On Convex Optimization with Semi-Sensitive Features

Badih Ghazi BADIHGHAZI@GMAIL.COM
Google Research, Mountain View, USA

Pritish Kamath PRITISH@ALUM.MIT.EDU
Google Research, Mountain View, USA

Ravi Kumar RAVI.K53@GMAIL.COM
Google Research, Mountain View, USA

Pasin Manurangsi PASIN@GOOGLE.COM
Google Research, Thailand

Raghu Meka RAGHUM@CS.UCLA.EDU
University of California, Los Angeles, USA

Chiyuan Zhang CHIYUAN@GOOGLE.COM
Google Research, Mountain View, USA

Editors: Shipra Agrawal and Aaron Roth

Abstract

We study the differentially private (DP) empirical risk minimization (ERM) problem under the *semi-sensitive DP* setting where only some features are sensitive. This generalizes the Label DP setting where only the label is sensitive. We give improved upper and lower bounds on the excess risk for DP-ERM. In particular, we show that the error only scales polylogarithmically in terms of the sensitive domain size, improving upon previous results that scale polynomially in the sensitive domain size (Ghazi et al., 2021).

Keywords: Differential Privacy, Semi-sensitive Features, Label Differential Privacy, Convex Optimization

1. Introduction

In empirical risk minimization (ERM) problem, we are given a dataset $D = \{x_i\}_{i\in[n]} \in \mathcal{X}^n$ and a loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ and the goal is to find $w \in \mathcal{W}$ that minimizes the empirical risk $\mathcal{L}(w; D) := \frac{1}{n}\sum_{i\in[n]} \ell(w; x_i)$. The excess risk is defined as $\mathcal{L}(w; D) - \min_{w'\in\mathcal{W}} \mathcal{L}(w'; D)$. Often, the dataset might contain sensitive data, and to provide privacy protection, we will use the notion of differential privacy (DP) (Dwork et al., 2006). ERM is among the most well-studied problems in the DP literature and tight excess risk bounds are known under assumptions such as Lipschitzness, convexity, and strong convexity (e.g., Chaudhuri et al. (2011); Kifer et al. (2012); Bassily et al. (2014); Wang et al. (2017); Bassily et al. (2019); Feldman et al. (2020); Gopi et al. (2022)).

In most of these studies, each x_i is assumed to be sensitive. However, in several applications, such as online advertising, it can be the case that x_i consists of both sensitive and non-sensitive attributes. This can be modeled by letting $\mathcal{X} = \mathcal{X}^{\text{pub}} \times \mathcal{X}^{\text{priv}}$ where $\mathcal{X}^{\text{pub}}$ is the domain of the *non-sensitive* features and $\mathcal{X}^{\text{priv}}$ is the domain of the *sensitive* features. We will also write each example x as $(x^{\text{pub}}, x^{\text{priv}})$ where $x^{\text{pub}} \in \mathcal{X}^{\text{pub}}$ and $x^{\text{priv}} \in \mathcal{X}^{\text{priv}}$. Here, our only aim is to protect x^{priv}; in terms of DP, this means that we allow two neighboring datasets to differ only on the sensitive

features of a single example. We refer to this DP notion as *semi-sensitive DP*. (See Section 2.1 for a more formal definition.) Our definition is identical to the ones considered by Chua et al. (2024); Shen et al. (2023); a related notion has been considered recently as well (Krichene et al., 2024). To avoid confusion, we refer to the standard DP notion (where a single entire example can be changed in neighboring datasets) as *full DP*. Throughout, we use k to denote the size of the domain for private features, i.e., $k = |\mathcal{X}^{\text{priv}}|$.

The semi-sensitive DP model generalizes the so-called *label DP* (Ghazi et al., 2023) where the only sensitive "feature" is the label.[1] In our language, Ghazi et al. (2021) give an ε-semi-sensitive DP algorithm for convex ERM (under Lipschitzness assumption) that yields an expected excess risk of $\tilde{O}\left(\frac{k}{\varepsilon\sqrt{n}}\right)$; for (ε, δ)-semi-sensitive DP, they achieve an expected excess risk of $\tilde{O}\left(\frac{\sqrt{k\log(1/\delta)}}{\varepsilon\sqrt{n}}\right)$. Complementing these upper bounds, they also provide a lower bound of $\Omega\left(\frac{1}{\varepsilon\sqrt{n}}\right)$ against any (ε, δ)-semi-sensitive DP. An interesting aspect of these bounds is that they are dimension-independent; meanwhile, for full DP, it is known that the expected excess risk grows (polynomially) with the dimension of $\mathcal{W}$ (Bassily et al., 2014). Despite this, the results from Ghazi et al. (2021) leave a rather large gap in terms of k: the upper bounds have a polynomial dependence on k that is not captured by the lower bound.

1.1. Our Contributions

The main contribution of our paper is to (nearly) close this gap. In particular, we show that the dependency on k is polylogarithmic rather than polynomial, as stated below.

Theorem 1 (Informal; see Theorem 18) *For $\varepsilon \le O(\log 1/\delta)$, there is an (ε, δ)-semi-sensitive DP algorithm for ERM w.r.t. any G-Lipschitz convex loss function with domain radius R that has expected excess empirical risk $\tilde{O}\left(RG \cdot \frac{\sqrt[4]{\log(1/\delta)\cdot\log k}}{\sqrt{\varepsilon n}}\right)$.*

Theorem 2 (Informal; see Theorem 33) *For $\varepsilon \le O(\log k)$, any $(\varepsilon, o(1/k))$-semi-sensitive DP algorithm for ERM w.r.t. any G-Lipschitz convex loss function with domain radius R has expected excess empirical risk at least $\Omega\left(RG \cdot \min\left\{1, \frac{\sqrt{\log k}}{\sqrt{\varepsilon n}}\right\}\right)$.*

Notice that the dependency on k is essentially tight for $\delta = 1/k^{1+\Theta(1)}$. It remains an interesting open question to tighten the bound for a wider regime of δ values.

When the loss function is further assumed to be strongly convex and smooth, we can improve on the above excess risk and also provide a nearly tight bound in this case.

Theorem 3 (Informal; see Theorem 19) *For $\varepsilon \le O(\log 1/\delta)$, there is an (ε, δ)-semi-sensitive DP algorithm for ERM w.r.t. any G-Lipschitz μ-strongly convex λ-smooth loss function that has expected excess empirical risk $\tilde{O}\left(\frac{G^2}{\mu} \cdot \frac{\sqrt{\log(1/\delta)\cdot\log k\cdot\log(\lambda/\mu)}}{\varepsilon n}\right)$.*

Theorem 4 (Informal; see Theorem 34) *For $\varepsilon \le O(\log k)$, any $(\varepsilon, o(1/k))$-semi-sensitive DP algorithm for ERM w.r.t. any G-Lipschitz μ-strongly convex μ-smooth loss function has expected excess empirical risk at least $\Omega\left(\frac{G^2}{\mu} \cdot \min\left\{1, \frac{\log k}{\varepsilon n}\right\}\right)$.*

1. In our formulation of ERM, there is no distinction between a label and a feature; indeed, the two models are equivalent.

Finally, our techniques are sufficient for solving *multiple* convex ERM problems on the same input dataset, where the error grows only polylogarithmic in the number of ERM problems:

Theorem 5 (Informal; see Theorem 18) *For $\varepsilon \leq O(\log 1/\delta)$, there is an (ε, δ)-semi-sensitive DP algorithm for m ERM problems w.r.t. any G-Lipschitz convex loss function with domain radius R that has expected excess empirical risk $\tilde{O}\left(RG \cdot \frac{\sqrt[4]{\log(1/\delta)\cdot\log k}\sqrt{\log m}}{\sqrt{\varepsilon n}}\right)$.*

Theorem 6 (Informal; see Theorem 19) *For $\varepsilon \leq O(\log 1/\delta)$, there is an (ε, δ)-semi-sensitive DP algorithm for m ERM problems w.r.t. any G-Lipschitz μ-strongly convex λ-smooth loss function that has expected excess empirical risk $\tilde{O}\left(\frac{G^2}{\mu} \cdot \frac{\sqrt{\log(1/\delta)\cdot\log k}\cdot\log(m\lambda/\mu)}{\varepsilon n}\right)$.*

In the full DP setting, this problem was first studied by Ullman (2015). The error bound was improved by Feldman et al. (2017), but their bound still depends polylogarithmically on the dimension d. By removing this dependency, our theorem above improves upon the bounds of Feldman et al. (2017). We note, however, that Feldman et al. (2017) also give bounds for the ℓ_p-bounded setting for any $p \neq 2$, but, for simplicity, we do not consider this in our work.

1.2. Technical Overview

In this section, we briefly discuss the techniques used in our work.

Answer Linear Vector Queries. The key ingredient of our work is an algorithm that can answer online linear *vector* queries. Such a query is of the form $f : \mathcal{X} \to \mathcal{B}_2^d(1)$ where $\mathcal{B}_2^d(1)$ denotes the (Euclidean) unit ball in d dimensions, and our goal is to approximate $f(D) := \frac{1}{n}\sum_{i\in[n]} f(x_i)$. There can be up to T online queries (i.e., we have to answer the previous query before receiving the next).

The case $d = 1$ is often referred to as linear queries. In this case, in the full DP setting, Hardt and Rothblum (2010) introduced the "Private Multiplicative Weights" algorithm that has error $\mathrm{polylog}(|\mathcal{X}|, \frac{1}{\delta})/\sqrt{\varepsilon n}$. We extend their algorithm in two crucial aspects:

(i) We adapt the algorithm to the semi-sensitive DP setting and show that we can improve on the error in this setting: the $|\mathcal{X}|$ term (size of the entire input domain) becomes $k = |\mathcal{X}^{\mathrm{priv}}|$ (size of the sensitive features domain).

(ii) We show a natural way to handle $d > 1$. In this case, the error is now measured in the ℓ_2-error of the vector. Interestingly, we show that the error remains (roughly) the same in this setting and is in fact dimension-independent. This is crucial for achieving dimension-independent bounds in our theorems.

The algorithm of Hardt and Rothblum (2010) works by maintaining a distribution over all the domain $\mathcal{X}$. For each query, we (privately) check whether the current distribution is sufficiently accurate to answer the query. If so, we answer using the current distribution. Otherwise, we apply a multiplicative weight update (MWU) rule to update the distribution. The MWU rule depends on the privatized true answer and the answer computed using the current distribution, where the former is achieved via, e.g., adding Laplace noise. The crux of the analysis is that the privacy budget is only charged when an update occurs. Finally, a standard analysis of MWU shows that there cannot be too many updates.

To achieve (i), our algorithm maintains, for each example x_i, a distribution over x_i^{priv} and applies a multiplicative weight update. For (ii), we modify the update as follows. First, we privatize the true

answer using the Gaussian mechanism. Then, we apply the MWU rule based on the dot product of this privatized true answer and the value of each example. Crucially, our analysis shows that, even though the total norm of the noise can be very large (growing with the dimension), it does not interfere too much with the update as only the noise in a few directions is relevant.

