Journal of Privacy and Confidentiality
Vol. 9 (1) 2019 TPDP 2016
Submitted June 2, 2018
Published March 2019

# PER-INSTANCE DIFFERENTIAL PRIVACY

YU-XIANG WANG

UC Santa Barbara, Santa Barbara, CA 93106

*e-mail address*: yuxiangw@cs.ucsb.edu

Abstract. We consider a refinement of differential privacy — per instance differential privacy (pDP), which captures the privacy of a specific individual with respect to a fixed data set. We show that this is a strict generalization of the standard DP and inherits all its desirable properties, e.g., composition, invariance to side information and closure to postprocessing, except that they all hold for every instance separately. When the data is drawn from a distribution, we show that moments of per-instance DP imply generalization. Moreover, we provide explicit calculations of the per-instance DP for the output perturbation on a class of smooth learning problems. The result reveals an interesting and intuitive fact that an individual has stronger privacy if he/she has small "leverage score" with respect to the data set and if he/she can be predicted more accurately using the leave-one-out data set. Simulations show several orders-of-magnitude more favorable privacy and utility trade-off when we consider the privacy of only the users in the data set. In a case study on differentially private linear regression, we provide a novel analysis of the One-Posterior-Sample (OPS) estimator and show that when the data set is well-conditioned it provides $(\epsilon, \delta)$-pDP for any target individuals and matches the exact lower bound up to a $1 + \tilde{O}(n^{-1}\epsilon^{-2})$ multiplicative factor. We also demonstrate how we can use a "pDP to DP conversion" step to design AdaOPS which uses adaptive regularization to achieve the same results with $(\epsilon, \delta)$-DP.

## 1. Introduction

While modern statistics and machine learning had seen amazing success, their applications to sensitive domains involving personal data remain challenging due to privacy issues. Differential privacy (DP) [Dwork et al., 2006] is a mathematical definition that provides strong provable protection to individuals and prevents them from being identified by an arbitrarily powerful adversary. DP has been increasingly popular within the machine learning community as a solution to the aforementioned problem [McSherry and Mironov, 2009, Chaudhuri et al., 2011, Liu et al., 2015, Abadi et al., 2016]. The strong privacy protection, however, comes with a steep price to pay. Differential privacy often leads to a substantial

*Key words and phrases:* differential privacy, data-dependent, personalized, adaptive, leverage score.
The work was partially written while the author was a PhD student at Carnegie Mellon University.

JPrivacy Confidentiality
www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.662

and sometimes unacceptable drop in utility, e.g., in contingency tables [Fienberg et al., 2010] and in genome-wide association studies [Yu et al., 2014]. This motivated a large body of research to focus on making differential privacy more practical [Nissim et al., 2007, Dwork and Lei, 2009, Sheffet, 2017, Wang et al., 2015, Dwork and Rothblum, 2016, Bun and Steinke, 2016, Foulds et al., 2016] by exploiting local structures and/or revising the privacy definition.

The majority of these approaches adopt a "privacy-centric" model, which involves theoretically proving that an algorithm is differentially private for any data sets (within a data domain), then carefully analyzing the utility of the algorithm under additional assumptions on the data set. For instance, in statistical estimation, it is often assumed that the data is drawn i.i.d. from a family of distributions. In nonparametric statistics and statistical learning, the data are often assumed to having specific deterministic/structural conditions, e.g., smoothness, incoherence, eigenvalue conditions, low-rank, sparsity and so on. While these assumptions are strong and sometimes unrealistic, they are often necessary for a model to work correctly, even without privacy constraints. Take high-dimensional statistics for example, "sparsity" is never really true in applications. However, if the true model is dense and unstructured, then information-theoretically no method can estimate the true model in the "small $n$ large $d$" regime. That is why Friedman et al. [2001] argued that one should "bet on sparsity" regardless and only expect the method to work when the input data is well-approximated by a sparse model. Behind this informal "bet on sparsity" principle, is the pursuit for *adaptivity* in modeling and algorithm design. We say an algorithm is *adaptive*[1] if it can automatically *adapt* to favorable properties of each input data set and perform better.

These conditions on the data set can also have a profound impact on a DP algorithm's privacy guarantee. As we know, DP quantifies its privacy guarantee with a single nonnegative number $\epsilon$ — the privacy loss. Smaller $\epsilon$ implies stronger privacy guarantee. DP algorithms are designed to calibrate itself according to a prescribed budget of $\epsilon$, and to achieve $\epsilon$-DP (or $(\epsilon, \delta)$-DP) regardless of what the input data is. This concise data-independent privacy loss $\epsilon$ is one of DP's most attractive feature as it makes DP universally applicable. But in many real-world applications, $\epsilon$ is often an *overly simplified summary* and a *crude upper bound* of the *actual* privacy loss incurred to individuals in the data set. An algorithm $\mathcal{A}$ that is calibrated to achieve a privacy loss of $\epsilon = 10$ on the worst pair of adjacent data sets, could imply a much stronger privacy level of $\epsilon' = 0.1$ when $\mathcal{A}$ is applied to a particular data set in which some of the additional assumptions are true. In this case, it will be too conservative to quantify the actual privacy loss with just $\epsilon = 10$.

The extent to which DP is conservative, however, is highly problem-dependent. In cases such as releasing counting queries, DP's $\epsilon$ clearly measures the correct information leakage, since the sensitivity of such queries do not change with respect to the two adjacent data sets; however, in the context of machine learning and statistical estimation (as we will show later), the $\epsilon$ of DP can be orders of magnitude larger than the actual amount of information leakage

[1]This is not to be confused with the "adaptivity" as in "adaptive composition"[Dwork et al., 2010] and "adaptive data analysis"[Dwork et al., 2015] commonly seen in the differential privacy literature. The latter is about how a sequence of actions can be chosen as a function of the outcomes to all previous actions, while the notion of "adaptivity" that we considered is the same as that in "adaptive estimation", "adaptive algorithm design" [see, e.g., Bickel, 1982], which is about the extent to which algorithms can exploits properties in the data sets without knowing that they exists a priori.

that comes with the release. That is part of the reason why in practice, it is challenging even for experts of differential privacy to provide a consistent recommendation on standard questions such as:

"What is the value of privacy budget $\epsilon$ I should set in my application?"

In this paper, we take a new "algorithm-centric" approach of analyzing privacy and address a related but different question:

"What is the privacy loss $\epsilon$ incurred to an individual $z$ when $\mathcal{A}(Z)$ is released?"

Instead of designing algorithms that take the privacy budget $\epsilon$ as an input, we start with a fixed randomized algorithm $\mathcal{A}$[2] and then analyze its privacy protection for every pair of adjacent data sets separately. This is equivalent to treating $\epsilon$ as a function parameterized by each problem instance — a Dataset-Target pair. $\epsilon$ as a function provides a more fine-grained description of the randomized algorithm compared to using only the privacy loss $\epsilon$ of its DP guarantee, i.e., the maximum of $\epsilon(\text{Dataset}, \text{Target})$ over all pairs of datasets and individuals.

Our contribution is threefold.

(1) First, we develop per-instance differential privacy as a strict generalization of the standard pure and approximate DP. It provides a more fine-grained description of the privacy protection for each target individual and a fixed data set. We show that it inherits many desirable properties of differential privacy and can easily recover differential privacy for a given class of data and target users.

(2) Secondly, we quantify the per-instance sensitivity in a class of smooth learning problems including linear and kernel machines. The result allows us to explicitly calculate per-instance DP of a multivariate Gaussian mechanism. For an appropriately chosen noise covariance, the per-instance DP is proportional to the norm of the pseudo-residual in norm specified by the Hessian matrix. In particular, in linear regression, the per-instance sensitivity for a data point is proportional to its square root statistical leverage score (predictive variance) and its leave-one-out prediction error (predictive bias).

(3) Lastly, we analyze the procedure of releasing one sample from the posterior distribution (the OPS estimator) for linear and ridge regression as an output perturbation procedure with a data-dependent choice of the covariance matrix. We show using the pDP technique that, when conditioning on a data set drawn from the linear regression model or having a well-conditioned design matrix, OPS achieves $(\epsilon, \delta)$-pDP for while matching the Cramer-Rao lower bound up to a $1+\tilde{O}(n^{-1}\epsilon^{-2})$ multiplicative factor. OPS, unfortunately, cannot achieve DP with a constant $\epsilon$ while remaining asymptotically efficient. We fixed that by a new algorithm called AdaOPS, which provides $(\epsilon, \delta)$-DP and $1+\tilde{O}(n^{-1}\epsilon^{-2})$-statistical efficiency at the same time.

To avoid any confusion, we also highlight a few things that this paper is *not* about. First of all, pDP is *not* a replacement of DP. It is rather an analytical tool for us to understand the adaptivity in privacy loss and to design more data-dependent DP algorithms. For instance, you can use pDP to describe the *actually incurred privacy loss* to an individual when an $\epsilon$-DP algorithm is applied to a given data set, rather than just covering everything

[2] $\mathcal{A}$ can be a DP algorithm but it does not have to be.

under a blanket statement that: "All I know is that it's smaller than $\epsilon$." Secondly, for output perturbation algorithms, we do *not* advocate calibrating the noise to per-instance sensitivity for setting pDP to a prescribed budget, because the per-instance sensitivity itself is a data-dependent quantity. This is not an intended use for pDP. In fact, if you insist on doing that, then you essentially changed the algorithm all together and the corresponding per-instance sensitivity will change as well. Thirdly, pDP privacy loss is something that the curator can calculate and keep as a confidential certificate, but *not* shared publicly or even with data contributors, since pDP itself contains private information about the entire data set. Publishing pDP differentially privately is an important problem and it is part of an ongoing future work.

1.1. **Symbols and notations.** Throughout the paper, we will use the standard notation in statistical learning. Data point $z \in \mathcal{Z}$. In supervised learning setting, $z = (x, y) \in \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$. We use $\theta \in \Theta$ to denote either the predictive function $\mathcal{X} \to \mathcal{Y}$ or the parameter vector that specifies such a function. $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$ to denote the loss function or in a statistical model, $\ell$ represents the negative log-likelihood $-\log p_\theta(z)$. For example, in linear regression, $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y} \subset \mathbb{R}$, $\Theta \subset \mathbb{R}^d$ and $\ell(\theta, (x, y)) = (y - x^T\theta)^2$. Capital $Z$ denotes a data set of an unspecified size, i.e., $Z \in \mathcal{Z}^* = \cup_{n=0,1,2,3,...} \mathcal{Z}^n$. We use $\mathcal{A} : \mathcal{Z}^* \to P_\Theta$ to denote a randomized algorithm that takes in a data set and outputs a draw from a distribution defined on a model space. In particular, $\mathcal{A}(Z)$ is used to denote both a random variable and its distribution, so that we can say $\theta \sim \mathcal{A}(Z)$. $\epsilon$ and $\epsilon(Z, z)$ will be reserved to denote privacy loss, and $Z, Z' \in \mathcal{Z}^*$ are reserved to denote the two adjacent data set. In particular, unless we specify otherwise, $Z'$ will be either adding $z$ to $Z$ or removing $z$ from $Z$ depending on whether $Z$ contains $z$ or not. The notation $Z \overset{z}{\sim} Z'$ is used to explicitly say that $Z$ and $Z'$ differ by a single data point $z$.

1.2. **Related work.** This paper is related to the existing work in relaxing DP[Hall et al., 2013, Barber and Duchi, 2014, Wang et al., 2016, Dwork and Rothblum, 2016, Bun and Steinke, 2016, Mironov, 2017], personalizing DP[Ghosh and Roth, 2015, Ebadi et al., 2015, Liu et al., 2015], post hoc calculation of privacy guarantee [Abadi et al., 2016, Rogers et al., 2016, Ligett et al., 2017, Balle and Wang, 2018], as well as in analytical frameworks for designing data-adaptive DP algorithms [Nissim et al., 2007, Dwork and Lei, 2009]. We provide more details below.

**Relaxing and personalizing DP.** Recall that the pure differential privacy loss can be defined as

$$\epsilon = \sup_{Z \overset{z}{\sim} Z'} \sup_{\theta \in \Theta} \log \frac{p_{\mathcal{A}(Z)}(\theta)}{p_{\mathcal{A}(Z')}(\theta)}.$$

The effort in relaxing differential privacy mostly consider relaxing the $\sup_\theta$ part of the definition, by either using a different divergence measure [Barber and Duchi, 2014] or explicitly treating $\epsilon(\theta) = \log \frac{p_{\mathcal{A}(Z)}(\theta)}{p_{\mathcal{A}(Z')}(\theta)}$ as a random variable induced by $\theta \sim \mathcal{A}(Z)$. The privacy random variable point of view connects $(\epsilon, \delta)$-DP to concentration inequalities and in particular, it produces the advanced composition of privacy losses via Martingale concentration [Dwork et al., 2010]. More recently, the idea is extended to define weaker notions of privacy such as concentrated-DP [Dwork and Rothblum, 2016, Bun and Steinke,

2016] and Rényi-DP [Mironov, 2017]. They allow for tighter accounting of the privacy losses through the moment generating function of the privacy random variable.

Our work is complementary to this line of work, as we relax the $\sup_{Z \overset{z}{\sim} Z'}$ part of the definition and consider the adaptivity of $\epsilon$ to a fixed pair of data set $Z$ and privacy target $z$. In some cases, we consider $\epsilon$ to be a random variable jointly parameterized by $Z, z$ and $\theta$.

The closest existing definition to ours is perhaps the personalized-DP, first seen in Ghosh and Roth [2015] for the problem of selling privacy in auctions and reinvented by Ebadi et al. [2015], Liu et al. [2015] in the context of private database queries and private recommendation systems respectively. They also try to capture a personalized level of privacy for each individual $z$. The difference is that personalized-DP considers adding or removing $z$ from all data sets, while we consider adding $z$ or removing $z$ from a fixed $Z$.

Finally, pDP is related to random differential privacy [Hall et al., 2013] and on-average KL-privacy [Wang et al., 2016]. They respectively measure the high-probability and expected privacy loss when $z$ and the data points in $Z$ are drawn i.i.d. from a distribution $\mathcal{D}$, while we consider a fixed $(Z, z)$ pair that is not necessarily random.

We summarize these definitions in Table 1. It is clear from the table that if we ignore the differences in the probability metric used, per-instance DP is arguably the most general, and adaptive, since it depends on specific $(Z, z)$ pairs.

