Privately Publishable Per-instance Privacy

Rachel Redberg
Department of Computer Science
UC Santa Barbara
Santa Barbara, CA 93106
rredberg@ucsb.edu

Yu-Xiang Wang
Department of Computer Science
UC Santa Barbara
Santa Barbara, CA 93106
yuxiangw@cs.ucsb.edu

Abstract

We consider how to privately share the personalized privacy losses incurred by objective perturbation, using per-instance differential privacy (pDP). 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.

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. 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 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!

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.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

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 paper, 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 2.

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. 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., 2018), which conditions on a fixed dataset; personalized DP (Ghosh & 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 & 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 by making distributional assumptions about the sensitive data. The Rényi-DP-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.

2 Preliminaries

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 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 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. (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 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 $\mathcal{A}$.*

Following the proof of Theorem 22, 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 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 $\hat{\theta}^P$ 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.

2.3 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 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 $\Theta = \mathbb{R}^d$.

2.4 Objective Perturbation

The objective perturbation algorithm solves

$$\hat{\theta}^P = \underset{\theta \in \Theta}{\operatorname{argmin}}\, L(\theta; D) + r(\theta) + \frac{\lambda}{2}||\theta||_2^2 + b^T\theta, \tag{1}$$

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 (1).

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

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. 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_1)-differential privacy.*

The privacy guarantees stated in Theorem 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.

3 Privately Publishable pDP

3.1 pDP Analysis of Objective Perturbation

Our goal in this section is to derive the personalized privacy losses (under Definition 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 6 (*ex-post* pDP loss of objective perturbation for a convex loss function)**.** *Let $J(\theta; D) = L(\theta; D) + r(\theta) + \frac{\lambda}{2}||\theta||_2^2$ 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 $\hat{\theta}^P = \operatorname{argmin}_{\theta \in \mathbb{R}^d} J(\theta; D) + b^T\theta$ satisfies $\epsilon_1(\hat{\theta}^P, D, D_{\pm z})$-ex-post per-instance differential privacy with*

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

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 E. □

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 $J(\theta; D) = L(\theta; D) + r(\theta) + \frac{\lambda}{2}||\theta||_2^2$ 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 $\hat{\theta}^P = \operatorname{argmin}_{\theta \in \mathbb{R}^d} J(\theta; D) + b^T\theta$ satisfies $\epsilon_1(\hat{\theta}^P, D, D_{\pm z})$-ex-post per-instance differential privacy with*

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

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.

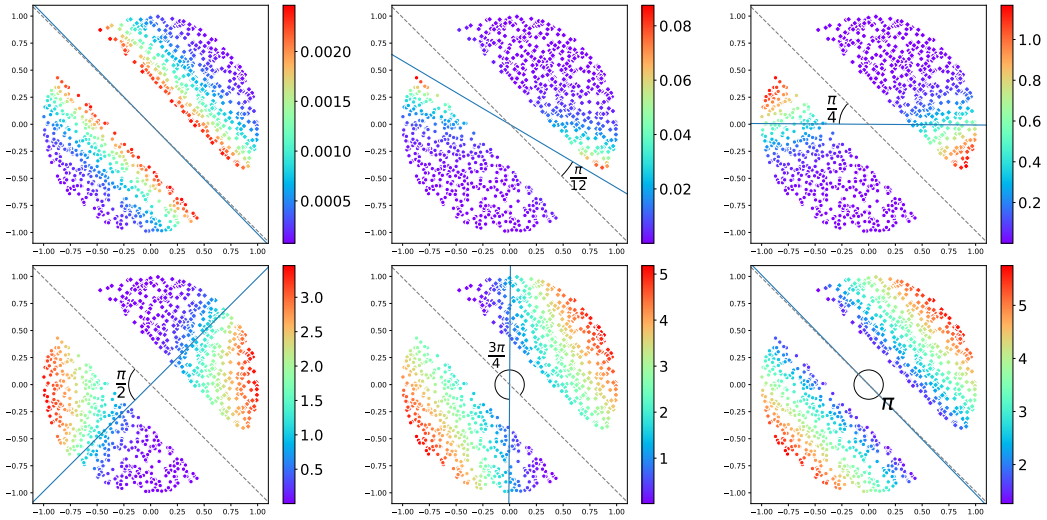


Figure 1: Visualization of *ex-post* pDP losses for logistic regression ($n = 1000, d = 2$).

