Privately Publishable Per-instance Privacy

Anonymous Author(s)
Affiliation
Address
email

Abstract

We consider how to release personalized privacy losses using per-instance differential privacy (pDP), focusing on private empirical risk minimization over the class of generalized linear models. Standard differential privacy (DP) 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 paper, we 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 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 & 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.

While the data-independent formulation of standard DP is inarguably advantageous, in many cases the bound on the worst-case privacy loss guaranteed by DP is overly conservative for particular individuals with respect to a fixed dataset. On the flip side, a data curator might spend a large portion of the privacy budget protecting outliers in the dataset at the cost of model accuracy.

The privacy-utility trade-off challenges the practicality of differential privacy. 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 & 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.

Moreover, differentially private applications in practice commonly use high values of ϵ. A study of Apple's deployment of differential privacy 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 (Tang et al., 2017) – offering only scant privacy protection!

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Per-instance differential privacy shows promise as a means of navigating the privacy-utility trade-off. 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 paper, we propose a methodology to privately release the personalized privacy losses associated with private empirical risk minimization. Our contributions are as follows:

- We present a novel analysis of the per-instance privacy losses incurred by the objective perturbation mechanism, demonstrating that these 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 pDP losses. In the particular case of generalized linear models, we show that we can accurately publish the private pDP losses using a dimension- and dataset-independent bound. Furthermore, we present an alternative approach to privately release data-dependent bounds that provide a provably tight multiplicative approximation to the pDP losses and low privacy-loss overhead.

1.1 Related Work

This paper builds upon (Wang, 2019), which proposed the per-instance DP framework and left as an open question the matter of publishing the pDP losses. Notably, we embellish the pDP framework 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.

Our paper joins a body of work that seeks to provide stronger privacy guarantees by taking into account properties of the data. (Wang, 2019) gives an overview of several of these methodologies, including propose-test-release (Dwork & Lei, 2009) and local sensitivity (Nissim et al., 2007). In addition, Bayesian differential privacy (Triastcyn & Faltings, 2020) provides data-dependent privacy guarantees that afford strong protection to "typical" data (drawn from the same distribution as the dataset). The pDP definition that we use in this paper differs from the Bayesian DP formulation in that we condition on the differing individual's data, rather than making distributional assumptions about the sensitive data. The Rényi-based privacy filters of (Feldman & 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. Our privacy losses also depend on the output of the computations, but we choose to analyze this through the lens of *ex-post* differential privacy (Ligett et al., 2017). The work of (Papernot et al., 2018) initiated the problem of privately publishing data-dependent privacy losses but considered neither *per-instance* nor *ex-post* settings.

2 Preliminaries

2.1 Notation

We use conventional notation for common statistical objects: $\Pr[\cdot]$ for probability, $\mathbb{E}[\cdot]$ for expectation, etc. Adopting a standard abuse of 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 refer to both an individual and their data; i.e., individual z holds data $z = (x, y)$. $D_{\pm z} \in \mathcal{Z}^*$ denotes the fixed dataset $D = \{z_1, \ldots, z_n\} \in \mathcal{Z}^*$ with the data point z removed if $z \in D$, and the point z added 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 $z \notin D$, subtract otherwise".

Our paper considers a perturbed version of the following optimization problem:

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}}\, L(\theta; D) + r(\theta),$$

where $L(\theta; D) = \sum_{i=1}^{n} \ell(\theta; z_i)$. Throughout, we assume that $\ell(\theta; z)$ and $r(\theta)$ are convex and twice-differentiable.

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 Differential Privacy

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

Definition 1. (Differential privacy) A randomized algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{R}$ satisfies (ϵ, δ)-DP if for all datasets $D \in \mathcal{Z}^*$ and data points $z \in \mathcal{Z}$, and for all measurable sets $S \subset \mathcal{R}$, it holds that

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

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

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

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 more 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 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 $o \in \mathrm{Range}(\mathcal{A})$ 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 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 $\mathcal{A}$.*

The output o of algorithm $\mathcal{A}$ is distributed as $o \sim \mathcal{N}(Q(D), \sigma^2)$. So the ex-post *pDP can be 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 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 solution to a private ERM problem. Any individual (not just those whose data is contained in the dataset) can plug in 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.

2.3 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

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}}\, L(\theta; D) + r(\theta),$$

where $r(\theta)$ is a convex and twice-differentiable regularizer and $L(\theta; D) = \sum_{i=1}^{n} \ell(\theta; z_i)$ is a loss function. 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 $y \in \mathcal{Y} \subseteq \mathbb{R}$. $\Theta \subseteq \mathbb{R}^d$ is a convex domain.

Our assumptions on the data distribution are fairly mild. We posit only that $x \in \mathcal{X}$ are sampled from some distribution on the unit ball, so that $||x|| \leq 1$ for all $x \in \mathcal{X}$. We also assume that $|y| \leq 1$ for all $y \in \mathcal{Y}$.

Our pDP analysis will consider objective perturbation, a well-known approach for privacy-preserving ERM. We review this algorithm and its privacy guarantees in the next section.

2.4 Objective Perturbation

The objective perturbation algorithm solves

$$\hat{\theta}^P = \underset{\theta \in \Theta}{\operatorname{argmin}}\, L(\theta; D) + r(\theta) + \frac{\lambda}{2}||\theta||^2 + b^T\theta, \tag{1}$$

where $b \sim \mathcal{N}(0, \sigma^2 I_d)$ .Below, we present a simplified version of the objective perturbation algorithm and state its privacy guarantees.

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

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 $r(\theta)$.
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 (1).

Theorem 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. Let β be an upper bound on the eigenvalues of the Hessian $\nabla^2 \ell(\theta; z_i)$, for all $z_i \in \mathcal{X} \times \mathcal{Y}$ and for all $\theta \in \Theta$. Let $\ell(\cdot)$ have a bounded gradient such that $||\nabla \ell(\theta; z_i)|| \leq \xi$ for all $z_i \in \mathcal{X} \times \mathcal{Y}$ and for all $\theta \in \Theta$. For $\lambda \geq \frac{2\beta}{\epsilon_1}$ and $\sigma = \frac{\xi^2(8\log(2/\delta) + 4\epsilon_1)}{\epsilon_1^2}$, Algorithm 1 satisfies (ϵ_1, δ_1)-differential privacy.*

The objective perturbation algorithm chooses a regularization strength to ensure that the objective function is strongly convex, and also requires smoothness. This means that the objective function is well-conditioned and therefore robust to small perturbations. In other words, for sufficiently large λ, the minimizer of an objective function that is λ-strongly convex will be insensitive to any particular data point.

3 Privately Publishable pDP

3.1 pDP Analysis of Objective Perturbation

Our goal in this section is to derive the personalized privacy losses associated with observing the output $\hat{\theta}^P$ of objective perturbation under Definition 3. As it turns out, this *ex-post* perspective is not only more adaptive, but also more 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 6 (*ex-post* pDP loss of objective perturbation for a convex loss function). *Let $J(\theta; D) = L(\theta; D) + r(\theta) + \frac{\lambda}{2}||\theta||^2$ 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 $b \sim \mathcal{N}(0, \sigma^2 I)$. Then for every privacy target $z = (x, y)$, releasing $\hat{\theta}^P = \operatorname{argmin}_{\theta \in \mathbb{R}^d} J(\theta; D) + b^T\theta$ satisfies $\epsilon_1(\hat{\theta}^P, D, D_{\pm z})$-ex-post per-instance differential privacy with*

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

where $\mu_j = \lambda_j u_j^T \left(\nabla \mathbf{b}(\hat{\theta}^P; D) \mp \sum_{k=1}^{j-1} \lambda_k u_k u_k^T\right)^{-1} u_j$ 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 added noise 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. Next we borrow a trick from (Kifer et al., 2012), observing that since $\hat{\theta}^P$ is the minimizer of the private objective function $J(\theta; D) + b^T\theta$, we can set its gradient to 0 and 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 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 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 7 (*ex-post* pDP loss of objective perturbation for GLMs)**.** *Let* $J(\theta; D) = L(\theta; D) + r(\theta) + \frac{\lambda}{2}||\theta||^2$ *such that* $L(\theta; D) = \sum_i \ell_i(\theta)$ *is a linear loss function, and sample* $b \sim \mathcal{N}(0, \sigma^2 I)$*. Then for every privacy target* $z = (x, y)$*, releasing* $\hat{\theta}^P = \text{argmin}_{\theta\in\Theta} J(\theta; D) + b^T\theta$ *satisfies* $\epsilon_1(\hat{\theta}^P, D, D_{\pm z})$*-ex-post per-instance differential privacy with*

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

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* $f''(\cdot) = f''(x^T\hat{\theta}^P; y)$*.*

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.

