as a case study. Additionally, we apply our algorithm to solve an open problem from "Private Aggregation of Teacher Ensembles (PATE)" [Papernot et al., 2017, 2018] — privately releasing the entire model with a delicate data-dependent analysis.

1.2.3 Improving Objective Perturbation

In the arena of privacy-preserving machine learning, differentially private stochastic gradient descent (DP-SGD) has outstripped the objective perturbation mechanism in popularity and interest. Though unrivaled in versatility, DP-SGD requires a non-trivial privacy overhead (for privately tuning the model's hyperparameters) and a computational complexity which might be extravagant for simple models such as linear and logistic regression. Chapter 4 revamps the objective perturbation mechanism with tighter privacy analyses and new computational tools that boost it to perform competitively with DP-SGD on unconstrained convex generalized linear problems.

Chapter 2

Privately Publishable Per-instance Privacy

2.1 Introduction

An explosion of data has fueled innovation in machine learning applications and demanded, in equal turn, privacy protection for the sensitive data with which machine learning practitioners train and evaluate models.

Differential privacy (DP) (Dwork et al., 2006, 2014a) has become a mainstay of privacy-preserving data analysis, replacing less robust privacy definitions such as k-anonymity which fail to protect against sufficiently powerful de-anonymization attacks (Narayanan and Shmatikov, 2008). In contrast, DP offers provable privacy guarantees that are robust against an arbitrarily strong adversary.

The data curator could trivially protect against privacy loss by reporting a constant function, or by releasing only data-independent noise. The key challenge of DP is to release privatized output that retains utility to the data analyst.

A desired level of utility in a machine learning application might necessitate a high

value of ϵ, but the privacy guarantees degrade quickly past $\epsilon = 1$. Triastcyn and Faltings (2020) construct an example whereby a differentially private algorithm with $\epsilon = 2$ allows an attacker to use a maximum-likelihood estimate to conclude with up to 88% accuracy that an individual is in a dataset. For $\epsilon = 5$, the theoretical upper bound on the accuracy of an optimal attack is 99.3%.

Moreover, practical applications of differential privacy commonly use large values of ϵ. A study of Apple's deployment of differential privacy (Tang et al., 2017) revealed that the overall daily privacy loss permitted by the system was as high as $\epsilon = 6$ for Mac OS 10.12.3 and $\epsilon = 14$ for iOS 10.1.1 – offering only scant privacy protection!

Recent work (Yu et al., 2021) has empirically justified large privacy parameters by conducting membership inference attacks to demonstrate that these seemingly tenuous privacy guarantees are actually much stronger in practice. These results are unsurprising from the perspective that DP gives a data-independent bound on the worst-case privacy loss which is likely to be a conservative estimate of the risk to a particular individual when a DP algorithm is applied to a particular input dataset.

Per-instance differential privacy provides a theoretically sound alternative to the empirical approach for revealing the gap between the worst-case DP bound and the actual privacy loss in practice. The privacy loss to a particular individual relative to a fixed dataset might be orders of magnitude smaller than the worst-case bound guaranteed by standard DP. In this case, an algorithm meeting a desired level of utility but providing weak DP guarantees may, for the same level of utility, achieve drastically more favorable *per-instance* DP guarantees.

The remaining challenge is that the per-instance privacy loss is a function of the entire dataset; publishing it directly would negate the purpose of privately training a model in the first place! In this chapter, we propose a methodology to privately release the per-instance privacy losses associated with private empirical risk minimization. Our

contributions are as follows:

- We introduce *ex-post* per-instance differential privacy to provide a sharp characterization of the privacy loss to a particular individual that adapts to both the input dataset and the algorithm's output.
- We present a novel analysis of the *ex-post* per-instance privacy losses incurred by the objective perturbation mechanism, demonstrating that these *ex-post* pDP losses are orders of magnitude smaller than the worst-case guarantee of differential privacy.
- We propose a group of methods to privately and accurately release the *ex-post* pDP losses. In the particular case of generalized linear models, we show that we can accurately publish the private *ex-post* pDP losses using a dimension- and dataset-independent bound.
- One technical result of independent interest is a new DP mechanism that releases the Hessian matrix by adding a *Gaussian Orthogonal Ensemble* matrix, which improves the classical "AnalyzeGauss" (Dwork et al., 2014b) by roughly a constant factor of two.

2.1.1 Related Work

This chapter builds upon Wang (2019), which proposed the per-instance DP framework and left as an open question the matter of publishing the pDP losses. We extend the pDP framework to an *ex-post* setting to provide privacy guarantees that adapt even more fluidly to data-dependent properties of our algorithms. Another fundamental ingredient in our privacy analysis is the objective perturbation algorithm (`Obj-Pert`) of Chaudhuri et al. (2011), further analyzed by Kifer et al. (2012), which privately releases the minimizer of an empirical risk by adding a linear perturbation to the objective function before optimizing.

Per-instance DP and *ex-post* per-instance DP belong to a growing family of DP definitions that provide a more fine-grained characterization of the privacy loss. Among these are data-dependent DP (Papernot et al., 2018b), which conditions on a fixed dataset; personalized DP (Ghosh and Roth, 2011; Ebadi et al., 2015; Liu et al., 2015), which conditions on a fixed individual's datapoint; and *ex-post* DP (Ligett et al., 2017), which conditions on the realized output of the algorithm. Per-instance DP conditions on both a fixed dataset and a fixed individual's datapoint, and *ex-post* per-instance DP adapts even further to the realized output of the algorithm. A more detailed comparison of these DP variants is included in the supplementary materials.

Other data-adaptive methodologies include propose-test-release (Dwork and Lei, 2009) and local sensitivity (Nissim et al., 2007). In addition, Bayesian differential privacy (Triastcyn and Faltings, 2020) provides data-dependent privacy guarantees that afford strong protection to "typical" data by making distributional assumptions about the sensitive data. The Rényi-DP-based privacy filters of Feldman and Zrnic (2020) are also closely related to our work; the authors study composition of personalized (but not per-instance) privacy losses using adaptively-chosen privacy parameters.

2.2 Preliminaries

2.2.1 Symbols and Notation

We write the output of a randomized algorithm $\mathcal{A}$ as $\mathcal{A}(\cdot)$, and for continuous distributions we take $\Pr[\mathcal{A}(D) = o]$ to be the value of the probability density function at output o.

We will let $z \in \mathcal{Z}$ refer to both an individual and their data; for example, individual z holds data $z = (x, y)$ in a supervised learning problem. We take $\mathcal{Z}^* = \cup_{n=0}^{\infty} \mathcal{Z}^n$ to be the

space of datasets with an unspecified number of data points. $D_{\pm z} \in \mathcal{Z}^*$ denotes the fixed dataset $D = \{z_1, \ldots, z_n\} \in \mathcal{Z}^*$ with the data point z removed from D if $z \in D$, or added to D if $z \notin D$. In our mathematical expressions, we use "$\pm$" to mean "add if $z \notin D$, subtract otherwise". Similarly. "$\mp$" means "subtract if $z \notin D$, add otherwise".

We distinguish between ϵ as fixed input to a DP algorithm, and $\epsilon(\cdot)$ as a function parameterized according to a particular DP relaxation — e.g., $\epsilon(o, D, D_{\pm z})$ means the *ex-post* per-instance privacy loss conditioned on output o, dataset D, and data point z.

2.2.2 Differential Privacy

Let $\mathcal{Z}$ denote the data domain, and $\mathcal{R}$ the set of all possible outcomes of algorithm $\mathcal{A}$. Fix $\epsilon, \delta \geq 0$.

Definition 2.2.1. (Differential privacy) A randomized algorithm $\mathcal{A} : \mathcal{Z}^* \to \mathcal{R}$ satisfies (ϵ, δ)-DP if for all datasets $D \in \mathcal{Z}^*$ and data points $z \in \mathcal{Z}$, and for all measurable sets $S \subset \mathcal{R}$,

$$\Pr\big[\mathcal{A}(D) \in S\big] \leq e^{\epsilon} \Pr\big[\mathcal{A}(D_{\pm z}) \in S\big] + \delta.$$

Differential privacy guarantees that the presence or absence of any particular data record has little impact on the output distribution of a randomized algorithm. In this paper we use the "add/remove" notion of DP, by which we construct neighboring dataset $D_{\pm z}$ by adding or removing an individual z from dataset D.

DP is powerful and universal in that its guarantee applies to any D, z and set of output events. However, there are often situations where the privacy losses of $\mathcal{A}$ vary drastically depending on its input data, and the privacy loss bound ϵ (protecting even the worst-case pair of neighboring datasets) may not be informative of the privacy loss

incurred to individuals when the input to $\mathcal{A}$ is typical. This motivated Wang (2019) to consider a per-instance version of the DP definition.

Definition 2.2.2. (Per-instance differential privacy) A randomized algorithm $\mathcal{A} : \mathcal{Z}^* \to \mathcal{R}$ satisfies $\big(\epsilon(D, D_{\pm z}), \delta\big)$-pDP if for dataset D and data point z, and for all measurable sets $S \subset \mathcal{R}$,

$$\Pr\big[\mathcal{A}(D) \in S\big] \leq e^{\epsilon}\Pr\big[\mathcal{A}(D_{\pm z}) \in S\big] + \delta,$$
$$\Pr\big[\mathcal{A}(D_{\pm z}) \in S\big] \leq e^{\epsilon}\Pr\big[\mathcal{A}(D) \in S\big] + \delta.$$

The pDP definition can be viewed as using a function $\epsilon(D, D_{\pm z})$ that more precisely describes the privacy guarantee in protecting a fixed data point z when $\mathcal{A}$ is applied to dataset D.

As it turns out, it is most convenient for us to work with an even more *instance-specific* description of the privacy loss that is further parameterized by the realized output of $\mathcal{A}$ *ex-post* — after the random coins of $\mathcal{A}$ are flipped and the outcome released.

Definition 2.2.3. (*Ex-post* per-instance differential privacy) A randomized algorithm $\mathcal{A}$ satisfies $\epsilon(\cdot)$-*ex-post* per-instance differential privacy for an individual z and a fixed dataset D at an outcome $o \in \text{Range}(\mathcal{A})$ if

$$\left|\log\left(\frac{\Pr\big[\mathcal{A}(D) = o\big]}{\Pr\big[\mathcal{A}(D_{\pm z}) = o\big]}\right)\right| \leq \epsilon(o, D, D_{\pm z}).$$

This definition generalizes the *ex-post* DP definition (Ligett et al., 2017) (introduced for a different purpose) to a *per-instance* version that depends on a given pair of neighboring datasets. The above quantity is essentially the absolute value of the log-odds ratio, used extensively in hypothesis testing. Intuitively, the *ex-post* per-instance privacy loss $\epsilon(o, D, D_{\pm z})$ describes how confidently an attacker could infer, given the output of

algorithm $\mathcal{A}$, whether or not individual z is in dataset D.

Despite (or perhaps because of) its precise accounting for privacy, *ex-post* pDP could reveal sensitive information about the dataset, as the following example explicitly illustrates.

Example 2.2.4 (The privacy risk of exposing *ex-post* pDP)**.** *Consider a standard Gaussian mechanism $\mathcal{A}$ that adds noise to a counting query Q applied to dataset D, i.e. $\mathcal{A}(D) = Q(D) + \mathcal{N}(0, \sigma^2)$. Q has global sensitivity $\Delta_Q = 1$. We will show that an attacker, knowing only the output o of algorithm $\mathcal{A}$, her* ex-post *pDP loss and that her individual data is not contained in dataset D, can conclusively uncover the sensitive quantity $Q(D)$ protected by algorithm $\mathcal{A}$.*

Following the proof of Theorem A.6.1, the ex-post *pDP can be directly calculated as*

$$\epsilon(o, D, D_{\pm z}) = \frac{|Q(D) - Q(D_{\pm z})||2o - Q(D) - Q(D_{\pm z})|}{2\sigma^2}.$$

Enter attacker z, who has auxiliary information: she knows that her own individual data is not contained in D. After algorithm $\mathcal{A}$ is applied to D, attacker z receives output $o = 1$ and is informed of her ex-post *pDP $\epsilon(o, D, D_{+z})$. Since $Q(D_{+z}) = Q(D) + 1$ is known, attacker z can solve for $Q(D)$ and obtain $Q(D) = o - 0.5 \pm \sigma^2\epsilon(o, D, D_{+z})$. With probability 1, only one of the two possibilities is an integer*[1]*. Therefore, exposing* ex-post *pDP in this case completely reveals $Q(D)$.*

Problem statement. The lesson of Example 2.2.4 is that we cannot directly reveal the *ex-post* pDP losses without potentially nullifying the algorithm's privacy benefits. How, then, can we privately and accurately publish the *ex-post* pDP losses?

The goal of this paper is to develop an algorithm that publishes a *function* $\tilde{\epsilon} : \mathcal{Z} \to \mathbb{R}$ whose output estimates the *ex-post* pDP loss to an individual z of releasing the output $\hat{\theta}^P$

[1]Take $Q(D) = 0$ and $o = 0.1$ as an example, the two possibilities are 0 and -0.8.

from the objective perturbation mechanism. Any individual (not just those whose data is contained in the dataset) can plug her own data z into this function in order to receive a high-probability bound on her *ex-post* pDP loss which does not depend directly on any sensitive data except her own.

This requirement offers the same type of privacy protection as joint differential privacy (Kearns et al., 2014), which relaxes the standard DP definition by allowing an algorithm's output to individual z to be sensitive only in her own private data. Our notion of privacy is slightly more general in that it holds for individuals both in and out of the dataset. The difference lies in how the algorithm's output space is defined; whereas a joint DP algorithm produces a fixed-length tuple partitioning the output to each individual in the dataset, our algorithm outputs a function whose domain includes any data point $z \in \mathcal{Z}$. As a result, our methods are robust against collusion by arbitrary coalitions of adversaries, allowing repeated queries by any group of individuals without invalidating the privacy guarantees promised by the pDP losses.

2.2.3 Problem Setting

We consider a general family of problems known as *private empirical risk minimization* (ERM), which aim to approximate the solution to an ERM problem while preserving privacy. That is, we wish to privately solve optimization problems of the form

$$\hat{\theta} = \arg\min_{\theta \in \Theta} L(\theta; D) + r(\theta),$$

where $r(\theta)$ is a regularizer and $L(\theta; D) = \sum_{i=1}^{n} \ell(\theta; z_i)$ a loss function. Throughout, we assume that $\ell(\theta; z)$ and $r(\theta)$ are convex and twice-differentiable with respect to θ. Dataset D is given by $D = \{z_i\}_{i=1}^{n}$, and $z_i = (x_i, y_i)$ for $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y \in \mathcal{Y} \subseteq \mathbb{R}$, where $||x||_2 \leq 1$ and $|y| \leq 1$. We consider only unconstrained optimization over $\Theta = \mathbb{R}^d$.

2.2.4 Objective Perturbation

The objective perturbation algorithm solves

$$\hat{\theta}^P = \arg\min_{\theta \in \Theta} L(\theta; D) + r(\theta) + \frac{\lambda}{2}||\theta||_2^2 + b^T\theta, \tag{2.2.1}$$

where $b \sim \mathcal{N}(0, \sigma^2 I_d)$ and parameters σ, λ are chosen according to a desired (ϵ, δ)-DP guarantee.

Algorithm 1 Release $\hat{\theta}^P$ via `Obj-Pert` (Kifer et al., 2012)

Input: Dataset D, noise parameter σ, regularization parameter λ, loss function $L(\theta; D) = \sum_i \ell(\theta; z_i)$, convex and twice-differentiable regularizer $r(\theta)$, convex set Θ.

Output: $\hat{\theta}^P$, the minimizer of the perturbed objective.

Draw noise vector $b \sim \mathcal{N}(0, \sigma^2 I)$.

Compute $\hat{\theta}^P$ according to (2.2.1).

Theorem 2.2.5 (Privacy guarantees of Algorithm 1 (Kifer et al., 2012))**.** *Consider dataset $D = \{z_i\}_{i=1}^n$; loss function $L(\theta; D) = \sum_i \ell(\theta; z_i)$; convex regularizer $r(\theta)$; and convex domain Θ. Assume that $\nabla^2 \ell(\theta; z_i) \prec \beta I_d$ and $||\nabla \ell(\theta; z_i)||_2 \leq \xi$ for all $z_i \in \mathcal{X} \times \mathcal{Y}$ and for all $\theta \in \Theta$. For $\lambda \geq 2\beta/\epsilon_1$ and $\sigma = \xi^2(8\log(2/\delta) + 4\epsilon_1)/\epsilon_1^2$, Algorithm 1 satisfies (ϵ_1, δ)-differential privacy.*

The privacy guarantees stated in Theorem 2.2.5 apply even when θ is constrained to a closed convex set, but for ease of our per-instance privacy analysis we will require $\Theta = \mathbb{R}^d$ from this point on.

2.3 Privately Publishable pDP

2.3.1 pDP Analysis of Objective Perturbation

Our goal in this section is to derive the personalized privacy losses (under Definition 2.2.3) associated with observing the output $\hat{\theta}^P$ of objective perturbation. This *ex-post* perspective is highly adaptive and also convenient for our analysis of Algorithm 1, whose privacy parameters are a function of the data. Since we are analyzing the per-instance privacy cost of *releasing* $\hat{\theta}^P$, it makes perfect sense to condition the pDP loss on the privatized output of the computation.

Our first technical result is a precise calculation of the *ex-post* pDP loss of objective perturbation.

Theorem 2.3.1 (*ex-post* pDP loss of objective perturbation for a convex loss function)**.** *Let* $J(\theta; D) = L(\theta; D) + r(\theta) + \frac{\lambda}{2}||\theta||_2^2$ *such that* $L(\theta; D) + r(\theta) = \sum_i \ell(\theta; z_i) + r(\theta)$ *is a convex and twice-differentiable regularized loss function, and sample* $b \sim \mathcal{N}(0, \sigma^2 I_d)$*. Then for every privacy target* $z = (x, y)$*, releasing* $\hat{\theta}^P = \arg\min_{\theta \in \mathbb{R}^d} J(\theta; D) + b^T\theta$ *satisfies* $\epsilon_1(\hat{\theta}^P, D, D_{\pm z})$*-ex-post per-instance differential privacy with*

$$\epsilon_1(\hat{\theta}^P, D, D_{\pm z}) = \left| -\log \prod_{j=1}^{d} \left(1 \mp \mu_j\right) + \frac{1}{2\sigma^2}||\nabla \ell(\hat{\theta}^P; z)||_2^2 \pm \frac{1}{\sigma^2} \nabla J(\hat{\theta}^P; D)^T \nabla \ell(\hat{\theta}^P; z) \right|,$$

where $\mu_j = \lambda_j u_j^T \left(\nabla b(\hat{\theta}^P; D) \mp \sum_{k=1}^{j-1} \lambda_k u_k u_k^T\right)^{-1} u_j$ *according to the eigendecomposition* $\nabla^2 \ell(\theta; z) = \sum_{k=1}^{d} \lambda_k u_k u_k^T$*.*

Proof sketch. Following the analysis of Chaudhuri et al. (2011), we establish a bijection between the mechanism output $\hat{\theta}^P$ and the noise vector b, and use a change-of-variables defined by the Jacobian mapping between $\hat{\theta}^P$ and b in order to rewrite the log-probability ratio in terms of the probability density function of b. First-order conditions then allow

us to solve directly for the distribution of b. To calculate the first term of the above equation, we use the eigendecomposition of the Hessian $\nabla^2 \ell(\hat{\theta}^P; z)$ and recursively apply the matrix determinant lemma. The rest of the proof is straightforward algebra. The full proof is given in Appendix A.5. □

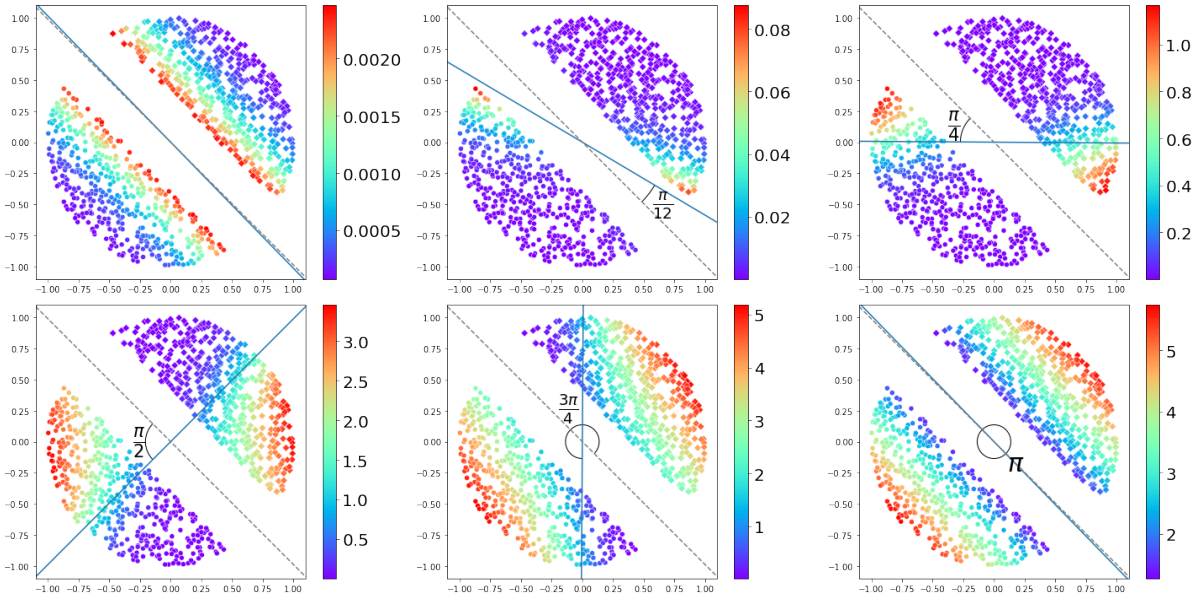
The above expression holds for any convex loss function, but is a bit unwieldy. The calculation becomes much simpler when we assume $\ell(\cdot)$ to be a generalized linear loss function, with inner-product form $\ell(\theta; z) = f(x^T\theta; y)$. For the sake of interpretability, we will defer further discussion of the *ex-post* pDP loss of objective perturbation until after presenting the following corollary.

Corollary 2.3.2 (*ex-post* pDP loss of objective perturbation for GLMs)**.** *Let* $J(\theta; D) = L(\theta; D) + r(\theta) + \frac{\lambda}{2}||\theta||_2^2$ *such that* $L(\theta; D) = \sum_i \ell(\theta; z_i)$ *is a linear loss function, and sample* $b \sim \mathcal{N}(0, \sigma^2 I_d)$*. Then for every privacy target* $z = (x, y)$*, releasing* $\hat{\theta}^P = \arg\min_{\theta \in \mathbb{R}^d} J(\theta; D) + b^T\theta$ *satisfies* $\epsilon_1(\hat{\theta}^P, D, D_{\pm z})$*-ex-post per-instance differential privacy with*

$$\epsilon(\hat{\theta}^P, D, D_{\pm z}) \leq \left| -\log\left(1 \pm f''(\cdot)\mu(x)\right) + \frac{1}{2\sigma^2}||\nabla \ell(\hat{\theta}^P; z)||_2^2 \pm \frac{1}{\sigma^2}\nabla J(\hat{\theta}^P; D)^T \nabla \ell(\hat{\theta}^P; z) \right|,$$

where $\mu(x) = x^T\left(\nabla^2 J(\hat{\theta}^P; D)\right)^{-1}x$*,* $\nabla \ell(\hat{\theta}^P; z) = f'(x^T\hat{\theta}^P; y)x$ *and* $f''(\cdot)$ *is shorthand for* $f''(\cdot) = f''(x^T\hat{\theta}^P; y)$*. The notation* $b(\hat{\theta}^P; D)$ *means the realization of the noise vector* b *for which the output of Algorithm 1 will be* $\hat{\theta}^P$ *when the input dataset is* D*.*

Note that the quantity $\mu(x)$ in the first term is the *generalized leverage score* (Wei et al., 1998), quantifying the influence of a data point on the model fit. The second and third terms are a function of the gradient of the loss function and provide a complementary measure of how well the fitted model predicts individual z's data.



Since the *ex-post* pDP is a function of $\hat{\theta}^P$, we don't even need to run Algorithm 1 to calculate *ex-post* pDP losses – we can plug in directly to Corollary 2.3.2 in order to calculate the pDP distribution induced by any hypothetical $\hat{\theta}^P$. For Figure 2.3.1, we use a synthetic dataset D sampled from the unit ball with two linearly separable classes separated by margin $m = 0.4$. Then we solve for $\hat{\theta} = \arg\min J(\theta; D)$ with $\lambda = 1$ to minimize the logistic loss, and directly perturb the output by rotating it by angle $\omega \in [0, \frac{\pi}{12}, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \pi]$. We then denote $\hat{\theta}^P := \theta_{+\omega}$ to mean θ rotated counter-clockwise by angle ω. The color scale is a function of the *ex-post* pDP loss of data point z.

Figure 2.3.1 illustrates how the mechanism output $\hat{\theta}^P$ affects the *ex-post* pDP distribution of objective perturbation for our logistic regression problem. For $\omega \in [0, \frac{\pi}{12}]$, the data points closest to the decision boundary have the highest *ex-post* pDP loss. These data points have a strong effect on the learned model and would therefore have high *leverage scores*, making the first term dominate. As the perturbation (and model error) increases, the second and third terms dominate; the more badly a model predicts a data point, the less protection this data point has.

Hidden in this analysis are the δ's of Theorem 2.2.5, which along with the choice of σ and λ could affect which of the three terms is dominant. Fortunately, the probability of outputting something like $\hat{\theta}^P = \theta_{+\pi}$ is astronomically low for any reasonable privacy setting!

2.3.2 Releasing the pDP losses

Next we consider: after having released $\hat{\theta}^P$ and calculated the per-instance privacy losses of doing so, how do we privately release these pDP losses? Our goal is to allow any individual $z \in \mathcal{Z}$ (in the dataset or not) to know her privacy loss while preserving the privacy of others in the dataset.

Observe that the expression from Theorem 2.3.1 depends on the dataset D only through two quantities: the leverage score $\mu(x) = x^T\big(\nabla^2 J(\hat{\theta}^P; D)\big)^{-1}x$ and the inner product $\nabla J(\hat{\theta}^P; D)^T \nabla \ell(\hat{\theta}^P; z)$. As a result, if we can find a data-independent bound for these two terms, or privately release them with only a small additional privacy cost, then we are done.

Data-independent bound of *ex-post* pDP losses

Below, we present a pair of lemmas which will allow us to find a high-probability, data-independent bound on the *ex-post* pDP loss.

Theorem 2.3.3. *Suppose $\ell(\cdot)$ is a function with continuous second-order partial derivatives. Then*

$$\left| -\log \prod_{j=1}^{d} \Big(1 \mp \mu_j\Big) \right| \leq -\sum_{j=1}^{d} \log(1 - \frac{\lambda_j}{\lambda}),$$

where $\mu_j = \lambda_j u_j^T \Big(\nabla \mathbf{b}(\hat{\theta}^P; D) \mp \sum_{k=1}^{j-1} \lambda_k u_k u_k^T\Big)^{-1} u_j$ according to the eigendecomposi-

tion $\nabla^2 \ell(\hat{\theta}^P; z) = \sum_{k=1}^{d} \lambda_k u_k u_k^T$. *When specializing to linear loss functions such that* $\ell(\theta; z) = f(x^T\theta; y)$, $\lambda_j = 0$ *for all* $j > 1$ *and the above bound can be simplified to* $-\log\left(1 - f''(x^T\hat{\theta}^P; y)||x||_2^2/\lambda\right)$.

Theorem 2.3.4. *Let* $\hat{\theta}^P$ *be a random variable such that* $\hat{\theta}^P = \arg\min\left(J(\theta; D) + b^T\theta\right)$ *as in (2.2.1), where* $b \sim \mathcal{N}(0, \sigma^2 I_d)$ *and* $\ell(\theta; z)$ *is a convex and twice-differentiable loss function. Then for* $z \in \mathcal{Z}$, *the following holds with probability* $1 - \rho$*:*

$$\left|\nabla J(\hat{\theta}^P; D)^T \nabla \ell(\hat{\theta}^P; z)\right| \leq \sigma\sqrt{2\log(2d/\rho)}\|\nabla \ell(\hat{\theta}^P; z)\|_1.$$

For linear loss functions the bound can be substantially strengthened to

$$\left|\nabla J(\hat{\theta}^P; D)^T \nabla \ell(\hat{\theta}^P; z)\right| \leq f'(x^T\hat{\theta}^p; y)\sigma||x||_2\sqrt{2\log(2/\rho)}.$$

We make a few observations on the bounds. First, the general bound in Theorem 2.3.4 holds simultaneously for all z and it depends only logarithmically in dimension when the features are *sparse*. Second, the bound for a linear loss function is dimension-free and somewhat surprising because we are actually bounding an inner product of two *dependent* random vectors (both depend on $\hat{\theta}^P$).

Finally, we remark that the bounds in this section are data-independent in that they do not depend on the rest of the dataset beyond already released information $\hat{\theta}^P$. It allows us to reveal a pDP bound of each individual when she plugs in her own data without costing any additional privacy budget!

2.3.3 The privacy report

For certain regimes, we may wish to consider privatizing the data-dependent quantities of the *ex-post* pDP losses, at an additional privacy cost, as an alternative to using data-

independent bounds. Of course, it only makes sense to do so if we can show that (a) these data-dependent estimates are more accurate than the data-independent bounds; (b) the overhead of releasing additional quantities (the additional privacy cost in terms of both DP and pDP) is not too large; and (c) we can share the pDP losses of the private reporting algorithm using data-independent bounds (so we do not have to recursively publish such reports).

Full details are in the appendix. We show that by adding slightly more regularization than required by `Obj-Pert` (i.e., making λ just a bit larger so that the minimum eigenvalue of the Hessian $H = \nabla^2 J$ is above a certain threshold), we can find a multiplicative bound that estimates $\mu(x) = x^T H^{-1} x$ uniformly for all x. We do so by adding noise to the Hessian using a natural variant of "Analyze Gauss" (Dwork et al., 2014b), hence privately releasing $\overline{\mu^P} : \mathcal{X} \to \mathbb{R}$. See Algorithm 2 for details.

For brevity, we use the short-hands $f'(\cdot) := f'(x^T\hat{\theta}^P; y)$ and $f''(\cdot) := f''(x^T\hat{\theta}^P; y)$, where $\ell(\theta; z) = f(x^T\theta; y)$ for GLMs. $F^{-1}_{\mathcal{N}(0,1)}$ is the inverse CDF of the standard normal distribution, and $F^{-1}_{GOE(d)}$ is the inverse CDF of the largest eigenvalue of the Gausian Orthogonal Ensemble (GOE) matrix, whose distribution is calculated exactly by Chiani (2014). Algorithm 2 specializes to GLMs for clarity of presentation, but we could adapt it to any convex loss function by replacing the GLM-specific bounds with the more general ones.

We implicitly assume that the data analyst has already decided the privacy budgets ϵ_2 and ϵ_3 for the data-dependent release of the gradient (third term of $\epsilon_1(\cdot)$) and of the Hessian (first term of $\epsilon_1(\cdot)$). Inputs σ_2 and σ_3 are then calibrated to achieve (ϵ_2, ρ)-DP and (ϵ_3, ρ)-DP, respectively.

Algorithm 2 Privacy report for `Obj-Pert` on GLMs

Input: $\hat{\theta}^p \in \mathbb{R}^d$ from `Obj-Pert`, noise parameter $\sigma, \sigma_2, \sigma_3$; regularization parameter λ; Hessian $H := \sum_i \nabla^2 \ell(\hat{\theta}^p; z_i) + \lambda I_d$, Boolean B $\in$ [`DATA-INDEP`, `DATA-DEP`], failure probability ρ

Require: $\lambda \geq 2\sigma_3 F^{-1}_{\lambda_1(\mathrm{GOE}(d))}(1-\rho/2)$

Output: Reporting function $\tilde{\epsilon} : (x, y), \delta \to \mathbb{R}^3_+$

if B = `DATA-INDEP` **then**

Set $\epsilon_2(\cdot) := 0, \epsilon_3(\cdot) := 0$.

Set $\overline{g^P}(z) := \sigma ||f'(\cdot)x||_2 F^{-1}_{\mathcal{N}(0,1)}(1-\rho/2)$ and set $\overline{\mu^p}(x) := \frac{\|x\|^2}{\lambda}$.

else if B = `DATA-DEP` **then**

Privately release $\hat{g}^p$ by Algorithm 10 with parameter σ_2.

Set $\epsilon_2(\cdot)$ according to Theorem A.3.4.

Set $\overline{g^P}(z) := \min \begin{cases} f'(\cdot)[\hat{g}^P(z)]^T x + \sigma_2 ||f'(\cdot)x||_2 F^{-1}_{\mathcal{N}(0,1)}(1-\rho/2), \\ \sigma ||f'(\cdot)x||_2 F^{-1}_{\mathcal{N}(0,1)}(1-\rho/2). \end{cases}$

Privately release $\hat{H}^p$ by a variant of "Analyze Gauss"[2]with parameter σ_3.

Set $\epsilon_3(\cdot)$ according to Statement 2 of Theorem 2.3.5.

Set $\overline{\mu^p}(x) = \frac{3}{2} x^T [\hat{H}^p]^{-1} x$.

end if

Set $\overline{\epsilon_1^p}(z) := \left| -\log\left(1 - f''(\cdot)\overline{\mu^p}(x)\right)\right| + \frac{||f'(\cdot)x||_2^2}{2\sigma^2} + \frac{\left|\overline{g^P}(z)\right|}{\sigma^2}$.

Output the function $\tilde{\epsilon}(z) := \left(\overline{\epsilon_1^p}(z), \epsilon_2(z), \epsilon_3(z)\right)$.

Note that the pDP functions $\epsilon_2(\cdot)$ and $\epsilon_3(\cdot)$ – which we use to report the additional pDP losses of releasing the private estimates of the gradient and the Hessian – do not depend on the dataset, and thus are not required to be separately released. The privately

[2]Instead of adding "analyze-gauss" noise, we sample from the Gaussian Orthogonal Ensemble (GOE) distribution to obtain a random matrix (Appendix A.4). Under this model we show that τ is on the order of $O(\sqrt{d}(1 + \log(C/\rho)^{3/2}))$.

released pDP functions depend on $\hat{\theta}^P$; to reduce clutter, we omit this parameter in our presentation of Algorithm 2.

Theorem 2.3.5. *There is a universal constant C such that if $\lambda > C\sigma_2\sqrt{d}(1+(\log(1/\rho))^{2/3})$, then Algorithm 2 satisfies the following properties*

1. $(\frac{\xi^2}{2\sigma_2^2} + \frac{\beta^2}{4\sigma_3^2} + \sqrt{\frac{\xi^2}{\sigma_2^2} + \frac{\beta^2}{2\sigma_3^3}}\sqrt{2\log(1/\delta)}, \delta)$*-DP*
2. $(\frac{f'(\hat{\theta}^p;z)^2\|x\|^2}{2\sigma_2^2} + \frac{f''(\hat{\theta}^p;z)^2\|x\|^4}{4\sigma_3^2} + \sqrt{\frac{f'(\hat{\theta}^p;z)^2\|x\|^2}{\sigma_2^2} + \frac{f''(\hat{\theta}^p;z)^2\|x\|^4}{2\sigma_3^2}}\sqrt{2\log(1/\delta)}, \delta)$*-pDP for all $x \in \mathcal{X}$ and $0 \leq \delta < 1$.*
3. *For a fixed input z and D, and all $\rho > 0$, the privately released privacy report $\tilde{\epsilon}(\cdot)$ satisfies that $\epsilon_1(\hat{\theta}^p, D, D_{\pm z}) \leq \overline{\epsilon_1^p}(z) \leq 12\epsilon_1(\hat{\theta}^p, D, D_{\pm z}) + \frac{|f'(\cdot)|\|x\|}{\sigma_2}\sqrt{2\log(2/\rho)}$ with probability $1 - 3\rho$ where $\epsilon_1(\cdot)$ is the expression from Theorem 2.3.1.*

Accurate approximation with low privacy cost. This theorem shows that if we use a slightly larger λ in ObjPert then we get an upper bound of the pDP for each individual z up to a multiplicative and an additive factor. The multiplicative factor is coming from a multiplicative approximation of $-\log\left(1 \pm f''(\cdot)\mu(x)\right)$ and the additive error is due to the additional noise added for releasing the third term $\frac{1}{\sigma^2}\nabla J(\hat{\theta}^P; D)^T\nabla\ell(\hat{\theta}^P; z)$. The additional DP and pDP losses for releasing H and g are comparable to the DP and pDP losses in Objective Perturbation itself if $\sigma_2 \asymp \sigma_3 \asymp \sigma$.

Moreover, while using a large λ may appear to introduce additional bias, the required choice of $\lambda \asymp \sqrt{d}\sigma$ is actually exactly the choice to obtain the minimax rate in general convex private ERM (Bassily et al., 2014) (Figure 2.1 demonstrates the impact of increasing λ).

Joint DP interpretation. Finally, we can also interpret our results from a joint-DP perspective (Kearns et al., 2014). Given any realized output $\hat{\theta}^p \in \mathbb{R}^d$, the tuple of

$\{\tilde{\epsilon}(z_1, \hat{\theta}^p), ..., \tilde{\epsilon}(z_1, \hat{\theta}^p)\}$ satisfies joint DP with the same ϵ parameter as in Theorem 2.3.5. This follows from the billboard lemma (Hsu et al., 2016).

2.4 Experiments

Here we evaluate our methods to release the pDP losses using logistic regression as a case study. In Section 2.4.1, we demonstrate that the stronger regularization required by Algorithm 2 does not affect the utility of the model. In Section 2.4.2 we show that by carefully allocating the privacy budget of the data-dependent release, we can achieve a more accurate estimate of the *ex-post* pDP losses of Algorithm 1 compared to the data-independent release, with reasonable overhead (same overall DP budget and only a slight uptick in the overall pDP losses).

Experiments with linear regression, with additional datasets and with alternative privacy budget allocation schemes are included in the supplementary materials.

2.4.1 Stronger regularization does not worsen model utility

In this experiment we use a synthetic dataset generated by sampling $x_i, \theta \sim \mathcal{N}(0, I_d)$ and normalizing each $x_i \in X$ so that $||x_i||_2 = 1$. Then we rescale $Y = X\theta$ to ensure $y_i \in [0, 1]$ for each $y_i \in Y$.

Algorithm 2 requires a larger λ than suggested by Theorem 2.2.5 in order to achieve a uniform multiplicative approximation of $\mu(\cdot)$. We investigate the effect of stronger regularization on the utility of a private logistic regression model applied to a synthetic dataset ($n = 1000, d = 50$), for several settings of ϵ_1.

E.g., for logistic regression the objective perturbation mechanism requires $\lambda \geq \frac{1}{2\epsilon}$, and so in Figure 2.1 a λ-inflation value of 10 means that we set $\lambda = \frac{5}{\epsilon}$. For each λ-inflate value c , we run Algorithm 1 with $\lambda = c\lambda_{\texttt{Obj-Pert}}$. In particular, the star symbol marks

the level of λ-inflation enforced by Algorithm 2.

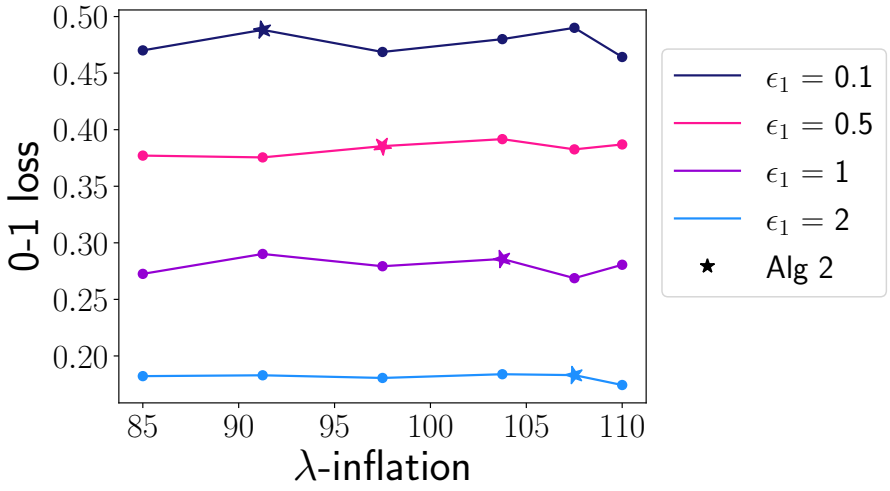


Figure 2.1: Utility of Obj-Pert with larger λ.

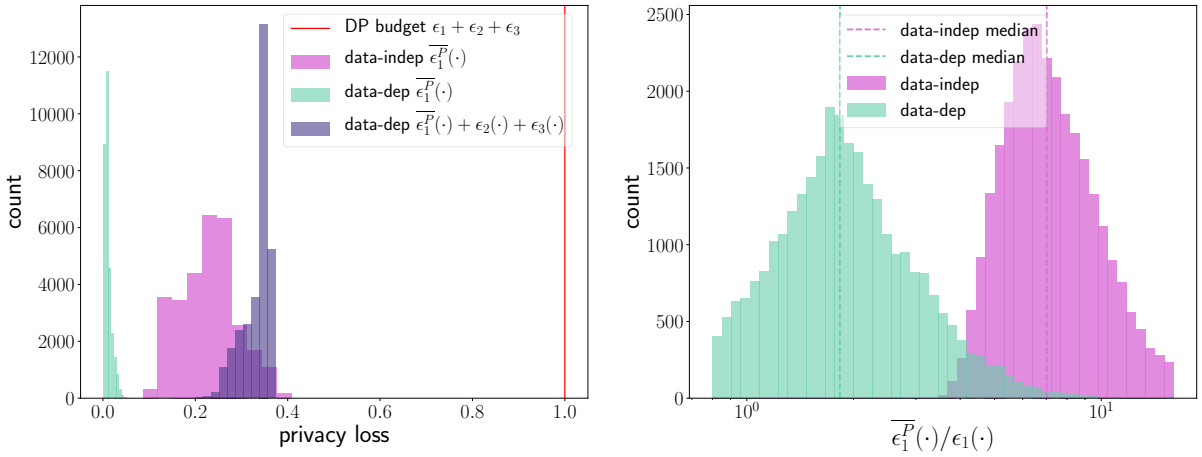
The experimental results summarized in Figure 2.1 show that the performance of the private logistic regression model (as measured by the 0-1 loss) remains roughly constant across varying scales of λ.

2.4.2 Comparison of data-independent and data-dependent bounds

The following experiments feature the credit card default dataset ($n = 30000, d = 21$) (Yeh and Lien, 2009) from the UCI Machine Learning Repository. We privately train a binary classifier to predict whether or not a credit card client z defaults on her payment (Algorithm 1), and calculate the true pDP loss $\epsilon_1(\cdot)$ as well as the data-independent and -dependent estimates $\overline{\epsilon_1^P}(\cdot)$ for each z in the training set (Algorithm 2).

The failure probabilities for both Algorithms 1 and 2 are set as $\delta = \rho = 10^{-6}$. Our choices of σ and λ depend on ϵ_1 and follows the requirements stated in Theorem 2.2.5 to achieve DP. We don't use any additional regularization, i.e. $r(\theta) = 0$. For the data-

dependent release, the noise parameters σ_2, σ_3 are each calibrated according to the analytic Gaussian mechanism of Balle and Wang (2018).



((a)) Distribution of privately released *ex-post* pDP losses.

((b)) Distribution of ratios between the privately released *ex-post* pDP losses $\overline{\epsilon_1^P}(\cdot)$ and their true values $\epsilon_1(\cdot)$.

Figure 2.2: True and privately released pDP losses when the total privacy budget is $\epsilon = 1$. For the data-independent release we use the entire privacy budget on releasing $\hat{\theta}^P$ ($\epsilon_1 = 1$). For the data-dependent release we reserve some of the privacy budget for releasing $\overline{\mu^P}(\cdot)$ and $\overline{g^P}(\cdot)$ ($\epsilon_1 = 0.2, \epsilon_2 = 0.7, \epsilon_3 = 0.1$).

Using $\epsilon = 1$ as a DP budget, we investigate how to allocate the privacy budget among the components of the data-dependent release ($\epsilon = \epsilon_1 + \epsilon_2 + \epsilon_3$) to achieve a favorable comparison with the data-independent release which requires no additional privacy cost ($\epsilon = \epsilon_1$). The configuration described in Figure A.3, which skews the data-dependent privacy budget toward more accurately releasing $\overline{\epsilon_1^P}(\cdot)$, was empirically chosen as an example where the sum $\overline{\epsilon_1^P}(\cdot) + \epsilon_2(\cdot) + \epsilon_3(\cdot)$ of privately released pDP losses of the data-dependent approach are comparable to the privately released *ex-post* pDP loss $\overline{\epsilon_1^P}(\cdot)$ of

the data-independent approach. Note that $\epsilon_2(\cdot)$ and $\epsilon_3(\cdot)$ aren't *ex-post* in the traditional sense; however, we feel comfortable summing $\overline{\epsilon_1^P}(\cdot) + \epsilon_2(\cdot) + \epsilon_3(\cdot)$ since all three terms are a function of $\hat{\theta}^P$ and individual z's data. Note also that since the total budget ϵ is the same for both the data-independent and -dependent releases, ϵ_1 differs between them. Therefore Figure 2.2(b) compares the accuracy of both approaches using the ratio between $\overline{\epsilon_1^P}(\cdot)$ and $\epsilon_1(\cdot)$ rather than their raw values.

When including the additional privacy budget incurred by the data-dependent approach, the data-dependent approach loses its competitive edge over the data-independent approach. Note that setting $\epsilon_2 = \epsilon_3 = 0$ would reduce the data-dependent approach to the data-independent one. The real advantage of the data-dependent approach can be best seen by allotting only a small portion of the overall privacy budget to Algorithm 1; then we can release $\hat{\theta}^P$ and $\overline{\epsilon_1^P}(\cdot)$ with reasonable overhead while achieving tighter and more accurate upper bounds for $\overline{\mu^P}(\cdot)$ and $\overline{g^P}(\cdot)$. By suffering a small additional *ex-post* pDP loss (Figure 2.2(a)), we can release the *ex-post* pDP losses of Algorithm 1 much more accurately (Figure 2.2(b)). The downside to this is that reducing ϵ_1 reduces the accuracy of the output $\hat{\theta}^P$. Deciding how to allocate the privacy budget between ϵ_1, ϵ_2 and ϵ_3 thus requires weighing the importance of an accurate $\hat{\theta}^P$ against the importance of an accurate $\overline{\epsilon_1^P}(\cdot)$.

2.5 Conclusion

In this chapter we have derived the *ex-post* per-instance privacy losses of objective perturbation, and shown how to privately and accurately publish them. These data-dependent privacy losses are significantly smaller than the worst-case DP bound, demonstrating that the large values of ϵ commonly used in practice may offer stronger protection to most data than is implied by the loose privacy guarantee.

Publishing these data-dependent privacy losses, however, does not improve the privacy guarantee based on the worst-case DP bound. In the next chapter we will introduce an adaptive algorithm which uses data-dependent privacy losses in order to improve the privacy-utility trade-off.

Chapter 3

Generalized Propose-Test-Release

3.1 Introduction

The guarantees of differential privacy (DP) (Dwork et al., 2006) are based on worst-case outcomes across all possible datasets. A common paradigm is therefore to add noise scaled by the *global sensitivity* of a query f, which measures the maximum change in f between any pair of neighboring datasets.

A given dataset X might have a *local sensitivity* $\Delta_{LS}(X)$ that is much smaller than the global sensitivity Δ_{GS}, in which case we can hope to add a smaller amount of noise (calibrated to the local rather than global sensitivity) while achieving the same privacy guarantee. This must not be undertaken naïvely; the local sensitivity is a dataset-dependent function and so calibrating noise to the local sensitivity could leak information about the dataset (Nissim et al., 2007).

The "Propose-Test-Release" (PTR) framework (Dwork and Lei, 2009) resolves this issue by introducing a test to privately check whether a proposed bound on the local sensitivity is valid. Only if the test "passes" is the output released with noise calibrated to the proposed bound on the local sensitivity.

PTR is a powerful tool for designing data-adaptive DP algorithms, but it has several limitations. First, it applies only to noise-adding mechanisms which calibrate noise according to the sensitivity of a query. Second, the test in "Propose-Test-Release" is computationally expensive for all but a few simple queries such as privately releasing the median or mode. Third, while some existing works (Decarolis et al., 2020; Kasiviswanathan et al., 2013; Liu et al., 2021) follow the adaptive approach of privately testing properties of an input dataset for "niceness"[1], there has not been a systematic recipe for *discovering* which properties should be tested.

In this paper, we propose a generalization of PTR which addresses these limitations. The centerpiece of our framework is a differentially private test on the *data-dependent privacy loss*. This test does not directly consider the local sensitivity of a query and is therefore not limited to additive noise mechanisms. Moreover, in many cases the test can be efficiently implemented by privately releasing a high-probability upper bound, thus avoiding the need to search an exponentially large space of datasets. Furthermore, the derivation of the test itself often spells out exactly what properties of the input dataset need to be checked, which streamlines the design of data-adaptive DP algorithms.

