Generalized PTR: User-Friendly Recipes for Data-Adaptive Algorithms with Differential Privacy

Anonymous Author
Anonymous Institution

Abstract

The "Propose-Test-Release" (PTR) framework [4] is a classic recipe for designing differentially private (DP) algorithms that are data-adaptive, i.e. those that add less noise when the input dataset is "nice". We extend PTR to a more general setting by privately testing *data-dependent privacy losses* rather than *local sensitivity*, hence making it applicable beyond the standard noise-adding mechanisms, e.g. to queries with unbounded or undefined sensitivity. We demonstrate the versatility of generalized PTR using private linear regression as a case study. Additionally, we apply our algorithm to solve an open problem from "Private Aggregation of Teacher Ensembles (PATE)" [18, 19] — privately releasing the entire model with a delicate data-dependent analysis.

1 Introduction

The guarantees of differential privacy (DP) [5] are based on worst-case outcomes across all possible datasets. A common paradigm is therefore to add noise scaled by the *global sensitivity* of a query f, i.e. the maximum change in f between any pair of neighboring datasets.

A given dataset X might have a *local sensitivity* $\Delta_{LS}(X)$ that is much smaller than the global sensitivity Δ_{GS}, in which case we can hope to add a smaller amount of noise (calibrated to the local rather than global sensitivity) while achieving the same privacy guarantee. However, this must not be undertaken naïvely – the local sensitivity is a dataset-dependent function and so calibrating noise to the local sensitivity could leak information about the dataset [16].

The "Propose-Test-Release" (PTR) framework [4] resolves this issue by introducing a test to privately check whether a proposed bound on the local sensitivity is valid. Only if the test "passes" is the output released with noise calibrated to the proposed bound on the local sensitivity.

PTR is a powerful and flexible tool for designing data-adaptive DP algorithms, but it has several limitations. First, it applies only to noise-adding mechanisms which calibrate noise according to the sensitivity of a query. Second, the test in "Propose-Test-Release" is computationally expensive for all but a few simple queries such as privately releasing the median or mode. Third, while some existing works [3, 9, 12] follow the approach of testing "nice" properties of a dataset before exploiting these properties in a private release to PTR [1], there has not been a systematic recipe for *discovering* which properties should be tested.

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

Our contributions are summarized as follows:

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

2. We demonstrate that many existing examples of PTR and PTR-like algorithms can be unified under the generalized PTR framework, sometimes resulting in a tighter analysis (see an example of report-noisy-max in Section 9.1).

Preliminary work. Under review by AISTATS 2023. Do not distribute.

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

3. We show that one can publish a DP model through privately upper-bounding a one-dimensional statistic — no matter how complex the output space of the mechanism is. We apply this result to solve an open problem from PATE [18, 19].

4. Our results broaden the applicability of private hyper-parameter tuning [11, 17] in enabling joint-parameter selection of DP-specific parameters (e.g., noise level) and native parameters of the algorithm (e.g., learning rate, regularization weight), which may jointly affect the data-dependent DP losses.

2 Related Work

Data-dependent DP algorithms. Privately calibrating noise to the local sensitivity is a well-studied problem. One approach is to add noise calibrated to the smooth sensitivity [16], an upper bound on the local sensitivity which changes slowly between neighboring datasets. An alternative to this – and the focus of our work – is Propose-Test-Release (PTR) [4], which works by calculating the distance $\mathcal{D}_\beta(X)$ to the nearest dataset to X whose local sensitivity violates a proposed bound β. The PTR algorithm then adds noise to $\mathcal{D}_\beta(X)$ before testing whether this privately computed distance is large enough to permit releasing the output with noise calibrated to β.

PTR spin-offs abound. Notable examples include stability-based methods [22] (stable local sensitivity of 0 near the input data) and privately releasing upper bounds of local sensitivity [9, 12, 3]. We refer readers to Chapter 3 of Vadhan [23] for a concise summary of these classical results. Recent work [24] has provided Rényi DP bounds [14] for PTR and demonstrated its applications to robust DP-SGD. Our work (see Section 5.2) also considers applications of PTR in data-adaptive private deep learning: Instead of testing the local sensitivity of each gradient step as in Wang et al. [24], our PTR-based PATE algorithm tests the data-dependent privacy loss as a whole.

Liu et al. [12] proposed a new variant called High-dimensional Propose-Test-Release (HPTR). HPTR provides a systematic way of solving DP statistical estimation problems by using the exponential mechanism (EM) with carefully constructed scores based on certain one-dimensional robust statistics, which have stable local sensitivity bounds. HPTR focuses on designing data-adaptive DP mechanisms from scratch; our method, in contrast, converts existing randomized algorithms (including EM and even some that do not satisfy DP) into those with formal DP guarantees. Interestingly, our proposed method also depends on a one-dimensional statistic of direct interest: the data-dependent privacy loss.

Data-dependent DP losses. The flip side of data-dependent DP algorithms is the study of data-dependent DP losses [19, 21, 25], which fix the randomized algorithm but parameterize the resulting privacy loss by the specific input dataset. For example: In the simple mechanism that adds Laplace noise with parameter b, data-dependent DP losses are $\epsilon(X) = \Delta_{LS}(X)/b$. The data-dependent DP losses $\epsilon(X)$ are often much smaller than the DP loss ϵ, but they themselves depend on the data and thus may reveal sensitive information; algorithms satisfying a data-dependent privacy guarantee are not formally DP with guarantees any smaller than that of the worst-case. Existing work has considered privately publishing these data-dependent privacy losses [19, 20], but notice that privately publishing these losses does not improve the DP parameter of the given algorithm. Part of our contribution is to resolve this conundrum by showing that a simple post-processing step of the privately released upper bound of $\epsilon(X)$ gives a formal DP algorithm.

Private hyper-parameter tuning. Our work has a nice connection with private hyper-parameter tuning. Prior work [11, 17] requires each candidate configuration to be released with the same DP (or Rényi DP) parameter set. Another hidden assumption is that the parameters must not be privacy-correlated (i.e., parameter choice will not change the privacy guarantee). Otherwise we need to use the largest DP bound across all candidates. For example, Liu and Talwar [11] show that if each mechanism (instantiated with one group of hyper-parameters) is $(\epsilon, 0)$-DP, then running a random number of mechanisms and reporting the best option satisfies $(3\epsilon, 0)$-DP. Our work directly generalizes the above results by (1) considering a wide range of hyper-parameters, either privacy-correlated or not; and (2) requiring only that individual candidates have a *testable* data-dependent DP.

3 Preliminaries

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

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

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

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

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

Definition 3.1 (Differential privacy [5])**.** Fix $\epsilon, \delta \geq 0$. A randomized algorithm $\mathcal{M} : \mathcal{X} \to \mathcal{R}$ satisfies (ϵ, δ)-DP if for all neighboring datasets $X \simeq X'$ and for all measurable sets $S \subseteq \mathcal{R}$,

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

Definition 3.2 (Sensitivity)**.** The global ℓ_*-sensitivity of a

function f is defined as

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

and its local sensitivity at dataset X is

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

Theorem 3.3 (Noise-adding mechanisms)**.** *Consider a real-valued function $f : \mathcal{X} \to \mathbb{R}$ with global ℓ_1-sensitivity Δ_1 and global ℓ_2-sensitivity Δ_2.*

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

The Gaussian mechanism $\mathcal{M}(X) = f(X) + \mathcal{N}(0, \sigma^2)$ satisfies (ϵ, δ)-differential privacy with noise parameter $\sigma = \Delta_2\sqrt{2\log(1.25/\delta)}/\epsilon$.

3.1 Propose-Test-Release

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

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

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

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

Algorithm 1 Propose-Test-Release [4]

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

Theorem 3.4. *Algorithm 1 satisfies $(2\epsilon, \delta)$-DP. [4]*

Rather than proposing an arbitrary bound β on $\Delta_{LS}(X)$, one can also privately release an upper bound of the local sensitivity and calibrate noise according to this upper bound. This was used for node DP in graph statistics [9], and for fitting topic models using spectral methods [3].

4 Generalized PTR

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

Definition 4.1 (Data-dependent privacy)**.** Suppose we have $\delta > 0$ and a function $\epsilon : \mathcal{X} \to \mathbb{R}^+$. We say that mechanism $\mathcal{M}$ satisfies $(\epsilon(X), \delta)$ data-dependent DP[2] for dataset X if for all possible output sets S and neighboring datasets X',

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

In generalized PTR, we propose a value ϕ for the randomized algorithm $\mathcal{M}$, which could be a noise scale or regularization parameter – or a set including both. For example, the parameter set is $\phi = (\lambda, \gamma)$ in Example 4.4. We then say that $\mathcal{M}_\phi$ is the mechanism $\mathcal{M}$ parameterized by ϕ, with $\epsilon_\phi(X)$ its data-dependent DP.

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

Example 4.2. *(*Data-dependent DP of Laplace Mechanism.*) Given a function $f : \mathcal{X} \to \mathbb{R}$, we will define*

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

We then have

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

Maximizing the above calculation over all possible outputs y yields an equality between the two expressions. So using Definition 4.1,

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

The data-dependent DP $\epsilon_\phi(X)$ is a function of both the dataset X and the parameter ϕ. Maximizing $\epsilon_\phi(X)$ over X recovers the standard DP guarantee of running $\mathcal{M}$ with parameter ϕ.

Algorithm 2 distills the generalized PTR framework into a simple procedure: we run mechanism $\mathcal{M}$ with proposed parameter ϕ only if the test $\mathcal{T}$ "passes".

Let's suppose that our privacy budget for mechanism $\mathcal{M}_\phi$ is (ϵ, δ); that our test $\mathcal{T}$ satisfies $(\hat{\epsilon}, \hat{\delta})$-DP; and that $\mathcal{T}$ has a "false positive" rate δ', meaning $\mathcal{T}$ passes an insufficient proposal ϕ (where $\mathcal{M}_\phi$ exceeds its privacy budget) with probability at most δ'. Theorem 4.3 states the privacy guarantee of generalized PTR under these assumptions.

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

Algorithm 2 Generalized Propose-Test-Release

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

Theorem 4.3 (Privacy guarantee of generalized PTR)**.** *Consider a proposal ϕ and a data-dependent DP function $\epsilon_\phi(X)$ w.r.t. δ. Suppose that we have an $(\hat{\epsilon}, \hat{\delta})$-DP test $\mathcal{T} : \mathcal{X} \to \{0, 1\}$ such that when $\epsilon_\phi(X) > \epsilon$,*

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

Then Algorithm 2 satisfies $(\epsilon + \hat{\epsilon}, \delta + \hat{\delta} + \delta')$-DP.

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

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

Full details of the proof are provided in the appendix. □

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

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

Notice that our choice of parameterization is $\phi = \frac{\beta}{\epsilon}$, where ϕ is the scale of the Laplace noise. In other words, we know from Example 4.2 that $\epsilon_\phi(X) > \epsilon$ exactly when $\Delta_{LS}(X) > \beta$.

For noise-adding mechanisms such as the Laplace mechanism, the sensitivity is proportional to the privacy loss (in both the global and local sense, i.e. $\Delta_{GS} \propto \epsilon$ and $\Delta_{LS} \propto \epsilon(X)$). Therefore for these mechanisms the only difference between privately testing the local sensitivity (Algorithm 1) and privately testing the data-dependent DP (Theorem 4.3) is a change of parameterization.

4.1 Limitations of local sensitivity

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

The following example illustrates this issue.

Example 4.4 (Private posterior sampling)**.** *Let $\mathcal{M} : \mathcal{X} \times \mathcal{Y} \to \Theta$ be a private posterior sampling mechanism [13, 27, 8] for approximately minimizing $F_X(\theta)$.*

$\mathcal{M}$ samples $\theta \sim P(\theta) \propto e^{-\gamma(F_X(\theta) + 0.5\lambda||\theta||^2)}$ with parameters γ, λ. Note that γ, λ cannot be appropriately chosen for this mechanism to satisfy DP without going through a sensitivity calculation of $\arg\min F_X(\theta)$. In fact, the global and local sensitivity of the minimizer is unbounded even in linear regression problems, i.e when $F_X(\theta) = \frac{1}{2}||y - X\theta||^2$.