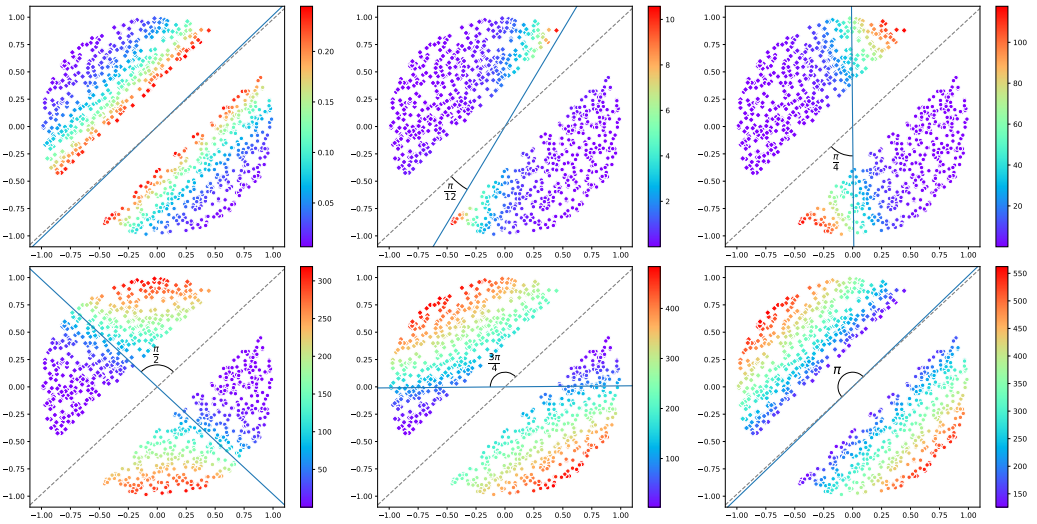

Figure 1: Visualization of *ex-post* pDP losses for logistic regression ($n = 1000, d = 2$).

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 7 in order to calculate the pDP distribution induced by any hypothetical $\hat{\theta}^P$. For Figure 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 $\hat{\theta} = \text{argmin}\, J(\theta; D)$ with $\lambda = 1$ to minimize the logistic loss, and directly perturb the output by rotating it by angle $\omega \in [0, \frac{\pi}{12}, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \pi]$. The color scale is a function of the *ex-post* pDP loss of data point z.

Figure 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 Algorithm 1, 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!

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?

Observe that the expression from Theorem 6 depends on the dataset D only through two quantities: the leverage score $\mu = x^T \left(\nabla^2 J(\hat{\theta}^P; D)\right)^{-1} x$ 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.

3.2.1 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 8. *Let $\ell(\cdot)$ be β-smooth, such that $\nabla^2 \ell(\theta; z) \prec \beta I_d$ for all $\theta \in \mathbb{R}^d$ and $z \in \mathcal{Z}$.*

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

where $\mu_j = \lambda_j u_j^T \left(\nabla \mathbf{b}(\hat{\theta}^P; D) \mp \sum_{k=1}^{j-1} \lambda_k u_k u_k^T\right)^{-1} u_j$ according to the eigendecomposition $\nabla^2 \ell(\hat{\theta}^P; z) = \sum_{k=1}^{d} \lambda_k u_k u_k^T$. 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 $-\log\left(1 - \frac{f''(x^T \hat{\theta}^P)||x||^2}{\lambda}\right)$.

Theorem 9. *Let $\hat{\theta}^P$ be a random variable such that $\hat{\theta}^P = \operatorname{argmin}\left(J(\theta; D) + b^T \theta\right)$ as in (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

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

We make a few observations on the bounds. First, the general bound in Theorem 9 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 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!

3.3 The privacy report

For certain regimes, we may wish to consider privatizing the data-dependent quantities of the 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 $f''(\cdot) := f''(x^T\hat{\theta}^P)$, where $\ell(\theta; z) = f(x^T\theta; y)$.

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

Input: $\hat{\theta}^p$ from `Obj-Pert`, noise parameter σ, σ_2; 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
Output: Reporting function $\tilde{\epsilon} : (x, y), \delta \to \mathbb{R}^2_+$
if B = `DATA-INDEP` **then**
  Set $\epsilon_2(\cdot) := 0$.
  Set $\overline{\mu^p}(x) := \frac{\|x\|^2}{\lambda}$.
else if B = `DATA-DEP` **then**
  Privately release $\hat{H}^p$ by a variant of "Analyze Gauss"[2] with parameter σ_2.
  Set $\epsilon_2(\cdot)$ according to Statement 2 of Theorem 10.
  Set $\tau = F^{-1}_{\lambda_1(\mathrm{GOE}(d))}(1 - \rho/2)$.

  if $\lambda \geq 2\tau\sigma_2$ **then**
    Set $\overline{\mu^p}(x) = \frac{3}{2}x^T[\hat{H}^p]^{-1}x$.
  end if
end if
Set $\overline{\epsilon^p_1}(z) := \max\left\{\left|-\log\left(1 \pm f''(\cdot)\overline{\mu^p}(x)\right) + \frac{|f'(\cdot)|^2||x||^2}{2\sigma^2} \pm \frac{|f'(\cdot)|||x||F^{-1}_{\mathcal{N}(0,1)}(1-\rho/2)}{\sigma}\right|\right\}$.
Output the function $\tilde{\epsilon}(z) := (\overline{\epsilon^p_1(z)}, \epsilon_2(z))$.

Note that the pDP function $\epsilon_2(\cdot)$ – which we use to report the additional pDP loss of releasing $\hat{H}^P$, the private estimate of the Hessian – does not depend on the dataset, and thus is not required to be separately released.

Theorem 10. *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. $(\frac{\beta^2}{4\sigma_2^2} + \frac{\beta\sqrt{\log(1/\delta)}}{\sigma_2}, \delta)$*-DP*
2. $(\frac{f''(x,\hat{\theta}^p)^2\|x\|^4}{4\sigma_2^2} + \frac{f''(x,\hat{\theta}^p)\|x\|^2\sqrt{\log(1/\delta)}}{\sigma_2}, \delta)$*-pDP for all $x \in \mathcal{X}$ and $0 \leq \delta < 1$.*
3. *For a fixed input z, and all $\rho > 0$, the privately released privacy report $\tilde{\epsilon}(\cdot)$ satisfies that $\epsilon_1(z, \hat{\theta}^p) \leq \overline{\epsilon^p}_1(z, \hat{\theta}^p) \leq 12\epsilon_1(z, \hat{\theta}^p)$ with probability $1 - 2\rho$ where ϵ_1 is the expression from Theorem 6.*

Constant approximation with low privacy cost. This theorem suggests that if we choose $\sigma_2 \asymp \sigma$ and use a slightly large λ in ObjPert we could obtain *constant multiplicative approximation* of the per-instance privacy loss for all individuals, with only a constant blow-up in both the DP and pDP losses. 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). We discuss a more adaptive algorithm in the appendix that adapts based on a well-conditioned H matrix that avoids using a large λ and achieves a stronger approximation.