Our contributions are summarized as follows:

1. We propose a generalization of PTR which can handle algorithms beyond noise-adding mechanisms. Generalized PTR allows us to plug in *any* data-dependent DP analysis to construct a high-probability DP test that adapts to favorable properties of the input dataset, without painstakingly designing each test from scratch.

2. We show that many existing examples of PTR and PTR-like methods can be unified under the generalized PTR framework, sometimes resulting in a tighter analysis (see an example of report-noisy-max in Section B.3.1).

[1]We refer to these as PTR-like algorithms.

3. We demonstrate that one can publish a DP model through privately upper-bounding a one-dimensional statistic — no matter how complex the output space of the mechanism is. We apply this result to solve an open problem from PATE (Papernot et al., 2017, 2018a).

4. Our results broaden the applicability of private hyperparameter tuning (Liu and Talwar, 2019; Papernot and Steinke, 2021) in enabling joint selection of DP-specific parameters (e.g., noise level) and native parameters of the algorithm (e.g., regularization).

3.2 Related Work

Data-dependent DP algorithms. Privately calibrating noise to the local sensitivity is a well-studied problem. One approach is to add noise calibrated to the smooth sensitivity (Nissim et al., 2007), an upper bound on the local sensitivity which changes slowly between neighboring datasets. An alternative to this — and the focus of our work — is Propose-Test-Release (PTR) (Dwork and Lei, 2009), which works by calculating the distance $\mathcal{D}_\beta(X)$ to the nearest dataset to X whose local sensitivity violates a proposed bound β. The PTR algorithm then adds noise to $\mathcal{D}_\beta(X)$ before testing whether this privately computed distance is large enough to permit releasing the output with noise calibrated to β.

PTR spin-offs abound. Notable examples include stability-based methods (Thakurta and Smith, 2013) (stable local sensitivity of 0 near the input data) and privately releasing upper bounds of local sensitivity (Kasiviswanathan et al., 2013; Liu et al., 2021; Decarolis et al., 2020). We refer readers to Chapter 3 of Vadhan (2017) for a concise summary of these classic results. More recently, Wang et al. (2022) have provided Rényi DP bounds

(Mironov, 2017) for PTR and demonstrated its applications to robust DP-SGD. Our work (Section 3.6.2) also considers applications of PTR in data-adaptive private deep learning: Instead of testing the local sensitivity of each gradient step as in Wang et al. (2022), our PTR-based PATE algorithm tests the data-dependent privacy loss as a whole.

Liu et al. (2021) proposed the High-dimensional Propose-Test-Release (HPTR) framework. HPTR provides a systematic way of solving DP statistical estimation problems by using the exponential mechanism (EM) with carefully constructed scores based on certain one-dimensional robust statistics, which have stable local sensitivity bounds. HPTR focuses on designing data-adaptive DP mechanisms from scratch; our method, in contrast, converts existing randomized algorithms (including EM and even some that do not satisfy DP) into those with formal DP guarantees. Interestingly, our proposed method also depends on a one-dimensional statistic of direct interest: the data-dependent privacy loss.

Data-dependent DP losses. The flip side of data-dependent DP algorithms is the study of data-dependent DP losses (Papernot et al., 2018a; Soria-Comas et al., 2017; Wang, 2017), which fix the randomized algorithm but parameterize the resulting privacy loss by the specific input dataset. For example: In the simple mechanism that adds Laplace noise with parameter b, data-dependent DP losses are $\epsilon(X) = \Delta_{LS}(X)/b$. The data-dependent DP losses $\epsilon(X)$ are often much smaller than the DP loss ϵ, but they themselves depend on the data and thus may reveal sensitive information; algorithms satisfying a data-dependent privacy guarantee are not formally DP with guarantees any smaller than that of the worst-case. Existing work has considered privately publishing these data-dependent privacy losses (Papernot et al., 2018a; Redberg and Wang, 2021), but notice that privately publishing these losses does not improve the DP parameter of the given algorithm. Part of our contribution is to resolve this conundrum by showing that a simple post-processing step of the privately released upper bound of $\epsilon(X)$ gives a formal DP algorithm.

Private hyperparameter tuning. Our work has a nice connection with private hyperparameter tuning. Prior work (Liu and Talwar, 2019; Papernot and Steinke, 2021) requires each candidate configuration to be released with the same DP (or Rényi DP) parameter set. Another hidden assumption is that the parameters must not be privacy-correlated (i.e., parameter choice will not change the privacy guarantee). Otherwise we need to use the largest DP bound across all candidates. For example, Liu and Talwar (2019) show that if each mechanism (instantiated with one group of hyperparameters) is $(\epsilon, 0)$-DP, then running a random number of mechanisms and reporting the best option satisfies $(3\epsilon, 0)$-DP. Our work directly generalizes the above results by (1) considering a wide range of hyperparameters, either privacy-correlated or not; and (2) requiring only that individual candidates have a *testable* data-dependent DP.

3.3 Preliminaries

Datasets $X, X' \in \mathcal{X}$ are neighbors if they differ by no more than one datapoint; we say $X \simeq X'$ if $d(X, X') \leq 1$.

We measure the distance $d(\cdot)$ between same-sized datasets $X = \{x_i\}_{i=1}^n$ and $\tilde{X} = \{\tilde{x}_i\}_{i=1}^n$ as the number of coordinates that differ between them:

$$d(X, \tilde{X}) = \#\{i \in [n] : x_i \neq \tilde{x}_i\}.$$

We use $||\cdot||$ to denote the radius of the smallest Euclidean ball that contains the input set, e.g. $||\mathcal{X}|| = \sup_{x \in \mathcal{X}} ||x||$.

For mechanisms with continuous output space, the probability density of $\mathcal{M}(X)$ at y is denoted $\Pr[\mathcal{M}(X) = y]$.

Definition 3.3.1 (Differential privacy (Dwork et al., 2006))**.** Fix $\epsilon, \delta \geq 0$. A randomized

algorithm $\mathcal{M} : \mathcal{X} \to \mathcal{R}$ satisfies (ϵ, δ)-DP if for all neighboring datasets $X \simeq X'$ and for all measurable sets $S \subseteq \mathcal{R}$,

$$\Pr\big[\mathcal{M}(X) \in S\big] \leq e^{\epsilon}\Pr\big[\mathcal{M}(X') \in S\big] + \delta.$$

Definition 3.3.2 (Sensitivity)**.** The global ℓ_*-sensitivity of a function f is defined as

$$\Delta_{GS} = \max_{X,X':X\simeq X'} ||f(X) - f(X')||_*$$

and its local sensitivity at dataset X is

$$\Delta_{LS}(X) = \max_{X\simeq X'} ||f(X) - f(X')||_*.$$

Theorem 3.3.3 (Noise-adding mechanisms (Dwork et al., 2006; Balle and Wang, 2018))**.** *Consider a real-valued function $f : \mathcal{X} \to \mathbb{R}$ with global ℓ_1-sensitivity Δ_1 and global ℓ_2-sensitivity Δ_2.*

The Laplace mechanism $\mathcal{M}(X) = f(X) + Lap\,(\Delta_1/\epsilon)$ satisfies ϵ-differential privacy.

The Gaussian mechanism $\mathcal{M}(X) = f(X) + \mathcal{N}(0, \sigma^2)$ satisfies $(\epsilon, \delta(\epsilon))$-differential privacy for all $\epsilon \geq 0$ with $\delta(\epsilon) = \Phi(\frac{\Delta_2}{2\sigma} - \frac{\epsilon\sigma}{\Delta_2}) - e^{\epsilon}\Phi(-\frac{\Delta_2}{2\sigma} + \frac{\epsilon\sigma}{\Delta_2})$, where Φ is the cumulative density function of a standard normal distribution.

3.3.1 Propose-Test-Release

Calibrating the noise level to the local sensitivity $\Delta_{LS}(X)$ of a function would allow us to add less noise and therefore achieve higher utility for releasing private queries. However, the local sensitivity is a data-dependent function and naïvely calibrating the noise level to $\Delta_{LS}(X)$ will not satisfy DP.

PTR resolves this issue in a three-step procedure: **propose** a bound on the local sensitivity, privately **test** that the bound is valid (with high probability), and if so calibrate noise according to the bound and **release** the output.

PTR privately computes the distance $\mathcal{D}_\beta(X)$ between the input dataset X and the nearest dataset X'' whose local sensitivity exceeds the proposed bound β:

$$\mathcal{D}_\beta(X) = \min_{X''}\{d(X, X'') : \Delta_{LS}(X'') > \beta\}.$$

Algorithm 3 Propose-Test-Release (Dwork and Lei, 2009)

1: **Input**: Dataset X; privacy parameters ϵ, δ; proposed bound β; query function $f : \mathcal{X} \to \mathbb{R}$.
2: **if** $\mathcal{D}_\beta(X) + \text{Lap}\left(\frac{1}{\epsilon}\right) \leq \frac{\log(1/\delta)}{\epsilon}$ **then** output $\perp$,
3: **else** release $f(X) + \text{Lap}\left(\frac{\beta}{\epsilon}\right)$.

Theorem 3.3.4 (PTR (Dwork and Lei, 2009))**.** *Algorithm 3 satisfies $(2\epsilon, \delta)$-DP.*

Rather than proposing an arbitrary bound β on $\Delta_{LS}(X)$, one can also privately release an upper bound of the local sensitivity and calibrate noise according to this upper bound. This was used for node DP in graph statistics (Kasiviswanathan et al., 2013), and for fitting topic models using spectral methods (Decarolis et al., 2020).

3.4 Related Work

Data-dependent DP algorithms. Privately calibrating noise to the local sensitivity is a well-studied problem. One approach is to add noise calibrated to the smooth sensitivity (Nissim et al., 2007), an upper bound on the local sensitivity which changes slowly between

neighboring datasets. An alternative to this — and the focus of our work — is Propose-Test-Release (PTR) (Dwork and Lei, 2009), which works by calculating the distance $\mathcal{D}_\beta(X)$ to the nearest dataset to X whose local sensitivity violates a proposed bound β. The PTR algorithm then adds noise to $\mathcal{D}_\beta(X)$ before testing whether this privately computed distance is large enough to permit releasing the output with noise calibrated to β.

PTR spin-offs abound. Notable examples include stability-based methods (Thakurta and Smith, 2013) (stable local sensitivity of 0 near the input data) and privately releasing upper bounds of local sensitivity (Kasiviswanathan et al., 2013; Liu et al., 2021; Decarolis et al., 2020). We refer readers to Chapter 3 of Vadhan (2017) for a concise summary of these classic results. More recently, Wang et al. (2022) have provided Rényi DP bounds (Mironov, 2017) for PTR and demonstrated its applications to robust DP-SGD. Our work (Section 3.6.2) also considers applications of PTR in data-adaptive private deep learning: Instead of testing the local sensitivity of each gradient step as in Wang et al. (2022), our PTR-based PATE algorithm tests the data-dependent privacy loss as a whole.

Liu et al. (2021) proposed the High-dimensional Propose-Test-Release (HPTR) framework. HPTR provides a systematic way of solving DP statistical estimation problems by using the exponential mechanism (EM) with carefully constructed scores based on certain one-dimensional robust statistics, which have stable local sensitivity bounds. HPTR focuses on designing data-adaptive DP mechanisms from scratch; our method, in contrast, converts existing randomized algorithms (including EM and even some that do not satisfy DP) into those with formal DP guarantees. Interestingly, our proposed method also depends on a one-dimensional statistic of direct interest: the data-dependent privacy loss.

Data-dependent DP losses. The flip side of data-dependent DP algorithms is the study of data-dependent DP losses (Papernot et al., 2018a; Soria-Comas et al., 2017; Wang, 2017), which fix the randomized algorithm but parameterize the resulting privacy

loss by the specific input dataset. For example: In the simple mechanism that adds Laplace noise with parameter b, data-dependent DP losses are $\epsilon(X) = \Delta_{LS}(X)/b$. The data-dependent DP losses $\epsilon(X)$ are often much smaller than the DP loss ϵ, but they themselves depend on the data and thus may reveal sensitive information; algorithms satisfying a data-dependent privacy guarantee are not formally DP with guarantees any smaller than that of the worst-case. Existing work has considered privately publishing these data-dependent privacy losses (Papernot et al., 2018a; Redberg and Wang, 2021), but notice that privately publishing these losses does not improve the DP parameter of the given algorithm. Part of our contribution is to resolve this conundrum by showing that a simple post-processing step of the privately released upper bound of $\epsilon(X)$ gives a formal DP algorithm.

Private hyperparameter tuning. Our work has a nice connection with private hyperparameter tuning. Prior work (Liu and Talwar, 2019; Papernot and Steinke, 2021) requires each candidate configuration to be released with the same DP (or Rényi DP) parameter set. Another hidden assumption is that the parameters must not be privacy-correlated (i.e., parameter choice will not change the privacy guarantee). Otherwise we need to use the largest DP bound across all candidates. For example, Liu and Talwar (2019) show that if each mechanism (instantiated with one group of hyperparameters) is $(\epsilon, 0)$-DP, then running a random number of mechanisms and reporting the best option satisfies $(3\epsilon, 0)$-DP. Our work directly generalizes the above results by (1) considering a wide range of hyperparameters, either privacy-correlated or not; and (2) requiring only that individual candidates have a *testable* data-dependent DP.

3.5 Generalized PTR

This section introduces the generalized PTR framework. We first formalize the notion of *data-dependent* differential privacy that conditions on an input dataset X.

Definition 3.5.1 (Data-dependent privacy)**.** Suppose we have $\delta > 0$ and a function $\epsilon : \mathcal{X} \to \mathbb{R}^+$. We say that mechanism $\mathcal{M}$ satisfies $(\epsilon(X), \delta)$ data-dependent DP[2] for dataset X if for all possible output sets S and neighboring datasets X',

$$\Pr\big[\mathcal{M}(X) \in S\big] \leq e^{\epsilon(X)} \Pr\big[\mathcal{M}(X') \in S\big] + \delta,$$
$$\Pr\big[\mathcal{M}(X') \in S\big] \leq e^{\epsilon(X)} \Pr\big[\mathcal{M}(X) \in S\big] + \delta.$$

In generalized PTR, we propose a value (or set of values) ϕ with which to parameterize mechanism $\mathcal{M}_\phi$. For instance, in Example 3.5.5 we might propose $\phi = (\gamma, \lambda)$ as a parameter set that includes the noise scale and regularization strength. For a given δ, we then say that mechanism $\mathcal{M}_\phi$ satisfies $\epsilon_\phi(X)$ data-dependent DP for dataset X.

The following example illustrates how to derive the data-dependent DP for a familiar friend – the Laplace mechanism.

Example 3.5.2. *(*Data-dependent DP of Laplace Mechanism.*) Given a function $f : \mathcal{X} \to \mathbb{R}$, we will define*

$$\mathcal{M}_\phi(X) = f(X) + Lap\left(\phi\right).$$

We then have

$$\log \frac{\Pr[\mathcal{M}_\phi(X) = y]}{\Pr[\mathcal{M}_\phi(X') = y]} \leq \frac{|f(X) - f(X')|}{\phi}.$$

Maximizing over all possible outputs y yields an equality between the two expressions above.

[2]We will sometimes write that $\mathcal{M}(X)$ satisfies $\epsilon(X)$ data-dependent DP w.r.t. δ.

Using Definition 3.5.1,

$$\epsilon_\phi(X) = \max_{X':X\simeq X'} \frac{|f(X) - f(X')|}{\phi} = \frac{\Delta_{LS}(X)}{\phi}.$$

Maximizing $\epsilon_\phi(X)$ over X recovers the standard DP guarantee of running $\mathcal{M}$ with parameter ϕ.

Algorithm 4 distills the generalized PTR framework into a simple procedure: we run mechanism $\mathcal{M}$ with proposed parameter ϕ only if the test $\mathcal{T}$ "passes".

Let's suppose that our privacy budget for mechanism $\mathcal{M}_\phi$ is (ϵ, δ); that our test $\mathcal{T}$ satisfies $(\hat{\epsilon}, \hat{\delta})$-DP; and that $\mathcal{T}$ has a "false positive" rate δ', meaning $\mathcal{T}$ passes an insufficient proposal ϕ (where $\mathcal{M}_\phi$ exceeds its privacy budget) with probability at most δ'. Theorem 3.5.3 states the privacy guarantee of generalized PTR under these assumptions.

Algorithm 4 Generalized Propose-Test-Release

1: **Input**: Dataset X; mechanism $\mathcal{M}_\phi : \mathcal{X} \to \mathcal{R}$ and its privacy budget ϵ, δ; $(\hat{\epsilon}, \hat{\delta})$-DP test $\mathcal{T}$; false positive rate $\leq \delta'$; data-dependent DP function $\epsilon_\phi(\cdot)$ w.r.t. δ.
2: **if not** $\mathcal{T}(X)$ **then** output $\perp$,
3: **else** release $\theta = \mathcal{M}_\phi(X)$.

Theorem 3.5.3 (Privacy guarantee of generalized PTR)**.** *Consider a proposal ϕ and a data-dependent DP function $\epsilon_\phi(X)$ w.r.t. δ. Suppose that we have an $(\hat{\epsilon}, \hat{\delta})$-DP test $\mathcal{T} : \mathcal{X} \to \{0, 1\}$ such that when $\epsilon_\phi(X) > \epsilon$,*

$$\mathcal{T}(X) = \begin{cases} 0 & \textit{with probability } 1 - \delta', \\ 1 & \textit{with probability } \delta'. \end{cases}$$

Then Algorithm 4 satisfies $(\epsilon + \hat{\epsilon}, \delta + \hat{\delta} + \delta')$-DP.

Proof sketch. We can split the possible input datasets X into two main cases based on the data-dependent DP for a given δ: $\epsilon_\phi(X) > \epsilon$ and $\epsilon_\phi(X) \leq \epsilon$. At a high level, we can analyze both cases using the composition property of DP (that ϵ's and δ's "add up") and then combine them by taking an upper bound of the maximum value of the ϵ's and δ's between the two cases.

By the "false positive" assumption on the test $\mathcal{T}$, the first case can be viewed as a composition of an $(\hat{\epsilon}, \hat{\delta})$-DP mechanism and a $(0, \delta')$-DP mechanism. The second case, when the data-dependent DP is at most ϵ, is a composition of an $(\hat{\epsilon}, \hat{\delta})$-DP mechanism and an (ϵ, δ)-DP mechanism.

Full details of the proof are provided in the appendix.

□

Remark 3.5.4. The appendix (Section B.2.4) also includes an RDP (Mironov, 2017) analysis of Algorithm 4, where we demonstrate that by assuming a data-independent RDP bound of $\mathcal{M}_\phi$, it is possible to replace DP budgets and tests from Algorithm 4 with their RDP counterparts. The overall RDP guarantee can then be amplified by the false positive rate δ'.

Generalized PTR is a *strict* generalization of Propose-Test-Release. For some function f, define $\mathcal{M}_\phi$ and $\mathcal{T}$ as follows:

$$\mathcal{M}_\phi(X) = f(X) + \text{Lap}(\phi);$$

$$\mathcal{T}(X) = \begin{cases} 0 & \text{if } \ \mathcal{D}_\beta(X) + \text{Lap}\left(\frac{1}{\epsilon}\right) > \frac{\log(1/\delta)}{\epsilon}, \\ 1 & \text{otherwise.} \end{cases}$$

Notice that our choice of parameterization is now $\phi = \frac{\beta}{\epsilon}$, where ϕ is the scale of the

Laplace noise. In other words, we know from Example 3.5.2 that $\epsilon_\phi(X) > \epsilon$ exactly when $\Delta_{LS}(X) > \beta$.

For noise-adding mechanisms such as the Laplace mechanism, the sensitivity is proportional to the privacy loss in both the global and local sense: $\Delta_{GS} \propto \epsilon$ and $\Delta_{LS}(X) \propto \epsilon(X)$. Therefore for these mechanisms the only difference between privately testing the local sensitivity (Algorithm 3) and privately testing the data-dependent DP (Theorem 3.5.3) is a change of parameterization.

3.5.1 Limitations of local sensitivity

Why do we want to generalize PTR beyond noise-adding mechanisms? Compared to classic PTR, the generalized PTR framework allows us to be more flexible in both the type of test conducted and also the type of mechanism whose output we wish to release. For many mechanisms, the local sensitivity either does not exist or is only defined for specific data-dependent quantities (e.g., the sensitivity of the score function in the exponential mechanism) rather than the mechanism's output.

The following example illustrates this issue.

Example 3.5.5 (Private posterior sampling)**.** *Let $\mathcal{M} : \mathcal{X} \times \mathcal{Y} \to \Theta$ be a private posterior sampling mechanism (Minami et al., 2016; Wang et al., 2015; Gopi et al., 2022) for approximately minimizing $F_X(\theta)$.*

$\mathcal{M}$ samples $\theta \sim P(\theta) \propto e^{-\gamma(F_X(\theta)+\lambda/2||\theta||_2^2)}$ with parameters γ, λ. Note that γ, λ cannot be appropriately chosen for this mechanism to satisfy DP without calculating the sensitivity of $\arg\min F_X(\theta)$*, which in many cases (e.g., logistic regression) lacks a closed-form solution. In fact, the global and local sensitivity of the minimizer is unbounded even in linear regression problems, i.e when $F_X(\theta) = \frac{1}{2}||y - X\theta||_2^2$.*

Output perturbation algorithms do work for the above problem when we regularize,

but they are known to be suboptimal in theory and in practice (Chaudhuri et al., 2011). In Section 3.6.1 we demonstrate how to apply generalized PTR to achieve a data-adaptive posterior sampling mechanism.

Even in the cases of noise-adding mechanisms where PTR seems to be applicable, it does not lead to a tight privacy guarantee. Specifically, by an example of privacy amplification by post-processing (Example B.3.1 in the appendix), we demonstrate that the local sensitivity does not capture all sufficient statistics for data-dependent privacy analysis and thus is loose.

3.5.2 Which ϕ to propose

A limitation of generalized PTR (inherited from its predecessor) is that one needs to "propose" a good guess of parameter ϕ. Take the example of ϕ being the noise level in a noise-adding mechanism. Choosing too small a ϕ will result in a useless output $\perp$, while choosing too large a ϕ will add more noise than necessary. Finding this 'Goldilocks' ϕ might require trying out many different possibilities – each of which will consume privacy budget.

This section introduces a method to jointly tune privacy parameters (e.g., noise scale) along with parameters related only to the utility of an algorithm (e.g., learning rate or batch size in stochastic gradient descent) — while avoiding the $\perp$ output.

Algorithm 5 takes a list of parameters as input, runs generalized PTR with each of the parameters, and returns the output with the best utility. We show that the privacy guarantee with respect to ϵ is independent of the number of ϕ that we try.

Formally, let $\phi_1, ..., \phi_k$ be a set of hyperparameters and $\tilde{\theta}_i \in \{\perp, \text{Range}(\mathcal{M})\}$ the output of running generalized PTR with ϕ_i on dataset X. Let X_{val} be a public validation set and $q(\tilde{\theta}_i)$ be the score of evaluating $\tilde{\theta}_i$ with X_{val} (e.g., validation accuracy). The goal

is to select a pair $(\tilde{\theta}_i, \phi_i)$ such that DP model $\tilde{\theta}_i$ maximizes the validation score.

The generalized PTR framework with privacy calibration is described in Algorithm 5; its privacy guarantee is an application of Liu and Talwar (2019).

Algorithm 5 PTR with hyperparameter selection

1: **Input**: Privacy budget per PTR algorithm (ϵ^*, δ^*), cut-off T, parameters $\phi_{1:k}$, flipping probability τ and validation score function $q(\cdot)$.
2: Initialize the set $S = \varnothing$.
3: Draw G from a geometric distribution $\mathcal{D}_\tau$ and let $\hat{T} = \min(T, G)$.
4: **for** i = 1 ,..., $\hat{T}$ **do**
5: pick a random ϕ_i from $\phi_{1:k}$.
6: evaluate ϕ_i: $(\tilde{\theta}_i, q(\tilde{\theta}_i)) \leftarrow$ Algorithm 4$(\phi_i, (\epsilon^*, \delta^*))$.
7: $S \leftarrow S \cup \{\tilde{\theta}_i, q(\tilde{\theta}_i)\}$.
8: **end for**
9: Output the highest scored candidate from S.

Theorem 3.5.6 (Theorem 3.4 (Liu and Talwar, 2019))**.** *Fix any $\tau \in [0, 1], \delta_2 > 0$ and let $T = \frac{1}{\tau} \log \frac{1}{\delta_2}$. If each oracle access to Algorithm 4 is (ϵ^*, δ^*)-DP, then Algorithm 5 is $(3\epsilon^* + 3\sqrt{2\delta^*}, \sqrt{2\delta^*}T + \delta_2)$-DP.*

The theorem implies that one can try a random number of ϕ while paying a constant ϵ. In practice, we can roughly set $\tau = \frac{1}{10k}$ so that the algorithm is likely to test all k parameters. We emphasize that the privacy and the utility guarantee (stated in the appendix) is not our contribution. But the idea of applying generalized PTR to enforce a uniform DP guarantee over all choices of parameters with a data-dependent analysis is new.

In the appendix (Section B.2.3), we also show how to avoid hyperparameter selection by directly tuning (rather than proposing) ϕ using a uniform bound of $\epsilon_\phi(X)$. We use

this technique to tune γ in Example 3.6.2.

3.5.3 Construction of the DP test

Classic PTR uses the Laplace mechanism to construct a differentially private upper bound of $\mathcal{D}_\beta(X)$, the distance from input dataset X to the closest dataset whose local sensitivity exceeds the proposed bound β. The tail bound of the Laplace distribution then ensures that if $\mathcal{D}_\beta(X) = 0$ (that is, if $\Delta_{LS}(X) > \beta$), then the output will be released with only a small probability δ.

The following theorem shows that we could instead use a differentially private upper bound of the data-dependent DP $\epsilon_\phi(X)$ in order to test whether to run the mechanism $\mathcal{M}_\phi$.

Theorem 3.5.7 (Generalized PTR with private upper bound)**.** *Suppose we have a differentially private upper bound of $\epsilon_\phi(X)$ w.r.t. δ such that with probability at least $1 - \delta'$, $\epsilon_\phi^P(X) > \epsilon_\phi(X)$. Further suppose we have an $(\hat{\epsilon}, \hat{\delta})$-DP test $\mathcal{T}$ such that*

$$T(X) = \begin{cases} 1 & \text{if } \epsilon_\phi^P(X) < \epsilon, \\ 0 & \text{otherwise.} \end{cases}$$

Then Algorithm 4 is $(\epsilon + \hat{\epsilon}, \delta + \hat{\delta} + \delta')$-DP.

In Section 3.6.2, we demonstrate how to upper-bound the data-dependent DP through a modification of the smooth sensitivity framework applied on $\epsilon_\phi(X)$. In Section 3.6.1 we provide a direct application of Theorem 3.5.7 with private linear regression by making use of the per-instance DP technique (Wang, 2017).

The applications in Section 3.6 are illustrative of two distinct approaches to constructing the DP test for generalized PTR:

1. Private sufficient statistics release (used in the private linear regression example of Section 3.6.1) specifies the data-dependent DP as a function of the dataset and privately releases each data-dependent component.

2. The second approach (used in the PATE example of Section 3.6.2) uses the smooth sensitivity framework to privately release the data-dependent DP as a whole, and then construct a high-confidence test using the Gaussian mechanism.

These two flavors cover most of the scenarios arising in data-adaptive analysis. For example, in the appendix we demonstrate the merits of generalized PTR in handling data-adaptive private generalized linear models (GLMs) using private sufficient statistics release. Moreover, sufficient statistics release together with our private hyperparameter tuning (Algorithm 5) can be used to construct data-adaptive extensions of DP-PCA and Sparse-DP-ERM (see details in the future work section).

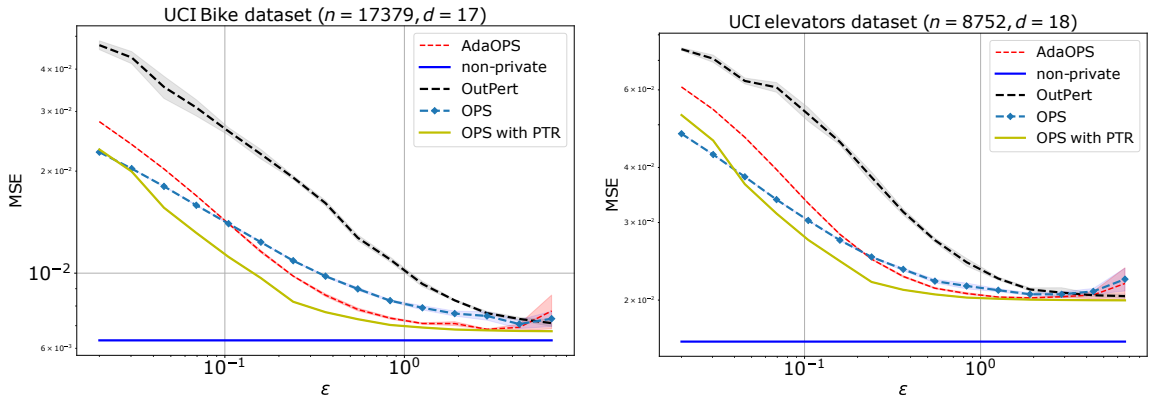
3.6 Applications

In this section, we put into action our approaches to construct the DP test and provide applications in private linear regression and PATE.

3.6.1 Private Linear Regression

Theorem 3.6.1 ((Wang, 2017)). *For input data $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, define the following:*

- $\lambda_{\min}(X)$ *denotes the smallest eigenvalue of* $X^T X$*;*
- $||\theta^*_\lambda||$ *is the magnitude of the solution* $\theta^*_\lambda = (X^T X + \lambda I)^{-1} X^T Y$*;*
- *and* $L(X, Y) := ||\mathcal{X}||\big(||\mathcal{X}||\,||\theta^*_\lambda|| + ||\mathcal{Y}||\big)$ *is the local Lipschitz constant, denoted* L *in brief.*



(a) Bike dataset (b) Elevator dataset

Figure 3.1: Differentially private linear regression algorithms on UCI datasets. y-axis reports the MSE error with confidence intervals. ϵ is evaluated with $\delta = 1e^{-6}$.

For brevity, denote $\lambda^* = \lambda + \lambda_{\min}(X)$. *The algorithm used in Example 3.5.5 with proposal* $\phi = (\lambda, \gamma)$ *obeys* $(\epsilon_\phi(Z), \delta)$ *data-dependent DP for each dataset* $Z = (X, Y)$ *with* $\epsilon_\phi(Z)$ *equal to*

$$\sqrt{\frac{\gamma L^2 \log(2/\delta)}{\lambda^*}} + \frac{\gamma L^2}{2(\lambda^* + ||\mathcal{X}||^2)} + \frac{1 + \log(2/\delta)||\mathcal{X}||^2}{2\lambda^*}.$$

Notice that $\epsilon_\phi(Z)$ is a function of the data-dependent quantities $\lambda_{\min}(X)$ and L (which is itself a function of $||\theta^*_\lambda||$). Could we privately release $\epsilon_\phi(Z)$ and tune the privacy parameters $\phi = (\lambda, \gamma)$ based on the sanitized data-dependent DP? Unfortunately in this case, $||\theta^*_\lambda||$ is a complicated function of λ and it is not clear how to choose an optimal λ.

The calibration of γ, however, is fairly straightforward from the expression for $\epsilon_\phi(Z)$ given in Theorem 3.6.1. We can apply the generalized PTR framework to the private posterior sampling problem described in Example 3.5.5 by proposing $\phi = \lambda$ as the regularization parameter; releasing a high-probability upper bound $\epsilon^P_\lambda(Z)$ of the data-dependent DP, as a function of γ; and tuning the noise scale γ to achieve the desired utility under the constraint $\epsilon^P_\lambda(Z) \leq \epsilon$.

Example 3.6.2 (OPS for linear regression with PTR)**.** *Consider the posterior sampling*

mechanism described in Example 3.5.5 and the expression $\epsilon_\phi(Z)$ given in Theorem 3.6.1. Suppose we have a quality score $q(\cdot)$ that measures the utility of the input parameter, e.g. $q(\gamma) = \gamma$ for the inverse noise scale. We can apply generalized PTR as follows.

- *Given a proposed value $\phi = \lambda$, privately release $\lambda_{\min}(X)$ and L with combined privacy budget $(\hat{\epsilon}, \hat{\delta})$ in order to obtain $\epsilon_\lambda^P(Z)$ such that with probability $1-\delta'$, $\epsilon_\lambda^P(Z) \leq \epsilon_\lambda(Z)$.*
- *Calibrate $\gamma^* = \sup_{q(\gamma)}\{\gamma \mid \epsilon_\lambda^P(Z) \leq \epsilon\}$.*
- *Output $\theta \sim e^{-\frac{\gamma^*}{2}\left(||Y-X\theta||_2^2+\lambda||\theta||_2^2\right)}$ if γ^* exists; else output $\perp$.*

In the appendix, we provide full details of the above algorithm and show that it satisfies $(\epsilon + \hat{\epsilon}, \delta + \hat{\delta} + \delta')$-DP.

The main idea of the above algorithm boils down to privately releasing all data-dependent quantities in data-dependent DP, constructing high-probability confidence intervals of these quantities, and then deciding whether to run the mechanism $\mathcal{M}$ with the proposed parameters. In Example 3.5.5, we need only propose λ as we can tune γ directly based on $\epsilon_\lambda^P(Z)$.

Remark 3.6.3. Tuning λ is even more troublesome for generalized linear models (GLMs) beyond linear regression. The data-dependent DP there involves a local strong-convexity parameter that is a complex function of the regularizer λ and for which we only have zeroth-order access. In the appendix, we demonstrate how to apply generalized PTR to provide a generic solution to a family of private GLMs where the link function satisfies a self-concordance assumption.

We next apply Algorithm 5 for Example 3.6.2 with UCI regression datasets. Standard z-scoring is applied and each data point is normalized with a Euclidean norm of 1. We consider $(60\%, 10\%, 30\%)$ splits for the train, validation and test sets.

Baselines

- Output Perturbation (Outpert) (Chaudhuri et al., 2011): $\theta = (X^T X + \lambda I)^{-1} X^T \mathbf{y}$. Release $\hat{\theta} = \theta + \mathbf{b}$ with an appropriate λ, where $\mathbf{b}$ is a Gaussian random vector.
- Posterior sampling (OPS). Sample $\hat{\theta} \sim P(\theta) \propto e^{-\gamma(F(\theta)+0.5\lambda||\theta||^2)}$ with parameters γ, λ.
- Adaptive posterior sampling (AdaOPS) (Wang, 2018). Run OPS with (λ, γ) chosen adaptively according to the dataset.

Outpert and OPS serve as two non-adaptive baselines. In particular, we consider OPS-Balanced (Wang, 2018), which chooses λ to minimize a data-independent upper bound of empirical risk and dominates other OPS variants. AdaOPS is one state-of-the-art algorithm for adaptive private regression, which automatically chooses λ by minimizing an upper bound of the data-dependent empirical risk.

We implement OPS-PTR as follows: propose a list of λ through grid search (we choose $k = 30$ and λ ranges from $[2.5, 2.5^{10}]$ on a logarithmic scale); instantiate Algorithm 5 with $\tau = 0.05/k$, $T = \frac{1}{\tau}\log(1/\delta_2)$ and $\delta_2 = 1/2\delta$; calibrate the per-PTR privacy budget (ϵ^*, δ^*) according to Theorem 3.5.6; set $\epsilon = \hat{\epsilon} = 0.5\epsilon^*$ and $\delta = 1/6\delta^*, \delta' = 1/2\delta^*, \hat{\delta} = 1/3\delta^*$; calibrate γ to meet the privacy requirement for each λ; sample $\hat{\theta}$ using (λ, γ) and return the one with the best validation accuracy.

Figure 3.1 demonstrates how the MSE error of the linear regression algorithms varies with the privacy budget ϵ. OutPert suffers from the large global sensitivity of output θ. OPS performs well but does not benefit from the data-dependent quantities. AdaOPS is able to adaptively choose (λ, γ) based on the dataset, but suffers from the estimation error of the data-dependent empirical risk. On the other hand, OPS-PTR selects a (λ, γ) pair that minimizes the empirical error on the validation set directly, and the privacy parameter γ adapts to the dataset thus achieving the best result.

3.6.2 PATE

In this section, we apply generalized PTR to solve an open problem from Private Aggregation of Teacher Ensembles (PATE) (Papernot et al., 2017, 2018a) — privately publishing the entire model through sanitizing the data-dependent DP losses. Our algorithm uses of smooth sensitivity (Nissim et al., 2007) and the Gaussian mechanism to construct a high-probability test of the data-dependent DP. Data-dependent DP is one-dimensional, enabling efficient computation under the smooth sensitivity framework. This approach is thus generally applicable for private data-adaptive analyses beyond PATE.

PATE is a knowledge transfer framework for model-agnostic private learning. In this framework, an ensemble of teacher models is trained on the disjoint private data and uses the teachers' aggregated consensus answers to supervise the training of a "student" model agnostic to the underlying machine-learning algorithms. By publishing only the aggregated answers and by the careful analysis of the "consensus", PATE has become a practical technique in recent private model training.

The tight privacy guarantee of PATE heavily relies on a delicate data-dependent DP analysis, for which the authors of PATE use the smooth sensitivity framework to privately publish the data-dependent privacy cost. However, it remains an open problem to show that the released model is DP under data-dependent analysis. Our generalized PTR resolves this gap by carefully testing a private upper bound of the data-dependent privacy cost. Our algorithm is fully described in Algorithm 6, where the modification over the original PATE framework is highlighted in blue.

Algorithm 6 takes the input of privacy budget $(\epsilon', \hat{\epsilon}, \delta)$, unlabeled public data $x_{1:T}$ and K teachers' predictions on these data. The parameter ϵ denotes the privacy cost of publishing the data-dependent DP and ϵ' is the predefined privacy budget for testing.

Algorithm 6 PATE with generalized PTR

Input: Unlabeled public data $x_{1:T}$, aggregated teachers prediction $n(\cdot)$, privacy parameter $\hat{\epsilon}, \epsilon', \delta$, noisy parameter σ_1.
Set $\alpha = \frac{2\log(2/\delta)}{\hat{\epsilon}} + 1$, $\sigma_s = \sigma_2 = \sqrt{\frac{3\alpha+2}{\hat{\epsilon}}}, \delta_2 = \delta/2$, smoothness parameter $\beta = \frac{0.2}{\alpha}$.
Compute noisy labels: $y_i{}^p \leftarrow \text{argmax}_{j\in[C]}\{n_j(x_i) + \mathcal{N}(0, \sigma_1^2)\}$ for all $i \in [1:T]$.
${}_{\sigma_1}(\alpha, X) \leftarrow$ data-dependent RDP at the α-th order.
$SS_\beta(X) \leftarrow$ the smooth sensitivity of ${}^{\text{upper}}_{\sigma_1}(\alpha, X)$.
Privately release $\mu := \log(SS_\beta(X)) + \beta \cdot \mathcal{N}(0, \sigma_2^2) + \sqrt{2\log(2/\delta_2)} \cdot \sigma_2 \cdot \beta$
${}^{\text{upper}}_{\sigma_1}(\alpha) \leftarrow$ an upper bound of data-dependent RDP through Lemma 3.6.5.
$\epsilon_{\sigma_1} \leftarrow$ DP guarantee converted from ${}^{\text{upper}}_{\sigma_1}(\alpha)$.
If $\epsilon' \geq \epsilon_{\sigma_1}$ **return** a student model trained using $(x_{1:T}; y^p_{1:T})$.
Else return $\perp$.

$n_j(x_i)$ denotes the the number of teachers that agree on label j for x_i and C denotes the number of classes. The goal is to privately release a list of plurality outcomes — $\text{argmax}_{j\in[C]} n_j(x_i)$ for $i \in [T]$ — and use these outcomes to supervise the training of a "student" model in the public domain. The parameter σ_1 denotes the noise scale for the vote count.

In their privacy analysis, Papernot et al. (2018a) compute the data-dependent ${}_{\sigma_1}(\alpha, X)$ of labeling the entire group of student queries. ${}_{\sigma_1}(\alpha, X)$ can be orders of magnitude smaller than its data-independent version if there is a strong agreement among teachers. Note that ${}_{\sigma_1}(\alpha, X)$ is a function of the RDP order α and the dataset X, analogous to our Definition 3.5.1 but subject to RDP (Mironov, 2017).

Theorem 3.6.4 ((Papernot et al., 2018a)). *If the top three vote counts of x_i are $n_1 > n_2 > n_3$ and $n_1 - n_2, n_2 - n_3 \gg \sigma_1$, then the data-dependent RDP of releasing $argmax_j\{n_j + \mathcal{N}(0, \sigma_1^2)\}$ satisfies $(\alpha, \exp\{-2\alpha/\sigma_1^2\}/\alpha)$-RDP and the data-independent RDP (using the Gaussian mechanism) satisfies $(\alpha, \frac{\alpha}{\sigma_1^2})$-RDP.*

However, ${}_{\sigma_1}(\alpha, X)$ is data-dependent and thus cannot be revealed. The authors therefore privately publish the data-dependent RDP using the smooth sensitivity framework (Nissim et al., 2007). The smooth sensitivity calculates a smooth upper bound

on the local sensitivity of ${}_{\sigma_1}(\alpha, X)$, denoted as $SS_\beta(X)$, such that $SS_\beta(X) \leq e^\beta SS_\beta(X')$ for any neighboring dataset X and X'. By adding Gaussian noise scaled by the smooth sensitivity (i.e., releasing $\epsilon_{\sigma_1}(\alpha, X) + SS_\beta(X) \cdot \mathcal{N}(0, \sigma_s^2)$), the privacy cost can be safely published.

Unlike most noise-adding mechanisms, the standard deviation σ_s cannot be published since $SS_\beta(X)$ is a data-dependent quantity. Moreover, this approach fails to provide a valid privacy guarantee of the noisy labels obtained through the PATE algorithm, as the published privacy cost could be smaller than the real privacy cost. Our solution in Algorithm 6 looks like the following:

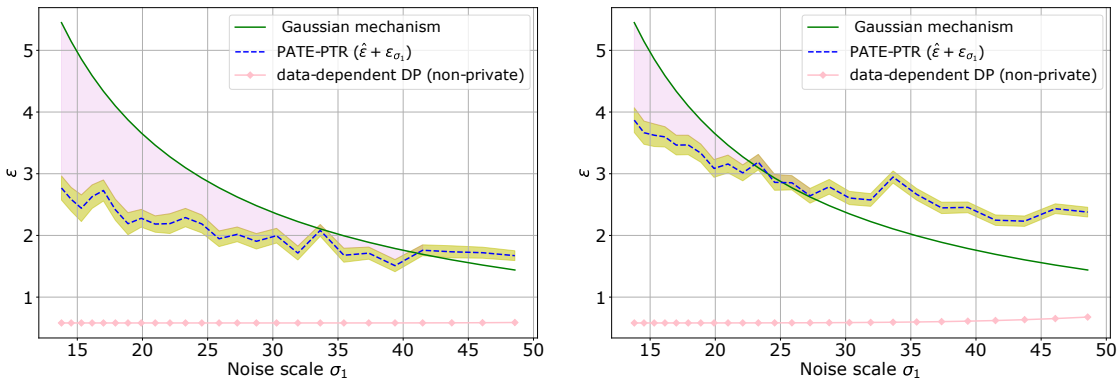
- Privately release an upper bound of the smooth sensitivity $SS_\beta(X)$ with e^μ.
- Conditioned on a high-probability event of e^μ, publish the data-dependent RDP with ${}^{\text{upper}}_{\sigma_1}(\alpha)$.
- Convert ${}^{\text{upper}}_{\sigma_1}(\alpha)$ back to the standard DP guarantee using RDP to DP conversion at $\delta/2$.
- Test if the converted DP is above the predefined budget ϵ'.

The following lemma states that ${}^{\text{upper}}_{\sigma_1}(\alpha)$ is a valid upper bound of the data-dependent RDP.

Lemma 3.6.5 (Private upper bound of data-dependent RDP)**.** *We are given a RDP function (α, X) and a β-smooth sensitivity bound $SS(\cdot)$ of (α, X). Let μ (defined in Algorithm 6) denote the private release of $\log(SS_\beta(X))$. Let the $(\beta, \sigma_s, \sigma_2)$-GNSS mechanism be*

$$^{upper}(\alpha) := (\alpha, X) + SS_\beta(X) \cdot \mathcal{N}(0, \sigma_s^2) + \sigma_s \sqrt{2\log(\tfrac{2}{\delta_2})} e^\mu$$

Then, the release of $^{upper}(X)$ satisfies $(\alpha, \frac{3\alpha+2}{2\sigma_s^2})$-RDP for all $1 < \alpha < \frac{1}{2\beta}$; w.p. at least $1 - \delta_2$, $^{upper}(\alpha)$ is an upper bound of (α, X).



(a) High consensus and strong data-dependent DP

(b) Low consensus and low data-dependent DP

Figure 3.2: Privacy and utility tradeoffs with PATE. When σ_1 is aligned, the three algorithms provide the same utility if the privacy budget of PATE-PTR is chosen from the purple region. y-axis plots the privacy cost of labeling $T = 200$ public data with $\delta = 10^{-5}$. The left figure considers the high-consensus case, where the data-adaptive analysis is preferred.

The proof (deferred to the appendix) makes use of the facts that: (1) the log of $SS_\beta(X)$ has a bounded global sensitivity β through the definition of smooth sensitivity; (2) releasing ${}_{\sigma_1}(\alpha, X) + SS_\beta(X) \cdot \mathcal{N}(0, \sigma_s^2)$ is $(\alpha, \frac{\alpha+1}{\sigma_s^2})$-RDP (Theorem 23 from Papernot et al. (2018a)).

Now we can state the privacy guarantee of Algorithm 6.

Theorem 3.6.6. *Algorithm 6 satisfies $(\epsilon' + \hat{\epsilon}, \delta)$-DP.*

In the proof, the choice of α ensures that the cost of the $\delta/2$ contribution (used in the RDP-to-DP conversion) is roughly $\hat{\epsilon}/2$. Then the release of ${}^{\text{upper}}_{\sigma_1}(\alpha)$ with $\sigma_s = \sqrt{\frac{2+3\alpha}{\hat{\epsilon}}}$ accounts for another cost of $(\epsilon/2, \delta/2)$-DP.

Empirical results. We next empirically evaluate Algorithm 6 (PATE-PTR) on the MNIST dataset. Following the experimental setup from Papernot et al. (2018a), we consider the training set to be the private domain, and the testing set is used as the public domain. We first partition the training set into 400 disjoint sets and 400 teacher models,

each trained individually. Then we select $T = 200$ unlabeled data from the public domain, with the goal of privately labeling them. To illustrate the behaviors of algorithms under various data distributions, we consider two settings of unlabeled data, high-consensus and low-consensus. In the low-consensus setting, we choose T unlabeled data such that there is no high agreement among teachers, so the advantage of data-adaptive analysis is diminished. We provide further details on the distribution of these two settings in the appendix.

Baselines. We consider the Gaussian mechanism as a data-independent baseline, where the privacy guarantee is valid but does not take advantage of the properties of the dataset. The data-dependent DP (Papernot et al. (2018a)) serves as a non-private baseline, which requires further sanitation. Note that these two baselines provide different privacy analyses of the same algorithm (see Theorem 3.6.4).

Figure 3.2 plots privacy-utility tradeoffs between the three approaches by varying the noise scale σ_1. The purple region denotes a set of privacy budget choices ($\hat{\epsilon} + \epsilon'$ used in Algorithm 6) such that the utility of the three algorithms is aligned under the same σ_1. In more detail, the purple region is lower-bounded by $\hat{\epsilon} + \epsilon_{\sigma_1}$. We first fix $\sigma_s = \sigma_2 = 15$ such that $\hat{\epsilon}$ is fixed. Then we empirically calculate the average of ϵ_{σ_1} (the private upper bound of the data-dependent DP) over 10 trials. Running Algorithm 6 with any choice of $\hat{\epsilon} + \epsilon'$ chosen from the purple region implies $\epsilon' > \epsilon_{\sigma_1}$. Therefore, PATE-PTR will output the same noisy labels (with high probability) as the two baselines.

Observation As σ_1 increases, the privacy loss of the Gaussian mechanism decreases, while the data-dependent DP curve does not change much. This is because the data-dependent DP of each query is a complex function of both the noise scale and the data and does not monotonically decrease when σ_1 increases (see more details in the appendix). However, the data-dependent DP still dominates the Gaussian mechanism for a wide range of σ_1. Moreover, PATE-PTR nicely interpolates between the data-independent DP

guarantee and the non-private data-adaptive DP guarantee. In the low-consensus case, the gap between the data-dependent DP and the DP guarantee of the Gaussian mechanism unsurprisingly decreases. Meanwhile, PATE-PTR (the purple region) performs well when the noise scale is small but deteriorates when the data-independent approach proves more advantageous. This example demonstrates that using PTR as a post-processing step to convert the data-dependent DP to standard DP is effective when the data-adaptive approach dominates others.