Output perturbation algorithms do work for the above problem when we regularize, but they are known to be suboptimal in theory and in practice [2]. In Section 5.1 we demonstrate how to apply generalized PTR to achieve a data-adaptive posterior sampling mechanism.

Even in the cases of noise-adding mechanisms where PTR seems to be applicable, it does not lead to a tight privacy guarantee. Specifically, by an example of privacy amplification by post-processing (Example 9.1 in the appendix), we demonstrate that the local sensitivity does not capture all sufficient statistics for data-dependent privacy analysis and thus is loose.

4.2 Which ϕ to propose

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

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

Algorithm 3 takes a list of parameters as input, runs generalized PTR with each of the parameters, and returns the output with the best utility. We show that the privacy guarantee

with respect to ϵ is independent of the number of ϕ that we try.

Formally, let $\phi_1, ..., \phi_k$ be a set of hyper-parameters and $\tilde{\theta}_i \in \{\perp, \text{Range}(\mathcal{M})\}$ denotes the output of running generalized PTR on a private dataset X with ϕ_i. Let X_{val} be a public validation set and $q(\tilde{\theta}_i)$ be the score of evaluating $\tilde{\theta}_i$ with X_{val} (e.g., validation accuracy). The goal is to select a pair $(\tilde{\theta}_i, \phi_i)$ such that DP model $\tilde{\theta}_i$ maximizes the validation score.

The generalized PTR framework with privacy calibration is described in Algorithm 3; its privacy guarantee is an application of Liu and Talwar [11].

Algorithm 3 PTR with hyper-parameter selection

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

Theorem 4.5 (Theorem 3.4 Liu and Talwar [11]). *Fix any $\tau \in [0, 1], \delta_2 > 0$ and let $T = \frac{1}{\tau} \log \frac{1}{\delta_2}$. If each oracle access to Algorithm 2 is (ϵ^*, δ^*)-DP, then Algorithm 3 is $(3\epsilon^* + 3\sqrt{2\delta^*}, \sqrt{2\delta^* T} + \delta_2)$-DP.*

The theorem implies that one can try a random number of ϕ while paying a constant ϵ. In practice, we can roughly set $\tau = \frac{1}{10k}$ so that the algorithm is likely to test all k parameters. We emphasize that the privacy and the utility guarantee (stated in the appendix) is not our contribution. But the idea of applying generalized PTR to enforce a uniform DP guarantee over all choices of parameters with a data-dependent analysis is new, and in our opinion, significantly broadens the applicability to generic hyper-parameter tuning machinery from Liu and Talwar [11].

4.3 Construction of the DP test

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

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

Theorem 4.6 (Generalized PTR with private upper bound)**.** *Suppose we have a differentially private upper bound of $\epsilon_\phi(X)$ w.r.t. δ such that with probability at least $1 - \delta'$, $\epsilon_\phi^P(X) > \epsilon_\phi(X)$. Further suppose we have an $(\hat{\epsilon}, \hat{\delta})$-DP test $\mathcal{T}$ such that*

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

Then Algorithm 2 is $(\epsilon + \hat{\epsilon}, \delta + \hat{\delta} + \delta')$-DP.

In Section 5.2, we demonstrate that one can upper bound the data-dependent DP through a modification of the smooth sensitivity framework applied on $\epsilon_\phi(X)$. Moreover, in Section 5.1 we provide a direct application of Theorem 4.6 with private linear regression by making use of the per-instance DP technique [25].

The applications in Section 5 are illustrative of two distinct approaches to constructing the DP test for generalized PTR:

1. Private sufficient statistics release (used in the private linear regression example of Section 5.1) specifies the data-dependent DP as a function of the dataset and privately releases each data-dependent component.

2. The second approach (used in the PATE example of Section 5.2) uses the smooth sensitivity framework to privately release the data-dependent DP as a whole, and then construct a high-confidence test using the Gaussian mechanism.

These two approaches cover most of the scenarios arising in data-adaptive analysis. For example, in the appendix we demonstrate the merits of generalized PTR in handling data-adaptive private generalized linear models (GLMs) using private sufficient statistics release. Moreover, sufficient statistics release together with our private hyper-parameter tuning (Algorithm 3) can be used to construct data-adaptive extensions of DP-PCA and Sparse-DP-ERM (see details in the future work section).

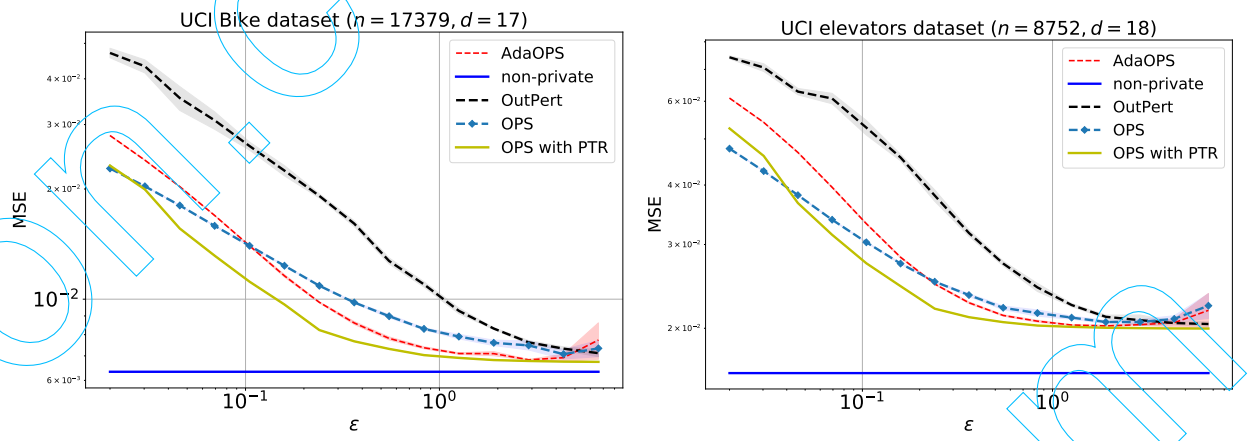
5 Applications

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

5.1 Private Linear Regression

Theorem 5.1 ([25])**.** *For input data $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, define the following:*

- $\lambda_{\min}(X)$ *denotes the smallest eigenvalue of $X^T X$;*



(a) Bike dataset

(b) Elevators dataset

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

- $||\theta^*_\lambda||$ *is the magnitude of the solution* $\theta^*_\lambda = (X^TX + \lambda I)^{-1}X^TY$*;*
- *and* $L(X, \mathbf{y}) := ||\mathcal{X}||(||\mathcal{X}||||\theta^*_\lambda|| + ||\mathcal{Y}||)$ *is the local Lipschitz constant, denoted* L *in brief.*

For brevity, denote $\lambda^* = \lambda + \lambda_{\min}(X)$*. The algorithm used in Example 4.4 with parameter* $\phi = (\lambda, \gamma)$ *obeys* $(\epsilon_\phi(Z), \delta)$ *data-dependent DP for each dataset* $Z = (X, Y)$ *with* $\epsilon_\phi(Z)$ *equal to*

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

Notice that the data-dependent DP is a function of $(\lambda_{\min}, L, ||\theta^*_\lambda||, \lambda, \gamma)$, where $(\lambda_{\min}, L, ||\theta^*_\lambda||)$ are data-dependent quantities. One can apply the generalized PTR framework as in the following example.

Example 5.2 (OPS with PTR)**.** *We demonstrate here how to apply generalized PTR to the one-posterior sample (OPS) algorithm, a differentially private mechanism which outputs one sample from the posterior distribution of a Bayesian model with bounded log-likelihood.*

- *Propose* $\phi = (\lambda, \gamma)$*.*
- *Based on* (λ, γ)*, differentially privately release* $\lambda_{min}, ||\theta^*_\lambda||, L$ *with privacy budget* $(\epsilon, \delta/2)$*.*
- *Condition on a high probability event (with probability at least* $1 - \delta/2$*) of* $\lambda_{min}, ||\theta^*_\lambda||, L$*, test if* $\epsilon^P_\phi(X)$ *is smaller than the predefined privacy budget* $(\hat{\epsilon}, \hat{\delta})$*, where* $\epsilon^P_\phi(X)$ *denotes the sanitized data-dependent DP.*
- *Based on the outcome of the test, decide whether to release* $\theta \propto e^{-\frac{\gamma}{2}||Y - X\theta||^2 + \lambda||\theta||^2}$*.*

Theorem 5.3. *The algorithm outlined in Example 5.2 satisfies* $(\epsilon + \hat{\epsilon}, \delta + \hat{\delta})$*-DP.*

The main idea of the above algorithm boils down to privately releasing all data-dependent quantities in data-dependent DP, constructing high-probability confidence intervals of these quantities, and then deciding whether to run the mechanism $\mathcal{M}$ with the proposed parameters. We defer the details of the privacy calibration of data-dependent quantities to the appendix.

One may ask why we cannot directly tune privacy parameters (λ, γ) based on the sanitized data-dependent DP. This is because, in many scenarios, data-dependent quantities depend on the choice of privacy parameters, e.g., $||\theta^*_\lambda||$ is a complicated function of λ. Thus, the optimization on λ becomes a circular problem — to solve λ, we need to sanitize $||\theta^*_\lambda||$, which needs to choose a λ to begin with. Alternatively, generalized PTR provides a clear and flexible framework to test the validity of privacy parameters adapted to the dataset.

Remark 5.4. The above "circular" issue is even more serious for generalized linear models (GLMs) beyond linear regression. The data-dependent DP there involves a local strong-convexity parameter, a complex function of the regularizer λ and we only have zeroth-order access to. In the appendix, we demonstrate how to apply generalized PTR to provide a generic solution to a family of private GLMs where the link function satisfies a self-concordance assumption.

We next apply Algorithm 3 for Example 5.2 with UCI regression datasets. Standard z-scoring is applied and each data point is normalize with a Euclidean norm of 1. We consider $(60\%, 10\%, 30\%)$ splits for training, validation and testing test.

Baselines

- Output Perturbation (Outpert) [2]: $\theta = (X^T X + \lambda I)^{-1} X^T \mathbf{y}$. Release $\hat{\theta} = \theta + \mathbf{b}$ with an appropriate λ, where $\mathbf{b}$ is a Gaussian random vector.
- Posterior sampling (OPS). Sample $\hat{\theta} \sim P(\theta) \propto e^{-\gamma(F(\theta)+0.5\lambda||\theta||^2)}$ with parameters γ, λ.
- Adaptive posterior sampling (AdaOPS) [26]. Run OPS with (λ, γ) chosen adaptively according to the dataset.

Outpert and OPS serve as two non-adaptive baselines. In particular, we consider OPS-Balanced [26], which chooses λ to minimize a data-independent upper bound of empirical risk and dominates other OPS variants. AdaOPS is one state-of-the-art algorithm for adaptive private regression, which automatically chooses λ by minimizing an upper bound of the data-dependent empirical risk.

We implement OPS-PTR as follows: propose a list of λ through grid search (we choose $k = 30$ and λ ranges from $[2.5, 2.5^{10}]$ on a logarithmic scale); instantiate Algorithm 3 with $\tau = 0.1k$, $T = \frac{1}{\tau}\log(1/\delta_2)$ and $\delta_2 = 1/2\delta$; calibrate γ to meet the privacy requirement for each λ. sample $\hat{\theta}$ using (λ, γ) and return the one with the best validation accuracy. Notice that we use a "no $\perp$" variant of Algorithm 2 as the calibration of γ is clear given a fixed λ and privacy budget (see more details in the appendix). We can propose various combinations of (λ, γ) for more general applications.

Figure 1 demonstrates how the MSE error of the linear regression algorithms varies with the privacy budget ϵ. OutPert suffers from the large global sensitivity of output θ. OPS performs well but does not benefit from the data-dependent quantities. AdaOPS is able to adaptively choose (λ, γ) based on the dataset, but suffers from the estimation error of the data-dependent empirical risk. On the other hand, OPS-PTR selects a (λ, γ) pair that minimizes the empirical error on the validation set directly, and the privacy parameter γ adapts to the dataset thus achieving the best result.