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 $\hat{\theta} = \operatorname{argmin} J(\theta; D)$ with $\lambda = 1$ to minimize the logistic loss, and directly perturb the output by rotating it by angle $\omega \in [0, \frac{\pi}{12}, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \pi]$. 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 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 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!

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 6 depends on the dataset D only through two quantities: the leverage score $\mu(x) = x^T \left(\nabla^2 J(\hat{\theta}^P; D)\right)^{-1} x$ 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. *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 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 $-\log\Big(1 - f''(x^T\hat{\theta}^P; y)||x||_2^2/\lambda\Big)$.

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

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

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 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!

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 $f''(\cdot) := f''(x^T\hat{\theta}^P; y)$, 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 5 with parameter σ_2.
  Set $\epsilon_2(\cdot)$ according to Theorem 14.
  Set $\overline{g^P}(z) := \min\left\{ f'(\cdot)[\hat{g}^P(z)]^T x + \sigma_2 ||f'(\cdot)x||_2 F^{-1}_{\mathcal{N}(0,1)}(1-\rho/2),\ \sigma ||f'(\cdot)x||_2 F^{-1}_{\mathcal{N}(0,1)}(1-\rho/2) \right\}$.
  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 10.
  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 released pDP functions depend on $\hat{\theta}^P$; to reduce clutter, we omit this parameter in our presentation of Algorithm 2.

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. $(\frac{\xi^2}{2\sigma_2^2} + \frac{\beta^2}{4\sigma_3^2} + \sqrt{\frac{\xi^2}{\sigma_2^2} + \frac{\beta^2}{2\sigma_3^3}}\sqrt{2\log(1/\delta)}, \delta)$*-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 6.*

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 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 $\{\tilde{\epsilon}(z_1, \hat{\theta}^p), ..., \tilde{\epsilon}(z_1, \hat{\theta}^p)\}$

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

satisfies joint DP with the same ϵ parameter as in Theorem 10. This follows from the billboard lemma (Hsu et al., 2016).

4 Experiments

Here we evaluate our methods to release the pDP losses using logistic regression as a case study. In Section 4.1, we demonstrate that the stronger regularization required by Algorithm 2 does not affect the utility of the model. In Section 4.2, we compare the data-independent and data-dependent approaches of Algorithm 2 under two different privacy budget allocation schemes.

Experiments with linear regression and with additional datasets are included in the supplementary materials.

4.1 Stronger regularization does not worsen model utility

In this section 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 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. Since each value of ϵ_1 demands a different minimum value of λ in order to achieve (ϵ_1, δ)-differential privacy, we compare via "λ-inflation": a measure of how many times larger we set λ than its minimum value required to achieve the worst-case DP bound of objective perturbation.

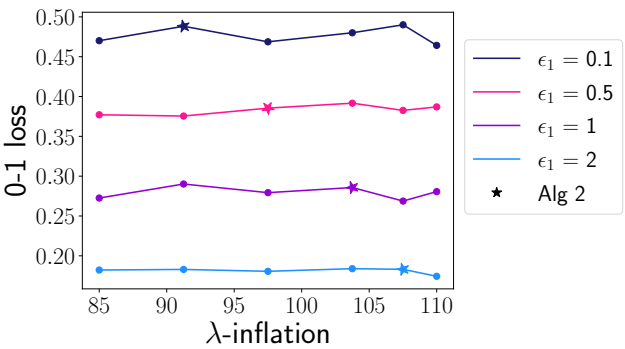


Figure 2: Utility of Obj-Pert with larger λ.

E.g., for logistic regression the objective perturbation mechanism requires $\lambda \geq \frac{1}{2\epsilon}$, and so in Figure 2 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. The experimental results summarized in Figure 2 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.