3.7 Limitations and Future Work

One weakness of generalized PTR is that it requires a case-specific privacy analysis. Have we simply exchanged the problem of designing a data-adaptive DP algorithm with the problem of analyzing the data-dependent privacy loss? We argue that this limitation is inherited from classic PTR. In situations where classic PTR is not applicable, we've outlined several approaches to constructing the DP test for our framework (see Sections 3.5.3 and 3.6.2).

Furthermore, the data-dependent privacy loss is often more straightforward to compute than local sensitivity, and often exists in intermediate steps of classic DP analysis already. Most DP analysis involves providing a high-probability tail bound of the privacy loss random variable. If we stop before taking the max over the input dataset, then we get a data-dependent DP loss right away (as in Example 3.5.2).

There are several exciting directions for applying generalized PTR to more problems. Sufficient statistics release with private hyperparameter tuning can be used to construct data-adaptive extensions of DP-PCA (Dwork et al., 2014b) and Sparse-DP-ERM (Kifer et al., 2012). For DP-PCA we could use Algorithm 5 to tune the variance of the noise added to the spectral gap; for Sparse-DP-ERM we would test the restricted strong

convexity parameter (RSC) and not add additional regularization if the RSC is already large.

3.8 Conclusion

Generalized PTR extends the classic "Propose-Test-Release" framework to a more general setting by testing the data-dependent privacy loss of an input dataset, rather than its local sensitivity. In this paper we've provided several examples – private linear regression with hyperparameter selection and PATE – to illustrate how generalized PTR can enhance DP algorithm design via a data-adaptive approach.

Chapter 4

Improving the Privacy and Practicality of Objective Perturbation

4.1 Introduction

The rise of deep neural networks has transformed the study of differentially private learning no less than any other area of machine learning. Differentially private stochastic gradient descent (DP-SGD) (Song et al., 2013; Bassily et al., 2014; Abadi et al., 2016) has thus gained widespread appeal as a versatile framework for privately training deep learning models.

How does DP-SGD fare on simpler models such as linear and logistic regression? The verdict is unclear. Clearly an algorithm capable of privately optimizing non-convex functions represented by millions of parameters is up to the computational task of fitting a linear model. A more pressing concern is that DP-SGD is up to *too* much. Look, for example, at the algorithm's computational complexity: DP-SGD requires $O(n^2)$ steps to achieve the optimal excess risk bounds for DP convex empirical risk minimization (Bassily et al., 2014).

DP-SGD furthermore takes after its non-private counterpart in sensitivity to hyperparameters. A poor choice of learning rate or batch size, for instance, could lead to suboptimal performance or slow convergence. There are well-established procedures for hyperparameter optimization that typically involve evaluating the performance of the model trained using different sets of candidate hyperparameters. But with privacy constraints, there is a catch: tuning hyperparameters requires multiple passes over the training dataset and thereby constitutes a privacy cost.

At best, existing work tends to circumvent this obstacle by optimistically assuming the availability of a public auxiliary dataset for hyperparameter tuning. More often the procedure for private hyperparameter selection is left largely to the reader's imagination. Only recently have Liu and Talwar (2019) and subsequently Papernot and Steinke (2021) studied how to obtain tighter privacy loss bounds for this task beyond standard composition theorems.

In the meantime, objective perturbation (Chaudhuri et al., 2011; Kifer et al., 2012) has been to some extent shelved as a historical curiosity. Sifting through the literature, we find that opinions are divided: some tout objective perturbation as "[o]ne of the most effective algorithms for differentially private learning and optimization" (Neel et al., 2020), whereas other works (Wang et al., 2017) dismiss objective perturbation as being impractical and restrictive. Some empirical evaluations (Yu et al., 2019; McKenna et al., 2021) suggest that DP-SGD often achieves better utility in practice than does objective perturbation; others (Iyengar et al., 2019) report the opposite.

Our goal in this chapter is to lay some of this debate to rest and demonstrate that for generalized linear problems in particular, objective perturbation can outshine DP-SGD.

4.1.1 Our Contributions

- **We establish an improved (ϵ, δ)-DP bound for objective perturbation via privacy profiles, a modern tool for privacy accounting that bounds the hockey-stick divergence.** The formula can be computed numerically using only calls to Gaussian CDFs. We further obtain a *dominating pair* of distributions as defined by Zhu et al. (2022) which enables tight composition and amplification by subsampling of the privacy profiles.

- **We present a novel Rényi differential privacy (RDP) (Mironov, 2017) analysis of the objective perturbation mechanism. Using this analysis, we show empirically that objective perturbation performs competitively against DP-SGD with "honestly" [1]-tuned hyperparameters.** The tightest analyses to date of private hyperparameter tuning are the RDP bounds derived in Papernot and Steinke (2021). This tool allows us to empirically evaluate objective perturbation against DP-SGD on a level playing field (Section 4.5).

- **We fix a decade-old oversight in the privacy analysis of objective perturbation.** Existing literature overlooks a nuanced argument in the privacy analysis of objective perturbation, which requires a careful treatment of the dependence between the noise vector and the private minimizer. Without assuming GLM structure, the privacy bound of objective perturbation is subject to a dimensional dependence that has gone unacknowledged in previous work[2].

- **We introduce computational tools that expand the applicability of objective perturbation to a broader range of loss functions.** The privacy guarantees of objective perturbation require the loss function to have bounded

[1] "Honest" hyperparameter tuning is a term coined by Mohapatra et al. (2022).
[2] The concurrent work of Agarwal et al. (2023) has independently identified this bug as well.

gradient. Our proposed framework extends the Approximate Minima Perturbation framework of Iyengar et al. (2019) to take any smooth loss function as a black-box, then algorithmically ensure that it has bounded gradient. We also provide a **computational guarantee** $O(n \log n)$ on the running time of this algorithm, in contrast to the $O(n^2)$ complexity of DP-SGD for achieving information-theoretic limits.

4.1.2 A Short History of DP Learning

Differentially private learning dates back to Chaudhuri et al. (2011), which extended the output perturbation method of Dwork et al. (2006) to classification algorithms and also introduced objective perturbation. In its first public appearance, objective perturbation required gamma-distributed noise; Kifer et al. (2012) provided a refined analysis of the mechanism with Gaussian noise, which is the entry point into our work.

Differentially private stochastic gradient descent (DP-SGD) (Song et al., 2013; Bassily et al., 2014; Abadi et al., 2016) brought DP into the fold of modern machine learning, allowing private training of models with arbitrarily complex loss landscapes that can scale to enormous datasets. DP-SGD adds Gaussian noise at every iteration to an aggregation of clipped gradients, and thus privacy analysis for DP-SGD often boils down to finding tight composition bounds (of the subsampled Gaussian mechanism).

The initial version of DP-SGD based on standard strong composition (Bassily et al., 2014) is not quite practical; but that has changed, thanks to a community-wide effort over the past few years in developing modern numerical privacy accounting tools. These include the moments accountant which composes Rényi DP functions (Abadi et al., 2016; Wang et al., 2019; Mironov et al., 2019) and the Fourier accountant (also known as PLV or PLD accountant) which directly composes the *privacy profile* (Sommer et al., 2019;

Koskela et al., 2020; Gopi et al., 2021; Zhu et al., 2022) of a mechanism. It is safe to conclude that the numerically computed privacy loss of DP-SGD using these modern tools is now very precise.

Because DP-SGD releases each intermediate model, the algorithm can stop after any number of iterations and simply accumulates privacy loss as it goes. In contrast, the privacy guarantees of objective perturbation hold only when the output of the mechanism is the *exact* minima of the perturbed objective. This requirement is at odds with practical convex optimization frameworks which typically use first-order methods to search for an approximate solution.

To remedy this, Iyengar et al. (2019) proposed an approach to *approximately* minimize a perturbed objective function while maintaining privacy. Approximate Minima Perturbation (AMP) was introduced as a tractable alternative to objective perturbation whose privacy guarantees permit the output to be an approximate (rather than exact) solution to the perturbed minimization problem. In this paper we extend AMP to a broader range of loss functions, whose gradient can be unbounded; Algorithm 7 can be viewed as a special case of AMP with a transformation of the loss function.

4.2 Preliminaries

4.2.1 Differential Privacy

Differential privacy (DP) (Dwork et al., 2006) offers provable privacy protection by restricting how much the output of a randomized algorithm can leak information about a single data point.

DP requires a notion of how to measure similarity between datasets. We say that datasets Z and Z' are neighboring datasets (denoted $Z \simeq Z'$) if they differ by exactly

one datapoint z, i.e. $Z' = Z \cup \{z\}$ or $Z' = Z \setminus \{z\}$ for some data entry z.

Definition 4.2.1 (Differential privacy)**.** A mechanism $\mathcal{M} : \mathcal{Z} \to \mathcal{R}$ satisfies (ϵ, δ)-differential privacy if for all neighboring datasets $Z, Z' \in \mathcal{Z}$ and output sets $S \subseteq \mathcal{R}$,

$$\Pr\left[\mathcal{M}(Z) \in S\right] \leq e^{\epsilon}\Pr\left[\mathcal{M}(Z') \in S\right] + \delta.$$

When $\delta > 0$, $\mathcal{M}$ satisfies *approximate DP*. When $\delta = 0$, $\mathcal{M}$ satisfies the stronger notion of *pure DP*.

We say that $\mathcal{M}$ is tightly (ϵ, δ)-DP if there is no $\delta' < \delta$ for which $\mathcal{M}$ would be (ϵ, δ')-DP.

In what follows, we overview two different styles of achieving DP guarantees: one via hockey-stick divergence, and the other via Rényi divergence.

DP via hockey-stick divergence

Definition 4.2.2 (Hockey-stick divergence)**.** Denote $[x]_+ = \max\{0, x\}$ for $x \in \mathbb{R}$. For $\alpha > 0$ the hockey-stick divergence H_α from a distribution P to a distribution Q is defined as

$$H_\alpha(P||Q) = \int \left[P(x) - \alpha \cdot Q(x)\right]_+ \, \mathrm{d}x.$$

Now (with some abuse of notation) we will discuss how to bound the hockey-stick divergence between distributions $\mathcal{M}(Z)$ and $\mathcal{M}(Z')$ via the concept of *privacy profiles*.

Definition 4.2.3 (Privacy profiles Balle et al., 2018)**.** The privacy profile $\delta_{\mathcal{M}}(\epsilon)$ of a mechanism $\mathcal{M}$ is defined as

$$\delta_{\mathcal{M}}(\epsilon) := \max_{Z \simeq Z'} H_{\mathrm{e}^\epsilon}\big(\mathcal{M}(Z)||\mathcal{M}(Z')\big).$$

Tight (ϵ, δ)-DP bounds can then be obtained as follows.

Lemma 4.2.4 (Zhu et al., 2022, Lemma 5)**.** *Mechanism $\mathcal{M}$ satisfies (ϵ, δ)-DP* if any only if $\delta \geq \delta_{\mathcal{M}}(\epsilon)$.

Dominating pairs of distributions are useful for bounding the hockey-stick divergence $H_{\mathrm{e}^\epsilon}\big(\mathcal{M}(Z)||\mathcal{M}(Z')\big)$ accurately and, in particular, for obtaining tight bounds for compositions.

Definition 4.2.5 (Zhu et al. 2022)**.** A pair of distributions (P, Q) is a *dominating pair* of distributions for mechanism $\mathcal{M} : \mathcal{Z} \to \mathcal{R}$ if for all neighboring datasets Z and Z' and for all $\alpha > 0$,

$$H_\alpha(\mathcal{M}(Z)||\mathcal{M}(Z')) \leq H_\alpha(P||Q).$$

DP via Rényi divergence

Definition 4.2.6. (Rényi divergence.) Let $\alpha > 0$. For $\alpha \neq 1$, the Rényi divergence D_α from distribution P to distribution Q is defined as

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log \mathbb{E}_{x\sim Q}\left[\left(\frac{P(x)}{Q(x)}\right)^\alpha\right].$$

When $\alpha = 1$, Rényi divergence reduces to the Kullback–Leibler (KL) divergence:

$$D_1(P||Q) = \mathbb{E}_{x\sim P}\left[\log\left(\frac{P(x)}{Q(x)}\right)\right].$$

Rényi differential privacy (RDP) is a relaxation of pure DP ($\delta = 0$) based on Rényi divergence.

Definition 4.2.7 (Rényi differential privacy)**.** A mechanism $\mathcal{M} : \mathcal{Z} \to \mathcal{R}$ satisfies

(α, ϵ)-Rényi differential privacy if for all neighboring datasets $Z, Z' \in \mathcal{Z}$,

$$D_\alpha\big(\mathcal{M}(Z) \,||\, \mathcal{M}(Z')\big) \leq \epsilon,$$

RDP implies (ϵ, δ)-DP for any $0 < \delta \leq 1$ with $\epsilon = \min_{\alpha>1}\{\epsilon(\alpha) + \log(1/\delta)/(\alpha - 1)\}$. Tighter but more complex conversion formulae were derived by Balle et al. (2020) and Canonne et al. (2020), which we adopt numerically in our experiments whenever approximate DP is needed.

4.2.2 Differentially Private Empirical Risk Minimization

Given a dataset $Z \in \mathcal{Z}$ and a loss function $\ell(\theta; z)$, we want to solve problems of the form

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \sum_{z\in Z} \ell(\theta; z) + r(\theta),$$

where $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$ is a data point and $r(\theta)$ is a regularization term. The feature space is $\mathcal{X} \subseteq \mathbb{R}^d$ and the label space is $\mathcal{Y} \subseteq \mathbb{R}$. We will assume that $||x||_2 \leq 1$ and $|y| \leq 1$.

This work focuses on unconstrained convex generalized linear models (GLMs): we require that $\ell(\theta)$ and $r(\theta)$ are convex and twice-differentiable and that $\Theta = \mathbb{R}^d$. The loss function is assumed to have GLM structure of the form $\ell(\theta; z) = f(x^T\theta; y)$.

Objective Perturbation Construct the perturbed objective function by sampling $b \sim \mathcal{N}(0, \sigma^2 I_d)$:

$$\mathcal{L}^P(\theta; Z, b) = \sum_{z\in Z} \ell(\theta; z) + \frac{\lambda}{2}||\theta||_2^2 + b^T\theta.$$

The objective perturbation mechanism (ObjPert) outputs $\hat{\theta}^P(Z) = \arg\min_{\theta\in\Theta} \mathcal{L}^P(\theta; Z, b)$.

Theorem 4.2.8 (DP guarantees of objective perturbation (Kifer et al., 2012)). *Let $\ell(\theta; z)$ be convex and twice-differentiable such that $||\nabla \ell(\theta; z)||_2 \leq L$ and $\nabla^2 \ell(\theta; z) \prec \beta I_d$ for all $\theta \in \Theta$ and $z \in \mathcal{X} \times \mathcal{Y}$. Then objective perturbation satisfies (ϵ, δ)-DP when $\lambda \geq \frac{2\beta}{\epsilon}$ and $\sigma \geq \frac{L\sqrt{8\log(2/\delta)+4\epsilon}}{\epsilon}$.*

Differentially Private Gradient Descent DP-SGD is a differentially private version of stochastic gradient descent which ensures privacy by clipping the per-example gradients at each iteration before aggregating them and adding noise to the result. The update rule at iteration t is given by

$$\theta_{t+1} = \theta_t - \eta_t \left(\sum_{z \in B_t} \texttt{clip}\big(\nabla \ell(\theta_t; z)\big) + \mathcal{N}(0, \sigma^2 I_d) \right),$$

where η_t is the learning rate at iteration t, B_t is the current batch at iteration t, σ is the noise scale, and `clip` is a function that bounds the norm of the per-example gradients.

4.3 Analytical Tools

Existing privacy guarantees of the objective perturbation mechanism (Chaudhuri et al., 2011; Kifer et al., 2012) pre-date modern privacy accounting tools such as Rényi differential privacy and privacy profiles. In this section, we present two new privacy analyses of objective perturbation: an (ϵ, δ)-DP bound based on privacy profiles, and an RDP bound.

4.3.1 Approximate DP Bound

Theorem 4.3.1 (Approximate DP guarantees of objective perturbation for GLMs). *Consider a loss function $\ell(\theta; z) = f(x^T\theta; y)$ with GLM structure. Suppose that f is β-smooth and $||\nabla \ell(\theta; z)||_2 \leq L$ for all $\theta \in \mathcal{R}^d$ and $z \in \mathcal{X} \times \mathcal{Y}$. Fix $\lambda > \beta$. Let $\epsilon \geq 0$ and*

let $\tilde{\epsilon} = \epsilon - \log\left(1 - \frac{\beta}{\lambda}\right)$, $\hat{\epsilon} = \epsilon - \log\left(1 - \frac{\beta}{\lambda}\right) - \frac{L^2}{2\sigma^2}$, *and let* P *and* Q *be the density functions of* $\mathcal{N}(L, \sigma^2)$ *and* $\mathcal{N}(0, \sigma^2)$, *respectively. Objective perturbation satisfies* $\big(\epsilon,\delta(\epsilon)\big)$*-DP for*

$$\delta(\epsilon) = \begin{cases} 2 \cdot H_{e^{\tilde{\epsilon}}}\big(P||Q\big), & \text{if } \hat{\epsilon} \geq 0, \\ (1 - \mathrm{e}^{\hat{\epsilon}}) + \mathrm{e}^{\hat{\epsilon}} \cdot 2 \cdot H_{e^{\frac{L^2}{\sigma^2}}}\big(P||Q\big), & \text{otherwise.} \end{cases} \tag{4.3.1}$$

Notice that we can express (4.3.1) analytically using (C.2.1). To obtain the bound (4.3.1) we repeatedly use the fact that the privacy loss random variable (PLRV) determined by the distributions $\mathcal{N}(1, \sigma^2)$ and $\mathcal{N}(0, \sigma^2)$ is distributed as $\mathcal{N}(\frac{1}{2\sigma^2}, \frac{1}{\sigma^2})$. As the upper bound (4.3.1) is obtained using a PLRV that is a certain scaled and shifted half-normal distribution, we can also find certain scaled and shifted half-normal distributions P and Q which give the dominating pair of distributions for the objective perturbation mechanism such that the hockey-stick divergence between P and Q is exactly the upper bound (4.3.1) for all ϵ (shown in the appendix).

4.3.2 Rényi Differential Privacy Bound

If our sole objective is to obtain the tightest possible approximate DP bounds for objective perturbation, we can stop at Theorem 4.3.1! Directly calculating the privacy profiles of objective perturbation using the hockey-stick divergence, as in the previous section, will achieve this goal (until more privacy accounting tools come along).

In this section we turn instead to Rényi differential privacy, a popular relaxation of pure differential privacy ($\delta = 0$) which avoids the "catastrophic privacy breach" possibility permitted by approximate DP ($\delta > 0$). Below, we present an RDP guarantee for objective perturbation.

Theorem 4.3.2 (RDP guarantees of objective perturbation for GLMs)**.** *Consider a loss function* $\ell(\theta; z) = f(x^T\theta; y)$ *with GLM structure. Suppose that* f *is* β*-smooth and*

$||\nabla \ell(\theta; z)||_2 \leq L$ *for all* $\theta \in \mathcal{R}^d$ *and* $z \in \mathcal{X} \times \mathcal{Y}$. *Fix* $\lambda > \beta$. *Objective perturbation satisfies* (α, ϵ)*-RDP for any* $\alpha > 1$ *with*

$$\epsilon = -\log\left(1 - \frac{\beta}{\lambda}\right) + \frac{L^2}{2\sigma^2} + \frac{1}{\alpha - 1} \log \mathbb{E}_{X \sim \mathcal{N}\left(0, \frac{L^2}{\sigma^2}\right)} \left[e^{(\alpha-1)|X|}\right].$$

For $\alpha = 1$, *the RDP bound holds with*

$$\epsilon = -\log\left(1 - \frac{\beta}{\lambda}\right) + \frac{L^2}{2\sigma^2} + \log \mathbb{E}_{X \sim \mathcal{N}\left(0, \frac{L^2}{\sigma^2}\right)} \left[e^{|X|}\right].$$

One of our main motivations for improving the privacy analysis of objective perturbation comes from the observation that it can be competitive to DP-SGD when the privacy cost of hyperparameter tuning is included in the privacy budget. As the tightest results for DP hyperparameter tuning are stated in terms of RDP (Papernot and Steinke, 2021), in our experiments we use RDP bounds of objective perturbation to get a clear understanding of the differences in the privacy-utility trade-offs between these two approaches.

Remark 4.3.3. Privacy profile and RDP bounds (such as Theorems 4.3.1 and 4.3.2) are unified in the sense that they are both based on a certain bound of the PLRV $\epsilon_{Z,Z'}$ (Definition C.10.4) for a fixed pair of datasets Z, Z'. From Definitions 4.2.2 and 4.2.7 we see that for $\epsilon \in \mathbb{R}$, the hockey-stick divergence is

$$H_{\mathrm{e}^\epsilon}\big(\mathcal{M}(Z) \,||\, \mathcal{M}(Z')\big) = \mathbb{E}_{\theta \sim \mathcal{M}(Z)} \left[1 - \mathrm{e}^{\epsilon - \epsilon_{Z,Z'}(\theta)}\right]_+,$$

and for $\alpha > 1$ we have that the Rényi divergence is

$$D_\alpha\big(\mathcal{M}(Z) \,||\, \mathcal{M}(Z')\big) = \frac{1}{\alpha - 1} \log \mathbb{E}_{\theta \sim \mathcal{M}(Z)} \left[\mathrm{e}^{\alpha\, \epsilon_{Z,Z'}(\theta)}\right].$$

4.3.3 Distance to Optimality

How close to optimal are the bounds of Theorems 4.3.1 and 4.3.2? We can in fact show that the Gaussian mechanism is a special case of the objective perturbation mechanism — thereby providing a lower bound on its approximate DP and RDP.

Lemma 4.3.4. *Consider a loss function $\ell(\theta; z)$ with gradient norm bounded by $||\nabla\ell(\theta; z)||_2 \leq L$, and the objective perturbation mechanism $\hat{\theta}^P(Z)$ with noise parameter $\sigma > 0$ and regularization parameter $\lambda > 0$. For all $\alpha > 1$ and neighboring datasets $Z \simeq Z'$ we then have the following:*

$$H_\alpha\big(\hat{\theta}^P(Z) \,||\, \hat{\theta}^P(Z')\big) \geq H_\alpha\big(\mathcal{N}(0, \tfrac{\sigma^2}{\lambda^2}) \,||\, \mathcal{N}(\tfrac{L}{\lambda}, \tfrac{\sigma^2}{\lambda^2})\big) = H_\alpha\big(\mathcal{N}(0, \sigma^2) \,||\, \mathcal{N}(L, \sigma^2)\big).$$

Proof. Consider the loss function $\ell(\theta; x) = x^T\theta$ and choose neighboring datasets $X = \{x\}$ and $X' = \emptyset$, for some $x \in \mathbb{R}^d$. Fix $\lambda > 0$ and sample $b \sim \mathcal{N}(0, \sigma^2 I_d)$. Then the objective perturbation mechanism solves

$$\begin{aligned}\hat{\theta}^P(X) &= \arg\min_{\theta\in\mathbb{R}^d} x^T\theta + \frac{\lambda}{2}||\theta||_2^2 + b^T\theta = -\frac{1}{\lambda}(x + b),\\ \hat{\theta}^P(X') &= \arg\min_{\theta\in\mathbb{R}^d} \frac{\lambda}{2}||\theta||_2^2 + b^T\theta = -\frac{1}{\lambda}b.\end{aligned}$$

Observe that $\hat{\theta}^P(X) \sim \mathcal{N}(-\frac{1}{\lambda}x, \frac{\sigma^2}{\lambda^2}I_d)$ and $\hat{\theta}^P(X') \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda^2}I_d)$. Following the problem setting described in Theorem 4.2.8, we have that $||x||_2 = ||\nabla\ell(\theta; x)||_2 \leq L$. In this case, objective perturbation reduces directly to the Gaussian mechanism with sensitivity $\Delta_f = \frac{L}{\lambda}$ and noise scale $\frac{\sigma}{\lambda}$. Lemma 4.3.4 then holds due to the scaling invariance of the hockey-stick divergence. □

The argument works the same for the Rényi divergence D_α which is similarly invariant to scale. Lemma 4.3.4 implies that we can measure the tightness of the bounds given

in Theorems 4.3.1 and 4.3.2 by comparing them to the tight bounds of the Gaussian mechanism (C.2.1) with sensitivity $\Delta_f = L$ and noise scale σ.

This means that in Figure 4.1, the hockey-stick divergence of the Gaussian mechanism is a lower bound on the hockey-stick divergence for objective perturbation. While our hockey-stick divergence bound is unsurprisingly a bit tighter than the RDP bound for objective perturbation, we see that both significantly improve over the classic (ϵ, δ)-DP bounds of Kifer et al. (2012).

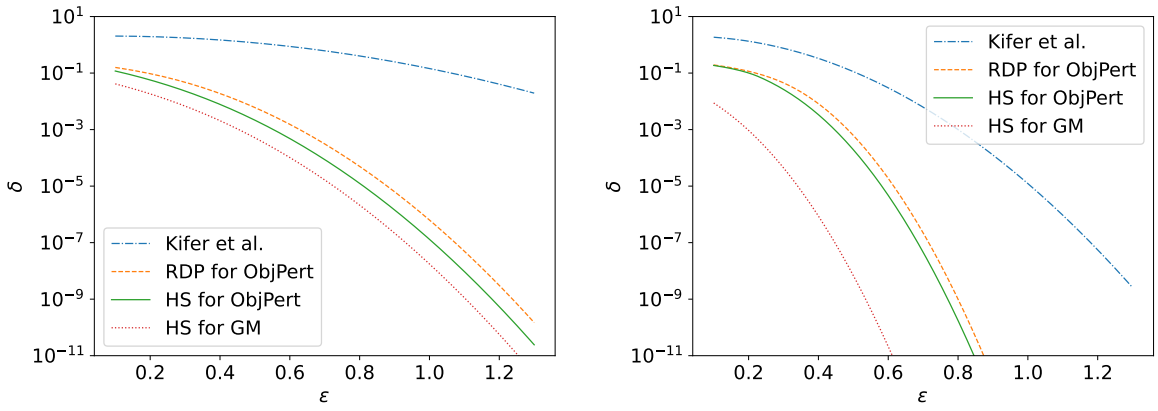


Figure 4.1: Comparison of different (ϵ, δ)-bounds for objective perturbation: the (ϵ, δ)-bound by Kifer et al. (2012) given in Thm. 4.2.8, the RDP bound of Thm. 4.3.2, the approximate DP bound of Thm. 4.3.1 using the hockey-stick divergence and the approximate DP lower bound obtained using the hockey-stick divergence and Cor. **??**. Left: $\sigma = 5$, $\beta = 1$ and $\lambda = 20$. Right: $\sigma = 10$, $\beta = 1$ and $\lambda = 5$.

Remark 4.3.5. The RDP and approximate DP bounds in this section require a careful analysis of the dependence between the noise vector b and the private minimizer θ^P. In the appendix, we show how the GLM assumption simplifies this issue.

4.4 Computational Tools

In this section we present Algorithm 7, which extends the Approximate Minima Perturbation of Iyengar et al. (2019) to handle loss functions with unbounded gradient.

Approximate minima The privacy guarantees of objective perturbation hold only when its output is an *exact* minimizer of the perturbed objective. Approximate Minima Perturbation (AMP) (Iyengar et al., 2019) addresses this issue by finding an approximate minimizer to the perturbed objective, then privately releases this approximate minimizer with the Gaussian mechanism.

Gradient clipping DP-SGD requires no *a priori* bound on the gradient of the loss function; at each iteration, the algorithm clips the per-example gradients above a pre-specified threshold in order to bound the gradient norms. We extend this same technique to objective perturbation.

Given a loss function $\ell(\theta; z)$ and a clipping threshold C, we can transform the gradient of $\ell(\theta; z)$ as follows:

$$\nabla \ell_C(\theta; z) = \begin{cases} \nabla \ell(\theta; z) & \text{if } ||\nabla \ell(\theta; z)||_2 \leq C, \\ \frac{C}{||\nabla \ell(\theta; z)||_2} \nabla \ell(\theta; z) & \text{if } ||\nabla \ell(\theta; z)||_2 > C. \end{cases}$$

Then we can define the aggregation of clipped gradients as $\nabla \mathcal{L}_C(\theta; Z) = \sum_{z \in Z} \nabla \ell_C(\theta; z)$.

The aggregation of clipped gradients $\nabla \mathcal{L}_C(\theta; Z)$ corresponds to an implicit "clipped-gradient" objective function $\mathcal{L}_C(\theta; Z)$. For convex GLMs, Song et al. (2020) define this function precisely and show that it retains the convexity and GLM structure of the original objective function $\mathcal{L}(\theta; Z)$. We furthermore demonstrate that this function retains the same bound β on the Lipschitz smoothness (Theorem C.6.3).

Algorithm 7 extends the privacy guarantees of AMP (Iyengar et al., 2019) to loss functions with unbounded gradient. Notice that for smooth loss functions with gradient norm bounded by L, we can set $C = L$ in order to recover Approximate Minima Perturbation.

Algorithm 7 Approximate Minima Perturbation with Gradient Clipping

Input: dataset Z; noise levels $\sigma, \sigma_{\texttt{out}}$; β-smooth loss function $\ell(\cdot)$; regularization strength λ; gradient norm threshold τ; clipping threshold C.

1. Construct the set of clipped-gradient loss functions $\{\ell_C(\theta; z) : z \in Z\}$.
2. Sample $b \sim \mathcal{N}(0, \sigma^2 I_d)$.
3. Let $\mathcal{L}_C^P(\theta; Z, b) = \sum_{z \in Z} \ell_C(\theta; z) + \frac{\lambda}{2}||\theta||_2^2 + b^T\theta$.
4. Solve for $\tilde{\theta}$ such that $||\nabla \mathcal{L}_C^P(\tilde{\theta}; Z)||_2 \leq \tau$.

Output: $\tilde{\theta}^P = \tilde{\theta} + \mathcal{N}(0, \sigma_{\texttt{out}}^2 I_d)$.

Theorem 4.4.1 (RDP guarantees of Algorithm 7). *Consider a loss function $\ell(\theta; z) = f(x^T\theta; y)$ with GLM structure, such that f is β-smooth. Fix $\lambda > \beta$. Algorithm 7 satisfies (α, ϵ)-RDP for any $\alpha > 1$ with*

$$\epsilon \leq -\log\left(1 - \frac{\beta}{\lambda}\right) + \frac{C^2}{2\sigma^2} + \frac{1}{\alpha - 1} \log \mathbb{E}_{X \sim \mathcal{N}\left(0, \frac{C^2}{\sigma^2}\right)} \left[e^{(\alpha-1)|X|}\right] + \frac{2\tau^2\alpha}{\sigma_{\texttt{out}}^2 \lambda^2}.$$

Remark 4.4.2. Gradient clipping aside, our proof of Theorem 4.4.1 takes a different tack than the proof of Theorem 1 (for AMP) in Iyengar et al. (2019). We observe that Algorithm 7 is essentially an adaptive composition of the objective perturbation mechanism and the Gaussian mechanism. We can write $\tilde{\theta} = \theta^P + (\tilde{\theta} - \theta^P)$ to see that we are releasing two quantities: θ^P (with objective perturbation) and the difference $\tilde{\theta} - \theta^P$ (with the Gaussian mechanism). Algorithm 7 stops iterating only after the gradient norm $||\nabla \mathcal{L}_C^P(\tilde{\theta}; Z)||_2$ is below the threshold τ. This along with the λ-strong convexity of the objective function $\nabla \mathcal{L}_C^P(\theta; Z)$ ensures a bound on the ℓ_2-sensitivity of the difference $\tilde{\theta} - \theta^P$, so that we can apply the Gaussian mechanism.

4.4.1 Computational Guarantee

To achieve the optimal excess risk bounds for DP-ERM in the convex setting, DP-SGD clocks in at a hefty $O(n^2)$ gradient evaluations (Bassily et al., 2014). It has been an open problem to obtain an optimal DP-ERM algorithm that runs in subquadratic time (Kulkarni et al., 2021). One of our contributions is to show that when we further restrict to smooth GLM losses (so that ObjPert is applicable) Algorithm 7 can achieve the same optimal rate with only $O(n \log n)$ gradient evaluations.

A formal claim and proof that Algorithm 7 — with appropriately chosen parameters — achieves the optimal rate is deferred to Appendix C.8. The analysis is largely the same as that in Kifer et al. (2012) but with the bug fixed (details in Appendix C.5) by adding a GLM assumption.

The improved computational complexity is due to the fact that we can apply any off-the-shelf optimization algorithm to solve Step 4 of Algorithm 7. Observing that $\mathcal{L}^P(\theta; Z, b)$ has a finite-sum structure, we can employ the Stochastic Averaged Gradient (SAG) method (Schmidt et al., 2017) which halts in $O(n \log n)$ with high probability. Details are provided in Appendix C.7.

4.5 Empirical Evaluation

In this section we evaluate Algorithm 7 against two baselines: "dishonest" DP-SGD and "honest" DP-SGD. Dishonest DP-SGD does not account for the privacy cost of hyperparameter tuning; honest DP-SGD follows the private selection algorithm and RDP bounds from Papernot and Steinke (2021).

Our experimental design includes some guidelines in order to make it a fair fight. One of the strengths of Algorithm 7 that we advocate for is its blackbox optimization. Whereas DP-SGD consumes privacy budget for testing each set of hyperparameter candidates,

an advantage of approximate minima perturbation is that the privacy guarantees are independent of the choice of optimizer used to solve for $\tilde{\theta}$. We can therefore test out any number of optimization hyperparameters for Algorithm 7 *at no additional privacy cost*, provided that these parameters are independent of the privacy guarantee (e.g. learning rate, batch size). More specifically, once the loss function is perturbed with the noise b in Algorithm 7, any $\tilde{\theta}$ that satisfies the convergence guarantees with the tolerance parameter τ will have the RDP guarantees of Theorem 4.4.1 and therefore we are free to also carry out tuning of the optimization algorithm without an additional privacy cost.

Because we are interested in measuring the effect of the privacy cost of hyperparameter tuning, we tune only the learning rate which does not affect the privacy guarantee of the base algorithm. This isolates the effect of hyperparameter tuning as we will need to appeal to Papernot and Steinke (2021) to get valid DP bounds for DP-SGD, but Algorithm 7 enjoys hyperparameter tuning "for free".

The following table summarizes the optimization-related parameters for all three methods.

	Dishonest DP-SGD	Honest DP-SGD	Algorithm 7
clipping	$C = \sqrt{2}$	$C = \sqrt{2}$	$C = \sqrt{2}$
learning rate	$\log(10^{-8}, 10^{-1})$	$\log(10^{-8}, 10^{-1})$	linear$(.08, .5)$
grid size	$s = 10$	$\mu \approx 15.4.$	$s = 10$
optimizer	Adam	Adam	L-BFGS
convergence	after T iterations	after T iterations	$\|\|\nabla\mathcal{L}(\tilde{\theta})\|\|_2 \leq \tau$

The choice of $C = \sqrt{2}$ is a natural value for logistic regression in that $||\nabla\ell(\theta, z)|| \leq \sqrt{2}$ for all θ, z due to data-preprocessing and the bias term. Dishonest DP-SGD selects $s = 10$ learning rate candidates evenly log-spaced from the range of values between 10^{-8} and 10^{-1}. Honest DP-SGD selects learning rate candidates from the same range of values, but with granularity determined by a random variable K sampled from the Poisson distribution

Poisson(μ). We select μ so that with 90% probability, K is larger than the grid size s used for dishonest DP-SGD, resulting in $\mu \approx 15.4$.

We use the Adam optimizer for both honest and dishonest DP-SGD. For Algorithm 7 we use the L-BFGS optimizer whose second-order behavior allows us to get within a smaller distance to optimal (as required by Algorithm 7).

For DP-SGD we set the subsampling ratio such that the expected batch size is 256 and we run for 60 "epochs". We calculate the number of iterations as $T = 60 \cdot \texttt{num\_batches}$, where `num_batches` is the number of batches in the training dataset (we pass the train loader through the `Opacus` privacy engine, so the size of each batch is random). To calibrate the scale of the noise for DP-SGD, we use the analytical moments accountant (Wang et al., 2019) with Poisson sampling (Zhu and Wang, 2019; Mironov et al., 2019).

The parameters specific to AMP are σ_{out}, the noise scale for the output perturbation step; and τ, the gradient norm bound. A larger τ will improve our computational cost, though our approximate minimizer $\tilde{\theta}$ will be farther away from the true minimizer θ^P. We can achieve a smaller τ by choosing a larger σ_{out}, but this will mean that our release of $\tilde{\theta}$ will be noisier. In our experiments we fix $\tau = 0.01$ and $\sigma_{out} = 0.15$.

The privacy parameters of objective perturbation are the noise scale σ and the regularization strength λ. Balancing these parameters is a classic exercise in bias-variance trade-off. A larger σ will allow us to use less regularization, but if σ is too large then we risk adding too much noise to the objective function and hurting utility.

Our strategy is to find the smallest possible λ such that σ isn't too large. To quantify when σ is "too large", we use the Gaussian mechanism as a reference point: the noise scale σ for objective perturbation shouldn't be too much larger than the noise scale σ_G for the Gaussian mechanism. Let's say that the Gaussian mechanism with noise scale σ_G satisfies (ϵ, δ)-DP, then we want our σ for (ϵ, δ)-DP objective perturbation to satisfy $\sigma \leq f\sigma_G$ for some small constant factor f. In our experiments, we set $f = 1.3$.

For objective perturbation, we can thus select the privacy parameters σ and λ using fixed values (e.g., $\epsilon, \delta, \sigma_G$) that are independent of the data. Likewise, the choices of σ_{out} and τ are fixed across all datasets. This is noteworthy since $\sigma, \lambda, \sigma_{out}$ and τ each have an effect on the privacy guarantee, outside of the blackbox algorithm. Tuning these parameters on the data would require us to use the same private selection algorithm as we need for honest DP-SGD.

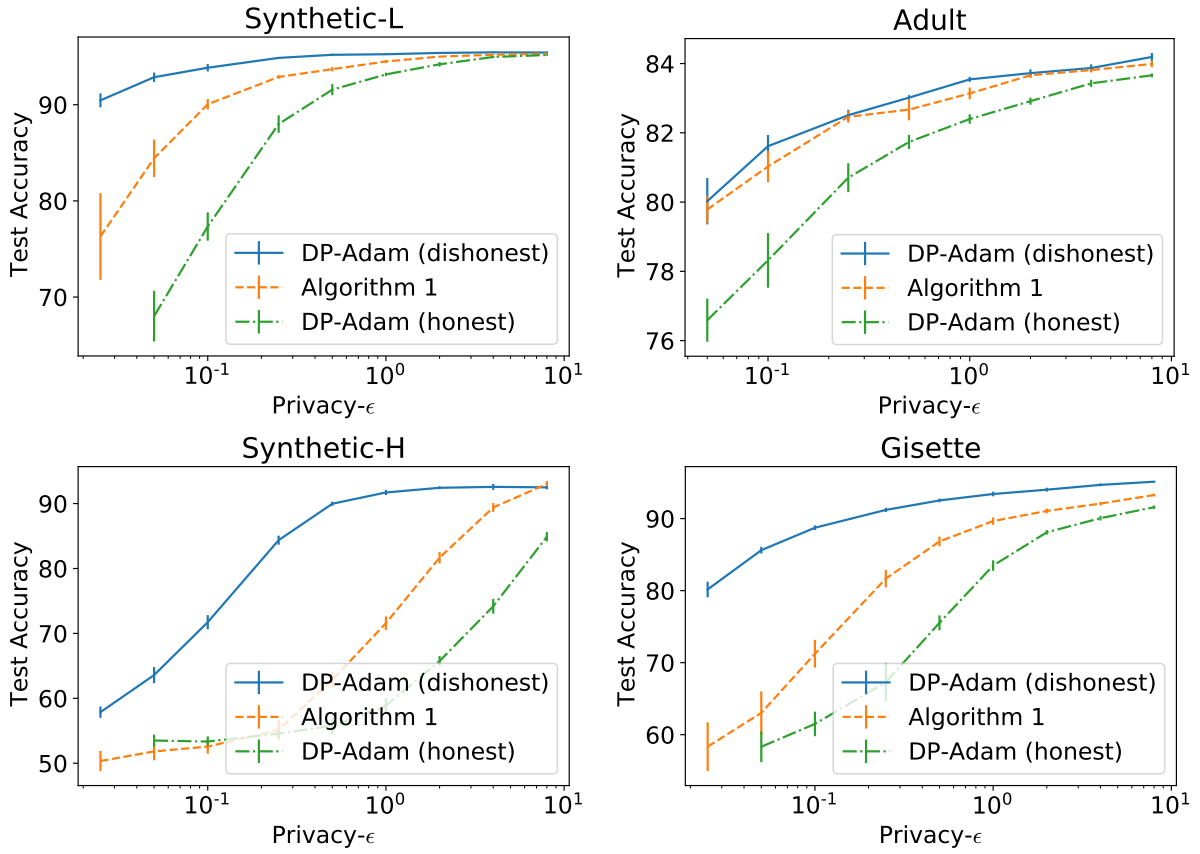


Figure 4.2: Comparison of Algorithm 7 against honest and dishonest DP-SGD baselines, varying $\epsilon \in \{0.025, 0.05, 0.1, 0.25, 0.5, 1.0, 2.0, 4.0, 8.0\}$ and fixing $\delta = 10^{-5}$. On all three methods, we train the model for each learning rate on its grid (see Table 4.5) and report the test accuracy for the best learning rate on the grid. Results are averaged over 10 trials and the error bars on both sides of the mean values depict 1.96 times the standard error, giving the asymptotic 95% coverage.

We evaluate our methods for binary classification on the Adult, Synthetic-L, Synthetic-

H and Gisette datasets provided by Iyengar et al. (2019). We normalize each row x_i to have unit ℓ_2-norm. Note that the assignment $x_i \leftarrow \frac{x_i}{||x_i||_2}$ doesn't require expending any privacy budget as each data point is transformed only by its own per-sample norm.

Table 4.1: Synthetic-L

	Dishonest	Alg 7	Honest
$\epsilon = 0.1$	93.85%	90.05%	77.34%
$\epsilon = 1$	95.25%	94.50%	93.15%
$\epsilon = 8$	95.43%	95.30%	95.17 %

Table 4.2: Adult

	Dishonest	Alg 7	Honest
$\epsilon = 0.1$	81.61%	81.37%	78.32%
$\epsilon = 1$	83.54%	83.18%	82.40%
$\epsilon = 8$	84.19%	83.99%	83.66%

Results from Figure 4.5 in numerical format for the low-dimensional datasets, Synthetic-L and Adult. For $\epsilon = 0.1$, these can be cross-referenced with the results in Fig. 3 from Iyengar et al. (2019).

The experimental results shown in Figure 4.5, Table 4.1 and Table 4.2 paint a consistent picture. While dishonest DP-SGD is clearly the best-performing algorithm, when we account for the cost of hyperparameter tuning then Algorithm 7 can compete with and often best honest DP-SGD. This effect is especially pronounced under small ϵ, for which diverting some of the limited privacy budget to hyperparameter tuning could be more impactful.

Is it fair? Our experimental design aims to fairly compare ObjPert to DP-SGD. One limitation, however, is that the state-of-the-art tools for private hyperparameter tuning from Papernot and Steinke (2021) are RDP bounds — and RDP is *not* state-of-the-art for DP-SGD privacy accounting. At this moment, the tighest privacy accounting tools for DP-SGD belong to a family of work (Koskela et al., 2020; Gopi et al., 2021; Zhu et al., 2022) which numerically computes its privacy curve. These are the counterparts to our privacy profiles analysis for ObjPert (Theorem 4.3.1). Unfortunately, even though dishonest DP-SGD would benefit from using these numerical accountants, for private hyperparameter tuning we would then have to use the sub-optimal private selection

bounds for approximate DP from Liu and Talwar (2019). In our experiments we therefore use RDP-based privacy accounting for both ObjPert and DP-SGD. Comparing DP-SGD with numerical accounting of privacy profiles against ObjPert with Theorem 4.3.1 will have to wait until more private selection tools are available.

One might also object that by tuning only the learning rate for DP-SGD, we didn't explore the full range of hyperparameters relevant to DP-SGD's performance. While tuning additional hyperparameters such as the batch size and number of epochs could benefit dishonest DP-SGD, it would likely worsen the privacy-utility tradeoff for honestly-tuned DP-SGD due to the increased privacy cost of the hyperparameter tuning algorithm from Papernot and Steinke (2021).

4.6 Conclusion

One point that we really wanted to drive home is that while DP-SGD works extraordinarily well across a wide variety of problem settings, it's not necessarily the best solution for *every* problem setting. But at the same time, DP-SGD has received the benefit of an enormous amount of attention that other DP learning algorithms haven't received. A goal of our paper was to hone in on a particular problem setting and give a different algorithm the same star treatment.

Objective perturbation now boasts two new privacy analyses. One is an improved (ϵ, δ)-DP analysis based on bounding the hockey-stick divergence. The other is an RDP analysis which allows us to fairly compare objective perturbation against DP-SGD — the workhorse of differentially private learning — with honest hyperparameter tuning. We've also expanded the approximate minima perturbation algorithm of Iyengar et al. (2019) in order to encompass a broader range of loss functions which need not have bounded gradient. Our algorithm moreover can be used in conjunction with SVRG to guarantee a

running time of $O(n \log n)$ to achieve the optimal excess risk bounds, improving on the $O(n^2)$ computational guarantee of DP-SGD.

Appendix A

Supplementary Material for Chapter 2

A.1 DP Variants

Algorithm design is a typical use case for differential privacy: given a privacy budget of ϵ, the data curator would like to add noise calibrated to meet the privacy demands. Our work in Chapter 2 concerns the converse problem of how to calculate and report the *incurred* privacy loss to an individual after a randomized algorithm is run on a fixed dataset. The table below summarizes the relevant variations of the DP definition which characterize the privacy loss with varying degrees of granularity.

Let P, Q be distributions over Ω, taking $p(\omega)$ and $q(\omega)$ to be the probability density/mass function of each at ω. Then the probability metrics used in the table are defined as follows:

- $D_\infty(P||Q) = \sup_{S \subset \Omega} \left(\log \frac{P(S)}{Q(S)} \right)$ (max divergence)
- $D_\infty^\delta(P||Q) = \sup_{S \subset \Omega : P(\omega) \geq \delta} \left(\log \frac{P(S) - \delta}{Q(S)} \right)$ (δ-approximate max divergence),
- $D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log \mathbb{E}_{\omega \sim Q} \left[\left(\frac{p(\omega)}{q(\omega)} \right)^\alpha \right]$ (Rényi divergence).

Pure DP	$\sup_{D} \sup_{z, D' : D' \simeq_z D} D_\infty\big(\mathcal{A}(D) \| \mathcal{A}(D')\big) \leq \epsilon$
Approximate DP	$\sup_{D} \sup_{z, D' : D' \simeq_z D} D_\infty^\delta\big(\mathcal{A}(D) \| \mathcal{A}(D')\big) \leq \epsilon$
Rényi DP	$\sup_{D} \sup_{z, D' : D' \simeq_z D} D_\alpha\big(\mathcal{A}(D) \| \mathcal{A}(D')\big) \leq \epsilon$
Data-dependent DP	$\sup_{z, D' : D' \simeq_z D} D_\alpha\big(\mathcal{A}(D) \| \mathcal{A}(D')\big) \leq \epsilon(D)$
Personalized DP	$\sup_{D, D' : D' \simeq_z D} \max \left\{ \begin{matrix} D_\infty^\delta\big(\mathcal{A}(D) \| \mathcal{A}(D')\big), \\ D_\infty^\delta\big(\mathcal{A}(D') \| \mathcal{A}(D)\big) \end{matrix} \right\} \leq \epsilon(z)$
Per-instance DP	$\max \left\{ \begin{matrix} D_\infty^\delta\big(\mathcal{A}(D) \| \mathcal{A}(D')\big), \\ D_\infty^\delta\big(\mathcal{A}(D') \| \mathcal{A}(D)\big) \end{matrix} \right\} \leq \epsilon(D, z)$
***Ex-post* per-instance DP**	$\left\lvert \log \frac{\Pr\big[\mathcal{A}(D) = o\big]}{\Pr\big[\mathcal{A}(D') = o\big]} \right\rvert \leq \epsilon(o, D, D') \quad \text{where } D' \simeq_z D$

A.2 Additional Experiments

A.2.1 Varying dimension and dataset size

Our first experiment uses a synthetic dataset for logistic regression as described in the experiments section of the main paper. Figure A.1 illustrates how the worst-case pDP loss over all individuals in the dataset – i.e., $\max_{z \in D} \epsilon_1(\hat{\theta}^P, D, D_{\pm z})$ – changes as a function of the dataset size (number of individuals in the dataset) n, compared to the worst-case pDP bounds given by the data-independent and data-dependent approaches. We fix $d = 50$ and vary n from $n = 100$ to $n = 10000$.

Figure A.1 illustrates how the worst-case pDP loss and bounds change as a function of the data dimension d. We fix $n = 1000$ and vary d from $d = 1$ to $d = 60$. Figures A.1 and A.2 demonstrate that for GLMs, the strength of our *ex-post* pDP bounds $\epsilon_1^P(\cdot)$ does not depend on the size of the dataset or the dimensionality of the data.