5.2 PATE

nIn this section, we apply the generalized PTR framework to solve an open problem from the Private Aggregation of Teacher Ensembles (PATE) [18, 19] — privately publishing the entire model through privately releasing data-dependent DP losses. Our algorithm makes use of the smooth sensitivity framework [16] and the Gaussian mechanism to construct a high-probability test of the data-dependent DP. The one-dimensional statistical nature of data-dependent DP enables efficient computations under the smooth sensitivity framework. Thus, this approach is generally applicable for other private data-adaptive analysis beyond PATE.

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

The tight privacy guarantee of PATE heavily relies on a delicate data-dependent DP analysis, for which the authors of PATE use the smooth sensitivity framework to privately publish the data-dependent privacy cost. However, it remains an open problem to show that the released model is DP under data-dependent analysis. Our generalized PTR resolves this gap by carefully testing a private upper bound of the data-dependent privacy cost. Our algorithm is fully described in Algorithm 4, where the modification over the original PATE framework is highlighted in blue.

Algorithm 4 takes the input of privacy budget $(\epsilon', \hat{\epsilon}, \delta)$, unlabeled public data $x_{1:T}$ and K teachers' predictions on these data. The parameter ϵ denotes the privacy cost of publishing the data-dependent DP and ϵ' is the predefined privacy budget for testing. $n_j(x_i)$ denotes the the number of teachers that agree on label j for x_i and C denotes the number of classes. The goal is to privately release a list of plurality outcomes — $\text{argmax}_{j\in[C]} n_j(x_i)$ for $i \in [T]$ — and use these outcomes to supervise the training of a "student" model in the public domain. The parameter σ_1 denotes the noise scale for the vote count.

In their privacy analysis, Papernot et al. [19] compute the data-dependent $\text{RDP}_{\sigma_1}(\alpha, X)$ of labeling the entire group of student queries. $\text{RDP}_{\sigma_1}(\alpha, X)$ can be orders of magnitude smaller than its data-independent version if there is a strong agreement among teachers. Note that $\text{RDP}_{\sigma_1}(\alpha, X)$ is a function of the RDP order α and the dataset X, analogous to our Definition 4.1 but subject to RDP [14].

Theorem 5.5 ([19]). *If the top three vote counts of x_i are $n_1 > n_2 > n_3$ and $n_1 - n_2, n_2 - n_3 \gg \sigma_1$, then the data-dependent RDP of releasing $\text{argmax}_j\{n_j + \mathcal{N}(0, \sigma_1^2)\}$ satisfies $(\alpha, \exp\{-2\alpha/\sigma_1^2\}/\alpha)$-RDP and the data-independent RDP (using the Gaussian mechanism) satisfies $(\alpha, \frac{\alpha}{\sigma_1^2})$-RDP.*

Algorithm 4 PATE with generalized PTR

Input: Unlabeled public data $x_{1:T}$, aggregated teachers prediction $n(\cdot)$, privacy parameter $\hat{\epsilon}, \epsilon', \delta$, noisy parameter σ_1.
Set $\alpha = \frac{2\log(2/\delta)}{\hat{\epsilon}} + 1$, $\sigma_s = \sigma_2 = \sqrt{\frac{3\alpha+2}{\hat{\epsilon}}}$, $\delta_2 = \delta/2$, smoothness parameter $\beta = \frac{0.2}{\alpha}$.
Compute noisy labels: $y_i{}^p \leftarrow \text{argmax}_{j\in[C]}\{n_j(x_i) + \mathcal{N}(0,\sigma_1^2)\}$ for all $i \in [1:T]$.
$\text{RDP}_{\sigma_1}(\alpha, X) \leftarrow$ data-dependent RDP at the α-th order.
$SS_\beta(X) \leftarrow$ the smooth sensitivity of $\text{RDP}^{\text{upper}}_{\sigma_1}(\alpha, X)$.
Privately release $\mu := \log(SS_\beta(X)) + \beta \cdot \mathcal{N}(0,\sigma_2^2) + \sqrt{2\log(2/\delta_2)} \cdot \sigma_2 \cdot \beta$
$\text{RDP}^{\text{upper}}_{\sigma_1}(\alpha) \leftarrow$ an upper bound of data-dependent RDP through Lemma 5.6.
$\epsilon_{\sigma_1} \leftarrow$ DP guarantee converted from $\text{RDP}^{\text{upper}}_{\sigma_1}(\alpha)$.
If $\epsilon' \geq \epsilon_{\sigma_1}$ **return** a student model trained using $(x_{1:T}; y^p_{1:T})$.
Else return $\perp$.

However, $\text{RDP}_{\sigma_1}(\alpha, X)$ is data-dependent and thus cannot be revealed. The authors therefore privately publish the data-dependent RDP using the smooth sensitivity framework [16]. The smooth sensitivity calculates a smooth upper bound on the local sensitivity of $\text{RDP}_{\sigma_1}(\alpha, X)$, denoted as $SS_\beta(X)$, such that $SS_\beta(X) \leq e^\beta SS_\beta(X')$ for any neighboring dataset X and X'. By adding Gaussian noise scaled by the smooth sensitivity (i.e., releasing $\epsilon_{\sigma_1}(\alpha, X) + SS_\beta(X) \cdot \mathcal{N}(0,\sigma_s^2)$), the privacy cost can be safely published.

Unlike most noise-adding mechanisms, the standard deviation σ_s cannot be published since $SS_\beta(X)$ is a data-dependent quantity. Moreover, this approach fails to provide a valid privacy guarantee of the noisy labels obtained through the PATE algorithm, as the published privacy cost could be smaller than the real privacy cost. Our solution in Algorithm 4 looks like the following:

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

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

Lemma 5.6 (Private upper bound of data-dependent RDP)**.** *We are given a RDP function* $\text{RDP}(\alpha, X)$ *and a* β*-smooth sensitivity bound* $SS(\cdot)$ *of* $\text{RDP}(\alpha, X)$*. Let* μ *(defined in Algorithm 4) denote the private release of* $\log(SS_\beta(X))$*. Let the* $(\beta, \sigma_s, \sigma_2)$*-GNSS mechanism be*

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

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

The proof (deferred to the appendix) makes use of the facts that: (1) the log of $SS_\beta(X)$ has a bounded global sensitivity β through the definition of smooth sensitivity; (2) releasing $\text{RDP}_{\sigma_1}(\alpha, X) + SS_\beta(X) \cdot \mathcal{N}(0, \sigma_s^2)$ is $(\alpha, \frac{\alpha+1}{\sigma_s^2})$-RDP (Theorem 23 from Papernot et al. [19]).

Now, we are ready to state the privacy guarantee of Algorithm 4.

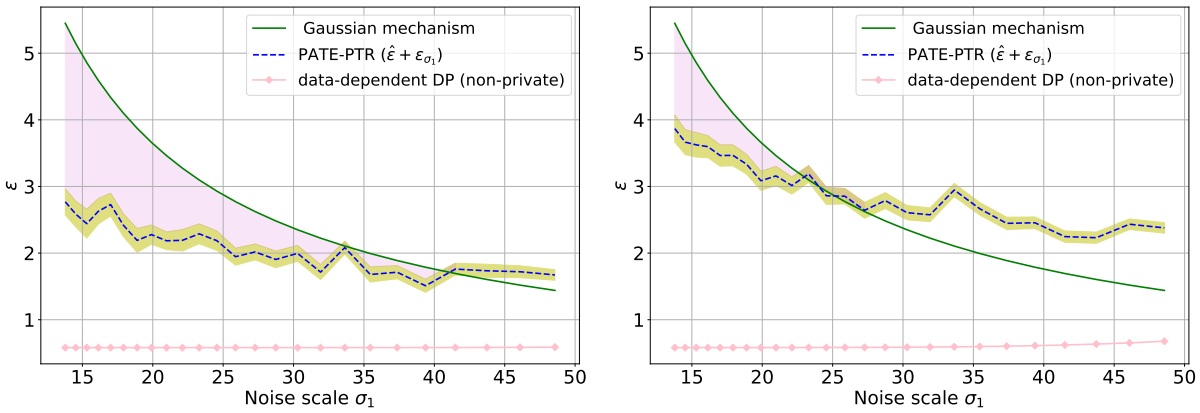
Theorem 5.7. *Algorithm 4 satisfies* $(\epsilon' + \hat{\epsilon}, \delta)$*-DP.*

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

Empirical results. We next empirically evaluate Algorithm 4 (PATE-PTR) on the MNIST dataset. Following the experimental setup from Papernot et al. [19], we consider the training set to be the private domain, and the testing set is used as the public domain. We first partition the training set into 400 disjoint sets and 400 teacher models, each trained individually. Then we select $T = 200$ unlabeled data from the public domain, with the goal of privately labeling them. To illustrate the behaviors of algorithms under various data distributions, we consider two settings of unlabeled data, high-consensus and low-consensus. In the low-consensus setting, we choose T unlabeled data such that there is no high agreement among teachers, so the advantage of data-adaptive analysis is diminished. We provide further details on the distribution of these two settings in the appendix.

Baselines. We consider the Gaussian mechanism as a data-independent baseline, where the privacy guarantee is valid but does not take advantage of the properties of the dataset. The data-dependent DP (Papernot et al. [19]) serves as a non-private baseline, which requires further sanitation. Note that these two baselines provide different privacy analyses of the same algorithm (see Theorem 5.5).

Figure 2 plots privacy-utility tradeoffs between the three approaches by varying the noise scale σ_1. The purple region denotes a set of privacy budget choices ($\hat{\epsilon} + \epsilon'$ used in Algorithm 4) such that the utility of the three algorithms is aligned under the same σ_1. In more detail, the purple region is lower-bounded by $\hat{\epsilon} + \epsilon_{\sigma_1}$. We first fix $\sigma_s = \sigma_2 = 15$ such that $\hat{\epsilon}$ is fixed. Then we empirically calculate the average of ϵ_{σ_1} (the private upper bound of the data-dependent DP) over



(a) High consensus and strong data-dependent DP

(b) Low consensus and low data-dependent DP

Figure 2: Privacy and utility tradeoffs with PATE. When σ_1 is aligned, three algorithms provide the same utility. y-axis plots the privacy cost of labeling $T = 200$ public data with $\delta = 10^{-5}$. The left figure considers the high-consensus case, where the data-adaptive analysis is preferred.

10 trials. Running Algorithm 4 with any choice of $\hat{\epsilon} + \epsilon'$ chosen from the purple region implies $\epsilon' > \epsilon_{\sigma_1}$. Therefore, PATE-PTR will output the same noisy labels (with high probability) as the two baselines.

Observation As σ_1 increases, the privacy loss of the Gaussian mechanism decreases, while the data-dependent DP curve does not change much. This is because the data-dependent DP of each query is a complex function of both the noise scale and the data and does not monotonically decrease when σ_1 increases (see more details in the appendix). However, the data-dependent DP still dominates the Gaussian mechanism for a wide range of σ_1. Moreover, PATE-PTR nicely interpolates between the data-independent DP guarantee and the non-private data-adaptive DP guarantee. In the low-consensus case, the gap between the data-dependent DP and the DP guarantee of the Gaussian mechanism unsurprisingly decreases. Meanwhile, PATE-PTR (the purple region) performs well when the noise scale is small but deteriorates when the data-independent approach proves more advantageous. This example demonstrates that using PTR as a post-processing step to convert the data-dependent DP to standard DP is effective when the data-adaptive approach dominates others.

6 Limitations and Future Work

One weakness of generalized PTR is that it requires a case-specific privacy analysis. Have we simply exchanged the problem of designing a data-adaptive DP algorithm with the problem of analyzing the data-dependent privacy loss? We argue that this limitation is inherited from classic PTR. In situations where classic PTR is not applicable, we've outlined several approaches to constructing the DP test for our framework (see Sections 4.3 and 5.2).

Furthermore, the data-dependent privacy loss is often more straightforward to compute than local sensitivity, and often exists in intermediate steps of classic DP analysis already. Most DP analysis involves providing a high-probability tail bound of the privacy loss random variable. If we stop before taking the max over the input dataset, then we get a data-dependent DP loss right away (as in Example 4.2).

There are several exciting directions for applying generalized PTR to more problems. Sufficient statistics release and our private hyperparameter tuning (Algorithm 3) can be used to construct data-adaptive extensions of DP-PCA [7] and Sparse-DP-ERM [10]. For DP-PCA we could use our Algorithm 3 to tune the variance of the noise added to the spectral gap; for Sparse-DP-ERM we would test the restricted strong convexity parameter (RSC), i.e. not adding additional regularization if the RSC is already large.