Table 1: Comparing variants of differential privacy.

| | Data set | private target | probability metric | parametrized by |
|---|---|---|---|---|
| Pure-DP | $\sup_Z$ | $\sup_z$ | $D_\infty(P\|Q)$ | $\mathcal{A}$ only |
| Approx-DP | $\sup_Z$ | $\sup_z$ | $D_\infty^\delta(P\|Q)$ | $\mathcal{A}$ only |
| (z/m)-CDP | $\sup_Z$ | $\sup_z$ | $D_{\text{subG}}(P\|Q)$ | $\mathcal{A}$ only |
| Rényi-DP | $\sup_Z$ | $\sup_z$ | $D_\alpha(P\|Q)$ | $\mathcal{A}$ only |
| Personal-DP | $\sup_Z$ | fixed $z$ | $D_\infty^\delta(P\|Q)$ | $\mathcal{A}$ and $z$ |
| TV-privacy | $\sup_Z$ | $\sup_z$ | $\|P-Q\|_{TV}$ | $\mathcal{A}$ only |
| KL-privacy | $\sup_Z$ | $\sup_z$ | $D_{KL}(P\|Q)$ | $\mathcal{A}$ only |
| On-Avg KL-privacy | $\mathbb{E}_{Z\sim\mathcal{D}^n}$ | $\mathbb{E}_{z\sim\mathcal{D}}$ | $D_{KL}(P\|Q)$ | $\mathcal{A}$ and $\mathcal{D}$ |
| Random-DP | $1-\delta$ | $1-\delta$ | $D_\infty^\delta(P\|Q)$ | $\mathcal{A}$ and $\mathcal{D}$ |
| Per-instance DP | fixed $Z$ | fixed $z$ | $D_\infty^\delta(P\|Q)$ | $\mathcal{A}$, $Z$ and $z$ |

**Post hoc calculation of privacy loss.** The idea of calculating the privacy loss after running a fixed randomized algorithm is not new. It is the inverse problem of the typical task of calibrating noise to meet a prescribed privacy requirement and is often used as an intermediate step in the analysis of the latter [Dwork et al., 2006].

More recently, the post hoc view is adopted in the design of privacy odometer that tracks the post hoc overall privacy loss of a list of sequentially-chosen privacy parameters [Rogers et al., 2016]. Their analysis stays at an abstract-level as it describes an algorithm solely by its $(\epsilon, \delta)$-DP guarantee. It is also used in a more refined algorithm-specific privacy analysis for noisy SGD [Abadi et al., 2016] and Gaussian noise adding [Balle and Wang, 2018], which simultaneously ensures $(\epsilon(\delta), \delta)$-DP for all $0 < \delta < 1$ with a monotonically decreasing function $\epsilon(\delta)$. However, they do not adapt to the given input data set.

Ligett et al. [2017] define "ex-post privacy loss" (Definition 2.2.) to be the realized privacy loss random variable $\epsilon$(Outcome), but also do not adapt to the given input data set. In fact, a direct comparison of their analysis of linear/ridge regression (Theorem 3.1) to our case study reveals that our pDP analysis with the subsequent pDP-to-DP conversion allows us to come up with an algorithm that exploits the strong convexity that comes from the data set, and hence a more favorable bias-variance trade-off.

Data-dependent post hoc privacy analysis is relatively recent and was discussed in [Papernot et al., 2016] in the same flavor of "pDP for all", except that it is done with Renyi DP. They did not consider more fine-grained pDP which can be different for every individual.

**Frameworks for data-dependent DP algorithms.** Data-dependent DP algorithms were investigated under the classical framework of smooth sensitivity [Nissim et al., 2007] and propose-test-release (PTR) [Dwork and Lei, 2009]. The focus of these popular frameworks is on how to calibrate noise to local sensitivity, rather than how to calculate the data-dependent privacy loss after running a fixed randomized algorithm. Note that the algorithm under consideration needs not be differentially private and we do not propose to calibrate an algorithm based on pDP. The purpose of pDP analysis is to provide a more precise privacy loss summary of a given randomized algorithm, even though it might not be DP in the worst case.

Building upon the pDP analysis, we demonstrated that sometimes one can use it to design data-dependent DP algorithms. The approach is closely related to the PTR framework but has a more systematic way of identifying the key quantities that contribute to the sensitivity. Also, instead of proposing and testing an (often exponentially long) sequence of criteria, our approach involves directly releasing differentially private high-confidence bounds of certain key quantities. In particular, the proposed data-dependent differentially private linear regression estimator by [Dwork and Lei, 2009] runs in time exponential in the dimension, while our proposed AdaOPS is polynomial in all parameters.

While the manuscript is under peer-review, an anonymous reviewer brought to our attention the independent work of Cummings and Durfee [2018]. They consider an alternative framework for data-dependent DP algorithm design in which they use a notion called "individual sensitivity", which measures the maximum perturbation of a function when we add an individual $z$ to any data set $Z$ that does not contain $z$. This is different from either local sensitivity or per-instance sensitivity (that we will discuss in the next section) but is almost identical to the "personalized sensitivity" used in [Ghosh and Roth, 2015, Liu et al., 2015, Ebadi et al., 2015]. For many problems, e.g., linear regression, personalized sensitivity can be unbounded for all $z$, while per-instance sensitivity remains finite provided that the design matrix is not singular. In addition, their approach runs in exponential time in general and seems to apply only to output perturbation algorithms. pDP, on the other hand, is well-defined for any randomized algorithm.

## 2. Per-instance differential privacy

In this section, we define per-instance differential privacy, and derive its properties. We begin by parsing the standard definition of differential privacy.

**Definition 2.1** (Differential privacy [Dwork et al., 2006])**.** We say a randomized algorithm $\mathcal{A}$ satisfies $(\epsilon, \delta)$-DP if, for *all* data set $Z$ and data set $Z'$ that can be constructed by adding or removing one data point $z$ from $Z$,

$$\mathbb{P}_{\theta\sim\mathcal{A}(Z)}(\theta \in \mathcal{S}) \leq e^{\epsilon}\mathbb{P}_{\theta\sim\mathcal{A}(Z')}(\theta \in \mathcal{S}) + \delta, \quad \forall \text{ measurable set } \mathcal{S}.$$

When $\delta = 0$, this is also known as pure differential privacy.

It is helpful to understand what differential privacy is protecting against — a powerful adversary that knows everything in the entire universe, except one bit of information: whether a target $z$ is in the data set or not in the data set. The optimal strategy for such an adversary is to conduct a likelihood ratio test (or posterior inference) on this bit, and differential privacy uses randomization to limit the probability of success of such test [Wasserman and Zhou, 2010].

Note that the adversary always knows $Z$ and has a clearly defined target $z$, and it is natural to evaluate the winnings and losses of the "player", the data curator by conditioning on the same data set and privacy target. This gives rise to the following generalization of DP.

**Definition 2.2** (Per-instance Differential Privacy)**.** For a fixed data set $Z$ and a fixed data point $z$. We say a randomized algorithm $\mathcal{A}$ satisfy $(\epsilon, \delta)$-per-instance-DP for $(Z, z)$ if, for all measurable set $\mathcal{S} \subset \Theta$, it holds that

$$P_{\theta\sim\mathcal{A}(Z)}(\theta \in \mathcal{S}) \leq e^{\epsilon}P_{\theta\sim\mathcal{A}([Z,z])}(\theta \in \mathcal{S}) + \delta,$$
$$P_{\theta\sim\mathcal{A}([Z,z])}(\theta \in \mathcal{S}) \leq e^{\epsilon}P_{\theta\sim\mathcal{A}(Z)}(\theta \in \mathcal{S}) + \delta.$$

This definition is different from DP primarily because DP is the property of the $\mathcal{A}$ only and pDP is the property of both $\mathcal{A}, Z$ and $z$. If we take supremum over all $Z \in \mathcal{Z}^n$ and $z \in \mathcal{Z}$, then it recovers the standard differential privacy.

Similarly, we can define per-instance sensitivity for $(Z, z)$.

**Definition 2.3** (per-instance sensitivity)**.** Let $\mathcal{H} = \mathbb{R}^d$, for a fixed $Z$ and $z$. The per-instance $\|\cdot\|_*$ sensitivity of a function $f : \text{Data} \to \mathbb{R}^d$ is defined as $\|f(Z) - f([Z, z])\|_*$, where $\|\cdot\|_*$ could be $\ell_p$ norm or $\|\cdot\|_A = \sqrt{(\cdot)^T A(\cdot)}$ defined by a positive definite matrix $A$.

This definition also generalizes quantities in the classic DP literature. If we fix $Z$ but maximize over all $z \in Z$, we get local-sensitivity [Nissim et al., 2007]. If we maximize over both $Z \in \mathcal{Z}^*$ and $z \in \mathcal{Z}$, we get global sensitivity [Dwork et al., 2014, Definition 3.1]. These two are often infinite in real-life problems, but for a fixed data set $Z$ and target $z$ to be protected, we could still get meaningful per-instance sensitivity.

Immediately, the per-instance sensitivity implies pDP for a noise adding procedure.

**Lemma 2.4** (Multivariate Gaussian mechanism)**.** *Let $\hat{\theta}$ be a deterministic map from a data set to a point in $\Theta$, e.g., a deterministic learning algorithm, and let the $A$-norm per-instance sensitivity $\Delta_A(Z, z)$ be $\|\hat{\theta}([Z, z]) - \hat{\theta}(Z)\|_A$. Then adding noise with covariance matrix $A^{-1}/\gamma$ obeys $(\epsilon(Z, z), \delta)$-pDP for any $\delta > 0$ with*

$$\epsilon(Z, z) = \gamma\Delta_A(Z, z)\sqrt{\log(1.25/\delta)}.$$

The proof, which is standard and we omit, simply verifies the definition of $(\epsilon, \delta)$-pDP by calculating a tail bound of the privacy loss random variable and invokes Lemma B.6.

2.1. **Basic properties of pDP.** We now describe properties of per-instance DP, which mostly mirror those of DP.

**Fact 2.5** (Strong protection against identification)**.** Let $\mathcal{A}$ obeys $(\epsilon, \delta)$-pDP for $(Z, z)$, then for any measurable set $\mathcal{S} \subset \Theta$ where $\min\{\mathbb{P}_{\theta\sim\mathcal{A}(Z)}(\theta \in \mathcal{S}), \mathbb{P}_{\theta\sim\mathcal{A}(Z)}(\theta \in \mathcal{S})\} \geq \delta/\epsilon$ then given any side information $\mathsf{aux}$

$$-2\epsilon \leq \log \frac{\mathbb{P}_{\theta\sim\mathcal{A}(Z)}(\theta \in \mathcal{S}|\mathsf{aux})}{\mathbb{P}_{\theta\sim\mathcal{A}([Z,z])}(\theta \in \mathcal{S}|\mathsf{aux})} \leq 2\epsilon.$$

*Proof.* Note that after fixing $Z$, $\theta$ is a fresh sample from $\mathcal{A}(Z)$, as a result, $\theta \perp\!\!\!\perp \mathsf{aux}|Z$. The claimed fact then directly follows from the definition. □

Note that the log-odds ratio measures how likely one is able to tell one distribution from another based on side information and an event $\mathcal{S}$ of the released result $\theta$. When the log-odds ratio is close to 0, the outcome $\theta$ is equally likely to be drawn from either distribution.

**Fact 2.6** (Convenient properties directly inherited from DP)**.** For each $(Z, z)$ separately we have:

(1) Simple composition: Let $\mathcal{A}$ and $\mathcal{B}$ be two randomized algorithms, satisfying $(\epsilon_1, \delta_1)$-pDP, $(\epsilon_2, \delta_2)$-pDP, then $(\mathcal{A}, \mathcal{B})$ jointly is $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-pDP.

(2) Advanced composition: Let $\mathcal{A}_1, ..., \mathcal{A}_k$ be a sequence of randomized algorithms, where $\mathcal{A}_i$ could depend on the realization of $\mathcal{A}_1(Z), ..., \mathcal{A}_i(Z)$, each with $(\epsilon, \delta)$-pDP, then jointly $\mathcal{A}_{1:k}$ obeys $O(\sqrt{k\log(1/\delta)}\epsilon), O(k\delta)$-pDP. The same claim also holds for algorithm-specific advanced composition via concentrated DP and Renyi DP.

(3) Closedness to post-processing: If $\mathcal{A}$ satisfies $(\epsilon_1, \delta_1)$-pDP, for any function $f$, $f(\mathcal{A}(\cdot))$ also obeys $(\epsilon_1, \delta_1)$-pDP.

(4) Group privacy: If $\mathcal{A}$ obeys $(\epsilon, \delta)$-pDP with $\epsilon, \delta$ parameterized by (Data, Target), then

$$P_{\theta\sim\mathcal{A}(Z)}(\theta \in S) \leq e^{\epsilon(Z,z_1)+\epsilon([Z,z_1],z_2)+...+\epsilon([Z,z_{1:k-1}],z_k)} P_{\theta\sim\mathcal{A}([Z,z_{1:k}])}(\theta \in S) + \tilde{\delta}.$$
$$P_{\theta\sim\mathcal{A}([Z,z_{1:k}])}(\theta \in S) \leq e^{\epsilon(Z,z_1)+\epsilon([Z,z_1],z_2)+...+\epsilon([Z,z_{1:k-1}],z_k)} P_{\theta\sim\mathcal{A}(Z)}(\theta \in S) + \tilde{\delta}.$$

for $\tilde{\delta} = \sum_{i=1:k} \left[\delta([Z, z_{1:i-1}], z_i) \prod_{j=1:i-1} e^{\epsilon([Z,z_{1:j-1}],z_j)}\right]$.

*Proof.* These properties all directly follow from the proof of these properties for differential privacy (see e.g., [Dwork et al., 2014]), as the uniformity over data sets is never used in the proof. The group privacy is more involved since the size of the data set changes as the size of the privacy target (now a fixed group of people) gets larger, group privacy statement follows from a simple calculation that repeatedly applies the definition of pDP for a different data set. □

2.2. **The distribution and moments of pDP.** One useful notion to consider in practice is to understand exactly how much privacy loss is incurred for those who participated in the data set. This is practically relevant, because if a cautious individual decides to not submit his/her data, he/she would necessarily do it by rejecting a data-usage agreement and therefore the data collector is not legally obligated to protect this person and in fact does not have access to his/her data in the first place.