From Answering Linear Vector Queries to Convex Optimization. By letting each query f be the gradient of the loss function, our aforementioned algorithm allows one to construct an approximate gradient oracle. By leveraging existing results in the optimization literature (d'Aspremont, 2008; Devolder et al., 2014, 2013), we immediately arrive at the claimed bounds.

Comparison to Previous Work. Feldman et al. (2017) show that approximate gradient oracle can be accomplished via Statistical Queries (SQs). For the purpose of our high-level discussion, one can think of SQs as just linear queries. Using this, they observe that the Hardt and Rothblum (2010) algorithm can be used to solve convex optimization problem(s) with low error. We note that this approach can be used in our setting, too, once we extend the Hardt–Rothblum algorithm to the semi-sensitive DP setting with decreased error (i). However, this alone does *not* yield a dimension-independent bound since the number of linear queries required still depends on the dimension. As such, we still require (ii) to achieve the results stated here. Finally, we remark that vector versions of MWU have been used in the DP literature before (e.g., Ullman (2015)). However, we are not aware of its study with respect to the effect of Gaussian noise; in particular, to the best of our knowledge, the fact that we still have a dimension-independent bound even after applying noise is novel.

Lower Bounds. Suppose for simplicity that $\varepsilon = \Theta(1)$. For the lower bound, we first recall the construction from previous work (Ghazi et al., 2021), which is a reduction from (vector) mean estimation. Roughly speaking, they let the ith example contribute only to the ith coordinate and let the sensitive feature (which is binary in Ghazi et al. (2021)) determine whether this coordinate should be +1 or -1. They argue that any (ε, δ)-semi-sensitive DP algorithm must make an error in determining the sign of $\Omega(1)$ fraction of the coordinates; this results in the $\Omega\left(\frac{1}{\sqrt{n}}\right)$ error for mean estimation, which can then be converted to a lower bound for convex ERM via standard techniques (Bassily et al., 2014). We extend this lower bound by grouping together $O(\log k)$ examples and assign k common coordinates for them. The examples in each group share the same sensitive feature, and it determines which of the k coordinates the examples contribute to. In other words, each group is a hard instance of the so-called selection problem. This helps increase the error to $\Omega\left(\sqrt{\frac{\log k}{n}}\right)$.

2. Preliminaries

We use $\mathcal{B}_2^d(R)$ to denote the Euclidean ball of radius R in d dimensions, i.e., $\{y \in \mathbb{R}^d \mid \|y\|_2 \leq R\}$.

2.1. Differential Privacy

We recall the definition of differential privacy below.

Definition 7 (Differential Privacy, Dwork et al. (2006)) *For $\varepsilon, \delta \geq 0$, a mechanism A is said to be (ε, δ)-differentially private ((ε, δ)-DP) with respect to a certain neighboring relationship iff, for every pair D, D' of neighboring datasets and every set S of outputs, we have $\Pr[\mathrm{A}(D) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathrm{A}(D') \in S] + \delta$.*

In this paper, we consider datasets consisting of examples with a sensitive and non-sensitive part. More precisely, each dataset D is $\{x_i\}_{i\in[n]}$ where $x_i = (x_i^{\text{pub}}, x_i^{\text{priv}}) \in \mathcal{X}^{\text{pub}} \times \mathcal{X}^{\text{priv}}$. Two datasets $D = \{(x_i^{\text{pub}}, x_i^{\text{priv}})\}_{i\in[n]}, D' = \{(x_i'^{\text{pub}}, x_i'^{\text{priv}})\}_{i\in[n]}$ are *neighbors* if they differ on a single example's sensitive part. I.e., $x_i^{\text{pub}} = x_i'^{\text{pub}}$ for all $i \in [n]$ and there exists $i' \in [n]$ such that $x_i^{\text{priv}} = x_i'^{\text{priv}}$ for all $i \in [n] \setminus \{i'\}$. We often use the prefix "semi-sensitive" (e.g., semi-sensitive DP) to signify that we are working with this neighboring relationship notion. Note that, the lemmas below that are stated without such a prefix, hold for any neighboring relationship.

For the purpose of privacy accounting, it will be convenient to work with the zero-concentrated DP (zCDP) notion.

Definition 8 (Dwork and Rothblum (2016); Bun and Steinke (2016)) *For $\rho > 0$, an algorithm A is said to be ρ-*zero concentrated DP (ρ-zCDP) *with respect to a certain neighboring relationship iff, for every pair D, D' of neighboring datasets and every $\alpha > 1$, we have $D_\alpha(\mathrm{A}(D)\|A(D')) \leq \rho \cdot \alpha$, where $D_\alpha(P\|Q)$ denotes the α-Renyi divergence between P and Q.*

We will use the following results from Bun and Steinke (2016) in the privacy analysis.

Lemma 9 (ρ-zCDP vs (ε, δ)-DP) *(i) For any $\varepsilon > 0$, any ε-DP mechanism is $(0.5\varepsilon^2)$-zCDP. (ii) For any $\rho > 0$ and $\delta \in (0, 1/2)$, a ρ-zCDP mechanism is $\left(\rho + 2\sqrt{\rho \ln(1/\delta)}, \delta\right)$-DP.*

Lemma 10 (zCDP composition) *If $\mathcal{M}$ is a mechanism is a (possibly adaptive) composition of mechanisms $\mathcal{M}_1, \ldots, \mathcal{M}_T$, where $\mathcal{M}_i$ is ρ_i-zCDP, then $\mathcal{M}$ is $(\rho_1 + \cdots + \rho_T)$-zCDP.*

2.2. Assumptions on the Loss Function

Throughout this work, we assume that the loss function ℓ is convex and subdifferentiable (in the first parameter). Furthermore, we assume that it is G-Lipschitz; that is, $|\ell(w) - \ell(w')| \leq G \cdot \|w - w'\|_2$.

There are also two additional assumptions that we use in our second result (Theorem 19):

- μ-strong convexity: $\ell(w) \geq \ell(w') + \langle \nabla\ell(w'), w - w' \rangle + \frac{\mu}{2}\|w - w'\|_2^2$.
- λ-smoothness: $\nabla\ell$ is λ-Lipschitz, implying, $\ell(w) \leq \ell(w') + \langle \nabla\ell(w'), w - w' \rangle + \frac{\lambda}{2}\|w - w'\|_2^2$.

2.3. Concentration Bounds

We will now prove a lemma with respect to a "clipped" distribution. To do this, let us define the clipping operation as follows. For $\varphi \in \mathbb{R}^d$ and $c \in \mathbb{R}_{>0}$, we let $\mathrm{clip}_{\varphi,c} : \mathbb{R}^d \to \mathbb{R}^d$ be defined as[2]

$$\mathrm{clip}_{\varphi,c}(u) = \begin{cases} u \cdot \min\{1, c/|\langle \varphi, u \rangle|\} & \text{if } \langle \varphi, u \rangle \neq 0 \\ u & \text{if } \langle \varphi, u \rangle = 0, \end{cases}$$

For convenience, for $c > 0$, we also define $\mathrm{trunc}_c : \mathbb{R} \to \mathbb{R}$ to denote[3] the function $\mathrm{trunc}_c(b) := b \cdot \min\{1, c/|b|\}$, i.e., a rescaling of b so that its absolute value is at most c.

The desired lemma is stated below. Although it might seem overly specific at the moment, we state it in this form as it is most convenient for our usage in the accuracy analysis later (without specifying too many extra parameters). Its proof is deferred to Appendix B.

2. In other words, u is scaled so that its φ-semi-norm is at most c.
3. Note that this coincides with $\mathrm{clip}_{1,c}$ but we keep a separate notation for brevity and clarity.

Lemma 11 *Let $\mathcal{P}$ be any distribution over $\mathcal{B}_2^d(1)$ and $\mu_{\mathcal{P}} := \mathbb{E}_{U\sim\mathcal{P}}[U]$. Let Z be drawn from $\mathcal{N}(\mu_Z, \sigma_Z^2 I_d)$ for some $\sigma_Z \in (0,1], \mu_Z \in \mathcal{B}_2^d(2)$. Then, we have*

$$\Pr_Z\left[\left|\left\langle Z, \mathbb{E}_{U\sim\mathcal{P}}[\mathrm{clip}_{Z,3}(U)] - \mu_{\mathcal{P}}\right\rangle\right| > 2\exp(-0.1/\sigma_Z^2)\right] < 2\exp(-0.1/\sigma_Z^2).$$

3. Answering Linear Vector Queries with Semi-Sensitive DP

As mentioned in the introduction, we consider a setting similar to Hardt and Rothblum (2010) but with two main changes: (i) we support semi-sensitive DP and (ii) each query in the family is allowed to be vector-valued (instead of scalar-valued). We describe this setting in more detail below.

A (bounded ℓ_2-norm) *linear vector query* is a function $f : \mathcal{X} \to \mathcal{B}_2^d(1)$, where $d \in \mathbb{N}$. The value of the function on a dataset $D = \{x_i\}_{i\in[n]}$ is defined as $f(D) := \frac{1}{n}\sum_{i\in[n]} f(x_i)$.

Online Linear Vector Query problem. In the *Online Linear Vector Query (OLVQ)* problem, the interaction proceeds in T rounds. At the beginning, the algorithm receives the dataset D as the input. In round t, the analyzer (aka adversary) selects some linear vector query $f_t : \mathcal{X} \to \mathcal{B}_2^{d_t}(1)$. The algorithm has to output an estimate e_t of $f_t(D)$. We say that the algorithm is (α,β)-*accurate* if, with probability $1-\beta$, $\|e_t - f_t(D)\|_2 \le \alpha$ for all $t \in [T]$. Finally, we say that the algorithm satisfies (ε,δ)-semi-sensitive DP iff the transcript of the interaction satisfies (ε,δ)-semi-sensitive DP.

Our Algorithm. The rest of this section is devoted to presenting (and analyzing) our semi-sensitive DP algorithm for OLVQ. The guarantee of the algorithm is stated formally below.