7 Conclusion

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

Acknowledgments

We thank the anonymous reviewers and area chair for helpful input. The work was partially supported by NSF Award # 2048091 and the Google Research Scholar Award. Yuqing was supported by a Google PhD Fellowship.

Contents

1 Introduction **1**

2 Related Work **2**

3 Preliminaries **2**

3.1 Propose-Test-Release . . . 3

4 Generalized PTR **3**

4.1 Limitations of local sensitivity . . . 4

4.2 Which ϕ to propose . . . 4

4.3 Construction of the DP test . . . 5

5 Applications **5**

5.1 Private Linear Regression . . . 5

5.2 PATE . . . 7

6 Limitations and Future Work **9**

7 Conclusion **9**

8 Omitted proofs **11**

9 Omitted examples in the main body **11**

9.1 Limits of the classic PTR in private binary voting . . . 12

9.2 Self-concordant generalized linear model (GLM) . . . 13

9.3 Differentially privately release $\lambda_{min}\left(\nabla^2 F(\theta)\right)$ . . . 16

9.4 Other applications of generalized PTR . . . 17

10 Omitted proofs in Section 4 **17**

11 Experimental details **18**

11.1 Experimental details in private linear regression . . . 18

11.2 Details of PATE case study . . . 19

12 Omitted proofs in private GLM **20**

12.1 Per-instance DP of GLM . . . 20

8 Omitted proofs

Proof of Theorem 4.3. The proof of our main privacy result relies on two central properties of differential privacy: composition and immunity to post-processing. We review these below.

Theorem 8.1 (Composition [6])**.** *For $i \in [k]$, let $\mathcal{M}_i : \mathcal{Z} \to \mathcal{R}_i$ be a randomized algorithm satisfying (ϵ_i, δ_i)-DP. Define the mechanism $\mathcal{M} : \mathcal{Z} \to \prod_{i=1}^k \mathcal{R}_i$ as $\mathcal{M}(Z) = (\mathcal{M}_1(Z), \mathcal{M}_2(Z), \ldots, \mathcal{M}_k(Z))$. Then $\mathcal{M}$ satisfies $\left(\sum_{i=1}^k \epsilon_i, \sum_{i=1}^k \delta_i\right)$-DP.*

Theorem 8.2 (Closure under post-processing [6])**.** *Consider a mechanism $\mathcal{M} : \mathcal{Z} \to \mathcal{R}$ that satisfies (ϵ, δ)-DP. Let $f : \mathcal{R} \to \mathcal{R}'$ be a data-independent (randomized or deterministic) mapping. Then $f \circ \mathcal{M}$ satisfies (ϵ, δ)-DP.*

We consider two cases:

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

Let E be the event $\mathcal{T}(X) = 1$ and consider a possible output set $S \subseteq \mathcal{R} \cup \{\perp\}$. Recall that the test $\mathcal{T}$ satisfies $(\hat{\epsilon}, \hat{\delta})$-DP.

When $\perp \in S$,

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

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

From the assumption of Theorem 4.3 on the test $\mathcal{T}$, $\Pr[E] = \Pr[\mathcal{T}(X) = 1] \leq \delta'$. So

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

Putting these together, we have

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

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

Since $\mathcal{M}_\phi$ satisfies $(\epsilon_\phi(X), \delta)$ data-dependent DP for dataset X, for any neighboring dataset X' and output set $\Theta \subseteq \mathcal{R}$ we have

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

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

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

Running $\mathcal{M}$ is therefore a composition of a $(\hat{\epsilon}, \hat{\delta})$-DP mechanism and a (ϵ, δ)-DP mechanism, and by basic composition properties satisfies $(\epsilon + \hat{\epsilon}, \delta + \hat{\delta})$-DP.

Taking Cases I and II together, mechanism $\mathcal{M}$ therefore satisfies $(\epsilon + \hat{\epsilon}, \delta + \delta + \delta')$-DP.

□

9 Omitted examples in the main body

In this appendix, we provide more examples to demonstrate the merits of generalized PTR. We focus on a simple example of post-processed Laplace mechanism in Section 9.1 and then an example on differentially private learning of generalized linear models in Section 4. In both cases, we observe that generalized PTR provides data-adaptive algorithms with formal DP guarantees, that are simple, effective and not previously proposed in the literature (to the best of our knowledge).

9.1 Limits of the classic PTR in private binary voting

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

Example 9.1. *Consider a binary class voting problem: n users vote for a binary class $\{0, 1\}$ and the goal is to output the class that is supported by the majority. Let n_i denote the number of people who vote for the class i. We consider the report-noisy-max mechanism:*

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

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

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

We first explicitly state the data-dependent DP.

Theorem 9.2. *The data-dependent DP of the above example is*

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

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

In Figure 3(a), we empirically compare the above data-dependent DP with the Laplace mechanism by varying the gap between the two vote counts $|n_0(X) - n_1(X)|$. The noise scale is fixed to $\epsilon = 10$. The data-dependent DP substantially improves over the standard DP if the gap is large. However, the data-dependent DP is a function of the dataset. We next demonstrate how to apply generalized PTR to exploit the data-dependent DP.

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

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

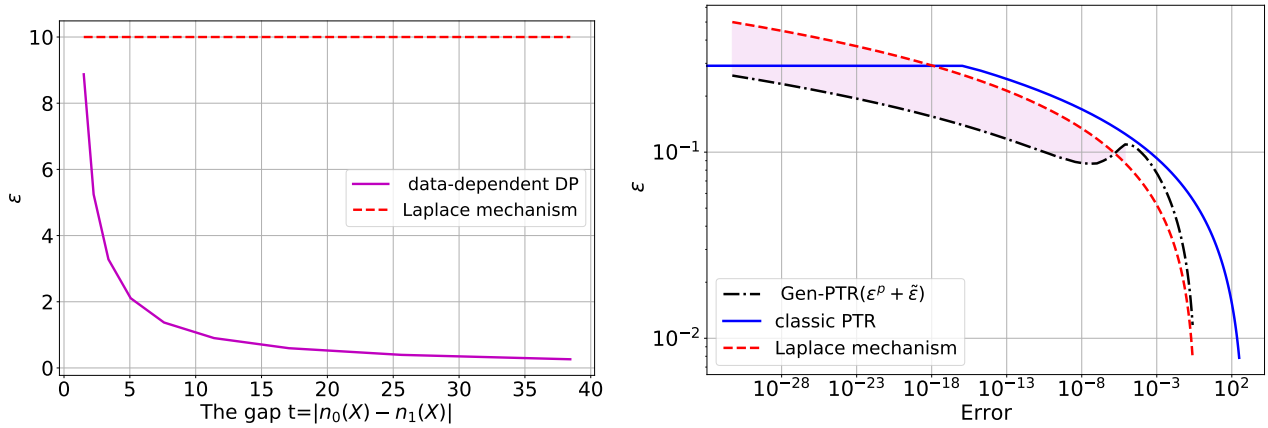
Similarly, $p' = \frac{2 + \epsilon \cdot (t+\ell)}{4\exp(\epsilon \cdot (t+\ell))}$, where $\ell \in [-1, 1]$ denotes adding or removing one data point to construct the neighboring dataset X'. Therefore, we can upper bound $\log(p/p')$ by

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

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

On the other hand, the local sensitivity $\Delta_{LS}(X)$ of this noise-adding mechanism is 0 if $t > 1$. Specifically, if the gap is larger than one, adding or removing one user will not change the result. To apply classic PTR, we let $\gamma(X)$ denote the distance to the nearest dataset X'' such that $\Delta_{LS} > 0$ and test if $\gamma(X) + \text{Lap}(1/\epsilon) > \frac{\log(1/\delta)}{\epsilon}$. Notice in this example that $\gamma(X) = \max(t - 1, 0)$ can be computed efficiently. We provide the detailed implementation of these approaches.

1. Gen PTR: lower bound t with $t^p = t - \frac{log(1/\delta)}{\tilde{\epsilon}} + \text{Lap}(1/\tilde{\epsilon})$. Calculate an upper bound of data-dependent DP ϵ^p using Theorem 9.2 with t^p. The algorithm then tests if ϵ^p is within an predefined privacy budget ϵ'. If the test passes, the algorithm returns $\text{argmax}_{i\in[0,1]} n_i(X) + Lap(1/\epsilon)$ satisfies $(\tilde{\epsilon} + \epsilon', \delta)$-DP.



(a) data-dependent DP vs Laplace mechanism

(b) Privacy-utility tradeoff between three approaches.

Figure 3: In Figure 3(a), we compare the privacy guarantee by varying the gap. In Figure 3(b) We fix $t = n_0(X) - n_1(X) = 100$ and compare privacy cost when the accuracy is aligned. Gen-PTR with any choice of privacy budget $(\tilde{\epsilon} + \epsilon')$ chosen from the purple region would achieve the same utility as Laplace mechanism but with a smaller privacy cost. The curve of Gen-PTR is always below than that of the classic PTR, which implies that Gen-PTR can result a tighter privacy analysis when the utility is aligned.

2. classic PTR: lower bound t with $t^p = t - \frac{log(1/\delta)}{\tilde{\epsilon}} + \text{Lap}(1/\tilde{\epsilon})$. If $t^p > 1$, classic PTR outputs the ground-truth result else returns a random class. This algorithm satisfies $(\tilde{\epsilon}, \delta)$-DP.
3. Laplace mechanism. $\mathcal{M}(X) : \text{argmax}_{i\in[0,1]} n_i(X) + Lap(1/\epsilon)$. $\mathcal{M}$ is (ϵ, δ)-DP.

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

To confirm this conjecture, Figure 3(b) plots a privacy-utility trade-off curve between these three approaches. We consider a voting example with $n_0(X) = n_1(X) + 100$ and $t = 100$, chosen such that the data-adaptive analysis is favorable.

In Figure 3(b), we vary the noise scale $b = 1/\epsilon$ between $[0, 0.5]$. For each choice of b, we plot the privacy guarantee of three algorithms when the error rate is aligned. For Gen-PTR, we set $\tilde{\epsilon} = \frac{1}{2b}$ and empirically calculate ϵ^p over 100000 trials.

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

9.2 Self-concordant generalized linear model (GLM)

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

- Introduce a family of GLMs with the self-concordance property.
- Introduce a general output perturbation algorithm for private GLMs.

- Analyze the data-dependent DP of GLMs with the self-concordance property.
- Provide an example of applying our generalized PTR framework to logistic regression.

Consider the empirical risk minimization problem of the generalized linear model

$$\theta^* = \text{argmin}_\theta \sum_{i=1^n} l_i(\theta) + r(\theta),$$

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

We now define the self-concordance property.

Definition 9.3 (Generalized self-concordance [1])**.** A convex and three-times differentiable function $f : \Theta \to \mathbb{R}$ is R-generalized-self-concordant on an open nonempty convex set $\Theta^* \subset \Theta$ with respect to norm $\|\cdot\|$ if for all $u \in \Theta^*$ and all $v \in \mathbb{R}^d$,

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

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

Lemma 9.4 (Stability of Hessian[15, Theorem 2.1.1], [1, Proposition 1])**.** *Let $H_\theta := \nabla^2 F_s(\theta)$. If F_s is R-self-concordant at θ, then for any v such that $R\|v\|_{H_\theta} < 1$, we have that*

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

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

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

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

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

Example 9.5 (Binary logistic regression)**.** *Assume $\|x\|_2 \leq 1$ for all $x \in \mathcal{X}$ and $y \in \{-1, 1\}$. Then binary logistic regression with datasets in $\mathcal{X} \times \mathcal{Y}$ has a log-likelihood of $F(\theta) = \sum_{i=1}^n \log(1+e^{-y_i x_i^T\theta})$. The univariate function $l := \log(1+\exp(\cdot))$ satisfies*

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

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

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

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

Theorem 9.6 (Data-dependent DP of GLM)**.** *Denote the smooth part of the loss function $F_s = \sum_{i=1}^n l(y_i, < x_i, \cdot >) + r_s(\cdot)$. Assume the following:*

1. *The GLM loss function l is convex, three-times continuously differentiable and R-generalized-self-concordant w.r.t. $\|\cdot\|_2$,*

2. *F_s is locally α-strongly convex w.r.t. $\|\cdot\|_2$,*
3. *and in addition, denote $L := \sup_{\theta\in[\theta^*,\tilde{\theta}^*]} |l'(y, x^T\theta)|$, $\beta := \sup_{\theta\in[\theta^*,\tilde{\theta}^*]} |l''(y, x^T\theta)|$. That is, $\ell(\cdot)$ is L-Lipschitz and β-smooth.*

We then have the data-dependent DP

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

The proof follows by taking an upper bound of the per-instance DP loss (Theorem 12.1) $\epsilon(Z, z)$ over $z = (x, y) \in (\mathcal{X}, \mathcal{Y})$.