Joint DP interpretation. Finally, we can also interpret our results from a joint-DP perspective (Kearns et al., 2014). Given any realized output $\tilde{\theta}^p \in \mathbb{R}^d$, the tuple of $\{\tilde{\epsilon}(z_1, \hat{\theta}^p), ..., \tilde{\epsilon}(z_1, \hat{\theta}^p)\}$ satisfies joint-DP and joint-pDP with the same ϵ parameter as in Theorem 10.

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

4 Experiments

We evaluate our methods to release the pDP losses using both real-word and synthetic data, focusing for now on logistic regression.

4.1 True pDP loss ϵ_1 vs. released bound $\tilde{\epsilon_1}$

In this experiment we use a synthetic dataset, sampling $\theta \sim \mathcal{N}(0, 1)$, sampling $X \in \mathbb{R}^{n \times d}$ from the d-dimensional unit ball. Y is then a deterministic function of X and θ such that $\Pr[Y = 1] = \texttt{sigmoid}(X^T\theta) + E$ for noise matrix E. Figure 2 plots the pDP distributions for the true ϵ_1 and its data-independent bound $\tilde{\epsilon}$ calculated according to Algorithm 2, demonstrating both that the true *ex-post* pDP loss is an order of magnitude improvement compared to the worst case DP bound ϵ, and that the data-independent bound provides a good approximation to the true pDP loss.

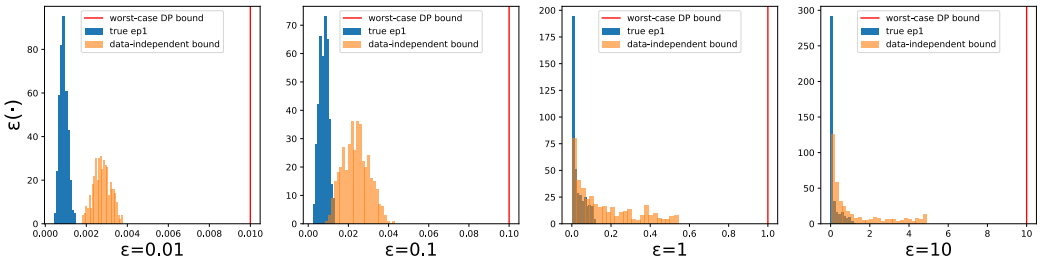

Figure 2: pDP distributions for the true ϵ_1 and its data-independent bound.

4.2 Dimension

We generate synthetic data as described in the previous experiment, this time varying $d \in [1, 5, 10, 20, 30, 40, 50, 60]$. We choose $\epsilon = 0.5$ as the DP bound and $\delta = 10^{-3}$, and set σ, λ according to Theorem 5. Figure 3 plots the true worst-, best- and median-case *ex-post* pDP loss over all individuals z against its data-independent counterpart (i.e., the worst-case pDP loss is calculated as $\max_{z \in D} \epsilon(\cdot, D_{\pm z})$ and the worst-case data-independent bound as $\max_{z \in D} \tilde{\epsilon}(\cdot, z)$).

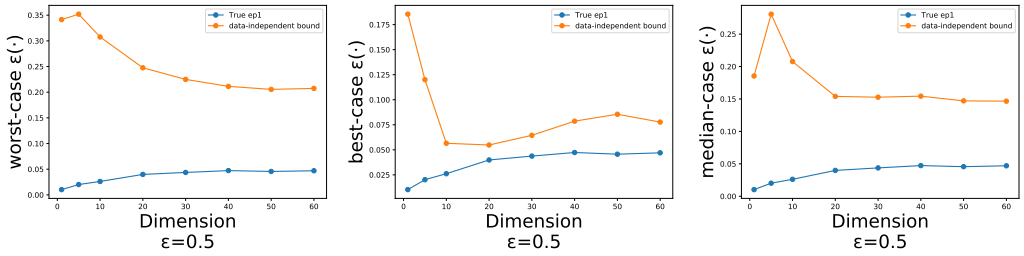

Figure 3: pDP losses across varying data dimensions.

5 Conclusion

In this paper we show how to privately release the personalized privacy losses incurred by the objective perturbation mechanism. We present both a simple data-independent bound on the pDP losses, as well as a data-dependent approach which provides stronger per-instance privacy guarantees under certain regimes. Our theoretical and experimental results show that our methods provide strong, adaptive privacy guarantees and that we can release privatized pDP losses which are good approximations of the true data-dependent pDP losses.

Our framework applies to a wide range of learning problems, but our pDP analysis limits us to unconstrained optimization. In future work we will explore how to extend our approach to optimization over any convex domain.

References

Balle, B. and Wang, Y.-X. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pp. 394–403. PMLR, 2018.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Symposium on Foundations of Computer Science*, pp. 464–473. IEEE, 2014.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Chiani, M. 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.

Dwork, C. and Lei, J. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 371–380, 2009.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014a.

Dwork, C., Talwar, K., Thakurta, A., and Zhang, L. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 11–20, 2014b.

Feldman, V. and Zrnic, T. Individual privacy accounting via a renyi filter. *arXiv preprint arXiv:2008.11193*, 2020.

Kearns, M., Pai, M., Roth, A., and Ullman, J. Mechanism design in large games: Incentives and privacy. In *Conference on Innovations in theoretical computer science (ITCS-14)*, pp. 403–410, 2014.

Kifer, D., Smith, A., and Thakurta, A. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pp. 25–1. JMLR Workshop and Conference Proceedings, 2012.

Ligett, K., Neel, S., Roth, A., Waggoner, B., and Wu, Z. S. Accuracy first: Selecting a differential privacy level for accuracy-constrained erm. *Advances in Neural Information Processing Systems*, 2017:2567–2577, 2017.

Narayanan, A. and Shmatikov, V. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125. IEEE, 2008.

Nissim, K., Raskhodnikova, S., and Smith, A. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 75–84, 2007.

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., and Úlfar Erlingsson. Scalable private learning with pate. In *International Conference on Learning Representations (ICLR-18)*, 2018.

Rudelson, M. and Vershynin, R. 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*, pp. 1576–1602. World Scientific, 2010.

Stewart, G. W. Matrix perturbation theory. 1990.

Tang, J., Korolova, A., Bai, X., Wang, X., and Wang, X. Privacy loss in apple's implementation of differential privacy on macos 10.12. *arXiv preprint arXiv:1709.02753*, 2017.

Triastcyn, A. and Faltings, B. Bayesian differential privacy for machine learning. In *International Conference on Machine Learning*, pp. 9583–9592. PMLR, 2020.

Wang, Y.-X. Per-instance differential privacy. *Journal of Privacy and Confidentiality*, 9(1), 2019.

Wei, B.-C., Hu, Y.-Q., and Fung, W.-K. Generalized leverage and its applications. *Scandinavian Journal of statistics*, 25(1):25–37, 1998.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Part

Appendix

Table of Contents

A Additional Experiments **12**
A.1 The effect of stronger regularization 12
A.2 Comparing data-independent and data-dependent bounds 13

B Even Stronger Privacy Report **15**
B.1 More Accurate Privacy Report by Adapting to the Data 15
B.2 Dataset-Dependent Privacy report for general smooth learning problems 18
B.3 Uniform Privacy Report and Privacy Calibration 18

C Improved "Analyze Gauss" with Gaussian Orthogonal Ensembles **19**
C.1 Exact statistical inference with the Gaussian Orthogonal Ensemble 21

D Omitted Proofs **21**
D.1 Proofs for the pDP analysis of objective perturbation 22
D.2 Proofs for the Privacy Report in the main paper 24

E pDP Analysis of the Gaussian mechanism **24**

F Technical Lemmas **25**

A Additional Experiments

A.1 The effect of stronger regularization

The data-independent bound introduced in Section 3.2.1 provide stronger privacy guarantees for choices of larger λ; likewise, the data-dependent bounds of Section 3.3 require that λ be above a certain threshold. We will now show that enforcing more regularization will not hurt the utility of the algorithm.