Theorem 12 *For all $\delta,\beta \in (0,1/2)$ and $\varepsilon \in (0, \sqrt{\ln(1/\delta)})$, there is an (ε,δ)-semi-sensitive DP algorithm for OLVQ that is (α,β)-accurate for $\alpha = O\left(\frac{\sqrt[4]{\ln k\cdot\ln(1/\delta)}\cdot\sqrt{\ln(Tn/\beta)+\ln\ln k+\ln\left(\frac{\sqrt{\ln(1/\delta)}}{\varepsilon}\right)}}{\sqrt{\varepsilon n}}\right)$.*

As mentioned earlier, it will be slightly more convenient to work with the zCDP definition instead of DP for composition theorems. In zCDP terms, our algorithm gives the following guarantee:

Theorem 13 *For every $\rho \in (0,1), \beta \in (0,1/2)$, there is a ρ-semi-sensitive zCDP algorithm for OLVQ that is (α,β)-accurate for $\alpha = O\left(\frac{\sqrt[4]{\frac{\ln k}{\rho}}\cdot\sqrt{\ln(Tn/\beta)+\ln\ln k+\ln(1/\rho)}}{\sqrt{n}}\right)$.*

Note that Theorem 12 follows from Theorem 13 by setting $\rho = \frac{0.1\varepsilon^2}{\log(1/\delta)}$ and applying Lemma 9(ii). The presentation below follows that of Dwork and Roth (2014, Section 4.2) which is based on the original paper of Hardt and Rothblum (2010) and the subsequent work of Gupta et al. (2012). We use the presentation from the Dwork and Roth's book as it uses a more modern privacy analysis through the sparse vector technique, whereas Hardt and Rothblum (2010); Gupta et al. (2012) use a more direct privacy analysis.

3.1. Linear Vector Query Multiplicative Update

First, we present the analysis of the multiplicative weight update (MWU) step for linear vector query. This generalizes the standard analysis for scalar-valued query to a vector-valued one. Note that this subsection does not contain any privacy statements, as those will be handled later.

The algorithm takes as input a "synthetic" (belief) distribution of the sensitive features for each of the n examples. We write p_i^ℓ to denote the distribution for x_i^{priv}. Furthermore, we write $p_i^\ell(y)$ to denote the probability that $x_i^{\text{priv}} = y$ under p_i^ℓ. For $\mathbf{p}^\ell = (p_i^\ell)_{i\in[n]}$ and a linear vector query f, we write $f(\mathbf{p}^\ell; D)$ as a shorthand for $\frac{1}{n}\sum_{i\in[n]}\sum_{y\in\mathcal{X}^{\text{priv}}} p_i^\ell(y)\cdot f(x_i^{\text{pub}}, y)$. We may drop D for brevity when it is clear from the context. The update is based on the difference between the estimated value (which will be set as a noised version of the true answer $f(D)$) and $f(\mathbf{p}^\ell)$. Since the noise can have unbounded value, we "truncate" the dot product when using it to simplify the analysis (recall the notion trunc_c from Section 2.3). The full update is stated in Algorithm 1.

Algorithm 1 $\text{MWU}_{\eta,c}(\mathbf{p}^{\ell-1}, f, v, \iota; D)$: MULTIPLICATIVE WEIGHT UPDATE (MWU) RULE

Input: Dataset $D = \{x_i\}_{i\in[n]}$, $\mathbf{p}^{\ell-1} = (p_i^{\ell-1})_{i\in[n]}$, a linear vector query f, estimated value v, norm bound $\iota > 0$

Parameters: Learning rate $\eta > 0$ and truncation bound c.

$\overline{\varphi} \leftarrow v - f(\mathbf{p}^{\ell-1})$ ▷ Difference between evaluated and estimated value

$\varphi \leftarrow \overline{\varphi}/\iota$

for $i \in [n]$ **do**

  for $y \in \mathcal{X}^{\text{priv}}$ **do**

    $p_i^\ell(y) \leftarrow \frac{p_i^{\ell-1}(y)\cdot\exp\left(\eta\cdot\text{trunc}_c\left(\left\langle\varphi, f(x_i^{\text{pub}}, y)\right\rangle\right)\right)}{\sum_{y'\in\mathcal{X}^{\text{priv}}} p_i^{\ell-1}(y')\cdot\exp\left(\eta\cdot\text{trunc}_c\left(\left\langle\varphi, f(x_i^{\text{pub}}, y')\right\rangle\right)\right)}$ ▷ Multiplicative Weight Update

  end

end

return $\mathbf{p}^\ell = (p_i^\ell)_{i\in[n]}$

We now analyze this update rule. To do so, recall the notion clip from Section 2.3; it will be convenient to also define the following additional notation:

$$f^{\text{clip},\varphi,c}(x^{\text{pub}}, y) := \text{clip}_{\varphi,c}(f(x^{\text{pub}}, y)), \qquad f^{\text{clip},\varphi,c}(D) := \frac{1}{n}\sum_{i\in[n]} f^{\text{clip},\varphi,c}(x_i),$$

$$f^{\text{clip},\varphi,c}(\mathbf{p}^\ell; D) := \frac{1}{n}\sum_{i\in[n]}\sum_{y\in\mathcal{X}^{\text{priv}}} p_i^\ell(y)\cdot f^{\text{clip},\varphi,c}(x_i^{\text{pub}}, y).$$

For readability, we sometimes drop φ and c from the notations above when it is clear from context.

For convenience, we separate the requirement for the MWU analysis into the following condition. The first item states that the error is sufficiently large, the second that the noise added to v is sufficiently small, the next two assert that clipping does not change the function value too much (for the true answer and that evaluated from the synthetic data $\mathbf{p}^{\ell-1}$, respectively), and the remaining two state that ι is a good estimate for $\|f(D) - f(\mathbf{p}^{\ell-1})\|_2$.

Condition 14 *Suppose that $\eta \le \frac{1}{c}$ and the following hold:*

(i) $\|f(D) - f(\mathbf{p}^{\ell-1})\|_2 \ge (2c^2+7)\eta$,

(ii) $\left\langle f(D) - v, f(\mathbf{p}^{\ell-1}) - f(D)\right\rangle \le \eta\cdot\|f(D) - f(\mathbf{p}^{\ell-1})\|_2$,

(iii) $\left|\left\langle v - f(\mathbf{p}^{\ell-1}), f^{\text{clip}}(D) - f(D)\right\rangle\right| \le \eta^2$,

(iv) $\left|\left\langle v - f(\mathbf{p}^{\ell-1}), f^{\text{clip}}(\mathbf{p}^{\ell-1}) - f(\mathbf{p}^{\ell-1})\right\rangle\right| \le \eta^2$,

(v) $\iota \ge \eta$,

(vi) $\iota \le 2 \cdot \|f(D) - f(\mathbf{p}^{\ell-1})\|_2$.

Under the above conditions, we show that the update cannot be applied too many times:

Theorem 15 (MWU Utility Analysis) *Suppose that* $\mathrm{MWU}_{\eta,c}(\mathbf{p}^{\ell-1}, f, v, \iota; D)$ *is applied for* $\ell = 1, \dots, L$ *with the initial distribution being the uniform distribution (i.e.,* $p_i^0(y) = \frac{1}{k}$ *for all* $i \in [n]$ *and* $y \in \mathcal{X}^{\mathrm{priv}}$*) such that Condition 14 holds for all* $\ell \in [L]$*. Then, it must be that* $L < \ln k/\eta^2$.

Let the potential be $\Psi^\ell := \frac{1}{n}\sum_{i\in[n]} \ln\left(\frac{1}{p_i^\ell(x_i^{\mathrm{priv}})}\right)$. The main lemma underlying the proof of Theorem 15 is that the potential always decreases under Condition 14, which immediately implies the proof since the potential satisfies $\Psi^0 = \ln k$ and $\Psi^L > 0$.

Lemma 16 *Assuming that Condition 14 holds, then* $\Psi^{\ell-1} - \Psi^\ell \ge \eta^2$.

To prove Lemma 16, we use the following two simple facts.

Fact 17 *(i) For all* $x \in \mathbb{R}$, $1 + x \le \exp(x)$*. (ii) For all* $x \in (-\infty, 1]$, $\exp(x) \le 1 + x + x^2$.

Proof of Lemma 16 From the definition of clip and trunc, we have $\mathrm{trunc}_c\left(\left\langle \varphi, f(x_i^{\mathrm{pub}}, y)\right\rangle\right) = \left\langle \varphi, f^{\mathrm{clip}}(x_i^{\mathrm{pub}}, y)\right\rangle$. In other words, the update rule can be rewritten as

$$p_i^\ell(y) \leftarrow \frac{p_i^{\ell-1}(y) \cdot \exp\left(\eta \cdot \left\langle \varphi, f^{\mathrm{clip}}(x_i^{\mathrm{pub}}, y)\right\rangle\right)}{\sum_{y'\in\mathcal{X}^{\mathrm{priv}}} p_i^{\ell-1}(y') \cdot \exp\left(\eta \cdot \left\langle \varphi, f^{\mathrm{clip}}(x_i^{\mathrm{pub}}, y')\right\rangle\right)}.$$

For brevity, let γ_i^ℓ be the normalization factor $\sum_{y'\in\mathcal{X}^{\mathrm{priv}}} p_i^{\ell-1}(y') \cdot \exp\left(\eta \cdot \left\langle \varphi, f^{\mathrm{clip}}(x_i^{\mathrm{pub}}, y')\right\rangle\right)$ for all $i \in [n]$. We have

$$\begin{aligned}\Psi^{\ell-1} - \Psi^\ell &= \frac{1}{n}\sum_{i\in[n]} \ln\left(\frac{p_i^\ell(x_i^{\mathrm{priv}})}{p_i^{\ell-1}(x_i^{\mathrm{priv}})}\right) = \frac{1}{n}\sum_{i\in[n]}\left(\eta \cdot \left\langle \varphi, f^{\mathrm{clip}}(x_i)\right\rangle - \ln\gamma_i^\ell\right)\\ &= \eta \cdot \left\langle \varphi, f^{\mathrm{clip}}(D)\right\rangle - \frac{1}{n}\sum_{i\in[n]} \ln\gamma_i^\ell. \qquad (1)\end{aligned}$$

By definition, $\left\langle \varphi, f^{\mathrm{clip}}(x_i, y')\right\rangle \le c$. Thus, by our assumption that $\eta \le 1/c$, we can bound the normalization factor γ_i^ℓ as follows:

$$\begin{aligned}\gamma_i^\ell &= \sum_{y'\in\mathcal{X}^{\mathrm{priv}}} p_i^{\ell-1}(y') \cdot \exp\left(\eta \cdot \left\langle \varphi, f^{\mathrm{clip}}(x_i, y')\right\rangle\right)\\ \text{(Fact 17(ii))} &\le \sum_{y'\in\mathcal{X}^{\mathrm{priv}}} p_i^{\ell-1}(y')\left(1 + (\eta \cdot \left\langle \varphi, f^{\mathrm{clip}}(x_i, y')\right\rangle) + \left(\eta \cdot \left\langle \varphi, f^{\mathrm{clip}}(x_i, y')\right\rangle\right)^2\right)\\ &\le \sum_{y'\in\mathcal{X}^{\mathrm{priv}}} p_i^{\ell-1}(y')\left(1 + (\eta \cdot \left\langle \varphi, f^{\mathrm{clip}}(x_i, y')\right\rangle) + c^2\eta^2\right)\end{aligned}$$

$$= 1 + \eta \cdot \left\langle \varphi, \sum_{y' \in \mathcal{X}^{\text{priv}}} p_i^{\ell-1}(y') \cdot f^{\text{clip}}(x_i, y') \right\rangle + c^2\eta^2.$$