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

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

Remark 9.7 (Logistic regression)**.** For logistic regression, we know $L \leq 1$, $\beta \leq 1/4$ and if $\|x\|_2 \leq 1$, it is 1-generalized self-concordant. For any dataset $Z = (X, y)$, the data-dependent DP $\epsilon(X)$ w.r.t. δ can be simplified to:

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

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

Example 9.8. *We demonstrate here how we apply generalized PTR to output perturbation of the logistic regression problem.*

1. *Take an exponential grid of parameters $\{\lambda\}$ and propose each λ.*
2. *Solve for $\theta^*_\lambda = argmin_\theta F(\theta) + \lambda\|\theta\|^2/2$*
3. *Calculate the smallest eigenvalue $\lambda_{\min}(\nabla^2 F(\theta^*_\lambda))$ (e.g., using power method).*
4. *Differentially privately release $\lambda_{\min}$ with $\lambda^p_{\min} := \max\{\lambda_{\min} + \frac{\sqrt{\log(4/\delta)}}{\epsilon/2}\cdot\Delta_{GS}\cdot Z - \frac{\sqrt{2\log(4/\delta)\cdot\log(1/\delta)}\Delta_{GS}}{\epsilon/2}, 0\}$, where Δ_{GS} denote the global sensitivity of $\lambda_{\min}$ using Theorem 9.11.*
5. *Let $\epsilon^p(\cdot)$ be instantiated with $\epsilon(X)$ w.r.t. δ from Remark 9.7, where $\alpha = \lambda^p_{\min} + \lambda$. Then, conditioned on a high probability event, $\epsilon^p(\cdot)$ (a function of γ) is a valid DP bound that holds for all datasets and all parameters γ.*
6. *Calculate the maximum γ such that $\epsilon^p_{\delta/2}(\gamma) \leq \epsilon/2$.*
7. *Release $\hat{\theta} \sim \mathcal{N}(\theta^*_\lambda, \gamma^{-1}\nabla^2 F_s(\theta^*_\lambda)^{-1})$.*
8. *Evaluate the utility on the validation set and return the (λ, γ) pair that leads to the highest utility.*

Theorem 9.9. *For each proposed λ, the algorithm that releases $\hat{\theta} \sim \mathcal{N}(\theta^*_\lambda, \gamma^{-1}\nabla^2 F_s(\theta^*_\lambda)^{-1})$ is $(\epsilon, 2\delta)$-DP.*

Proof. The proof follows the recipe of generalized PTR with private upper bound (Example 4.6). First, the release of $\lambda_{\min}(\nabla^2 F(\theta^*_\lambda))$ is $(\epsilon/2, \delta/2)$-DP. Then, with probability at least $1-\delta$, $\epsilon^p_\delta(\cdot) > \epsilon_\delta(X)$ holds for all X and γ. Finally, γ is chosen such that the valid upper bound is $(\epsilon/2, \delta/2)$-DP. □

For the hyper-parameter tuning on λ (Steps 1 and 8), we can use Algorithm 3 to evaluate each λ.

Unlike Example 5.2, the $\lambda_{min}(\nabla^2 F(\theta^*_\lambda))$ is a complicated data-dependent function of λ. Thus, we cannot privately release the data-dependent quantity $\lambda_{min}(\nabla^2 F(\theta^*_\lambda))$ without an input λ. The PTR approach allows us to test a number of different λ and hence get a more favorable privacy-utility trade-off.

An interesting perspective of this algorithm for logistic regression is that increasing the regularization α is effectively increasing the number of data points within the soft "margin"[3] of separation, hence a larger contribution to the Hessian from the loss function.

Remark 9.10. The PTR solution for GLMs follows a similar recipe: propose a regularization strength λ; construct a lower bound of the strong convexity α at the optimal solution θ^*_λ; and test the validity of data-dependent DP using Theorem 12.1.

Before moving on to other applications of generalized PTR, we will show how to differentially privately release λ_{min} according to the requirements of the logistic regression example.

9.3 Differentially privately release $\lambda_{min}\left(\nabla^2 F(\theta)\right)$

To privately release $\lambda_{min}\nabla^2 F(\theta)$, we first need to compute its global sensitivity. Once we have that then we can release it differentially privately using either the Laplace mechanism or the Gaussian mechanism.

Theorem 9.11 (Global sensitivity of the minimum eigenvalue at the optimal solution)**.** *Let $F(\theta) = \sum_{i=1}^n f_i(\theta) + r(\theta)$ and $\tilde{F}(\theta) = F(\theta) + f(\theta)$ where $f_1, ..., f_n$ are loss functions corresponding to a particular datapoint x. Let $\theta^* = argmin_\theta F(\theta)$ and $\tilde{\theta}^* = argmin_\theta \tilde{F}(\theta)$. Assume f is L-Lipschitz and β-smooth, $r(\theta)$ is λ-strongly convex, and F and $\tilde{F}$ are R-self-concordant. If in addition, $\lambda \geq RL$, then we have*

$$\sup_{X,x}(\lambda_{min}(\nabla^2 F(\theta^*_\lambda)) - \lambda_{min}(\nabla^2 \tilde{F}(\tilde{\theta}^*_\lambda))) \leq 2RL + \beta.$$

Proof.

$$\begin{aligned}
&\lambda_{min}(\nabla^2 F(\theta^*_\lambda)) - \lambda_{min}(\nabla^2 \tilde{F}(\tilde{\theta}^*_\lambda)) \\
&\quad = (\lambda_{min}(\nabla^2 F(\theta^*_\lambda)) - \lambda_{min}(\nabla^2 \tilde{F}(\theta^*_\lambda))) \\
&\quad + (\lambda_{min}(\nabla^2 \tilde{F}(\theta^*_\lambda)) - \lambda_{min}(\nabla^2 \tilde{F}(\tilde{\theta}^*_\lambda))).
\end{aligned} \tag{1}$$

We first bound the part on the left. By applying Weyl's lemma $\lambda(X+E) - \lambda(X) \leq ||E||_2$, we have

$$\sup_x ||\nabla^2 F(\theta^*_\lambda) - \nabla^2 F(\tilde{\theta}^*_\lambda)||_2 = ||\nabla^2 f(\theta^*_\lambda)||_2 \leq \beta \tag{2}$$

In order to bound the part on the right, we apply the semidefinite ordering using self-concordance, which gives

$$e^{-R\|\tilde{\theta}^*_\lambda - \theta^*_\lambda\|}\nabla^2 \tilde{F}(\tilde{\theta}^*_\lambda) \prec \nabla^2 \tilde{F}(\theta^*_\lambda) \prec e^{R\|\tilde{\theta}^*_\lambda - \theta^*_\lambda\|}\nabla^2 \tilde{F}(\tilde{\theta}^*_\lambda).$$

By the Courant-Fischer Theorem and the monotonicity theorem, we also have that for the smallest eigenvalue

$$\begin{aligned}
e^{-R\|\tilde{\theta}^*_\lambda - \theta^*_\lambda\|}\lambda_{\min}\left(\nabla^2 \tilde{F}(\tilde{\theta}^*_\lambda)\right) &\leq \lambda_{\min}\left(\nabla^2 \tilde{F}(\theta^*_\lambda)\right) \\
&\leq e^{R\|\tilde{\theta}^*_\lambda - \theta^*_\lambda\|}\lambda_{\min}\left(\nabla^2 \tilde{F}(\tilde{\theta}^*_\lambda)\right).
\end{aligned} \tag{3}$$

Moreover by Proposition 12.2, we have that

$$\|\tilde{\theta}^*_\lambda - \theta^*_\lambda\|_2 \leq \frac{\|\nabla f(\tilde{\theta^*}_\lambda)\|}{\lambda_{\min}\left(\nabla^2 \tilde{F}(\tilde{\theta}^*_\lambda)\right)} \leq \frac{L}{\lambda_{\min}\left(\nabla^2 \tilde{F}(\tilde{\theta}^*_\lambda)\right)}.$$

If $\lambda_{\min}\left(\nabla^2 \tilde{F}(\tilde{\theta}^*_\lambda)\right) \geq RL$, then use that $e^x - 1 \leq 2x$ for $x \leq 1$. Substituting the above bound to (3) then to (1) together with (2), we get a data-independent global sensitivity bound of

$$\lambda_{min}(\nabla^2 F(\theta^*_\lambda)) - \lambda_{min}(\nabla^2 \tilde{F}(\tilde{\theta}^*_\lambda)) \leq 2RL + \beta$$

as stated. □

[3] If we think of logistic regression as a smoothed version of SVM, then increasing α leads to more support vectors. The "margin" is "softer" in logistic regression, but qualitatively the same.

Proposition 9.12. *Let $\|\cdot\|$ be a norm and $\|\cdot\|_*$ be its dual norm. Let $F(\theta)$, $f(\theta)$ and $\tilde{F}(\theta) = F(\theta) + f(\theta)$ be proper convex functions and θ^* and $\tilde{theta}^*$ be their minimizers, i.e., $0 \in \partial F(\theta^*)$ and $0 \in \partial \tilde{F}(\tilde{theta}^*)$. If in addition, $F, \tilde{F}$ is $\alpha, \tilde{\alpha}$-strongly convex with respect to $\|\cdot\|$ within the restricted domain $\theta \in \{t\theta^* + (1-t)\tilde{\theta}^* \mid t \in [0,1]\}$. Then there exists $g \in \partial f(\theta^*)$ and $\tilde{g} \in \partial f(\tilde{\theta}^*)$ such that*

$$\|\theta^* - \tilde{\theta}^*\| \le \min\left\{\frac{1}{\alpha}\|\tilde{g}\|_*, \frac{1}{\tilde{\alpha}}\|g\|_*\right\}.$$

Proof. Apply the first order condition to F restricted to the line segment between $\tilde{\theta}^*$ and θ^*, we get

$$F(\tilde{\theta}^*) \ge F(\theta^*) + \langle \partial F(\theta^*), \tilde{\theta}^* - \theta^* \rangle + \frac{\alpha}{2}\|\tilde{\theta}^* - \theta^*\|^2 \tag{4}$$

$$F(\theta^*) \ge F(\tilde{\theta}^*) + \langle \partial F(\tilde{\theta}^*), \theta^* - \tilde{\theta}^* \rangle + \frac{\alpha}{2}\|\tilde{\theta}^* - \theta^*\|^2 \tag{5}$$

Note by the convexity of F and f, $\partial\tilde{F} = \partial F + \partial f$, where $+$ is the Minkowski Sum. Therefore, $0 \in \partial\tilde{F}(\tilde{\theta}^*)$ implies that there exists $\tilde{g}$ such that $\tilde{g} \in \partial f(\tilde{\theta}^*)$ and $-\tilde{g} \in \partial F(\tilde{\theta}^*)$. Take $-\tilde{g} \in \partial F(\tilde{\theta}^*)$ in Equation 10 and $0 \in \partial F(\theta^*)$ in Equation 9 and add the two inequalities, we obtain

$$\begin{aligned} 0 &\ge \langle -\tilde{g}, \theta^* - \tilde{\theta}^* \rangle + \alpha\|\tilde{\theta}^* - \theta^*\|^2 \\ &\ge -\|\tilde{g}\|_*\|\theta^* - \tilde{\theta}^*\| + \alpha\|\tilde{\theta}^* - \theta^*\|^2. \end{aligned}$$

For $\|\tilde{\theta}^* - \theta^*\| = 0$ the claim is trivially true; otherwise, we can divide both sides of the above inequality by $\|\tilde{\theta}^* - \theta^*\|$ and get $\|\theta^* - \tilde{\theta}^*\| \le \frac{1}{\alpha}\|\tilde{g}\|_*$.