It is debatable whether it is as important to protect individuals who are not in the data set as those who are. We will illustrate this point with an example. Suppose the Federal government is to decide on a potential funding support based on whether a township's average household income qualifies for it. Household income is clearly considered sensitive information, so the census bureau decides to add a Laplace noise to the average income to prevent privacy risk. The noise level is chosen independently of the data such that it will not make the funding decision impossible. If we consider the richest person on earth, it is possible that his/her household income is larger than the total GDP of this township. The noise-adding algorithm does not provide a meaningful DP guarantee to him/her, as it will be straightforward to infer with high confidence that this person does not live in this township. But does it matter? It is unlikely that the richest person on earth would consider this kind of inference about him/her a breach of privacy.

pDP provides analytical tools to formally study the privacy of only those people in a data set. It offers a natural way to analyze and also empirically estimate any statistics of the pDP losses over a distribution of data points corresponding to a fixed randomized algorithm $\mathcal{A}$.

**Definition 2.7** (Moment pDP for a distribution)**.** Let $(Z, z)$ be drawn from some distribution (not necessarily a product distribution) $\mathcal{P}$, it induces a distribution of $\epsilon(Z, z)$. Then we say that the distribution obeys $k$th moment per-instance DP with parameter vector $(\mathbb{E}\epsilon, \mathbb{E}[\epsilon^2], ..., \mathbb{E}[\epsilon^k], \delta)$.

The moments of pDP and the corresponding view of pDP's privacy loss as a random variable is a powerful idea and it enables flexible and comprehensive descriptions of the privacy *footprint* of a randomized algorithm on a set of targets subject to a constraint or a distribution of the input data set. We give a few examples below.

**pDP of a data set:** Let $Z$ be a fixed data set, when we choose $\mathcal{P}$ to be a discrete uniform distribution supported on $\{(Z_{-i}, z_i)\}_{i=1}^n$ with probability $1/n$ for each $i$. Then taking $k = 2$ allows us to calculate the mean and variance of the privacy loss of individuals in a data set, and taking higher order $k$ allows us to produce quantile estimates and high probability tail bounds of the random-variable of an average user in the data set.

**pDP for all:** When we fix $Z$ and but allow $z$ to be drawn from any distribution defined on $\mathcal{Z}$, then this becomes a much stronger notion of privacy that protects all individual $z \in \mathcal{Z}$ provided that the data set is $Z$. This is closely related to local sensitivity [Nissim et al., 2007, Dwork and Lei, 2009].

**pDP for one:** When we fix $z$ but allow $Z$ to be drawn from any distributions defined on a collection of data sets, then this becomes the worst case privacy loss that can happen to a given individual $z$. This notion is closely related to the personalized DP [Ghosh and

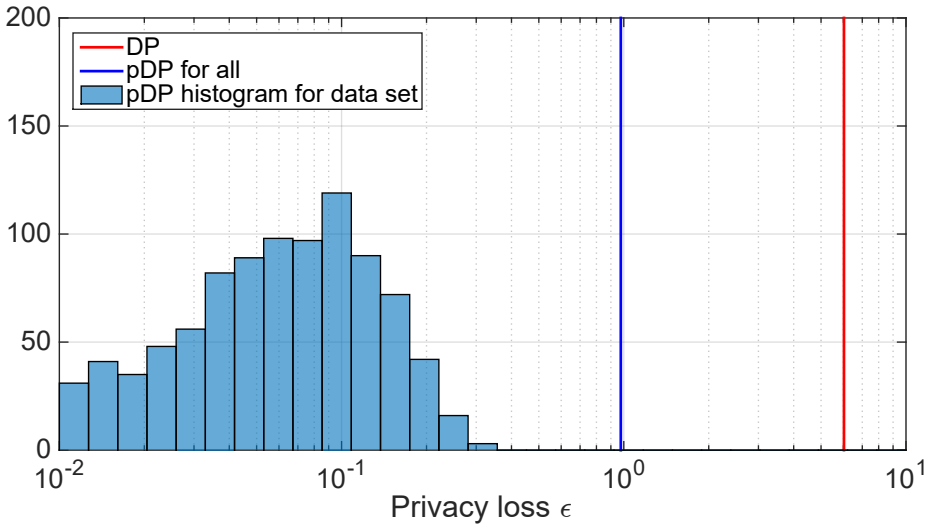


Figure 1: Illustration of the privacy loss $\epsilon$ of an output perturbation algorithm under DP, pDP for all, as well as the distribution of pDP's privacy loss for data points in the data set. The data set is generated by a linear Gaussian model, where the design matrix is normalized such that each row has Euclidean norm 1 and $y$ is also clipped at $[-1, 1]$. The output perturbation algorithm releases $\hat{\theta} \sim \mathcal{N}((X^T X + I)^{-1} X\mathbf{y}, \sigma^2 I)$ with $\sigma = 4$. Our choice of $\delta = 10^{-6}$.

Roth, 2015, Ebadi et al., 2015, Liu et al., 2015] and could be useful when individuals have different sensitivity.

**pDP with assumptions on data sets:** As we mentioned in the introduction, a branch of modern machine learning focuses on finding reasonable assumptions on the data sets which reduces the computation and sample complexity of a problem. In this case, $Z$ could be drawn from any distributions such that these assumptions are true with probability 1, in other words, we can take advantage of these assumptions when calculating the privacy loss of an individual $z$.

**pDP with a data set prior:** When we take $Z \sim \pi$ for some prior distribution $\pi$, then the moments of pDP make it possible to take advantage of that prior distribution to describe the privacy of an individual $z$ as a distribution over the possible privacy loss.

As an illustration, we compare pDP of a data set, pDP for all and the classical differential privacy using a simulated experiment. The results are shown in Figure 1. As we can see, the more fine-grained per-instance DP reveals more than an order of magnitude stronger privacy protection for all users in the data set, and six times better privacy protection for all users in the entire universe, than the standard DP's characterization. We will revisit some of these notions in our case study for linear regressions in Section 4 with concrete bounds.

2.3. **Generalization and domain adaptation.** Assume that the data set is drawn iid from some unknown distribution $\mathcal{D}$ — a central assumption in statistical learning theory — then we can take $\mathcal{P} = \mathcal{D}^{n-1} \times \mathcal{D}$. This allows us to use the moment of pDP losses to capture

on average how well data points drawn from $\mathcal{D}$ are protected. It also controls generalization error, and more generally cross-domain generalization.

**Definition 2.8** (On-average generalization)**.** Under the standard notations of statistical learning, the on-average generalization error of an algorithm $\mathcal{A}$ is defined as

$$\mathrm{Gen}(\mathcal{A}, \mathcal{D}, n) = \left| \mathbb{E}_{Z\sim\mathcal{D}^n, z\sim\mathcal{D}} \mathbb{E}_{\theta\sim\mathcal{A}(Z)} \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, z_i) - \ell(\theta, z) \right|.$$

**Proposition 2.9** (Moment pDP implies generalization)**.** *Assume bounded loss function $0 \le \ell(\theta, z) \le 1$. Then the on-average generalization is smaller than*

$$\mathbb{E}_{Z\sim\mathcal{D}^n}(\mathbb{E}_{z\sim\mathcal{D}}[e^{\epsilon(Z,z)}|Z])^2 - 1 + \mathbb{E}_{Z\sim\mathcal{D}^n, z\sim\mathcal{D}}\delta(Z, z) + (\mathbb{E}_{Z\sim\mathcal{D}^n}\mathbb{E}_{z\sim\mathcal{D}}[e^{\epsilon(Z,z)}|Z]\mathbb{E}_{z\sim\mathcal{D}}[\delta(Z, z)|Z].$$

Note that this can also be used to capture the privacy and generalization of transfer learning (also known as domain adaptation) with a fixed data set or a fixed distribution. Let the training distribution be $\mathcal{D}$ and target distribution be $\mathcal{D}'$,

Take $\mathcal{P} = \mathcal{D}^n \otimes \mathcal{D}'$ or $\mathcal{P} = \delta_Z \otimes \mathcal{D}'$. In practice, this allows us to upper bound the generalization to the Asian demographics group, when the training data is drawn from a distribution that is dominated by white males (e.g., the current DNA sequencing data set). We formalize this idea as follows.

**Definition 2.10** (Cross-domain generalization)**.** Assume $0 \le \ell(\theta, z) \le 1$. The on-average cross-domain generalization with base distribution $\mathcal{D}$ to target distribution $\mathcal{D}'$ is defined as:

$$\mathrm{Gen}(\mathcal{A}, \mathcal{D}, \mathcal{D}', n) \le \left| \mathbb{E}_{Z\sim\mathcal{D}^n, z\sim\mathcal{D}'} \mathbb{E}_{\theta\sim\mathcal{A}(Z)} \left[ \frac{1}{n} \sum_{i=1}^{n} \rho_i \ell(\theta, z_i) - \ell(\theta, z) \right] \right|.$$

where $\rho_i = \mathcal{D}'(z_i)/\mathcal{D}(z_i)$ is the inverse propensity (or importance weight) to account for the differences in the two domains.

**Proposition 2.11.** *The cross-domain on-average generalization can be bounded as follows:*

$$Gen(\mathcal{A}, \mathcal{D}, \mathcal{D}', n) = \mathbb{E}_{Z\sim\mathcal{D}^{n-1}, \{z'\}\sim\mathcal{D}, z''\sim\mathcal{D}'}[(e^{\epsilon(Z,z')+\epsilon(Z,z'')} - 1) + \delta(Z, z') + \epsilon(Z, z')\delta(Z, z'')].$$

The expressions in Proposition 2.9 and 2.11 are a little complex, we will simplify them to make it more readable.

**Corollary 2.12.** *Let $\sup_{Z,z} \delta(Z, z) \le \delta$, and $\mathbb{E}_{\mathcal{D}}[e^{2\epsilon(Z,z)}] \le 1$ and for simplicity, we write $\mathbb{E}_{Z\sim\mathcal{D}^n, z\sim\mathcal{D}}\epsilon(Z, z) = \mathbb{E}_{\mathcal{D}} f$ and $\mathbb{E}_{Z\sim\mathcal{D}^n, z\sim\mathcal{D}'}\epsilon(Z, z) = \mathbb{E}_{\mathcal{D}'} f$. Then the cross-domain on-average generalization is smaller than*

$$\frac{1}{2}[\mathbb{E}_{\mathcal{D}} e^{2\epsilon} + \mathbb{E}_{\mathcal{D}'} e^{2\epsilon}] - 1 + 2\delta = \frac{1}{2}\left[ \sum_{i=1}^{\infty} \frac{2^i}{i!} \mathbb{E}_{\mathcal{D}} \epsilon^i + \mathbb{E}_{\mathcal{D}'} \epsilon^i \right] + 2\delta.$$

It will be interesting to compare the quantity to Rényi-DP which also uses the moment generating function of the privacy random variable. The difference is that in Rényi-DP, the privacy random variable is induced by the distribution of the output, while here it is induced by the distribution of the data set and the target.

## 3. Per-instance sensitivity in smooth learning problems

In this section, we present our main results and give concrete examples in which per-instance sensitivity (hence per-instance privacy) can be analytically calculated. Specifically, we consider following regularized empirical risk minimization form:

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \sum_i \ell(\theta, z_i) + r(\theta), \tag{3.1}$$

or in the non-convex case, finding a local minimum. $\ell(\theta, z)$ and $r(\theta)$ are the loss functions and regularization terms. We make the following assumptions:

A.1. $\ell$ and $r$ are differentiable in argument $\theta$.

A.2. The partial derivatives are absolute continuous, i.e., they are twice differentiable almost everywhere and the second order partial derivatives are Lebesgue integrable.

Our results under these assumptions will cover learning problems such as linear and kernel machines as well as some neural network formulations (e.g., multilayer perceptron and convolutional net with sigmoid/tanh activation), but not non-smooth problems like lasso, $\ell_1$-SVM or neural networks ReLU activation. We also note that these conditions are implied by standard assumptions of strong smoothness (gradient Lipschitz) and do not require the function to be twice differentiable everywhere. For instance, the results will cover the case when either $\ell$ or $r$ is a Huber function, which is not twice differentiable.

Technically, these assumptions allow us to take Taylor expansion and have an integral form of the remainder, which allows us to prove the following stability bound.

**Lemma 3.1.** *Assume $\ell$ and $r$ satisfy Assumption A.1 and A.2. Let $\hat{\theta}$ be a stationary point of $\sum_i \ell(\theta, z_i) + r(\theta)$, $\hat{\theta}'$ be a stationary point $\sum_i \ell(\theta, z_i) + \ell(\theta, z) + r(\theta)$ and in addition, let $\eta_t = t\hat{\theta} + (1-t)\hat{\theta}'$ denotes the interpolation of $\hat{\theta}$ and $\hat{\theta}'$. Then the following identity holds:*

$$\begin{aligned}\hat{\theta} - \hat{\theta}' &= \left[\int_0^1 \left(\sum_i \nabla^2 \ell(\eta_t, z_i) + \nabla^2 \ell(\eta_t, z) + \nabla^2 r(\eta_t)\right) dt\right]^{-1} \nabla \ell(\hat{\theta}, z) \\ &= -\left[\int_0^1 \left(\sum_i \nabla^2 \ell(\eta_t, z_i) + \nabla^2 r(\eta_t)\right) dt\right]^{-1} \nabla \ell(\hat{\theta}', z).\end{aligned}$$

The proof uses first order stationarity condition of the optimal solution and apply Taylor's theorem on the gradient.

The lemma is very interpretable. It says that the perturbation of adding or removing a data point can be viewed as a one-step quasi-newton update to the parameter. Also note that $\nabla \ell(\hat{\theta}', z)$ is the "score function" in parametric statistical models, and when $\ell(\hat{\theta}', (x, y)) = \ell(f_{\hat{\theta}}(x), y)$, it is the product of the "pseudo-residual" $\frac{\partial \ell}{\partial f}$ and the gradient direction $\nabla f$ in gradient boosting [see e.g., Friedman et al., 2001, Chapter 10].