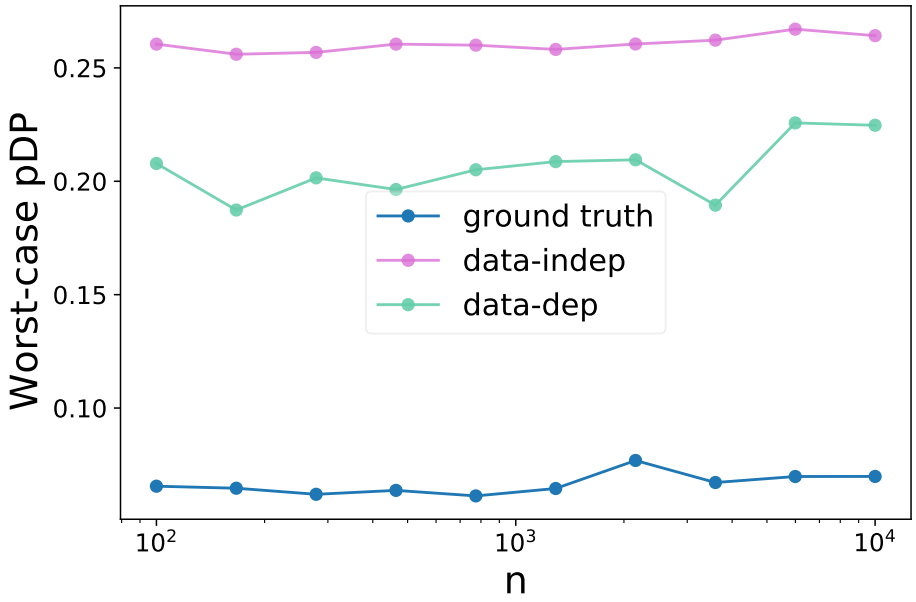


Figure A.1: Worst-case pDP while varying n.

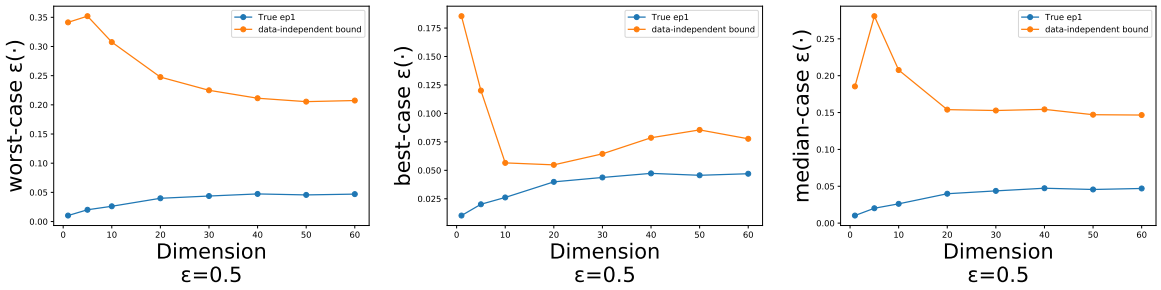


Figure A.2: Worst-case pDP while varying d.

A.2.2 Privacy budget allocation

Here we investigate how to distribute the privacy budget between the components of Algorithm 1 and Algorithm 2, with the same experimental setup as in Section 2.4.2. As before, we use the UCI credit default dataset. Our experiments show that a careful allocation of the privacy budget is essential to reaping the benefits of the data-dependent approach to releasing the *ex-post* pDP losses.

The plots in Figures A.3 and A.4 are ordered by increasing ϵ_1^{DEP}. $\epsilon_1^{INDEP} = 1$ is fixed, as are (implicitly) $\epsilon_2^{INDEP} = \epsilon_3^{INDEP} = 0$. We see that as ϵ_1^{DEP} approaches the total privacy budget of $\epsilon_1^{INDEP} = 1$, leaving less budget for ϵ_2^{DEP} and ϵ_3^{DEP}, the data-dependent release is little better than the data-independent release – worse, even, because we've expended additional privacy cost without significantly boosting the accuracy of the release.

Deciding between the data-independent or data-dependent approach is a delicate choice which depends on the particular problem setting. However, based on our theoretical and experimental results we can offer some loose guidelines:

- For non-GLMs, the data-independent bound has a dimension dependence. Therefore in the high-dimensional case, we recommend the data-dependent approach for generic convex loss functions and the data-independent approach for GLMs.

- For GLMs, the data-independent approach gives tight bounds without any overhead. The only reason to use the data-dependent approach for GLMs would be to gain an even more accurate estimate of the *ex-post* pDP losses, in which case it would be necessary to either suffer an additional privacy cost, or maintain the privacy cost by suffering a less accurate estimate of $\hat{\theta}^P$.

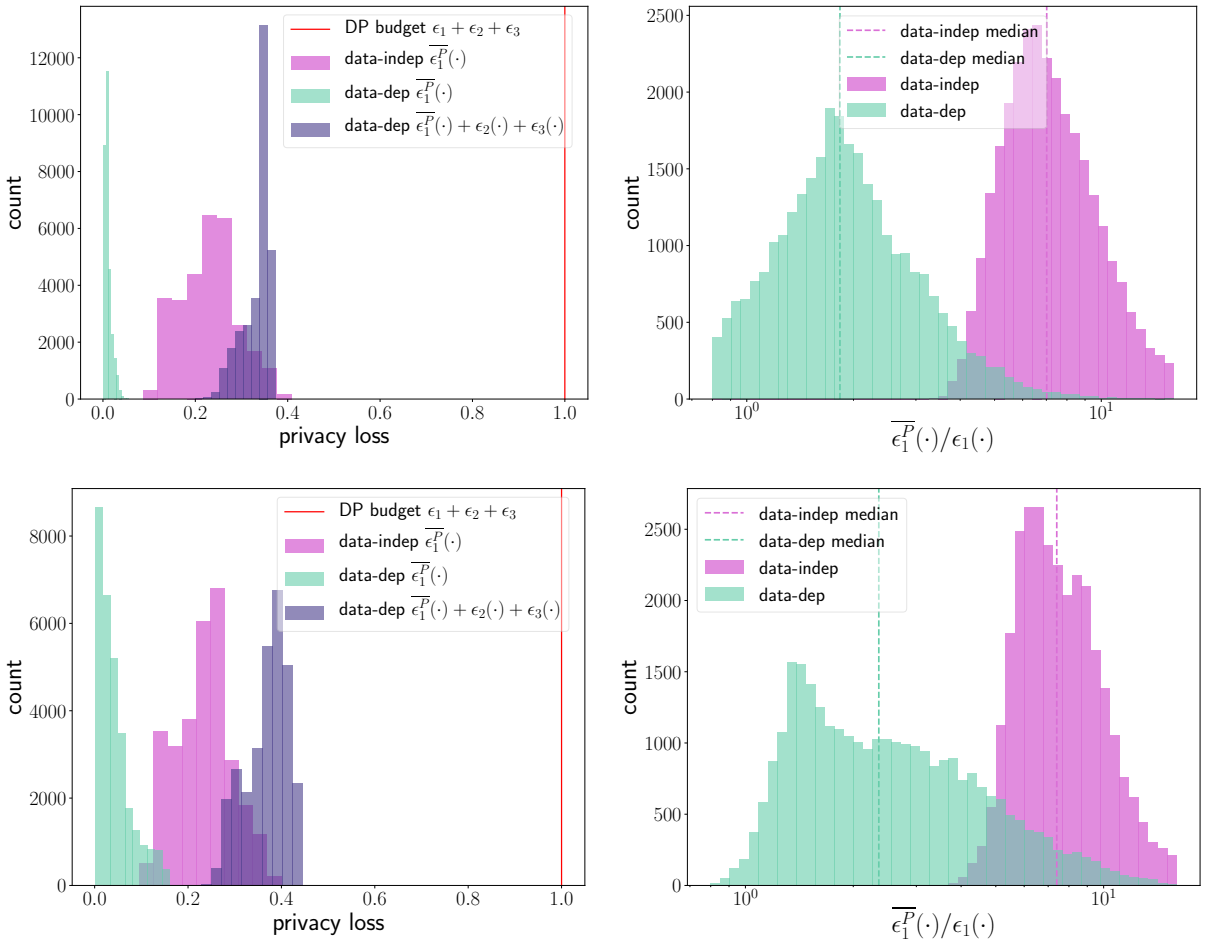


Figure A.3: Data-independent release uses a privacy budget $\epsilon_1 = 1$ for each plot. Left: $\epsilon_1 = 0.2, \epsilon_2 = 0.7, \epsilon_3 = 0.1$. Right: $\epsilon_1 = 0.4, \epsilon_2 = 0.5, \epsilon_3 = 0.1$.

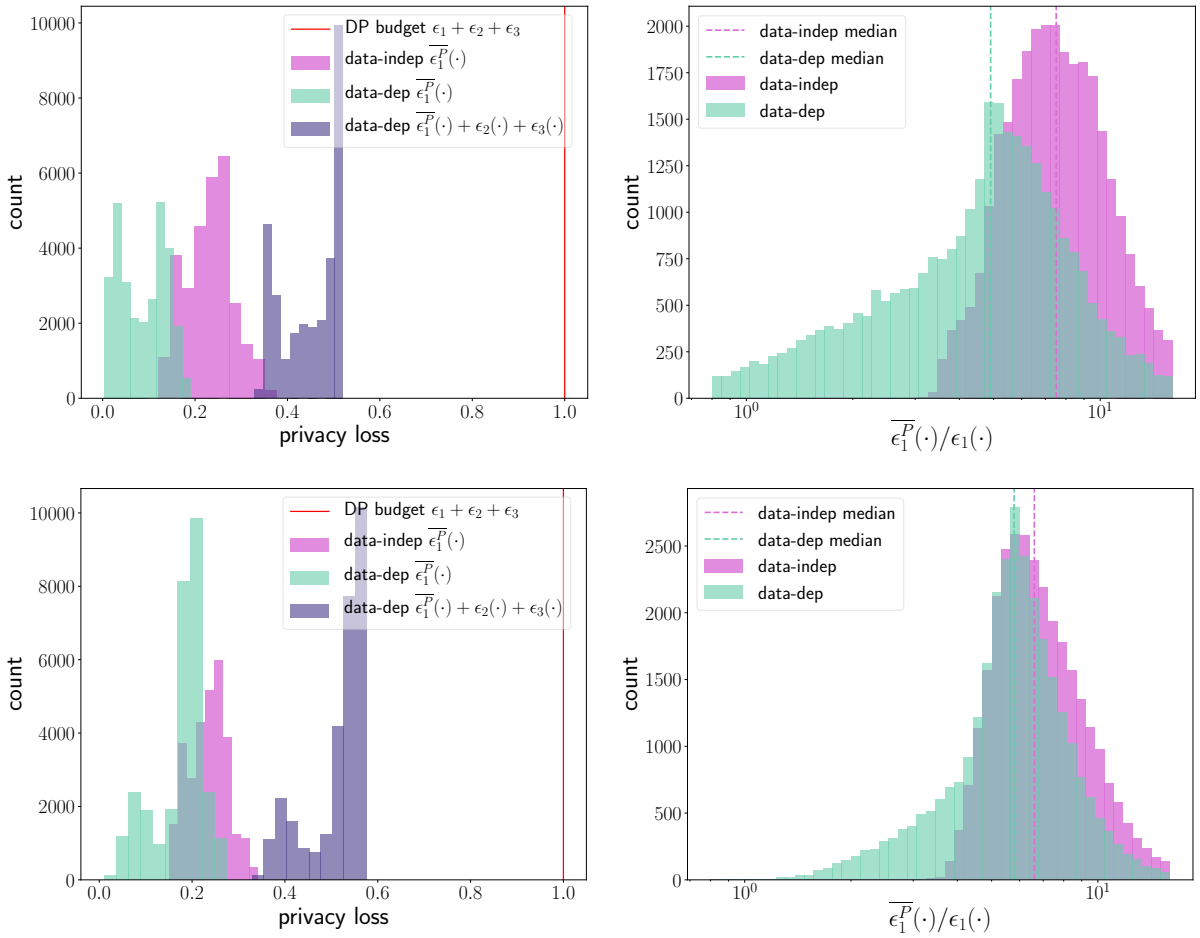


Figure A.4: Data-independent release uses a privacy budget $\epsilon_1 = 1$ for each plot. Left: $\epsilon_1 = 0.5, \epsilon_2 = 0.25, \epsilon_3 = 0.25$;. Right: $\epsilon_1 = 0.8, \epsilon_2 = 0.1, \epsilon_3 = 0.1$.

A.2.3 Comparison of pDP losses and private upper bounds

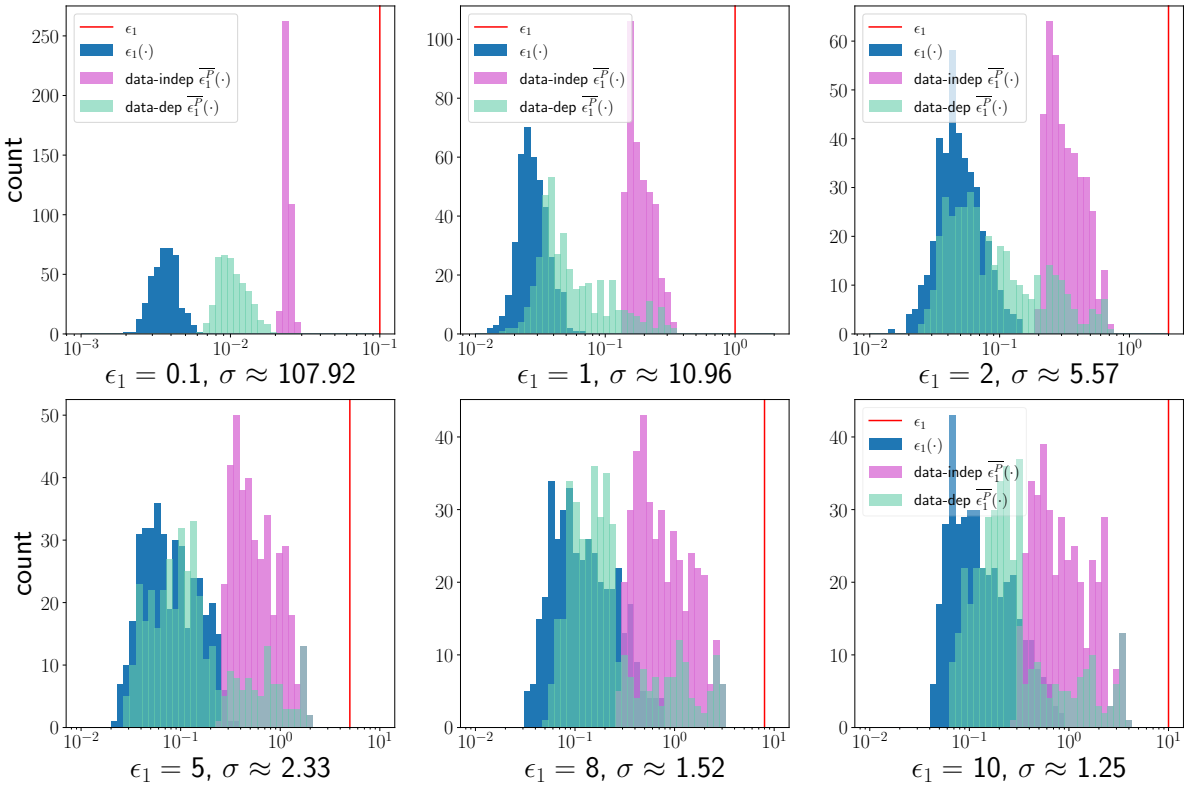


Figure A.5: pDP losses $\epsilon_1(\cdot)$ and upper bounds $\overline{\epsilon_1^P}(\cdot)$ for private logistic regression applied to the UCI kidney dataset. DP budget for releasing $\hat{\theta}^P$ is $\epsilon = 1$, marked in red.

We run both the data-independent and -dependent variations of Algorithm 1 as described in the experimental setup. Note that in this experiment the additional DP budget for the data-dependent release is $\epsilon_2 = \epsilon_3 = 1$, i.e. the privacy budget for the data-dependent release is three times the DP budget for the data-independent release. Figures A.5 and A.6 compare the pDP losses $\epsilon_1(\cdot)$ and private upper bounds $\overline{\epsilon_1^P}$ with ϵ_1 (indicated by the vertical red line), the DP budget for Algorithm 1. Figure A.5 shows results for private logistic regression on the UCI kidney dataset; Figure A.6 shows results for private linear regression on the UCI wine quality dataset (Dua and Graff, 2017). Our

experimental results indicate that for smaller $\epsilon_1 << 1$ (larger σ), the data-dependent approach provides a markedly tighter bound on $\epsilon_1(\cdot)$.

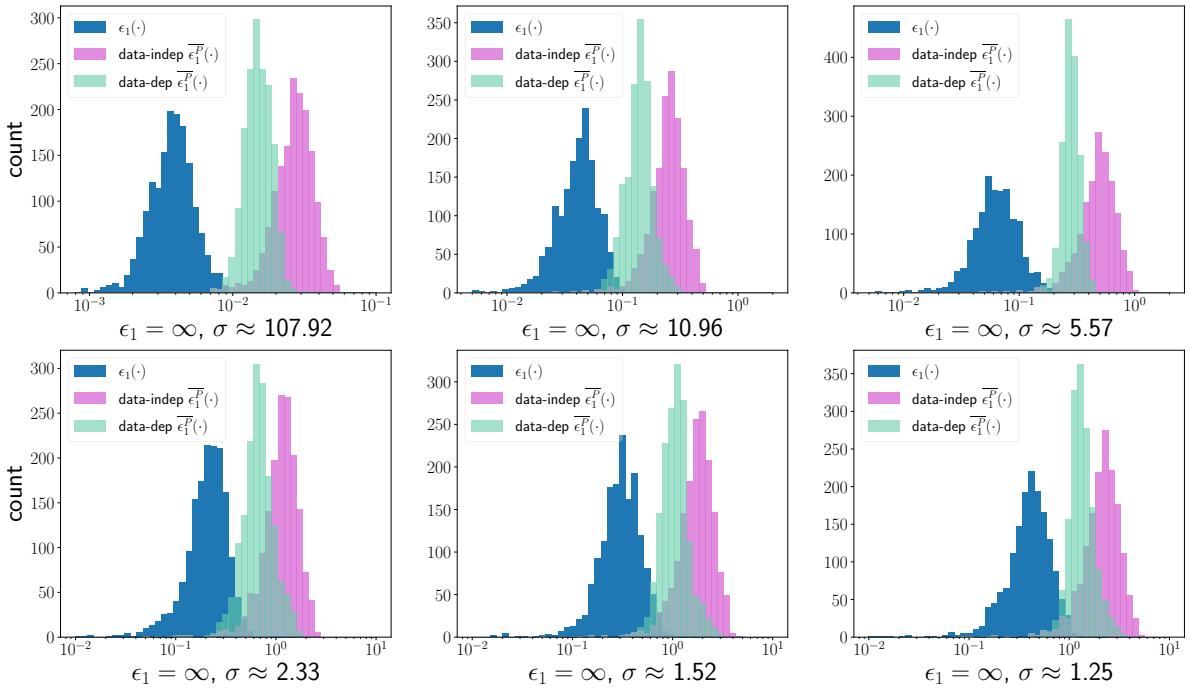


Figure A.6: pDP losses $\epsilon_1(\cdot)$ and upper bounds $\overline{\epsilon_1^P}(\cdot)$ for private linear regression applied to the UCI wine quality dataset. Since we are dealing with an unbounded domain $\mathbb{R}^d$, the algorithm does not satisfy worst-case DP for any $\epsilon < \infty$.

Figures A.7 and A.8 plot the ratio of the private upper bound $\epsilon_1^P(\cdot)$ for both the data-independent and -dependent approaches to the true pDP loss $\epsilon_1(\cdot)$. This illustrates the relative accuracy of the pDP estimates $\epsilon_1^P(\cdot)$. For both logistic regression on the UCI kidney dataset (Figure A.7) and linear regression on the UCI wine quality dataset (Figure A.8), the data-dependent approach provides a more accurate estimate of the pDP loss $\epsilon_1(\cdot)$, especially for logistic regression on the kidney dataset.

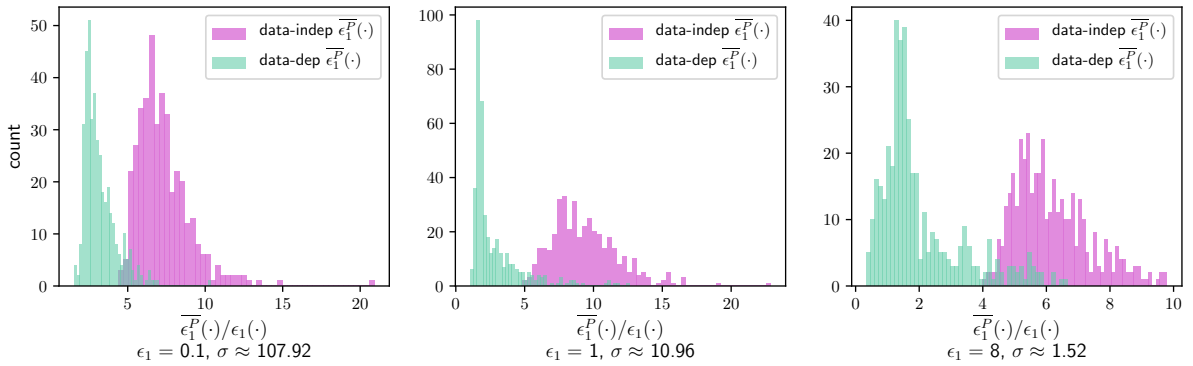


Figure A.7: Ratio of private upper bound $\overline{\epsilon_1^P}(\cdot)$ to actual pDP loss $\epsilon_1(\cdot)$ for private logistic regression applied to the UCI kidney dataset.

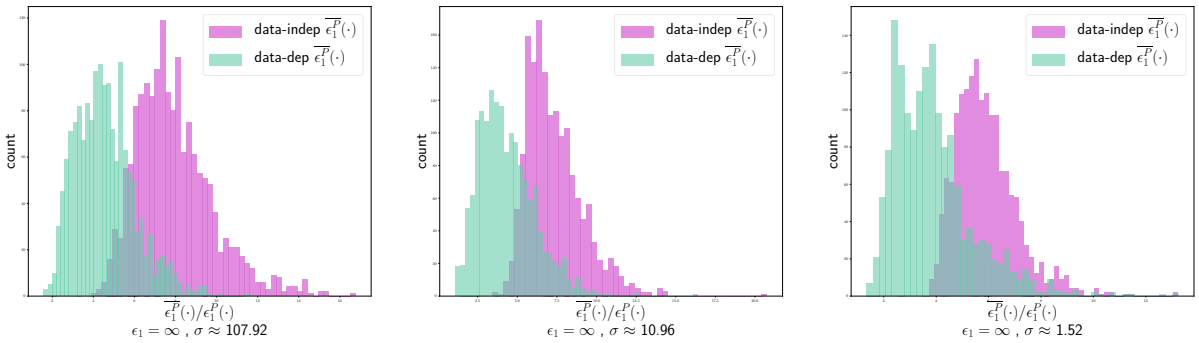


Figure A.8: Ratio of private upper bound $\overline{\epsilon_1^P}(\cdot)$ to actual pDP loss $\epsilon_1(\cdot)$ for private linear regression applied to the UCI wine quality dataset.

A.3 Even Stronger Privacy Report

A.3.1 More Accurate Privacy Report by Adapting to the Data

We now present a more adaptive version of Algorithm 2 that could be even more accurate depending on the intrinsic stability of the dataset itself. The key technical components include:

- Adapting to a well-conditioned H by releasing $\lambda_{\min}$.
- A "regularized" construction of $\hat{\mu}^P(\cdot)$ that provides valid upper bounds of $\mu(\cdot)$ for all choices of $\lambda > 0$.

Algorithm 9 makes use of a subroutine to add noise to the smallest eigenvalue of H, presented below along with its privacy guarantees.

Algorithm 8 Releasing the smallest eigenvalue of H

Input: Dataset D, noise parameter σ_4, $\lambda_{\min}$ denoting the smallest eigenvalue of H.

Output: $\hat{\lambda}^P_{\min}$.

Output $\hat{\lambda}^P_{\min} = \lambda_{\min} + \mathcal{N}(0, \sigma_4^2)$.

Theorem A.3.1. *Algorithm 8 satisfies pDP with*

$$\epsilon_4(\cdot) = \frac{f''(\cdot)^2\|x\|^4}{2\sigma_4^2} + \frac{f''(\cdot)\|x\|^2\sqrt{2\log(1/\delta)}}{\sigma_4},$$

and if $f''(\cdot)\|x\|^2 \leq \beta$ *for all* x *then Algorithm 8 also satisfies* (ϵ, δ)*-DP with* $\epsilon = \frac{\beta^2}{2\sigma_4^2} + \frac{\beta\sqrt{2\log(1/\delta)}}{\sigma_4}$.

Proof. Algorithm 8 is a standard Gaussian mechanism. By Weyl's lemma, the smallest singular value satisfies a perturbation bound of $f''(\hat{\theta}^P; z)\|xx^T\|_2 = f''(\hat{\theta}^P; z)\|x\|^2$ from adding or removing one individual data point. The stated result follows from the theorem

of the Gaussian mechanism with per-instance (and global) sensitivity set as the above perturbation bound. □

In the more general smooth-loss case we can simply replace $f''(\hat{\theta}^P; z)\|x\|^2$ with $\|\nabla^2 \ell(\hat{\theta}^P; z)\|_F$.

Algorithm 9 More adaptive privacy report for `Obj-Pert`

Input: $\hat{\theta}^P$ from `Obj-Pert`, noise parameter $\sigma, \sigma_2, \sigma_3, \sigma_4$; regularization parameter λ; Hessian $H := \sum_i \nabla^2 \ell(\hat{\theta}^P; z_i) + \lambda I_d$, failure probability ρ.

Output: Reporting function $\tilde{\epsilon} : (x, y), \rho \to \mathbb{R}^2_+$.

Privately release $\hat{g}^P$ by Algorithm 10 with parameter σ_2.

Set $\epsilon_2(\cdot)$ according to Theorem A.3.4.

Set $\overline{g^P}(z) := f'(\cdot)[\hat{g}^P]^T x + \sigma_2 ||f'(\cdot)x||_2 F^{-1}_{\mathcal{N}(0,1)}(1 - \rho/2)$.

Set $\tau = F^{-1}_{\lambda_1(\mathrm{GOE}(d))}(1 - \rho/2)$.

Privately release $\hat{H}^P$ by Algorithm 11 with parameter σ_3.

Set $\epsilon_3(\cdot)$ according to Theorem A.4.1

Privately release $\hat{\lambda}^P_{\min} = \lambda_{\min} + \mathcal{N}(0, \sigma_4^2)$ (Algorithm 8).

Set $\epsilon_4(\cdot)$ according to Theorem A.3.1.

Set $\underline{\hat{\lambda}^P_{\min}} := \max\{\lambda, \hat{\lambda}^P_{\min} - \sigma_4 F^{-1}_{\mathcal{N}(0,1)}(1 - \rho/2)\}$.

if $\underline{\hat{\lambda}^P_{\min}} \geq 2\tau\sigma_3$ **then**

Set $\overline{\mu^P}(x) = \min\left\{\frac{\hat{\lambda}^P_{\min} + \tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}} x^T (\hat{H}^P)^{-1} x, \frac{\|x\|^2}{\underline{\hat{\lambda}^P_{\min}}}\right\}$. (use the standard estimator)

else

Set $\overline{\mu^P}(x) = \min\left\{\frac{\hat{\lambda}^P_{\min} + 2\tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}} x^T (\hat{H}^P + \tau\sigma_3 I_d)^{-1} x, \frac{\|x\|^2}{\underline{\hat{\lambda}^P_{\min}}}\right\}$. (use the regularized estimator)

end if

Set $\overline{\epsilon_1^P}(\cdot) := \left|-\log\left(1 - f''(\cdot)\overline{\mu^P}(x)\right)\right| + \frac{||f'(\cdot)x||_2^2}{2\sigma^2} + \frac{|\overline{g^P}(z)|}{\sigma^2}$.

Return the "privacy report" function $\tilde{\epsilon} = (\overline{\epsilon_1^P}, \epsilon_2 + \epsilon_3 + \epsilon_4)$, i.e., the *ex-post* pDP of Algorithm 1 and the pDP of Algorithm 9 (i.e., overhead).

This algorithm allows any choice of λ to be used in ObjPert, so that the privacy report is non-intrusive and can be attached to an existing workflow without changing the main algorithm at all. The following proposition shows that $\overline{\mu}^P(x)$ is always a valid upper bound of the leverage score $\mu(x)$ and it is accurate if $\lambda_{\min}$ is large (from either the Hessian or the regularization).

Proposition A.3.2 (Uniform multiplicative approximation). *Let* $\underline{\hat{\lambda}^P_{\min}}$ *and* $\hat{H}^P$ *be constructed as in Algorithm 9. Then with probability* $1-2\rho$*,*

$$\lambda_{\min} - \sigma_4 F^{-1}_{\mathcal{N}(0,1)}(1-\rho/2) \le \hat{\lambda}^P_{\min} \le \lambda_{\min} + \sigma_4 F^{-1}_{\mathcal{N}(0,1)}(1-\rho/2)$$

and for all $x \in \mathbb{R}^d$ *simultaneously, the regularized estimator obeys that*

$$x^T(\hat{H}^P + \tau\sigma_3 I_d)^{-1}x \le \mu(x) \le \frac{\underline{\hat{\lambda}^P_{\min}} + 2\tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}} x^T(\hat{H}^P + \tau\sigma_3 I_d)^{-1}x.$$

Moreover, under the same high-probability event, if $\underline{\hat{\lambda}^P_{\min}} \ge 2\tau\sigma_3$*, then the standard estimator obeys that*

$$\frac{\underline{\hat{\lambda}^P_{\min}} - \tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}} x^T(\hat{H}^P)^{-1}x \le x^T H^{-1} x \le \frac{\underline{\hat{\lambda}^P_{\min}} + \tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}} x^T(\hat{H}^P)^{-1}x.$$

Proof. By Lemma A.4.4, if we choose $\tau = F^{-1}_{\lambda_1(\mathrm{GOE}(d))}(1-\rho/2)$, then with probability $1-\rho$, the GOE noise matrix G satisfies that $\|G\|_2 \prec \tau$, the following holds: $-\tau I_d \prec G \prec \tau I_d$.

Next, by the definition of Gaussian CDF, with probability $1-\rho$,

$$\lambda_{\min} - \sigma_4 F^{-1}_{\mathcal{N}(0,1)}(1-\rho/2) \le \hat{\lambda}^P_{\min} \le \lambda_{\min} + \sigma_4 F^{-1}_{\mathcal{N}(0,1)}(1-\rho/2)$$

which implies that $\lambda_{\min} \geq \underline{\hat{\lambda}^P_{\min}}$, i.e.,

$$H - \underline{\hat{\lambda}^P_{\min}} I_d \succ 0$$

Therefore with probability $1 - 2\rho$,

$$\begin{aligned} H \prec H + G + \tau\sigma_3 I_d \prec H + 2\tau\sigma_3 I = H - \underline{\hat{\lambda}^P_{\min}} I_d + \underline{\hat{\lambda}^P_{\min}} I_d + 2\tau\sigma_3 I_d \\ \prec \frac{\underline{\hat{\lambda}^P_{\min}} + 2\tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}} (H - \underline{\hat{\lambda}^P_{\min}} I_d + \underline{\hat{\lambda}^P_{\min}} I_d) = \frac{\underline{\hat{\lambda}^P_{\min}} + 2\tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}} H, \end{aligned}$$

where the first semidefinite inequality uses that $H - \underline{\hat{\lambda}^P_{\min}} I_d$ is positive semi-definite.

Taking the inverse on both sides, we get

$$\frac{\underline{\hat{\lambda}^P_{\min}}}{\underline{\hat{\lambda}^P_{\min}} + 2\tau\sigma_3} H^{-1} \prec (\hat{H}^P + \tau\sigma_3 I_d)^{-1} \prec H^{-1}.$$

Thus for all $x \in \mathbb{R}^d$, $x^T(\hat{H}^P + \tau\sigma_3 I_d)^{-1} x \leq x^T H^{-1} x \leq \frac{\underline{\hat{\lambda}^P_{\min}} + 2\tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}} x^T(\hat{H}^P + \tau\sigma_3 I_d)^{-1} x$, which finishes the proof for the regularized estimator.

Now we turn to the standard (unregularized) estimator. Under the same high-probability event:

$$\begin{aligned} H + G \prec & H + \tau\sigma_3 I = H - \underline{\hat{\lambda}^P_{\min}} I_d + \underline{\hat{\lambda}^P_{\min}} I_d + \tau\sigma_3 I_d \\ & \prec \frac{\underline{\hat{\lambda}^P_{\min}} + \tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}} (H - \underline{\hat{\lambda}^P_{\min}} I_d + \underline{\hat{\lambda}^P_{\min}} I_d) = \frac{\underline{\hat{\lambda}^P_{\min}} + \tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}} H. \end{aligned}$$

Similarly,

$$H + G \succ H - \tau\sigma_3 I_d \succ H - \underline{\hat{\lambda}^P_{\min}} I_d + \underline{\hat{\lambda}^P_{\min}} I_d - \tau\sigma_3 I_d$$
$$\succ \frac{\underline{\hat{\lambda}^P_{\min}} - \tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}}(H - \underline{\hat{\lambda}^P_{\min}} I_d + \underline{\hat{\lambda}^P_{\min}} I_d) = \frac{\underline{\hat{\lambda}^P_{\min}} - \tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}} H.$$

Together the above two inequalities give

$$\frac{\underline{\hat{\lambda}^P_{\min}} - \tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}} H \prec H + G \prec \frac{\underline{\hat{\lambda}^P_{\min}} + \tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}} H.$$

Take the inverse on both sides we get

$$\frac{\underline{\hat{\lambda}^P_{\min}}}{\underline{\hat{\lambda}^P_{\min}} + \tau\sigma_3} H^{-1} \prec (\hat{H}^P)^{-1} \prec \frac{\underline{\hat{\lambda}^P_{\min}}}{\underline{\hat{\lambda}^P_{\min}} - \tau\sigma_3} H^{-1},$$

which implies that for all $x \in \mathbb{R}^d$, $\frac{\underline{\hat{\lambda}^P_{\min}} - \tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}} x^T (\hat{H}^P)^{-1} x \leq x^T H^{-1} x \leq \frac{\underline{\hat{\lambda}^P_{\min}} + \tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}} x^T (\hat{H}^P)^{-1} x$ as stated in the proposition. □

The privacy (DP and pDP) of Algorithm 9 is a composition of the stated results in Theorem 2.3.5 with the the privacy guarantees stated in Theorem A.3.1. Observe that if we choose $\sigma_3 = \sigma_1$ then the additional DP and pDP losses are smaller than those of the main algorithm, i.e., we have a constant overhead in terms of the privacy loss.

The next theorem shows that when $\lambda_{\min}(H) \to +\infty$ as the number of data points $n \to +\infty$, we could improve the leverage score part of the pDP losses from a multiplicative factor of 12 to $1 + o(1)$.

Theorem A.3.3 (Utility of Adaptive privacy report.). *Assume* $\lambda_{\min}(H) \geq \max\{2\beta, 2\tau\sigma_3\}$. *There is a universal constant* $0 < C \leq 4\tau\sigma_3 + 2\beta$ *such that for a fixed* $z \in \mathcal{X} \times \mathcal{Y}$, *and all*

$\rho > 0$, the privately released privacy report $\overline{\epsilon_1^P}(\cdot)$ from Algorithm 9 obeys that

$$\epsilon_1(\cdot) \leq \overline{\epsilon_1^P}(\cdot) \leq (1 + \frac{C}{\lambda_{\min}})\epsilon_1(\cdot) + \frac{|f'(\cdot)|\|x\|}{\sigma_2}\sqrt{2\log(2/\rho)}$$

with probability $1 - 3\rho$ where ϵ_1 is the expression from Theorem 2.3.1.

Proof of Theorem A.3.3. Similar to the proof of Theorem 2.3.5, it suffices to consider the approximation of the first term when we replace μ with $\overline{\mu^P}$. First of all, by a union bound, the high probability event in Proposition A.3.2 and the high probability event in Theorem 2.3.4 (to bound the third term in the *ex-post* pDP of `ObjPert`) holds simultaneously with probability at least $1 - 3\rho$. The remainder of the proof conditions on this event.

Observe that it suffices to construct a multiplicative approximation bound for the first term $\log(1 + f''(\cdot)\mu)$ or $-\log(1 - f''(\cdot)\mu)$.

By our assumption that $\lambda > 2\beta$, as well as the pointwise minimum in the construction of $\overline{\mu^P}$ from Algorithm 9, we know that $\overline{\mu^P} \leq 1/2$ and $\log(1 - f''(\cdot)\overline{\mu^P})$ is well-defined.

Using the fact that for all $a \geq -1$, $\frac{a}{1+a} \leq \log(1 + a) \leq a$, we will now derive the multiplicative approximation for both $\log(1 + f''(\cdot)\mu)$ or $-\log(1 - f''(\cdot)\mu)$ using the plug-ins: $\log(1 + f''(\cdot)\overline{\mu^P})$ or $-\log(1 - f''(\cdot)\overline{\mu^P})$.

For brevity, in the subsequent derivation we will be using a to denote $f''(\cdot)\mu(x)$ and $\hat{a}$ to denote $f''(x^T\hat{\theta}^P; y)\overline{\mu^P}(x)$.

Thus

$$\begin{aligned}\log(1 + a) \leq &\log(1 + \hat{a}) \leq \hat{a} \leq (1 + \frac{2\tau\sigma_3}{\underline{\hat{\lambda}_{\min}^P}})a \leq (1 + \frac{2\tau\sigma_3}{\underline{\hat{\lambda}_{\min}^P}})(1 + a)\log(1 + a) \\ \leq &(1 + \frac{4\tau\sigma_3}{\lambda_{\min}})(1 + \frac{\beta}{\lambda_{\min}})\log(1 + a) \leq (1 + \frac{C}{\lambda_{\min}})\log(1 + a)\end{aligned}$$

where C can be taken as $4\tau\sigma_3 + 2\beta$, by our assumption on $\lambda_{\min}$ and a high probability

bound under which $\underline{\hat{\lambda}^P_{\min}} \geq \lambda_{\min}/2$.

Similarly,

$$-\log(1-a) \leq \frac{a}{1-a} \leq \frac{\hat{a}}{1-a} \leq \frac{(1+\frac{2\tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}})a}{1-a} \leq \frac{(1+\frac{2\tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}})}{1-\frac{\beta}{\lambda_{\min}}}(-\log(1-a))$$

where

$$\frac{(1+\frac{2\tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}})}{1-\frac{\beta}{\lambda_{\min}}} = 1 + \frac{2\tau\sigma_3}{\underline{\hat{\lambda}^P_{\min}}} + \frac{\beta/\lambda_{\min}}{1-\beta/\lambda_{\min}} \leq 1 + \frac{4\tau\sigma_3 + 2\beta}{\lambda_{\min}}$$

under our assumption for $\lambda_{\min}$, β. The additive error term in the third term follows from the same bound as in the non-adaptive result without any changes.

The version for the standard (non-regularized) version is similar and is left as an exercise. □

A.3.2 Dataset-Dependent Privacy report for general smooth learning problems

So far, we have focused on generalized linear losses. Most of our results can be extended to general smooth learning problems.

For the third term in the pDP bound of Theorem 2.3.5, the challenge is that the two vectors are now nontrivially coupled with each other via $\hat{\theta}^P$. For this reason we propose to privately release the gradient at $\hat{\theta}^P$, which helps to decouple the dependence and allow a tighter approximation at a small cost of accuracy and additional privacy budget.

For convenience, we will denote $g = \nabla J(\hat{\theta}^P; D)^T \nabla \ell(\hat{\theta}^P; z)$. Below, we present an algorithm that outputs g^P (a private approximation of g) as well as the additional privacy cost $\epsilon_4(\cdot)$ of outputting g^P.

Algorithm 10 Release g^P, a private approximation of $g = \nabla J(\hat{\theta}^P; D)^T \nabla \ell(\hat{\theta}^P; z)$

Input: Dataset D, privatized output $\hat{\theta}^P$, noise parameter σ_2, linear loss function $L(\theta; D) = \sum_i \ell(\theta; z_i)$, regularization parameter λ, convex and twice-differentiable regularizer r.

Output: $g^P(\cdot), \epsilon_2(\cdot)$.

Construct noise vector $e \sim \mathcal{N}(0, \sigma_2^2 I)$.

Set $J^P := \nabla L(\hat{\theta}^P; D) + \nabla r(\theta) + \lambda \hat{\theta}^P + e$.

Set $g^P(\cdot)$ s.t. $g^P(z) = (J^P)^T \nabla \ell_z(\hat{\theta}^P; z)$.

Set $\epsilon_2(\cdot)$ s.t. $\epsilon_2(z) = \frac{\|\nabla \ell(\hat{\theta}^P; z)\|^2}{2\sigma_2^2} + \frac{\|\nabla \ell(\hat{\theta}^P; z)\| \sqrt{2 \log(2/\delta)}}{\sigma_2}$.

Theorem A.3.4. *Let $\hat{\theta}^P$ be fixed, Algorithm 10 satisfies*

1. *$(\epsilon_2(D, D_{\pm z}), \delta)$-pDP, with*

$$\epsilon_2(D, D_{\pm z}) = \frac{\|\nabla \ell(\hat{\theta}^P; z)\|^2}{2\sigma_2^2} + \frac{\|\nabla \ell(\hat{\theta}^P; z)\| \sqrt{2 \log(1/\delta)}}{\sigma_2}.$$

2. *$\epsilon_2(o, D, D_{\pm z})$-ex post pDP with probability $1 - \rho$,*

$$\epsilon_2(o, D, D_{\pm z}) = \frac{\|\nabla \ell(\hat{\theta}^P; z)\|^2}{2\sigma_2^2} + \frac{\|\nabla \ell(\hat{\theta}^P; z)\| \sqrt{2 \log(2/\rho)}}{\sigma_2}.$$

Proof. This is a Gaussian mechanism and the proof follows from Corollary A.6.2. □

The theorem avoids an additional dependence in d from the ℓ_1-norm $\|\nabla \ell(\hat{\theta}^P; z)\|_1$ in the dataset-independent bound.

We remark that Algorithm 10's pDP loss is dataset-independent and if we choose $\sigma_2 = \sigma_1$, the pDP losses for running Algorithm 10 are on the same order as those of the main algorithm. Thus the additional overhead is on the same order and no recursive privacy reporting is needed.

For the first term, our release of H and $\lambda_{\min}$ extends without any changes to the more general case. The estimator of the leverage score needs to be modified accordingly. Add the plug-in estimator that replaces H with $\hat{H}^P$ in the general case here.

We defer the analysis of how accurately this estimator approximates the first term of $\epsilon_1(\cdot)$ to a longer version of the paper.

A.3.3 Uniform Privacy Report and Privacy Calibration

The "privacy report" algorithm (Algorithm 2) that we presented in the main paper and the "adaptive privacy report" (Algorithm 9 is straightforward and omitted. focus on releasing a reporting function $\tilde{\epsilon}$ that is accurate with high probability for every fixed input.

Sometimes there is a need to ensure that with high probability, $\tilde{\epsilon}$ is accurate *simultaneously* for all $z_1, ..., z_n$ in the dataset, or even for all $z \in \mathcal{Z}$ for a data domain $\mathcal{Z}$. The following theorem shows that this is possible at a mild additional cost in the accuracy. These results are stated for Algorithm 2), but extensions to that of Algorithm 9 is straightforward and thus omitted.

Proposition A.3.5 (Uniform privacy report). *With probability* $1-2\rho$, simultaneously *for all n users in the dataset, the output of Algorithm 2 obeys that* $\epsilon_1(\hat{\theta}^P, D, D_{\pm z}) \leq \overline{\epsilon_1^P}(\hat{\theta}^P, z) \leq 12\epsilon_1(\hat{\theta}^P, D, D_{\pm z}) + \frac{|f'(\cdot)|\|x\|}{\sigma_2}\sqrt{2\log(n/\rho)}$.

If we, instead, use the data-independent bound $\frac{|f'(x^T\hat{\theta}^P;y)|\|x\|_1\sqrt{2\log(2d/\rho)}}{\sigma}$ *to replace the third-term in* $\overline{\epsilon_1^P}(\cdot)$, *then with probability* $1-2\rho$, simultaneously *for all* $x \in \mathcal{X}$, *the ex-post pDP report* $\overline{\epsilon_1^P}$ *from Algorithm 2 satisfies that*

$$\epsilon_1(\cdot) \leq \overline{\epsilon_1^P}(z, \hat{\theta}^P) \leq 12\epsilon_1(\cdot) + \frac{|f'(\cdot)|\|x\|_1\sqrt{2\log(2d/\rho)}}{\sigma}.$$

Proof. We note that the approximation of μ_x is uniform for all x. It remains to consider

a uniform bound for the third term over the randomness of ObjPert. The first statement follows by taking a union bound. The second result is achieved by Holder's inequality, the concentration of max of i.i.d. Gaussians. □

Sometimes it is desirable to calibrate the noise-level to a prescribed "worst-case" DP parameter ϵ, δ. The following corollary explains that the additional DP loss and pDP losses when we calibrate Algorithm 2 with the same privacy parameter as those in Algorithm 1 will yield a total DP and pDP that are at most twice as large under an additional condition that $f'' \leq f'$.

Corollary A.3.6 (The additional privacy cost). *If we calibrate σ_2 such that the Algorithm 2 satisfies the same (ϵ, δ)-DP as Algorithm 1, i.e., when $\epsilon < 1$, we could choose $\sigma_2 = \frac{\rho_{\max}}{\epsilon}\sqrt{2\log(1.25/\delta)}$. Then Algorithm 2 satisfies $(\epsilon(\cdot), \delta)$-pDP with*

$$\epsilon(\cdot) = \frac{\epsilon^2 (f''(\cdot))^2 \|x\|^4}{8\rho_{\max}^2 \log(1.25/\delta)} + \frac{\epsilon(f''(\cdot))\|x\|^2}{\rho_{\max}\sqrt{2}}.$$

For those cases when $\frac{(f''(\cdot))\|x\|^2}{\rho_{\max}} \leq \frac{|f'(\cdot)|\|x\|}{\beta}$ (which is the case in logistic regression for all x s.t., $\|x\| \leq 1$), the additional overhead in releasing a dataset-dependent pDP is smaller than the ex post pDP bound in Theorem 2.3.1.

A.4 Improved "Analyze Gauss" with Gaussian Orthogonal Ensembles

In this section we propose a differentially private mechanism that releases a matrix H when

$$H = \sum_{i=1}^{n} H_x$$

where $H_x \in \mathbb{R}^{d\times d}$ is a symmetric matrix computed from individual data point x.

Examples of this include

1. (unnormalized / uncentered) sample covariance $H_x = xx^T$
2. Empirical Fisher information $H_x = \nabla\ell(\theta; x)\nabla\ell(\theta; x)T$ where ℓ is the log-likelihood and θ is the true parameter;
3. Hessian of a generalized linear loss function $H_x = f''(x, \theta)xx^T$.
4. Hessian of a smooth loss function $H_x = \nabla^2\ell(x, \theta)$.

In the first three cases H_x is a rank-1 matrix and our use case in this paper is the third and fourth example. Throughout this section we assume $\|H_x\|_F \leq \beta$ for all $x \in \mathcal{X}$.

The mechanism we propose is a variant of "Analyze-Gauss" (Dwork et al., 2014b) but it reduces the required variance of the added noise by a factor of 2 in almost all coordinates hence resulting in higher utility.

The standard "Analyze-Gauss" leverages the symmetry of H and uses the standard Gaussian mechanism to release the upper triangular region (including the diagonal) of the matrix H with an ℓ_2-sensitivity upper bound:

$$\|\text{UpperTriangle}(H) - \text{UpperTriangle}(H')\|_2 \leq \|H_x\|_F \leq \beta.$$

where $\text{UpperTriangle}(H) \in \mathbb{R}^{d^2/2+d/2}$ is the vector that enumerates the elements of the upper-triangular region of H. The resulting Gaussian noise is distributed i.i.d as $\mathcal{N}(0, \sigma_3^2)$ and it satisfies (ϵ, δ)-DP with

$$\epsilon = \frac{\beta^2}{2\sigma_3^2} + \frac{\beta\sqrt{2\log(1/\delta)}}{\sigma_3}.$$

The alternative that we propose also adds a symmetric noise but doubles the variance on the diagonal elements.

Algorithm 11 Release H (a natural variant of "Analyze-Gauss")

Input: Dataset D, noise parameter σ_3, $H = \sum_{i=1}^{n} \nabla^2 \ell(z_i, \hat{\theta}^P) + \lambda I_d$.

Output: $\hat{H}^P$.

Draw a Gaussian random matrix $Z \in \mathbb{R}^{d \times d}$ with $Z_{i,j} \sim \mathcal{N}(0, \sigma_3^2)$ independently.