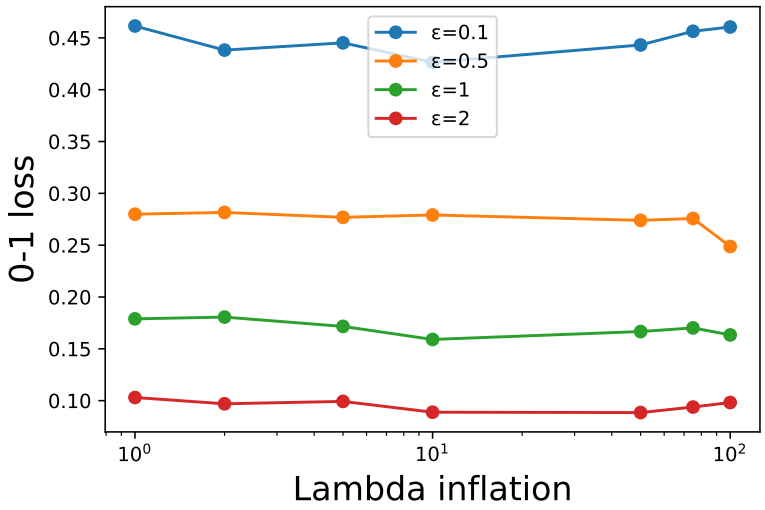

Figure 4: Utility across varying λ.

In Figure 4, we plot the results of running Algorithm 1 for logistic regression on a synthetic dataset generated according to the scheme described in Section 4. Our fixed experimental parameters are $n = 500, d = 50$ and $\delta = 10^{-3}$. Since each value of ϵ demands a different minimum value of λ in order to achieve (ϵ, δ)-differential privacy, we plot the "lambda inflation" on the x-axis: a measure of how many times larger we set λ than its minimum value required by DP. E.g., for logistic regression we require $\lambda \geq \frac{1}{2\epsilon}$, and so in Figure 4 a λ-inflation value of 10 means that we set $\lambda = \frac{5}{\epsilon}$. Figure 4 clearly demonstrates that we do not sacrifice the utility of our algorithms by specifying larger values of λ.

A.2 Comparing data-independent and data-dependent bounds

Next, we compare how the data-independent and data-dependent bounds on $\mu(\cdot)$ affect how accurately we can release $-\log(1 - f''(\cdot)\mu(\cdot))$, an upper bound on the first term of the *ex-post* pDP loss $\epsilon_1(\cdot)$. We use synthetic data on a linear regression task, with $n = 500, d = 50$ and $\delta = 10^{-3}$. Note that we normalize X so that $||x|| = 1$ for all $x \in X$. We set $\tau = 1.7\sqrt{d} \approx 12$ and $\sigma_2 = 0.1\sigma$ for the data-dependent bounds. We set $\lambda = \max(2, 2\tau\sigma_2)$. Figure 5 plots the distributions of the ground-truth value $\mu(\cdot)$ and the data-dependent bound $\overline{\mu^P}_{\texttt{DEP}}$, as well as the vertical line representing the data-independent bound $\overline{\mu^P}_{\texttt{INDEP}}$ which is identical for each x since each data point is normalized to 1.

Figure 5 corroborates the findings of Theorem 13, showing that the data-dependent private estimate of $\mu(\cdot)$ provides a constant multiplicative approximation to the true value. We observe that the data-dependent approximation fares better than the data-independent bound for smaller σ, σ_2, but are comparable for larger noise scales – for $\sigma = 10, \sigma_2 = 1$, the distribution of the data-dependent bound on $\mu(\cdot)$ collapses almost fully into the constant data-independent bound.

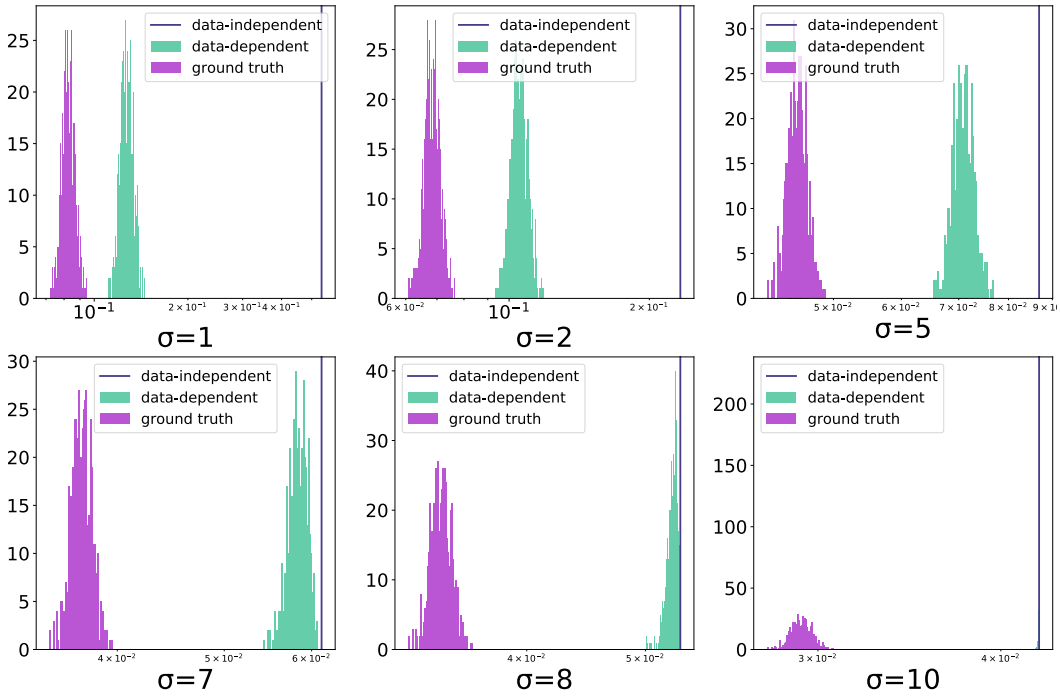

Figure 5: Data-independent and data-dependent bounds on $-\log(1 - f''(\cdot)\mu(\cdot))$, compared to its true value.

Figure 6 offers a different view of the same experimental results illustrated in Figure 5. We plot the distribution of the ratio $\frac{-\log(1-f''(\cdot)\overline{\mu(\cdot)^P}}{-\log(1-f''(\cdot)\mu(\cdot)}$ for the data-independent and data-dependent approximation $\overline{\mu(\cdot)^P}$. For $\sigma \in \{1, 2\}$, we see that the distribution of the data-dependent bound

is an almost perfectly multiplicative approximation of the true $\mu(\cdot)$. As σ and σ_2 increase, the data-dependent and data-independent approximations become virtually identical.