The result implies that the per-instance sensitivity in $\|\cdot\|_A$ for some p.d. matrix $A$ can be stated in terms of a certain norm of the "score function" specified by a quadratic form $H^{-1}AH^{-1}$, and therefore by Lemma 2.4, the output perturbation algorithm:

$$\tilde{\theta} \sim \mathcal{N}(\hat{\theta}(X), A^{-1}/\gamma), \tag{3.2}$$

obeys $(\epsilon, \delta)$-pDP for any $\delta > 0$ and

$$\epsilon(Z, z) = \sqrt{\nabla \ell(\hat{\theta}', z)^T H^{-1} A H^{-1} \nabla \ell(\hat{\theta}', z) \log(1.25/\delta)}. \tag{3.3}$$

This is interesting because for most loss functions the "score function" is often proportional to the prediction error of the fitted model $\hat{\theta}'$ on data point $z$ and this result suggests that the more accurately a model predicts a data point, the more private this data point is. This connection is made more explicit when we specialize to linear regression and the per-instance sensitivity

$$\begin{aligned} \Delta_A(Z, z) &= |y - x^T \hat{\theta}| \sqrt{x^T ([X']^T X')^{-1} A ([X']^T X')^{-1} x} \\ &= |y - x^T \hat{\theta}'| \sqrt{x^T (X^T X)^{-1} A (X^T X)^{-1} x}. \end{aligned} \tag{3.4}$$

is clearly proportional to prediction error. In addition, when we choose $A \approx X^T X$, the second term becomes either $\mu := x^T ([X, x]^T [X, x])^{-1} x$ or $\mu' := x^T (X^T X)^{-1} x$, which are "in-sample" and "out-of-sample" *statistical leverage scores* of $x$.

Leverage score measures the importance/uniqueness of a data point relative to the rest of the data set and it is used extensively in regression analysis [Chatterjee and Hadi, 1986] (for outlier detection and experiment design), graph sparsification (for adaptive sampling) [Spielman and Srivastava, 2011] and numerical linear algebra (for fast matrix computation)[Drineas et al., 2012]. To the best of our knowledge, this is the first time leverage scores are shown to be connected to differential privacy.

## 4. Case study: pDP analysis in linear regression

So far we have described output perturbation algorithms with a fixed noise adding procedure. However in practice it is not known ahead of time how to choose $A$. Assume all $x$ are normalized to $\|x\| = 1$[3], denote $\mu_2(x) := x^T (X^T X)^{-2} x$, $\mu_1(x) := x^T (X^T X)^{-1} x$. We discuss the pros and cons of the three natural choices.

- $A \approx \lambda_{\min} I$: This corresponds to the standard $\ell_2$-sensitivity and it adds an isotropic noise and provides a uniform guarantee for all data-target pairs where $X^T X$ has smallest eigenvalue $\lambda_{\min}$, because $\sup_x \sqrt{\mu_2(x)} \leq 1/\lambda_{\min}$, but it adds more noise than necessary for those with much smaller $\mu_2(x)$.
- $A \approx (X^T X)^2$: We call this the "democratic" choice conditioned on the data set, as it homogenizes the "leverage" part of the per-instance sensitivity of points to $\|x\| = 1$ so any $x$ gets about the same level of privacy. It, however, is not robust if our data-independent choice of $A$ is in fact far away from the actual $(X^T X)^2$.
- $A \approx X^T X$: We call this the "Fisher"-choice, because the covariance matrix will be proportional to the inverse Fisher information, which is the natural estimation error of $\hat{\theta}$ under the linear regression assumption. The advantage of this choice is that conducting

[3] This assumption simplifies the presentation. In general, we can conduct pDP analysis to any randomized algorithm on any data set. However, as Cummings et al. [2015] has shown, if the data domain is not bounded, then no algorithm can differentially privately release linear regression coefficients with non-trivial accuracy in general.

statistical inference, e.g., t-test and ANOVA for linear regression coefficients would be trivial.

Interestingly, for linear and ridge regression, the second and third choices are closely related to popular algorithms studied before. In fact, taking $A = (X^TX)^2$ recovers the objective perturbation (ObjPert) method[Chaudhuri et al., 2011, Kifer et al., 2012]:

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\Theta} \|\mathbf{y} - X\theta\|^2 + \langle z, \theta\rangle, \quad \theta \sim \mathcal{N}(0, \sigma^2 I). \tag{4.1}$$

while taking $A = X^TX$ recovers the one-posterior-sampling (OPS) mechanism proposed in [Dimitrakakis et al., 2014, Wang et al., 2015], which outputs

$$\hat{\theta} \sim P(\theta|X, \mathbf{y}) \propto e^{-\gamma\|\mathbf{y}-X\theta\|^2}. \tag{4.2}$$

An important difference is that in ObjPert and OPS, $A$ is not fixed, but rather depends on the data. As a result, we cannot use Lemma 2.4 to calculate the pDP. In fact, the data-independent choice of $A$ could imply an unbounded $\epsilon$ (consider an arbitrarily near singular $X$ and $x$ in its null space).

Not surprisingly, existing analyses of ObjPert and OPS require additional assumptions. Kifer et al. [2012] adds an additional $\lambda\|\theta\|^2$ to (4.1), while Wang et al. [2015] assumes that the loss function is bounded (by modifying it or constraining the domain $\Theta$) so that the exponential mechanism [McSherry and Talwar, 2007] would apply. It was later pointed out in [Foulds et al., 2016] that OPS is not asymptotically efficient in that it has an asymptotic relative efficiency (ARE) inversely proportional to $\epsilon$, while simple sufficient statistics perturbation can achieve asymptotic efficiency comparable to [Smith, 2008].

In the remainder of the section, we will first zoom into the OPS and propose a direct analysis of pDP using Lemma 3.1, then we will describe how to use the pDP analysis to obtain an extension of OPS that obeys $(\epsilon, \delta)$-DP and asymptotically efficient under the same data assumption in [Foulds et al., 2016]. We will see that OPS effectively converges to the "Fisher"-choice of noise adding in the same asymptotic regime and offers dimension and condition number independent expected pDP loss.

### 4.1. **pDP analysis of OPS.** The first result calculates the pDP loss of OPS.

**Theorem 4.1** (The adaptivity of OPS in Linear/Ridge Regression)**.** *Consider the* OPS *algorithm that samples from*

$$p(\theta|X, \mathbf{y}) \propto e^{-\frac{\gamma}{2}\left(\|\mathbf{y}-X\theta\|^2+\lambda\|\theta\|^2\right)}.$$

*Let $\hat{\theta}$ and $\hat{\theta}'$ be the ridge regression estimate with data set $X \times \mathbf{y}$ and $[X, x] \times [\mathbf{y}, y]$ and defined the out of sample leverage score $\mu := x^T(X^TX + \lambda I)^{-1}x = x^TH^{-1}x$ and in-sample leverage score $\mu' := x^T[(X')^TX' + \lambda I]^{-1}x = x^T(H')^{-1}x$. Then for every $\delta > 0$, privacy*

*target* $(x, y)$*, the algorithm is* $(\epsilon, \delta)$*-pDP with*

$$\epsilon(Z, z) \leq \frac{1}{2}\left|-\log(1+\mu) + \frac{\gamma\mu}{(1+\mu)}(y - x^T\hat{\theta})^2\right| + \frac{\mu}{2}\log(2/\delta) + \sqrt{\gamma\mu\log(2/\delta)}|y - x^T\hat{\theta}| \tag{4.3}$$

$$= \frac{1}{2}\left|-\log(1-\mu') - \frac{\gamma\mu'}{1-\mu'}(y - x^T\hat{\theta}')^2\right| + \frac{\mu'}{2}\log(2/\delta) + \sqrt{\gamma\mu'\log(2/\delta)}|y - x^T\hat{\theta}'|. \tag{4.4}$$

The proof is given in the appendix.

The two equivalent upper bounds are both useful. (4.3) is ideal for calculating pDP when $x$ is not in the data set and (4.4) is perfect for the case when $x$ is in the data set.

**Remark 4.2.** The bound (4.3) can be simplified to

$$\frac{\mu}{2}(1 + \log(2/\delta)) + \frac{1}{2}\gamma\min(\mu, 1)|y - x^T\hat{\theta}'|^2 + \sqrt{\gamma\mu\log(2/\delta)}|y - x^T\hat{\theta}'|.$$

If $\mu = o(\log(2/\delta))$[4] and we choose $\gamma$ such that $\sqrt{\gamma\mu'\log(2/\delta)}|y - x^T\hat{\theta}'| \leq 1$, then the bound can be simplified to

$$\epsilon(Z, z) \leq 2\sqrt{\gamma\mu\log(2/\delta)}|y - x^T\hat{\theta}| + o(1).$$

This matches the order of Gaussian mechanism with a fixed (data-independent) covariance matrix.

The results in [Foulds et al., 2016] are stated for general exponential family models under a set of assumptions that translate into the following for linear regression:

(a) data $x_1, ..., x_n$ is drawn i.i.d. from $\mathcal{D}$ supported on $\mathcal{X}$ where $\mathcal{X} \subset \mathcal{B}_{\|\cdot\|_2}(1)$.

(b) population covariance matrix $\frac{m}{d}I \preceq \mathbb{E}_{\mathcal{D}}xx^T \preceq \frac{M}{d}I$ for constant $m$ and $M$,

(c) $y_i \sim \mathcal{N}(x_i^T\theta_0, \sigma^2)$ for some $\theta_0$.

To simplify the presentation, we also assume $n$ scales with respect to $d$ such that

(d) with high probability, $XX^T \succ \frac{\alpha n}{2d}I$.

The last assumption measures how quickly the empirical covariance matrix $\frac{1}{n}XX^T$ concentrates to $\mathbb{E}_{x\sim\mathcal{D}}xx^T$. It can be shown that if $X$ is an appropriately scaled subgaussian random matrix, this happens with probability $1 - n^{-10}$ whenever $n > \max(10d, 10d^{-2/3}\log n)$.

**Proposition 4.3.** *The sequence of OPS algorithm with parameter* $\gamma_n$*,* $\lambda_n$ *obeys the following properties.*

(1) ***pDP and DP in the agnostic setting.*** *Assume* $\|x\| \leq 1$ *for every* $x \in \mathcal{X}$*. The algorithm obeys* $(\epsilon_n, \delta)$*-pDP, for each data set* $(X, \mathbf{y})$ *and all target* $(x, y)$*,*

$$\epsilon_n = \sqrt{\frac{\gamma_n\log(2/\delta)}{\lambda_n + \lambda_{\min}}}|y - x^T\hat{\theta}| + \frac{\gamma_n|y - x^T\hat{\theta}|^2}{2\max\{\lambda_n + \lambda_{\min}, 1\}} + \frac{\gamma_n(1 + \log(2/\delta))}{2(\lambda_n + \lambda_{\min})}. \tag{4.5}$$

[4] This is not an unrealistic assumption because $\mu$ and $\mu'$ are $o(1)$ as long as $x$ is bounded and the minimum eigenvalue of $X^TX + \lambda I$ is $\omega(1)$. This is required for (agnostic) linear regression to be consistent and is implied by the condition that the population covariance matrix $\frac{1}{n}\mathbb{E}X^TX$ is full rank.

*If we further assume* $|y|<1$, *then* $\sup_{(X,\mathbf{y}),(x,y)}|y-x^T\hat{\theta}| = 1+n^{1/2}\lambda_n^{-1/2}$ *and the algorithm obeys* $(\epsilon_n,\delta)$*-DP with*

$$\epsilon_n = \sqrt{\frac{2(n+\lambda_n)\gamma_n\log(2/\delta)}{\lambda_n^2}} + \frac{2(n+\lambda_n)\gamma_n}{\lambda_n\max\{1,\lambda_n\}} + \frac{\gamma_n(1+\log(2/\delta))}{2\lambda_n}. \tag{4.6}$$

(2) ***pDP under model assumption.*** *Assume conditions (a)(b)(c)(d) above are true, and also* $\gamma_n=\omega(1)$, $\lambda_n=o(\sqrt{n})$. *Then with high probability over the joint distribution of* $(X,\mathbf{y})$, *the algorithm with* $\gamma_n \le \frac{4n\log(2/\delta)}{\max\{d,(1+\log(2/\delta))^2\}}$ *obeys* $(\epsilon_n,\delta)$*-pDP with*

$$\epsilon_n = \begin{cases} O\left(\sqrt{\frac{(1+\|\theta_0\|)^2 d\gamma_n}{\alpha n}\log(\frac{2}{\delta})}\right) & \text{for all (x,y) satisfying } \|x\|=O(1) \text{ and } y=O(1). \\ O\left(\sqrt{\frac{\sigma^2 d\gamma_n}{\alpha n}\log(\frac{2}{\delta})\log(\frac{2}{\delta'})}\right) & \text{for any } x \text{ with probability } 1-\delta' \text{ over } y\sim\mathcal{N}(\theta_0^T x,\sigma^2). \end{cases}$$

*Moreover, with probability* $1-n\delta'$ *over the conditional distribution* $\mathbf{y}|X$, *the privacy loss of* $(x_1,y_1),...,(x_n,y_n)$ *obeys*

$$\frac{1}{n}\sum_{i=1}^n \epsilon_n\big((X,\mathbf{y}),(x_i,y_i)\big)^2 = O\Big(\frac{\sigma^2 d\gamma_n}{n}\log(2/\delta)\log(2/\delta')\Big),$$

*which does not depend on* $\alpha$ *— the smallest eigenvalue of* $dX^TX/n$.

(3) ***Statistical efficiency.*** *for every realization of data set* $X$ *such that* $n>d$ *and let the smallest eigenvalue of* $X^TX$ *be* $\lambda_{\min}$, *then*

$$\mathbb{E}_{\mathbf{y}\sim\mathcal{N}(X\theta_0,\sigma^2 I_n)}\left[\|\tilde{\theta}-\theta_0\|^2\Big|X\right] = \sigma^2\mathrm{tr}[(X^TX+\lambda_n I)^{-1}](1+\gamma_n^{-1}) + \lambda_n^2\|(X^TX+\lambda_n I)^{-1}\theta_0\|^2.$$

*If* $\lambda_{\min}=\Omega(d/n)$ *( this is true with high probability under assumption (b)(d)), then*

$$\mathbb{E}_{\mathbf{y}\sim\mathcal{N}(X\theta_0,\sigma^2 I_n)}\left[\|\tilde{\theta}-\theta_0\|^2\Big|X\right] = \sigma^2\mathrm{tr}[(X^TX+\lambda_n I)^{-1}](1+\gamma_n^{-1}) + O(\frac{\lambda_n^2 d^2\|\theta_0\|^2}{n^2}).$$

*In other words, the estimator is asymptotically efficient, for all* $\lambda_n=o(n^{1/2})$ *and* $\gamma_n=\omega(1)$.