Output $\hat{H}^P = H + \frac{1}{\sqrt{2}}(Z + Z^T)$.

The symmetric random matrix $\frac{1}{\sqrt{2}}(Z + Z^T)$ is known as the Gaussian Orthogonal Ensemble (GOE) and well-studied in the random matrix theory. We will first show this this mechanism obeys DP and pDP.

Theorem A.4.1. *Algorithm 11 satisfies pDP with*

$$\epsilon(\cdot) = \frac{\|H_x\|_F^2}{4\sigma_3^2} + \frac{\|H_x\|_F \sqrt{2\log(1/\delta)}}{\sqrt{2}\sigma_3},$$

and $\hat{H}^p$ satisfies ex post pDP of the same ϵ with probability $1 - 2\delta$. If in addition $\sup_{x \in \mathcal{X}} \|H_x\|_F \leq \beta$ then, Algorithm 11 satisfies (ϵ, δ)-DP with

$$\epsilon \leq \frac{\beta^2}{4\sigma_3^2} + \frac{\beta\sqrt{2\log(1/\delta)}}{\sqrt{2}\sigma_3}.$$

Improvements over "Analyze Gauss". Notice that if we choose σ_3 to be $1/\sqrt{2}$ of the noise scale with used in the standard "Analyze Gauss", we will be adding the same amount of noise on the diagonal, achieve the same DP and pDP bounds, while adding noise with only half the variance in the off-diagonal elements. The idea is to add noise with respect to the natural geometry of the sensitivity, as we illustrate in the proof.

Proof. Algorithm 11 is equivalent to releasing the vector $[f_1, f_2]$ using a standard Gaussian mechanism with $\mathcal{N}(0, \sigma_3^2 I_{\frac{d^2}{2}+d/2})$, where $f_1 \in \mathbb{R}^d$ is the diagonal of $H/\sqrt{2}$ and $f_2 \in \mathbb{R}^{(d^2-d)/2}$ is the vectorized the strict upper triangular part of H.

The per-instance ℓ_2-sensitivity of $[f_1, f_2]$ is

$$\begin{aligned}\|\Delta_x\|_2 &= \sqrt{\sum_{1\le i<j\le d} H_x[i,j]^2 + \sum_{k=1^d} H_x[k,k]^2(1/\sqrt{2})^2} \\ &= \sqrt{\frac{1}{2}\left(\sum_{1\le i<j\le d} H_x[i,j]^2 + \sum_{1\le j<i\le d} H_x[i,j]^2 + \sum_{k=1^d} H_x[k,k]^2\right)} \\ &= \frac{1}{\sqrt{2}}\|H_x\|_F\end{aligned}$$

The result then follows from an application of the pDP computation of the Gaussian mechanism. □

A.4.1 Exact statistical inference with the Gaussian Orthogonal Ensemble

Besides a constant improvement in the required noise, another major advantage of using the Gaussian Orthogonal Ensemble is that we know the exact distribution of its eigenvalues (Chiani, 2014) which makes statistical inference, e.g., constructing confidence intervals, easy and constant-tight.

Lemma A.4.2 (Largest singular value of Gaussian random matrix (Rudelson and Vershynin, 2010, Equation (2.4))). *Let $A \in \mathbb{R}^{d\times d}$ be a random matrix with i.i.d. σ^2-subgaussian entries, then there exists universal constants C, c such that for all $t > 0$*

$$\mathbb{P}[s_{\max}(A) \ge (2+t)\sqrt{d\sigma^2}] \le Ce^{-cdt^{3/2}}.$$

i.e., with probability $1-\delta$

$$\|A\|_2 \le \left(2 + \left(\frac{(\log(C/\delta))}{cd}\right)^{2/3}\right)\sqrt{d\sigma^2}.$$

Notice that the symmetric matrix, i.e., Gaussian orthogonal ensemble is identically distributed to $\frac{1}{\sqrt{2}}(Z + Z^T)$ where Z is a iid Gaussian random matrix, thus by triangular inequality, we have

Lemma A.4.3 (Largest eigenvalue of Gaussian orthogonal ensemble). *Let A be a Gaussian orthogonal ensemble (i.e., a symmetric random matrix with $\mathcal{N}(0, \sigma^2)$ on the off-diagonal and $\mathcal{N}(0, 2\sigma^2)$ on the diagonal), with probability $1 - \delta$,*

$$\|A\|_2 \leq \sqrt{2}\left(2 + \left(\frac{(\log(C/\delta))}{cd}\right)^{2/3}\right)\sqrt{d\sigma^2}.$$

Proof. The proof follows from triangular inequality of the spectral norm. □

The above bound is asymptotic and we will use it for deriving the theoretical results. For practical computation, the the exact formula of the CDF of the largest eigenvalue of GOE matrices is given by (Chiani, 2014, Theorem 2). We could use this to bound the spectral norm of the noise added to Algorithm 11.

Lemma A.4.4. *Let A be described as in Lemma A.4.3.*

$$\|A\|_2 \leq \sigma F^{-1}_{\lambda_1 \text{ of } GOE}(1 - \rho/2)$$

where $F_{\lambda_1 \text{ of } GOE}$ is the CDF of the largest eigenvalue of the standard GOE matrix with constructed by $\frac{1}{\sqrt{2}}(Z + Z^T)$ where each element of matrix Z is drawn i.i.d. from a standard gaussian.

Proof. Notice that the GOE matrix is symmetric, so the largest eigenvalue λ_1 and the negative of the smallest eigenvalue $-\lambda_d$ are identically distributed. Thus the operator norm $\|A\|_2 \leq \max\{|\lambda_1|, |\lambda_d|\} \leq F^{-1}_{\lambda_1 \text{ of GOE}}(1 - \rho/2)$ with probability $1 - \rho$. □

Numerical computation: Chiani (2014, Theorem 2) characterized the distribution of λ_1 and provided an exact analytical formula with stable numerical implementation to compute $F_{\lambda_1 \text{ of GOE}}$. Thus $F^{-1}_{\lambda_1 \text{ of GOE}}$ can be evaluated using a binary search.

Using the Mathematica implementation provided by (Chiani, 2014), we find that $F^{-1}_{\lambda_1 \text{ of GOE}(50)}(1-\rho/2) = 12$ for $\rho = 8.465 \times 10^{-6}$. Therefore in our experiments with $d = 50$, we choose $\tau \approx 12$.

Explain how we got that $1.7\sqrt{d}$ constant here (what is the corresponding ρ when we choose 1.7?). The author has also provided a Mathematica implementation. I don't have one installed but I think UCSB has it for free for students. I think it will be an ease of mind if we can try running the authors' implementation to figure out and comment here that we got this number not just from the figure but also from the author's provided code. You may find the code here `https://sites.google.com/site/marcochianigroup/articles`

This is of a slightly lower-priority, but I think we should provide an implementation of this in python and make it available. A general inference tool for the GOE-Analyze-Gauss is very useful.

A.5 Omitted Proofs

With the two technical components presented, we are now ready to present the detailed proofs of our main results: Theorem 2.3.1 and Theorem 2.3.5.

A.5.1 Proofs for the pDP analysis of objective perturbation

appendix/proofs/pdp$_a lg1_p roof$

Proof of Theorem 2.3.3. Using the eigendecomposition $\nabla^2 \ell(\hat{\theta}^P; z) = \sum_{k=1}^{d} \lambda_k u_k u_k^T$, for

$0 \leq j \leq d$ we have that

$$\mu_j(x) = \begin{cases} \lambda_j u_j^T \Big(-\nabla^2 L(\hat{\theta}^P; D) - \lambda I_d - \nabla^2 r(\hat{\theta}^P) - \sum_{k=1}^{j-1} \lambda_k u_k u_k^T \Big)^{-1} u_j & \text{if } z \notin D \\ \lambda_j u_j^T \Big(-\sum\limits_{\substack{z_i \in D \\ z_i \neq z}} \nabla^2 \ell(\hat{\theta}^P; z_i) - \lambda I_d - \nabla^2 r(\hat{\theta}^P) - \sum_{k=j}^{d} \lambda_k u_k u_k^T \Big)^{-1} u_j & \text{if } z \in D. \end{cases}$$

$$:= \begin{cases} \lambda_j u_j^T H_{+z}^{-1} u_j & \text{if } z \notin D \\ \lambda_j u_j^T H_{-z}^{-1} u_j & \text{if } z \in D. \end{cases}$$

The second equality introduces the shorthand $\mu_j(x) := \lambda_j u_j^T H_{\pm z}^{-1} u_j$. Observe that $\nabla^2 \ell(\hat{\theta}^P; z_i), \nabla^2 r(\hat{\theta}^P) \in \mathbb{R}^{d \times d}$ are positive semi-definite, since $\ell(\cdot)$ and $r(\theta)$ by assumption are convex functions with continuous second-order partial derivatives. Since $\nabla^2 \ell(\hat{\theta}^P; z_i)$ is PSD, its eigenvalues are non-negative and so $\lambda_k \geq 0$ for all $0 \leq k \leq d$. Then for any $x \in \mathbb{R}^d$, $x^T u_k u_k^T x = (x^T u_k)^2 \geq 0$. So $u_k u_k^T$ is also PSD, and we then have that $H_{+z} + \lambda I_d$ and $H_{-z} + \lambda I_d$ are both negative semi-definite. Therefore, $H_{\pm z} \prec -\lambda I_d$ and after taking the inverse, we see that $\mu_j(x) \leq -\frac{\lambda_j}{\lambda} \leq 0$ or equivalently $-\mu_j(x) \geq \frac{\lambda_j}{\lambda} \geq 0$.

For $-1 < \mu_j(x) \leq 0$, we have that

$$\begin{aligned} \big| -\log(1 - \mu_j(x)) \big| &= \log(1 + (-\mu_j(x))) \\ &\leq -\mu_j(x) \\ &\leq -\log(1 + \mu_j(x)) \\ &= \big| -\log(1 + \mu_j(x)) \big| \\ &\leq -\log(1 - \frac{\lambda_j}{\lambda}). \end{aligned}$$

The rest of the proof follows from converting the log-product into a sum of logs. For a linear loss function $\ell(\theta; z) = f(x^T \theta; y)$, the simplified bound can be achieved due to the rank-one

Hessian $\nabla^2 \ell(\hat{\theta}^P; z) = f''(x^T\theta; y)xx^T$ whose only eigenvalue is $\lambda_1 = f''(x^T\hat{\theta}^P; y)||x||_2^2$. □

Proof of Theorem 2.3.4. By Holder's inequality,

$$\left|\nabla J(\hat{\theta}^P; D)\nabla \ell(\hat{\theta}^P; z)\right| \leq ||\nabla J(\hat{\theta}^P; D)||_\infty ||\ell(\hat{\theta}^P; z)||_1.$$

Recall from (**??**) that $\nabla J(\hat{\theta}^P; D) = -b(\hat{\theta}^P; D)$. Therefore $||\nabla J(\hat{\theta}^P)||_\infty = \max_{i\in[d]} |b_i|$, where $b_i \sim \mathcal{N}(0, \sigma^2)$. Applying a union bound and using the standard Gaussian tail bound,

$$\begin{aligned}
\Pr\left[\max_{i\in[d]} |b_i| \geq t\right] &= \Pr\left[\bigcup_i |b_i| \geq t\right] \\
&\leq \sum_{i\in[d]} \Pr\big[|b_i| \geq t\big] \\
&\leq 2de^{-\frac{t^2}{2\sigma^2}}.
\end{aligned}$$

So with probability $1-\rho$, we have $||\nabla^J(\hat{\theta}^P; D)||_\infty \leq \sigma\sqrt{2\log(2d/\rho)}$. The stronger bound for linear loss functions comes from substituting $||\nabla \ell(\hat{\theta}^P)||_1 = f'(x^T\theta; y)||x||_1$. □

A.5.2 Proofs for the Privacy Report in the main paper

The proof of Theorem 2.3.5 relies on the following intermediate result.

Proposition A.5.1 (Uniform multiplicative approximation). *If* $\lambda_{\min}(H) \geq 2\sigma_2 F^{-1}_{\lambda_1(\mathrm{GOE}(d))}(1-\rho/2)$*, then with probability* $1-\rho$*, for all* $x \in \mathbb{R}^d$ *simultaneously*

$$\frac{1}{2}x^T(\hat{H}^P)^{-1}x \leq x^T H^{-1} x \leq \frac{3}{2}x^T(\hat{H}^P)^{-1}x.$$

Proof. By the choice of $\tau = F^{-1}_{\lambda_1(\mathrm{GOE}(d))}(1-\rho/2)$, with probability $1-\rho$, the noise matrix Z from the release of $\hat{H}^P$ satisfies that $\|Z\|_2 \leq \sigma_2\tau \leq \lambda_{\min}/2$. Thus $-\frac{H}{2} \prec -\frac{\lambda_{\min}}{2}I_d \prec$

$Z \prec \frac{\lambda_{\min}}{2} I_d \prec \frac{H}{2}$. Adding H on both sides

$$\frac{H}{2} \prec H + Z \prec \frac{3H}{2}$$

which implies that

$$\frac{2}{3} H^{-1} \leq (H+Z)^{-1} \prec 2H^{-1}.$$

By definition of semidefinite ordering, for all $x \in \mathbb{R}^d$

$$\frac{2}{3} x^T H^{-1} x \leq x^T (H+Z)^{-1} x \leq 2 x^T H^{-1} x.$$

In other word, $\frac{1}{2}\hat{\mu}_1^p(x) \leq \mu_1(x) \leq \frac{3}{2}\hat{\mu}_1^p(x)$. □

Proof of Theorem 2.3.5. The privacy guarantees (Statement 1-3) follow directly from the pDP analysis in Theorem A.4.1 that analyzes the release of H by adding a GOE noise matrix and the Gaussian mechanism that releases g.

By the result follows from Proposition A.5.1 we know that with probability $1-\rho$, for all x

$$\mu(x) \leq \frac{3}{2}\hat{\mu}^p(x) \leq 3\mu(x)$$

For all $a \geq -1$ $\frac{a}{1+a} \leq \log(1+a) \leq a$. Recall that $\beta \geq \sup_z \|\nabla^2 \ell(\hat{\theta}^p; z)\|2$. By our condition that $\lambda > 2\beta$, as well as the pointwise minimum in the construction of $\overline{\mu^p}$, we have that $f''\overline{\mu^p} \leq \frac{1}{2}$ and

$$\frac{f''\overline{\mu^p}}{2} \leq \max\{\log(1+f''\overline{\mu^p}), -\log(1-f''\overline{\mu^p}\} \leq 2f''\overline{\mu^p}.$$

Thus

$$\log(1+f''\mu) \leq f''\mu \leq f''\overline{\mu^p} \leq 2\log(1+f''\overline{\mu^p}) \leq 2f''\overline{\mu^p} \leq 3f''\hat{\mu}^p \leq 6f''\mu \leq 12\log(1+f''\mu),$$

and similarly

$$-\log(1-f''\mu) \le 2f''\mu \le 2\overline{f''\mu^p} \le -2\log(1-f''\overline{\mu^p}) \le 4f''\overline{\mu^p} \le 6f''\hat{\mu}^p \le 12f''\mu \le -12\log(1-f''\mu).$$

This concludes the factor 12 multiplicative approximation in the first term of $\epsilon_1(\cdot)$. The second term of $\epsilon_1(\cdot)$ does not involve an approximation. The third term of $\epsilon_1(\cdot)$ is random and the bound is off by an additive factor of $\min\{\sigma, \sigma_2\}|f'(\cdot)|\|x\|_2\sqrt{2\log(2/\rho)}$ — via the smaller of the data-dependent bound and the data-independent bound, each holds with probability $1-\rho/2$. □

A.6 pDP Analysis of the Gaussian mechanism

Theorem A.6.1 (*ex-post* pDP of Gaussian mechanism). *Let $Q : \mathcal{Z}^* \to \mathbb{R}^d$ be a function of the data. Let $|Q(D_{\pm z}) - Q(D)| \le \Delta_z$. Then the Gaussian mechanism that releases $o \sim Q(D) + \mathcal{N}(0, \sigma^2 I_d)$ obeys* ex-post *pDP with*

$$\epsilon(o, D, D_z) = \left| \frac{\|\Delta_z\|^2}{2\sigma^2} - \frac{\Delta_z^T(o - Q(D))}{\sigma^2} \right|.$$

Proof. We can directly calculate the log-odds ratio:

$$\begin{aligned}
&\frac{1}{2\sigma^2}\left(\|o - Q(D)\|^2 - \|o - Q(D_{\pm z})\|^2\right) \\
=&\frac{1}{2\sigma^2}\left((Q(D_{\pm z}) - Q(D))^T(2o - Q(D) - Q(D_{\pm z}))\right) \\
=&\frac{1}{2\sigma^2}\left(\Delta_z^T(2o - 2Q(D) - \Delta_z)\right) \\
=&\frac{-\|\Delta_z\|^2}{2\sigma^2} + \frac{\Delta_z^T(o - Q(D))}{\sigma^2}.
\end{aligned}$$

The proof is complete by taking the absolute value. □

Corollary A.6.2 (pDP bound and high-probability ex-post pDP of Gaussian mechanism). *Let* Φ *be the cumulative distribution function (CDF) of a standard normal random variable. The Gaussian mechanism that releases* $o \sim Q(D)+\mathcal{N}(0, \sigma^2 I_d)$ *satisfies dataset independent pDP bound with*

$$\epsilon(D, D_{\pm z}) \leq \frac{\|\Delta_z\|^2}{2\sigma^2} + \frac{\|\Delta_z\|\Phi^{-1}(1-\delta)}{\sigma} \leq \frac{\|\Delta_z\|^2}{2\sigma^2} + \frac{\|\Delta_z\|\Phi^{-1}(1-\delta)}{\sigma}.$$

Moreoever, with probability at least $1-\rho$ *over the distribution of the randomized output* o*, the Gaussian mechanism satisfies obeys the following dataset-independent* ex post *pDP bound*

$$\epsilon(o, D, D_{\pm z}) \leq \frac{\|\Delta_z\|^2}{2\sigma^2} + \frac{\|\Delta_z\|\Phi^{-1}(1-\rho/2)}{\sigma} \leq \frac{\|\Delta_z\|^2}{2\sigma^2} + \frac{\|\Delta_z\|\sqrt{2\log(2/\rho)}}{\sigma}. \quad \text{(A.6.1)}$$

Proof. Since $o \sim Q(D)+\mathcal{N}(0, \sigma^2 I_d)$, we have $\Delta_z^T(o-Q(D)) \sim \mathcal{N}(0, \sigma^2\|\Delta_z\|^2)$. The results of pDP follows from the tailbound of the privacy loss random variable and Lemma A.7.4.

For the high-probability bound of the *ex post* pDP, we need to bound both sides of the privacy loss random variable. It suffices to show that the absolute value of the added noise is bounded with a union bound on the two-sided tails, each with probability $1-\rho/2$. □

A tighter pDP bound can be obtained using the analytical Gaussian mechanism (Balle and Wang, 2018). We choose to present the tail bound-based formula above for the interpretability of the results.

A.7 Technical Lemmas

Lemma A.7.1 (Sherman-Morrison-Woodbury Formula). *Let A, U, C, V be matrices of compatible size. Assuming A, C and $C^{-1} + VA^{-1}U$ are all invertible, then*

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

Lemma A.7.2 (Determinant of Rank-1 perturbation). *For invertible matrix A and vector c, d of compatible dimension*

$$\det(A + cd^T) = \det(A)(1 + d^T A^{-1}c).$$

Lemma A.7.3 (Gaussian tail bound). *Let $X \sim \mathcal{N}(0, \sigma^2)$. Then*

$$\mathbb{P}(X > \sigma\epsilon) \leq \frac{e^{-\epsilon^2/2}}{\epsilon}.$$

A convenient alternative representation (slightly weaker) is

$$\mathbb{P}(X > \sigma\sqrt{2\log(1/\delta)}) \leq \delta,$$

and

$$\mathbb{P}(|X| > \sigma\sqrt{2\log(2/\delta)}) \leq \delta.$$

for all $\delta > 0$.

Lemma A.7.4 (Tail bound to (ϵ, δ)-DP conversion). *Let $\epsilon(o) = \log(\frac{p(o)}{p'(o)})$ where p and p' are densities of θ. If*

$$\mathbb{P}_p(\epsilon(o) > \epsilon) \leq \delta$$

then for any measurable set $\mathcal{S}$

$$\mathbb{P}_p(\theta \in \mathcal{S}) \leq e^{\epsilon}\mathbb{P}_{p'}(\theta \in \mathcal{S}) + \delta.$$

Two useful applications of this result for DP are:

1. *if* $\mathbb{P}_p(\epsilon(o) > \epsilon) \leq \delta$ *for all pairs of neighboring dataset* D, D' *such that* $p = \mathcal{A}(D), p' = \mathcal{A}(D')$ *then* $\mathcal{A}$ *is* (ϵ, δ)*-DP.*

2. *If* $D' = D_{\pm z}$, $p = \mathcal{A}(D), p' = \mathcal{A}(D_{\pm z})$ *and that* $\mathbb{P}_p(\epsilon(o) > \epsilon) \leq \delta$ *and* $\mathbb{P}_{p'}(-\epsilon(o) < -\epsilon) \leq \delta$*, then* $\mathcal{A}$ *satisfies* (ϵ, δ)*-pDP for individual* z *and dataset* D*.*

Proof. Let E be the event that $|\epsilon(\theta)| > t$, by definition it implies that for any $\tilde{E} \subset E$, $\mathbb{P}_p(\theta \in \tilde{E}) \leq e^t\mathbb{P}_{p'}(\theta \in \tilde{E})$. Now consider any measurable set $\mathcal{S}$:

$$\begin{aligned}\mathbb{P}_p(\theta \in \mathcal{S}) &= \mathbb{P}_p(\theta \in \mathcal{S} \cap E^c) + \mathbb{P}_p(\theta \in \mathcal{S} \cap E) \\ &\leq \mathbb{P}_{p'}(\theta \in \mathcal{S} \cap E^c)e^t + \mathbb{P}_p(\theta \in E) \leq e^t\mathbb{P}_{p'}(\theta \in \mathcal{S}) + \delta.\end{aligned}$$

The two applications follow directly from the definitions of (ϵ, δ)-DP and pDP. □

Lemma A.7.5 (maximum of subgaussian). *Let* $X_1, ..., X_n$ *be iid* σ^2*-subgaussian random variables.*

$$\mathbb{P}[\max_i X_i \geq \sqrt{2\sigma^2(\log n + t)}] \leq e^{-t}.$$

Proof. The proof is by standard subgaussian concentration and union bound. □

Lemma A.7.6 (Weyl's theorem; Theorem 4.11, p. 204 in Stewart (1990)). . *Let* A, E *be given* $m \times n$ *matrices with* $m \geq n$*, then*

$$\max_{i \in [n]} |\sigma_i(A) - \sigma_i(A + E)| \leq \|E\|_2 \tag{A.7.1}$$

Lemma A.7.7 ("Change-of-variables" for density functions). *Let $g : \mathbb{R}^d \to \mathbb{R}^d$ be a bijective and differentiable function, and let X, Y be continuous random variables in $\mathbb{R}^d$ related by the transformation $Y = g(X)$. Then the probability density of Y is*

$$f_Y(y) = f_X(g^{-1}(y)) \left| det \left[\frac{\partial g^{-1}(y)}{\partial y} \right] \right|,$$

with $\left[\frac{\partial g^{-1}(y)}{\partial y} \right]$ denoting the $d \times d$ Jacobian matrix of the mapping $X = g^{-1}(Y)$.

Lemma A.7.8. (Billboard lemma) *Suppose $\mathcal{A} : \mathcal{D} \to \mathcal{R}$ satisfies (ϵ, δ) differential privacy. Consider any set of functions $f_i : \mathcal{D}_i \times \mathcal{R} \to \mathcal{R}$, where $\mathcal{D}_i$ is the portion of the dataset containing individual i's data. The composition $\{f_i(\Pi_i D, \mathcal{A}(D))\}$ satisfies (ϵ, δ)-joint differential privacy, where $\Pi_i : \mathcal{D} \to \mathcal{D}_i$ is the projection to individual i's data.*

Appendix B

Supplementary Material for Chapter 3

B.1 Summary of PTR Variants

	PTR	Generalized PTR
Private Test	Test $\Delta_{\text{LS}} \leq \beta$ for a proposed bound β, then add noise $\propto \beta$ if the test passes (Vadhan, 2017, Sec 3.2).	Test $\epsilon_\phi \leq \epsilon$ for a proposed parameter ϕ, then run $\mathcal{M}_\phi$ if the test passes (Alg 4)
Private pointwise bounds	no analogous algorithm	Release $\bar{\epsilon}$ s.t. $\epsilon_\phi \leq \bar{\epsilon}$ for a *fixed* ϕ w.h.p. for general randomized mechanism $\mathcal{M}_\phi$, then run $\mathcal{M}_\phi$ if $\bar{\epsilon} \leq \epsilon$ (Alg 4).
Private uniform bounds	Release $\bar{\Delta}$ s.t. $\Delta_{\text{LS}} \leq \bar{\Delta}$ w.h.p for a noise-adding mechanism with noise $\propto \bar{\Delta}$ (Vadhan, 2017, Sec 3.4). (Choose appropriate noise level σ, no $\perp$.)	Release $\bar{\epsilon}_\phi$ s.t. $\epsilon_\phi \leq \bar{\epsilon}_\phi$ for *all* ϕ w.h.p. for general randomized mechanism $\mathcal{M}_\phi$ (Choose appropriate ϕ, no $\perp$, as in Alg 12)
Stability-based	Test $\Delta_{\text{LS}} = 0$ before releasing stable numerical value deterministically (Vadhan, 2017, Sec 3.3).	Test $\epsilon_\phi = 0$ before releasing stable general output deterministically (special case of Alg 4).
What to propose?	Select $\beta \in \{\beta_1, ..., \beta_M\}$ s.t. $\Delta_{\text{LS}} \leq \beta$ passes the test (using e.g. AboveThreshold)[1]	Select $\phi \in \{\phi_1, ..., \phi_M\}$, s.t. ϵ_ϕ passes the test (using private selection as in Alg 5).

The table above compares our generalization to the standard variants of PTR. Vanilla PTR, typically implemented using a distance test, was proposed originally in Dwork and Lei (2009). The stability-based argument was originally proposed by Thakurta and Smith (2013). We are citing the book of Vadhan (2017) for a clean treatment to these PTR-like mechanisms. The corresponding generalized version are from this paper.

B.2 Omitted algorithms and proofs in Section 3.5

B.2.1 Main privacy result of Theorem 3.5.3

Proof of Theorem 3.5.3. The proof of our main privacy result relies on two central properties of differential privacy: composition and immunity to post-processing. We review these below.

Theorem B.2.1 (Composition (Dwork et al., 2014a)). *For $i \in [k]$, let $\mathcal{M}_i : \mathcal{Z} \to \mathcal{R}_i$ be a randomized algorithm satisfying (ϵ_i, δ_i)-DP. Define the mechanism $\mathcal{M} : \mathcal{Z} \to \prod_{i=1}^{k} \mathcal{R}_i$ as $\mathcal{M}(Z) = (\mathcal{M}_1(Z), \mathcal{M}_2(Z), \ldots, \mathcal{M}_k(Z))$. Then $\mathcal{M}$ satisfies $\left(\sum_{i=1}^{k} \epsilon_i, \sum_{i=1}^{k} \delta_i\right)$-DP.*

Theorem B.2.2 (Closure under post-processing (Dwork et al., 2014a)). *Consider a mechanism $\mathcal{M} : \mathcal{Z} \to \mathcal{R}$ that satisfies (ϵ, δ)-DP. Let $f : \mathcal{R} \to \mathcal{R}'$ be a data-independent (randomized or deterministic) mapping. Then $f \circ \mathcal{M}$ satisfies (ϵ, δ)-DP.*

Let $\mathcal{M}$ denote the mechanism described in Algorithm 4. We split the input space $\mathcal{X}$ into two cases.

Case I: $\epsilon_\phi(X) > \epsilon$

We restrict the input space to $\tilde{\mathcal{X}} = \{X \in \mathcal{X} \mid \epsilon_\phi(X) > \epsilon\}$, for $\mathcal{M} : \tilde{\mathcal{X}} \to \mathcal{R} \cup \{\perp\}$. Let E be the event $\mathcal{T}(X) = 1$ and consider a possible output set $S \subseteq \mathcal{R} \cup \{\perp\}$. Recall that the

[1]This is probably folklore. We could not find the particular approach with AboveThreshold presented in the literature — the original PTR work by Dwork and Lei (2009) uses composition, thus depends on poly(M), while using AboveThreshold (or our approach with general DP selection) incurs only log(M).

test $\mathcal{T}$ satisfies $(\hat{\epsilon}, \hat{\delta})$-DP.

When $\perp \in S$,

$$\begin{aligned}\Pr\left[\mathcal{M}(X) \in S \cap E^C\right] &= \Pr\left[\mathcal{T}(X) = 0\right] \\ &\leq e^{\hat{\epsilon}} \Pr\left[\mathcal{T}(X') = 0\right] + \hat{\delta} \\ &= e^{\hat{\epsilon}} \Pr\left[\mathcal{M}(X') \in S \cap E^C\right] + \hat{\delta}.\end{aligned}$$

This inequality also holds true when $\perp \notin S$, in which event $\Pr\left[\mathcal{M}(X) \in S \cap E^C\right] = \Pr\left[\mathcal{M}(X') \in S \cap E^C\right] = 0$.

From the assumption of Theorem 3.5.3 on the test $\mathcal{T}$, $\Pr[E] = \Pr\left[\mathcal{T}(X) = 1\right] \leq \delta'$. So

$$\Pr\left[\mathcal{M}(X) \in S \cap E\right] \leq \Pr[E] \leq \delta'.$$

Putting these together, we have

$$\begin{aligned}\Pr\left[\mathcal{M}(X) \in S\right] &= \Pr\left[\mathcal{M}(X) \in S \cap E^C\right] + \Pr\left[\mathcal{M}(X) \in S \cap E\right] \\ &\leq e^{\hat{\epsilon}} \Pr\left[\mathcal{M}(X') \in S \cap E^C\right] + \hat{\delta} + \delta' \\ &\leq e^{\hat{\epsilon}} \Pr\left[\mathcal{M}(X') \in S\right] + \hat{\delta} + \delta'.\end{aligned}$$

Case II: $\epsilon_\phi(X) \leq \epsilon$

Consider $\mathcal{M} : \tilde{\tilde{\mathcal{X}}} \to \mathcal{R} \cup \{\perp\}$, with the input space restricted to $\tilde{\tilde{\mathcal{X}}} = \{X \in \mathcal{X} \,|\, \epsilon_\phi(X) \leq \epsilon\}$. Since $\mathcal{M}_\phi$ satisfies $(\epsilon_\phi(X), \delta)$ data-dependent DP for dataset X, for any neighboring dataset X' and output set $\Theta \subseteq \mathcal{R}$ we have

$$\begin{aligned}\Pr\left[\mathcal{M}_\phi(X) \in \Theta\right] &\leq e^{\epsilon_\phi(X)} \Pr\left[\mathcal{M}_\phi(X') \in \Theta\right] + \delta, \\ \Pr\left[\mathcal{M}_\phi(X') \in \Theta\right] &\leq e^{\epsilon_\phi(X)} \Pr\left[\mathcal{M}_\phi(X) \in \Theta\right] + \delta.\end{aligned}$$

By the assumption $\epsilon_\phi(X) \leq \epsilon$,

$$\Pr\left[\mathcal{M}_\phi(X) \in \Theta\right] \leq e^\epsilon \Pr\left[\mathcal{M}_\phi(X') \in \Theta\right] + \delta,$$
$$\Pr\left[\mathcal{M}_\phi(X') \in \Theta\right] \leq e^\epsilon \Pr\left[\mathcal{M}_\phi(X) \in \Theta\right] + \delta.$$

Therefore the mechanism $\mathcal{M}_\phi : \tilde{\tilde{X}} \to \mathcal{R}$ satisfies (ϵ, δ)-DP.

Now consider an "expanded" mechanism $\mathcal{M}^*(X) = (\mathcal{T}(X), \mathcal{M}_\phi(X))$ that differs from $\mathcal{M}$ by releasing both the test output and the parameterized mechanism output. Instead of post-processing the test output to determine whether to run $\mathcal{M}_\phi$, the mechanism $\mathcal{M}^*$ runs $\mathcal{M}_\phi(X)$ regardless of the outcome of $\mathcal{T}(X)$. Define a post-processing function $\mathcal{P} : \{0, 1\} \times \mathcal{R} \to \mathcal{R} \cup \{\perp\}$ as follows:

$$\mathcal{P}(T, \theta) = \begin{cases} \perp & \text{if } T = 0, \\ \theta & \text{if } T = 1. \end{cases}$$

By composition (Theorem B.2.1), the expanded mechanism $\mathcal{M}^*$ satisfies $(\hat{\epsilon} + \epsilon, \hat{\delta} + \delta)$-DP. Writing $\mathcal{M} = \mathcal{P} \circ \mathcal{M}^*$, by closure to post-processing (Theorem B.2.2) we see that $\mathcal{M}$ also satisfies $(\hat{\epsilon} + \epsilon, \hat{\delta} + \delta)$-DP. △

To complete the proof, we recall that $\mathcal{X} = \tilde{X} \cup \tilde{\tilde{X}}$. So combining the two cases (and restoring the input space), the mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{R} \cup \{\perp\}$ satisfies (ϵ^*, δ^*)-DP for $\epsilon^* = \max(\hat{\epsilon}, \epsilon + \hat{\epsilon}) = \epsilon + \hat{\epsilon}$ and $\delta^* = \max(\hat{\delta} + \delta', \delta + \hat{\delta}) = \hat{\delta} + \max(\delta', \delta) \leq \delta + \hat{\delta} + \delta'$. □

B.2.2 Utility guarantee of Algorithm 5

The utility of Algorithm 5 depends on how many rounds that Algorithm 4 is invoked. We next provide the utility guarantee of Algorithm 5, which follows a simplification of the result in the Section A.2 of Papernot and Steinke (2021).

Theorem B.2.3. *Suppose applying Algorithm 4 with each ϕ_i has an equal probability to achieve the highest validation score. Let $\hat{T}$ denotes the number of invocation of Algorithm 4, where $\hat{T}$ follows a truncated geometric distribution. Then the expected quantile of the highest score candidate is given by $\mathbb{E}_{\hat{T}}\left[1 - \frac{1}{\hat{T}+1}\right]$.*

In practice, we can roughly set $\tau = \frac{1}{10k}$ so that the algorithm is likely to test all k parameters.

Proof. Suppose each oracle access to $Q(X)$ has a probability $1/k$ of achiving the best validation accuracy. Let β denote the probability that $\mathcal{A}$ (shorthand for Algorithm 5) outputs the best choice of ϕ_i.

$$\begin{aligned}\beta &= 1 - [\mathcal{A}(X)\text{is not best}] \\ &= 1 - \mathbb{E}_{\hat{T}}\left[[Q(X)\text{is not best}]^{\hat{T}}\right] \\ &= 1 - \mathbb{E}_{\hat{T}}\left[(1 - \frac{1}{k})^{\hat{T}}\right].\end{aligned}$$

Let $f(x) = \mathbb{E}[x^{\hat{T}}]$. Applying a first-order approximation on $f(1 - \frac{1}{k})$, we have $f(1 - \frac{1}{k}) \approx f(1) - f'(1) \cdot \frac{1}{k} = 1 - \mathbb{E}[\hat{T}]/k$. Then, if k is large and we choose $\tau = 0.1/k$, $\mathcal{A}$ can roughly return the best ϕ_i. □

B.2.3 Avoid hyperparameter selection with a uniform bound

The sufficient statistics of $\epsilon_\phi(X)$ are sometimes independent to ϕ. In this case, the resulting $\epsilon_\phi^P(X)$ from Approach 1 above is a valid upper bound of $\epsilon_\phi(X)$ for all ϕ *simultaneously* with high probability. In this case, we can directly choose a valid ϕ using the uniform upper bound, rather than proposing one as Algorithm 4, while avoiding hyperparameter selection all together as in Algorithm 5. The procedure is summarized

in Algorithm 12. This subsumes the classical procedure for privately releasing an upper bound of the local sensitivity (Vadhan, 2017, Section 3.4).

There are also cases where the bound is only partially uniform over some coordinates of ϕ. In these cases, Algorithm 12 can be used to reduce the dimension of the search space in Algorithm 5 (e.g., Example 3.6.2).

Algorithm 12 Generalized PTR with Uniform bound

1: **Input**: Dataset X; mechanism $\mathcal{M}_\phi : \mathcal{X} \to \mathcal{R}$ and its privacy budget ϵ, δ; $(\hat{\epsilon}, \hat{\delta})$-DP algorithm $\mathcal{A}$ that outputs $\bar{\epsilon}(\cdot)$ such that $\bar{\epsilon}(\phi) \geq \epsilon_\phi(X) \forall \phi$ with probability $1 - \delta'$, where $\epsilon_\phi(X)$ is the data-dependent DP w.r.t. δ.
2: Release $\bar{\epsilon}(\cdot) = \mathcal{A}(X)$.
3: Choose ϕ such that $\bar{\epsilon}(\phi) \leq \epsilon$
4: Release $\theta = \mathcal{M}_\phi(X)$.

Theorem B.2.4. *Algorithm 12 satisfies* $(\epsilon + \hat{\epsilon}, \delta + \hat{\delta} + \delta')$*-DP.*

Proof. Let $\mathcal{M} : \mathcal{X} \to \mathcal{R}$ denote the mechanism described in Algorithm 12.

We view $\mathcal{M}$ as a composition of two parts: an $(\hat{\epsilon}, \hat{\delta})$-DP algorithm $\mathcal{A}$ and the release of $\theta = \mathcal{M}_\phi(X)$. First of all, note that $\mathcal{A}$ outputs $\bar{\epsilon}(\cdot)$ such that for all ϕ, $\bar{\epsilon} \geq \epsilon_\phi(X)$ with probability at least $1 - \delta'$. Let E denote the event that $\epsilon_\phi(X) \leq \bar{\epsilon}$, and observe that $[E^C] \leq \delta'$.

Then we have

$$[\mathcal{M}_\phi(X) \in S] = [\mathcal{M}_\phi(X) \in S \mid E] + [\mathcal{M}_\phi(X) \in S \mid E^c] \tag{B.2.1}$$

$$\leq [\mathcal{M}_\phi(X) \in S \mid E] + \delta' \tag{B.2.2}$$

$$\leq e^{\bar{\epsilon}(\phi)}[\mathcal{M}_\phi(X') \in S \mid E] + \delta' + \delta \tag{B.2.3}$$

$$\leq e^{\bar{\epsilon}(\phi)}[\mathcal{M}_\phi(X') \in S] + \delta' + \delta \tag{B.2.4}$$

$$\leq e^{\epsilon}[\mathcal{M}_\phi(X') \in S] + \delta' + \delta. \tag{B.2.5}$$

The inequality holds for both directions (i.e., we can swap X and X').

The second inequality comes from the definition of $\bar{\epsilon}(\phi)$ and the last inequality is because we have conditioned on event E which ensures that $\bar{\epsilon}(\phi) \leq \bar{\epsilon}$.

Finally, by the composition theorem over algorithm $\mathcal{A}$ and $\mathcal{M}_\phi$, we have that $\mathcal{M}$ satisfies $(\epsilon + \hat{\epsilon}, \hat{\delta} + \delta + \delta')$-DP. □

B.2.4 RDP analysis of generalized PTR

Algorithm 13 Generalized Propose-Test-Release with RDP

1: **Input**: Dataset X; the RDP parameter α, mechanism $\mathcal{M}_\phi : \mathcal{X} \to \mathcal{R}$ that satisfies $(\alpha, \tilde{\epsilon}(\alpha))$-RDP and its RDP budget $\epsilon(\alpha)$; An $(\alpha, \hat{\epsilon}(\alpha))$-RDP test $\mathcal{T}$; false positive rate $\leq \delta'$; data-dependent RDP function $\epsilon_\phi(\alpha, X)$ w.r.t. α.

2: **if not** $\mathcal{T}(X)$ **then** output $\perp$,

3: **else** release $\theta = \mathcal{M}_\phi(X)$.

Theorem B.2.5 (Privacy guarantee of generalized PTR with RDP). *Consider a proposal ϕ and a data-dependent RDP function $\epsilon_\phi(\alpha, X)$ w.r.t. α. Suppose that $\mathcal{M}_\phi$ satisfies*

$(\alpha, \tilde{\epsilon}(\alpha))$-RDP for every dataset and we have an $(\alpha, \hat{\epsilon}(\alpha))$-RDP test $\mathcal{T} : \mathcal{X} \to \{0, 1\}$ such that when $\epsilon_\phi(\alpha, X) > \epsilon(\alpha)$,

$$\mathcal{T}(X) = \begin{cases} 0 & \textit{with probability } 1 - \delta', \\ 1 & \textit{with probability } \delta'. \end{cases}$$

Then Algorithm 4 satisfies $(\alpha, \hat{\epsilon}(\alpha) + \frac{1}{\alpha-1} \log \left(\delta' e^{(\alpha-1)\tilde{\epsilon}(\alpha)} + (1 - \delta') e^{(\alpha-1)\epsilon(\alpha)} \right)$-RDP.

Proof. We can view Algorithm 4 as a composition of two part: an $(\alpha, \hat{\epsilon})$-RDP test and a decision of whether or not running $\theta = \mathcal{M}_\phi(X)$ based on the output of the test. Let $\mathcal{M} : \mathcal{X} \to \{\mathcal{R}, \perp\}$ denote the randomized algorithm of the second part, where we use P, Q to denote the distribution of $\mathcal{M}(X)$ and $\mathcal{M}(X')$ respectively. Let E denote the false positive event of the test $\mathcal{T}$: the test passes but $\epsilon(\alpha, X) > \epsilon(\alpha)$.

We have

$$\begin{aligned} \mathbb{E}_Q[(dP/dQ)^\alpha] &= E_Q[(dP/dQ)^\alpha | E] \mathbb{P}_Q[E] + E_Q[(dP/dQ)^\alpha | E^c] \mathbb{P}_Q(E^c) \\ &\leq \delta' e^{(\alpha-1)\tilde{\epsilon}(\alpha)} + (1 - \delta') e^{(\alpha-1)\epsilon(\alpha)} \end{aligned}$$

The inequality uses the fact that $\mathcal{M}_\phi(\cdot)$ satisfies $(\alpha, \tilde{\epsilon}(\alpha))$-RDP for all datasets and includes the event of E. Therefore, $\mathcal{M}$ satisfies $(\alpha, \frac{1}{\alpha-1} \log \left(\delta' e^{(\alpha-1)\tilde{\epsilon}(\alpha)} + (1 - \delta') e^{(\alpha-1)\epsilon(\alpha)} \right)$-RDP. Finally, we conclude the proof using the composition rule of RDP over two parts. □

B.3 Omitted examples in the main body

In this section, we provide more examples to demonstrate the merits of generalized PTR. We focus on a simple example of post-processed Laplace mechanism in Section B.3.1 and then an example on differentially private learning of generalized linear models in

Section 3.5. In both cases, we observe that generalized PTR provides data-adaptive algorithms with formal DP guarantees that are simple, effective and not previously proposed in the literature (to the best of our knowledge).

B.3.1 Limits of the classic PTR in private binary voting

The following example demonstrates that classic PTR does not capture sufficient data-dependent quantities even when the local sensitivity exists and can be efficiently tested.

Example B.3.1. *Consider a binary class voting problem: n users vote for a binary class $\{0, 1\}$ and the goal is to output the class that is supported by the majority. Let n_i denote the number of people who vote for the class i. We consider the report-noisy-max mechanism:*

$$\mathcal{M}(X) : argmax_{i\in[0,1]} n_i(X) + Lap(b),$$

where $b = 1/\epsilon$ denotes the scale of Laplace noise.

In the example, we will (1) demonstrate the merit of data-dependent DP; and (2) empirically compare classic PTR with generalized PTR.

We first explicitly state the data-dependent DP.

Theorem B.3.2. *The data-dependent DP of the above example is*

$$\epsilon(X) := \max_{X'}\{|\log\frac{p}{p'}|, |\log\frac{1-p}{1-p'}|\},$$

where $p := \Pr[n_0(X) + Lap(1/\epsilon) > n_1(X) + Lap(1/\epsilon)]$ and $p' := \Pr[n_0(X') + Lap(1/\epsilon) > n_1(X') + Lap(1/\epsilon)]$. There are four possible neighboring datasets $X' : n_0(X') = \max(n_0(X) \pm 1, 0), n_1(X') = n_1(X)$ or $n_0(X') = n_0(X), n_1(X') = \max(n_1(X) \pm 1, 0)$.

In Figure B.1 (a), we empirically compare the above data-dependent DP with the Laplace mechanism by varying the gap between the two vote counts $|n_0(X) - n_1(X)|$. The noise scale is fixed to $\epsilon = 10$. The data-dependent DP substantially improves over the standard DP if the gap is large. However, the data-dependent DP is a function of the dataset. We next demonstrate how to apply generalized PTR to exploit the data-dependent DP.

Notice that the probability $n_0(X) + Lap(1/\epsilon) > n_1(X) + Lap(1/\epsilon)$ is equal to the probability that a random variable $Z := X - Y$ exceeds $\epsilon(n_1(X) - n_0(X))$, where X, Y are two independent Lap(1) distributions. We can compute the pdf of Z through the convolution of two Laplace distributions, which implies $f_{X-Y}(z) = \frac{1+|z|}{4e^{|z|}}$. Let t denote the difference between $n_1(X)$ and $n_0(X)$, i.e., $t = n_1(X) - n_0(X)$. Then we have

$$p = [Z > \epsilon \cdot t] = \frac{2 + \epsilon \cdot t}{4 \exp(\epsilon \cdot t)}$$

Similarly, $p' = \frac{2 + \epsilon \cdot (t+\ell)}{4 \exp(\epsilon \cdot (t+\ell))}$, where $\ell \in [-1, 1]$ denotes adding or removing one data point to construct the neighboring dataset X'. Therefore, we can upper bound $\log(p/p')$ by

$$\begin{aligned}
\log \frac{p}{p'} &= \frac{2 + \epsilon \cdot t}{4 \exp(\epsilon \cdot t)} \cdot \frac{4 \exp(\epsilon(t+\ell))}{2 + \epsilon \cdot (t+\ell)} \\
&\leq \epsilon \cdot \log \left(\frac{2 + \epsilon t}{2 + \epsilon(t+1)} \right) \\
&= \epsilon \log \left(1 - \frac{\epsilon}{2 + \epsilon(t+1)} \right)
\end{aligned}$$

Then we can apply generalized PTR by privately lower-bounding t.

On the other hand, the local sensitivity $\Delta_{LS}(X)$ of this noise-adding mechanism is 0 if $t > 1$. Specifically, if the gap is larger than one, adding or removing one user will

not change the result. To apply classic PTR, we let $\gamma(X)$ denote the distance to the nearest dataset X'' such that $\Delta_{LS} > 0$ and test if $\gamma(X) + \text{Lap}(1/\epsilon) > \frac{\log(1/\delta)}{\epsilon}$. Notice in this example that $\gamma(X) = \max(t-1, 0)$ can be computed efficiently. We provide the detailed implementation of these approaches.

1. Gen PTR: lower bound t with $t^p = t - \frac{log(1/\delta)}{\tilde{\epsilon}} + \text{Lap}(1/\tilde{\epsilon})$. Calculate an upper bound of data-dependent DP ϵ^p using Theorem B.3.2 with t^p. The algorithm then tests if ϵ^p is within an predefined privacy budget ϵ'. If the test passes, the algorithm returns $\text{argmax}_{i\in[0,1]} n_i(X) + Lap(1/\epsilon)$ satisfies $(\tilde{\epsilon} + \epsilon', \delta)$-DP.

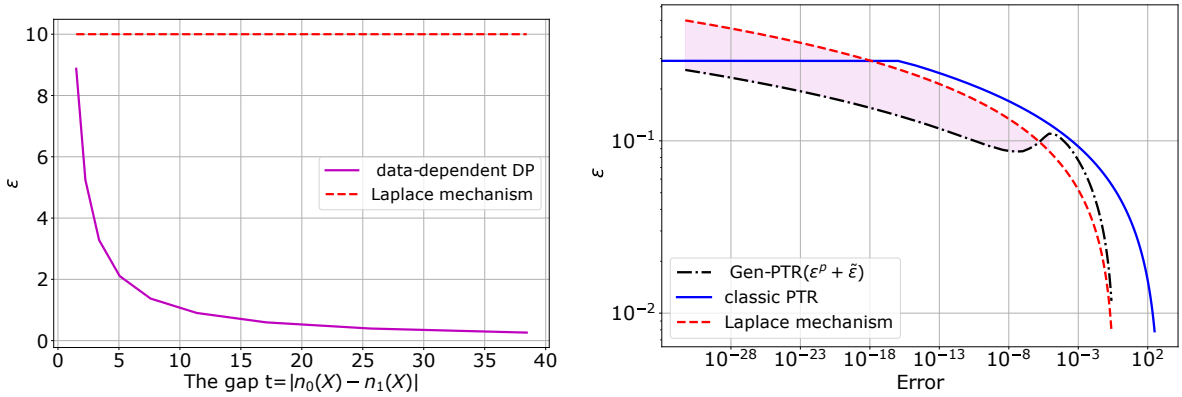
2. classic PTR: lower bound t with $t^p = t - \frac{log(1/\delta)}{\tilde{\epsilon}} + \text{Lap}(1/\tilde{\epsilon})$. If $t^p > 1$, classic PTR outputs the ground-truth result else returns a random class. This algorithm satisfies $(\tilde{\epsilon}, \delta)$-DP.