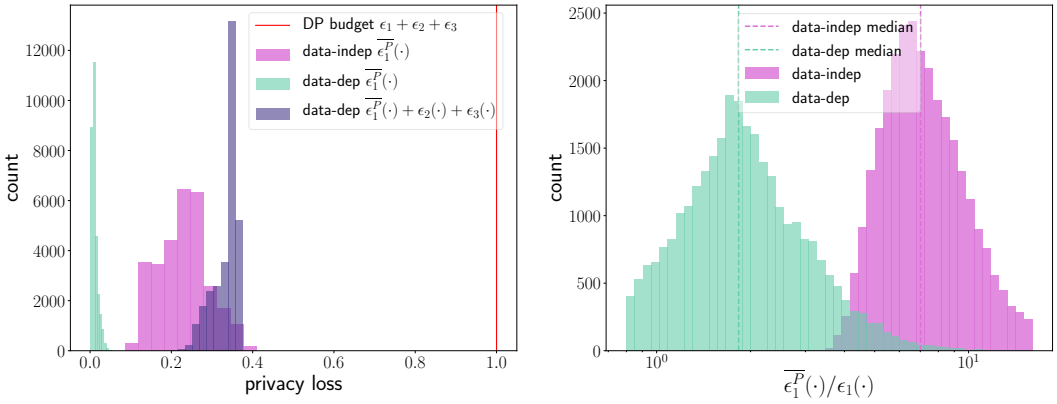
4.2 Comparison of data-independent and data-dependent bounds

The following experiments feature the credit card default dataset ($n = 30000, d = 21$) (Yeh & 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 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 & Wang, 2018).

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 6, 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 3b compares the accuracy of both approaches using the ratio between $\overline{\epsilon_1^P}(\cdot)$ and $\epsilon_1(\cdot)$ rather than their raw values.



(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 3: 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 $\overline{g^P}(\cdot)$ ($\epsilon_1 = 0.2, \epsilon_2 = 0.7, \epsilon_3 = 0.1$).

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 $\hat{\theta}^P$ 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 3a), we can release the *ex-post* pDP losses of Algorithm 1 much more accurately (Figure 3b). 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 $\overline{\epsilon_1^P}(\cdot)$.

5 Conclusion

We precisely calculate the privacy loss that an individual z suffers *after* the objective perturbation algorithm is run on a specific dataset. The *ex-post* pDP loss function in DP-ERM can be accurately released to all individuals with little or no additional privacy cost. In particular, we present a *data-independent* bound which empirically provides a reasonably accurate estimate of the *ex-post* pDP loss while requiring no further privatization step. Reserving some of the privacy budget allows us to alternatively release a tighter *data-dependent* bound.

An important next step is to extend the per-instance DP analysis to the setting of constrained optimization. There are many promising future directions including using publishable pDP losses for designing more data-adaptive DP algorithms.

Acknowledgments

The work was partially supported by NSF CAREER Award # 2048091, Google Research Scholar Award and a gift from Evidation Health.

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.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

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.

Ebadi, H., Sands, D., and Schneider, G. Differential privacy: Now it's getting personal. *ACM Sigplan Notices*, 50(1):69–81, 2015.

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

Ghosh, A. and Roth, A. Selling privacy at auction. In *Proceedings of the 12th ACM conference on Electronic commerce*, pp. 199–208, 2011.

Hsu, J., Huang, Z., Roth, A., Roughgarden, T., and Wu, Z. S. Private matchings and allocations. *SIAM Journal on Computing*, 45(6):1953–1984, 2016.

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.

Liu, Z., Wang, Y.-X., and Smola, A. Fast differentially private matrix factorization. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pp. 171–178, 2015.

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.

Yeh, I.-C. and Lien, C.-h. 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.

Yu, D., Zhang, H., Chen, W., Yin, J., and Liu, T. Large scale private learning via low-rank reparametrization. In *International Conference on Machine Learning (ICML-21)*, volume 139, pp. 12208–12218. PMLR, 2021.