It remains to show that $\|\theta^* - \tilde{\theta}^*\| \le \frac{1}{\tilde{\alpha}}\|g\|_*$. This can be obtained by exactly the same arguments above but applying strong convexity to $\tilde{F}$ instead. Note that we can actually get something slightly stronger than the statement because the inequality holds for all $g \in \partial f(\theta^*)$. □

9.4 Other applications of generalized PTR

Besides one-posterior sampling for GLMs, there are plenty of examples that our generalized-PTR could be applied, e.g., DP-PCA [7] and Sparse-DP-ERM [10] (when the designed matrix is well-behaved).

[7] provides a PTR style privacy-preserving principle component analysis (PCA). The key observation of [7] is that the local sensitivity is quite "small" if there is a large eigengap between the k-th and the $k+1$-th eigenvalues. Therefore, their approach (Algorithm 2) chooses to privately release a lower bound of the k-th eigengap (k is fixed as an input) and use that to construct a high-confidence upper bound of the local sensitivity.

For noise-adding mechanisms, the local sensitivity is proportional to the data-dependent loss and generalized PTR is applicable. We can formulate the data-dependent DP of DP-PCA as follows:

Theorem 9.13. *For a given matrix $A \in \mathcal{R}^{m\times n}$, assume each row of A has a bounded ℓ_2 norm being 1. Let V_k denotes the top k eigenvectors of A^TA and d_k denotes the gap between the k-th and the $k+1$-th eigenvalue. Then releasing $V_kV_k^T + E$, where $E \in \mathcal{R}^{n\times n}$ is a symmetric matrix with the upper triangle is i.i.d samples from $\mathcal{N}(0,\sigma^2)$ satisfies $(\epsilon(A), \delta)$ data-dependent DP and $\epsilon(A) = \frac{2\sqrt{\log(1.25/\delta)}}{\sigma(d_k-2)}$.*

The proof is based on the local sensitivity result from [7] and the noise calibration of Gaussian mechanism.

We can combine Theorem 9.13 with our Algorithm 3 to instantiate the generalized PTR framework. The improvement over Dwork et al. [7] will be to allow joint tuning of the parameter k and the noise variance (added to the spectral gap d_k).

10 Omitted proofs in Section 4

The utility of Algorithm 3 depends on how many rounds that Algorithm 2 is invoked. We next provide the utility guarantee of Algorithm 3, which follows a simplification of the result in the Section A.2 of Papernot and Steinke [17].

Theorem 10.1. *Suppose applying Algorithm 2 with each ϕ_i has an equal probability to achieve the highest validation score. Let $\hat{T}$ denotes the number of invocation of Algorithm 2, where $\hat{T}$ follows a truncated geometric distribution. Then the expected quantile of the highest score candidate is given by $\mathbb{E}_{\hat{T}}\left[1 - \frac{1}{\hat{T}+1}\right]$.*

Algorithm 5 OPS-PTR: One-Posterior Sample with propose-test-release (no-"perp" version)

1: **Input**: Data $X, \mathbf{y}$. Private budget : ϵ, δ, proposed regularizer λ.
2: Calculate the minimum eigenvalue $\lambda_{\min}(X^TX)$.
3: Sample $Z \sim \mathcal{N}(0,1)$ and privately release $\tilde{\lambda}_{\min} = \max\left\{\lambda_{\min} + \frac{\sqrt{\log(6/\delta)}}{\epsilon/4}Z - \frac{\sqrt{2\log(6/\delta)\cdot\log(2/\delta)}}{\epsilon/4}, 0\right\}$
4: Calculate $\hat{\theta} = (X^TX + \lambda I)^{-1}X^Ty$.
5: Sample $Z \sim \mathcal{N}(0,1)$ and privately release $\Delta = \log(||\mathcal{Y}|| + ||\mathcal{X}||||\hat{\theta}||) + \frac{\log(1+||\mathcal{X}||^2/(\lambda+\tilde{\lambda}_{\min}))}{\epsilon/(4\sqrt{6/\delta})}Z + \frac{\log(1+||\mathcal{X}||^2/(\lambda+\tilde{\lambda}_{\min}))}{\epsilon/(4\sqrt{2\log(6/\delta)\log(2/\delta)})}$.
6: Set the local Lipschitz $\tilde{L} := ||X||e^{\Delta}$.
7: Calibrate γ with Theorem 5.1$(\delta/3, \epsilon/2.)$
8: Output $\tilde{\theta} \sim p(\theta|X,\mathbf{y}) \propto \mathbf{e}^{-\frac{\gamma}{2}||\mathbf{y}-\mathbf{X}\theta||^2+\lambda||\theta||^2}$

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

Proof. Suppose each oracle access to $Q(X)$ has a probability $1/k$ of achiving the best validation accuracy. Let β denote the probability that $\mathcal{A}$ (shorthand for Algorithm 3) outputs the best choice of ϕ_i.

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

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

11 Experimental details

11.1 Experimental details in private linear regression

We start with the privacy calibration of the OPS-PTR algorithm.

Algorithm 5 provides the detailed privacy calibration of the private linear regression problem.

Theorem 11.1. *Algorithm 5 is $(\epsilon, 2\delta)$-DP.*

Proof. There are three data-dependent quantities in Theorem 5.1: $\lambda_{\min}, ||\theta^*_\lambda||$ and L. First, notice that $\lambda_{\min}$ has a global sensitivity of $||\mathcal{X}||^2$ by Weyl's lemma. Under the assumption $||\mathcal{X}||^2 \leq 1$, we privately release $\lambda_{\min}$ using $(\epsilon/4, \delta/3)$ in Step 3. Notice that with probability at least $1-\delta/2$, $\tilde{\lambda}_{\min}$ is a lower bound of $\lambda_{\min}$.

Then, we apply Lemma 11.2 from Wang [26] to privately release $\log(||\mathcal{Y}|| + ||\mathcal{X}||||\hat{\theta}||)$ using $(\epsilon/4, \delta/3)$. Note that both the local Lipschitz constant L and the norm $||\theta^*_\lambda||$ are functions of $\log(||\mathcal{Y}|| + ||\mathcal{X}||||\hat{\theta}||)$. Thus, we can construct a private upper bound of these by post-processing of Δ.

Then, with probability at least $1-\delta$ (by a union bound over $\tilde{\lambda}_{\min}$ and Δ), instantiating Theorem 5.1 with $\tilde{\lambda}_{\min}$ and $\tilde{L}$ provides a valid upper bound of the data-dependent DP. We then tune the parameter γ using the remaining privacy budget $(\epsilon/2, \delta/3)$. □

Lemma 11.2 (Lemma 12 [26])**.** *Let θ^*_λ be the ridge regression estimate with parameter λ and the smallest eigenvalue of X^TX be λ_{min}, then the function $\log(||\mathcal{Y} + ||\mathcal{X}||||\theta^*_\lambda||)$ has a local sensitivity of $\log(1 + \frac{||\mathcal{X}||^2}{\lambda_{min}+\lambda})$.*

11.2 Details of PATE case study

Definition 11.3 (Renyi DP [14])**.** We say a randomized algorithm $\mathcal{M}$ is $(\alpha, \epsilon_{\mathcal{M}}(\alpha))$-RDP with order $\alpha \geq 1$ if for neighboring datasets X, X'

$$\mathbb{D}_\alpha(\mathcal{M}(X)||\mathcal{M}(X')) := \frac{1}{\alpha-1}\log \mathbb{E}_{o\sim\mathcal{M}(X')}\left[\left(\frac{\Pr[\mathcal{M}(X)=o]}{\Pr[\mathcal{M}(X')=o]}\right)^\alpha\right] \leq \epsilon_{\mathcal{M}}(\alpha).$$

At the limit of $\alpha \to \infty$, RDP reduces to $(\epsilon, 0)$-DP. We now define the data-dependent Renyi DP that conditioned on an input dataset X.

Definition 11.4 (Data-dependent Renyi DP [19])**.** We say a randomized algorithm $\mathcal{M}$ is $(\alpha, \epsilon_{\mathcal{M}}(\alpha, X))$-RDP with order $\alpha \geq 1$ for dataset X if for neighboring datasets X'

$$\mathbb{D}_\alpha(\mathcal{M}(X)||\mathcal{M}(X')) := \frac{1}{\alpha-1}\log \mathbb{E}_{o\sim\mathcal{M}(X')}\left[\left(\frac{\Pr[\mathcal{M}(X)=o]}{\Pr[\mathcal{M}(X')=o]}\right)^\alpha\right] \leq \epsilon_{\mathcal{M}}(\alpha, X).$$

RDP features two useful properties.

Lemma 11.5 (Adaptive composition)**.** $\epsilon_{(\mathcal{M}_1,\mathcal{M}_2)} = \epsilon_{\mathcal{M}_1}(\cdot) + \epsilon_{\mathcal{M}_2}(\cdot)$.

Lemma 11.6 (From RDP to DP)**.** *If a randomized algorithm $\mathcal{M}$ satisfies $(\alpha, \epsilon(\alpha))$-RDP, then $\mathcal{M}$ also satisfies $(\epsilon(\alpha) + \frac{\log(1/\delta)}{\alpha-1}, \delta)$-DP for any $\delta \in (0, 1)$.*

Definition 11.7 (Smooth Sensitivity)**.** Given the smoothness parameter β, a β-smooth sensitivity of $f(X)$ is defined as

$$SS_\beta(X) := \max_{d\geq 0} e^{-\beta d} \cdot \max_{\tilde{X}':dist(X,\tilde{X}')\leq d} \Delta_{LS}(\tilde{X}')$$

Lemma 11.8 (Private upper bound of data-dependent RDP, Restatement of Theorem 5.6)**.** *] Given a RDP function $\mathrm{RDP}(\alpha, X)$ and a β-smooth sensitivity bound $SS(\cdot)$ of $\mathrm{RDP}(\alpha, X)$. Let μ (defined in Algorithm 4) denote the private release of $\log(SS_\beta(X))$. Let $(\beta, \sigma_s, \sigma_2)$-GNSS mechanism be*

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

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

Proof sketch. We first show that releasing the smooth sensitivity SS_β with e^μ satisfies $(\alpha, \frac{\alpha}{2\sigma_2^2})$-RDP. Notice that the log of $SS_\beta(X)$ has a bounded global sensitivity β (Definition 11.7 implies that $|\log SS_\beta(X) - \log SS_\beta(X')| \leq \beta$ for any neighboring dataset X, X'). By Gaussian mechanism, scaling noise with $\beta\sigma_2$ to $\log SS_\beta(X)$ is $(\alpha, \frac{\alpha}{2\sigma_2^2})$-RDP. Therefore, the release of $\mathrm{RDP}(\alpha, X)$ is $(\alpha, \epsilon_s(\alpha) + \frac{\alpha}{2\sigma_2^2})$-RDP. Since the release of $f(X) + SS_\beta(X)\cdot\mathcal{N}(0,\sigma_s^2)$ is $(\alpha, \frac{\alpha+1}{\sigma_s^2})$-RDP (Theorem 23 from Papernot et al. [19]) for $\alpha < \frac{1}{2\beta}$, we have $\epsilon_s(\alpha) + \frac{\alpha}{2\sigma_2^2} = \frac{3\alpha+2}{2\sigma_s^2}$.