(4) ***Optimization error.*** *Let* $F(\theta)=0.5\|\mathbf{y}-X\theta\|^2+\lambda\|\theta\|^2$ *and* $\hat{\theta}=\mathrm{argmin}\,F(\theta)$, *then*

$$\mathbb{E}F(\tilde{\theta})-F(\hat{\theta}) = d/\gamma_n,$$

*and also with probability at least* $1-\delta$ *over* $P(\tilde{\theta}|Z)$

$$F(\tilde{\theta})-F(\hat{\theta}) \le d\log(d/\delta)/\gamma_n.$$

*With* $\gamma_n=\frac{\epsilon^2\alpha n}{d\log(2/\delta)}$, *the result matches the information-theoretic lower bound for differentially private empirical risk minimization [Bassily et al., 2014]*[5].

[5]Note that in [Bassily et al., 2014], the strong convexity parameter $\Delta$ is assumed for each loss function, therefore it maps into our $\alpha$ as $\Delta\asymp\alpha/d$. Also, we used that in [Bassily et al., 2014]'s setting the loss function is Lipschitz within the bounded domain, which ensures $|y-x^T\theta|=O(1)$.

We now discuss a few aspects of the above results.

**pDP vs DP in the agnostic setting.** Firstly, it highlights the key advantage of pDP over DP. DP is not able to take advantage of desirable structures in the data set, while pDP provides a principled framework to handle them.

In particular, let us compare the pDP and DP in the agnostic setting, for the OPS that uses the same randomization. DP measures something that is completely data independent and corresponds specifically to a contrivedly constructed data set $(X, \mathbf{y})$ such that $\mathbf{y}$ is an eigenvector of $XX^T$ corresponding to a specific eigenvalue of magnitude $\sqrt{\lambda_n}$, this makes $\|\hat{\theta}\|_2$ as large as $\sqrt{n}/\sqrt{\lambda_n}$. Moreover, a target data point is chosen so that $x$ match the direction of $\hat{\theta}$. While this is a legitimate construction in theory, it does not directly correspond to the specific data set that a statistician just spent two years collecting, and it is unreasonable that he/she will have to calibrate the amount of noise to inject to provide more reasonable protection to a pathological case that has nothing to do with the reality.

pDP, on the other hand, makes it possible for the statistician to condition on the data set. If the statistician finds out that $\|\hat{\theta}\|_2 = O(1)$, then the pDP loss is as small as $\sqrt{\gamma_n \log(2/\delta)/\lambda_n}$ for *everyone* in the population. With $\gamma_n = n^{\alpha/2}$ and $\lambda_n = n^{1/2-\alpha/2}$ for any $\alpha > 0$, the algorithm remains to be statistically efficient with an ARE of $(1 + n^{-\alpha})$ yet can provide a strong privacy guarantee of $\epsilon_n = n^{-1/4+\alpha/2}$. If in addition, the statistician realized that the data set is *well-conditioned*, that is, the maximum and minimum eigenvalue of $X^T X$ are on the same order of $n/d$, then we can further improve the bound by replacing $\lambda_n$ with $\lambda_{\min} + \lambda_n$. The statistician can happily get away with the same privacy guarantee ($\epsilon_n = n^{-1/4}$) while not having to add too much noise or even regularize at all (setting $\gamma_n = n^{1/2}$ and $\lambda_n = 0$). Note that the condition number is a desirable property that governs how reliably one can hope to estimate the linear regression coefficients using the given data set.

We would like to emphasize that the pDP guarantee in the two cases we discussed above applies to everyone in the population $\{(x, y) | \|x\| \leq 1, |y| \leq 1\}$, therefore such $(\epsilon, \delta)$-pDP guarantee is as powerful as $(\epsilon, \delta)$-DP after the data set is collected.

**pDP-for-all vs average pDP on the data set.** Secondly, unlike DP which always provides a crude upper bound for everyone, pDP is able to reflect the differences in the protection of different target person. Under the model assumption, the average privacy loss of people in the data set is scale-invariant and interestingly, also independent of the condition number (smallest eigenvalue). It is a factor of $(1 + |\theta_0|)^2/m$ times smaller than the pDP guarantee for everyone in the population. This is significant for finite sample performance since $(1 + \|\theta_0\|)/m$ (although they do not change with $n$), can be quite large.

**pDP under covariate shift.** Lastly, if we consider a setting in between the above two, where the target $x$ can be drawn from any distribution defined on $\mathcal{X}$ that could be arbitrarily different from the training data distribution, then the scale-invariant property remains (the factor of $(1 + \|\theta_0\|)$ is dropped). This is relevant in causal learning when the $\mathbb{E}(y|x)$ is specified by some physical principles that are invariant to the distribution of $x$. In this case, the moments of the pDP would imply a much stronger notion of cross-domain generalization than what we show in Proposition 2.11 since it does not depend on the target distribution of interest.

**Improved DP guarantee for OPS.** The proposition also improves the existing analysis for the OPS algorithm as a byproduct. The first statement shows that OPS preserves a meaningful (almost constant) differential privacy when $\gamma_n = 1$ and $\lambda_n = \sqrt{n/d}$ without requiring a constant boundedness in the domain $\Theta$ or clipping the loss function like in Wang et al. [2015]. As a matter of fact, the ridge regression solution $\hat{\theta}$ could be in a ball of radius $\Theta(n^{1/4})$, and even if we impose the smallest domain bound that covers $\hat{\theta}$, by exponential mechanism, the algorithm only obeys a pure $O(n^{1/2})$-DP, in contrast to the $(O(\log(1/\delta)), \delta)$-DP that we showed in the proposition above.

Despite the improvement, the DP guarantee is still a little unsatisfactory. If we require $(\epsilon, \delta)$-DP with constant $\epsilon$, then the OPS algorithm with $\lambda_n = \sqrt{n}$ is not asymptotically efficient (although it does achieve the optimal $O(1/n)$ rate).

Meanwhile, there are algorithms that attain asymptotic efficiency either by subsample-and-aggregate [Smith, 2008] or by simply adding noise to the sufficient statistics [Dwork and Smith, 2010, Foulds et al., 2016]. So the question becomes: can we modify OPS such that it becomes asymptotically efficient with $(\epsilon_n, \delta)$-differentially private with $\epsilon_n = o(1)$?

We address this issue next.

4.2. **"pDP to DP conversion" and AdaOPS.** In this section, we resolve the dilemma described earlier by using the idea of Dwork and Lei [2009]. The new algorithm, which we call AdaOPS, adaptively and differentially privately chooses the tuning parameter $\lambda_n$ and $\gamma_n$ according to properties of the data set and privacy requirement. A pseudocode of AdaOPS is given in Algorithm 1. We acknowledge that the same idea of adaptively adding regularization terms is not new and has been used by Kifer et al. [2012], Blocki et al. [2012], Sheffet [2017] for analyzing other related differentially private algorithms. Our contribution here is only to assemble the ideas together into a working algorithm and illustrate how pDP analysis can help us design data-dependent DP algorithm that takes a prescribed $(\epsilon, \delta)$ budget as an input.

**Algorithm 1:** AdaOPS: One-Posterior Sample with adaptive regularization

**input** Data $X$, $\mathbf{y}$. Privacy budget: $\epsilon$, $\delta$. Parameter $\kappa$ satisfying $0 \le \kappa \le \frac{n\epsilon}{4d(1+\log(4/\delta))}$

1. Calculate the minimum eigenvalue $\lambda_{\min}(X^T X)$.
2. Private release $\tilde{\lambda}_{\min} = \lambda_{\min} + \frac{\sqrt{\log(4/\delta)}}{\epsilon/2} Z$, where $Z \sim \mathcal{N}(0,1)$.
3. Get one sample
$$\tilde{\theta} \sim \mathbb{P}(\theta | X, \mathbf{y}) \propto e^{-\frac{\gamma_n}{2}\left(\|\mathbf{y} - X\theta\|^2 + \lambda_n \|\theta\|^2\right)}.$$
with parameter
$$\lambda_n = \min\left\{0, \frac{n}{d\kappa} - \tilde{\lambda}_{\min} + \frac{\log(4/\delta)}{\epsilon/2}\right\},$$
$$\gamma_n = \min\left\{\frac{n\epsilon^2}{16\kappa^2 d^2 \log(4/\delta)}, \frac{n\epsilon}{8\kappa^2 d^2}\right\}.$$

**output** $\tilde{\theta}$

The $\kappa$ parameter is the largest acceptable condition number in the data set. Often it can be determined independently of the data. It is used in the algorithm to rule out the pathological case of a possibly near-singular design matrix. We now analyze the properties of AdaOPS.

**Proposition 4.4.** (1) *Assume data domain is* $\|x\|_2 \le 1$ *and* $|y| \le 1$. *The* AdaOPS *estimator preserves* $(\epsilon, \delta)$*-DP.*

(2) *If Assumption (a)(b)(c) are true and in addition for the specific realization of* $X$*,*

$$\lambda_{\min}(X^T X) > \frac{n}{\kappa d} + \frac{\sqrt{\log(10n)\log(4/\delta)}}{\epsilon/2},$$

*then, we have*

$$\mathbb{E}[\|\tilde{\theta} - \theta_0\|^2 | X] = [1 + \gamma_n]\, \sigma^2 \mathrm{tr}[(X^T X)^{-1}] + O(n^{-10})\|\theta_0\|^2.$$

*In other words, since* $\gamma_n \le \min\{\frac{\kappa^2 d^2 \log(4/\delta)}{n\epsilon^2}, \frac{\kappa^2 d^2}{n\epsilon}\}$*, the* AdaOPS *estimator achieves asymptotic efficiency whenever* $\epsilon$ *obeys that* $\min\{n\epsilon^2, s\epsilon\} = o(\kappa^2 d^2 \log(4/\delta)/n)$.

This proposition reveals that AdaOPS improves over previous results in the literature [Smith, 2008, Foulds et al., 2016] in several ways. First of all, we only need $n\epsilon^2 = o(1)$ to achieve asymptotic efficiency. In contrast, [Foulds et al., 2016] does not provide non-asymptotic results with explicit dependence and [Smith, 2008]'s bound for the subsample-and-aggregate method requires $n^{-1/5}\epsilon^{-6/5} = o(1)$ to achieve asymptotic efficiency.

Secondly, our bound has explicit dimension dependence while [Foulds et al., 2016] and [Smith, 2008] treat $d$ as a constant. In particular, our bound on the additive difference from exactly matching the Cramer-Rao lower bound of $(\sigma^2 \mathrm{tr}((X^T X)^{-1}))$ translates into $\sigma^2 \mathrm{tr}[(X^T X)^{-1}] + d^3/(n^2\epsilon^2)$.

The extension from OPS to AdaOPS is a good example of what we call "pDP to DP conversion", which follows the following procedures:

(1) Start with a fixed randomized algorithm of interest $\mathcal{A}$ (e.g.,OPS).

(2) Calculate its pDP analytically (Proposition 4.1).

(3) Inspect to identify key quantities ( in our case it is the strong convexity parameters).

(4) Differentially privately release high-probability confidence intervals of these key quantities (by releasing the smallest eigenvalue) and enforce the properties when needed (add regularization.)

"pDP to DP conversion" uses the high-level idea of the Propose-Test-Release framework [Dwork and Lei, 2009], which involves testing a sequence of conditions on key data-dependent quantities of the problem. Our approach is different because we propose to directly release these key quantities and intervene (regularize) if necessary. On the meta-level, a careful pDP analysis allows us to identify what these key quantities are and how they contribute to the sensitivity. Compared to the "robust linear regression" approach [Dwork and Lei, 2009, Section 4], AdaOPS avoids the need to discretize $\Theta$, hence does not require a runtime that is exponential in dimension $d$.

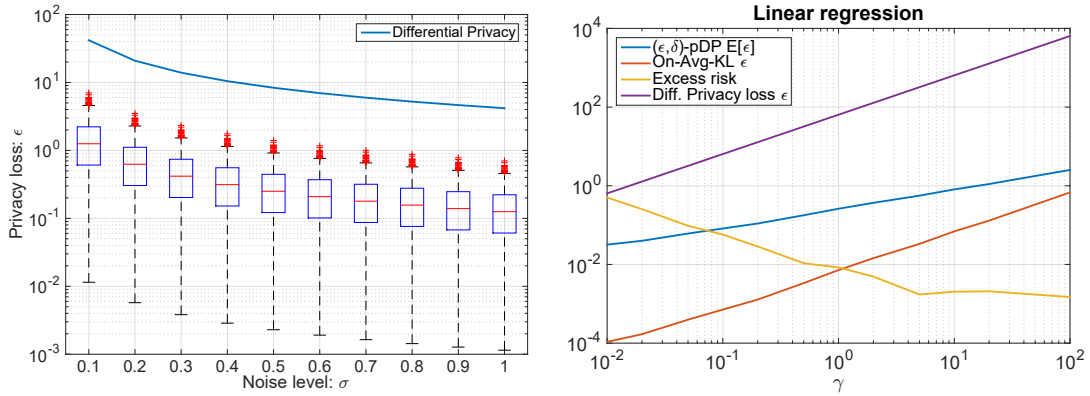


Figure 2: **Left:** $(\epsilon, \delta)$-DP and distribution of $(\epsilon(z, Z), \delta)$-pDP data points in linear regression with isotropic Gaussian noise adding. **Right:** Comparing the pDP privacy loss to the $\epsilon$-DP obtained through exponential mechanism [Wang et al., 2015] using the same posterior sampling algorithm. In both experiment $\delta = 1e - 6$.

4.3. **Simulation.** We conclude the case study with two simulated experiments (shown in the two panes of Figure 2). In the first experiment, we consider the algorithm of adding isotropic Gaussian noise to linear regression coefficients and then compare the worst-case DP and the distribution of per-instance DP for points in the data set (illustrated as box plots). In the second experiment, we compare different notions of privacy to utility (measured as excess risk) of the fixed algorithm that samples from a scaled posterior distribution. In both cases, the average per-instance differential privacy over the data sets is several orders of magnitude smaller than the worst-case differential privacy.