Part
Appendix

Table of Contents

A DP Variants **14**

B Additional Experiments **15**
- B.1 Varying dimension and dataset size 15
- B.2 Privacy budget allocation 15
- B.3 Comparison of pDP losses and private upper bounds 17

C Even Stronger Privacy Report **19**
- C.1 More Accurate Privacy Report by Adapting to the Data 19
- C.2 Dataset-Dependent Privacy report for general smooth learning problems 22
- C.3 Uniform Privacy Report and Privacy Calibration 23

D Improved "Analyze Gauss" with Gaussian Orthogonal Ensembles **24**
- D.1 Exact statistical inference with the Gaussian Orthogonal Ensemble 25

E Omitted Proofs **26**
- E.1 Proofs for the pDP analysis of objective perturbation 26
- E.2 Proofs for the Privacy Report in the main paper 29

F pDP Analysis of the Gaussian mechanism **30**

G Technical Lemmas **31**

A 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 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	$\sup_{D, D' : D' \simeq_z D} \max\Big(D_\infty^\delta\big(\mathcal{A}(D) \| \mathcal{A}(D')\big), D_\infty^\delta\big(\mathcal{A}(D') \| \mathcal{A}(D)\big)\Big) \leq \epsilon(z)$
Per-instance DP	$\max\Big(D_\infty^\delta\big(\mathcal{A}(D) \| \mathcal{A}(D')\big), D_\infty^\delta\big(\mathcal{A}(D') \| \mathcal{A}(D)\big)\Big) \leq \epsilon(D, z)$
***Ex-post* per-instance DP**	$\left\lvert \log \frac{\Pr\big[\mathcal{A}(D) = o\big]}{\Pr\big[\mathcal{A}(D') = o\big]} \right\rvert \leq \epsilon(o, D, D')$ where $D' \simeq_z D$

B Additional Experiments

B.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 4 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 4 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 4 and 5 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.

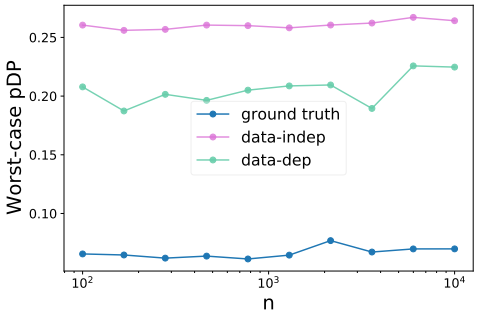


Figure 4: Worst-case pDP while varying n.

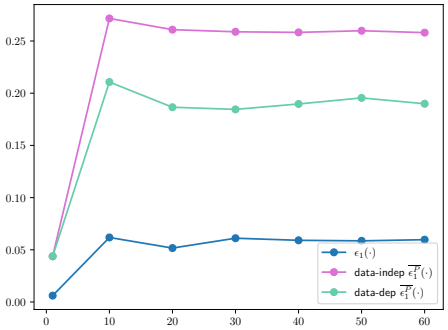


Figure 5: Worst-case pDP while varying d.

B.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 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 Figure 6 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 $\epsilon_3^{DEP} = 1$, 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 $\hat{\theta}^P$.