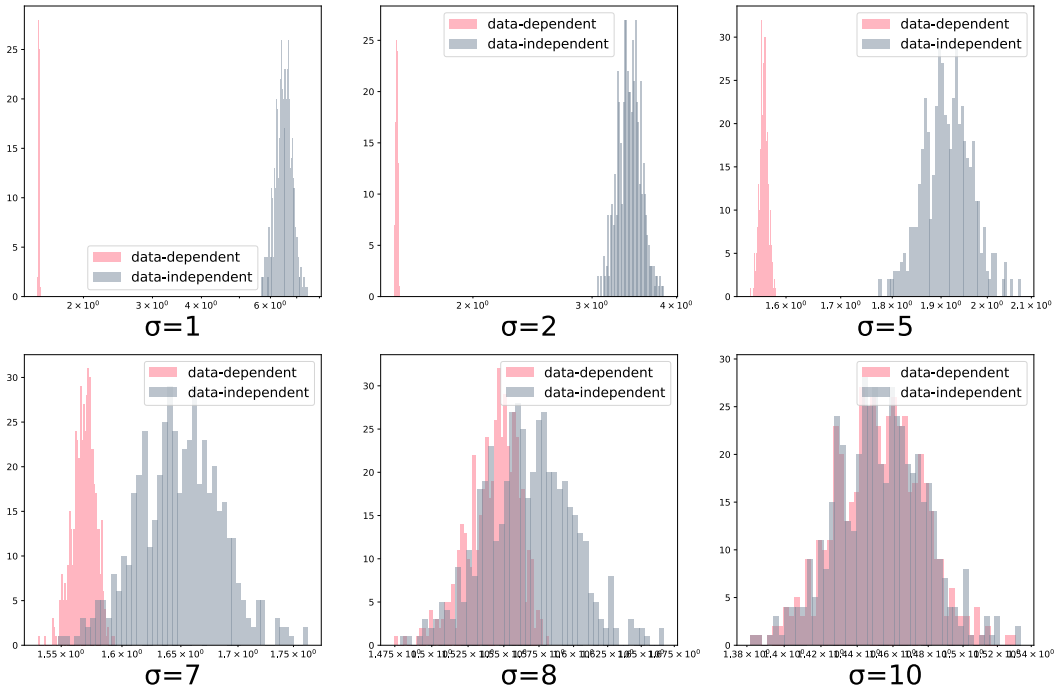

Figure 6: Ratio of $\frac{-\log(1-f''(\cdot)\overline{\mu(\cdot)^P}}{-\log(1-f''(\cdot)\mu(\cdot)}$ for data-independent and data-dependent $\overline{\mu(\cdot)^P}$.

Next, we run objective perturbation for linear regression on the real-world UCI "wine quality" dataset, varying σ (and varying λ as a function of σ). Our experimental parameters are as described in the previous experiments ($n = 500, d = 50$ and $\delta = 10^{-3}$, $\tau = 1.7\sqrt{d} \approx 12$ and $\sigma_2 = 0.1\sigma$).

Note that for linear regression with a well-conditioned $X^T X$, we will have a much better approximation with the "adaptive" algorithm presented in the next section. Empirical evaluation of the adaptive algorithm is left to a longer version of the paper.

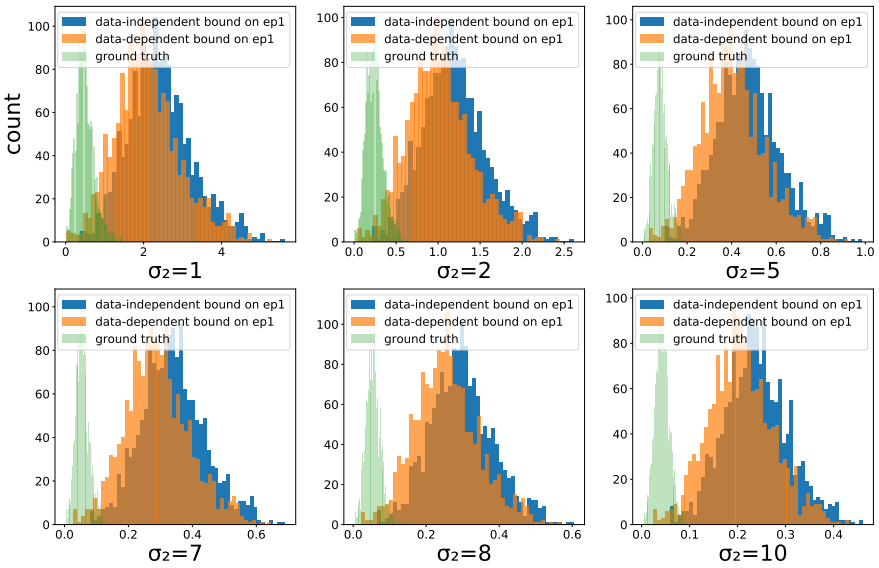

Figure 7: Distribution of pDP losses (both true and the data-dependendent and -independent bounds) for linear regression on the UCI "wine quality" dataset. Technically speaking, since we are dealing with an unbounded domain $\mathbb{R}^d$, the algorithm does not satisfy worst-case DP for any $\epsilon < \infty$.

B Even Stronger Privacy Report

B.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 4 makes use of a subroutine to add noise to the smallest eigenvalue of H, presented below along with its privacy guarantees.

Algorithm 3 Releasing the smallest eigenvalue of H

Input: Dataset D, noise parameter σ_3, $\lambda_{\min}$ denoting the smallest eigenvalue of H.
Output: $\hat{\lambda}^P_{\min}$.
Output $\hat{\lambda}^P_{\min} = \lambda_{\min} + \mathcal{N}(0, \sigma_3^2)$.

Theorem 11. *Algorithm 3 satisfies pDP with*

$$\epsilon_3(\cdot) = \frac{(f''(x^T\hat{\theta}^p))^2\|x\|^4}{2\sigma_3^2} + \frac{f''(x^T\hat{\theta}^p)\|x\|^2\sqrt{2\log(1/\delta)}}{\sigma_3},$$

and if $f''(x, \hat{\theta}^p)\|x\|^2 \leq \beta$ for all x then Algorithm 3 also satisfies (ϵ, δ)-DP with $\epsilon = \frac{\beta^2}{2\sigma_3^2} + \frac{\beta\sqrt{2\log(1/\delta)}}{\sigma_3}$.

Proof. Algorithm 3 is a standard Gaussian mechanism. By Weyl's lemma, the smallest singular value satisfies a perturbation bound of $f''(x,\hat{\theta}^p)\|xx^T\|_2 = f''(x,\hat{\theta}^p)\|x\|^2$ 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 $f''(x,\hat{\theta}^p)\|x\|^2$ with $\|\nabla^2\ell(\hat{\theta}^p,z)\|_F$.

Algorithm 4 More adaptive privacy report for `Obj-Pert`

Input: $\hat{\theta}^p$ from `Obj-Pert`, noise parameter $\sigma_1,\sigma_2,\sigma_3$; regularization parameter λ; Hessian $H := \sum_i \nabla^2\ell(\hat{\theta}^p;z_i) + \lambda I_d$.
Output: Reporting function $\tilde{\epsilon}:(x,y),\rho \to \mathbb{R}^2_+$.
Set $\tau = F^{-1}_{\lambda_1(\mathrm{GOE}(d))}(1-\rho/2)$.
Privately release $\hat{H}^p$ by Algorithm 6 with parameter σ_2.
Set $\epsilon_2(\cdot)$ according to Theorem 17
Privately release $\hat{\lambda}^p_{\min} = \lambda_{\min} + \mathcal{N}(0,\sigma_3^2)$ (Algorithm 3).
Set $\epsilon_3(\cdot)$ according to Theorem 11.
Set $\underline{\hat{\lambda}^p_{\min}} := \max\{\lambda, \hat{\lambda}^p_{\min} - \sigma_3 F^{-1}_{\mathcal{N}(0,1)}(1-\rho/2)\}$.
Set $\overline{\mu^p}(x) = \min\{\frac{\hat{\lambda}^p_{\min}+2\tau\sigma_2}{\underline{\hat{\lambda}^p_{\min}}}x^T(\hat{H}^p+\tau\sigma_2 I_d)^{-1}x, \frac{\|x\|^2}{\underline{\hat{\lambda}^p_{\min}}}\}$.
Set $\overline{\epsilon^p_1}(\cdot) := \max\left\{\left|-\log\left(1 \pm f''(z,\hat{\theta}^p)\overline{\mu^p}(x)\right) + \frac{|f'(z,\hat{\theta}^p)|^2\|x\|^2}{2\sigma_1^2} \pm \frac{|f'(z,\hat{\theta}^p)|\|x\|F^{-1}_{\mathcal{N}(0,1)}(1-\rho/2)}{\sigma_1}\right|\right\}$.
Return the "privacy report" function $\tilde{\epsilon} = (\overline{\epsilon^p_1}, \epsilon_2+\epsilon_3)$, i.e., the ex post pDP of Algorithm 1 and the pDP of Algorithm 4 (i.e., overhead).