## 5. Concluding discussion

In this paper, we proposed to use per-instance differential privacy (pDP) for quantifying the fine-grained privacy loss of a fixed individual against randomized data analysis conducted on a fixed data set. We analyzed its properties and showed that pDP is proportional to well-studied quantities, e.g., leverage scores, residual and pseudo-residual in statistics and statistical learning theory. This formalizes the intuitive idea that the more one can "blend into the crowd" like a chameleon, the more privacy one gets; and that the better a model fits the data, the easier it is to learn the model differentially privately. Moreover, the new notion allows us to conduct statistical learning and inference and take advantage of desirable structures of the data sets to gain orders-of-magnitude more favorable privacy guarantee than the worst case. This makes it highly practical in applications.

Specifically, we conducted a detailed case-study on linear regression to illustrate how pDP can be used. The pDP analysis allows us to identify and account for key properties of the data set, like the well-conditionedness of the feature matrix and the magnitude of the fitted coefficient vector, thereby provides strong uniform differential privacy coverage to everyone in the population whenever such structures exist. As a byproduct, the analysis also leads to an improved differential privacy guarantee for the OPS algorithm [Dimitrakakis et al., 2014, Wang et al., 2015] and also a new algorithm called AdaOPS that adaptively chooses

the regularization parameters and improves the guarantee further. In particular, AdaOPS achieves asymptotic statistical efficiency and differential privacy at the same time with stronger parameters than known before.

The introduction of pDP also raises many open questions for future research. First of all, how do we tell individuals what their $\epsilon$s and $\delta$s of pDP are? This is tricky because the pDP loss itself is a function of the data, thus needs to be privatized against possible malicious dummy users. Secondly, the problem gets substantially more interesting when we start to consider the economics of private data collection. For instance, what happens if what we tell the individuals would affect their decision on whether they will participate in the data set? In fact, it is unclear how to provide an estimation of pDP in the first place if we are not sure what would the data be at the end of the day. Thirdly, from the data collector's point of view, the data is going to be "easier" and the model will have a better "goodness-of-fit" on the collected data, but that will be falsely so to some extent, due to the bias incurred during data collection according to pDP. How do we correct for such bias and estimate the real performance of a model on the population of interest? Addressing these problems thoroughly would require the joint effort of the community and we hope the exposition in this paper will encourage researchers to play with pDP in both theory and practical applications.

## Acknowledgment

The author thanks Steve Fienberg, Jing Lei, Ryan Tibshirani, Adam Smith, Jennifer Chayes and Christian Borgs for useful and inspiring discussions that motivate the work. We also thank the journal editor and anonymous reviewers for their helpful feedbacks that lead to significant improvements of the paper.

## References

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security (CCS-16)*, pages 308–318. ACM, 2016.

B. Balle and Y.-X. Wang. Improving gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. *International Conference in Machine Learning (ICML-18)*, 2018.

R. F. Barber and J. C. Duchi. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451*, 2014.

R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Foundations of Computer Science (FOCS-14)*, pages 464–473. IEEE, 2014.

P. J. Bickel. On adaptive estimation. *The Annals of Statistics*, pages 647–671, 1982.

J. Blocki, A. Blum, A. Datta, and O. Sheffet. The johnson-lindenstrauss transform itself preserves differential privacy. In *Foundations of Computer Science (FOCS-12)*, pages 410–419. IEEE, 2012.

M. Bun and T. Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

S. Chatterjee and A. S. Hadi. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, pages 379–393, 1986.

K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *The Journal of Machine Learning Research*, 12:1069–1109, 2011.

R. Cummings and D. Durfee. Individual sensitivity preprocessing for data privacy. *arXiv preprint arXiv:1804.08645*, 2018.

R. Cummings, S. Ioannidis, and K. Ligett. Truthful linear regression. In *Conference on Learning Theory*, pages 448–483, 2015.

C. Dimitrakakis, B. Nelson, A. Mitrokotsa, and B. I. Rubinstein. Robust and private Bayesian inference. In *Algorithmic Learning Theory*, pages 291–305. Springer, 2014.

P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13 (Dec):3475–3506, 2012.

C. Dwork and J. Lei. Differential privacy and robust statistics. In *ACM symposium on Theory of computing (STOC-09)*, pages 371–380. ACM, 2009.

C. Dwork and G. N. Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

C. Dwork and A. Smith. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2):2, 2010.

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

C. Dwork, G. N. Rothblum, and S. Vadhan. Boosting and differential privacy. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 51–60. IEEE, 2010.

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

C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth. Preserving statistical validity in adaptive data analysis. In *ACM symposium on Theory of computing (STOC-15)*, pages 117–126. ACM, 2015.

H. Ebadi, D. Sands, and G. Schneider. Differential privacy: Now it's getting personal. In *ACM SIGPLAN Notices*, volume 50, pages 69–81. ACM, 2015.

S. E. Fienberg, A. Rinaldo, and X. Yang. Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In *International Conference on Privacy in Statistical Databases*, pages 187–199. Springer, 2010.

J. Foulds, J. Geumlek, M. Welling, and K. Chaudhuri. On the theory and practice of privacy-preserving Bayesian data analysis. In *Conference on Uncertainty in Artificial Intelligence (UAI-16)*, pages 192–201. AUAI Press, 2016.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

A. Ghosh and A. Roth. Selling privacy at auction. *Games and Economic Behavior*, 91: 334–346, 2015.

R. Hall, L. Wasserman, and A. Rinaldo. Random differential privacy. *Journal of Privacy and Confidentiality*, 4(2), 2013.

D. Kifer, A. Smith, and A. Thakurta. Private convex empirical risk minimization and high-dimensional regression. *Journal of Machine Learning Research*, 1:41, 2012.

K. Ligett, S. Neel, A. Roth, B. Waggoner, and S. Z. Wu. Accuracy first: Selecting a differential privacy level for accuracy constrained erm. In *Advances in Neural Information*

*Processing Systems*, pages 2566–2576, 2017.

Z. Liu, Y.-X. Wang, and A. Smola. Fast differentially private matrix factorization. In *ACM Conference on Recommender Systems (RecSys-15)*, pages 171–178. ACM, 2015.

L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, J. A. Tropp, et al. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42(3): 906–945, 2014.

F. McSherry and I. Mironov. Differentially private recommender systems: building privacy into the netflix price contenders. In *International conference on Knowledge discovery and data mining (KDD-09)*, pages 627–636. ACM, 2009.

F. McSherry and K. Talwar. Mechanism design via differential privacy. In *Foundations of Computer Science (FOCS-07)*, pages 94–103. IEEE, 2007.

I. Mironov. Rényi differential privacy. In *Computer Security Foundations Symposium (CSF), 2017 IEEE 30th*, pages 263–275. IEEE, 2017.

K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *ACM symposium on Theory of computing (STOC-07)*, pages 75–84. ACM, 2007.

N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *ICLR*, 2016.

R. M. Rogers, A. Roth, J. Ullman, and S. Vadhan. Privacy odometers and filters: Pay-as-you-go composition. In *Advances in Neural Information Processing Systems*, pages 1921–1929, 2016.

O. Sheffet. Differentially private ordinary least squares. In *International Conference on Machine Learning (ICML-17)*, pages 3105–3114, 2017.

A. Smith. Efficient, differentially private point estimators. *arXiv preprint arXiv:0809.4794*, 2008.

D. A. Spielman and N. Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.

G. W. Stewart. Perturbation theory for the singular value decomposition. Technical report, 1998.

Y.-X. Wang, S. Fienberg, and A. Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning (ICML-15)*, pages 2493–2502, 2015.

Y.-X. Wang, J. Lei, and S. E. Fienberg. On-average kl-privacy and its equivalence to generalization for max-entropy mechanisms. In *International Conference on Privacy in Statistical Databases*, pages 121–134. Springer, 2016.

L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

F. Yu, M. Rybar, C. Uhler, and S. E. Fienberg. Differentially-private logistic regression for detecting multiple-snp association in gwas databases. In *International Conference on Privacy in Statistical Databases*, pages 170–184. Springer, 2014.

## Appendix A. Proofs of technical results

*Proof of Proposition 2.9.* We first show that the first moment of pDP implies on-average stability and then on-average stability implies on-average generalization.

Let $Z' = [Z, z']$, $Z'' = [Z, z'']$ and fix $z$. We first prove stability. Let $S = \theta | p(\theta) \geq p'(\theta)$

$$\begin{aligned}
&\left|\mathbb{E}_{\theta\sim\mathcal{A}(Z')}\ell(\theta, z) - \mathbb{E}_{\theta\sim\mathcal{A}(Z'')}\ell(\theta, z)\right| \\
=&\sup_{\theta,z} \ell(\theta, z)[P_{Z'}(\theta \in S) - P_{Z''}(\theta \in S)] \\
\leq& e^{\epsilon(Z,z')} P_Z(\theta \in S) + \delta((Z, z')) - P_{Z''}(\theta \in S) \\
\leq& (e^{\epsilon(Z,z')+\epsilon(Z,z'')} - 1) P_{Z''}(\theta \in S) + \delta(Z, z') + \epsilon(Z, z')\delta(Z, z'') \\
\leq& (e^{\epsilon(Z,z')+\epsilon(Z,z'')} - 1) + \delta(Z, z') + \epsilon(Z, z')\delta(Z, z'')
\end{aligned}$$

Note that the bound is independent to $z$.

Now we will show stability implies generalization using a "ghost sample" trick in which we resample $Z' \sim \mathcal{D}^n$ and construct $Z^{(i)}$ by replacing the $i$th data point from the $i$th data point of $Z'$.

$$\begin{aligned}
&\left|\mathbb{E}_{Z\sim\mathcal{D}^n}\left(\mathbb{E}_{\theta\sim\mathcal{A}(Z)}\mathbb{E}_{z\sim\mathcal{D}}\ell(\theta, z) - \mathbb{E}_{\theta\sim\mathcal{A}(Z)}\frac{1}{n}\sum_{i=1}^n \ell(\theta, z_i)\right)\right| \\
=&\left|\mathbb{E}_{Z\sim\mathcal{D}^n,\{z_1',\ldots,z_n'\}\sim\mathcal{D}^n}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\theta\sim\mathcal{A}(Z)}\ell(\theta, z_i') - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\theta\sim\mathcal{A}(Z^{(i)})}\ell(\theta, z_i')\right)\right| \\
\leq&\mathbb{E}_{Z\sim\mathcal{D}^n,\{z_1',\ldots,z_n'\}\sim\mathcal{D}^n}\frac{1}{n}\sum_{i=1}^n \left|\mathbb{E}_{\theta\sim\mathcal{A}(Z)}\ell(\theta, z_i') - \mathbb{E}_{\theta\sim\mathcal{A}(Z^{(i)})}\ell(\theta, z_i')\right| \\
\leq&\mathbb{E}_{Z\sim\mathcal{D}^{n-1},\{z',z''\}\sim\mathcal{D}^2}[(e^{\epsilon(Z,z')+\epsilon(Z,z'')} - 1) + \delta(Z, z') + \epsilon(Z, z')\delta(Z, z'')]
\end{aligned}$$

The last step simply substitutes the stability bound. Take expectation on both sides, we get a generalization upper bound of form:

$$\xi = \mathbb{E}_{Z\sim\mathcal{D}^n}(\mathbb{E}_{z\sim\mathcal{D}}[e^{\epsilon(Z,z)}|Z])^2 - 1 + \mathbb{E}_{Z\sim\mathcal{D}^n, z\sim\mathcal{D}}\delta(Z, z) + (\mathbb{E}_{Z\sim\mathcal{D}^n}\mathbb{E}_{z\sim\mathcal{D}}[e^{\epsilon(Z,z)}|Z]\mathbb{E}_{z\sim\mathcal{D}}[\delta(Z, z)|Z].$$

□

*Proof of Proposition 2.11.* The stability argument remains the same, because it is applied to a fixed pair of $(Z, z)$. We will modify the ghost sample arguments with and additional change of measure.

$$\begin{aligned}
&\left|\mathbb{E}_{Z\sim\mathcal{D}^n}\left(\mathbb{E}_{\theta\sim\mathcal{A}(Z)}\mathbb{E}_{z\sim\mathcal{D}'}\ell(\theta, z) - \mathbb{E}_{\theta\sim\mathcal{A}(Z)}\frac{1}{n}\sum_{i=1}^n \rho(z_i)\ell(\theta, z_i)\right)\right| \\
=&\left|\mathbb{E}_{Z\sim\mathcal{D}^n}\left(\mathbb{E}_{\theta\sim\mathcal{A}(Z)}\mathbb{E}_{z\sim\mathcal{D}}\rho(z)\ell(\theta, z) - \mathbb{E}_{\theta\sim\mathcal{A}(Z)}\frac{1}{n}\sum_{i=1}^n \rho(z_i)\ell(\theta, z_i)\right)\right| \\
=&\left|\mathbb{E}_{Z\sim\mathcal{D}^n,\{z_1',\ldots,z_n'\}\sim\mathcal{D}^n}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\theta\sim\mathcal{A}(Z)}\rho(z_i')\ell(\theta, z_i') - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\theta\sim\mathcal{A}(Z^{(i)})}\rho(z_i')\ell(\theta, z_i')\right)\right| \\
\leq&\mathbb{E}_{Z\sim\mathcal{D}^n,\{z_1',\ldots,z_n'\}\sim\mathcal{D}^n}\frac{1}{n}\sum_{i=1}^n \rho(z_i')\left|\mathbb{E}_{\theta\sim\mathcal{A}(Z)}\ell(\theta, z_i') - \mathbb{E}_{\theta\sim\mathcal{A}(Z^{(i)})}\ell(\theta, z_i')\right| \\
\leq&\mathbb{E}_{Z\sim\mathcal{D}^{n-1},\{z',z''\}\sim\mathcal{D}^2}\rho(z'')[(e^{\epsilon(Z,z')+\epsilon(Z,z'')} - 1) + \delta(Z, z') + \epsilon(Z, z')\delta(Z, z'')] \\
=&\mathbb{E}_{Z\sim\mathcal{D}^{n-1},z'\sim\mathcal{D},z''\sim\mathcal{D}'}[(e^{\epsilon(Z,z')+\epsilon(Z,z'')} - 1) + \delta(Z, z') + \epsilon(Z, z')\delta(Z, z'')].
\end{aligned}$$

□

*Proof of Corollary 2.12.*

$$\mathbb{E}\left[\mathbb{E}_{\mathcal{D}}[e^{\epsilon(Z,z)}|Z]\mathbb{E}_{\mathcal{D}'}[e^{\epsilon(Z,z)}|Z]\right] - 1 + \delta(1 + \mathbb{E}[e^{\epsilon}(Z,z)]) \leq \sqrt{\mathbb{E}_{\mathcal{D}}e^{2\epsilon}\mathbb{E}_{\mathcal{D}'}e^{2\epsilon}} - 1 + 2\delta.$$

The inequality uses Jensen's inequality $\mathbb{E}\left[\mathbb{E}[e^{\epsilon(Z,z)}|Z]^2\right] \leq \mathbb{E}e^{2\epsilon(Z,z)}$ and the monotonicity of moment generating function on non-negative random variables. The statement is obtained by Taylor's series on $\mathbb{E}e^{2\epsilon(Z,z)}$. Lastly, we use the algebraic mean to upper bound the geometric mean in the first term and then use Taylor expansion. □

*Proof of Lemma 3.1.* By the stationarity of $\hat{\theta}$

$$\sum_i \nabla\ell(\hat{\theta}, z_i) + \nabla r(\hat{\theta}) = 0$$

Add and subtract $\ell(\hat{\theta}, z)$ and apply first order Taylor's Theorem centered at $\hat{\theta}'$ on $\sum_i \nabla\ell(\hat{\theta}, z_i) + \nabla\ell(\hat{\theta}, z_i) + \nabla r(\hat{\theta})$, we get

$$\sum_i \nabla\ell(\hat{\theta}', z_i) + \nabla\ell(\hat{\theta}', z_i) + \nabla r(\hat{\theta}') + R - \nabla\ell(\hat{\theta}, z) = 0.$$

where if we define $\eta_t = (1-t)\hat{\theta}' + t\hat{\theta}$, the remainder term $R \in \mathbb{R}^d$ can be explicitly written as

$$R = \left[\int_0^1 \left(\sum_i \nabla^2\ell(\eta_t, z_i) + \nabla^2\ell(\eta_t, z) + \nabla^2 r(\eta_t)\right) dt\right](\hat{\theta} - \hat{\theta}').$$

By the mean value theorem for Frechet differentiable functions, we know there is a $t$ such that we can take $\eta_t$ such that the integrand is equal to the integral.