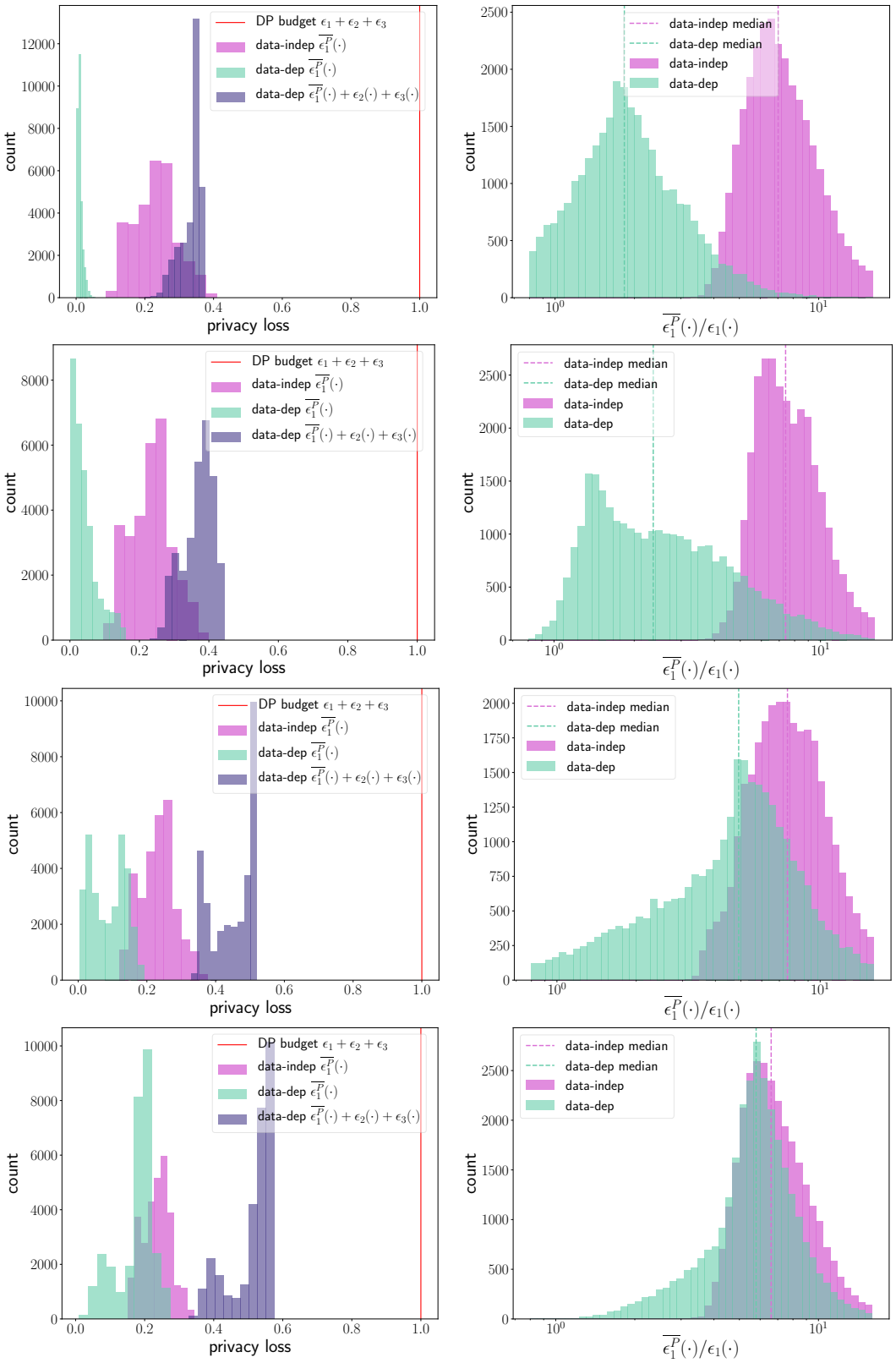


Figure 6: Data-independent release uses a privacy budget $\epsilon_1 = 1$ for each plot. From top to bottom, the budgets for the data-dependent release are $\epsilon_1 = 0.2, \epsilon_2 = 0.7, \epsilon_3 = 0.1$; $\epsilon_1 = 0.4, \epsilon_2 = 0.5, \epsilon_3 = 0.1$; $\epsilon_1 = 0.5, \epsilon_2 = 0.25, \epsilon_3 = 0.25$; and $\epsilon_1 = 0.8, \epsilon_2 = 0.1, \epsilon_3 = 0.1$.