We next prove the second statement. First, notice that with probability at least $1 - \delta_2/2$, $e^\mu \geq SS_\beta(X)$ using the standard Gaussian tail bound. Let E denote the event that $e^\mu \geq SS_\beta(X)$.

$$\begin{aligned}
&\Pr\left[\mathrm{RDP}^{\text{upper}}(\alpha) \leq \mathrm{RDP}(\alpha, X)\right] \\
&= \Pr\left[\mathrm{RDP}^{\text{upper}}(\alpha) \leq \mathrm{RDP}(\alpha, X)|E\right] + \Pr\left[\mathrm{RDP}^{\text{upper}}(\alpha) \leq \mathrm{RDP}(\alpha, X)|E^c\right] \\
&\leq \Pr\left[\mathrm{RDP}^{\text{upper}}(\alpha) \leq \mathrm{RDP}(\alpha, X)|E\right] + \delta_2/2 \\
&= \underbrace{\Pr\left[\mathcal{N}(0,\sigma_s^2)\cdot SS_{\beta(X)} \geq \sigma_s\cdot\sqrt{2\log(2/\delta_2)}e^\mu|E\right]}_{\text{denoted by}(*)} + \delta_2/2
\end{aligned}$$

Condition on the event E, e^{μ} is a valid upper bound of $SS_\beta(X)$, which implies

$$(*) \leq \Pr[\mathcal{N}(0, \sigma_s^2) \cdot SS_\beta(X) \geq \sigma_s \cdot \sqrt{2\log(2/\delta_2)} SS_\beta(X) | E] \leq \delta_2/2$$

Therefore, with probability at least $1-\delta_2$, $\mathrm{RDP}^{\text{upper}}(\alpha) \geq \mathrm{RDP}(\alpha, X)$. □

Theorem 11.9 (Restatement of Theorem 5.7)**.** *Algorithm 4 satisfies $(\epsilon' + \hat{\epsilon}, \delta)$-DP.*

Proof. The privacy analysis consists of two components — the privacy cost of releasing an upper bound of data-dependent RDP ($\epsilon_{\text{upper}}(\alpha) := \epsilon_s(\alpha) + \frac{\alpha}{2\sigma_2^2}$ and the valid upper bound $\epsilon^p_{\sigma_1}(\alpha)$. First, set $\alpha = \frac{2\log(2/\delta)}{\epsilon} + 1$ and use RDP to DP conversion with $\delta/2$ ensures that the cost of $\delta/2$ contribution to be roughly $\epsilon/2$ (i.e., $\frac{\log(2/\delta)}{\alpha-1} = \epsilon/2$). Second, choosing $\sigma_s = \sqrt{\frac{2+3\alpha}{\epsilon}}$ gives us another $\epsilon/2$. □

Experimental details $K = 400$ teacher models are trained individually on the disjoint set using AlexNet model. We set $\sigma_2 = \sigma_s = 15.0$. Our data-dependent RDP calculation and the smooth-sensitivity calculation follow Papernot et al. [19]. Specifically, we use the following theorem (Theorem 6 from Papernot et al. [19]) to compute the data-dependent RDP of each unlabeled data x from the public domain.

Theorem 11.10 (data-dependent RDP Papernot et al. [19])**.** *Let $\tilde{q} \geq \Pr[\mathcal{M}(X) \neq Argmax_{j\in[C]} n_j(x)]$, i.e., an upper bound of the probability that the noisy label does not match the majority label. Assume $\alpha \leq \mu_1$ and $\tilde{q} \leq e^{(\mu_2-1)\epsilon_2} / \left(\frac{\mu_1}{\mu_1-1} \cdot \frac{\mu_2}{\mu_2-1}\right)^{\mu_2}$, then we have:*

$$\epsilon_{\mathcal{M}}(\alpha, X) \leq \frac{1}{\alpha-1} \log\left((1-\tilde{q}) \cdot A(\tilde{q}, \mu_2, \epsilon_2)^{\alpha-1} + \tilde{q} \cdot B(\tilde{q}, \mu_1, \epsilon_1)^{\alpha-1}\right)$$

where $A(\tilde{q}, \mu_2, \epsilon_2) := (1-\tilde{q}) / \left(1 - (\tilde{q}e^{\epsilon_2})^{\frac{\mu_2-1}{\mu_2}}\right)$, $B(\tilde{q}, \mu_1, \epsilon_1) = e^{\epsilon_1} / \tilde{q}^{\frac{1}{\mu_1-1}}$, $\mu_2 = \sigma_1 \cdot \sqrt{\log(1/\tilde{q})}$, $\mu_1 = \mu_2 + 1$, $\epsilon_1 = \mu_1/\sigma_1^2$ and $\epsilon_2 = \mu_2/\sigma_2^2$.

In the experiments, the non-private data-dependent DP baseline is also based on the above theorem. Notice that the data-dependent RDP of each query is a function of $\tilde{q}$, where $\tilde{q}$ denotes an upper bound of the probability where the plurality output does not match the noisy output. $\tilde{q}$ is a complex function of both the noisy scale and data and is not monotonically decreasing when σ_1 is increasing.

Simulation of two distributions. The motivation of the experimental design is to compare three approaches under different data distributions. Notice that there are $K = 400$ teachers, which implies the number of the vote count for each class will be bounded by 400. In the simulation of high-consensus distribution, we choose $T = 200$ unlabeled public data such that the majority vote count will be larger than 150 (i.e., $\max_{j\in[C]} n_j(x) > 150$). For the low-consensus distribution, we choose to select T unlabeled data such that the majority vote count will be smaller than 150.

12 Omitted proofs in private GLM

12.1 Per-instance DP of GLM

Theorem 12.1 (Per-instance differential privacy guarantee)**.** *Consider two adjacent data sets Z and $Z' = [Z, (x, y)]$, and denote the smooth part of the loss function $F_s = \sum_{i=1}^n l(y_i, \langle x_i, \cdot\rangle) + r_s(\cdot)$ (thus $\tilde{F}_s = F_s + l(y, \langle x, \cdot\rangle)$. Let the local neighborhood be the line segment between θ^* and $\tilde{\theta}^*$. Assume*

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

Then the algorithm obeys (ϵ, δ)-pDP for Z and $z = (x, y)$ with any $0 < \delta < 2/e$ and

$$\epsilon \leq \epsilon_0(1 + \log(2/\delta)) + e^{\frac{RL\|x\|_2}{\alpha}} \left[\frac{\gamma L^2 \|x\|^2_{H^{-1}}}{2} + \sqrt{\gamma L^2 \|x\|^2_{H^{-1}} \log(2/\delta)} \right]$$

where $\epsilon_0 \leq e^{\frac{RL\|x\|_2}{\alpha}} - 1 + 2\beta\|x\|^2_{H_1^{-1}} + 2\beta\|x\|^2_{\tilde{H}_2^{-1}}$. If we instead assume that l is R-self concordant. Then the same results hold, but with all $e^{\frac{RL\|x\|_2}{\alpha}}$ replaced with $(1 - RL\|x\|_{H^{-1}})^2$.

Under the stronger three-times continuous differentiable assumption, by mean value theorem, there exists ξ on the line-segment between θ^* and $\tilde{\theta}^*$ such that

$$H = \left[\int_{t=0}^{1} \nabla^2 F_s((1-t)\theta^* + t\tilde{\theta}^*) dt \right] = \nabla^2 F_s(\xi).$$

The two distributions of interests are $\mathcal{N}(\theta^*, [\gamma \nabla^2 F_s(\theta^*)]^{-1})$ and $\mathcal{N}(\tilde{\theta}^*, [\gamma \nabla^2 F_s(\tilde{\theta}^*) + \nabla^2 l(y, x^T \tilde{\theta}^*)]^{-1})$. Denote $[\nabla^2 F_s(\theta^*)]^{-1} =: \Sigma$ and $[\nabla^2 F_s(\tilde{\theta}^*) + \nabla^2 l(y, x^T \tilde{\theta}^*)]^{-1} =: \tilde{\Sigma}$. Both the means and the covariance matrices are different, so we cannot use multivariate Gaussian mechanism naively. Instead we will take the tail bound interpretation of (ϵ, δ)-DP and make use of the per-instance DP framework as internal steps of the proof.

First, we can write down the privacy loss random variable in analytic form

$$\log \frac{|\Sigma|^{-1/2} e^{-\frac{\gamma}{2}\|\theta - \theta^*\|^2_{\Sigma^{-1}}}}{|\tilde{\Sigma}|^{-1/2} e^{-\frac{\gamma}{2}\|\theta - \tilde{\theta}^*\|^2_{\tilde{\Sigma}^{-1}}}} = \underbrace{\frac{1}{2} \log \left(\frac{|\Sigma^{-1}|}{|\tilde{\Sigma}^{-1}|} \right)}_{(*)} + \underbrace{\frac{\gamma}{2} \left[\|\theta - \theta^*\|^2_{\Sigma^{-1}} - \|\theta - \tilde{\theta}^*\|^2_{\tilde{\Sigma}^{-1}} \right]}_{(**)}$$

The general idea of the proof is to simplify the expression above and upper bounding the two terms separately using self-concordance and matrix inversion lemma, and ultimately show that the privacy loss random variable is dominated by another random variable having an appropriately scaled shifted χ-distribution, therefore admits a Gaussian-like tail bound.

To ensure the presentation is readable, we define a few short hands. We will use H and $\tilde{H}$ to denote the Hessian of F_s and $F_s + f$ respectively and subscript 1 2 indicates whether the Hessian evaluated at at θ^* or $\tilde{\theta}^*$. H without any subscript or superscript represents the Hessian of F_s evaluated at ξ as previously used.

$$(*) = \frac{1}{2} \log \frac{|H_1|}{|H|} \frac{|H|}{|H_2|} \frac{|H_2|}{|\tilde{H}_2|} \leq \frac{1}{2} \left[\log \frac{|H_1|}{|H|} + \log \frac{|H|}{|H_2|} + \log \frac{|H_2|}{|\tilde{H}_2|} \right]$$

By the R-generalized self-concordance of F_s, we can apply Lemma 12.3,

$$-\|\theta^* - \xi\|_2 R \leq \log \frac{|H_1|}{|H|} \leq R\|\theta^* - \xi\|_2, \quad -R\|\xi - \tilde{\theta}^*\|_2 \leq \log \frac{|H|}{|H_2|} \leq R\|\xi - \tilde{\theta}^*\|_2.$$

The generalized linear model ensures that the Hessian of f is rank-1:

$$\nabla^2 f(\tilde{\theta}^*) = l''(y, x^T \tilde{\theta}^*) x x^T$$

and we can apply Lemma **??** in both ways (taking $A = H_2$ and $A = \tilde{H}_2$) and obtain

$$\frac{|H_2|}{|\tilde{H}_2|} = \frac{1}{1 + l''(y, x^T \tilde{\theta}^*) x^T H_2^{-1} x} = 1 - l''(y, x^T \tilde{\theta}^*) x^T \tilde{H}_2 x$$

Note that $l''(y, x^T \tilde{\theta}^*) x^T \tilde{H}_2^{-1} x$ is the in-sample leverage-score and $l''(y, x^T \tilde{\theta}^*) x^T H_2^{-1} x$ is the out-of-sample leverage-score of the locally linearized problem at $\tilde{\theta}^*$. We denote them by μ_2 and μ_2' respectively (similarly, for the consistency of notations, we denote the in-sample and out of sample leverage score at θ^* by μ_1 and μ_1').

Combine the above arguments we get

$$(*) \leq R\|\theta^* - \xi\|_2 + R\|\xi - \tilde{\theta}^*\|_2 + \log(1 - \mu_2) \leq R\|\theta^* - \tilde{\theta}^*\|_2 + \log(1 - \mu_2) \tag{6}$$

$$(*) \geq -R\|\theta^* - \tilde{\theta}^*\|_2 - \log(1 - \mu_2). \tag{7}$$

We now move on to deal with the second part, where we would like to express everything in terms of $\|\theta - \theta^*\|_{H_1}$, which we know from the algorithm is χ-distributed.

$$(**) = \frac{\gamma}{2}\left[\|\theta-\theta^*\|_{H_1}^2 - \|\theta-\theta^*\|_{H_2}^2 + \|\theta-\theta^*\|_{H_2}^2 - \|\theta-\tilde{\theta}^*\|_{H_2}^2 + \|\theta-\tilde{\theta}^*\|_{H_2}^2 - \|\theta-\tilde{\theta}^*\|_{\tilde{H}_2}^2\right]$$

By the generalized self-concordance at θ^*

$$e^{-R\|\theta^*-\tilde{\theta}^*\|_2}\|\cdot\|_{H_1}^2 \le \|\cdot\|_{H_2}^2 \le e^{R\|\theta^*-\tilde{\theta}^*\|_2}\|\cdot\|_{H_1}^2$$

This allows us to convert from $\|\cdot\|_{H_2}$ to $\|\cdot\|_{H_1}$, and as a consequence:

$$\left|\|\theta-\theta^*\|_{H_1}^2 - \|\theta-\theta^*\|_{H_2}^2\right| \le [e^{R\|\theta^*-\tilde{\theta}^*\|_2} - 1]\|\theta-\theta^*\|_{H_1}^2.$$