3. Laplace mechanism. $\mathcal{M}(X) : \text{argmax}_{i\in[0,1]} n_i(X) + Lap(1/\epsilon)$. $\mathcal{M}$ is (ϵ, δ)-DP.

We argue that though the Gen-PTR and the classic PTR are similar in privately lower-bounding the data-dependent quantity t, the latter does not capture sufficient information for data-adaptive analysis. That is to say, only testing the local sensitivity restricts us from learning helpful information to amplify the privacy guarantee if the test fails. In contrast, our generalized PTR, where privacy parameters and the local sensitivity parameterize the data-dependent DP, can handle those failure cases nicely.

To confirm this conjecture, Figure B.1 (b) plots a privacy-utility trade-off curve between these three approaches. We consider a voting example with $n_0(X) = n_1(X) + 100$ and $t = 100$, chosen such that the data-adaptive analysis is favorable.

In Figure B.1 (b), we vary the noise scale $b = 1/\epsilon$ between $[0, 0.5]$. For each choice of b, we plot the privacy guarantee of three algorithms when the error rate is aligned. For Gen-PTR, we set $\tilde{\epsilon} = \frac{1}{2b}$ and empirically calculate ϵ^p over 100000 trials.



(a) data-dependent DP vs Laplace mechanism (b) Privacy-utility tradeoff between three approaches.

Figure B.1: Left: We compare the privacy guarantee by varying the gap. Right: We fix $t = n_0(X) - n_1(X) = 100$ and compare the privacy cost when the accuracy is aligned. Gen-PTR with any choice of privacy budget $(\tilde{\epsilon} + \epsilon')$ chosen from the purple region would achieve the same utility as Laplace mechanism but with a smaller privacy cost. The curve of Gen-PTR is always below than that of the classic PTR, which implies that Gen-PTR can result a tighter privacy analysis when the utility is aligned.

In the plot, when $\epsilon \ll \frac{\log(1/\delta)}{t}$, the classic PTR is even worse than the Laplace mechanism. This is because the classic PTR is likely to return $\perp$ while the Laplace mechanism returns $\text{argmax}_{i\in[0,1]} n_i(X) + \text{Lap}(1/\epsilon)$, which contains more useful information. Compared to the Laplace mechanism, Gen-PTR requires an extra privacy allocation $\tilde{\epsilon}$ to release the gap t. However, it still achieves an overall smaller privacy cost when the error rate $\leq 10^{-5}$ (the purple region). Meanwhile, Gen-PTR dominates the classic PTR (i.e., the dashed black curve is always below the blue curve). Note that the classic PTR and the Gen-PTR utilize the gap information differently: the classic PTR outputs $\perp$ if the gap is not sufficiently large, while the Gen-PTR encodes the gap into the data-dependent DP function and tests the data-dependent DP in the end. This empirical result suggests that testing the local sensitivity can be loosely compared to testing the data-dependent DP. Thus, Gen-PTR could provide a better privacy-utility trade-off.

B.3.2 Self-concordant generalized linear model (GLM)

In this section, we demonstrate the effectiveness and flexibility of generalized PTR in handling a family of GLMs where the link function satisfies a self-concordance assumption. This section is organized as follows:

- Introduce a family of GLMs with the self-concordance property.
- Introduce a general output perturbation algorithm for private GLMs.
- Analyze the data-dependent DP of GLMs with the self-concordance property.
- Provide an example of applying our generalized PTR framework to logistic regression.

Consider the empirical risk minimization problem of the generalized linear model

$$\theta^* = \arg\min_{\theta} \sum_{i=1^n} l_i(\theta) + r(\theta),$$

where $l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ belongs to a family of convex GLMs: $l_i(\theta) = l(y, x_i^T\theta)$. Let $r : \mathbb{R}^d \to \mathbb{R}$ be a regularization function.

We now define the self-concordance property.

Definition B.3.3 (Generalized self-concordance (Bach, 2010)). A convex and three-times differentiable function $f : \Theta \to \mathbb{R}$ is R-generalized-self-concordant on an open nonempty convex set $\Theta^* \subset \Theta$ with respect to norm $\|\cdot\|$ if for all $u \in \Theta^*$ and all $v \in \mathbb{R}^d$,

$$\nabla^3 f(u)[v, v, v] \leq 2R\|v\|(\nabla^2 f(u)[v, v]).$$

The closer R is to 0, the "nicer" — more self-concordant — the function is. A consequence of (generalized) self-concordance is the spectral (multiplicative) stability of Hessian to small perturbations of parameters.

Lemma B.3.4 (Stability of Hessian(Nesterov and Nemirovskii, 1994, Theorem 2.1.1), (Bach, 2010, Proposition 1)). *Let* $H_\theta := \nabla^2 F_s(\theta)$. *If* F_s *is* R*-self-concordant at* θ*, then for any* v *such that* $R\|v\|_{H_\theta} < 1$*, we have that*

$$\begin{aligned}(1 - R\|v\|_{H_\theta})^2 \nabla^2 F_s(\theta) &\prec \nabla^2 F_s(\theta + v) \\ &\prec \frac{1}{(1 - R\|v\|_{H_\theta})^2} \nabla^2 F_s(\theta).\end{aligned}$$

If instead we assume F_s *is* R*-generalized-self-concordant at* θ *with respect to norm* $\|\cdot\|$*, then*

$$e^{-R\|v\|} \nabla^2 F_s(\theta) \prec \nabla^2 F_s(\theta + v) \prec e^{R\|v\|} \nabla^2 F_s(\theta)$$

The two bounds are almost identical when $R\|v\|$ and $R\|v\|_\theta$ are close to 0. In particular, for $x \leq 1/2$, we have that $e^{-2x} \leq 1 - x \leq e^{-x}$.

In particular, the loss function of binary logistic regression is 1-generalized self-concordant.

Example B.3.5 (Binary logistic regression). *Assume* $\|x\|_2 \leq 1$ *for all* $x \in \mathcal{X}$ *and* $y \in \{-1, 1\}$. *Then binary logistic regression with datasets in* $\mathcal{X} \times \mathcal{Y}$ *has a log-likelihood of* $F(\theta) = \sum_{i=1}^n \log(1 + e^{-y_i x_i^T \theta})$. *The univariate function* $l := \log(1 + \exp(\cdot))$ *satisfies*

$$|l'''| = \left| \frac{\exp(\cdot)(1 - \exp(\cdot))}{(1 + \exp(\cdot))^3} \right| \leq \frac{\exp(\cdot)}{(1 + \exp(\cdot))^2} := l''.$$

We next apply the modified output perturbation algorithm to privately release θ^*. The algorithm is simply:

1. Solve

$$\theta^* = \arg\min_\theta \sum_{i=1}^n l_i(\theta) + r(\theta).$$

2. Release

$$\hat{\theta} = \theta^* + Z,$$

where $\gamma > 0$ is a tuning parameter and $Z \sim \mathcal{N}(0, \gamma^{-1}(\sum_{i=1}^{n} \nabla^2 l_i(\theta) + \nabla^2 r(\theta))^{-1})$.

The data-dependent DP of the above procedure is stated as follows.

Theorem B.3.6 (Data-dependent DP of GLM). *Denote the smooth part of the loss function* $F_s = \sum_{i=1}^{n} l(y_i, < x_i, \cdot >) + r_s(\cdot)$. *Assume the following:*

1. *The GLM loss function l is convex, three-times continuously differentiable and R-generalized-self-concordant w.r.t. $\|\cdot\|_2$,*
2. *F_s is locally α-strongly convex w.r.t. $\|\cdot\|_2$,*
3. *and in addition, denote $L := \sup_{\theta \in [\theta^*, \tilde{\theta}^*]} |l'(y, x^T\theta)|$, $\beta := \sup_{\theta \in [\theta^*, \tilde{\theta}^*]} |l''(y, x^T\theta)|$. That is, $\ell(\cdot)$ is L-Lipschitz and β-smooth.*

We then have the data-dependent DP

$$\epsilon(Z) \leq \frac{R(L+\beta)}{\alpha}(1 + \log(2/\delta)) + \frac{\gamma L^2}{\alpha} + \sqrt{\frac{\gamma L^2}{\alpha} \log(2/\delta)}.$$

The proof follows by taking an upper bound of the per-instance DP loss (Theorem B.3.6) $\epsilon(Z, z)$ over $z = (x, y) \in (\mathcal{X}, \mathcal{Y})$.

Notice that the Hessians can be arbitrarily singular and α could be 0, which leads to an infinite privacy loss without additional assumptions. Thus, we will impose an additional regularization of form $\frac{\lambda}{2}||\theta||^2$, which ensures that for any dataset F_S is λ-strongly convex.

This is not yet DP because it is still about a fixed dataset. We also need a pre-specified privacy budget (ϵ, δ). We next demonstrate how to apply the generalized PTR to provide a general solution to the above GLM, using logistic regression as an example.

Remark B.3.7 (Logistic regression). For logistic regression, we know $L \leq 1$, $\beta \leq 1/4$ and if $\|x\|_2 \leq 1$, it is 1-generalized self-concordant. For any dataset $Z = (X, y)$, the data-dependent DP $\epsilon(X)$ w.r.t. δ can be simplified to:

$$\frac{1.25}{\alpha}(1 + \log(2/\delta)) + \frac{\gamma}{\alpha} + \sqrt{\frac{\gamma}{\alpha}\log(2/\delta)}$$

Now, the data-dependent DP is a function of α and γ, where α denotes the local strong convexity at θ^*_λ and γ controls the noise scale. We next show how to select these two parameters adapted to the dataset.

Example B.3.8. *We demonstrate here how we apply generalized PTR to output perturbation of the logistic regression problem.*

1. *Take an exponential grid of parameters* $\{\lambda\}$ *and propose each* λ.
2. *Solve for* $\theta^*_\lambda = \arg\min_\theta F(\theta) + \lambda\|\theta\|^2/2$
3. *Calculate the smallest eigenvalue* $\lambda_{\min}(\nabla^2 F(\theta^*_\lambda))$ *(e.g., using power method).*
4. *Differentially privately release* $\lambda_{\min}$ *with* $\lambda^p_{\min} := \max\{\lambda_{\min} + \frac{\sqrt{\log(4/\delta)}}{\epsilon/2} \cdot \Delta_{GS} \cdot Z - \frac{\sqrt{2\log(4/\delta)\cdot\log(1/\delta)}\Delta_{GS}}{\epsilon/2}, 0\}$, *where* Δ_{GS} *denote the global sensitivity of* $\lambda_{\min}$ *using Theorem B.3.11.*
5. *Let* $\epsilon^p(\cdot)$ *be instantiated with* $\epsilon(X)$ *w.r.t.* δ *from Remark B.3.7, where* $\alpha = \lambda^p_{\min} + \lambda$. *Then, conditioned on a high probability event,* $\epsilon^p(\cdot)$ *(a function of* γ*) is a valid DP bound that holds for all datasets and all parameters* γ.
6. *Calculate the maximum* γ *such that* $\epsilon^p_{\delta/2}(\gamma) \leq \epsilon/2$.
7. *Release* $\hat{\theta} \sim \mathcal{N}(\theta^*_\lambda, \gamma^{-1}\nabla^2 F_s(\theta^*_\lambda)^{-1})$.

8. *Evaluate the utility on the validation set and return the (λ, γ) pair that leads to the highest utility.*

Theorem B.3.9. *For each proposed λ, the algorithm that releases $\hat{\theta} \sim \mathcal{N}(\theta^*_\lambda, \gamma^{-1}\nabla^2 F_s(\theta^*_\lambda)^{-1})$ is $(\epsilon, 2\delta)$-DP.*

Proof. The proof follows the recipe of generalized PTR with private upper bound (Example 3.5.7). First, the release of $\lambda_{\min}(\nabla^2 F(\theta^*_\lambda))$ is $(\epsilon/2, \delta/2)$-DP. Then, with probability at least $1-\delta$, $\epsilon^p_\delta(\cdot) > \epsilon_\delta(X)$ holds for all X and γ. Finally, γ is chosen such that the valid upper bound is $(\epsilon/2, \delta/2)$-DP. □

For the hyperparameter tuning on λ (Steps 1 and 8), we can use Algorithm 5 to evaluate each λ.

Unlike Example 3.6.2, the $\lambda_{min}(\nabla^2 F(\theta^*_\lambda))$ is a complicated data-dependent function of λ. Thus, we cannot privately release the data-dependent quantity $\lambda_{min}(\nabla^2 F(\theta^*_\lambda))$ without an input λ. The PTR approach allows us to test a number of different λ and hence get a more favorable privacy-utility trade-off.

An interesting perspective of this algorithm for logistic regression is that increasing the regularization α is effectively increasing the number of data points within the soft "margin"[2] of separation, hence a larger contribution to the Hessian from the loss function.

Remark B.3.10. The PTR solution for GLMs follows a similar recipe: propose a regularization strength λ; construct a lower bound of the strong convexity α at the optimal solution θ^*_λ; and test the validity of data-dependent DP using Theorem B.3.6.

Before moving on to other applications of generalized PTR, we will show how to differentially privately release λ_{min} according to the requirements of the logistic regression example.

[2]If we think of logistic regression as a smoothed version of SVM, then increasing α leads to more support vectors. The "margin" is "softer" in logistic regression, but qualitatively the same.

B.3.3 Differentially privately release $\lambda_{min}\left(\nabla^2 F(\theta)\right)$

To privately release $\lambda_{min}\nabla^2 F(\theta)$, we first need to compute its global sensitivity. Once we have that then we can release it differentially privately using either the Laplace mechanism or the Gaussian mechanism.

Theorem B.3.11 (Global sensitivity of the minimum eigenvalue at the optimal solution). *Let* $F(\theta) = \sum_{i=1}^{n} f_i(\theta) + r(\theta)$ *and* $\tilde{F}(\theta) = F(\theta) + f(\theta)$ *where* $f_1, ..., f_n$ *are loss functions corresponding to a particular datapoint* x. *Let* $\theta^* = \arg\min_\theta F(\theta)$ *and* $\tilde{\theta}^* = \arg\min_\theta \tilde{F}(\theta)$. *Assume* f *is L-Lipschitz and* β*-smooth,* $r(\theta)$ *is* λ*-strongly convex, and* F *and* $\tilde{F}$ *are R-self-concordant. If in addition,* $\lambda \geq RL$, *then we have*

$$\sup_{X,x}(\lambda_{min}(\nabla^2 F(\theta_\lambda^*)) - \lambda_{min}(\nabla^2 \tilde{F}(\tilde{\theta_\lambda^*}))) \leq 2RL + \beta.$$

Proof.

$$\begin{aligned} &\lambda_{min}(\nabla^2 F(\theta_\lambda^*)) - \lambda_{min}(\nabla^2 \tilde{F}(\tilde{\theta_\lambda^*})) \\ &\quad = (\lambda_{min}(\nabla^2 F(\theta_\lambda^*)) - \lambda_{min}(\nabla^2 \tilde{F}(\theta_\lambda^*))) \\ &\quad + (\lambda_{min}(\nabla^2 \tilde{F}(\theta_\lambda^*)) - \lambda_{min}(\nabla^2 \tilde{F}(\tilde{\theta_\lambda^*}))). \end{aligned} \tag{B.3.1}$$

We first bound the part on the left. By applying Weyl's lemma $\lambda(X + E) - \lambda(X) \leq ||E||_2$, we have

$$\sup_x ||\nabla^2 F(\theta_\lambda^*) - \nabla^2 F\tilde{(}\theta_\lambda^*)||_2 = ||\nabla^2 f(\theta_\lambda^*)||_2 \leq \beta \tag{B.3.2}$$

In order to bound the part on the right, we apply the semidefinite ordering using self-concordance, which gives

$$e^{-R\|\tilde{\theta_\lambda^*} - \theta_\lambda^*\|}\nabla^2 \tilde{F}(\tilde{\theta_\lambda^*}) \prec \nabla^2 \tilde{F}(\theta_\lambda^*) \prec e^{R\|\tilde{\theta_\lambda^*} - \theta_\lambda^*\|}\nabla^2 \tilde{F}(\tilde{\theta_\lambda^*}).$$

By the Courant-Fischer Theorem and the monotonicity theorem, we also have that for the smallest eigenvalue

$$\begin{aligned} e^{-R\|\tilde{\theta}^*_\lambda - \theta^*_\lambda\|}\lambda_{\min}\left(\nabla^2\tilde{F}(\tilde{\theta}^*_\lambda)\right) &\leq \lambda_{\min}\left(\nabla^2\tilde{F}(\theta^*_\lambda)\right) \\ &\leq e^{R\|\tilde{\theta}^*_\lambda - \theta^*_\lambda\|}\lambda_{\min}\left(\nabla^2\tilde{F}(\tilde{\theta}^*_\lambda)\right). \end{aligned} \tag{B.3.3}$$

Moreover by Proposition B.3.12, we have that

$$\|\tilde{\theta}^*_\lambda - \theta^*_\lambda\|_2 \leq \frac{\|\nabla f(\tilde{\theta^*}_\lambda)\|}{\lambda_{\min}\left(\nabla^2\tilde{F}(\tilde{\theta}^*_\lambda)\right)} \leq \frac{L}{\lambda_{\min}\left(\nabla^2\tilde{F}(\tilde{\theta}^*_\lambda)\right)}.$$

If $\lambda_{\min}\left(\nabla^2\tilde{F}(\tilde{\theta}^*_\lambda)\right) \geq RL$, then use that $e^x - 1 \leq 2x$ for $x \leq 1$. Substituting the above bound to (B.3.3) then to (B.3.1) together with (B.3.2), we get a data-independent global sensitivity bound of

$$\lambda_{min}(\nabla^2 F(\theta^*_\lambda)) - \lambda_{min}(\nabla^2\tilde{F}(\tilde{\theta}^*_\lambda)) \leq 2RL + \beta$$

as stated. □

Proposition B.3.12. *Let $\|\cdot\|$ be a norm and $\|\cdot\|_*$ be its dual norm. Let $F(\theta)$, $f(\theta)$ and $\tilde{F}(\theta) = F(\theta) + f(\theta)$ be proper convex functions and θ^* and $\tilde{theta}^*$ be their minimizers, i.e., $0 \in \partial F(\theta^*)$ and $0 \in \partial\tilde{F}(\tilde{theta}^*)$. If in addition, $F, \tilde{F}$ is $\alpha, \tilde{\alpha}$-strongly convex with respect to $\|\cdot\|$ within the restricted domain $\theta \in \{t\theta^* + (1-t)\tilde{\theta}^* \mid t \in [0,1]\}$. Then there exists $g \in \partial f(\theta^*)$ and $\tilde{g} \in \partial f(\tilde{\theta}^*)$ such that*

$$\|\theta^* - \tilde{\theta}^*\| \leq \min\left\{\frac{1}{\alpha}\|\tilde{g}\|_*, \frac{1}{\tilde{\alpha}}\|g\|_*\right\}.$$

Proof. Apply the first order condition to F restricted to the line segment between $\tilde{\theta}^*$ and

θ^*, we get

$$F(\tilde{\theta}^*) \geq F(\theta^*) + \langle \partial F(\theta^*), \tilde{\theta}^* - \theta^* \rangle + \frac{\alpha}{2}\|\tilde{\theta}^* - \theta^*\|^2 \tag{B.3.4}$$

$$F(\theta^*) \geq F(\tilde{\theta}^*) + \langle \partial F(\tilde{\theta}^*), \theta^* - \tilde{\theta}^* \rangle + \frac{\alpha}{2}\|\tilde{\theta}^* - \theta^*\|^2 \tag{B.3.5}$$

Note by the convexity of F and f, $\partial \tilde{F} = \partial F + \partial f$, where $+$ is the Minkowski Sum. Therefore, $0 \in \partial \tilde{F}(\tilde{\theta}^*)$ implies that there exists $\tilde{g}$ such that $\tilde{g} \in \partial f(\tilde{\theta}^*)$ and $-\tilde{g} \in \partial F(\tilde{\theta}^*)$. Take $-\tilde{g} \in \partial F(\tilde{\theta}^*)$ in Equation B.3.5 and $0 \in \partial F(\theta^*)$ in Equation B.3.4 and add the two inequalities, we obtain

$$\begin{aligned} 0 &\geq \langle -\tilde{g}, \theta^* - \tilde{\theta}^* \rangle + \alpha\|\tilde{\theta}^* - \theta^*\|^2 \\ &\geq -\|\tilde{g}\|_*\|\theta^* - \tilde{\theta}^*\| + \alpha\|\tilde{\theta}^* - \theta^*\|^2. \end{aligned}$$

For $\|\tilde{\theta}^* - \theta^*\| = 0$ the claim is trivially true; otherwise, we can divide both sides of the above inequality by $\|\tilde{\theta}^* - \theta^*\|$ and get $\|\theta^* - \tilde{\theta}^*\| \leq \frac{1}{\alpha}\|\tilde{g}\|_*$.

It remains to show that $\|\theta^* - \tilde{\theta}^*\| \leq \frac{1}{\tilde{\alpha}}\|g\|_*$. This can be obtained by exactly the same arguments above but applying strong convexity to $\tilde{F}$ instead. Note that we can actually get something slightly stronger than the statement because the inequality holds for all $g \in \partial f(\theta^*)$. □

B.3.4 Other applications of generalized PTR

Besides one-posterior sampling for GLMs, there are plenty of examples that our generalized-PTR could be applied, e.g., DP-PCA (Dwork et al., 2014b) and Sparse-DP-ERM (Kifer et al., 2012) (when the designed matrix is well-behaved).

(Dwork et al., 2014b) provides a PTR style privacy-preserving principle component analysis (PCA). The key observation of (Dwork et al., 2014b) is that the local sensitivity

is quite "small" if there is a large eigengap between the k-th and the $k+1$-th eigenvalues. Therefore, their approach (Algorithm 2) chooses to privately release a lower bound of the k-th eigengap (k is fixed as an input) and use that to construct a high-confidence upper bound of the local sensitivity.

For noise-adding mechanisms, the local sensitivity is proportional to the data-dependent loss and generalized PTR is applicable. We can formulate the data-dependent DP of DP-PCA as follows:

Theorem B.3.13. *For a given matrix $A \in \mathcal{R}^{m\times n}$, assume each row of A has a bounded ℓ_2 norm being* 1. *Let V_k denotes the top k eigenvectors of A^TA and d_k denotes the gap between the k-th and the $k+1$-th eigenvalue. Then releasing $V_kV_k^T + E$, where $E \in \mathcal{R}^{n\times n}$ is a symmetric matrix with the upper triangle is i.i.d samples from $\mathcal{N}(0,\sigma^2)$ satisfies $(\epsilon(A),\delta)$ data-dependent DP and $\epsilon(A) = \frac{2\sqrt{\log(1.25/\delta)}}{\sigma(d_k-2)}$.*

The proof is based on the local sensitivity result from (Dwork et al., 2014b) and the noise calibration of Gaussian mechanism.

We can combine Theorem B.3.13 with our Algorithm 5 to instantiate the generalized PTR framework. The improvement over Dwork et al. (2014b) will be to allow joint tuning of the parameter k and the noise variance (added to the spectral gap d_k).

B.4 Experimental details

B.4.1 Experimental details in private linear regression

Algorithm 14 OPS-PTR for linear regression (an extended version of Example 3.6.2)

Input: Data $Z = (X, Y)$; proposed regularization strength λ; failure probabilities $\delta', \delta'_1, \delta'_2$ such that $\delta'_1 + \delta'_2 = \delta'$; privacy budgets $(\hat{\epsilon}, \hat{\delta})$ and (ϵ, δ); quality score $q(\cdot)$.

Calculate the minimum eigenvalue $\lambda_{\min}(X)$ and the non-private solution $\theta^*_\lambda = (X^T X + \lambda I)^{-1} X^T Y$.

Release $\lambda_{\min}$ and $\Delta := \log(||\mathcal{Y}|| + ||\mathcal{X}||\,||\theta^*_\lambda||$ with privacy budget $(\hat{\epsilon}, \hat{\delta})$ such that

- $\lambda^P_{\min} \geq \lambda_{\min}$ with probability $1 - \delta'_1$; and
- $\Delta^P \leq \Delta$ with probability $1 - \delta'_2$.

Construct the private upper bound of the local Lipschitz constant:

$$\tilde{L} := ||\mathcal{X}||\, e^{(\Delta^P)}.$$

Construct the private upper bound of the data-dependent DP as a function of γ:

$$\overline{\epsilon}(\gamma) := \sqrt{\frac{\gamma \tilde{L}^2 \log(2/\delta)}{\lambda + \lambda^P_{\min}}} + \frac{\gamma \tilde{L}^2}{2(\lambda + \lambda^P_{\min} + ||\mathcal{X}||^2)} + \frac{1 + \log(2/\delta)||\mathcal{X}||^2}{2(\lambda + \lambda^P_{\min})}.$$

Calibrate $\gamma^* = \sup_{q(\gamma)}\{\gamma \mid \overline{\epsilon}(\gamma) \leq \epsilon\}$.

if $\gamma^* \geq 0$ **then**

  Output $\theta \sim e^{-\frac{\gamma^*}{2}\left(||Y - X\theta||_2^2 + \lambda||\theta||_2^2\right)}$,

else

  Output $\perp$.

end if

Algorithm 15 OPS-PTR: One-Posterior Sample with propose-test-release (no-"perp" version)

1: **Input**: Data $X, \mathbf{y}$. Private budget : ϵ, δ, proposed regularizer λ.

2: Calculate the minimum eigenvalue $\lambda_{\min}(X^T X)$.

3: Sample $Z \sim \mathcal{N}(0,1)$ and privately release $\tilde{\lambda}_{\min} = \max\left\{\lambda_{\min} + \frac{\sqrt{\log(6/\delta)}}{\epsilon/4} Z - \frac{\sqrt{2\log(6/\delta)\cdot\log(2/\delta)}}{\epsilon/4}, 0\right\}$

4: Calculate $\hat{\theta} = (X^T X + \lambda I)^{-1} X^T y$.

5: Sample $Z \sim \mathcal{N}(0,1)$ and privately release $\Delta = \log(||\mathcal{Y}|| + ||\mathcal{X}||||\hat{\theta}||) + \frac{\log(1+||\mathcal{X}||^2/(\lambda+\tilde{\lambda}_{\min}))}{\epsilon/(4\sqrt{6/\delta})} Z + \frac{\log(1+||\mathcal{X}||^2/(\lambda+\tilde{\lambda}_{\min}))}{\epsilon/(4\sqrt{2\log(6/\delta)\log(2/\delta)})}$.

6: Set the local Lipschitz $\tilde{L} := ||X|| e^{\Delta}$.

7: Calibrate γ with Theorem 3.6.1($\delta/3, \epsilon/2$.)

8: Output $\tilde{\theta} \sim p(\theta|X, \mathbf{y}) \propto \mathbf{e}^{-\frac{\gamma}{2}||\mathbf{y}-\mathbf{X}\theta||^2 + \lambda||\theta||^2}$

Algorithm 15 provides the detailed privacy calibration of the private linear regression problem.

Theorem B.4.1. *Algorithm 15 is* $(\epsilon + \hat{\epsilon}, \delta + \hat{\delta} + \delta')$*-DP.*

Proof. There are two data-dependent quantities in Theorem 3.6.1: $\lambda_{\min}$ and L, which is a function of $||\theta^*_\lambda||$.

First, we privately release $\lambda_{\min}$ and $\log(||\mathcal{Y}|| + ||\mathcal{X}||||\hat{\theta}||)$ using a combined privacy budget $(\hat{\epsilon}, \hat{\delta})$.

We apply Lemma B.4.2 from Wang (2018) to privately release $\log(||\mathcal{Y}|| + ||\mathcal{X}||||\hat{\theta}||)$, and then construct its private upper bound by post-processing of Δ. Specifically, the trick that we use is that $\log(||\mathcal{Y}|| + ||\mathcal{X}||||\hat{\theta}||)$ has a bounded local sensitivity for which we have an expression. Though Algorithm 15 leaves open-ended the question of *how* to release $\lambda_{\min}$ and L, the idea is that we could easily use the Gaussian mechanism to construct a high-probability upper bound of $\log(||\mathcal{Y}|| + ||\mathcal{X}||||\hat{\theta}||)$.

Notice that with probability at least $1 - \delta_1'$, $\lambda_{\min}$ is a lower bound of $\lambda_{\min}^P$. And with probability at least $1 - \delta_2'$, Δ is an upper bound of Δ^P. A union bound over these events then ensures that with probability $1 - \delta'$, $\bar{\epsilon}(\gamma) \leq \epsilon_\phi(X)$. That is, the expression given in Line 5 provides a valid upper bound of the data-dependent DP.

We then tune the parameter γ to satisfy the remaining privacy budget (ϵ, δ). □

Lemma B.4.2 (Lemma 12 (Wang, 2018)). *Let θ_λ^* be the ridge regression estimate with parameter λ and the smallest eigenvalue of $X^T X$ be λ_{min}, then the function* $\log(||\mathcal{Y} + ||\mathcal{X}||||\theta_\lambda^*||)$ *has a local sensitivity of* $\log(1 + \frac{||\mathcal{X}||^2}{\lambda_{min+\lambda}})$.

An idea on releasing $\lambda_{\min}$

We will state Weyl's lemma, which we could use to calculate the global sensitivity of $\lambda_{\min}$. Notice that $\lambda_{\min}$ has a global sensitivity of $||\mathcal{X}||^2$ by Weyl's lemma. This along with an assumption of $||\mathcal{X}||^2 \leq 1$ could allow us to release $\lambda_{\min}$ via the Gaussian mechanism.

Lemma B.4.3 (Weyl's theorem; Theorem 4.11, p. 204 in Stewart (1990)). . *Let A, E be given $m \times n$ matrices with $m \geq n$, then*

$$\max_{i \in [n]} |\sigma_i(A) - \sigma_i(A + E)| \leq ||E||_2 \tag{B.4.1}$$

Appendix C

Supplementary Material for Chapter 4

C.1 Notation

Denote the following:

- $\mathcal{L}(\theta) = \sum_{i=1}^{n} \ell_i(\theta)$,
- $\mathcal{L}_\lambda(\theta) = \mathcal{L}(\theta) + \frac{\lambda}{2}||\theta||_2^2$,
- $\mathcal{L}^P(\theta) = \mathcal{L}(\theta) + \frac{\lambda}{2}||\theta||_2^2 + b^T\theta, \quad b \sim \mathcal{N}(0, \sigma^2 I_d)$,
- $\theta^* = \underset{\theta \in \mathbb{R}^d}{\arg\min}\, \mathcal{L}(\theta)$,
- $\theta^*_\lambda = \underset{\theta \in \mathbb{R}^d}{\arg\min}\, \mathcal{L}_\lambda(\theta)$,
- $\theta^P = \underset{\theta \in \mathbb{R}^d}{\arg\min}\, \mathcal{L}^P(\theta)$,
- $\tilde{\theta}$ satisfies $||\nabla \mathcal{L}^P(\tilde{\theta})||_2 \leq \tau$,
- $\tilde{\theta}^P = \tilde{\theta} + b_2$, $b_2 \sim \mathcal{N}(0, \sigma_{out}^2 I_d)$.

We take $||\mathcal{X}||$ to be the size of the data domain $\mathcal{X}$, i.e. $||\mathcal{X}|| = \sup_{x\in\mathcal{X}} ||x||$. For conciseness of presentation we sometimes drop the dataset Z from the notation, e.g. we abbreviate $\mathcal{L}(\theta; Z)$ as $\mathcal{L}(\theta)$.

Sometimes (especially in the proofs that follow) we will abuse notation by overloading a function with its output, e.g. θ^P is the output of objective perturbation and $\theta^P(Z)$ is the objective perturbation mechanism.

C.2 Warm-up: Gaussian Mechanism

We will get started by reviewing the RDP and privacy profile bounds for the Gaussian mechanism. Once warmed up, we then present the RDP and privacy profile bounds for objective perturbation in Sections C.3 and C.4.

Consider the Gaussian mechanism defined by $\mathcal{M}(Z) = f(Z) + \mathcal{N}(0, \sigma^2 I_d)$, for a function $f : \mathcal{Z} \to \mathbb{R}^d$ with global sensitivity $\Delta_f = \max_{Z\simeq Z'} ||f(Z) - f(Z')||_2$.

C.2.1 Privacy profile of the Gaussian Mechanism

Analytic Gaussian mechanism (Balle and Wang, 2018).

Let P and Q be the density functions of $\mathcal{N}(\Delta_f, \sigma^2)$ and $\mathcal{N}(0, \sigma^2)$, respectively. Then (P, Q) is a dominating pair of distributions for $\mathcal{M}$, and $\mathcal{M}$ is tightly $(\epsilon, \delta(\epsilon))$-DP for

$$\delta(\epsilon) = H_{e^\epsilon}\big(P||Q\big) = \Phi\left(-\frac{\epsilon\sigma}{\Delta_f} + \frac{\Delta_f}{2\sigma}\right) - e^\epsilon \Phi\left(-\frac{\epsilon\sigma}{\Delta_f} - \frac{\Delta_f}{2\sigma}\right), \tag{C.2.1}$$

where Φ denotes the CDF of the standard univariate Gaussian distribution.

We can analytically express a tight upper bound for the Gaussian mechanism above, but in general numerical methods are needed to evaluate the hockey-stick divergence for dominating pairs of distributions. This is discussed with more detail in Section C.4.

C.2.2 RDP Analysis of the Gaussian Mechanism

Theorem C.2.1 (RDP guarantees of the Gaussian mechanism). *The Gaussian mechanism* $\mathcal{M}$ *satisfies* (α, ϵ)*-RDP for* $\alpha > 1$ *and* $\epsilon = \frac{\Delta_f^2 \alpha}{2\sigma^2}$.

C.3 RDP analysis of objective perturbation

In this section we present the proof of Theorem 4.3.2, one of our main privacy results: an RDP bound on the objective perturbation mechanism. Along the way we will also highlight the importance of the GLM assumption to the correctness of the proof.

Proof of Theorem 4.3.2. Recall from Definition 4.2.7 that the objective perturbation mechanism $\hat{\theta}^P : \mathcal{Z}^* \to \mathbb{R}^d$ satisfies $\epsilon(\alpha)$-Rényi differential privacy if for all neighboring datasets Z and Z',

$$\mathcal{D}_\alpha \left(\hat{\theta}^P(Z) \,||\, \hat{\theta}^P(Z') \right) \leq \epsilon(\alpha).$$

Assume that $Z \in \mathcal{Z}^n$ and construct $Z' \in \mathcal{Z}^{n+}$ by adding a datapoint z to Z. Note that this convention (while convenient for writing down the PLRV of objective perturbation) comes *with* loss of generality. As a consequence of asymmetry[1], the upper bound on the RDP must satisfy

$$\max \left(D_\alpha \left(\hat{\theta}^P(Z) \,||\, \hat{\theta}^P(Z') \right),\, D_\alpha \left(\hat{\theta}^P(Z') \,||\, \hat{\theta}^P(Z) \right) \right) \leq \epsilon(\alpha).$$

We will calculate the Rényi divergence $D_\alpha\big(\hat{\theta}^P(Z) \,||\, \hat{\theta}^P(Z')\big)$ of objective perturbation

[1]This is in contrast to the symmetry of the Gaussian mechanism $\mathcal{M}(Z) = f(Z) + \mathcal{N}(0, \sigma^2 I_d)$, in which case $\epsilon(\alpha)$ can be calculated exactly as $D_\alpha\big(\mathcal{N}(0, \sigma^2 I_d) \,||\, \mathcal{N}(\Delta_f, \sigma^2 I_d)\big) = \frac{\alpha \Delta_f^2}{2\sigma^2} = D_\alpha\big(\mathcal{N}(\Delta_f, \sigma^2 I_d) \,||\, \mathcal{N}(0, \sigma^2 I_d)\big)$.

under a change of measure:

$$D_\alpha(P \,||\, Q) = \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim Q} \left[\left(\frac{P(x)}{Q(x)} \right)^\alpha \right] = \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim P} \left[\left(\frac{P(x)}{Q(x)} \right)^{\alpha - 1} \right].$$

Let

$$R(\theta^P) := \frac{\Pr\left[\hat{\theta}^P(Z) = \theta^P\right]}{\Pr\left[\hat{\theta}^P(Z') = \theta^P\right]}$$

be shorthand for the probability density ratio at output θ^P, given a fixed pair of neighboring datasets Z and Z'. Then

$$\begin{aligned} D_\alpha(\hat{\theta}^P(Z) \,||\, \hat{\theta}^P(Z')) &= \frac{1}{\alpha - 1} \log \mathbb{E}_{\theta^P \sim \hat{\theta}^P(Z)} \left[R(\theta^P)^{(\alpha-1)} \right] \\ &= \frac{1}{\alpha - 1} \log \mathbb{E}_{\theta^P \sim \hat{\theta}^P(Z)} \left[e^{\log\left[R(\theta^P)^{(\alpha-1)}\right]} \right] \\ &= \frac{1}{\alpha - 1} \log \mathbb{E}_{\theta^P \sim \hat{\theta}^P(Z)} \left[e^{(\alpha-1)\log R(\theta^P)} \right] \\ &\leq \frac{1}{\alpha - 1} \log \mathbb{E}_{\theta^P \sim \hat{\theta}^P(Z)} \left[e^{(\alpha-1)\left|\log R(\theta^P)\right|} \right]. \end{aligned}$$

Denote $J(\theta; Z) = \sum_{z \in Z} \ell(\theta; z) + \frac{\lambda}{2}||\theta||_2^2$ and $\mu(\theta, Z, z) = x^T \left(\nabla^2 J(\theta; Z)\right)^{-1} x$.

From Lemma C.10.9 and the λ-strong convexity of $J(\theta; Z)$, we can bound $\mu(\theta, Z, z)$ by $\frac{||x||_2^2}{\lambda} \leq \frac{1}{\lambda}$.

We also know from the L-Lipschitzness of $\ell(\theta; z)$ and the β-smoothness of $f(x^T\theta; y)$ that $||\nabla \ell(\theta; z)||_2 \leq L$ and $f''(x^T\theta; y) \leq \beta$ for all $\theta \in \mathbb{R}^d$ and $z = (x, y) \in \mathcal{Z}$.

Abbreviate $f''(x^T\theta^P; y)\mu(\theta^P, Z, z)$ as $f''(\cdot)\mu(\cdot)$. Then using the GLM assumption, from Redberg and Wang (2021) we can bound the absolute value of the log-probability

ratio for any θ^P as

$$\begin{aligned}\left|\log R(\theta^P)\right| &\leq \left| -\log\left(1 - f''(\cdot)\mu(\cdot)\right) - \frac{1}{2\sigma^2}||\nabla\ell(\theta^P; z)||_2^2 - \frac{1}{\sigma^2}\nabla J(\theta^P; Z)^T\nabla\ell(\theta^P; z)\right| \\ &\leq \left|-\log\left(1 - f''(\cdot)\mu(\cdot)\right)\right| + \frac{1}{2\sigma^2}||\nabla\ell(\theta^P; z)||_2^2 + \frac{1}{\sigma^2}\left|\nabla J(\theta^P; Z)^T\nabla\ell(\theta^P; z)\right| \\ &\leq -\log\left(1 - \frac{\beta}{\lambda}\right) + \frac{L^2}{2\sigma^2} + \frac{1}{\sigma^2}\left|\nabla J(\theta^P; Z)^T\nabla\ell(\theta^P; z)\right|.\end{aligned}$$

It is more challenging to find a data-independent bound for the third term due to the shared dependence on θ^P.

Recall that $b \sim \mathcal{N}(0, \sigma^2 I_d)$ is the noise vector in the perturbed objective. By first-order conditions at the minimizer θ^P,

$$b = -\nabla J(\theta^P; Z).$$

If θ^P were fixed (or if θ^P were independent to b), the quantity $\nabla J(\theta^P; Z)^T\nabla\ell(\theta^P; z) = -b^T\nabla\ell(\theta^P; z)$ would have been distributed as a univariate Gaussian $\mathcal{N}(0, \sigma^2||\nabla\ell(\theta^P; z)||_2^2)$. Unfortunately in our case θ^P is a random variable, and consequently we don't have the tools to understand the distribution of $\nabla J(\theta^P; Z)^T\nabla\ell(\theta^P; z)$ for an arbitrary loss function.

But using the GLM assumption on the loss function, we can write

$$\nabla J(\theta; Z)^T\nabla\ell(\theta; z) = -b^T x f'(x^T\theta, y).$$

Observe that x is fixed w.r.t. b so that $-b^T x \sim \mathcal{N}(0, \sigma^2\|x\|_2^2)$, and $f'(x^T\theta, y)$ is a scalar. So while this scalar is a random variable that still depends on b in a complicated way, the worst possible dependence can be more easily quantified without incurring additional dimension dependence.

By the L-Lipschitz assumption and $|ab| \leq |a||b|$, we obtain the following bound:

$$\begin{aligned}\left|\nabla J(\theta^P; Z)^T \nabla \ell(\theta^P; z)\right| &= \left|f'(x^T\theta^P; y)\nabla J(\theta^P; Z)^T x\right| \\ &\leq L\left|\nabla J(\theta^P; Z)^T x\right|.\end{aligned}$$

This is much better! By first-order conditions we can then see

$$L\left|\nabla J(\theta^P; Z)^T x\right| = L\left|b^T x\right| = |\mathcal{N}(0, \sigma^2\|x\|^2 L^2)| \sim \text{Half-Normal}(\sigma L\|x\|).$$

Now we can bound

$$\begin{aligned}\left|\log R(\theta^P)\right| &\leq \left|-\log\left(1-\frac{\beta}{\lambda}\right)\right| + \frac{L^2}{2\sigma^2} + \frac{1}{\sigma^2}\left|\nabla J(\theta^P; Z)^T \nabla \ell(\theta^P; z)\right| \\ &\leq \left|-\log\left(1-\frac{\beta}{\lambda}\right)\right| + \frac{L^2}{2\sigma^2} + \frac{L}{\sigma^2}\left|\nabla J(\theta^P; Z)^T x\right|.\end{aligned}$$

Plugging in this bound on $\left|\log R(\theta^P)\right|$,

$$\begin{aligned}D_\alpha(\hat{\theta}^P(Z) \,||\, \hat{\theta}^P(Z')) &\leq \frac{1}{\alpha-1}\log \mathbb{E}_{\theta^P\sim\hat{\theta}^P(Z)}\left[e^{(\alpha-1)\left|\log R(\theta^P)\right|}\right] \\ &\leq \frac{1}{\alpha-1}\log \mathbb{E}_{\theta^P\sim\hat{\theta}^P(Z)}\left[e^{(\alpha-1)\left[\left|-\log(1-\frac{\beta}{\lambda})\right| + \frac{L^2}{2\sigma^2}\right]} e^{(\alpha-1)\cdot\frac{L}{\sigma^2}\left|\nabla J(\theta^P;Z)^T x\right|}\right] \\ &= \left|-\log\left(1-\frac{\beta}{\lambda}\right)\right| + \frac{L}{2\sigma^2} + \frac{1}{\alpha-1}\log \mathbb{E}_{\theta^P\sim\hat{\theta}^P(Z)}\left[e^{(\alpha-1)\cdot\frac{L}{\sigma^2}\left|\nabla J(\theta^P;Z)^T x\right|}\right].\end{aligned}$$

Let p_σ be the probability density function of $b \sim \mathcal{N}(0, \sigma^2 I_d)$, and p_Θ the probability density function of $\theta^P \sim \hat{\theta}^P(Z)$. We know from Lemmas C.10.6 and C.10.7 that $\partial\theta = \left|\det \frac{\partial\theta}{\partial b}\right| \partial b$, and $p_\Theta(\theta) = \left|\det \frac{\partial b}{\partial\theta}\right| p_\sigma(b)$.

We also know from Lemma C.10.8 that $\left|\det \frac{\partial\theta}{\partial b}\right| \cdot \left|\det \frac{\partial b}{\partial\theta}\right| = 1$.

Using the change of variables $b = -\nabla J(\theta^P; Z)$ and $b^T x = u \sim \mathcal{N}(0, \sigma^2 ||x||_2^2)$, we have

$$\begin{aligned}
\mathbb{E}_{\theta^P \sim \hat{\theta}^P(Z)} \left[e^{(\alpha-1)\cdot \frac{L}{\sigma^2} \left| \nabla J(\theta^P; Z)^T x \right|} \right] &= \int_{\mathbb{R}^d} e^{(\alpha-1)\cdot \frac{L}{\sigma^2} \left| \nabla J(\theta^P; Z)^T x \right|} p_\Theta \left(\theta^P \right) \partial \theta \\
&= \int_{\mathbb{R}^d} e^{(\alpha-1)\cdot \frac{L}{\sigma^2} \left| b^T x \right|} \left| \det \frac{\partial b}{\partial \theta} \right| p_\sigma(b) \left| \det \frac{\partial \theta}{\partial b} \right| \partial b \\
&= \int_{\mathbb{R}^d} e^{(\alpha-1)\cdot \frac{L}{\sigma^2} \left| b^T x \right|} p_\sigma(b) \partial b \\
&= \mathbb{E}_{b \sim \mathcal{N}(0, \sigma^2 I_d)} \left[e^{(\alpha-1)\cdot \frac{L}{\sigma^2} \left| b^T x \right|} \right] \\
&= \mathbb{E}_{u \sim \mathcal{N}(0, \sigma^2 ||x||_2^2)} \left[e^{(\alpha-1) \left| \frac{L}{\sigma^2} u^2 \right|} \right] \\
&\leq \mathbb{E}_{\zeta \sim \mathcal{N}\left(0, \frac{L^2}{\sigma^2}\right)} \left[e^{(\alpha-1)|\zeta|} \right].
\end{aligned}$$

In the last line, we applied our assumption that $\|x\| \leq 1$ and the fact that the MGF of a half-normal R.V. increases monotonically when its scale parameter gets larger.

The above bound holds for the reverse Rényi divergence $D_\alpha(\hat{\theta}^P(Z') \,||\, \hat{\theta}^P(Z))$. Observe that

$$D_\alpha(\hat{\theta}^P(Z') \,||\, \hat{\theta}^P(Z)) \leq \frac{1}{\alpha - 1} \log \mathbb{E}_{\theta^P \sim \hat{\theta}^P(Z')} \left[e^{(\alpha-1) \left| \log R(\theta^P) \right|} \right].$$

This is because $\log \frac{\Pr\left[\hat{\theta}^P(Z') = \theta^P\right]}{\Pr\left[\hat{\theta}^P(Z) = \theta^P\right]} = -\log R(\theta^P) \leq |\log R(\theta^P)|$. If we use the change of variables $b = -\nabla J(\theta^P; Z')$ for the reverse direction, the above calculation works out identically (the difference is that p_Θ and the bijection between b and θ^P are different under Z and Z' — but the determinant of the mapping cancels out with its inverse just the same).

We've shown $\max\left(D_\alpha\left(\hat{\theta}^P(Z) \,||\, \hat{\theta}^P(Z')\right),\ D_\alpha\left(\hat{\theta}^P(Z') \,||\, \hat{\theta}^P(Z)\right) \right) \leq \epsilon(\alpha)$ for any neighboring datasets Z and Z', where

$$\epsilon(\alpha) = -\log\left(1 - \frac{\beta}{\lambda}\right) + \frac{L}{2\sigma^2} + \mathbb{E}\left[e^{(\alpha-1) \left| \mathcal{N}\left(0, \frac{L^2}{\sigma^2}\right) \right|} \right].$$

□

C.3.1 Linearized RDP Bound for Objective Perturbation

In our calculation of the RDP for objective perturbation, we needed to take an absolute value of the privacy loss random variable in order to handle negative values. But in doing so we end up with a quantity that depends on the moments of the *half-normal* distribution rather than those of the normal distribution, which gives us a looser bound. Can we avoid having to make this compromise? In this section we demonstrate that a linearization of the first-order conditions on the perturbed and unperturbed objective functions provides a more precise analysis of the PLRV of objective perturbation, translating to a tighter RDP bound in some regimes.

Recall that the objective perturbation mechanism is given by

$$\hat{\theta}^P(Z) = \sum_{i=1}^{n} \ell(\theta; z_i) + \frac{\lambda}{2}||\theta||_2^2 + b^T\theta, \tag{C.3.1}$$

where $b \sim \mathcal{N}(0, \sigma^2 I_d)$.