Applying Fact 17(i), we can then conclude that

$$\ln \gamma_i^\ell \le \eta \cdot \left\langle \varphi, \sum_{y' \in \mathcal{X}^{\text{priv}}} p_i^{\ell-1}(y') \cdot f^{\text{clip}}(x_i, y') \right\rangle + c^2\eta^2.$$

Taking the average over all $i \in [n]$, we thus have

$$\frac{1}{n} \sum_{i \in [n]} \ln \gamma_i^\ell \le \eta \cdot \left\langle \varphi, f^{\text{clip}}(\mathbf{p}^{\ell-1}) \right\rangle + c^2\eta^2.$$

Plugging this back into Equation (1), we get

$$\begin{aligned}
&\Psi^{\ell-1} - \Psi^\ell \\
&\ge \eta \cdot \left\langle \varphi, f^{\text{clip}}(D) - f^{\text{clip}}(\mathbf{p}^{\ell-1}) \right\rangle - c^2\eta^2 \\
&= \frac{\eta}{\iota} \cdot \left(\left\langle \overline{\varphi}, f^{\text{clip}}(D) - f(D) \right\rangle + \left\langle \overline{\varphi}, f(\mathbf{p}^{\ell-1}) - f^{\text{clip}}(\mathbf{p}^{\ell-1}) \right\rangle + \left\langle \overline{\varphi}, f(D) - f(\mathbf{p}^{\ell-1}) \right\rangle \right) - c^2\eta^2 \\
&\overset{(\spadesuit)}{\ge} \frac{\eta}{\iota} \cdot \left\langle \overline{\varphi}, f(D) - f(\mathbf{p}^{\ell-1}) \right\rangle - (c^2+2)\eta^2 \\
&= \frac{\eta}{\iota} \cdot \left(\|f(D) - f(\mathbf{p}^{\ell-1})\|_2^2 - \left\langle f(D) - v, f(D) - f(\mathbf{p}^{\ell-1}) \right\rangle \right) - (c^2+2)\eta^2 \\
&\overset{(\blacksquare)}{\ge} \frac{\eta}{\iota} \cdot \left(\|f(D) - f(\mathbf{p}^{\ell-1})\|_2^2 - \eta \cdot \|f(D) - f(\mathbf{p}^{\ell-1})\|_2 \right) - (c^2+2)\eta^2 \\
&\overset{(\clubsuit)}{\ge} \frac{\eta}{2 \cdot \|f(D) - f(\mathbf{p}^{\ell-1})\|_2} \cdot \left((2c^2+7)\eta \cdot \|f(D) - f(\mathbf{p}^{\ell-1})\|_2 - \eta \cdot \|f(D) - f(\mathbf{p}^{\ell-1})\|_2 \right) \\
&\quad - (c^2+2)\eta^2 \\
&= \eta^2,
\end{aligned}$$

where ($\spadesuit$) follows from Condition 14(iii),(iv) and (v), ($\blacksquare$) follows from Condition 14(ii), and ($\clubsuit$) follows from Condition 14(i) and (vi). ■

3.2. The Algorithm

We are now ready to describe our algorithm and prove Theorem 13.

Proof of Theorem 13 Algorithm 2 contains the description of our algorithm.

Privacy Analysis. For each fixed ℓ, the mechanism is exactly a composition of the AboveThreshold mechanism[4] (Dwork and Roth, 2014, Algorithm 1), the Laplace mechanism with noise multiplier[5] $\frac{1}{\varepsilon'}$ and the Gaussian mechanism with noise multiplier[6] σ. The first is ε'-semi-sensitive DP

4. See Appendix A for more explanation on the AboveThreshold, Laplace, and Gaussian mechanisms.
5. The sensitivity of $\|f_t(\mathbf{p}^{\ell-1}) - f_t(D)\|_2$ with respect to semi-sensitive DP is $\frac{2}{n}$.
6. The ℓ_2-sensitivity of $f(D)$ with respect to semi-sensitive DP is $\frac{2}{n}$.

Algorithm 2 PRIVATE VECTOR MULTIPLICATIVE WEIGHT (PVMW)

Input: Dataset D, (online) stream of linear vector queries $f_1, \ldots, f_T$
Parameters: Privacy parameter $\rho > 0$, target accuracy τ, maximum number of MWU applications $L_{\max}$, truncation bound c, budget split parameter $\zeta \in (0, 1)$.

$\sigma \leftarrow \sqrt{\frac{2L_{\max}}{(1-\zeta)\rho}}$ ▷ Gaussian Noise Multiplier
$\varepsilon' \leftarrow \sqrt{\frac{\zeta\rho}{L_{\max}}}$ ▷ Privacy parameter for AboveThreshold and Norm Esimation
for $i = 1, \ldots, n$ **do**
  $p_i^0 \leftarrow$ uniform distribution over $\mathcal{X}^{\text{priv}}$ ▷ Initial distribution
end
$\ell \leftarrow 1$ ▷ Counter for # of updates performed
Sample $\chi_\ell \sim \text{Lap}\left(\frac{4}{\varepsilon' n}\right)$ ▷ Threshold noise for AboveThreshold
for $t = 1, \ldots, T$ **do**
  while $\ell < L_{\max}$ **do**
    Sample $\nu_{t,\ell} \sim \text{Lap}\left(\frac{8}{\varepsilon' n}\right)$ ▷ Query noise for AboveThreshold
    if $\|f_t(\mathbf{p}^{\ell-1}) - f_t(D)\|_2 + \nu_{t,\ell} \geq \tau + \chi_\ell$ **then**
      Sample $z^{\ell-1} \sim \mathcal{N}(0, (2\sigma/n)^2 I_d)$
      $v^{\ell-1} \leftarrow f_t(D) + z^{\ell-1}$
      Sample $\xi^\ell \sim \text{Lap}\left(\frac{2}{\varepsilon' n}\right)$
      $\iota^{\ell-1} \leftarrow \|f_t(\mathbf{p}^{\ell-1}) - f_t(D)\|_2 + \xi^{\ell-1}$ ▷ Difference Norm Estimation
      $\mathbf{p}^\ell \leftarrow \text{MWU}_{\eta,c}(\mathbf{p}^{\ell-1}, f_t, v^{\ell-1}, \iota^{\ell-1}; D)$ ▷ Multiplicative Weight Update
      $\ell \leftarrow \ell + 1$
      Sample $\chi_\ell \sim \text{Lap}\left(\frac{4}{\varepsilon' n}\right)$ ▷ Resample threshold noise
    end
    else
      Break ▷ Below threshold; estimated value is accurate enough
    end
  end
  if $\ell \geq L_{\max}$ **then**
    Halt and return "FAIL"
  end
  else
    return $f_t(\mathbf{p}^{\ell-1})$ ▷ Output estimate of $f_t(D)$
  end
end

(Dwork and Roth, 2014, Theorem 3.23) and, by Lemma 9(i) is thus $(0.5\varepsilon'^2)$-semi-sensitive zCDP; similarly, the Laplace mechanism is $(0.5\varepsilon'^2)$-semi-sensitive zCDP. Meanwhile, the Gaussian mechanism is $(2/\sigma^2)$-zCDP (Bun and Steinke, 2016). Thus, by the composition theorem (Lemma 10) for a fixed ℓ, the mechanism is $(0.5\varepsilon'^2) + (0.5\varepsilon'^2) + (2/\sigma^2) = (\rho/L_{\max})$-semi-sensitive zCDP. Thus, applying the composition theorem (Lemma 10) across all $L_{\max}$ iterations, the entire algorithm is ρ-semi-sensitive zCDP.

Utility Analysis. We set the parameters as follows: (i) $\zeta = \frac{1}{2}$, (ii) $c = 3$, (iii) η to be the smallest positive real number such that $\eta > \frac{1000\sqrt[4]{\frac{\ln k}{\rho}} \cdot \sqrt{\ln\left(\frac{nT\ln k}{\rho\beta\eta}\right)}}{\sqrt{n}}$, (iv) $\tau = 16\eta$, (v) $L_{\max} = 1 + \left\lfloor \frac{\ln k}{\eta^2} \right\rfloor$.

Note that we may assume w.l.o.g. that $\eta \leq 0.1$ as otherwise the desired guarantee is trivial (i.e., the algorithm can simply outputs zero always).

By the tail bound of Laplace noise (Lemma 31(i)), for a fixed $\ell \in [L_{\max}]$, the following holds with the probability at least $1 - \frac{0.1\beta}{L_{\max}}$:

$$|\chi_\ell| \le \frac{4}{\varepsilon' n} \cdot \ln\left(\frac{20 L_{\max}}{\beta}\right) \le \eta, \tag{2}$$

where the second inequality follows from our setting of parameters.

Similarly, for fixed $\ell \in [L_{\max}], t \in [T]$, the following holds with probability at least $1 - \frac{0.1\beta}{L_{\max} T}$:

$$|\nu_{t,\ell}| \le \frac{8}{\varepsilon' n} \cdot \ln\left(\frac{20 L_{\max} T}{\beta}\right) \le \eta, \tag{3}$$

and, for a fixed $\ell \in [L_{\max}]$, the following holds with probability at least $1 - \frac{0.1\beta}{L_{\max} T}$:

$$|\xi^{\ell-1}| \le \frac{4}{\varepsilon' n} \cdot \ln\left(\frac{20 L_{\max}}{\beta}\right) \le \eta. \tag{4}$$

Observe that, for a given $\mathbf{p}^{\ell-1}$, $\left\langle z^{\ell-1}, f_t(\mathbf{p}^{\ell-1}) - f_t(D)\right\rangle$ is distributed as $\mathcal{N}(0, (\sigma')^2)$ for $\sigma' = \frac{2\sigma}{n} \cdot \|f_t(\mathbf{p}^{\ell-1}) - f_t(D)\|_2$. Thus, by the Gaussian tail bound (Lemma 31(ii)) and our setting of parameters, the following holds with probability $1 - \frac{0.2\beta}{L_{\max} T}$ for fixed $\ell \in [L_{\max}], t \in [T]$:

$$\left\langle z^{\ell-1}, f_t(\mathbf{p}^{\ell-1}) - f_t(D)\right\rangle \le \sigma' \cdot \sqrt{2 \ln\left(\frac{10 L_{\max} T}{\beta}\right)} \le \eta \cdot \|f_t(\mathbf{p}^{\ell-1}) - f_t(D)\|_2. \tag{5}$$