Also,

$$\|\theta-\theta^*\|_{H_2}^2 - \|\theta-\tilde{\theta}^*\|_{H_2}^2 = \left\langle \tilde{\theta}^*-\theta^*, 2\theta - 2\theta^* + \theta^* - \tilde{\theta}^*\right\rangle_{H_2} = 2\langle \theta-\theta^*, \tilde{\theta}^*-\theta^*\rangle_{H_2} - \|\theta^*-\tilde{\theta}^*\|_{H_2}^2$$

Therefore

$$\begin{aligned}\left|\|\theta-\theta^*\|_{H_2}^2 - \|\theta-\tilde{\theta}^*\|_{H_2}^2\right| &\le 2\|\theta-\theta^*\|_{H_2}\|\theta^*-\tilde{\theta}^*\|_{H_2} + \|\theta^*-\tilde{\theta}^*\|_{H_2}^2 \\ &\le 2e^{R\|\tilde{\theta}^*-\theta^*\|_2}\|\theta-\theta^*\|_{H_1}\|\theta^*-\tilde{\theta}^*\|_H + e^{R\|\tilde{\theta}^*-\theta^*\|_2}\|\theta^*-\tilde{\theta}^*\|_H^2.\end{aligned}$$

Then lastly we have

$$\begin{aligned}0 \ge \|\theta-\tilde{\theta}^*\|_{H_2}^2 - \|\theta-\tilde{\theta}^*\|_{\tilde{H}_2}^2 &= -l''(y, x^T\tilde{\theta}^*)\left[\langle x, \theta-\theta^*\rangle + \langle x, \theta^*-\tilde{\theta}^*\rangle\right]^2 \\ &\ge -2\beta\|x\|_{H_1^{-1}}^2\|\theta-\theta^*\|_{H_1}^2 - 2\beta\|x\|_{H^{-1}}^2\|\theta^*-\tilde{\theta}^*\|_H^2\end{aligned}$$

$$\left|\|\theta-\tilde{\theta}^*\|_{H_2}^2 - \|\theta-\tilde{\theta}^*\|_{\tilde{H}_2}^2\right| \le 2\beta\|x\|_{H_1^{-1}}^2\|\theta-\theta^*\|_{H_1}^2 + 2\beta\|x\|_{H^{-1}}^2\|\theta^*-\tilde{\theta}^*\|_H^2$$

Combine the above derivations, we get

$$|(**)| \le \frac{\gamma}{2}\left[a\|\theta-\theta^*\|_{H_1}^2 + b\|\theta-\theta^*\|_{H_1} + c\right] \tag{8}$$

where

$$\begin{aligned}a &:= \left[e^{R\|\theta^*-\tilde{\theta}^*\|_2} - 1 + 2\beta\|x\|_{H_1^{-1}}^2\right] \\ b &:= 2e^{R\|\theta^*-\tilde{\theta}^*\|_2}\|\theta^*-\tilde{\theta}^*\|_H \\ c &:= (e^{R\|\theta^*-\tilde{\theta}^*\|_2} + 2\beta\|x\|_{H^{-1}}^2)\|\theta^*-\tilde{\theta}^*\|_H^2\end{aligned}$$

Lastly, by (6) and (8),

$$\left|\log\frac{p(\theta|Z)}{p(\theta|Z')}\right| \le R\|\theta^*-\tilde{\theta}^*\|_2 + \log(1-\mu_2) + \frac{\gamma}{2}[aW^2 + bW + c].$$

where according to the algorithm $W := \|\theta-\theta^*\|_{H_1}$ follows a half-normal distribution with $\sigma = \gamma^{-1/2}$.

By standard Gaussian tail bound, we have for all $\delta < 2/e$.

$$\mathbb{P}(|W| \le \gamma^{-1/2}\sqrt{\log(2/\delta)}) \le \delta.$$

This implies that a high probability upper bound of the absolute value of the privacy loss random variable $\log\frac{p(\theta|Z)}{p(\theta|Z')}$ under $p(\theta|Z)$. By the tail bound to privacy conversion lemma (Lemma **??**), we get that for any set $S \subset \Theta$ $\mathbb{P}(\theta \in S|Z) \le e^{\epsilon}\mathbb{P}(\theta \in S|Z') + \delta$ for any $0 < \delta < 2/e$ and

$$\epsilon = R\|\theta^*-\tilde{\theta}^*\|_2 + \log(1-\mu_2) + \frac{\gamma c}{2} + \frac{a}{2}\log(2/\delta) + \frac{\gamma^{1/2}b}{2}\sqrt{\log(2/\delta)}.$$

Denote $v := \theta^* - \tilde{\theta}^*$, by strong convexity

$$\|v\|_2 \leq \|\nabla l(y, x^T\theta)[\tilde{\theta}^*]\|_2/\alpha = |l'|\|x\|_2/\alpha \leq L\|x\|_2/\alpha$$

and

$$\|v\|_H \leq \|\nabla l(y, x^T\theta)[\tilde{\theta}^*]\|_{H^{-1}} = |l'|\|x\|_{H^{-1}} \leq L\|x\|_{H^{-1}}.$$

Also use the fact that $|\log(1-\mu_2)| \leq 2\mu_2$ for $\mu_2 < 0.5$ and $\mu_2 \leq \beta\|x\|^2_{\tilde{H}_2^{-1}}$, we can then combine similar terms and have a more compact representation.

$$\epsilon \leq \epsilon_0(1+\log(2/\delta)) + e^{\frac{RL\|x\|_2}{\alpha}}\left[\frac{\gamma L^2\|x\|^2_{H^{-1}}}{2} + \sqrt{\gamma L^2\|x\|^2_{H^{-1}}\log(2/\delta)}\right]$$

where

$$\epsilon_0 \leq e^{\frac{RL\|x\|_2}{\alpha}} - 1 + 2\beta\|x\|^2_{H_1^{-1}} + 2\beta\|x\|^2_{\tilde{H}_2^{-1}}$$

is the part of the privacy loss that does not get smaller as γ decreases.

Proposition 12.2. *Let $\|\cdot\|$ be a norm and $\|\cdot\|_*$ be its dual norm. Let $F(\theta)$, $f(\theta)$ and $\tilde{F}(\theta) = F(\theta) + f(\theta)$ be proper convex functions and θ^* and $\tilde{theta}^*$ be their minimizers, i.e., $0 \in \partial F(\theta^*)$ and $0 \in \partial\tilde{F}(\tilde{theta}^*)$. If in addition, $F, \tilde{F}$ is $\alpha, \tilde{\alpha}$-strongly convex with respect to $\|\cdot\|$ within the restricted domain $\theta \in \{t\theta^* + (1-t)\tilde{\theta}^* \mid t \in [0,1]\}$. Then there exists $g \in \partial f(\theta^*)$ and $\tilde{g} \in \partial f(\tilde{\theta}^*)$ such that*

$$\|\theta^* - \tilde{\theta}^*\| \leq \min\left\{\frac{1}{\alpha}\|\tilde{g}\|_*, \frac{1}{\tilde{\alpha}}\|g\|_*\right\}.$$

Proof. Apply the first order condition to F restricted to the line segment between $\tilde{\theta}^*$ and θ^*, there are we get

$$F(\tilde{\theta}^*) \geq F(\theta^*) + \langle \partial F(\theta^*), \tilde{\theta}^* - \theta^*\rangle + \frac{\alpha}{2}\|\tilde{\theta}^* - \theta^*\|^2 \tag{9}$$

$$F(\theta^*) \geq F(\tilde{\theta}^*) + \langle \partial F(\tilde{\theta}^*), \theta^* - \tilde{\theta}^*\rangle + \frac{\alpha}{2}\|\tilde{\theta}^* - \theta^*\|^2 \tag{10}$$

Note by the convexity of F and f, $\partial\tilde{F} = \partial F + \partial f$, where $+$ is the Minkowski Sum. Therefore, $0 \in \partial\tilde{F}(\tilde{\theta}^*)$ implies that there exists $\tilde{g}$ such that $\tilde{g} \in \partial f(\tilde{\theta}^*)$ and $-\tilde{g} \in \partial F(\tilde{\theta}^*)$. Take $-\tilde{g} \in \partial F(\tilde{\theta}^*)$ in Equation 10 and $0 \in \partial F(\theta^*)$ in Equation 9 and add the two inequalities, we obtain

$$0 \geq \langle -\tilde{g}, \theta^* - \tilde{\theta}^*\rangle + \alpha\|\tilde{\theta}^* - \theta^*\|^2 \geq -\|\tilde{g}\|_*\|\theta^* - \tilde{\theta}^*\| + \alpha\|\tilde{\theta}^* - \theta^*\|^2.$$

For $\|\tilde{\theta}^* - \theta^*\| = 0$ the claim is trivially true, otherwise, we can divide the both sides of the above inequality by $\|\tilde{\theta}^* - \theta^*\|$ and get $\|\theta^* - \tilde{\theta}^*\| \leq \frac{1}{\alpha}\|\tilde{g}\|_*$.

It remains to show that $\|\theta^* - \tilde{\theta}^*\| \leq \frac{1}{\tilde{\alpha}}\|g\|_*$. This can be obtained by exactly the same arguments above but applying strong convexity to $\tilde{F}$ instead. Note that we can actually get something slightly stronger than the statement because the inequality holds for all $g \in \partial f(\theta^*)$. □

A consequence of (generalized) self-concordance is the spectral (*multiplicative*) stability of Hessian to small perturbations of parameters.

Lemma 12.3 (Stability of Hessian[15, Theorem 2.1.1], [1, Proposition 1])**.** *Let $H_\theta := \nabla^2 F_s(\theta)$. If F_s is R-self-concordant at θ. Then for any v such that $R\|v\|_{H_\theta} < 1$, we have that*

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

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

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

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

References

[1] Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[2] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

[3] Chris Decarolis, Mukul Ram, Seyed Esmaeili, Yu-Xiang Wang, and Furong Huang. An end-to-end differentially private latent dirichlet allocation using a spectral algorithm. In *International Conference on Machine Learning*, pages 2421–2431. PMLR, 2020.

[4] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *ACM symposium on Theory of computing*, pages 371–380, 2009.

[5] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

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

[7] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 11–20, 2014.

[8] Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. Private convex optimization via exponential mechanism. *arXiv preprint arXiv:2203.00263*, 2022.

[9] Shiva Prasad Kasiviswanathan, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Analyzing graphs with node differential privacy. In *Theory of Cryptography Conference*, pages 457–476. Springer, 2013.

[10] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1. JMLR Workshop and Conference Proceedings, 2012.

[11] Jingcheng Liu and Kunal Talwar. Private selection from private candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 298–309, 2019.

[12] Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. *arXiv preprint arXiv:2111.06578*, 2021.

[13] Kentaro Minami, HItomi Arai, Issei Sato, and Hiroshi Nakagawa. Differential privacy without sensitivity. *Advances in Neural Information Processing Systems*, 29, 2016.

[14] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

[15] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

[16] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *ACM symposium on Theory of computing (STOC-07)*, pages 75–84. ACM, 2007.

[17] Nicolas Papernot and Thomas Steinke. Hyperparameter tuning with renyi differential privacy. *arXiv preprint arXiv:2110.03620*, 2021.

[18] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations (ICLR-17)*, 2017.

[19] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018.

[20] Rachel Redberg and Yu-Xiang Wang. Privately publishable per-instance privacy. *Advances in Neural Information Processing Systems*, 34, 2021.

[21] Jordi Soria-Comas, Josep Domingo-Ferrer, David Sánchez, and David Megías. Individual differential privacy: A utility-preserving formulation of differential privacy guarantees. *IEEE Transactions on Information Forensics and Security*, 12(6):1418–1429, 2017.

[22] Abhradeep Guha Thakurta and Adam Smith. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory*, pages 819–850. PMLR, 2013.

[23] Salil Vadhan. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer, 2017.

[24] Jiachen T Wang, Saeed Mahloujifar, Shouda Wang, Ruoxi Jia, and Prateek Mittal. Renyi differential privacy of propose-test-release and applications to private and robust machine learning. *arXiv preprint arXiv:2209.07716*, 2022.

[25] Yu-Xiang Wang. Per-instance differential privacy and the adaptivity of posterior sampling in linear and ridge regression. *arXiv preprint arXiv:1707.07708*, pages 48–71, 2017.

[26] Yu-Xiang Wang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. *arXiv preprint arXiv:1803.02596*, 2018.

[27] Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning*, pages 2493–2502. PMLR, 2015.