B.3 Comparison of pDP losses and private upper bounds

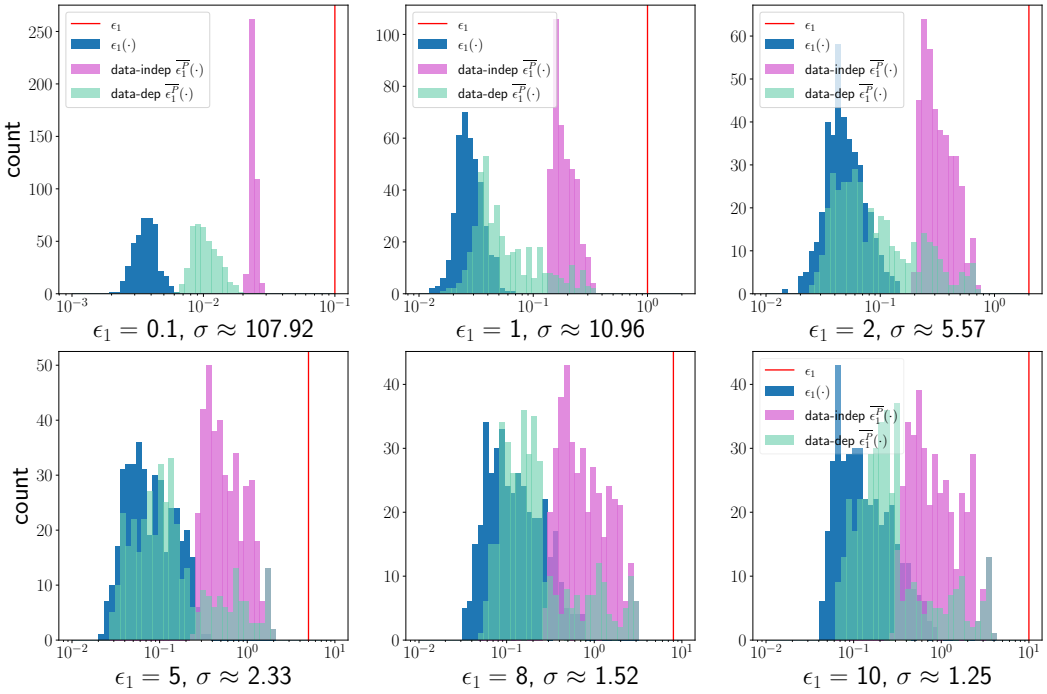


Figure 7: pDP losses $\epsilon_1(\cdot)$ and upper bounds $\overline{\epsilon_1^P}(\cdot)$ for private logistic regression applied to the UCI kidney dataset.

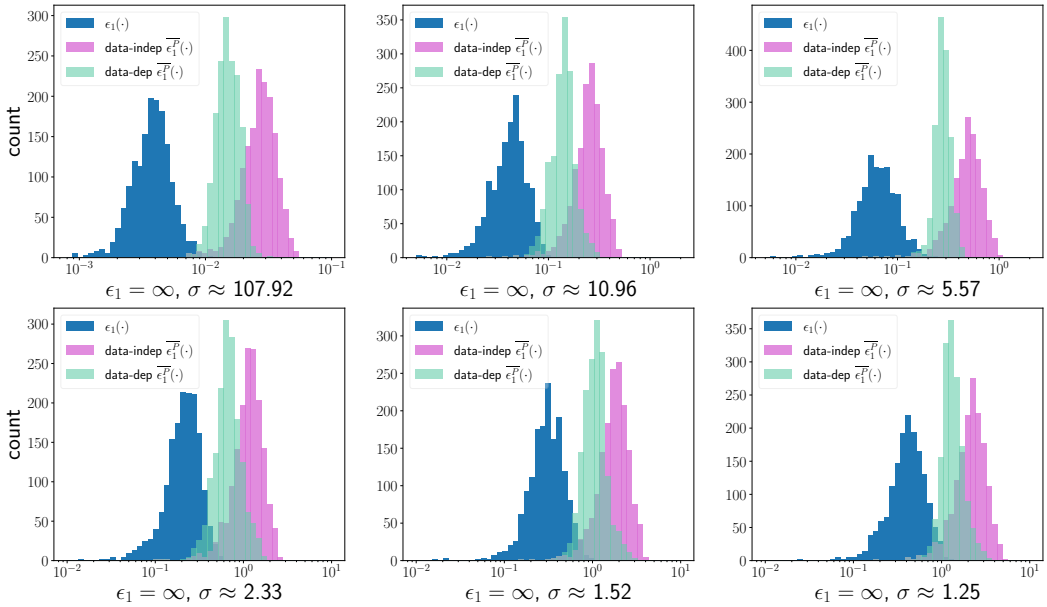


Figure 8: pDP losses $\epsilon_1(\cdot)$ and upper bounds $\overline{\epsilon_1^P}(\cdot)$ for private linear regression applied to the UCI wine quality dataset. Since we are dealing with an unbounded domain $\mathbb{R}^d$, the algorithm does not satisfy worst-case DP for any $\epsilon < \infty$.