From the non-linearized RDP calculation, we know that for any neighboring datasets Z and Z',

$$D_\alpha\left(\hat{\theta}^P(Z) \,||\, \hat{\theta}^P(Z')\right) \leq \left|-\log\left(1-\frac{\beta}{\lambda}\right)\right| + \frac{L^2}{2\sigma^2} + \frac{1}{\alpha-1}\log \mathbb{E}_{b\sim\mathcal{N}(0,\sigma^2 I_d)}\left[e^{(\alpha-1)b^T\nabla\ell(\theta^P)}\right],$$

where θ^P is the output of the objective perturbation mechanism given the noise vector b. We can write

$$b^T\nabla\ell(\theta^P) = b^T\nabla\ell(\theta^*_\lambda) + b^T\left[\nabla\ell(\theta^P) - \nabla\ell(\theta^*_\lambda)\right].$$

The first term $t_1 = b^T \nabla \ell(\theta^*_\lambda)$ is a univariate Gaussian $t_1 \sim \mathcal{N}(0, \sigma^2 ||\nabla \ell(\theta^*_\lambda)||_2^2)$ because θ^*_λ is fixed w.r.t. b. We can bound the second term $t_2 = b^T \left[\nabla \ell(\theta^P) - \nabla \ell(\theta^*_\lambda)\right]$ using our assumptions on the loss function $\ell(\theta)$.

By assumption, the loss function has GLM structure $\ell(\theta; z) = f(x^T\theta; y)$. We can therefore write

$$\begin{aligned} b^T \left[\nabla \ell(\theta^P) - \nabla \ell(\theta^*_\lambda)\right] &= b^T \left[f'(x^T\theta^P; y)x - f'(x^T\theta^*_\lambda; y)x\right] \\ &= b^T x \left(f'(x^T\theta^P; y) - f'(x^T\theta^*_\lambda; y)\right) \end{aligned}$$

We have furthermore assumed that the function f is β-smooth, so that for any $z = (x, y)$ and $\theta^P, \theta^*_\lambda \in \mathbb{R}^d$ we have

$$\begin{aligned} \left|f'(x^T\theta^P; y) - f'(x^T\theta^*_\lambda; y)\right| &\leq \beta \left|x^T\theta^P - x^T\theta^*_\lambda\right| \\ &= \beta \left|x^T \left[\theta^P - \theta^*_\lambda\right]\right|. \end{aligned}$$

We will next apply Taylor's Theorem to rewrite $\theta^P - \theta^*_\lambda$.

Recall that θ^P is the minimizer of the perturbed objective:

$$\theta^P = \arg\min \left(\sum_{i=1}^n \ell(\theta; z_i) + \frac{\lambda}{2}||\theta||_2^2 + b^T\theta \right); \tag{C.3.2}$$

and θ^*_λ is the minimizer of the (non-private) regularized objective:

$$\theta^*_\lambda = \arg\min \left(\sum_{i=1}^n \ell(\theta; z_i) + \frac{\lambda}{2}||\theta||_2^2 \right). \tag{C.3.3}$$

Parameterize the line segment between θ^P and θ^*_λ by $t \in [0, 1]$, i.e. the line segment is $t(\theta^P - \theta^*_\lambda) + \theta^P$. By Taylor's Theorem, there exists $\theta' = t'(\theta^P - \theta^*_\lambda) + \theta^P$ for some

$t' \in [0,1]$ such that

$$\nabla \ell(\theta^P) - \nabla \ell(\theta^*_\lambda) = \nabla^2 \ell(\theta')(\theta^P - \theta^*_\lambda).$$

By first-order conditions on Equations C.3.2 and C.3.3,

$$\nabla \mathcal{L}(\theta^P) + \lambda \theta^P + b = 0; \tag{C.3.4}$$

$$\nabla \mathcal{L}(\theta^*_\lambda) + \lambda \theta^*_\lambda = 0. \tag{C.3.5}$$

Then subtracting Equation C.3.5 from Equation C.3.4, we have that

$$\nabla \mathcal{L}(\theta^P) - \nabla \mathcal{L}(\theta^*_\lambda) + \lambda(\theta^P - \theta^*_\lambda) + b = 0. \tag{C.3.6}$$

Again applying Taylor's theorem, there exists $\theta'' = t''(\theta^P - \theta^*_\lambda) + \theta^P$ for some $t'' \in [0,1]$) such that

$$\nabla \mathcal{L}(\theta^P) - \nabla \mathcal{L}(\theta^*_\lambda) = \nabla^2 \mathcal{L}(\theta'')(\theta^P - \theta^*_\lambda). \tag{C.3.7}$$

Putting together Equations C.3.6 and C.3.7 we then have

$$\theta^P - \theta^*_\lambda = -\left(\nabla^2 \mathcal{L}(\theta'') + \lambda I_d\right)^{-1} b. \tag{C.3.8}$$

So we now have

$$\begin{aligned} b^T \left[\nabla \ell(\theta^P) - \nabla \ell(\theta^*_\lambda)\right] &\leq \beta \left|b^T x\right| \left|x^T \left(\theta^P - \theta^*_\lambda\right)\right| \\ &\leq \beta \left|b^T x\right| \left|x^T \left(\nabla^2 \mathcal{L}(\theta^*_\lambda) + \lambda I_d\right)^{-1} b\right|. \end{aligned} \tag{C.3.9}$$

Note that since $e^x > 0$ for all $x \in \mathbb{R}$, we have that $\mathbb{E}\left[|e^x|\right] = \mathbb{E}\left[e^x\right]$.

Let $a := \frac{1}{\sigma^2} b^T \nabla \ell(\theta^*)$ and $c := \frac{1}{\sigma^2} b^T \left[\nabla \ell(\theta^P) - \nabla \ell(\theta^*)\right]$. Then by Holder's inequality,

$$\begin{aligned}
\mathbb{E}\left[e^{(\alpha-1)a} e^{(\alpha-1)c}\right] &= \mathbb{E}\left[\left|e^{(\alpha-1)a} e^{(\alpha-1)c}\right|\right] \\
&\leq \mathbb{E}\left[\left|e^{(\alpha-1)a}\right|^p\right]^{\frac{1}{p}} \mathbb{E}\left[\left|e^{(\alpha-1)c}\right|^q\right]^{\frac{1}{q}} \\
&= \mathbb{E}\left[e^{(p\alpha-p)a}\right]^{\frac{1}{p}} \mathbb{E}\left[e^{(q\alpha-q)c}\right]^{\frac{1}{q}}.
\end{aligned}$$

By the GLM assumption, $b^T \nabla \ell(\theta^*_\lambda; z) = f'(x^T \theta^*_\lambda; y) b^T x$. Then

$$\begin{aligned}
\mathbb{E}_{b \sim \mathcal{N}(0, \sigma^2 I_d)}\left[e^{(p\alpha-p)\frac{1}{\sigma^2} b^T \nabla \ell(\theta^P)}\right] &= \mathbb{E}_{b \sim \mathcal{N}(0, \sigma^2 I_d)}\left[e^{(p\alpha-p)\frac{1}{\sigma^2} f'(x^T \theta^*_\lambda; y) b^T x}\right] \\
&= \mathbb{E}_{u_1 \sim \mathcal{N}\left(0, f'(x^T \theta^*_\lambda; y)^2 ||x||_2^2 \frac{1}{\sigma^2}\right)}\left[e^{(p\alpha-p)u_1}\right] \\
&\leq \mathbb{E}_{u_2 \sim \mathcal{N}\left(0, \frac{L^2}{\sigma^2}\right)}\left[e^{(p\alpha-p)u_2}\right].
\end{aligned}$$

Above, we've applied the assumption that $||x||_2 \leq 1$ and the fact that the MGF of a normal R.V. increases monotonically when its scale parameter gets larger (Lemma C.10.12). By Lemma C.10.9, we have that $x^T \left(\nabla^2 \mathcal{L}(\theta^*_\lambda) + I_d\right)^{-2} x \leq \frac{||x||_2^2}{\lambda^2}$. From C.3.9 we also have

$$\mathbb{E}_{b \sim \mathcal{N}(0, \sigma^2 I_d)}\left[e^{(q\alpha-q)\frac{1}{\sigma^2} b^T \left[\nabla \ell(\theta^P) - \nabla \ell(\theta^*_\lambda)\right]}\right] \leq \mathbb{E}_{b \sim \mathcal{N}(0, \sigma^2 I_d)}\left[e^{(q\alpha-q)\beta \left|b^T x\right| \left|x^T \left(\nabla^2 \mathcal{L}(\theta^*) + \lambda I_d\right)^{-1} b\right|}\right]$$

Define $z_1 := b^T x$ and $z_2 := x^T \left(\nabla^2 \mathcal{L}(\theta^*_\lambda) + \lambda I_d\right)^{-1} b$, and observe

$$z_1 \sim \mathcal{N}\left(0, \sigma^2 ||x||_2^2\right),$$
$$z_2 \sim \mathcal{N}\left(0, \sigma^2 x^T \left(\nabla^2 \mathcal{L}(\theta^*_\lambda) + I_d\right)^{-2} x\right).$$

Note that our approach below is agnostic to the relationship between $|z_1|$ and $|z_2|$; in reality, they depend on each other through the noise vector b. Again applying the assumption $||x||_2^2 \leq 1$ and Lemma C.10.12 (while not forgetting that the random variables z_1 and z_2 depend on each other through b), we get

$$\begin{aligned}
\mathbb{E}_{b\sim\mathcal{N}(0,\sigma^2 I_d)}\left[e^{(q\alpha-q)\beta\left|b^T x\right|\left|x^T\left(\nabla^2\mathcal{L}(\theta^*_\lambda)+\lambda I_d\right)^{-1} b\right|}\right] &= \mathbb{E}_{z_1,z_2}\left[e^{(q\alpha-q)\beta|z_1|\,|z_2|}\right] \\
&\leq \mathbb{E}_{z_3\sim\mathcal{N}(0,\sigma^2),z_4\sim\mathcal{N}(0,\frac{\sigma^2}{\lambda^2})}\left[e^{(q\alpha-q)\beta|z_3|\,|z_4|}\right] \\
&= \mathbb{E}_{z_3\sim\mathcal{N}(0,\sigma^2),z_5\sim\mathcal{N}(0,\sigma^2)}\left[e^{(q\alpha-q)\frac{\beta}{\lambda}|z_3|\,|z_5|}\right] \\
&\leq \mathbb{E}_{z\sim\mathcal{N}(0,\sigma^2)}\left[e^{(q\alpha-q)\frac{\beta}{\lambda}z^2}\right].
\end{aligned}$$

So altogether, for p, q such that $\frac{1}{p} + \frac{1}{q} = 1$, we get

$$\begin{aligned}
&D_\alpha\left(\hat{\theta}^P(Z) \,||\, \hat{\theta}^P(Z')\right) \leq \\
&\quad -\log\left(1-\frac{\beta}{\lambda}\right) + \frac{L^2}{2\sigma^2} + \frac{1}{\alpha-1}\log \mathbb{E}_{u\sim\mathcal{N}(0,\frac{L^2}{\sigma^2})}\left[e^{(p\alpha-p)u}\right]^{\frac{1}{p}} \mathbb{E}_{z\sim\mathcal{N}(0,\sigma^2)}\left[e^{(q\alpha-q)\frac{\beta}{\lambda}z^2}\right]^{\frac{1}{q}}.
\end{aligned}$$

C.3.2 Distance to Optimality

Consider the mechanism $\mathcal{M}(Z) = f(Z) + \mathcal{N}(0, \sigma^2 I_d)$, for a function $f : \mathcal{Z} \to \mathbb{R}^d$ with sensitivity $\Delta_f = L$. From Balle and Wang (2018), we know that for any neighboring datasets Z and Z', the privacy loss random variable of this mechanism is distributed

as $\mathcal{N}\left(\frac{\Delta^2_{Z,Z'}}{2\sigma^2}, \frac{\Delta^2_{Z,Z'}}{\sigma^2}\right)$. Maximizing the Rényi divergence $D_\alpha\left(\mathcal{M}(Z) \,||\, \mathcal{M}(Z')\right)$ over all neighboring datasets $Z \simeq Z'$ shows that the RDP for the Gaussian mechanism can be written as

$$\begin{aligned}\epsilon(\alpha) &= \frac{1}{\alpha-1}\log \mathbb{E}\left[e^{(\alpha-1)}\mathcal{N}\left(\frac{L^2}{2\sigma^2}, \frac{L^2}{\sigma^2}\right)\right] \\ &= \frac{L^2}{2\sigma^2} + \frac{1}{\alpha-1}\log \mathbb{E}\left[e^{(\alpha-1)}\mathcal{N}\left(0, \frac{L^2}{\sigma^2}\right)\right].\end{aligned}$$

Thus the main deviations between the RDP bound for objective perturbation and that of the Gaussian mechanism are 1) the leading term (a function of β and λ that vanishes as we increase the regularization) and 2) the moment-generating function of the *half-normal* (instead of normal) distribution.

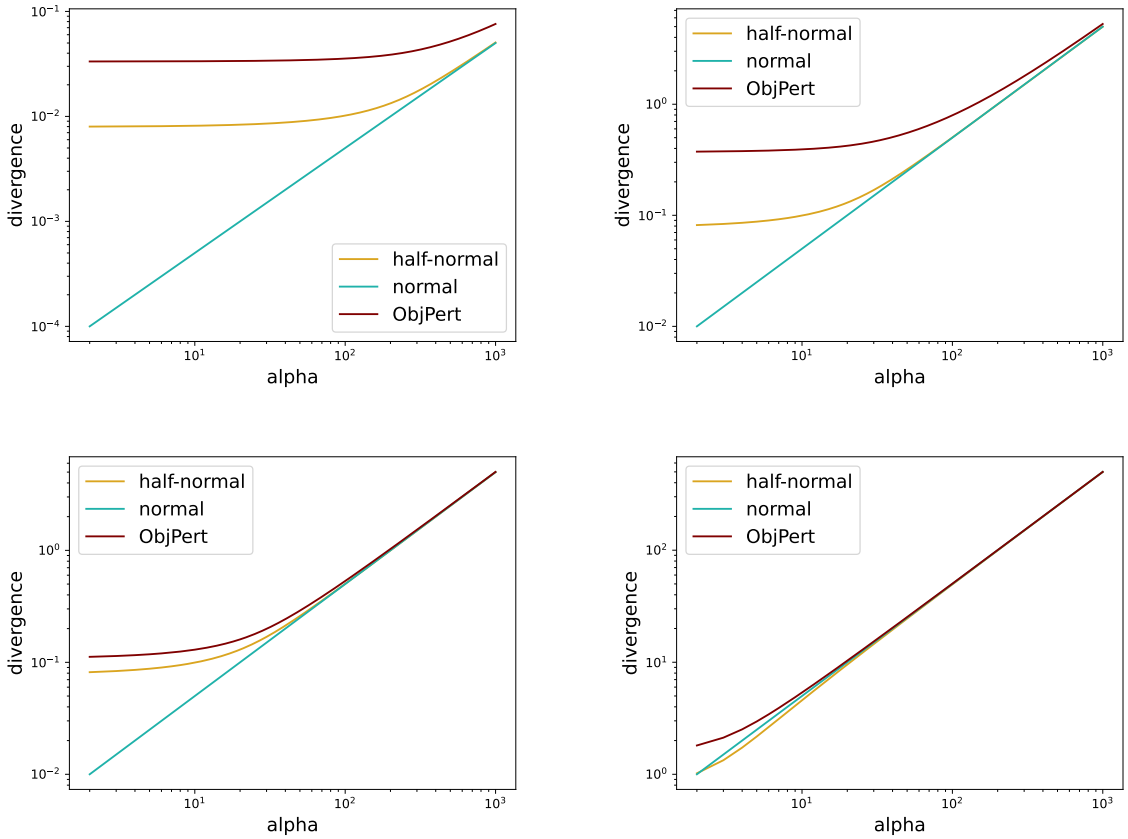


Figure C.1: RDP curves $\epsilon(\alpha)$ for the objective perturbation mechanism, the Gaussian mechanism, and the half-normal mechanism.

Figure C.3.2 plots the Rényi divergence $\epsilon(\alpha) := D_\alpha(\mathcal{M}(Z) \,||\, \mathcal{M}(Z'))$ for the Gaussian mechanism ("normal"), the objective perturbation mechanism ("ObjPert"), and the mechanism [2] obtained by adding noise from the half-normal distribution ("half-normal").

We consider several different regimes of interest by varying the noise scale σ and the regularization strength λ. There are several takeaways to observe:

1. The difference between the RDP for the half-normal mechanism and the RDP for the objective perturbation mechanism is due entirely to the leading term of the bound

[2] More formally, we say that the "half-normal" mechanism is $\mathcal{M}(Z) = f(Z) + \left|\mathcal{N}(0, \sigma^2)\right|$, where $f : \mathcal{Z} \to \mathbb{R}$ is the function we wish to release.

given in Theorem 4.3.2, which vanishes as λ increases (Figures **??** and **??**). For smaller λ there is a constant "start-up" gap between the half-normal and Objpert RDP curves (best displayed in Figure **??**) which disappears for larger α, where the moments of the half-normal distribution overwhelm the contribution of the leading term of the objective perturbation RDP.

2. As σ increases (e.g. between Figures **??** and **??**, and between Figures **??** and **??**), the half-normal curve – and therefore also the ObjPert curve – doesn't converge with the normal curve until larger α.

C.4 Hockey-stick Divergence Analysis of Objective Perturbation

C.4.1 Further Details on Hockey-stick Divergence Analysis

Using dominating pairs of distributions (Def. 4.2.5) for all the individual mechanisms in an adaptive composition, we can obtain accurate (ϵ, δ)-bounds for the whole composition. For this end we need the following result.

Theorem C.4.1 (Zhu et al. 2022). *If (P, Q) dominates $\mathcal{M}$ and (P', Q') dominates $\mathcal{M}'$ for all inputs of $\mathcal{M}'$, then $(P \times P', Q \times Q')$ dominates the adaptive composition $\mathcal{M} \circ \mathcal{M}'$.*

To get the hockey-stick divergence from $P \times P'$ to $Q \times Q'$ into an efficiently computable form, we express it using so called privacy loss random variables (recall Def. C.10.4). If P and Q are probability density functions, the privacy loss function $\mathcal{L}_{P/Q}$ is defined as

$$\mathcal{L}_{P/Q}(x) = \log \frac{P(x)}{Q(x)}$$

and the privacy loss random variable (PLRV) $\omega_{P/Q}$ as

$$\omega_{P/Q} = \mathcal{L}_{P/Q}(x), \quad x \sim P(x).$$

The $\delta(\epsilon)$-bounds can be represented using the following representation that involves the PLRV.

Theorem C.4.2 (Gopi et al. 2021). *We have:*

$$H_{\mathrm{e}^\epsilon}(P||Q) = \mathop{\mathbb{E}}_{x\sim P}\left[1 - \mathrm{e}^{\epsilon - \mathcal{L}_{P/Q}(x)}\right]_+ = \mathop{\mathbb{E}}_{s\sim\omega_{P/Q}}\left[1 - \mathrm{e}^{\epsilon - s}\right]_+. \tag{C.4.1}$$

Moreover, if $\omega_{P/Q}$ is the PLRV for the pair of distributions (P, Q) and $\omega_{P'/Q'}$ the PLRV for the pair of distributions (P', Q'), then the PLRV for the pair of distributions $(P\times P', Q\times Q')$ is given by $\omega_{P/Q} + \omega_{P'/Q'}$.

By Theorem C.4.2, to computing accurate (ϵ, δ)-bounds for compositions, it suffices that we can evaluate integrals of the form $\mathbb{E}_{s\sim\omega_1+\ldots+\omega_k}\left[1 - \mathrm{e}^{\epsilon - s}\right]_+$. For this we can use the Fast Fourier Transform (FFT)-based method by Koskela et al. (2021), where the distribution of each PLRV is truncated and placed on an equidistant numerical grid over an interval $[-L, L]$, where $L > 0$ is a pre-defined parameter. The distributions for the sums of the PLRVs are given by convolutions of the individual distributions and can be evaluated using the FFT algorithm. By a careful error analysis the error incurred by the numerical method can be bounded and an upper $\delta(\epsilon)$-bound obtained. For accurately carrying out this numerical computation one could also use, for example, the FFT-based method proposed by Gopi et al. (2021).

C.4.2 Proof of Theorem 4.3.1

Before giving a proof to Thm. 4.3.1, we first give the following bound which is a hockey-stick equivalent of the moment-generating function bound given in Thm. 4.3.2.

Lemma C.4.3. *Let $\epsilon \in \mathbb{R}$ and let the objective perturbation mechanism $\hat{\theta}^P$ be defined as in Section 4.2.2. Let $||\nabla \ell(\theta; z)||_2 \leq L$ and $\nabla^2 \ell(\theta; z) \prec \beta I_d$ for all $\theta \in \Theta$ and $z \in \mathcal{X} \times \mathcal{Y}$. Then, for any neighboring datasets Z and Z', we have:*

$$H_{e^\epsilon}\big(\hat{\theta}^P(Z)||\hat{\theta}^P(Z')\big) \leq \mathbf{E}_{s\sim\omega}[1 - e^{\epsilon - s}]_+, \tag{C.4.2}$$

where $\omega \sim \left|\log\left(1 - \frac{\beta}{\lambda}\right)\right| + \frac{L^2}{2\sigma^2} + \left|\mathcal{N}\left(\frac{\|x\|^2 L^2}{\sigma^2}\right)\right|$.

Proof. The proof goes analogously to the proof of Thm. 4.3.2. Let Z and Z' be any neighboring datasets. Following the proof of Thm. 4.3.2, denote the privacy loss

$$R(\theta) := \frac{\Pr\left[\hat{\theta}^P(Z) = \theta\right]}{\Pr\left[\hat{\theta}^P(Z') = \theta\right]}.$$

By Thm. C.4.2 and by using the reasoning of the proof of Thm. 4.3.2 for the moment-generating function, we have

$$\begin{aligned} H_{e^\epsilon}\big(\hat{\theta}^P(Z)||\hat{\theta}^P(Z')\big) &= \mathbf{E}_{\theta\sim\hat{\theta}^P(Z)}[1 - e^{\epsilon - \log R(\theta)}]_+ \\ &\leq \mathbf{E}_{\theta\sim\hat{\theta}^P(Z)}[1 - e^{\epsilon - |\log R(\theta)|}]_+ \\ &\leq \mathbf{E}_{s\sim\left|\mathcal{N}\left(\frac{\|x\|^2 L^2}{\sigma^2}\right)\right|}\left[1 - e^{\epsilon - \left|\log\left(1-\frac{\beta}{\lambda}\right)\right| - \frac{L^2}{2\sigma^2} - s}\right]_+, \end{aligned} \tag{C.4.3}$$

where the inequalities follow from the fact that the function $f(s) = [1 - e^{\epsilon - s}]_+$ is monotonically increasing function w.r.t. s for all $\epsilon \in \mathbb{R}$ and from the bound for $|R(\theta)|$ used in the proof of Thm. 4.3.2. □

Proof of Theorem 4.3.1. We use Lemma C.4.3 and simply upper bound the right-hand side of the inequality (C.4.2).

We first show that if $\|x\| \leq 1$, then for all $\epsilon \in \mathbb{R}$

$$\mathbf{E}_{s\sim\left|\log\left(1-\frac{\beta}{\lambda}\right)\right|+\frac{L^2}{2\sigma^2}\left|\mathcal{N}(0,\frac{\|x\|^2L^2}{\sigma^2})\right|}[1-e^{\epsilon-s}]_+ \leq \mathbf{E}_{s\sim\left|\log\left(1-\frac{\beta}{\lambda}\right)\right|+\frac{L^2}{2\sigma^2}\left|\mathcal{N}(0,\frac{L^2}{\sigma^2})\right|}[1-e^{\epsilon-s}]_+. \quad \text{(C.4.4)}$$

Denote $\hat{\epsilon} = \epsilon - \left|\log\left(1-\frac{\beta}{\lambda}\right)\right| - \frac{L^2}{2\sigma^2}$. Consider first the case $\hat{\epsilon} \geq 0$. Then, we have:

$$\begin{aligned}
\mathbf{E}_{s\sim\left|\mathcal{N}(0,\frac{\|x\|^2L^2}{\sigma^2})\right|}[1-e^{\epsilon-\left|\log\left(1-\frac{\beta}{\lambda}\right)\right|-\frac{L^2}{2\sigma^2}-s}]_+ &= 2\cdot\mathbf{E}_{s\sim\mathcal{N}(0,\frac{\|x\|^2L^2}{\sigma^2})}[1-e^{\hat{\epsilon}-s}]_+ \\
&\leq 2\cdot\mathbf{E}_{s\sim\mathcal{N}(0,\frac{L^2}{\sigma^2})}[1-e^{\hat{\epsilon}-s}]_+ \qquad \text{(C.4.5)} \\
&= \mathbf{E}_{s\sim\left|\mathcal{N}(0,\frac{L^2}{\sigma^2})\right|}[1-e^{\epsilon-\left|\log\left(1-\frac{\beta}{\lambda}\right)\right|-\frac{L^2}{2\sigma^2}-s}]_+,
\end{aligned}$$

where the first equality follows from the fact that for $s \geq \hat{\epsilon}$, the density function of the half-normal random variable is positive and 2 times the density of the corresponding normal distribution. The inequality follows from Lemma C.4.8, as

$$\mathbf{E}_{s\sim\mathcal{N}(0,\frac{\|x\|^2L^2}{\sigma^2})}[1-e^{\hat{\epsilon}-s}]_+ = \int_{\hat{\epsilon}}^{\infty} f_{0,\frac{\|x\|^2L^2}{\sigma^2}}(x)(1-\mathrm{e}^{\hat{\epsilon}-x})\,\mathrm{d}x.$$

Next, consider the case $\hat{\epsilon} < 0$. Then:

$$\begin{aligned}
\mathbf{E}_{s\sim\left|\mathcal{N}(0,\frac{\|x\|^2L^2}{\sigma^2})\right|}[1-e^{\epsilon-\left|\log\left(1-\frac{\beta}{\lambda}\right)\right|-\frac{L^2}{2\sigma^2}-s}]_+ &= 2\cdot\int_0^{\infty} f_{0,\frac{\|x\|^2L^2}{\sigma^2}}(x)(1-e^{\epsilon-\left|\log\left(1-\frac{\beta}{\lambda}\right)\right|-\frac{L^2}{2\sigma^2}-x})\,\mathrm{d}x \\
&\leq 2\cdot\int_0^{\infty} f_{0,\frac{L^2}{\sigma^2}}(x)(1-e^{\epsilon-\left|\log\left(1-\frac{\beta}{\lambda}\right)\right|-\frac{L^2}{2\sigma^2}-x})\,\mathrm{d}x \\
&= \mathbf{E}_{s\sim\left|\mathcal{N}(0,\frac{L^2}{\sigma^2})\right|}[1-e^{\epsilon-\left|\log\left(1-\frac{\beta}{\lambda}\right)\right|-\frac{L^2}{2\sigma^2}-s}]_+. \qquad \text{(C.4.6)}
\end{aligned}$$

where the inequality follows from Lemma C.4.8. Inequalities (C.4.5) and (C.4.6) together give (C.4.4).

Then, we show that for all $\epsilon \in \mathbb{R}$,

$$\mathbf{E}_{s\sim\left|\log\left(1-\frac{\beta}{\lambda}\right)\right|+\frac{L^2}{2\sigma^2}\left|\mathcal{N}(0,\frac{L^2}{\sigma^2})\right|}[1-e^{\epsilon-s}]_+ = \begin{cases} 2\cdot H_{e^{\tilde{\epsilon}}}\big(P||Q\big), & \text{if } \hat{\epsilon}\geq 0, \\ (1-\mathrm{e}^{\hat{\epsilon}})+\mathrm{e}^{\hat{\epsilon}}\cdot 2\cdot H_{e^{\frac{L^2}{\sigma^2}}}\big(P||Q\big), & \text{otherwise.} \end{cases}$$

Continuing from (C.4.5), by change of variables, we see that for $\hat{\epsilon}\geq 0$,

$$\begin{aligned} 2\cdot\mathbf{E}_{s\sim\mathcal{N}(0,\frac{L^2}{\sigma^2})}[1-e^{\epsilon-\left|\log\left(1-\frac{\beta}{\lambda}\right)\right|-\frac{L^2}{2\sigma^2}-s}]_+ &= 2\cdot\mathbf{E}_{s\sim\mathcal{N}(\frac{L^2}{2\sigma^2},\frac{L^2}{\sigma^2})}[1-e^{\epsilon-\left|\log\left(1-\frac{\beta}{\lambda}\right)\right|-s}]_+ \\ &= 2\cdot H_{e^{\tilde{\epsilon}}}\big(P||Q\big), \end{aligned}$$

where $\tilde{\epsilon}=\epsilon-\left|\log\left(1-\frac{\beta}{\lambda}\right)\right|$, P is the density function of $\mathcal{N}(L,\sigma^2)$ and Q the density function of $\mathcal{N}(0,\sigma^2)$. This follows from the fact that the PLRV determined by the pair (P,Q) is distributed as $\mathcal{N}(\frac{L^2}{2\sigma^2},\frac{L^2}{\sigma^2})$.

Continuing from (C.4.6), by change of variables (used after the third equality sign), we see that for $\hat{\epsilon}z0$,

$$\begin{aligned} 2\cdot&\int_0^\infty f_{0,\frac{L^2}{\sigma^2}}(x)(1-e^{\epsilon-\left|\log\left(1-\frac{\beta}{\lambda}\right)\right|-\frac{L^2}{2\sigma^2}-x})\,\mathrm{d}x \\ &= 2\cdot\int_0^\infty f_{0,\frac{L^2}{\sigma^2}}(x)\,\mathrm{d}x - 2\cdot\int_0^\infty f_{0,\frac{L^2}{\sigma^2}}(x)(1-e^{\hat{\epsilon}-x})\,\mathrm{d}x \\ &= (1-\mathrm{e}^{\hat{\epsilon}})\cdot 2\cdot\int_0^\infty f_{0,\frac{L^2}{\sigma^2}}(x)\,\mathrm{d}x + \mathrm{e}^{\hat{\epsilon}}\cdot 2\cdot\int_0^\infty f_{0,\frac{L^2}{\sigma^2}}(x)(1-e^{-x})\,\mathrm{d}x \\ &= (1-\mathrm{e}^{\hat{\epsilon}})+\mathrm{e}^{\hat{\epsilon}}\cdot 2\cdot\int_{\frac{L^2}{2\sigma^2}}^\infty f_{\frac{L^2}{2\sigma^2},\frac{L^2}{\sigma^2}}(x)(1-e^{\frac{L^2}{2\sigma^2}-x})\,\mathrm{d}x \\ &= (1-\mathrm{e}^{\hat{\epsilon}})+\mathrm{e}^{\hat{\epsilon}}\cdot 2\cdot\mathbf{E}_{s\sim\mathcal{N}(\frac{L^2}{2\sigma^2},\frac{L^2}{\sigma^2})}[1-e^{\frac{L^2}{2\sigma^2}-s}]_+ \\ &= (1-\mathrm{e}^{\hat{\epsilon}})+\mathrm{e}^{\hat{\epsilon}}\cdot 2\cdot H_{e^{\frac{L^2}{2\sigma^2}}}\big(P||Q\big), \end{aligned}$$

where we again use the fact that the PLRV of the Gaussian mechanism with sensitivity L and noise scale σ is distributed as $\mathcal{N}(\frac{L^2}{2\sigma^2},\frac{L^2}{\sigma^2})$. □

C.4.3 Dominating Pairs of Distributions for the Objective Perturbation Mechanism

From Lemma C.4.3 and the inequality (C.4.4) we have that for all $\epsilon \in \mathbb{R}$

$$H_{e^\epsilon}\big(\hat{\theta}^P(Z)||\hat{\theta}^P(Z')\big) \leq \mathbf{E}_{\omega\sim\left|\log\left(1-\frac{\beta}{\lambda}\right)\right|+\frac{L^2}{2\sigma^2}+\left|\mathcal{N}(0,\frac{L^2}{\sigma^2})\right|}[1-e^{\epsilon-\omega}]_+.$$

Thus, if we have distributions P and Q such that for all $\epsilon \in \mathbb{R}$

$$H_{e^\epsilon}\big(P||Q\big) = \mathbf{E}_{\omega\sim\left|\log\left(1-\frac{\beta}{\lambda}\right)\right|+\frac{L^2}{2\sigma^2}+\left|\mathcal{N}(0,\frac{L^2}{\sigma^2})\right|}[1-e^{\epsilon-\omega}]_+,$$

then the pair (P, Q) is a dominating pair of distributions for the objective perturbation mechanism. Then, by Theorem C.4.2, we can use this distribution ω also to compute (ϵ, δ)-bounds for compositions involving the objective perturbation mechanism. We give such a pair of distribution (P, Q) explicitly in Lemma C.4.6 below.

In the following, we denote the density of a discrete probability mass by a Dirac delta function and use the indicator function for the continuous part of the density. The following result is a straightforward calculation.

Lemma C.4.4. *Let $\sigma > 0$. Let P be the density function of $|\mathcal{N}(0, \sigma^2)|$, i.e.,*

$$P(x) = \frac{2}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{\frac{-x^2}{2\sigma^2}}\mathbf{1}_{[0,\infty)]}(x), \tag{C.4.7}$$

where $\mathbf{1}_A(x)$ denotes the indicator function, i.e., $\mathbf{1}_{[0,\infty)]}(x) = 1$ if $x \geq 0$, else $\mathbf{1}_{[0,\infty)]}(x) = 0$. Let $L > 0$ and let Q be a density function, where part of the mass of P is shifted to $-\infty$:

$$Q(x) = Q(-\infty)\cdot\delta_{-\infty}(x) + \frac{2}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{\frac{-(t+L)^2}{2\sigma^2}}\mathbf{1}_{[0,\infty)]}(x), \tag{C.4.8}$$

where

$$Q(-\infty) = 1 - \int_0^\infty \frac{2}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{\frac{-(t+L)^2}{2\sigma^2}} \, \mathrm{d}x.$$

Then, we have that the PLRV ω,

$$\omega = \log \frac{P(x)}{Q(x)}, \quad x \sim P, \tag{C.4.9}$$

is distributed as

$$\omega \sim \frac{L^2}{2\sigma^2} + \left| \mathcal{N}\left(0, \frac{L^2}{\sigma^2}\right) \right|.$$

Proof. As P has its support on $[0, \infty)$, we need to consider the values of the privacy loss function $\log \frac{P(x)}{Q(x)}$ only on $[0, \infty)$. We have, for all $x \geq 0$,

$$\log \frac{P(x)}{Q(x)} = \frac{L}{\sigma^2} \cdot x + \frac{L^2}{2\sigma^2}.$$

Since $x \sim P$, we see that $\frac{L}{\sigma^2} \cdot x \sim \left| \mathcal{N}\left(0, \frac{L^2}{\sigma^2}\right) \right|$ and the claim follows. □

Remark C.4.5. In Lemma C.4.4, instead of shifting part of the mass of P to $-\infty$ when forming Q, we could place this mass anywhere on the negative real axis. This would not affect the PLRV ω.

We can shift the PLRV ω given by Lemma C.4.4 by scaling the distribution Q. We get the following.

Lemma C.4.6. *Let $\sigma > 0$ and $L > 0$. Suppose P is the density function given in Eq.* (C.4.7) *and Q the density function*

$$Q(x) = Q(-\infty) \cdot \delta_{-\infty}(x) + \mathrm{e}^{-\left|\log\left(1-\frac{\beta}{\lambda}\right)\right|} \cdot \frac{2}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{\frac{-(t-L)^2}{2\sigma^2}} \mathbf{1}_{[0,\infty)]}(x), \tag{C.4.10}$$

where

$$Q(-\infty) = \left(1 - \mathrm{e}^{-\left|\log\left(1-\frac{\beta}{\lambda}\right)\right|} \cdot \int_0^\infty \frac{2}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{\frac{-(t-L)^2}{2\sigma}^2} \, \mathrm{d}x\right).$$

Then, the PLRV ω determined by P and Q is distributed as

$$\omega \sim \left|\log\left(1-\frac{\beta}{\lambda}\right)\right| + \frac{L^2}{2\sigma^2} + \left|\mathcal{N}\left(0, \frac{L^2}{\sigma^2}\right)\right|.$$

Proof. Showing this goes as the proof of Lemma C.4.4. We just now have that for all $x \geq 0$:

$$\log \frac{P(x)}{Q(x)} = \left|\log\left(1-\frac{\beta}{\lambda}\right)\right| + \frac{L^2}{2\sigma^2} + \frac{L}{\sigma^2} \cdot x.$$

□

As a corollary of Lemma C.4.6 and Thm. 4.2.4, we have:

Lemma C.4.7. *Let $k \in \mathbb{Z}_+$ and let for each $i \in [k]$*

$$\omega_i \sim \left|\log\left(1-\frac{\beta}{\lambda}\right)\right| + \frac{L^2}{2\sigma^2} + \left|\mathcal{N}(0, \frac{L^2}{\sigma^2})\right|,$$

such that ω_i's are independent. Then, the k-wise adaptive composition of $\hat{\theta}(Z)$ is $(\epsilon, \delta(\epsilon))$-DP for

$$\delta(\epsilon) = \mathbf{E}_{s\sim\omega_1+\ldots+\omega_k}[1 - e^{\epsilon - s}]_+. \tag{C.4.11}$$

Numerical Evaluation of (ϵ, δ)-Bounds for Compositions

Figure C.2 shows the result of applying the FFT-based numerical method of Koskela et al. (2021) for evaluating the expression (C.4.11). We compare the resulting approximate DP bounds to those obtained from the RDP bounds combined with standard composition results (Mironov, 2017).

Notice that we could also carry out tighter accounting of the approximative minima

perturbation (Section 4.4) by adding the PLRVs of the Gaussian mechanism to the total PLRV, similarly as RDP parameters of the Gaussian mechanism are added to the RDP guarantees of the objective perturbation mechanism (Theorem 4.4.1). Adding the Gaussian PLRV to the total PLRV using convolutions is straightforward using the method of Koskela et al. (2021).

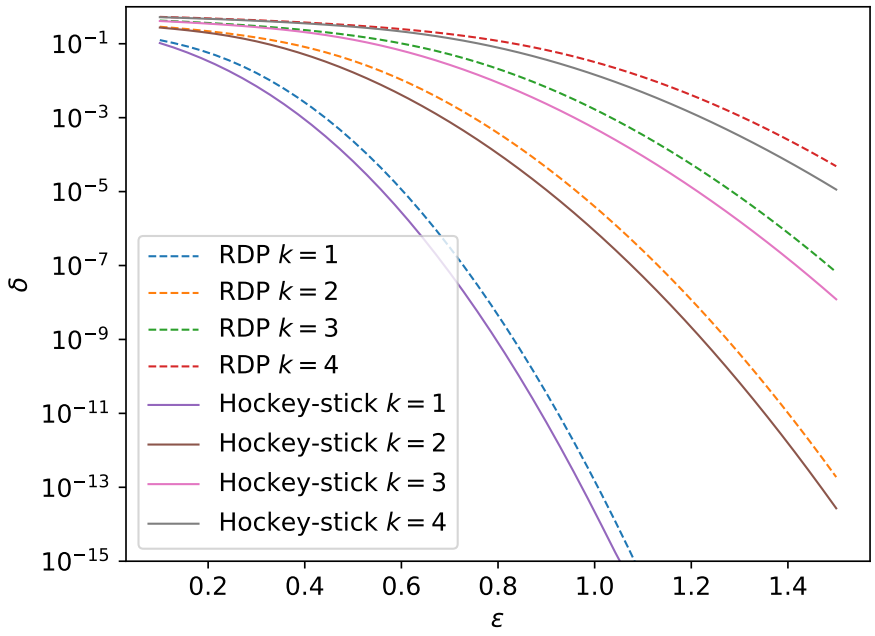


Figure C.2: Comparison of our RDP bound (implied (ϵ, δ)-DP bound) and our numerical PLRV bound (C.4.11) for different numbers of compositions k, when $\sigma = 8.0$, $\beta = 1.0$ and $\lambda = 10.0$.

C.4.4 Auxiliary Lemma

For Theorem 4.3.1, we need the following auxiliary result.

Lemma C.4.8. *Denote* $f_{\mu,\sigma^2}(x)$ *the density function of the normal distribution* $\mathcal{N}(\mu, \sigma^2)$ *and let* $c \geq \mu$. *Let* $g(x)$ *be a non-negative differentiable non-decreasing function on* $[c, \infty)$.

Then, if $\sigma_1 \leq \sigma_2$,

$$\int_c^\infty f_{\mu,\sigma_1^2}(x) \cdot g(x)\,\mathrm{d}x \leq \int_c^\infty f_{\mu,\sigma_2^2}(x) \cdot g(x)\,\mathrm{d}x.$$

Proof. By integration by parts, we have

$$\int_c^\infty f_{\mu,\sigma^2}(x) \cdot g(x)\,\mathrm{d}x = -\Phi_{\mu,\sigma^2}(c) \cdot g(c) - \int_c^\infty \Phi_{\mu,\sigma^2}(x) \cdot g'(x)\,\mathrm{d}x, \tag{C.4.12}$$

where $\Phi_{\mu,\sigma^2}(x)$ denotes the cdf of $\mathcal{N}(\mu,\sigma^2)$. A simple calculation shows that for all $x \in \mathbb{R}$,

$$\frac{\partial}{\partial\sigma}\Phi_{\mu,\sigma^2}(x) = -\frac{x-\mu}{\sigma^2} f_{\mu,\sigma^2}(x).$$

Thus, $\Phi_{\mu,\sigma^2}(c)$ is a non-increasing function of σ for all $c \geq \mu$. Furthermore, the first term in (C.4.12) is a non-decreasing function of σ since $g(c)$ is non-negative and the second term is a non-decreasing function of σ, since $g'(x)$ is non-negative for all $x \in [c,\infty)$. □

C.5 The GLM Bug

C.5.1 Discussion

Limiting our main results to generalized linear models might appear restrictive — but we argue that the GLM assumption is not specific to our paper, but rather has been lurking in the objective perturbation literature for some time now.

Let's first take a look at Section 3.3.2 of Chaudhuri et al. (2011): Lemma 10 requires that the matrix E have rank at most 2, but this is not necessarily true without assuming GLM structure. This is used to bound the determinant of the Jacobian, and corresponds to the first term of our bound in Theorem 4.3.2.

It is a similar story for bounding the log ratio / difference between the noise vector densities under neighboring datasets, corresponding to the second and third terms of our bound in Theorem 4.3.2. Let's also revisit this line from the proof of Lemma 17 of the Kifer et al. (2012) paper: "Note that Γ is independent of the noise vector." This is not true without assuming GLM structure! (In their proof, Γ is the difference between the noise vectors under neighboring distributions. From first-order conditions at the minimizer of the perturbed objective, we can see that $\Gamma = \nabla \ell(\theta^P)$, where θ^P is a function of the noise vector b.

In fact, to our knowledge, Iyengar et al. (2019) was the first work to acknowledge the GLM assumption on objective perturbation. But their privacy proof also fails to handle the dependence on the noise vector! In Theorems 4.3.2 and 4.3.1, we have included a careful analysis including a discussion on how the GLM assumption removes this dependence.

C.5.2 RDP bound for non-GLMs

In this section we generalize the RDP bound for objective perturbation to a general class of smooth convex losses.

Theorem C.5.1 (RDP bound for non-GLMS.). *Consider a loss function $\ell(\theta; z)$ such that $||\nabla \ell(\theta; z)||_2 \leq L$ and $\nabla^2 \ell(\theta; z) \prec \beta I_d$ for all $\theta \in \Theta$ and $z \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$. The objective perturbation mechanism which releases $\theta^P \sim \hat{\theta}^P(Z)$ satisfies (α, ϵ)-RDP for any $\alpha > 1$ with*

$$\epsilon \leq -d \log\left(1 - \frac{\beta}{\lambda}\right) + \frac{L}{2\sigma^2} + \frac{1}{\alpha - 1} \log \mathbb{E}_{Z \sim \chi_d}\left[e^{(\alpha-1)\frac{L}{\sigma}Z}\right],$$

where $Z \sim \chi_d$ if $Z = \sqrt{\sum_{i=1}^d X_i^2}$ and $X_i \sim \mathcal{N}(0, 1)$ for all $i \in [d]$.

Proof. Following the proof of Theorem 4.3.2, let

$$R(\theta^P; Z, Z') := \frac{\Pr\left[\hat{\theta}^P(Z) = \theta^P\right]}{\Pr\left[\hat{\theta}^P(Z') = \theta^P\right]}$$

be shorthand for the probability density ratio at output θ^P, given a fixed pair of neighboring datasets Z and Z'. Then we can state the following as a corollary to Theorem 6 of Redberg and Wang (2021).

Theorem C.5.2. *Given a dataset $Z \in \mathcal{Z}$ and a datapoint $z \in \mathcal{X} \times \mathcal{Y}$, construct neighboring dataset $Z' = Z \cup \{z\}$. Recall that $b(\theta^P; Z)$ is the bijection between the noise vector b and the output θ^P, satisfying $\theta^P = \arg\min \mathcal{L}(\theta; Z, b)$. Then for any dataset $Z \in \mathcal{Z}$, datapoint $z \in \mathcal{X} \times \mathcal{Y}$, and output $\theta^P \in \Theta$,*

$$\log R(\theta^P; Z, Z') = \left| -\log \prod_{j=1}^{d} \left(1 - \mu_j\right) + \frac{1}{2\sigma^2} ||\nabla \ell(\theta^P; z)||_2^2 + \frac{1}{\sigma^2} \nabla J(\theta^P; D)^T \nabla \ell(\theta^P; z) \right|,$$

where $\mu_j = \lambda_j u_j^T \left(\nabla b(\theta^P; D) \mp \sum_{k=1}^{j-1} \lambda_k u_k u_k^T\right)^{-1} u_j$ according to the eigendecomposition $\nabla^2 \ell(\theta^P; z) = \sum_{k=1}^{d} \lambda_k u_k u_k^T$.

To upper-bound $\log R(\theta^P; Z, Z')$, we first apply the triangle property so that we can bound the absolute value of each term individually.

We have assumed that $\nabla^2 \ell(\theta; z) \prec \beta I_d$ for all $\theta \in \Theta$ and $z \in \mathcal{Z}$. By Theorem 8 of Redberg and Wang (2021) and by applying this assumption, we have that

$$\left| -\log \prod_{j=1}^{d} (1 - \mu_j) \right| \leq -\sum_{j=1}^{d} \log\left(1 - \frac{\lambda_j}{\lambda}\right) \leq -d \log\left(1 - \frac{\beta}{\lambda}\right).$$

We now need to bound

$$\frac{1}{\alpha-1}\log\mathbb{E}\left[e^{(\alpha-1)\frac{1}{\sigma^2}\nabla J(\theta^P)^T\nabla\ell(\theta^P)}\right].$$

By the Cauchy-Schwarz inequality, we have for any $b\in\mathcal{R}^d$

$$-b^T\nabla\ell(\theta^P)\leq\left|-b^T\nabla\ell(\theta^P)\right|\leq L||b||_2.$$

Recall that $b\sim\mathcal{N}(0,\sigma^2I_d)$ and observe that by first-order conditions, $\nabla J(\theta^P;Z)=-b$. Using this change-of-variables we can show that

$$\frac{1}{\alpha-1}\log\mathbb{E}_{\theta^P\sim\hat{\theta}^P(Z)}\left[e^{(\alpha-1)\frac{1}{\sigma^2}\nabla J(\theta^P;Z)^T\nabla\ell(\theta^P;z)}\right]\leq\frac{1}{\alpha-1}\log\mathbb{E}_{Z\sim\chi_d}\left[e^{(\alpha-1)\frac{L}{\sigma}Z}\right].$$

□

C.6 RDP guarantee of Algorithm 7

In what follows, we will present a (corrected) privacy guarantee for Approximate Minima Perturbation (i.e., Algorithm 7 without gradient clipping). We will then demonstrate that the "clipped-gradient" function $\ell_C(\theta)$ not only bounds the per-example gradient norm by C, but also preserves other properties (i.e., β-smoothness and GLM structure) required for the privacy guarantees stated in Theorem 4.3.2.

C.6.1 Privacy Guarantee for Approximate Minima Perturbation

The proof of the privacy guarantee for Approximate Minima Perturbation (Iyengar et al., 2019), i.e. Algorithm 7 without gradient clipping, falls prey to the same trap as previous work on objective perturbation. In particular, we see that there is a mistake

in Lemma IV.1, with the assertion that "we get the statement of the lemma from the guarantees of the Gaussian mechanism." The Gaussian mechanism is inapplicable in Lemma IV.1 for similar reasons as discussed in Section C.5.

The proof of Theorem C.6.1 corrects this issue. We state it in terms of RDP, but it can also extend to approximate DP and other DP variants.

Algorithm 16 Approximate Minima Perturbation (Iyengar et al., 2019)

Input: dataset Z; noise levels $\sigma, \sigma_{\texttt{out}}$; β-smooth loss function $\ell(\cdot)$ with Lipschitz constant L; regularization strength λ; gradient norm threshold τ.

Sample $b \sim \mathcal{N}(0, \sigma^2 I_d)$.

Let $\mathcal{L}^P(\theta; Z) = \sum_{z \in Z} \ell(\theta; z) + \frac{\lambda}{2}||\theta||_2^2 + b^T\theta$.

Solve for $\tilde{\theta}$ such that $||\nabla \mathcal{L}_C^P(\tilde{\theta}; Z)||_2 \leq \tau$.

Output $\tilde{\theta}^P = \tilde{\theta} + \mathcal{N}(0, \sigma_{\texttt{out}}^2 I_d)$.