Observe also that $v^{\ell-1} - f_t(\mathbf{p}^{\ell-1}) = f_t(D) - f_t(\mathbf{p}^{\ell-1}) + z^{\ell-1}$ is distributed as $\mathcal{N}(f_t(D) - f_t(\mathbf{p}^{\ell-1}), (\sigma'')^2 I_d)$ where $\sigma'' = 2\sigma/n$. By our choice of parameters, we have $\sigma'' \le 0.1/\log\left(\frac{10 L_{\max}}{\beta\eta}\right)$ As a result, we can apply Lemma 11 with $Z = v^{\ell-1} - f_t(\mathbf{p}^{\ell-1})$ and $\mathcal{P}$ being the uniform distribution over $\{f_t(x_1), \ldots, f_t(x_n)\}$. This allows us to conclude that the following holds with probability at least $1 - \frac{0.2\beta}{L_{\max}}$ for every $\ell \in [L_{\max}]$:

$$\left|\left\langle v^{\ell-1} - f_t(\mathbf{p}^{\ell-1}), f_t^{\text{clip}}(D) - f_t(D)\right\rangle\right| \le 2\exp\left(-\frac{0.1}{(\sigma'')^2}\right) \le \eta^2. \tag{6}$$

Similarly, we can apply Lemma 11 with the same Z but with $\mathcal{P}$ being the distribution where each (x_i^{pub}, y') has probability mass $\frac{p_i^{\ell-1}(y')}{n}$ to conclude that the following holds with probability at least $1 - \frac{0.2\beta}{L_{\max}}$ for every $\ell \in [L_{\max}]$:

$$\left|\left\langle v^{\ell-1} - f_t(\mathbf{p}^{\ell-1}), f_t^{\text{clip}}(\mathbf{p}^{\ell-1}) - f_t(\mathbf{p}^{\ell-1})\right\rangle\right| \le 2\exp\left(-\frac{0.1}{(\sigma')^2}\right) \le \eta^2. \tag{7}$$

By a union bound, all of eqs. (2) to (7) hold for all $\ell \in [L_{\max}], t \in [T]$ with probability at least $1 - \beta$. We assume that these inequalities hold throughout the remainder of the analysis.

From eqs. (2) and (3), if we break the loop, we must have

$$\|f_t(\mathbf{p}^{\ell-1}) - f_t(D)\|_2 < \tau + \chi_\ell - \nu_{t,\ell} \le \tau + 2\eta \le 18\eta.$$

This means that whenever the algorithm outputs an estimate, it has ℓ_2-error of at most $18\eta \le O\left(\frac{\sqrt[4]{\frac{\ln k}{\rho}}\cdot\sqrt{\ln(Tn/\beta)+\ln\ln k+\ln(1/\rho)}}{\sqrt{n}}\right)$ as desired. As a result, it suffices to show that the algorithm never outputs "FAIL".

On the other hand, if MWU is called, we must have

$$\|f_t(\mathbf{p}^{\ell-1}) - f_t(D)\|_2 \ge \tau + \chi_\ell - \nu_{t,\ell} \ge \tau - 2\eta \ge 14\eta, \tag{8}$$

implying Condition 14(i) (for $v = v^{\ell-1}, f = f_t, c = 3$). Moreover, eqs. (5) to (7) are exactly equivalent to Condition 14(ii)(iii)(iv) respectively. Furthermore, by eqs. (4) and (8), we also have

$$\iota^{\ell-1} = \|f_t(\mathbf{p}^{\ell-1}) - f_t(D)\|_2 + \xi^{\ell-1} \ge 14\eta - \eta = 13\eta,$$

and

$$\iota^{\ell-1} = \|f_t(\mathbf{p}^{\ell-1}) - f_t(D)\|_2 + \xi^{\ell-1} \le \|f_t(\mathbf{p}^{\ell-1}) - f_t(D)\|_2 + \eta < 2\cdot\|f_t(\mathbf{p}^{\ell-1}) - f_t(D)\|_2,$$

which mean that Condition 14(v)(vi) hold, respectively. In other words, Condition 14 holds.

From this, we may apply Theorem 15 to conclude that the number of applications of MWU is less than $\frac{\ln k}{\eta^2} \le L_{\max}$. Thus, the algorithm never outputs "FAIL". This completes the proof of the accuracy guarantee. ■

4. From Online Linear Vector Queries to Convex Optimization

In this section, we prove our main results for convex optimization with semi-sensitive DP, as formalized below. We note that here we formulate it as the problem of solving m linear queries problems w.r.t. losses $\ell_1, \dots, \ell_m$. The expected excess risk guarantee is for all of these m problems[7].

Theorem 18 *Suppose that the loss functions $\ell_1, \dots, \ell_m$ are G-Lipschitz and $\mathcal{W}_1, \dots, \mathcal{W}_m \subseteq \mathcal{B}_2(R)$. For every $\delta \in (0, 1/2)$ and $\varepsilon \in (0, \ln(1/\delta))$, there is an (ε, δ)-semi-sensitive DP algorithm for ERM problems w.r.t. $\ell_1, \dots, \ell_m$ with expected excess risk*

$$O\left(RG\cdot\frac{\sqrt[4]{\ln k\cdot\ln(1/\delta)}\cdot\sqrt{\ln(mn)+\ln\ln k+\ln\left(\frac{\sqrt{\ln(1/\delta)}}{\varepsilon}\right)}}{\sqrt{\varepsilon n}}\right).$$

Theorem 19 *Suppose that the loss functions $\ell_1, \dots, \ell_m$ are G-Lipschitz, μ-strongly convex, and λ-smooth. For every $\delta \in (0, 1/2)$ and $\varepsilon \in (0, \ln(1/\delta))$, there is an (ε, δ)-semi-sensitive DP algorithm for ERM problems w.r.t. $\ell_1, \dots, \ell_m$ with expected excess risk*

$$O\left(\frac{G^2}{\mu}\cdot\frac{\sqrt{\ln k\cdot\ln(1/\delta)}\cdot\left(\ln(mn\lambda/\mu)+\ln\ln(G/\mu)+\ln\ln k+\ln\left(\frac{\sqrt{\ln(1/\delta)}}{\varepsilon}\right)\right)}{\varepsilon n}\right).$$

As stated in the Introduction, these results are shown via simple applications of known optimization algorithms with approximate gradients. The two cases use slightly different notion of approximate gradients, which we will explain below.

7. We say that the expected excess risk is e if, for all $i \in [m]$, we have $\mathbb{E}[\mathcal{L}_i(w_i; D) - \min_{w' \in \mathcal{W}_i} \mathcal{L}_i(w'; D)] \le e$ for all $i \in [m]$ where $\mathcal{L}_i$ denotes the empirical risk and w_i denotes the output of the algorithm for the ith instance.

4.1. Convex Case via Approximate Gradient Oracle

For the convex case, we use the following definition of approximate gradient oracle.

Definition 20 (Approximate Gradient Oracle, d'Aspremont (2008)) *For any convex function $F : \mathcal{W} \to \mathbb{R}$, an* ξ-approximate gradient oracle *of F provides $\tilde{g}(w)$ for any queried $w \in \mathcal{W}$ such that the following holds: $|\langle \tilde{g}(w) - \nabla F(w), y - u\rangle| \le \xi$ for all $y, u \in \mathcal{W}$.*

Under the above condition, standard gradient descent achieves a similar excess risk to the exact gradient case except that there is an extra ξ term:

Theorem 21 (Feldman et al. (2017, Theorem 4.5)) *For any G-Lipschitz convex function $F : \mathcal{W} \to \mathbb{R}$ with $\mathcal{W} \subseteq \mathcal{B}_2(D)$ and any $q \in \mathbb{N}$, there exists an algorithm that makes q queries to an ξ-approximate gradient oracle and achieves an excess risk of $O\left(\frac{DG}{\sqrt{q}} + \xi\right)$.*

We are now ready to prove Theorem 18.

Proof of Theorem 18 Let $q = n^2$ and $T = m \cdot q$. We simply run the algorithm from Theorem 21 for each $i \in [m]$ and, for the tth query to approximate gradient oracle, we invoke algorithm from Theorem 12 with $f_{q(i-1)+t}(x) = \frac{1}{G}\nabla \ell_i(w; x)$ and scale the answer back by a factor of G. From Theorem 12, with probability $1 - \beta$, this is an $(2RG \cdot \alpha)$-approximate gradient oracle for $\alpha = O\left(\frac{\sqrt[4]{\ln k \cdot \ln(1/\delta)} \cdot \sqrt{\ln(Tn/\beta) + \ln\ln k + \ln\left(\frac{\sqrt{\ln(1/\delta)}}{\varepsilon}\right)}}{\sqrt{\varepsilon n}}\right)$. When this occurs, Theorem 21 implies that the excess risk is at most $\frac{RG}{\sqrt{q}} + 2RG \cdot \alpha \le O(RG \cdot \alpha)$. With the remaining probability β, the excess risk is still at most RG. Substituting $\beta = 1/n$, we can conclude that the expected excess risk of this algorithm is at most $O\left(\frac{1}{n} \cdot RG + RG \cdot \alpha\right) \le O(RG \cdot \alpha)$. Finally, since the algorithm is simply a post-processing of the result from applying Theorem 12, it is (ε, δ)-semi-sensitive DP. ■

4.2. Strongly Convex and Smooth Case via Inexact Oracle

For the strongly convex and smooth case, we use the following notion called inexact oracle.

Definition 22 (Inexact Oracle, Devolder et al. (2014, 2013)) *For any convex function $F : \mathcal{W} \to \mathbb{R}$, a* first-order $(\upsilon, \tilde{\lambda}, \tilde{\mu})$-inexact oracle *of F provides $(\tilde{F}(w), \tilde{g}(w))$ for any queried $w \in \mathcal{W}$ such that the following holds for all $w, w' \in \mathcal{W}$:*

$$\frac{\tilde{\mu}}{2} \cdot \|w' - w\|_2^2 \le F(w') - \left(\tilde{F}(w) - \left\langle \tilde{g}(w), w' - w\right\rangle\right) \le \frac{\tilde{\lambda}}{2} \cdot \|w' - w\|_2^2 + \upsilon.$$

Note that if the gradient and function values are exact (i.e., $\tilde{F} = F, \tilde{g} = \nabla$), then the above condition holds for $\upsilon = 0$ when the function F is $\tilde{\mu}$-strongly convex and $\tilde{\lambda}$-smooth.

We use the following relation between the ℓ_2-error of the gradient estimate and inexact oracle.