We run both the data-independent and -dependent variations of Algorithm 1 as described in the experimental setup. Figures 7 and 8 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 7 shows results for private logistic regression on the UCI kidney dataset; Figure 8 shows results for private linear regression on the UCI wine quality dataset (Dua & Graff, 2017). Our experimental results indicate that for smaller $\epsilon_1 << 1$ (larger σ), the data-dependent approach provides a markedly tighter bound on $\epsilon_1()\cdot$.

Figures 9 and 10 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 9) and linear regression on the UCI wine quality dataset (Figure 10), the data-dependent approach provides a more accurate estimate of the pDP loss $\epsilon_1(\cdot)$, especially for logistic regression on the kidney dataset.

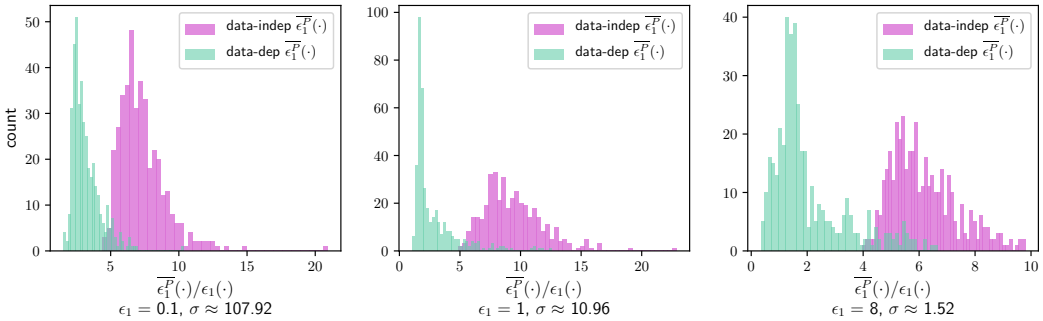


Figure 9: Ratio of private upper bound $\overline{\epsilon_1^P}(\cdot)$ to actual pDP loss $\epsilon_1(\cdot)$ for private logistic regression applied to the UCI kidney dataset.

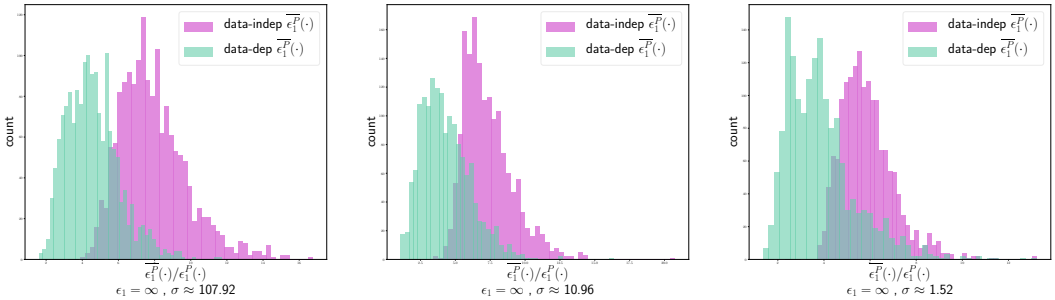


Figure 10: Ratio of private upper bound $\overline{\epsilon_1^P}(\cdot)$ to actual pDP loss $\epsilon_1(\cdot)$ for private linear regression applied to the UCI wine quality dataset.