This algorithm allows any choices 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 $\overline{\mu}^p(x)$ we used 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 12 (Uniform multiplicative approximation of a regularized estimator)**.** *Let $\underline{\hat{\lambda}^p_{\min}}$ and $\hat{H}^P$ be constructed as in Algorithm 4. Then with probability $1-2\rho$,*

$$\lambda_{\min} - \sigma_3 F^{-1}_{\mathcal{N}(0,1)}(1-\rho/2) \le \hat{\lambda}^p_{\min} \le \lambda_{\min} + \sigma_3 F^{-1}_{\mathcal{N}(0,1)}(1-\rho/2)$$

and for all $x \in \mathbb{R}^d$ simultaneously

$$x^T(\hat{H}^P+\tau\sigma_2 I_d)^{-1}x \le \mu(x) \le \frac{\hat{\lambda}^p_{\min}+2\tau\sigma_2}{\underline{\hat{\lambda}^p_{\min}}}x^T(\hat{H}^P+\tau\sigma_2 I_d)^{-1}x.$$

Proof. By Lemma 20, 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 $1-\rho$,

$$\lambda_{\min} - \sigma_3 F^{-1}_{\mathcal{N}(0,1)}(1-\rho/2) \le \hat{\lambda}^p_{\min} \le \lambda_{\min} + \sigma_3 F^{-1}_{\mathcal{N}(0,1)}(1-\rho/2)$$

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

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

Therefore with probability $1-2\rho$,

$$\begin{aligned} H &\prec H + G + \tau\sigma_2 I_d \prec H + 2\tau\sigma_2 I = H - \underline{\hat{\lambda}^p_{\min}} I_d + \underline{\hat{\lambda}^p_{\min}} I_d + 2\tau\sigma_2 I_d \\ &\prec \frac{\hat{\lambda}^p_{\min}+2\tau\sigma_2}{\underline{\hat{\lambda}^p_{\min}}}(H - \underline{\hat{\lambda}^p_{\min}} I_d + \underline{\hat{\lambda}^p_{\min}} I_d) = \frac{\hat{\lambda}^p_{\min}+2\tau\sigma_2}{\underline{\hat{\lambda}^p_{\min}}} H. \end{aligned}$$

Taking the inverse on both sides, we get

$$\frac{\underline{\hat{\lambda}^p_{\min}}}{\underline{\hat{\lambda}^p_{\min}} + 2\tau\sigma_2} H^{-1} \prec (\hat{H}^P + \tau\sigma_2 I_d)^{-1} \prec H^{-1}.$$

Thus for all $x \in \mathbb{R}^d$, $x^T(\hat{H}^P + \tau\sigma_2 I_d)^{-1}x \leq x^T H^{-1} x \leq \frac{\underline{\hat{\lambda}^p_{\min}} + 2\tau\sigma_2}{\underline{\hat{\lambda}^p_{\min}}} x^T(\hat{H}^P + \tau\sigma_2 I_d)^{-1}x$. □

The privacy (DP and pDP) of Algorithm 4 is a composition of the stated results in Theorem 10 with the the privacy guarantees stated in Theorem 11. Observe that if we choose $\sigma_3 = \sigma_1$ then the additional DP an pDP losses are smaller than that 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 release the pDP losses with a multiplicative factor of $1 + o(1)$.

Theorem 13 (Utility of Adaptive privacy report.)**.** *Assume $\lambda_{\min}(H) \geq \max\{2\beta, 2\tau\sigma_2\}$. There is a universal constant $0 < C \leq 4\tau\sigma_2 + 2\beta$ such that for a fixed $z \in \mathcal{X} \times \mathcal{Y}$, and all $\rho > 0$, the privately released privacy report $\tilde{\epsilon}(\cdot)$ from Algorithm 4 obeys that*

$$\epsilon_1(\cdot) \leq \overline{\epsilon^p}_1(\cdot) \leq (1 + \frac{C}{\lambda_{\min}})\epsilon_1(\cdot)$$

with probability $1 - 3\rho$ where ϵ_1 is the expression from Theorem 6.

Proof of Theorem 13. Similar to the proof of Theorem 10, 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 12 and the high probability event in Theorem 9 (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 $\log(1 + f''\mu)$ or $-\log(1 - f''\mu)$.

By our assumption that $\lambda > 2\beta$, as well as the pointwise minimum in the construction of $\overline{\mu^p}$ from Algorithm 4, 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 $\log(1 + f''\mu)$ or $-\log(1 - f''\mu)$ using the plug-ins: $\log(1 + f''\overline{\mu^p})$ or $-\log(1 - f''\overline{\mu^p})$

For brevity, in the subsequent derivation we will be using a to denote $f''(\cdot)\mu(x)$ and $\hat{a}$ to denote $f''(x, \hat{\theta}^p)\overline{\mu^p}(x)$.

Thus

$$\begin{aligned}\log(1 + a) \leq &\log(1 + \hat{a}) \leq \hat{a} \leq (1 + \frac{2\tau\sigma_2}{\underline{\hat{\lambda}^p_{\min}}})a \leq (1 + \frac{2\tau\sigma_2}{\underline{\hat{\lambda}^p_{\min}}})(1 + a)\log(1 + a) \\ \leq &(1 + \frac{4\tau\sigma_2}{\lambda_{\min}})(1 + \frac{\beta}{\lambda_{\min}})\log(1 + a) \leq (1 + \frac{C}{\lambda_{\min}})\log(1 + a)\end{aligned}$$

where C can be taken as $4\tau\sigma_2 + 2\beta$, by our assumption on $\lambda_{\min}$ and a high probability bound under which $\underline{\hat{\lambda}^p_{\min}} \geq \lambda_{\min}/2$.

Similarly,

$$-\log(1 - a) \leq \frac{a}{1 - a} \leq \frac{\hat{a}}{1 - a} \leq \frac{(1 + \frac{2\tau\sigma_2}{\underline{\hat{\lambda}^p_{\min}}})a}{1 - a} \leq \frac{(1 + \frac{2\tau\sigma_2}{\underline{\hat{\lambda}^p_{\min}}})}{1 - \frac{\beta}{\lambda_{\min}}}(-\log(1 - a))$$

where

$$\frac{(1 + \frac{2\tau\sigma_2}{\underline{\hat{\lambda}^p_{\min}}})}{1 - \frac{\beta}{\lambda_{\min}}} = 1 + \frac{2\tau\sigma_2}{\underline{\hat{\lambda}^p_{\min}}} + \frac{\beta/\lambda_{\min}}{1 - \beta/\lambda_{\min}} \leq 1 + \frac{4\tau\sigma_2 + 2\beta}{\lambda_{\min}}$$

under our assumption for $\lambda_{\min}$, β. □

B.2 Dataset-Dependent Privacy report for general smooth learning problems

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

For the third term in the pDP bound of Theorem 10, 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 5 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_4, 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 $e \sim \mathcal{N}(0, \sigma_4^2 I)$.
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 $\epsilon_4(\cdot)$ s.t. $\epsilon_4(z) = \frac{\|\nabla \ell_z(\hat{\theta}^P; z)\|^2}{2\sigma_4^2} + \frac{\|\nabla \ell_z(\hat{\theta}^P; z)\| \sqrt{2 \log(2/\delta)}}{\sigma_4}$.

Theorem 14. *Let $\hat{\theta}^P$ be fixed, Algorithm 5 satisfies*

1. *$(\epsilon_4(D, D_{\pm z}), \delta)$-pDP, with*

$$\epsilon_4(D, D_{\pm z}) = \frac{\|\nabla \ell_z(\hat{\theta}^P; z)\|^2}{2\sigma_4^2} + \frac{\|\nabla \ell_z(\hat{\theta}^P; z)\| \sqrt{2 \log(1/\delta)}}{\sigma_4}.$$

2. *$\epsilon_4(o, D, D_{\pm z})$-ex post pDP with probability $1 - \rho$,*

$$\epsilon_4(o, D, D_{\pm z}) = \frac{\|\nabla \ell_z(\hat{\theta}^P; z)\|^2}{2\sigma_4^2} + \frac{\|\nabla \ell_z(\hat{\theta}^P; z)\| \sqrt{2 \log(2/\rho)}}{\sigma_4}.$$

Proof. This is a Gaussian mechanism and the proof follows from Corollary 23. □

The theorem avoids an additional dependence in d from the L1-norm $\|\nabla \ell(\hat{\theta}^p; z)\|_1$ in the dataset-independent bound.