Lemma 23 (Devolder et al. (2013, Section 2.3)) *For any μ-strongly convex and λ-smooth $F : \mathcal{W} \to \mathbb{R}$, if $\tilde{g} : \mathcal{W} \to \mathbb{R}^d$ is an oracle such that $\|\tilde{g}(w) - \nabla g(w)\|_2 \le \xi$ for all $w \in \mathcal{W}$, then there exists $\tilde{F} : \mathcal{W} \to \mathbb{R}$ such that $(\tilde{F}, \tilde{g})$ is an $\left(\xi^2\left(\frac{1}{\mu} + \frac{1}{2\lambda}\right), 2\lambda, \mu/2\right)$-inexact oracle.*

It should be noted that we do not specify the exact $\tilde{F}$ precisely because the optimization algorithm we use does not need this either:

Theorem 24 (Feldman et al. (2017, Theorem 4.11)) *For any G-Lipschitz convex function $F : \mathcal{W} \to \mathbb{R}$ with $\mathcal{W} \subseteq \mathcal{B}_2(D)$ and any $q \in \mathbb{N}, \alpha > 0$, there exists an algorithm that makes q queries to a first-order $(\upsilon, \tilde{\lambda}, \tilde{\mu})$-inexact oracle and achieves an excess risk of $O\left(\frac{\tilde{\lambda}R^2}{2} \cdot \exp\left(-\frac{\tilde{\mu}}{\tilde{\lambda}} \cdot q\right) + \upsilon\right)$ where R denote the distance of the starting point to the optimum. Furthermore, the algorithm only uses the gradient estimate $\tilde{g}$ and does not use the function estimate $\tilde{F}$.*

The proof of Theorem 19 is almost the same as that of Theorem 18 except that we now use Theorem 24 (and Lemma 23) instead of Theorem 21.

Proof of Theorem 19 Note that from Lipschitzness and strong convexity, we can assume that the domain $\mathcal{W}_i$ is contained in $\mathcal{B}_2(R)$ for $R = O(G/\mu)$ for all $i \in [m]$. Let $q = \left\lceil \frac{40\lambda}{\mu} \cdot \ln\left(\frac{\lambda R^2}{n}\right) \right\rceil$ and $T = m \cdot q$. We simply run the algorithm from Theorem 24 for each $i \in [m]$, and for the tth query to approximate gradient oracle, we invoke the algorithm from Theorem 12 with $f_{q(i-1)+t}(x) = \frac{1}{G}\nabla \ell_i(w; x)$ and scale the answer back by a factor of G. From Theorem 12, with probability $1 - \beta$, this oracle has ℓ_2-error at most $G \cdot \alpha$ for $\alpha = O\left(\frac{\sqrt[4]{\ln k \cdot \ln(1/\delta)} \cdot \sqrt{\ln(Tn/\beta) + \ln\ln k + \ln\left(\frac{\sqrt{\ln(1/\delta)}}{\varepsilon}\right)}}{\sqrt{\varepsilon n}}\right)$. By Lemma 23, this[8] yields an $(\upsilon, 2\lambda, \mu/2)$-inexact oracle for $\upsilon = O\left((G\alpha)^2 \cdot \left(\frac{1}{\lambda} + \frac{1}{\mu}\right)\right)$. When this occurs, Theorem 24 implies that the excess risk is at most $O\left(\frac{\tilde{\lambda}R^2}{2} \cdot \exp\left(-\frac{\tilde{\mu}}{\tilde{\lambda}} \cdot T\right) + \upsilon\right) \leq O(\upsilon)$. With the remaining probability β, the excess risk is still at most $DG = O(G^2/\mu)$. Substituting $\beta = 1/n$, we can conclude that the expected excess risk of this algorithm is at most $O\left(\frac{1}{n} \cdot \frac{G^2}{\mu} + \upsilon\right) \leq O(\upsilon)$. Finally, since the algorithm is simply a post-processing of the result from applying Theorem 12, it is (ε, δ)-semi-sensitive DP. ■

5. Conclusion and Open Questions

We gave improved bounds for convex ERM with semi-sensitive DP; crucially they show that the dependency on k is only polylogarithmic instead of polynomial as in previous works. As an intermediate result, we give an algorithm for answering (online) linear vector queries. Given that linear queries are used well beyond convex optimization, we hope that this will find more applications.

An obvious open question is to close the gap between the upper and lower bounds. Another interesting question is to come up with a *pure*-DP algorithm with a similar bound as in Theorem 18. In particular, it is open if there is any pure-DP algorithm where the error depends only polylogarithmically on k.

8. Since the algorithm in Theorem 24 does not use the value from the oracle, we do not need to specify it explicitly.

References

Raef Bassily, Adam D. Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *FOCS*, pages 464–473, 2014.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In *NeurIPS*, pages 11279–11288, 2019.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *TCC*, pages 635–658, 2016.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 12:1069–1109, 2011.

Lynn Chua, Qiliang Cui, Badih Ghazi, Charlie Harrison, Pritish Kamath, Walid Krichene, Ravi Kumar, Pasin Manurangsi, Krishna Giri Narra, Amer Sinha, Avinash V. Varadarajan, and Chiyuan Zhang. Training differentially private ad prediction models with semi-sensitive features. *CoRR*, abs/2401.15246, 2024.

Alexandre d'Aspremont. Smooth optimization with approximate gradient. *SIAM J. Optim.*, 19(3): 1171–1183, 2008.

Olivier Devolder, François Glineur, Yurii Nesterov, et al. First-order methods with inexact oracle: the strongly convex case. *CORE Discussion Papers*, 2013016:47, 2013.

Olivier Devolder, François Glineur, and Yurii E. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Math. Program.*, 146(1-2):37–75, 2014.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

Cynthia Dwork and Guy N. Rothblum. Concentrated differential privacy. *CoRR*, abs/1603.01887, 2016.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.

Cynthia Dwork, Moni Naor, Omer Reingold, Guy N. Rothblum, and Salil P. Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *STOC*, pages 381–390, 2009.

Vitaly Feldman, Cristóbal Guzmán, and Santosh S. Vempala. Statistical query algorithms for mean vector estimation and stochastic convex optimization. In *SODA*, pages 1265–1277, 2017.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *STOC*, pages 439–449, 2020.

Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning with label differential privacy. In *NeurIPS*, pages 27131–27145, 2021.

Badih Ghazi, Pritish Kamath, Ravi Kumar, Ethan Leeman, Pasin Manurangsi, Avinash V. Varadarajan, and Chiyuan Zhang. Regression with label differential privacy. In *ICLR*, 2023.

Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. Private convex optimization via exponential mechanism. In *COLT*, pages 1948–1989, 2022.

Anupam Gupta, Aaron Roth, and Jonathan R. Ullman. Iterative constructions and private data release. In *TCC*, pages 339–356, 2012.

Moritz Hardt and Guy N. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *FOCS*, pages 61–70, 2010.

Daniel Kifer, Adam D. Smith, and Abhradeep Thakurta. Private convex optimization for empirical risk minimization with applications to high-dimensional regression. In *COLT*, pages 25.1–25.40, 2012.

Walid Krichene, Nicolas Mayoraz, Steffen Rendle, Shuang Song, Abhradeep Thakurta, and Li Zhang. Private learning with public features. In *AISTATS*, pages 4150–4158, 2024.

Zeyu Shen, Anilesh K. Krishnaswamy, Janardhan Kulkarni, and Kamesh Munagala. Classification with partially private features. *CoRR*, abs/2312.07583, 2023.

Jonathan R. Ullman. Private multiplicative weights beyond linear queries. In *PODS*, pages 303–312, 2015.

Salil P. Vadhan. The complexity of differential privacy. In Yehuda Lindell, editor, *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer International Publishing, 2017.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *NIPS*, pages 2722–2731, 2017.

Appendix A. Additional Preliminaries

In this section, we give some more background on the DP mechanisms from literature that we use as subroutines. We start with sensitivity, the Gaussian mechanism, and the Laplace mechanism.

Definition 25 (Sensitivity) *For any query $g : \mathcal{X}^n \to \mathbb{R}^d$ and $p \geq 1$, its* ℓ_p-sensitivity *is defined as $\Delta_p(g) := \max_{D,D'} \|g(D) - g(D')\|_p$ where the maximum is over all neighboring datasets D, D'.*

Definition 26 (Gaussian Mechanism) *The* Gaussian mechanism *for a function $g : \mathcal{X}^n \to \mathbb{R}^d$ simply outputs $g(D) + Z$ on input D where $Z \sim \mathcal{N}(0, \sigma^2 I_d)$.*

The zCDP property of Gaussian mechanism is well known:

Theorem 27 (Bun and Steinke (2016)) *The Gaussian mechanism is ρ-zCDP for $\rho = 0.5\Delta_2(g)^2/\sigma^2$.*

Definition 28 (Laplace Mechanism) *The* Laplace mechanism *for a function $g : \mathcal{X}^n \to \mathbb{R}^d$ simply outputs $g(D) + Z$ on input D where $Z \sim \mathrm{Lap}(a)^{\otimes d}$.*

The Laplace mechanism has been shown to be DP in the original work of Dwork et al. (2006).

Theorem 29 (Dwork et al. (2006)) *The Laplace mechanism is ε-DP for $\varepsilon = \Delta_1(g)/a$.*

Another tool we use is the so-called AboveThreshold mechanism, from the Sparse Vector Technique Dwork et al. (2009). This mechanism is shown in Algorithm 3, following the presentation in (Dwork and Roth, 2014, Algorithm 1).

Algorithm 3 ABOVETHRESHOLD

Input: Dataset D, threshold τ, (online) stream of queries $g_1, \ldots, g_T : \mathcal{X}^n \to \mathbb{R}$
Parameters: Privacy parameter ε, sensitivity bound Δ
Sample $\chi \sim \mathrm{Lap}\left(\frac{2\Delta}{\varepsilon}\right)$ ▷ Threshold noise
for $t = 1, \ldots, T$ **do**
  Sample $\nu_t \sim \mathrm{Lap}\left(\frac{4\Delta}{\varepsilon}\right)$ ▷ Query noise
  if $g_t(D) + \nu_t \geq \tau + \chi$ **then**
    Output $\top$ and halt
  end
  else
    Output $\perp$ and continue
  end
end

Despite the fact that we handle multiple queries, AboveThreshold only requires a constant amount of noise and satisfies pure-DP:

Theorem 30 ((Dwork and Roth, 2014, Theorem 3.23)) *Suppose that each of $g_1, \ldots, g_T$ has sensitivity at most Δ. Then,* ABOVETHRESHOLD *(Algorithm 3) is ε-DP.*

Appendix B. Proof of Lemma 11

We will use the following tail bounds for Laplace and Gaussian distributions.

Lemma 31 (Tail Bounds) *(i)* $\Pr_{X\sim\mathrm{Lap}(a)}[|X| \geq t] \leq 2\exp(-t/a)$. *(ii)* $\Pr_{X\sim\mathcal{N}(0,\sigma^2)}[|X| \geq t] \leq 2\exp(-0.5(t/\sigma)^2)$.

Using the above tail bound, it is relatively simple to show Lemma 11.