C Even Stronger Privacy Report

C.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_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 11. *Algorithm 3 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 3 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 3 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 4 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 5 with parameter σ_2.
Set $\epsilon_2(\cdot)$ according to Theorem 14.
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 6 with parameter σ_3.
Set $\epsilon_3(\cdot)$ according to Theorem 17
Privately release $\hat{\lambda}^p_{\min} = \lambda_{\min} + \mathcal{N}(0, \sigma_4^2)$ (Algorithm 3).
Set $\epsilon_4(\cdot)$ according to Theorem 11.
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 $\overline{\epsilon^P_1}(\cdot) := \left| -\log\left(1 - f''(\cdot)\overline{\mu^P}(x)\right)\right| + \frac{||f'(\cdot)x||_2^2}{2\sigma^2} + \frac{|\overline{g^P}(z)|}{\sigma^2}$.
Return the "privacy report" function $\tilde{\epsilon} = (\overline{\epsilon^P_1}, \epsilon_2 + \epsilon_3 + \epsilon_4)$, 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)**.** *Let $\underline{\hat{\lambda}^p_{\min}}$ and $\hat{H}^P$ be constructed as in Algorithm 4. Then with probability $1 - 2\rho$,*

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

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

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

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

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

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 $1 - \rho$,

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

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 $1-2\rho$,

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

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:

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

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 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 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 13 (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 $\overline{\epsilon^p}_1(\cdot)$ from Algorithm 4 obeys that*

$$\epsilon_1(\cdot) \leq \overline{\epsilon^p}_1(\cdot) \leq (1 + \frac{C}{\lambda_{\min}})\epsilon_1(\cdot) + \frac{|f'(\cdot)|\|x\|}{\sigma_2}\sqrt{2\log(2/\rho)}$$

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

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

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. □

C.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 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_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. $\epsilon_2(z) = \frac{\|\nabla \ell_z(\hat{\theta}^P; z)\|^2}{2\sigma_2^2} + \frac{\|\nabla \ell_z(\hat{\theta}^P; z)\| \sqrt{2\log(2/\delta)}}{\sigma_2}$.

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

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

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

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

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

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

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 5's pDP loss is dataset-independent and if we choose $\sigma_2 = \sigma_1$, the pDP losses for running Algorithm 5 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.

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

C.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 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 4 is straightforward and thus omitted.

Proposition 15 (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 $\frac{|f'(z,\hat{\theta}^p)|\|x\|_1\sqrt{2\log(2d/\rho)}}{\sigma}$ 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 $\overline{\epsilon^p}_1$ from Algorithm 2 satisfies that*

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

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 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 $\frac{(f'')\|x\|^2}{\rho_{\max}} \leq \frac{|f'|\|x\|}{\beta}$ (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 6.

D 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 6 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 17. *Algorithm 6 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 $\sup_{x\in\mathcal{X}} \|H_x\|_F \le \beta$ then, Algorithm 6 satisfies (ϵ, δ)-DP with

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

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 6 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. □

D.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 $t > 0$*

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

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 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.*

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

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 $\tau \approx 12$.

E Omitted Proofs

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

E.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 $\nu(b; \sigma) \propto e^{-\frac{||b||_2^2}{2\sigma^2}}$. 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

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

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,

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

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

$$
\begin{aligned}
(**) &= \frac{1}{2\sigma^2}\Big(||b(\hat{\theta}^P; D_{\pm z})||_2^2 - ||b(\hat{\theta}^P; D)||_2^2\Big) \\
&= \frac{1}{2\sigma^2}\Big[\mp \nabla \ell(\hat{\theta}^P; z)\Big]\Big[2b(\hat{\theta}^P; D) \mp \nabla \ell(\hat{\theta}^P; z)\Big] \\
&= \pm \frac{1}{\sigma^2}\big[\nabla J(\hat{\theta}^P; D)^T \nabla \ell(\hat{\theta}^P; z)\big] + \frac{1}{2\sigma^2}||\nabla \ell(\hat{\theta}^P; z)||_2^2.
\end{aligned}
$$

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. □

Proof of Theorem 8. 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

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

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

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 9. By Holder's inequality,

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

Recall from (3) 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 $||\nabla\ell(\hat{\theta}^P)||_1 = f'(x^T\theta; y)||x||_1$. □

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

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 and the Gaussian mechanism that releases g.

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 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 $1-\rho/2$. □

F 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*

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

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*

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

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.

G 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*

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

Lemma 26 (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 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.

Lemma 31. (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.*