Since $\hat{\theta}'$ is a stationary point, we have

$$\sum_i \nabla\ell(\hat{\theta}', z_i) + \nabla\ell(\hat{\theta}', z) + \nabla r(\hat{\theta}') = 0$$

and thus under the assumption that $\left[\int_0^1 \left(\sum_i \nabla^2\ell(\eta_t, z_i) + \nabla^2\ell(\eta_t, z) + \nabla^2 r(\eta_t)\right) dt\right]$ is invertible, we have

$$\hat{\theta} - \hat{\theta}' = \left[\int_0^1 \left(\sum_i \nabla^2\ell(\eta_t, z_i) + \nabla^2\ell(\eta_t, z) + \nabla^2 r(\eta_t)\right) dt\right]^{-1} \nabla\ell(\hat{\theta}, z).$$

The other equality follows by symmetry. □

*Proof of Theorem 4.1.* Let $X' = [X; x]$, $\mathbf{y}' = [\mathbf{y}; y]$. Denote $H := X^TX + \lambda I$, $H' := (X')^TX' + \lambda I$, $g := X^T\mathbf{y}$ and $g' := (X')^T\mathbf{y}'$. Correspondingly, the posterior mean $\hat{\theta} = H^{-1}g$ and $\hat{\theta}' = [H']^{-1}g'$.

The covariance matrix of the two distributions are $H/\gamma$ and $H'/\gamma$. Using the fact that the normalization constant is known for Gaussian, the log-likelihood ratio at output $\theta$ is

$$\begin{aligned}&\log \frac{|H^{-1}|^{-1/2}e^{-\frac{\gamma}{2}\|\theta-\hat{\theta}\|_H^2}}{|[H']^{-1}|^{-1/2}e^{-\frac{\gamma}{2}\|\theta-\hat{\theta}'\|_{H'}^2}}\\ =&\underbrace{\log\sqrt{\frac{|H|}{|H'|}}}_{(\#)} + \frac{\gamma}{2}\underbrace{\left[\|\theta-\hat{\theta}'\|_{H'}^2 - \|\theta-\hat{\theta}\|_H^2\right]}_{(*)}.\end{aligned}$$

Note that $H' = H + xx^T$. By Lemma B.3,

$$\frac{|H|}{|H'|} = \frac{|H|}{|H|(1+\mu)} = \frac{|H'|(1-\mu')}{|H'|},$$

so

$$(\#) = \log\sqrt{(1+\mu)^{-1}} = \log\sqrt{1-\mu'}.$$

The second term in the above equation can be expanded into

$$\begin{aligned}(*) =&\theta^T[H'-H]\theta + (\hat{\theta}')^TH'\hat{\theta}' - \hat{\theta}^TH\hat{\theta} - 2(\hat{\theta}')^TH'\theta + 2\hat{\theta}^TH\theta\\ =&(x^T\theta)^2 + \underbrace{(\mathbf{y}')^TX'[H']^{-1}X'^T\mathbf{y}' - \mathbf{y}^TX(H)^{-1}X^T\mathbf{y}}_{(**)} - 2y(x^T\theta) \qquad \text{(A.1)}\end{aligned}$$

$(**) = \left[(\mathbf{y}')^TX'[(X')^TX' + \lambda I]^{-1}X'^T\mathbf{y}' - \mathbf{y}^TX(X^TX+\lambda I)^{-1}X^T\mathbf{y}\right] = \left[(\mathbf{y}')^T\Pi'\mathbf{y}' - \mathbf{y}^T\Pi\mathbf{y}\right]$, where we denote the"hat" matrices $\Pi := X(X^TX+\lambda I)^{-1}X^T$ and $\Pi' = X'[(X')^TX' + \lambda I]^{-1}(X')^T$. Also define $v := X(X^TX+\lambda I)^{-1}x$. By Sherman-Morrison-Woodbury formula, we can write

$$\begin{aligned}\Pi' =& \begin{bmatrix}X\\ x^T\end{bmatrix}\left[H^{-1} - H^{-1}x(1+\mu)^{-1}x^TH^{-1}\right]\begin{bmatrix}X^T & x\end{bmatrix}\\ =&\begin{bmatrix}\Pi - (1+\mu)^{-1}vv^T, & v - \mu(1+\mu)^{-1}v\\ v^T - v^T(1+\mu)^{-1}\mu, & \mu - \mu^2(1+\mu)^{-1}\end{bmatrix}\end{aligned}$$

Note that $v^Ty = x^T\hat{\theta}$ and $1 - \mu(1+\mu)^{-1} = (1+\mu)^{-1}$, therefore

$$\begin{aligned}(**) =& -(1+\mu)^{-1}(x^T\hat{\theta})^2 + 2(1+\mu)^{-1}x^T\hat{\theta} + \mu(1+\mu)^{-1}y^2\\ =&-(1+\mu)^{-1}(y - x^T\hat{\theta})^2 + y^2.\end{aligned}$$

Substitute into (A.1), we get

$$(*) = (y - x^T\theta)^2 - (1+\mu)^{-1}(y-x^T\hat{\theta})^2.$$

And the log-probability ratio is

$$\log \frac{p(\theta|X,\mathbf{y})}{p(\theta|X',\mathbf{y}')} = \log\sqrt{(1+\mu)^{-1}} + \frac{\gamma}{2}\left[(y-x^T\theta)^2 - (1+\mu)^{-1}(y-x^T\hat{\theta})^2\right]$$
$$= \log\sqrt{(1+\mu)^{-1}} + \frac{\gamma}{2}\left[(x^T\hat{\theta}-x^T\theta)^2 + 2(y-x^T\hat{\theta})(x^T\hat{\theta}-x^T\theta) + \frac{\mu}{1+\mu}(y-x^T\hat{\theta})^2\right]$$

Under the distribution of $\theta$ when the data is $(X,\mathbf{y})$, $x^T\theta - x^T\hat{\theta}$ follows a univariate normal distribution with mean 0 and variance $\mu/\gamma$. By the standard tail probability of normal random variable,

$$\mathbb{P}\left(|x^T\theta - x^T\hat{\theta}| > \sqrt{\frac{\mu}{\gamma}\log(2/\delta)}\right) \le \frac{2e^{-\log(2/\delta)}}{\log(2/\delta)} = \frac{\delta}{\log(2/\delta)} \underset{\substack{\uparrow \\ \text{When } \delta < 2/e}}{\le} \delta.$$

we can calculate $(\epsilon,\delta)$-pDP for every $\delta > 0$. In particular, under $p(\theta|X,y)$

$$\mathbb{P}(\left|\log\frac{p(\theta|X,\mathbf{y})}{p(\theta|X',\mathbf{y}')}\right| \ge \epsilon) < \delta$$

for

$$\epsilon = \frac{1}{2}\left|-\log(1+\mu) + \frac{\mu\gamma}{(1+\mu)}(y-x^T\hat{\theta})^2\right| + \frac{\mu}{2}\log(2/\delta) + |y-x^T\hat{\theta}|\sqrt{\mu\gamma\log(2/\delta)}.$$

By Lemma B.6 this implies $(\epsilon,\delta)$-DP.

Now, we will work out an equivalent representation of the log-probability ratio that depends on $\hat{\theta}'$.

Let $\mu'$ be the in-sample leverage score of $x$ with respect to $X'$, namely, $\mu' := x^T[H']^{-1}x$. By Sherman-Morrison-Woodbury formula

$$H^{-1} = [H' - xx^T]^{-1} = [H']^{-1} + [H']^{-1}x(1-\mu')^{-1}x^T[H']^{-1}.. \tag{A.2}$$

Standard matrix algebra gives us

$$\mathbf{y}^T\Pi\mathbf{y} = (\mathbf{y}')^TX'H^{-1}(X')^T\mathbf{y}' - yx^TH^{-1}xy - 2yx^TH^{-1}X^T\mathbf{y}$$
$$= (\mathbf{y}')^TX'H^{-1}(X')^T\mathbf{y}' - 2yx^TH^{-1}(X')^T\mathbf{y}' + yx^TH^{-1}xy.$$

Substitute (A.2) into the above, we get

$$\mathbf{y}^T\Pi\mathbf{y} = (\mathbf{y}')^T\Pi'\mathbf{y}' + (1-\mu')^{-1}(x^T\hat{\theta}')^2 - 2yx^T\hat{\theta}'\left[1+\mu'(1-\mu')^{-1}\right] + y^2\mu' + y^2(\mu')^2(1-\mu')^{-1}$$
$$= (\mathbf{y}')^T\Pi'\mathbf{y}' + (1-\mu')^{-1}(x^T\hat{\theta}')^2 - 2yx^T\hat{\theta}'(1-\mu')^{-1} + y^2(1-\mu')^{-1} - y^2$$

Therefore,

$$(**) = -(y-x^T\hat{\theta}')^2(1-\mu')^{-1} + y^2,$$

and

$$(*) = (y-x^T\theta)^2 - (1-\mu')^{-1}(y-x^T\hat{\theta}')^2.$$

The corresponding log-probability ratio

$$\log\frac{p(\theta|X,\mathbf{y})}{p(\theta|X',\mathbf{y}')} = -\log(\sqrt{1-\mu'}) + \frac{\gamma}{2}\left[(y-x^T\theta)^2 - (1-\mu')^{-1}(y-x^T\hat{\theta}')^2\right]$$
$$= -\log(\sqrt{1-\mu'}) + \frac{\gamma}{2}\left[(x^T\hat{\theta}'-x^T\theta)^2 + 2(x^T\hat{\theta}'-x^T\theta)(y-x^T\hat{\theta}') - \frac{\mu'}{1-\mu'}(y-x^T\hat{\theta}')^2\right].$$

Under the posterior distribution of $(X',\mathbf{y}')$, the mean of $x^T\theta$ is centered at $x^T\hat{\theta}'$ with variance $\mu'/\gamma$. We can then derive a tail bound of the privacy loss random variable and it implies an $(\epsilon,\delta)-pDP$ guarantee by Lemma B.6. Specifically, it implies that the method is $(\epsilon,\delta)$-pDP with

$$\epsilon = \frac{1}{2}\left|-\log(1-\mu') - \frac{\gamma\mu'}{1-\mu'}(y-x^T\hat{\theta}')^2\right| + \frac{\mu'}{2}\log(2/\delta) + \sqrt{\gamma\mu'\log(2/\delta)}|y-x^T\hat{\theta}'|.$$

This complete the second statement of the proof. □

*Proof of Proposition 4.3.* The proof mostly involves applying Theorem 4.1 and substituting bounds over either a bounded domain assumption (typical for DP analysis), or a model assumption of how data are generated (typical for statistical analysis).

**Proof of Statement 1 in the agnostic setting.** For any $x\in\mathcal{X}$, and any data set $X$, using the choice of regularization term, we can bound $\mu = 1/\lambda_n$. Substitute that into Theorem 4.1, and use the inequality that $\log(1+x)\le x$ we get the first expression.

Now, restricting ourselves to the bounded domain. Under the choice of $\lambda_n$, we can choose an $X$,$\mathbf{y}$ with a singular value equal to $\sqrt{\lambda_n}$ and the corresponding singular vector $v\in\{-1,1\}^n$ such that the following upper bounds are attained

$$\|(X^TX+\lambda_nI)^{-1}X^T\| \le \frac{1}{2\sqrt{\lambda_n}}.$$
$$\|\hat{\theta}\| \le \|(X^TX+\lambda_nI)^{-1}X^T\|\|y\| \le \frac{\sqrt{n}}{2\sqrt{\lambda_n}}.$$

Now choose $(x,y)$ such that $|x^T\hat{\theta}| = \|x\|\|\hat{\theta}\|$, we get that $\sup_{(X,\mathbf{y}),(x,y)}|y-\hat{\theta}^Tx| = 1+\frac{\sqrt{n}}{2\sqrt{\lambda_n}}$. The DP claim follows by substituting the upper bound into the pDP's expression.

**Proof of Statement 2 under the model assumption.** To prove the second claim, note that by Assumption (b)(d), the smallest eigenvalue of $X^TX$ is lower bounded by $d/nm$. Also under the model assumption, the ridge regression estimator concentrates around $\theta_0$.