Proof of Lemma 11 By Markov's inequality, it suffices to show that

$$\mathbb{E}_Z\left[\left|\left\langle Z, \mathbb{E}_{U\sim\mathcal{P}}[\mathrm{clip}_{Z,3}(U)] - \mu_{\mathcal{P}}\right\rangle\right|\right] \leq 4\exp(-0.2/\sigma_Z^2). \tag{9}$$

To show this, first observe that

$$\begin{aligned}
\mathbb{E}_Z\left[\left|\left\langle Z, \mathbb{E}_{U\sim\mathcal{P}}[\mathrm{clip}_{Z,3}(U)] - \mu_{\mathcal{P}}\right\rangle\right|\right] &= \mathbb{E}_Z\left[\left|\mathbb{E}_{U\sim\mathcal{P}}[\left\langle Z, \mathrm{clip}_{Z,3}(U)\right\rangle - \langle Z, U\rangle]\right|\right] \\
&= \mathbb{E}_Z\left[\left|\mathbb{E}_{U\sim\mathcal{P}}\left[\mathrm{trunc}_3(\langle Z, U\rangle) - \langle Z, U\rangle\right]\right|\right] \\
&\leq \mathbb{E}_Z\left[\mathbb{E}_{U\sim\mathcal{P}}\left[\left|\mathrm{trunc}_3(\langle Z, U\rangle) - \langle Z, U\rangle\right|\right]\right] \\
&= \mathbb{E}_{U\sim\mathcal{P}}\left[\mathbb{E}_Z\left[\left|\mathrm{trunc}_3(\langle Z, U\rangle) - \langle Z, U\rangle\right|\right]\right] \\
&\leq \sup_{u\in\mathcal{B}_2^d(1)} \mathbb{E}_Z\left[\left|\mathrm{trunc}_3(\langle Z, u\rangle) - \langle Z, u\rangle\right|\right],
\end{aligned} \tag{10}$$

where the first inequality follows from Jensen.

Let us now fixed $u \in \mathcal{B}_2^d(1)$. Observe that

$$\begin{aligned}
&\mathbb{E}_Z\left[\left|\mathrm{trunc}_3(\langle Z, u\rangle) - \langle Z, u\rangle\right|\right] \leq \mathbb{E}_Z\left[\left|\langle Z, u\rangle\right| \cdot \mathbf{1}[|\left\langle Z, u\right\rangle| > 3]\right] \\
&\leq \sqrt{\mathbb{E}_Z\left[\left\langle Z, u\right\rangle^2\right] \cdot \mathbb{E}_Z\left[\mathbf{1}[|\left\langle Z, u\right\rangle| > 3]\right]} \leq \sqrt{\mathbb{E}_Z\left[\left\langle Z, u\right\rangle^2\right] \cdot \Pr_Z\left[\mathbf{1}[|\left\langle Z, u\right\rangle| > 3]\right]},
\end{aligned} \tag{11}$$

where the second inequality follows from Cauchy–Schwarz.

Notice further that $\langle Z, u\rangle$ is distributed as $\mathcal{N}(\mu', \sigma')$ for $\mu' = \langle \mu_Z, u\rangle$ and $\sigma' = \sigma_Z \cdot \|u\|_2 \leq \sigma_Z$. Moreover, since $\mu_Z \in \mathcal{B}_2^d(2)$, we have $|\mu'| \leq 2$. As a result, its second moment satisfies

$$\mathbb{E}_Z\left[\left\langle Z, u\right\rangle^2\right] = (\mu')^2 + (\sigma')^2 \leq 2 + \sigma_Z^2 \leq 3.$$

Moreover, applying Lemma 31(ii), we can conclude that

$$\Pr_Z[|\left\langle Z, u\right\rangle| > 3] \leq 2\exp\left(-0.5/\sigma_Z^2\right).$$

Plugging these back into (11), we get

$$\mathbb{E}_Z\left[\left|\mathrm{trunc}_3(\langle Z, u\rangle) - \langle Z, u\rangle\right|\right] \leq \sqrt{6\exp\left(-0.5/\sigma_Z^2\right)} \leq 4\exp(-0.2/\sigma_Z^2).$$

From this and (10), we can conclude that (9) holds as desired. ■

Appendix C. Excess Risk Lower Bound

In this section, we prove a nearly-matching lower bound on the excess risk. We will use "group privacy" bound in our lower bound proof. For any neighboring relationship $\sim$, we use $\sim_r$ to denote the relationship where two datasets D, D' are considered neighbors if there exists a sequence $D = D_0, D_1, \ldots, D_r = D'$ such that $D_{i-1} \sim D_i$ for all $i \in [r]$.

Lemma 32 (Group Privacy, e.g., Vadhan (2017, Lemma 2.2)) *Suppose that* A *is* (ε, δ)*-DP w.r.t.* $\sim$*, then it is* (ε', δ')*-DP w.r.t.* $\sim_r$ *for* $\varepsilon' = r\varepsilon$ *and* $\delta' = \frac{e^{r\varepsilon}-1}{e^{\varepsilon}-1} \cdot \delta$.

Our lower bounds are stated formally below.

Theorem 33 *For any* $\varepsilon, \delta, R, G > 0$ *and* $n, k \in \mathbb{N}$ *such that* $\varepsilon \leq \ln k$ *and* $\delta \leq \frac{0.4\varepsilon}{k}$*, there exists a* G*-Lipschitz convex loss function* $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$*, where* $\mathcal{W} \subseteq \mathcal{B}_2(R)$*, such that any* (ε, δ)*-semi-sensitive DP algorithm for ERM w.r.t on* ℓ *has expected excess empirical risk at least* $\Omega\left(DG \cdot \min\left\{1, \frac{\sqrt{\log k}}{\sqrt{\varepsilon n}}\right\}\right)$.

Theorem 34 *For any* $\varepsilon, \delta, G, \mu > 0$ *and* $n, k \in \mathbb{N}$ *such that* $\varepsilon \leq \ln k$*,* $\delta \leq \frac{0.4\varepsilon}{k}$*, there exists a* G*-Lipschitz* μ*-strongly convex* μ*-smooth loss function* $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ *such that any* (ε, δ)*-semi-sensitive DP algorithm for ERM w.r.t* ℓ *has expected excess empirical risk at least* $\Omega\left(\frac{G^2}{\mu} \cdot \min\left\{1, \frac{\log k}{\varepsilon n}\right\}\right)$.

C.1. Convex Case

To show the convex case (Theorem 33), it will in fact be convenient to first prove a (smaller) lower bound that holds even against very large ε (up to $O(\ln k)$), as stated more formally below.

Theorem 35 *For any* $\varepsilon, \delta, R, G > 0$ *and* $n, k \in \mathbb{N}$ *such that* $\frac{e^{\varepsilon}}{e^{\varepsilon}+k-1} + \delta < 0.99$*, there exists a* G*-Lipschitz convex loss function* $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$*, where* $\mathcal{W} \subseteq \mathcal{B}_2(R)$*, such that any* (ε, δ)*-semi-sensitive DP algorithm for ERM w.r.t* ℓ *has expected excess empirical risk at least* $\Omega\left(\frac{RG}{\sqrt{n}}\right)$.

Proof Let $\mathcal{X}^{\mathrm{pub}} = [n]$, $\mathcal{X}^{\mathrm{priv}} = [k]$, $d = nk$, and $\mathcal{W} = \mathcal{B}_2^d(R)$. For $x^{\mathrm{pub}} \in \mathcal{X}^{\mathrm{pub}}, y \in \mathcal{X}^{\mathrm{priv}}$, we write $j(x^{\mathrm{pub}}, y)$ as a shorthand for $k(x^{\mathrm{pub}} - 1) + y$. Finally, let $\ell : \mathcal{W} \times (\mathcal{X}^{\mathrm{pub}} \times \mathcal{X}^{\mathrm{priv}}) \to \mathbb{R}$ be

$$\ell(w, (x^{\mathrm{pub}}, y)) = -G \cdot \left\langle w, e_{j(x^{\mathrm{pub}}, y)} \right\rangle,$$

where $e_j \in \mathbb{R}^d$ denotes the jth vector in the standard basis for all $j \in [d]$.

Let $D = \{x_i\}_{i \in [n]}$ be the input dataset generated as follows:

- Sample $y_1, \ldots, y_n \sim [k]$ independently and uniformly at random and
- Let $x_i = (i, y_i)$ for all $i \in [n]$.

Let $\mathrm{A} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{W}$ denote any (ε, δ)-semi-sensitive DP algorithm.

For $i \in [n]$, we write $w(i)$ as a shorthand for $(w_{(i-1)k+1}, \ldots, w_{ik})$. We have[9]

$$\Pr_{D, \hat{w} \sim \mathrm{A}(D)}\left[\langle \hat{w}(i), e_{y_i} \rangle > \frac{1}{\sqrt{2}} \|\hat{w}(i)\|_2\right]$$

9. Here ties can be broken arbitrarily for argmax.

$$
\begin{aligned}
&\leq \Pr_{D,\hat{w}\sim \mathrm{A}(D)}\left[y_i = \operatorname{argmax}_{y\in[k]} \langle \hat{w}(i), e_y\rangle\right] \\
&\overset{(\blacksquare)}{=} \frac{1}{k}\sum_{y'\in[k]} \Pr_{D,\hat{w}\sim \mathrm{A}(D)}\left[y' = \operatorname{argmax}_{y\in[k]} \langle \hat{w}(i), e_y\rangle \mid y_i = y'\right] \\
&= \frac{1}{k}\sum_{y'\in[k]} \left(\frac{e^\varepsilon}{e^\varepsilon + k - 1}\cdot \Pr_{D,\hat{w}\sim \mathrm{A}(D)}\left[y' = \operatorname{argmax}_{y\in[k]} \langle \hat{w}(i), e_y\rangle \mid y_i = y'\right]\right. \\
&\qquad\qquad \left. + \frac{k-1}{e^\varepsilon + k - 1}\cdot \Pr_{D,\hat{w}\sim \mathrm{A}(D)}\left[y' = \operatorname{argmax}_{y\in[k]} \langle \hat{w}(i), e_y\rangle \mid y_i = y'\right]\right) \\
&\overset{(\clubsuit)}{\leq} \frac{1}{k}\sum_{y'\in[k]} \left(\frac{e^\varepsilon}{e^\varepsilon + k - 1}\cdot \Pr_{D,\hat{w}\sim \mathrm{A}(D)}\left[y' = \operatorname{argmax}_{y\in[k]} \langle \hat{w}(i), e_y\rangle \mid y_i = y'\right]\right. \\
&\qquad\qquad \left. + \frac{k-1}{e^\varepsilon + k - 1}\left(e^\varepsilon\cdot \Pr_{D,\hat{w}\sim \mathrm{A}(D)}\left[y' = \operatorname{argmax}_{y\in[k]} \langle \hat{w}(i), e_y\rangle \mid y_i \neq y'\right] + \delta\right)\right) \\
&\leq \frac{1}{k}\sum_{y'\in[k]} \left(\frac{e^\varepsilon}{e^\varepsilon + k - 1}\cdot \Pr_{D,\hat{w}\sim \mathrm{A}(D)}\left[y' = \operatorname{argmax}_{y\in[k]} \langle \hat{w}(i), e_y\rangle\right] + \delta\right) \\
&= \frac{e^\varepsilon}{e^\varepsilon + k - 1} + \delta,
\end{aligned}
$$

where ($\blacksquare$) follows from $y_i \sim y'$ and ($\clubsuit$) follows from the (ε, δ)-semi-sensitive DP guarantee of A.