We remark that Algorithm 5's pDP loss is dataset independent and if we choose $\sigma_4 = \sigma_1$, the pDP losses for running Algorithm 5 is on the same order as that of the main algorithm. Thus is 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.

We defer analysis of this estimator on how accurately it approximates the first term to a longer version of the paper.

B.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 4) 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.

Proposition 15 (Uniform privacy report)**.** *With probability* $1-2\rho$, simultaneously *for all* n *users in the dataset,* $\epsilon_1(z,\hat{\theta}^p) \le \overline{\epsilon_1}^P(z,\hat{\theta}^p) \le 12\sqrt{\log n}\epsilon_1(z,\hat{\theta}^p)$

If we, instead, use $\frac{|f'(z,\hat{\theta}^p)|\|x\|_1\sqrt{2\log(2d/\rho)}}{\sigma_1}$ *to replace the third-term in* $\overline{\epsilon^p}_1$*, then with probability* $1-2\rho$, simultaneously *for all* $x \in \mathcal{X}$*, the ex-post pDP report* $\overline{\epsilon^p}_1$ *satisfies that*

$$\epsilon_1(z,\hat{\theta}^p) \le \overline{\epsilon^p}_1(z,\hat{\theta}^p) \le 12\sqrt{d\log d}\epsilon_1(z,\hat{\theta}^p).$$

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, and that $\|x\|_1 \le \sqrt{d}\|x\|_2$. □

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 $f'' \le f'$.

Corollary 16 (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*

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

For those cases when $\frac{(f'')\|x\|^2}{\rho_{\max}} \le \frac{|f'|\|x\|}{\beta}$ *(which is the case in logistic regression for all* x *s.t.,* $\|x\| \le 1$*), the additional overhead in releasing a dataset-dependent pDP is smaller than the ex post pDP bound in Theorem 6.*

C 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 $\|H_x\|_F \le \beta$ 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:

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

where UpperTriangle$(H) \in \mathbb{R}^{d^2/2+d/2}$ is the vector that enumerates the elements of upper-triangular region of H. The resulting Gaussian noise is an i.i.d $\mathcal{N}(0, \sigma_2^2)$ and it satisfy (ϵ, δ)-DP with

$$\epsilon = \frac{\beta^2}{2\sigma_2^2} + \frac{\beta\sqrt{2\log(1/\delta)}}{\sigma_2}.$$

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

Algorithm 6 Release H (a natural variant of "Analyze-Gauss")

Input: Dataset D, noise parameter σ_2, $H = \sum_{i=1}^n \nabla^2 \ell(z_i, \hat{\theta}^P) + \lambda I_d$.
Output: $\hat{H}^P$.
Draw a Gaussian random matrix $Z \in \mathbb{R}^{d\times d}$ with $Z_{i,j} \sim \mathcal{N}(0, \sigma_2^2)$ 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 17. *Algorithm 6 satisfies pDP with*

$$\epsilon(\cdot) = \frac{\|H_x\|_F^2}{4\sigma_2^2} + \frac{\|H_x\|_F\sqrt{2\log(1/\delta)}}{\sqrt{2}\sigma_2},$$

and $\hat{H}^p$ satisfies ex post pDP of the same ϵ with probability $1-2\delta$. If in addition $\sup_{x\in\mathcal{X}} \|H_x\|_F \le \beta$ then, Algorithm 6 satisfies (ϵ, δ)-DP with

$$\epsilon \le \frac{\beta^2}{4\sigma_2^2} + \frac{\beta\sqrt{2\log(1/\delta)}}{\sqrt{2}\sigma_2}.$$

Improvements over "Analyze Gauss". Notice that if we choose σ_2 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 6 is equivalent to releasing the vector $[f_1, f_2]$ using a standard Gaussian mechanism with $\mathcal{N}(0, \sigma_2^2 I_{\frac{d^2}{2}+d/2})$, 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 L2-sensivitiy 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. □

C.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 18 (Largest singular value of Gaussian random matrix (Rudelson & 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 $t > 0$*

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

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

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

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 19 (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 6.

Lemma 20. *Let A be described as in Lemma 19.*

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

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 $\tau \approx 12$.

D Omitted Proofs

With the two technical components presented, we are now ready to present the detailed proofs of our mains results: Theorem 6 and Theorem 10.

D.1 Proofs for the pDP analysis of objective perturbation

Proof of Theorem 6. We calculate the *ex-post* pDP loss of Algorithm 1 as follows. Consider the perturbed objective function:

$$\hat{\theta}^P = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \hat{\mathcal{L}}(\theta; D) + r(\theta) + \frac{\lambda}{2}||\theta||_2^2 + b^T\theta. \tag{2}$$

Let $\mathcal{A}$ be the algorithm that outputs $\hat{\theta}^P$ as stated in 2. The *ex-post* per-instance privacy loss (with the abuse of notation discussed in Section 2.1) is then given by

$$\epsilon_1(\theta, D, D_{\pm z}) = \max\left(\log\frac{\Pr\big(\mathcal{A}(D) = \hat{\theta}^P\big)}{\Pr\big(\mathcal{A}(D_{\pm z}) = \hat{\theta}^P\big)}, \log\frac{\Pr\big(\mathcal{A}(D_{\pm z}) = \hat{\theta}^P\big)}{\Pr\big(\mathcal{A}(D) = \hat{\theta}^P\big)}\right),$$

Note that this characterization of *ex-post* per-instance DP is equivalent to that stated in Definition 3, since switching the numerator and denominator of a log ratio is the same as flipping its sign.

Since we can't easily calculate the distribution of $\hat{\theta}^P$, we will instead use the bijection between the output $\hat{\theta}^P$ and the noise vector b (observed in Chaudhuri et al. (2011)) to rewrite the log probability ratio more cleanly.

First-order conditions applied to (2) tell us that

$$b(\hat{\theta}^P; D) = -\left(\nabla\hat{\mathcal{L}}(\hat{\theta}^P; D) + \nabla r(\hat{\theta}^P) + \lambda\hat{\theta}^P\right). \tag{3}$$

Then taking the gradient of the noise vector, we have

$$\nabla b(\hat{\theta}^P; D) = -\left(\nabla^2\hat{\mathcal{L}}(\hat{\theta}^P; D) + \nabla^2 r(\hat{\theta}^P) + \lambda I_d\right). \tag{4}$$