Theorem C.6.1 (RDP guarantees of Approximate Minima Perturbation). *Consider the Approximate Minima Perturbation algorithm which satisfies (α, ϵ)-RDP for any $\alpha > 1$ with*

$$\epsilon \leq -\log\left(1 - \frac{\beta}{\lambda}\right) + \frac{L^2}{2\sigma^2} + \frac{1}{\alpha - 1}\log \mathbb{E}_{X \sim \mathcal{N}\left(0, \frac{L^2}{\sigma^2}\right)}\left[e^{(\alpha-1)|X|}\right] + \frac{\left(\frac{2\tau}{\lambda}\right)^2 \alpha}{2\sigma_{\texttt{out}}^2}.$$

Proof. Sample $b \sim \mathcal{N}(0, \sigma^2 I_d)$ and let $\mathcal{L}^P(\theta; Z, b) := \sum_{z \in Z} \ell(\theta; z) + \frac{\lambda}{2}\|\theta\|_2^2 + b^T\theta$, i.e., the perturbed and regularized objective function used by ObjPert.

Let $\theta^P = \arg\min \mathcal{L}^P(\theta; Z, b)$. From Chaudhuri et al. (2011); Kifer et al. (2012) we know that there is a bijection $b(\theta^P; Z)$ from the output θ^P to the noise vector b.

Consider a blackbox algorithm $\theta^A(Z, b)$ which returns θ such that $||\nabla \mathcal{L}^P(\theta; Z, b)|| \leq \tau$. Define query

$$q(Z, \theta^P) = \theta^A(Z) - \theta^P.$$

where $\theta^A(Z)$ is an abbreviation for $\theta^A(Z, b(\theta^P; Z))$.

We assume that q can recover b from the input θ^P via the bijection $b(\theta^P; Z)$, and hence has access to the perturbed objective function $\mathcal{L}^P(\theta; Z, b)$.

Notice that since $\mathcal{L}^P$ is λ-strongly convex, by applying the Cauchy-Schwarz inequality and by Definition C.10.3 we see that for any θ_1, θ_2,

$$\left|\left|\nabla\mathcal{L}^P(\theta_1) - \nabla\mathcal{L}^P(\theta_2)\right|\right|_2 ||\theta_1 - \theta_2||_2 \geq \left(\nabla\mathcal{L}^P(\theta_1) - \nabla\mathcal{L}^P(\theta_2)\right)^T (\theta_1 - \theta_2) \geq \lambda ||\theta_1 - \theta_2||_2^2.$$

Algorithm $\theta^A(Z, b)$ guarantees that its output θ satisfies $||\nabla\mathcal{L}^P(\theta)||_2 \leq \tau$ and by first-order conditions on the perturbed objective function, $\nabla\mathcal{L}^P(\theta^P; Z, b) = 0$. It follows that for any dataset Z and θ^P,

$$\left|\left|\theta^A(Z) - \theta^P\right|\right|_2 \leq \frac{\tau}{\lambda},$$

Since the algorithm $\theta^A(Z, b)$ guarantees that $\|\theta^A - \theta^P\|_2 \leq \gamma/\lambda$, then conditioning on θ^P, $q(Z, \theta^P)$ has a global sensitivity bounded by $2\gamma/\lambda$ since

$$\begin{aligned}
\|q(Z, \theta^P) - q(Z', \theta^P)\| &\leq \|(\theta^A(Z) - \theta^P) - (\theta^A(Z') - \theta^P)\|_2 \\
&\leq \|\theta^A(Z) - \theta^P\| + \|\theta^A(Z') - \theta^P\|_2 \\
&\leq \frac{2\gamma}{\lambda}.
\end{aligned}$$

Now, the algorithm that first draws b then outputs $\theta^A(Z, b) + \mathcal{N}(0, \sigma^2 I_d)$ is equivalent to

- First run ObjPert that returns θ^P.
- Release $\hat{\Delta} = q(Z, \theta^P) + \mathcal{N}(0, \sigma^2 I_d)$.
- Return $\theta^P + \hat{\Delta}$.

This is adaptive composition of ObjPert with the Gaussian mechanism. The third step is post processing.

The privacy guarantee stated in Theorem C.6.1 is thus achieved by combining the results of Theorem 4.3.2 (RDP of ObjPert), Theorem C.2.1 (RDP of the Gaussian mechanism) with $\Delta_q = \frac{2\tau}{\lambda}$, and Lemma C.10.5 (adaptive composition for RDP mechanisms).

□

C.6.2 The "Clipped-Gradient" Function

The RDP guarantees of objective perturbation (stated in Theorem 4.3.2) require several assumptions on the loss function $\ell(\theta; Z)$. If we can demonstrate that these properties are satisfied by the "clipped-gradient" loss function $\ell_C(\theta; Z)$, then the rest of the proof of Theorem 4.4.1 (the privacy guarantee of Algorithm 7) will follow directly from that.

In particular, we need to show:

1. That $\ell_C(\theta; z)$ retains the convex GLM structure of the original function $\ell(\theta; z)$.
2. That $\ell_C(\theta; z)$ satisfies $||\nabla \ell_C(\theta; z)||_2 \leq C$ for any θ, z.
3. That $\ell_C(\theta; z)$ has the same β-smoothness parameter as the original function $\ell(\theta; z)$.
4. That even though $\ell_C(\theta; z)$ is not twice-differentiable everywhere, the privacy guarantees of objective perturbation (whose proof involves a Jacobian mapping) still hold.

We will begin by stating a result from Song et al. (2020).

Theorem C.6.2 (Song et al., 2020, Lemma 5.1). *Let $f : \mathbb{R} \to \mathbb{R}$ be any convex function and let $C \in \mathbb{R}_+$ be any positive value. For any non-zero $x \in \mathbb{R}^d$, define*

$$U_L = \left\{ u : g < -\frac{C}{||x||_2} \ \ \forall g \in \partial f(u) \right\},$$
$$U_H = \left\{ u : g > \frac{C}{||x||_2} \ \ \forall g \in \partial f(u) \right\}.$$

If U_L is non-empty, let $u_L = \sup U_L$; otherwise $u_L = -\infty$. If U_H is non-empty, let $u_H = \inf U_H$; otherwise $u_H = \infty$. For any non-zero $x \in \mathbb{R}^d$, let

$$f_C(u) = \begin{cases} -\frac{C}{||x||_2}(u - u_L), & \text{for } u \in (-\infty, u_L) \\ f(u; y), & \text{for } u \in (u_L, u_H) \\ \frac{C}{||x||_2}(u - u_H), & \text{for } u \in (u_H, \infty) \end{cases}$$

Define $u_x(\theta) = x^T\theta$. Then the following holds.

1. *f_C is convex.*
2. *Let $\ell(\theta; (x, y)) = f(u_x(\theta); y)$ for any $\theta, z = (x, y)$. Then we have*

$$\partial_\theta \ell_C(\theta; z) = \left\{ \min\left\{1, \frac{C}{||u_x(\theta)||_2)}\right\} \cdot u : u \in \partial_\theta \ell(\theta; z) \right\}.$$

The first two desired properties of $\ell_C(\theta; z)$, i.e. GLM structure and gradient norm bound C, follow directly from the above theorem. Next, we will prove the third property of β-smoothness.

Theorem C.6.3. *For a data point $z = (x, y)$, consider a function f such that $\ell(\theta; z) = f(x^T\theta; y)$. Suppose that $f(x^T\theta; y)$ satisfies β-smoothness. Then the "clipped-gradient" function $f_C(x^T\theta; y)$ defined in Lemma 5.1 of Song et al. (2020) also satisfies β-smoothness.*

Proof. Because f is β-smooth by assumption, we know that $f(u)$ satisfies β-smoothness for all $u \in (u_L, u_H)$. When $u \in (-\infty, u_L)$ or when $u \in (u_H, \infty)$, the function $f_C(u)$ is linear in u and thus is 0-smooth (hence satisfying β-smoothness). □

Lastly, the proof of objective perturbation (see, e.g., Theorem 9 of Chaudhuri et al. (2011)) requires that the loss function be twice-differentiable. Even though $\ell_C(\theta; z)$ is not twice-differentiable everywhere, the privacy guarantees of objective perturbation still hold. To show this, we can invoke Corollary 13 of Chaudhuri et al. (2011). This corollary assumes the Huber loss; for brevity, we will leave it as an exercise for the reader to verify that the proof also carries through for the "clipped-gradient" loss.

C.7 Computational Guarantee of Algorithm 7

In this section, we provide a *computational guarantee* to Algorithm 7 in terms of the number of gradient evaluations on individual loss functions to compute the approximate minimizer for achieving (up to a constant) the information-theoretical limit.

Let $f(\theta) := \sum_i \ell_i(\theta) + \frac{\lambda}{2}\|\theta\|^2 + b^T\theta$, i.e., the perturbed and regularized objective function used by ObjPert. Let θ^{**} be the output returned by the blackbox algorithm $\theta^A(\cdot)$ described in Section C.6. Note the deviation from the notation used in the previous section.

Iyengar et al. (2019) proposed a procedure that keeps checking the gradients in an iterative optimization algorithm and stops when the gradient is smaller than τ. This always ensures that $\|\nabla f(\theta^{**})\| \leq \tau$.

And using tools from the next section, it can be proven that it implies that $\|\theta^{**} - \theta^*\|_2 \leq \tau/\lambda$ as was previously stated. But how many iterations it takes for this to happen for specific algorithms was not explicitly considered.

C.7.1 Tools from convex optimization

We will need a few tools from convex optimization.

Firstly, under our assumption that ℓ_i is β-smooth, f is $n\beta + \lambda$-smooth and λ-strongly convex. Let $L := n\beta + \lambda$ as a shorthand.

By L-smoothness (gradient Lipschitzness), and the optimality of θ^*, we have that for any θ

$$\|\nabla f(\theta)\| = \|\nabla f(\theta) - \nabla f(\theta^*)\| \leq L\|\theta - \theta^*\|_2. \tag{C.7.1}$$

By λ-strong convexity, we get

$$f(\theta) - f^* \geq \frac{\lambda}{2}\|\theta - \theta^*\|^2 \geq \frac{\lambda}{2L^2}\|\nabla f(\theta)\|^2 \tag{C.7.2}$$

By strong convexity also implies that

$$\|\nabla f(\theta)\| \geq \lambda\|\theta - \theta^*\| \tag{C.7.3}$$

which is the quantity used to establish the global sensitivity of $q(D, \theta^*)$ as we talked about earlier.

(C.7.2) and (C.7.3) sandwich $\|\theta - \theta^*\|$ in between by

$$\frac{\|\nabla f(\theta)\|}{L} \leq \|\theta - \theta^*\| \leq \frac{\|\nabla f(\theta)\|}{\lambda}.$$

C.7.2 Computational bounds for Stopping at small gradient

(C.7.1) and (C.7.2) together provides bounds for $\|\nabla f(\theta)\|$ using either objective function or argument convergence (in square ℓ_2.)

$$\|\nabla f(\theta)\|^2 \leq \min\left\{\frac{2L^2}{\lambda}(f(\theta) - f^*), L^2\|\theta - \theta^*\|^2\right\}.$$

Standard convergence results are often parameterized in terms of either suboptimality $f(\theta) - f^*$ or argument $\|\theta - \theta^*\|^2$. In the following we instantiate specific convergence bounds for deriving computation guarantees.

Gradient Descent. If we run gradient descent with learning rate $1/L$ for T iterations from θ_0, then

$$\|\theta_T - \theta^*\|^2 \leq (1 - \frac{\lambda}{L})^T\|\theta_0 - \theta^*\|^2$$

which implies that

$$\|\nabla f(\theta_T)\|^2 \leq L^2(1 - \frac{\lambda}{L})^T\|\theta_0 - \theta^*\|^2$$

This happens deterministically (with no randomness, or failure probability).

One may ask why are we not running a fixed number of iterations and directly applying the bound to $\|\theta_T - \theta^*\|$ in order to control the ℓ_2 sensitivity. That works fine, except that we have an unconstrained problem and θ^* can be anywhere, thus there might not be a fixed parameter T to provide a required bound for all input θ^*. We also do not know where θ^* is during the actual execution of the algorithm and thus cannot compute $\|\theta_0 - \theta^*\|$ directly.

The "gradient-norm check" as a stopping condition from Iyengar et al. (2019) is nice because it always ensures DP for any θ^* (at a price of sometimes running for a bit longer).

To ensure $\|\nabla f(\theta)\| \leq \gamma$, the number of iterations

$$T = \frac{\log(\frac{L^2\|\theta_0-\theta^*\|^2}{\gamma^2})}{\log(1+\frac{\lambda}{L-\lambda})} \leq \frac{2(L-\lambda)}{\lambda}\log(\frac{L^2\|\theta_0-\theta^*\|^2}{\gamma^2}) = \frac{2n\beta}{\lambda}\log(\frac{(n\beta+\lambda)^2\|\theta_0-\theta^*\|^2}{\gamma^2})$$

Since each gradient computation requires n incremental gradient evaluation, under the regime that λ is independent of n, under the choice that $\lambda = 1/\epsilon$ independent to n from the standard calibration, the total number of is therefore $O(n^2 \log n)$ for achieving $\gamma \leq n^{-v}$ for any constant $v > 0$.

The quadratic runtime is not ideal, but it can be improved using accelerated gradient descent which gives a convergence bound of

$$f(\theta_T) - f^* \leq (1 - \sqrt{\frac{\lambda}{L}})^T \|\theta_0 - \theta^*\|^2.$$

This would imply a computational guarantee of $O(n^{1.5} \log n)$. The overall computation bound depends on $\|\theta^*\|$ which is random (due to objective perturbation). The dependence on $\|\theta^*\|$ is only logarithmic though.

Finite Sum and SAG. The result can be further improved if we uses stochastic gradient methods. However, the sublinear convergence of the standard SGD or its averaged version makes the application of the above conversion rules somewhat challenging.

By taking advantage of the finite sum structure of $f(\theta)$ one can obtain faster convergence.

First of all, the finite sum structure says that $f(\theta) = \sum_{i=1}^n f_i(\theta)$. In our case, we can split the regularization and linear perturbation to the n data points, i.e.,

$$f_i(\theta) = \ell_i(\theta) + \frac{\lambda}{2n}\|\theta\|^2 + \frac{b^T\theta}{n}.$$

Check that it satisfies $\beta + \lambda/n$ smoothness.

There is a long list of methods that satisfy the faster convergence for finite sum problems, e.g., SAG, SVRG, SAGA, SARAH and so on (see, e.g., Nguyen et al., 2022, for a recent survey). Specifically, Stochastic Averaged Gradient (Schmidt et al., 2017) (and similarly others with slightly different parameters) satisfies

$$\mathbb{E}\left[f(\theta^T) - f^*\right] \leq (1 - \min\{\frac{\lambda}{16L}, \frac{1}{8n}\})^T \cdot (\frac{3n}{2}(f(\theta_0) - f^*) + 4L\|\theta_0 - \theta^*\|^2).$$

Therefore, by (C.7.2), we have

$$\mathbb{E}\left[\|\nabla f(\theta_T)\|^2\right] \leq (1 - \min\{\frac{\lambda}{16L}, \frac{1}{8n}\})^T \cdot \frac{L^2}{\lambda}(\frac{3n}{2}(f(\theta_0) - f^*) + 4L\|\theta_0 - \theta^*\|^2).$$

Note that each iteration costs just one incremental gradient evaluation, so to ensure $\mathbb{E}[\|\nabla f(\theta_T)\|^2] \leq \gamma^2$, the computational complexity is on the order of

$$\max\{n, \frac{L}{\lambda}\} \log\left(\frac{nL \max\{f(\theta_0) - f^*, \|\theta_0 - \theta^*\|^2\}}{\lambda\gamma}\right)$$

This is $O(n \log n)$ runtime to any $\gamma = n^{-s}$ for a constant $s > 0$.

On the other hand, the main difference from the gradient descent result is that we only get convergence in expectation. By Markov's inequality

$$\mathbb{P}\left[\|\nabla f(\theta_T)\|^2 > \gamma^2\right] \leq \frac{(1 - \min\{\frac{\lambda}{16L}, \frac{1}{8n}\})^T \cdot \frac{L^2}{\lambda}(\frac{3n}{2}(f(\theta_0) - f^*) + 4L\|\theta_0 - \theta^*\|^2)}{\gamma^2} := \delta,$$

which implies high probability convergence naturally.

Theorem C.7.1. *Assume $\lambda \geq \beta$. The algorithm that runs SAG and checks the stopping condition $\|\nabla f(\theta_T)\| \leq \gamma$ after every n iteration will terminate with probability at least*

$1-\delta$ *in less than*

$$C\max\{n, \frac{n\beta}{\lambda}\}\log\left(\frac{n\beta\max\{\|\theta_0-\theta^*\|, (f(\theta_0)-f^*)\}}{\gamma\delta}\right)$$

incremental gradient evaluations, where C is a universal constant.

How to set γ to achieve information-theoretic limit? The lower bounds for convex and smooth losses in differentially private ERM are well-known (Bassily et al., 2014) and it is known that among GLMs, θ^* from ObjPert achieves the lower bound with appropriate choices of λ, σ. Notably, $\lambda \asymp d/\epsilon$ for achieving an (ϵ, δ)-DP.

$$\mathbb{E}[\sum_i \ell_i(\theta^*)] - \min_\theta \sum_i \ell_i(\theta) \leq \text{MinimaxExcessEmpirialRisk}$$

Let $\hat{\theta} = \theta_T + N(0, \frac{\gamma^2}{2\lambda^2\rho}I)$ be the final output.

By the nG Lipschitzness of $\sum_i \ell_i$, with high probability over the Gaussian mechanism, we have that

$$\mathbb{E}[\sum_i \ell_i(\hat{\theta})] - \sum_i \ell_i(\theta^*) \leq nG(\|\theta_T-\theta^*\| + \|\hat{\theta}-\theta_T\|) \leq nG\gamma(1+\frac{\sqrt{\frac{d\log d}{\rho}}}{\lambda})$$

where ρ is the zCDP parameter for the Gausssian mechanism chosen to match the large α part of the ObjectivePerturbation's RDP bound, which increases the overall RDP by $\alpha\rho$.

Thus, it suffices to take $\gamma = \text{MinimaxExcessEmpirialRisk}/(nG(1+\frac{\sqrt{d\log d/\rho}}{\lambda}))$.

To conclude, the above results imply that the computationally efficient objective perturbation achieves the optimal rate under the same RDP guarantee with an algorithm that terminates in $O(n\log n)$ time with high probability.

C.8 Excess Empirical Risk of Algorithm 7

Our goal in this section is to find a bound on the excess empirical risk:

$$\mathbb{E}\left[\mathcal{L}(\tilde{\theta}^P; Z)\right] - \mathcal{L}(\theta^*; Z).$$

Theorem C.8.1. *Let $\tilde{\theta}^P$ be the output of Algorithm 7 and $\theta^* = \arg\min \mathcal{L}(\theta)$ the minimizer of the loss function $\mathcal{L}(\theta) = \sum_{i=1}^n \ell(\theta; z_i)$. Denote $||\mathcal{X}||$ as the diameter of the set $\mathcal{X}$. We have*

$$\mathbb{E}\left[\mathcal{L}(\tilde{\theta}^P; Z)\right] - \mathcal{L}(\theta^*; Z) \leq nL\left(\frac{\tau}{\lambda} + \sigma_{out}\sqrt{d}\right) + \frac{d\sigma^2}{2\lambda} + \frac{\lambda}{2}||\theta^*||_2^2.$$

Proof. Following the proof of Theorem 2 from Iyengar et al. (2019) (itself adapted from Kifer et al. (2012)), we write

$$\mathcal{L}(\tilde{\theta}^P) - \mathcal{L}(\theta^*) = \left(\mathcal{L}(\tilde{\theta}^P) - \mathcal{L}(\theta^P)\right) + \left(\mathcal{L}(\theta^P) - \mathcal{L}(\theta^*)\right).$$

By the λ-strong convexity of $\mathcal{L}^P$, for any $\tilde{\theta}, \theta^*$ we have

$$\left(\nabla\mathcal{L}^P(\tilde{\theta}) - \nabla\mathcal{L}^P(\theta^P)\right)^T\left(\tilde{\theta} - \theta^P\right) \geq \lambda||\tilde{\theta} - \theta^P||_2^2.$$

By first-order conditions, $\nabla\mathcal{L}^P(\theta^P) = 0$. Applying the Cauchy-Schwarz inequality along with our stopping criteria on the gradient norm, we then have

$$||\tilde{\theta} - \theta^P||_2 \leq \frac{1}{\lambda}||\nabla\mathcal{L}^P(\tilde{\theta})||_2 \leq \frac{\tau}{\lambda}.$$

Let $\xi \sim \mathcal{N}(0, \sigma_{out}^2 I_d)$. Because $\mathcal{L}$ is nL-Lipschitz continuous, we have

$$\begin{aligned}\left(\mathcal{L}(\tilde{\theta}^P) - \mathcal{L}(\theta^P)\right) &\leq nL||\tilde{\theta}^P - \theta^P||_2 \\ &= nL||\tilde{\theta} + \xi - \theta^P||_2 \\ &\leq nL\left(||\tilde{\theta} - \theta^P||_2 + ||\xi||_2\right) \\ &\leq nL\left(\frac{\tau}{\lambda} + ||\xi||_2\right).\end{aligned}$$

By Lemma C.10.11,

$$\mathbb{E}\left[nL\left(\frac{\tau}{\lambda} + ||\xi||_2\right)\right] \leq nL\left(\frac{\tau}{\lambda} + \sigma_{out}\sqrt{d}\right).$$

To bound the expectation of $\mathcal{L}(\theta^P) - \mathcal{L}(\theta^*)$, we can write

$$\mathcal{L}(\theta^P) - \mathcal{L}(\theta^*) = \left(\mathcal{L}(\theta^P) - \mathcal{L}_\lambda(\theta^P)\right) + \left(\mathcal{L}_\lambda(\theta^P) - \mathcal{L}_\lambda(\theta_\lambda^*)\right) + \left(\mathcal{L}_\lambda(\theta_\lambda^*) - \mathcal{L}(\theta^*)\right).$$

We can write $\mathcal{L}_\lambda(\theta_\lambda^*) - \mathcal{L}(\theta^*) = (\mathcal{L}_\lambda(\theta_\lambda^*) - \mathcal{L}_\lambda(\theta^*)) + (\mathcal{L}_\lambda(\theta^*) - \mathcal{L}(\theta^*))$ and observe that

$$\begin{aligned}&\mathcal{L}(\theta^P) - \mathcal{L}_\lambda(\theta^P) = -\frac{\lambda}{2}||\theta^P||_2^2 \leq 0, \\ &\mathcal{L}_\lambda(\theta^*) - \mathcal{L}(\theta^*) = \frac{\lambda}{2}||\theta^*||_2^2, \text{ and} \\ &\mathcal{L}_\lambda(\theta_\lambda^*) - \mathcal{L}_\lambda(\theta^*) \leq 0.\end{aligned}$$

The last inequality follows from the optimality condition $\theta_\lambda^* = \underset{\theta \in \mathbb{R}^d}{\arg\min}\, \mathcal{L}_\lambda(\theta)$. We then have

$$\mathcal{L}(\theta^P) - \mathcal{L}(\theta^*) \leq \mathcal{L}_\lambda(\theta^P) - \mathcal{L}_\lambda(\theta_\lambda^*) + \frac{\lambda}{2}||\theta^*||_2^2.$$

By Taylor's Theorem, for some $\theta' \in \left[\theta^P, \theta^*_\lambda\right]$ we can write

$$\mathcal{L}_\lambda(\theta^*_\lambda) - \mathcal{L}_\lambda(\theta^P) = \nabla\mathcal{L}_\lambda\left(\theta^P\right)^T\left(\theta^*_\lambda - \theta^P\right) + \tfrac{1}{2}||\theta^*_\lambda - \theta^P||^2_{\nabla^2\mathcal{L}_\lambda(\theta')}, \tag{C.8.1}$$

with norm $||\cdot||_A = \sqrt{(\cdot)^T A(\cdot)}$.

Then by the optimality condition $\nabla\mathcal{L}_\lambda(\theta^P) + b = 0$, we see that

$$\begin{aligned}\mathcal{L}_\lambda(\theta^*_\lambda) - \mathcal{L}_\lambda(\theta^P) + b^T\left(\theta^*_\lambda - \theta^P\right) &= \left(\nabla\mathcal{L}_\lambda(\theta^P) + b\right)^T\left(\theta^*_\lambda - \theta^P\right) + \tfrac{1}{2}||\theta^*_\lambda - \theta^P||^2_{\nabla^2\mathcal{L}_\lambda(\theta')} \\ &= \tfrac{1}{2}||\theta^*_\lambda - \theta^P||^2_{\nabla^2\mathcal{L}_\lambda(\theta')}\end{aligned}$$

After rearranging terms, we can use Lemma C.10.10 and then complete the square to see that

$$\begin{aligned}\mathcal{L}_\lambda(\theta^P) - \mathcal{L}_\lambda(\theta^*_\lambda) &= -\tfrac{1}{2}||\theta^*_\lambda - \theta^P||^2_{\mathcal{L}_\lambda(\theta')} - b^T\left(\theta^P - \theta^*_\lambda\right) \\ &\leq -\tfrac{\lambda}{2}||\theta^*_\lambda - \theta^P||_2^2 - b^T\left(\theta^P - \theta^*_\lambda\right) \\ &= -\left|\left|\sqrt{\tfrac{\lambda}{2}}\left(\theta^*_\lambda - \theta^P\right) + \sqrt{\tfrac{1}{2\lambda}}b\right|\right|_2^2 + \frac{||b||_2^2}{2\lambda} \\ &\leq \frac{||b||_2^2}{2\lambda}.\end{aligned}$$

By Lemma C.10.11,

$$\mathbb{E}\left[\mathcal{L}_\lambda(\theta^P)\right] - \mathcal{L}_\lambda(\theta^*_\lambda) \leq \frac{d\sigma^2}{2\lambda}$$

Putting together the pieces then completes the proof.

C.8.1 Optimal rates

Choose $\sigma \asymp \frac{L\sqrt{d\log(1/\delta)}}{\epsilon}$ and $\lambda = \frac{dL\sqrt{\log(1/\delta)}}{\epsilon||\theta^*||_2}$. If $\tau \approx 0$ and $\sigma_{out} \approx 0$, then

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}(\tilde{\theta}^P; Z)\right] - \mathcal{L}(\theta^*; Z) &\leq nL\left(\frac{\tau}{\lambda} + \sigma_{out}\sqrt{d}\right) + \frac{d\sigma^2}{2\lambda} + \frac{\lambda}{2}||\theta^*||_2^2 \\
&\asymp \frac{d^2L^2\log(1/\delta)}{\epsilon^2} \cdot \frac{\epsilon||\theta^*||_2}{2dL\log(1/\delta)} + \frac{dL\sqrt{\log(1/\delta)}||\theta^*||_2^2}{2\epsilon||\theta^*||_2} \\
&\asymp \frac{dL||\theta^*||_2}{\epsilon} + \frac{dL\sqrt{\log(1/\delta)}||\theta^*||_2}{\epsilon} \\
&\asymp \frac{dL||\theta^*||_2\sqrt{\log(1/\delta)}}{\epsilon}.
\end{aligned}$$

The optimal choice of τ is discussed in Section C.7. □

C.8.2 Generalized Linear Model

With some additional assumptions and restrictions, we can get a tighter bound on $\mathbb{E}\left[\mathcal{L}_\lambda(\theta^P)\right] - \mathcal{L}_\lambda(\theta^*_\lambda)$.

We will assume GLM structure on $\ell(\cdot)$, i.e. $\ell(\theta; z) = f(x^T\theta; y)$. We will further assume boundedness: $c \leq f(x^T\theta; y) \leq C$ for some universal constants $c, C \in \mathbb{R}$. Applying Taylor's Theorem (in an argument similar to Equation C.3.8), we can show that for some $\theta'' \in \left[\theta^P, \theta^*_\lambda\right]$

$$\theta^P - \theta^*_\lambda = \nabla^2\mathcal{L}_\lambda(\theta'')^{-1}b.$$

Note that by the GLM assumption, the eigendecomposition of $\nabla^2\mathcal{L}^*_\lambda(\theta)$ can be written as $X^T\Lambda(\theta)X$. Then plugging in from Equation C.8.1 and using the boundedness assumption

on the loss f,

$$\begin{aligned}
\mathcal{L}_\lambda(\theta^P) - \mathcal{L}_\lambda(\theta^*_\lambda) &= \frac{1}{2}||\theta^P - \theta^*_\lambda||^2_{\nabla^2\mathcal{L}_\lambda(\theta')} \\
&= b^T \left(\nabla^2\mathcal{L}_\lambda(\theta'')\right)^{-1} \nabla^2\mathcal{L}_\lambda(\theta') \left(\nabla^2\mathcal{L}_\lambda(\theta'')\right)^{-1} b \\
&= b^T \left(X^T\Lambda(\theta'')X + \lambda I_d\right)^{-1} \left(X^T\Lambda(\theta'')X + \lambda I_d\right) \left(X^T\Lambda(\theta')X + \lambda I_d\right)^{-1} b \\
&\leq b^T c^{-1} \left(X^TX + \lambda I_d\right)^{-1} C \left(X^TX + \lambda I_d\right) c^{-1} \left(X^TX + \lambda I_d\right)^{-1} b \\
&\leq \frac{C}{c^2}||b||^2_{(X^TX+\lambda I_d)^{-1}}
\end{aligned}$$

Then in expectation,

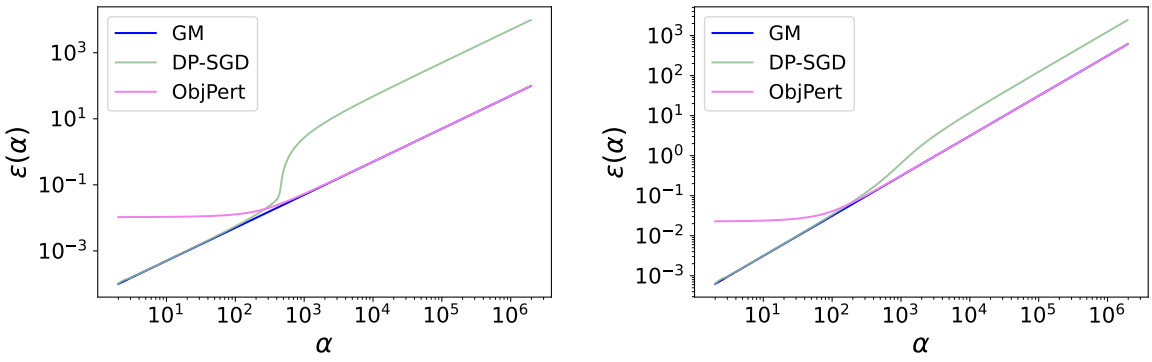
$$\mathbb{E}\left[\mathcal{L}_\lambda(\theta^P\right] - \mathcal{L}_\lambda(\theta^*_\lambda) \leq \frac{C\sigma^2}{c^2}\mathrm{tr}\left(\left(X^TX + \lambda I_d\right)^{-1}\right).$$

C.9 Bridging the Gap between Objective Perturbation and DP-SGD

C.9.1 RDP of Objective Perturbation vs DP-SGD

DP-SGD (for example, n^2 rounds of sampled Gaussian mechanism with Poisson sampling probability $1/n$) is known to experience a phase transition in its RDP curve: for smaller α, amplification by sampling is effective, and the RDP of a DP-SGD mechanism behaves like a Gaussian mechanism with $\epsilon(\alpha) = O(\frac{\alpha}{2\sigma^2})$, then it leaps up at a certain α and begins converging to $\epsilon(\alpha) = O(\frac{n^2\alpha}{2\sigma^2})$ that does not benefit from sampling at all (Wang et al., 2019; Bun et al., 2018). In contrast, the RDP curve for objective perturbation defined by Theorem 4.3.2 converges to the RDP curve of the Gaussian mechanism $\epsilon(\alpha) = O(\frac{\alpha}{2\sigma^2})$ *after* a certain point. Whereas DP-SGD offers stronger privacy parameters for small α, objective perturbation is stronger for large α, which offers stronger privacy protection for

lower-probability events (see, e.g., Mironov, 2017, Proposition 10).



(a) $\sigma_{\texttt{ObjPert}} = \sigma_{\texttt{GM}} = 100; p = 0.1, \sigma_{\texttt{DP-SGD}} = 10.$ (b) $\sigma_{\texttt{ObjPert}} = \sigma_{\texttt{GM}} = 40; p = 0.5, \sigma_{\texttt{DP-SGD}} = 20.$

Figure C.3: RDP curves of objective perturbation and DP-SGD.

C.9.2 A Spectrum of DP Learning Algorithms

We can also connect Algorithm 7 to *differentially private follow-the-regularized-leader* (DP-FTRL) (Kairouz et al., 2021), which uses a tree-based aggregation algorithm to privately release the gradients of the loss function as a prefix sum. This approach provides a competitive privacy/utility tradeoff without relying on privacy amplification or shuffling, which is often not possible in distributed settings. DP-FTRL differs from DP-SGD by adding *correlated* rather than *independent* noise at each iteration; Algorithm 7, in contrast, differs from both by adding *identical* noise at each iteration.

C.10 Technical Lemmas & Definitions

C.10.1 Convex Optimization

We will give a short review of relevant concepts from convex optimization.

Definition C.10.1 (ℓ_2-Lipschitz continuity). A function $f : \Theta \to \mathbb{R}$ is L-Lipschitz w.r.t. the ℓ_2-norm over $\Theta \subseteq \mathbb{R}^d$ if for all $\theta_1, \theta_2 \in \Theta$, the following holds: $|f(\theta_1) - f(\theta_2)| \leq L||\theta_1 - \theta_2||_2$.

Definition C.10.2 (β-smoothness). A differentiable function $f : \Theta \to \mathbb{R}$ is β-smooth over $\Theta \subseteq \mathbb{R}^d$ if its gradient ∇f is β-Lipschitz, i.e. $|\nabla f(\theta_1) - \nabla f(\theta_2)| \leq \beta||\theta_1 - \theta_2||_2$ for all $\theta_1, \theta_2 \in \Theta$.

Definition C.10.3 (Strong convexity). A differentiable function $f : \Theta \to \mathbb{R}$ is λ-strongly convex over $\Theta \subseteq \mathbb{R}^d$ if for all $\theta_1, \theta_2 \in \Theta$: $f(\theta_1) \geq f(\theta_2) + \nabla f(\theta_2)^T(\theta_1 - \theta_2) + \frac{\lambda}{2}||\theta_1 - \theta_2||_2^2$.

C.10.2 Differential Privacy

Definition C.10.4 (Privacy loss random variable). Let $\Pr[\mathcal{M}(Z) = \theta]$ denote the probability density of the random variable $\mathcal{M}(Z)$ at output θ. For a fixed pair of neighboring datasets Z and Z', the privacy loss random variable (PLRV) of mechanism $\mathcal{M} : \mathcal{Z} \to \Theta$ is defined as

$$\epsilon_{Z,Z'}(\theta) = \log \frac{\Pr[\mathcal{M}(Z) = \theta]}{\Pr[\mathcal{M}(Z') = \theta]},$$

for the random variable $\theta \sim \mathcal{M}(Z)$.

Lemma C.10.5 (Adaptive composition (RDP) (Mironov, 2017)). *Let $\mathcal{M}_1 : \mathcal{Z} \to \mathcal{R}_1$ be (α, ϵ_1)-RDP and $\mathcal{M}_2 : \mathcal{R}_1 \times \mathcal{Z} \to \mathcal{R}_2$ be (α, ϵ_2)-RDP. Then the mechanism $\mathcal{M} = (m_1, m_2)$, where $m_1 \sim \mathcal{M}_1(Z)$ and $m_2 \sim \mathcal{M}_2(Z, m_1)$, satisfies $(\alpha, \epsilon_1 + \epsilon_2)$-RDP*

Lemma C.10.6 (Change of coordinates). *Consider the map $g : \mathcal{X} \to \mathcal{Y}$ between $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}^d$ that transforms $y = g(x)$. Then*

$$\partial y = \left| \det \frac{\partial y}{\partial x} \right| \partial x,$$

where $\left|\det \frac{\partial y}{\partial x}\right|$ *is the absolute value of the determinant of the Jacobian of the map* g*:*

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \partial y_1/\partial x_1 & \dots & \partial y_1/\partial x_d \\ \vdots & \ddots & \vdots \\ \partial y_d/\partial x_1 & \dots & \partial y_d/\partial x_d \end{bmatrix}.$$

Lemma C.10.7 (Change of variables for probability density functions.). *Let* g *be a strictly monotonic function. Then for* $y = g(x)$*,*

$$p_X(x) = \left|\det \frac{\partial y}{\partial x}\right| p_Y(y).$$

Lemma C.10.8. *Let* A *be an invertible matrix. Then* $\det A^{-1} = \frac{1}{\det A}$.

Lemma C.10.9 (Maximum Rayleigh quotient). *For any symmetric matrix* $A \in \mathbb{R}^{d\times d}$*,*

$$\max_{v\in\mathbb{R}^d} \frac{v^T A v}{v^T v} = \lambda_{max},$$

where λ_{max} *is the largest eigenvalue of* A*.*

Lemma C.10.10 (Quadratic form inequalities). *Let* $A \in \mathcal{R}^{d\times d}$ *be a symmetric positive-definite matrix with smallest eigenvalue* $\lambda_{min}(A)$ *and largest eigenvalue* $\lambda_{max}(A)$*. Then for any vector* $x \in \mathcal{R}^d$*,*

$$\lambda_{min}(A)||x||_2^2 \leq x^T A x \leq \lambda_{max}(A)||x||_2^2.$$

Lemma C.10.11 (Bound on the expected norm of multivariate Gaussian with mean 0.).

Let $x = \mathcal{N}(0, \sigma^2 I_d)$. *Then*

$$\mathbb{E}\left[\, ||x||_2 \,\right] \leq \sigma\sqrt{d}.$$

Lemma C.10.12 (Gaussian MGF). *Let* $X \sim \mathcal{N}(\mu, \sigma^2)$ *for* $\mu \in \mathbb{R}, \sigma \in \mathbb{R}_{>0}$. *The moment generating function of order* t *is then* $MGF_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$. *So for any* $t \in \mathbb{R}_{\geq 0}$ *and* $\sigma_1 < \sigma_2$,

$$MGF_{X_1}(t) \leq MGF_{X_2}(t),$$

where $X_1 \sim \mathcal{N}(\mu, \sigma_1^2)$ *and* $X_2 \sim \mathcal{N}(\mu, \sigma_2^2)$.

Definition C.10.13 (Holder's Inequality). Let X, Y be random variables satisfying $\mathbb{E}\left[|X|^p\right] < \infty$, $\mathbb{E}\left[|X|^q\right] < \infty$ for $p > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$\mathbb{E}\left[|XY|\right] \leq \mathbb{E}\left[|X|^p\right]^{\frac{1}{p}} \mathbb{E}\left[|X|^q\right]^{\frac{1}{q}}.$$

Definition C.10.14 (Rényi Divergence). Let P, Q be distributions with probability density functions $P(x), Q(x)$. The Rényi divergence of order $\alpha > 1$ between P and Q is given by

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim Q}\left[\left(\frac{P(x)}{Q(x)}\right)^\alpha\right] = \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim P}\left[\left(\frac{P(x)}{Q(x)}\right)^{\alpha - 1}\right].$$

Lemma C.10.15 (MGF inequality for identical distributions). *Let* $x, y, z \sim \mathcal{N}(0, \sigma^2)$. *Then*

$$\mathbb{E}_{x,y}\left[e^{t|z_1|\,|z_2|}\right] \leq \mathbb{E}_z\left[e^{tz^2}\right].$$

Proof. Observe that for any $a, b \in \mathbb{R}$, it holds that $2ab \leq a^2 + b^2$. Applying this along with the Cauchy-Schwarz inequality,

$$\begin{aligned}
\mathbb{E}_{x,y}\left[e^{t|x|\,|y|}\right] &\leq \mathbb{E}\left[e^{\frac{t}{2}x^2}e^{\frac{t}{2}y^2}\right] \\
&\leq \sqrt{\mathbb{E}_x\left[e^{\frac{t}{2}x^2}\right]\mathbb{E}_y\left[e^{\frac{t}{2}y^2}\right]} \\
&= \mathbb{E}_z\left[e^{tz^2}\right].
\end{aligned}$$

□

Bibliography

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

Naman Agarwal, Satyen Kale, Karan Singh, and Abhradeep Thakurta. Differentially private and lazy online convex optimization. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4599–4632. PMLR, 2023.

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pages 394–403. PMLR, 2018.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Advances in Neural Information Processing Systems*, pages 6277–6287, 2018.

Borja Balle, Gilles Barthe, Marco Gaboardi, Justin Hsu, and Tetsuya Sato. Hypothesis testing interpretations and renyi differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pages 2496–2506. PMLR, 2020.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

Mark Bun, Cynthia Dwork, Guy N Rothblum, and Thomas Steinke. Composable and versatile privacy via truncated cdp. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 74–86, 2018.

Clément L Canonne, Gautam Kamath, and Thomas Steinke. The discrete gaussian for differential privacy. *Advances in Neural Information Processing Systems*, 33:15676–15688, 2020.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Marco Chiani. Distribution of the largest eigenvalue for real wishart and gaussian random matrices and a simple approximation for the tracy–widom distribution. *Journal of Multivariate Analysis*, 129:69–81, 2014.

Chris Decarolis, Mukul Ram, Seyed Esmaeili, Yu-Xiang Wang, and Furong Huang. An end-to-end differentially private latent dirichlet allocation using a spectral algorithm. In *International Conference on Machine Learning*, pages 2421–2431. PMLR, 2020.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380, 2009.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014a.

Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 11–20, 2014b.

Hamid Ebadi, David Sands, and Gerardo Schneider. Differential privacy: Now it's getting personal. *ACM Sigplan Notices*, 50(1):69–81, 2015.

Vitaly Feldman and Tijana Zrnic. Individual privacy accounting via a renyi filter. *arXiv preprint arXiv:2008.11193*, 2020.

Arpita Ghosh and Aaron Roth. Selling privacy at auction. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 199–208, 2011.

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. In *Advances in Neural Information Processing Systems*, 2021.

Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. Private convex optimization via exponential mechanism. *arXiv preprint arXiv:2203.00263*, 2022.

Justin Hsu, Zhiyi Huang, Aaron Roth, Tim Roughgarden, and Zhiwei Steven Wu. Private matchings and allocations. *SIAM Journal on Computing*, 45(6):1953–1984, 2016.

Roger Iyengar, Joseph P Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 299–316. IEEE, 2019.

Peter Kairouz, Brendan McMahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. Practical and private (deep) learning without sampling or shuffling. In *International Conference on Machine Learning*, pages 5213–5225. PMLR, 2021.

Shiva Prasad Kasiviswanathan, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Analyzing graphs with node differential privacy. In *Theory of Cryptography Conference*, pages 457–476. Springer, 2013.

Michael Kearns, Mallesh Pai, Aaron Roth, and Jonathan Ullman. Mechanism design in large games: Incentives and privacy. In *Conference on Innovations in theoretical computer science (ITCS-14)*, pages 403–410, 2014.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1. JMLR Workshop and Conference Proceedings, 2012.

Antti Koskela, Joonas Jälkö, and Antti Honkela. Computing tight differential privacy guarantees using FFT. In *International Conference on Artificial Intelligence and Statistics*, pages 2560–2569. PMLR, 2020.

Antti Koskela, Joonas Jälkö, Lukas Prediger, and Antti Honkela. Tight differential privacy for discrete-valued mechanisms and for the subsampled gaussian mechanism using FFT. In *International Conference on Artificial Intelligence and Statistics*, pages 3358–3366. PMLR, 2021.

Janardhan Kulkarni, Yin Tat Lee, and Daogao Liu. Private non-smooth erm and sco in subquadratic steps. *Advances in Neural Information Processing Systems*, 34:4053–4064, 2021.

Katrina Ligett, Seth Neel, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. Accuracy first: Selecting a differential privacy level for accuracy-constrained erm. *Advances in Neural Information Processing Systems*, 2017:2567–2577, 2017.

Jingcheng Liu and Kunal Talwar. Private selection from private candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 298–309, 2019.

Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. *arXiv preprint arXiv:2111.06578*, 2021.

Ziqi Liu, Yu-Xiang Wang, and Alexander Smola. Fast differentially private matrix factorization. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 171–178, 2015.

Steve Lohr. A $1 million research bargain for netflix, and maybe a model for others. *The New York Times*, 22, 2009.

Ryan McKenna, Hristo Paskov, and Kunal Talwar. A practitioners guide to differentially private convex optimization. 2021.

Kentaro Minami, HItomi Arai, Issei Sato, and Hiroshi Nakagawa. Differential privacy without sensitivity. *Advances in Neural Information Processing Systems*, 29, 2016.

Ilya Mironov. Rényi differential privacy. In *Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled Gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

Shubhankar Mohapatra, Sajin Sasy, Xi He, Gautam Kamath, and Om Thakkar. The role of adaptive optimizers for honest private hyperparameter selection. In *Proceedings of the aaai conference on artificial intelligence*, volume 36, pages 7806–7813, 2022.

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

Seth Neel, Aaron Roth, Giuseppe Vietri, and Steven Wu. Oracle efficient private non-convex optimization. In *International Conference on Machine Learning*, pages 7243–7252. PMLR, 2020.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

Lam M Nguyen, Marten van Dijk, Dzung T Phan, Phuong Ha Nguyen, Tsui-Wei Weng, and Jayant R Kalagnanam. Finite-sum smooth optimization with sarah. *Computational Optimization and Applications*, 82(3):561–593, 2022.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84, 2007.

Nicolas Papernot and Thomas Steinke. Hyperparameter tuning with renyi differential privacy. *arXiv preprint arXiv:2110.03620*, 2021.

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations (ICLR-17)*, 2017.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. 2018a.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. In *International Conference on Learning Representations (ICLR-18)*, 2018b.

Rachel Redberg and Yu-Xiang Wang. Privately publishable per-instance privacy. *Advances in Neural Information Processing Systems*, 34, 2021.

Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602. World Scientific, 2010.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112, 2017.

Ryan Singel. Netflix cancels recommendation contest after privacy lawsuit. *Wired Magazine*, 2010.

David M Sommer, Sebastian Meiser, and Esfandiar Mohammadi. Privacy loss classes: The central limit theorem in differential privacy. *Proceedings on Privacy Enhancing Technologies*, 2019(2):245–269, 2019.

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 245–248. IEEE, 2013.

Shuang Song, Thomas Steinke, Om Thakkar, and Abhradeep Thakurta. Evading curse of dimensionality in unconstrained private glms via private gradient descent. *arXiv preprint arXiv:2006.06783*, 2020.

Jordi Soria-Comas, Josep Domingo-Ferrer, David Sánchez, and David Megías. Individual differential privacy: A utility-preserving formulation of differential privacy guarantees. *IEEE Transactions on Information Forensics and Security*, 12(6):1418–1429, 2017.

Gilbert W Stewart. Matrix perturbation theory. 1990.

Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. Privacy loss in apple's implementation of differential privacy on macos 10.12. *arXiv preprint arXiv:1709.02753*, 2017.

Abhradeep Guha Thakurta and Adam Smith. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory*, pages 819–850. PMLR, 2013.

Clive Thompson. If you liked this, you're sure to love that. *The New York Times*, 21, 2008.

Aleksei Triastcyn and Boi Faltings. Bayesian differential privacy for machine learning. In *International Conference on Machine Learning*, pages 9583–9592. PMLR, 2020.

Salil Vadhan. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer, 2017.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30, 2017.

Jiachen T Wang, Saeed Mahloujifar, Shouda Wang, Ruoxi Jia, and Prateek Mittal. Renyi differential privacy of propose-test-release and applications to private and robust machine learning. *arXiv preprint arXiv:2209.07716*, 2022.

Yu-Xiang Wang. Per-instance differential privacy and the adaptivity of posterior sampling in linear and ridge regression. *arXiv preprint arXiv:1707.07708*, pages 48–71, 2017.

Yu-Xiang Wang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. *arXiv preprint arXiv:1803.02596*, 2018.

Yu-Xiang Wang. Per-instance differential privacy. *Journal of Privacy and Confidentiality*, 9(1), 2019.

Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning*, pages 2493–2502. PMLR, 2015.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1226–1235. PMLR, 2019.

Bo-Cheng Wei, Yue-Qing Hu, and Wing-Kam Fung. Generalized leverage and its applications. *Scandinavian Journal of statistics*, 25(1):25–37, 1998.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.

Da Yu, Huishuai Zhang, Wei Chen, Tie-Yan Liu, and Jian Yin. Gradient perturbation is underrated for differentially private convex optimization. *arXiv preprint arXiv:1911.11363*, 2019.

Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Large scale private learning via low-rank reparametrization. In *International Conference on Machine Learning (ICML-21)*, volume 139, pages 12208–12218. PMLR, 2021.

Yuqing Zhu and Yu-Xiang Wang. Poisson subsampled Rényi differential privacy. In *International Conference on Machine Learning*, pages 7634–7642, 2019.

Yuqing Zhu, Jinshuo Dong, and Yu-Xiang Wang. Optimal accounting of differential privacy via characteristic function. In *International Conference on Artificial Intelligence and Statistics*, pages 4782–4817. PMLR, 2022.