Now, let $I_{\hat{w},D}$ denote the set $\{i \in [n] \mid \langle \hat{w}(i), e_{y_i}\rangle > \frac{1}{\sqrt{2}}\|\hat{w}(i)\|_2\}$. The above inequality implies that

$$
\mathbb{E}_{D,\hat{w}\sim \mathrm{A}(D)}[|I_{\hat{w},D}|] \leq \left(\frac{e^\varepsilon}{e^\varepsilon + k - 1} + \delta\right) n \leq 0.99n, \tag{12}
$$

where the inequality is from our assumption on ε, δ, k.

Meanwhile, $|I_{\hat{w},D}|$ can be used to bound the loss function as follows.

$$
\begin{aligned}
\mathcal{L}(w; D) &= \frac{1}{n}\sum_{i\in[n]} \ell(w, (i, y_i)) \\
&= \frac{-G}{n}\sum_{i\in[n]} \left\langle w, e_{j(i,y_i)}\right\rangle \\
&= \frac{-G}{n}\sum_{i\in[n]} \langle w(i), e_{y_i}\rangle \\
&= \frac{-G}{n}\left[\left(\sum_{i\in I_{w,D}} \langle w(i), e_{y_i}\rangle\right) + \left(\sum_{i\notin I_{w,D}} \langle w(i), e_{y_i}\rangle\right)\right] \\
&\geq \frac{-G}{n}\left[\left(\sum_{i\in I_{w,D}} \|w(i)\|_2\right) + \left(\sum_{i\notin I_{w,D}} \frac{1}{\sqrt{2}}\|w(i)\|_2\right)\right]
\end{aligned}
$$

$$\begin{aligned}
&\overset{(\blacktriangle)}{\geq} \frac{-L}{n}\left[\sqrt{\left(\sum_{i\in I_{w,D}} 1\right)+\left(\sum_{i\notin I_{w,D}} \frac{1}{2}\right)}\cdot\sqrt{\left(\sum_{i\in I_{w,D}} \|w(i)\|_2^2\right)+\left(\sum_{i\notin I_{w,D}} \|w(i)\|_2^2\right)}\right]\\
&= \frac{-G}{n}\cdot\sqrt{n/2+|I_{w,D}|/2}\cdot\|w\|_2\\
&\geq \frac{-RG}{n}\cdot\sqrt{n/2+|I_{w,D}|/2},
\end{aligned}$$

where $(\blacktriangle)$ follows from Cauchy–Schwarz.

Note that, by picking $w^* = \frac{R}{\sqrt{n}}\sum_{i\in[n]} e_{j(i,y_i)}$, we have

$$\mathcal{L}(w^*; D) = -\frac{RG}{\sqrt{n}}.$$

Thus, the excess risk is

$$\mathcal{L}(w; D) - \mathcal{L}(w^*; D) \geq \frac{RG}{n}\left(\sqrt{n} - \sqrt{n/2+|I_{w,D}|/2}\right).$$

As a result, the expected excess risk of A is

$$\begin{aligned}
\mathbb{E}_{D,\hat{w}\sim\mathrm{A}(D)}[\mathcal{L}(\hat{w}, D) - \mathcal{L}(w^*, D)] &\geq \mathbb{E}_{D,\hat{w}\sim\mathrm{A}(D)}\left[\frac{RG}{n}\left(\sqrt{n} - \sqrt{n/2+|I_{\hat{w},D}|/2}\right)\right]\\
&\geq \frac{RG}{n}\left(\sqrt{n} - \sqrt{n/2+\mathbb{E}_{D,\hat{w}\sim\mathrm{A}(D)}[|I_{\hat{w},D}|]/2}\right)\\
&\overset{(12)}{\geq} \frac{RG}{n}\left(\sqrt{n} - \sqrt{0.995n}\right)\\
&\geq \Omega\left(\frac{RG}{\sqrt{n}}\right).
\end{aligned}$$

■

Theorem 33 now easily follows from applying the group privacy bound (Lemma 32).

Proof of Theorem 33 Let $r = \left\lfloor \frac{\ln k}{\varepsilon} \right\rfloor$. We will henceforth assume that $n \geq r$; otherwise, we can instead apply a lower bound for largest k' such that $\frac{\log k'}{\varepsilon} \leq n$ instead (which would give a lower bound of $\Omega(RG)$ already).

Suppose for the sake of contradiction that there is an (ε, δ)-semi-sensitive DP algorithm A that yields $o\left(\frac{RG\sqrt{\log k}}{\sqrt{\varepsilon n}}\right)$ excess risk in the aforementioned setting in the theorem statement. We assume w.l.o.g.[10] that n is divisible by r; let $n' = n/r$.

Let A' be an algorithm that takes in n' points, replicates each input datapoint r times and then runs A. From Lemma 32, A' is (ε', δ')-semi-sensitive DP for $\varepsilon' = r\varepsilon$ and $\delta' = \frac{e^{r\varepsilon}-1}{e^{\varepsilon}-1}\cdot\delta$. The expected excess risk of A' is

$$o\left(\frac{RG\sqrt{\log k}}{\sqrt{\varepsilon n}}\right) = o\left(\frac{RG}{\sqrt{n'}}\right).$$

10. Otherwise, we may simply add dummy input points with constant loss functions.

Furthermore, we have

$$\frac{e^{\varepsilon'}}{e^{\varepsilon'}+k-1}+\delta' = \frac{e^{r\varepsilon}}{e^{r\varepsilon}+k-1}+\frac{e^{r\varepsilon}-1}{e^{\varepsilon}-1}\cdot\delta \le \frac{k}{2k-1}+\frac{k}{\varepsilon}\cdot\delta \le 0.9.$$

This contradicts Theorem 35. ■

C.2. Strongly Convex (and Smooth) Case

The strongly convex and smooth case proceeds in very much the same way except we use the squared loss instead.

Theorem 36 *For any $\varepsilon, \delta, G, \mu > 0$ and $n, k \in \mathbb{N}$ such that $\frac{e^{\varepsilon}}{e^{\varepsilon}+k-1}+\delta < 0.99$, there exists an G-Lipschitz μ-strongly convex μ-smooth loss function $\ell : \mathcal{W}\times\mathcal{X}\to\mathbb{R}$ such that any (ε,δ)-semi-sensitive DP algorithm for ERM w.r.t ℓ has expected excess empirical risk at least $\Omega\left(\frac{G^2}{\mu n}\right)$.*

Proof We use the same notation as in the proof of Theorem 35 with $R = 0.5G/\mu$, except that we let the loss function be

$$\ell(w,(x,y)) = \frac{\mu}{2}\left\|w - R\cdot e_{j(x,y)}\right\|_2^2.$$

It is simple to see that ℓ is μ-strongly convex, μ-smooth, and G-Lipschitz.

Similar to the proof of Theorem 35, we can prove (12). From this, we rearrange the excess risk (where $w^* := \frac{R}{n}\sum_{i\in[n]} e_{j(i,y_i)}$) as follows:

$$\begin{aligned}
\mathcal{L}(w;D)-\mathcal{L}(w^*;D) &= \frac{\mu}{2}\|w-w^*\|_2^2 \\
&= \frac{\mu}{2}\sum_{i\in[n]}\left\|w(i)-\frac{R}{n}\cdot e_{y_i}\right\|_2^2 \\
&\ge \frac{\mu}{2}\sum_{i\notin I_{w,D}}\left\|w(i)-\frac{R}{n}\cdot e_{y_i}\right\|_2^2 \\
&= \frac{\mu}{2}\sum_{i\notin I_{w,D}}\left(\|w(i)\|^2-\frac{2R}{n}\langle w(i), e_{y_i}\rangle+\frac{R^2}{n^2}\right) \\
&\overset{(\spadesuit)}{\ge} \frac{\mu}{2}\sum_{i\notin I_{w,D}}\left(\|w(i)\|_2^2-\frac{R\sqrt{2}}{n}\cdot\|w(i)\|_2+\frac{R^2}{n^2}\right) \\
&= \frac{\mu}{2}\sum_{i\notin I_{w,D}}\left(\left(\|w(i)\|_2-\frac{R}{n\sqrt{2}}\right)^2+\frac{R^2}{2n^2}\right) \\
&\ge \frac{\mu R^2}{4n^2}\cdot(n-|I_{w,D}|),
\end{aligned}$$

where $(\spadesuit)$ follows from the definition of $I_{w,D}$.

Thus, we have

$$\mathbb{E}_{D,\hat{w}\sim \mathrm{A}(D)}[\mathcal{L}(\hat{w}, D) - \mathcal{L}(w^*, D)] \geq \frac{\mu R^2}{4n^2}\left(n - \mathbb{E}_{D,\hat{w}\sim \mathrm{A}(D)}[|I_{\hat{w},D}|]\right) \overset{(12)}{\geq} \Omega\left(\frac{\mu R^2}{n}\right) \geq \Omega\left(\frac{G^2}{\mu n}\right).$$

■

Again, Theorem 34 easily follows via group privacy.

Proof of Theorem 34 Let $r = \left\lfloor \frac{\ln k}{\varepsilon} \right\rfloor$. We will henceforth assume that $n \geq r$; otherwise, we can instead apply a lower bound for smallest k' such that $\frac{\log k'}{\varepsilon} \leq n$ instead (which gives a lower bound of $\Omega(G^2/\mu)$ already).

Suppose for the sake of contradiction that there is an algorithm A that yields $o\left(\frac{G^2}{\mu} \cdot \frac{\log k}{\varepsilon n}\right)$ excess risk in the aforementioned setting in the theorem statement. We assume w.l.o.g. that n is divisible by r; let $n' = n/r$.

Let A' be an algorithm that takes in n' points, replicates each input data point r times and then runs A. From Lemma 32, A' is (ε', δ')-semi-sensitive DP for $\varepsilon' = r\varepsilon$ and $\delta' = \frac{e^{r\varepsilon}-1}{e^{\varepsilon}-1} \cdot \delta$. The expected excess risk of A' is

$$o\left(\frac{G^2}{\mu} \cdot \frac{\log k}{\varepsilon n}\right) = o\left(\frac{G^2}{\mu} \cdot \frac{1}{n'}\right).$$

Similar to the calculation in the proof of Theorem 33, we have $\frac{e^{\varepsilon'}}{e^{\varepsilon'}+k-1} + \delta' \leq 0.9$. Thus, this contradicts Theorem 35. ■