Let $b \sim \mathcal{N}(0, \sigma^2 I_d)$, and denote $\nu(\cdot)$ as the probability density function of the normal distribution: i.e., the density at b is $\nu(b;\sigma) \propto e^{-\frac{||b||^2}{2\sigma^2}}$. Then since the objective function $J(\theta; D)$ is strictly convex in θ (implying as in Chaudhuri et al. (2011) that the mapping between $\hat{\theta}^P$ and b is bijective and monotonic), by Lemma 30 we can write

$$\begin{aligned}
\log\frac{\Pr\big(\mathcal{A}(D) = \hat{\theta}^P\big)}{\Pr\big(\mathcal{A}(D_{\pm z}) = \hat{\theta}^P\big)} &= \log\frac{\left|\det\Big(\nabla b(\hat{\theta}^P; D)\Big)\right|}{\left|\det\Big(\nabla b(\hat{\theta}^P; D_{\pm z})\Big)\right|}\frac{\nu(b(\hat{\theta}^P; D); \sigma)}{\nu(b(\hat{\theta}^P; D_{\pm z}); \sigma)} \\
&= \log\frac{\left|\det\Big(\nabla b(\hat{\theta}^P; D)\Big)\right|}{\left|\det\Big(\nabla b(\hat{\theta}^P; D_{\pm z})\Big)\right|} + \log\frac{e^{-\frac{1}{2\sigma^2}||b(\hat{\theta}^P; D)||_2^2}}{e^{-\frac{1}{2\sigma^2}||b(\hat{\theta}^P; D_{\pm z})||_2^2}} \\
&= \underbrace{\log\frac{\left|\det\Big(\nabla b(\hat{\theta}^P; D)\Big)\right|}{\left|\det\Big(\nabla b(\hat{\theta}^P; D_{\pm z})\Big)\right|}}_{(*)} + \underbrace{\frac{1}{2\sigma^2}\Big(||b(\hat{\theta}^P; D_{\pm z})||_2^2 - ||b(\hat{\theta}^P; D)||_2^2\Big)}_{(**)}.
\end{aligned}$$

Dealing first with the term (*), we observe that $\nabla b(\hat{\theta}^P; D_{\pm z}) = \nabla b(\hat{\theta}^P; D) \mp \nabla^2\ell(\hat{\theta}^P; z)$. The notation "$\mp$" means to subtract if $z \notin D$, and add if $z \in D$. Using the eigendecomposition $\nabla^2\ell(\hat{\theta}^P; z) = \sum_{k=1}^d \lambda_k u_k u_k^T$ and recursively applying the matrix determinant lemma, we have

$$
\begin{aligned}
\Big|\det\big(\nabla b(\hat{\theta}^P; D_{\pm z})\big)\Big| &= \Big|\det\Big(\nabla b(\hat{\theta}^P; D) \mp \nabla^2 \ell(\hat{\theta}^P; z)\Big)\Big| \\
&= \Big|\det\Big(\nabla b(\hat{\theta}^P; D) \mp \sum_{k=1}^{d} \lambda_k u_k u_k^T\Big)\Big| \\
&= \Big|\det\Big(\nabla b(\hat{\theta}^P; D) \mp \sum_{k=1}^{d-1} \lambda_k u_k u_k^T \mp \lambda_d u_d u_d^T\Big)\Big| \\
&= \Big|\det\Big(\nabla b(\hat{\theta}^P; D) \mp \sum_{k=1}^{d-1} \lambda_k u_k u_k^T\Big)\Big|\Big(1 \mp \lambda_d u_d^T \big(\nabla b(\hat{\theta}^P; D) \mp \sum_{k=1}^{d-1} \lambda_k u_k u_k^T\big)^{-1} u_d\Big) \\
&= \dots \\
&= \Big|\det\Big(\nabla b(\hat{\theta}^P; D)\Big)\Big| \prod_{j=1}^{d} (1 \mp \mu_j),
\end{aligned}
$$

where $\mu_j = \lambda_j u_j^T \Big(\nabla b(\hat{\theta}^P; D) \mp \sum_{k=1}^{j-1} \lambda_k u_k u_k^T\Big)^{-1} u_j$. Therefore,

$$
\begin{aligned}
(*) &= \log \frac{\Big|\det\big(\nabla b(\hat{\theta}^P; D)\big)\Big|}{\Big|\det\big(\nabla b(\hat{\theta}^P; D_{\pm z})\big)\Big|} \\
&= \log \frac{\Big|\det\big(\nabla b(\hat{\theta}^P; D)\big)\Big|}{\Big|\det\Big(\nabla b(\hat{\theta}^P; D)\Big)\Big| \prod_{j=1}^{d} (1 \mp \mu_j)} \\
&= \log \frac{1}{\prod_{j=1}^{d} (1 \mp \mu_j)} \\
&= -\log \prod_{j=1}^{d} (1 \mp \mu_j).
\end{aligned}
$$

We'll handle the second term (**) next. We have that

$$
\begin{aligned}
(**) &= \frac{1}{2\sigma^2}\Big(||b(\hat{\theta}^P; D_{\pm z})||^2 - ||b(\hat{\theta}^P; D)||^2\Big) \\
&= \frac{1}{2\sigma^2}\Big[\mp \nabla \ell(\hat{\theta}^P; z)\Big]\Big[2b(\hat{\theta}^P; D) \mp \nabla \ell(\hat{\theta}^P; z)\Big] \\
&= \pm \frac{1}{\sigma^2}\big[J(\hat{\theta}^P; D)^T \nabla \ell(\hat{\theta}^P; z)\big] + \frac{1}{2\sigma^2}||\nabla \ell(\hat{\theta}^P; z)||^2.
\end{aligned}
$$

The rest of the proof follows from adding together (*) and (**), and taking the absolute value. □

Proof of Corollary 7. By restricting to generalized linear models, we can give a more interpretable pDP result for Algorithm 1 with the main difference being a cleaner version of the generalized leverage score. In the case of GLMs, we have that $\nabla \ell(\cdot) = f'(\cdot)x$ and $\nabla^2 \ell(\cdot) = f''(\cdot)xx^T$. So $\ell(\cdot)$ has a rank-one Hessian with only one eigenvalue, and $\log \prod_{j=1}^{d} \Big(1 + \mu_j\Big) = \log(1 + \mu(x))$. Here μ_j is as defined in Theorem 6 and μ is defined as in Corollary 7. □

D.2 Proofs for the Privacy Report in the main paper

The proof of Theorem 10 relies on the following intermediate result.

Proposition 21 (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 $x \in \mathbb{R}^d$

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

In other word, $\frac{1}{2}\hat{\mu}^p_1(x) \leq \mu_1(x) \leq \frac{3}{2}\hat{\mu}^p_1(x)$. □

Proof of Theorem 10. The privacy guarantees (Statement 1-3) follow directly from the pDP analysis in Theorem 17 that analyzes the release of H by adding a GOE noise matrix.

By the result follows from Proposition 21 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

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

This concludes the factor 12 multiplicative approximation. □

E pDP Analysis of the Gaussian mechanism

Theorem 22 (*ex-post* pDP of Gaussian mechanism). *Let* $Q : \mathcal{Z}^* \to \mathbb{R}^d$ *be a function of the data. Let* $|Q(D_{\pm z}) - Q(D)| \leq \Delta_z$. *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 23 (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*

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

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*

$$\epsilon(o,D,D_{\pm z}) \le \frac{\|\Delta_z\|^2}{2\sigma^2}+\frac{\|\Delta_z\|\Phi^{-1}(1-\rho/2)}{\sigma} \le \frac{\|\Delta_z\|^2}{2\sigma^2}+\frac{\|\Delta_z\|\sqrt{2\log(2/\rho)}}{\sigma}. \tag{5}$$

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 27.

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 & Wang, 2018). We choose to present the tail bound-based formula above for the interpretability of the results.

F Technical Lemmas

Lemma 24 (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 25 (Determinant of Rank-1 perturbation)**.** *For invertible matrix A and vector c, d of compatible dimension*

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

Lemma 26 (Gaussian tail bound)**.** *Let $X \sim \mathcal{N}(0,\sigma^2)$. Then*

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

A convenient alternative representation (slightly weaker) is

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

and

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

for all $\delta > 0$.

Lemma 27 (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 28 (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 29 (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*

$$\max_{i \in [n]} |\sigma_i(A) - \sigma_i(A + E)| \leq \|E\|_2 \tag{6}$$

Lemma 30 ("Change-of-variables" for density functions)**.** *Let X be a continuous random variable with probability density function $f_X(x)$ for $x \in \mathbb{R}^d$. Consider monotonic transformation $y = g(x)$, where y has pdf $f_Y(y)$ and the function g is differentiable and strictly increasing or strictly decreasing. Then*

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

with $\left|\frac{\partial}{\partial y}g^{-1}(y)\right|$ denoting the Jacobian matrix of the mapping from y to x.