In particular, under the model assumption, the ridge regression estimate

$$\hat{\theta} = (X^TX+\lambda_nI)^{-1}X^Ty = (X^TX+\lambda_nI)^{-1}X^TX\theta_0 + (X^TX+\lambda_nI)^{-1}X^TZ \qquad (A.3)$$

With high probability over the distribution of $Z$

$$\|\hat{\theta}-\theta_0\|^2 = O(\frac{d\sigma^2\log(n)}{n} + \frac{\lambda_n^2d^2\|\theta_0\|^2}{n^2}),$$

thus for all $(x,y)$ satisfying $\|x\|\le 1$ $y\le 1$, we get

$$|y-x^T\hat{\theta}| \le |y-x^T\theta_0| + |x^T(\hat{\theta}-\theta_0)| = O(1+\|\theta_0\|).$$

Under the assumption that $n > 10d \log n$, $\|\theta_0\| = O(1)$ and $\sigma = O(1)$ this is effectively a constant.

For $x \in \mathcal{X}$ and $y \sim \mathcal{N}(\theta_0^T x, \sigma^2)$, using standard Gaussian tail bound, with high probability the perturbation is bounded, therefore $|y - x^T\theta_0|^2 \le \sigma^2 \log(2/\delta')$.

Lastly, we address the case of the average pDP loss over the empirical data distribution. Besides taking into the above bound on $|y - x^T\hat{\theta}|$, we further consider adding the different parts over the distributions. Since this is to deal with data points in the data set, we will instantiate the bound (4.4). Our assumption on $\gamma_n$, $\lambda_n$ ensures that the dominant term is the third term, thus

$$\frac{1}{n}\sum_{i=1}^{n} \epsilon_n((X, \mathbf{t}), (x_i, y_i))^2 \le C\gamma_n \frac{1}{n}\sum_{i=1}^{n} (y_i - x_i^T\hat{\theta}')^2 x_i^T (X^TX + \lambda_n I)^{-1} x_i.$$

Under the high probability event that the noise is bounded by $\sigma\sqrt{2/\delta'}$ for all data points, we can extract them out then note that

$$\frac{1}{n}\sum_{i=1}^{n} x_i^T (X^TX)^{-1} x_i = \frac{1}{n}\mathrm{tr}(\sum_{i=1}^{n} x_i x_i^T (X^TX + \lambda_n I)) \le \frac{1}{n}\mathrm{tr}(I) = \frac{d}{n}.$$

Substitute these bounds into Theorem 4.1, and we obtain the Statement 2.

**Proof of Statement 3 under the model assumption.** By (A.3) and the fact that OPS can be thought of as adding an independent multivariate Gaussian noise with covariance matrix $(X^TX + \lambda_n I)^{-1}X^TX(X^TX + \lambda_n I)^{-1}/\gamma_n$, we get

$$\tilde{\theta} = (X^TX + \lambda_n I)^{-1}X^TX\theta_0 + \sqrt{1+\gamma_n}(X^TX + \lambda_n I)^{-1}X^TZ.$$

By a bias-variance decomposition, we get

$$\begin{aligned}
&\mathbb{E}(\|\tilde{\theta} - \theta_0\|_2^2 | X) = \mathrm{Var}(\tilde{\theta}|X) + \|\mathbb{E}\tilde{\theta} - \theta_0\|^2 \\
=&(1+\gamma_n^{-1})\sigma^2 \mathrm{tr}\left[(X^TX + \lambda_n I)^{-1}X^TX(X^TX + \lambda_n I)^{-1}\right] + \left\|\left[I - (X^TX + \lambda_n I)^{-1}X^TX\right]\theta_0\right\|^2 \\
=&(1+\gamma_n^{-1})\sigma^2 \sum_{i=1}^{d} \frac{\sigma_i^2}{(\sigma_i^2 + \lambda_n)^2} + \lambda_n^2 \theta_0^T (X^TX + \lambda_n I)^{-2}\theta_0 \\
\le&(1+\gamma_n^{-1})\sigma^2 \mathrm{tr}(X^TX + \lambda_n I)^{-1} + \lambda_n^2 m^{-2} n^{-2} \|\theta_0\|^2
\end{aligned}$$

The proof is complete by substitute the values of $\gamma_n$ and $\lambda_n$ into the inequality and noting that $m = \Omega(1)$ and under the model assumption $\|\theta_0\|$ does not grow with $n$. Clearly, if $\gamma_n = \omega(1)$ and $\lambda_n = o(\sqrt{n})$, then the algorithm is asymptotically efficient. □

*Proof of Proposition 4.4.* We will first prove the claim on differential privacy and then analyze the statistical efficiency. **Proof of differential privacy.** First of all, by Weyl's theorem, and the assumption that $\|xx^T\|_2 \le 1$, we get that the global sensitivity of $\lambda_{\min}(X^TX)$ is 1. We will use $\lambda_{\min}$ as the short hand of $\lambda_{\min}(X^TX)$ in the rest of the proof. So releasing $\tilde{\lambda}_{\min}$ is $(\epsilon/2, \delta/2)$-DP using the standard Gaussian mechanism. Secondly, under the same event with probability at least $1 - \delta/2$, we have

$$\lambda_{\min} - \frac{\log(4/\delta)}{\epsilon} \le \tilde{\lambda}_{\min} \le \lambda_{\min} + \frac{\log(4/\delta)}{\epsilon}.$$

Therefore, by our selection rule of the regularization parameter $\lambda_n$,

$$\frac{n}{d\kappa} \le \lambda_{\min}(X^TX + \lambda_n I) \le \max\{\lambda_{\min}, \frac{n}{d\kappa} + \frac{\log(4/\delta)}{\epsilon}\}.$$

The lower bound implies that for any $(x, y)$ satisfying the condition, the out of sample leverage score

$$\mu = x^T(X^TX + \lambda_n I)^{-1}x \le \frac{\kappa d}{n}. \tag{A.4}$$

It also implies an upper bound on the prediction error:

$$|y - x^T\hat{\theta}| \le 1 + \|\hat{\theta}\| \le 1 + \|(X^TX + \lambda_n I)^{-1}X^T\|_2\|y\|_2 \le \min\sqrt{2d\kappa}. \tag{A.5}$$

We will prove the final inequality above using the following lemma with $h = n/d\kappa$ and then note that $\|y\|_2 \le \sqrt{n}$.

**Lemma A.1.** *For any matrix $X$, and any $\lambda \ge 0$. If $\lambda_{\min}(X^TX + \lambda I) \ge h$, then*

$$\|(X^TX + \lambda I)^{-1}X^T\| \le \sqrt{2/h}.$$

The proof is technical so we defer it to later.

Now combine (A.4)(A.5) with Theorem 4.1, we get that the OPS step which obeys $(\tilde{\epsilon}, \delta/2)$-pDP with

$$\begin{aligned}\tilde{\epsilon}((X, \mathbf{y}), (x, y)) &\le \frac{\mu}{2}(1 + \log(4/\delta)) + \frac{1}{2}\gamma_n \min(\mu, 1)(y - x^T\hat{\theta})^2 + \sqrt{\gamma\mu\log(4/\delta)}|y - x^T\hat{\theta}| \\ &\le \frac{\kappa d(1 + \log(4/\delta))}{2n} + \frac{\gamma_n}{2}\frac{\kappa d}{n}2\kappa d + \sqrt{\frac{\gamma_n}{2}\frac{\kappa d}{n}2\kappa d\log(4/\delta)} \\ &\le \epsilon/8 + \epsilon/8 + \epsilon/4 \le \epsilon/2.\end{aligned}$$

Note that in the last step, we made use of the choice of $\gamma_n$ and the condition that concerns $\epsilon$ and $\kappa$ as stated in the algorithm. Since this upper bound holds for all data set $(X, \mathbf{y})$ and all privacy target $(x, y)$. The OPS algorithm also satisfies $(\epsilon/2, \delta/2) - DP$.

The proof of the first claim is complete when we compose the two data access.

**Proof of the statistical efficiency.** Now we switch gears to analyze the estimation error bound. Let event $E$ be the event that $\tilde{\lambda}_{\min} > \lambda_{\min} - \frac{\sqrt{10\log(n)}\sqrt{\log(4/\delta)}}{\epsilon/2}$, which happens with probability $1 - n^{-10}$. Under $E$, we have $\lambda_0 = 0$. By our assumption, this happens with

Applying the third claim in Proposition 4.3, we get that

$$\mathbb{E}[\|\tilde{\theta} - \theta_0\|^2|X, E] \le (1 + \gamma_n^{-1})\sigma^2\mathrm{tr}((X^TX)^{-1}).$$

Under the small probability event $E^c$, we use a crude upper bound that takes the sum of the maximum square bias and maximum variance.

$$\mathbb{E}[\|\tilde{\theta} - \theta_0\|^2|X, E^c] \le (1 + \gamma_n^{-1})\sigma^2\mathrm{tr}((X^TX)^{-1}) + \|\theta_0\|^2$$

by law of total expectation, for an event $E$

$$\begin{aligned}&\mathbb{E}[\|\tilde{\theta} - \theta_0\|^2|X] = \mathbb{E}[\|\tilde{\theta} - \theta_0\|^2|X, E]\mathbb{P}(E|X) + \mathbb{E}[\|\tilde{\theta} - \theta_0\|^2|X, E^c]\mathbb{P}(E^c|X) \\ \le&(1 + \gamma_n^{-1})\sigma^2\mathrm{tr}((X^TX)^{-1}) + \mathbb{P}(E^c)\|\theta_0\|^2 = (1 + \gamma_n^{-1})\sigma^2\mathrm{tr}((X^TX)^{-1}) + O(n^{-10}).\end{aligned}$$

The proof is complete by substituting $\gamma_n$ into the bound. □

*Proof of Lemma A.1.* Take SVD of $X = U\Sigma V^T$, we can write

$$\|(X^TX + \lambda_n I)^{-1}X^T\|_2 = \max_{i\in[d]} \frac{\Sigma_{ii}}{\Sigma_{ii}^2 + \lambda_n}$$

We now discuss two cases. First, for those $i \in [d]$ such that $\Sigma_{ii}^2 \le \lambda_n$. In this case, adding $\lambda_n$ on both sides ensures that

$$h \le \lambda_{\min}(X^TX + \lambda_n I) = \lambda_{\min} + \lambda_n \le \Sigma_{ii}^2 + \lambda_n \le 2\lambda_n.$$

and therefore if $\Sigma_{ii} > 0$

$$\frac{\Sigma_{ii}}{\Sigma_{ii}^2 + \lambda_n} = \frac{1}{\Sigma_{ii} + \lambda_n/\Sigma_{ii}} \le \frac{1}{2\sqrt{\lambda_n}} \le \sqrt{1/(2h)}. \tag{A.6}$$

The final inequality is also true for $\Sigma_{ii} = 0$. If on the other hand, for those $i \in [d]$ such that, $\Sigma_{ii}^2 > \lambda_n$. This time by adding $\Sigma_{ii}^2$ on both sides, we get

$$2\Sigma_{ii}^2 > \lambda_n + \Sigma_{ii}^2 \ge \lambda_n + \lambda_{\min} = \lambda_{\min}(X^TX + \lambda_n I) \ge \frac{n}{\kappa}.$$

This implies that

$$\frac{\Sigma_{ii}}{\Sigma_{ii}^2 + \lambda_n} \le \frac{1}{\Sigma_{ii}} \le \sqrt{2/h}. \tag{A.7}$$

Combine (A.6) and (A.7) we get

$$\|(X^TX + \lambda_n I)^{-1}X^T\|_2 \le \sqrt{2/h}$$

□

## Appendix B. Technical lemmas

**Lemma B.1.** *Let $\hat{\theta}' = (X^TX + E_1)^{-1}(X\mathbf{y} + E_2)$ for any matrix $E_1$, $E_2$.*

$$\hat{\theta}' - \hat{\theta} = (X^TX + E_1)^{-1}(E_2 - E_1\hat{\theta})$$

*Proof.*

$$\begin{aligned}
\hat{\theta}' =& (X^TX + E_1)^{-1}(X^T\mathbf{y} + E_2) \\
=& (X^TX + E_1)^{-1}(X^TX)(X^TX)^{-1}X^T\mathbf{y} + (X^TX + E_1)^{-1}E_2 \\
=& \hat{\theta} + \left[(X^TX + E_1)^{-1}(X^TX + E_1) - (X^TX + E_1)^{-1}E_1 - I_d\right]\hat{\theta} + (X^TX + E_1)^{-1}E_2 \\
=& \hat{\theta} - (X^TX + E_1)^{-1}E_1\hat{\theta} + (X^TX + E_1)^{-1}E_2 \\
=& \hat{\theta} + (X^TX + E_1)^{-1}(E_2 - E_1\hat{\theta})
\end{aligned}$$

□

**Lemma B.2** (Sherman-Morrison-Woodbury Formula)**.** *Let $A, U, C, V$ be matrices of compatible size, assume $A, C$ and $C^{-1} + VA^{-1}U$ are all invertible, then*

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

**Lemma B.3** (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 B.4** (Weyl's eigenvalue bound [Stewart, 1998, Theorem 1])**.** *Let $X, Y, E \in \mathbb{R}^{m\times n}$, w.l.o.g., $m \geq n$. If $X - Y = E$, then $|\sigma_i(X) - \sigma_i(Y)| \leq \|E\|_2$ for all $i = 1, ..., n$.*

**Lemma B.5** (Gaussian tail bound)**.** *Let $X \sim \mathcal{N}(0, 1)$. Then*

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

**Lemma B.6** (Tail bound to $(\epsilon, \delta)$-DP conversion)**.** *Let $\epsilon(\theta) = \log(\frac{p(\theta)}{p'(\theta)})$ where $p$ and $p'$ are densities of $\theta$. If*

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

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

$$\mathbb{P}_p(\theta \in \mathcal{S}) \leq e^t \mathbb{P}_{p'}(\theta \in \mathcal{S}) + \delta.$$

*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 $\mathcal{S}$:

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

□

**Lemma B.7** (Matrix Hoeffding inequality [Mackey et al., 2014])**.** *Consider a finite sequence $X_1, ..., X_n$ of independent random and self-adjoint matrices with dimension $d$ and $A_1, ..., A_n$ be a sequence of fixed self-adjoint matrices. In addition, let $\mathbb{E}X_i = 0$ and $X_i^2 \preceq A_i^2$ almost surely for all $i = 1, ..., n$. Then, for all $t \geq 0$*

$$\mathbb{P}\left\{\lambda_{\max}(\sum_{i=1}^n X_i) \geq t\right\} \leq de^{-t^2/2\sigma^2}$$

*where $\sigma^2 \leq \|\sum_{i=1}^n A_i^2\|$.*