Output Perturbation for General Differentially Private Convex Optimization with Improved Population Loss Bounds, Runtimes and Applications to Private Adversarial Training

Andrew Lowy[1] **and Meisam Razaviyayn** [2]

[1, 2]University of Southern California

[1] lowya@usc.edu and [2] razaviya@usc.edu

Abstract

Finding efficient, easily implementable differentially private algorithms that offer strong excess risk bounds is an important problem in modern machine learning. To date, most work has focused on private empirical risk minimization (ERM) or private population loss minimization. However, there are often other objectives–such as fairness, adversarial robustness, or sensitivity to outliers–besides average performance that are not captured in the classical ERM setup. To this end, we study a completely general family of convex, Lipschitz loss functions and establish the first known differentially private excess risk and runtime bounds for optimizing this broad class. We provide similar bounds under additional assumptions of β-smoothness and/or strong convexity.

We also address private stochastic convex optimization (SCO). While (ϵ, δ)-differential privacy ($\delta > 0$) has been the focus of much recent work in private SCO, proving tight excess population loss bounds and runtime bounds for $(\epsilon, 0)$-differential privacy remains a challenging open problem. We provide *the tightest known* $(\epsilon, 0)$-differentially private population loss bounds and *fastest runtimes* under the presence of (or lack of) smoothness and strong convexity assumptions. Our methods extend to the $\delta > 0$ setting, where we offer the unique benefit of ensuring differential privacy for arbitrary $\epsilon > 0$ by incorporating *a new form of Gaussian noise* proposed in (Zhao et al. 2019). Our results are achieved using perhaps the simplest yet practical differentially private algorithm: output perturbation. Although this method is not novel conceptually, our analysis shows that the power of this method to achieve strong privacy, utility, and runtime guarantees has not been fully appreciated in prior works. Finally, we apply our theory to two learning frameworks, "tilted ERM" and "adversarial learning". In particular, our theory quantifies tradeoffs between adversarial robustness, privacy, and runtime.

1 Introduction

In recent years, big data has become more prolific and widely used, while at the same time there has been a growing desire among the public (and regulators) for guarantees that their data remains private. Thus, an important problem in modern machine learning is how to conceal individuals' sensitive information in a dataset while simultaneously training a useful model on that dataset. Differential privacy provides a rigorous guarantee that, with high probability, an adversary cannot discover an individual's data by observing the output of an algorithm. More precisely, a randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathbb{R}^d$ is said to be (ϵ, δ)-differentially private if for all measurable subsets $\mathcal{K} \subseteq \text{range}(\mathcal{A})$ and all n-element data sets $X, X' \in \mathcal{X}^n$ which differ by at most one observation (i.e. $|X \Delta X'| \leq 2$), we have

$$\mathbb{P}(\mathcal{A}(X) \in \mathcal{K}) \leq \mathbb{P}(\mathcal{A}(X') \in \mathcal{K})e^\epsilon + \delta,$$

where the probability is (solely) over the randomness of $\mathcal{A}$ (Dwork and Roth 2014). We refer to data sets differing in only one observation ($|X \Delta X'| \leq 2$) as "adjacent." If $\delta = 0$, we may say an algorithm is ϵ-differentially private. An (ϵ, δ)-differentially private algorithm is ϵ-differentially private with probability $1 - \delta$ (Dwork and Roth 2014, Lemma 3.17). Therefore, while large values of ϵ can still provide some privacy, $\delta \ll 1$ is necessary for meaningful privacy guarantees. In fact, $\delta \ll \frac{1}{n}$ is typically desirable: otherwise, a model may leak individuals' data and still satisfy the privacy constraint (Dwork and Roth 2014). For example, if $\delta > 0$, then we may have $\mathbb{P}(\mathcal{A}(X) \in \mathcal{K}) = 0$, and $\mathbb{P}(\mathcal{A}(X') \in \mathcal{K}) = \delta$ for some $\mathcal{K} \subseteq \text{range}(\mathcal{A})$, adjacent datasets $X, X' \in \mathcal{X}^n$. Therefore, if $\mathcal{K}$ occurs, then $\mathcal{A}$ completely reveals the underlying data set X'. While a lot of the literature has focused on efficient algorithms for (ϵ, δ)-differentially private algorithms, the important case of $\delta = 0$ has been neglected. The first contribution of this work is to fill this void.

Assume the parameters of a machine learning model are trained via solving the minimization problem:

$$\tilde{w}(X) \approx \arg\min_{w \in \mathbb{R}^d} F(w, X). \tag{1}$$

In the case of empirical risk minimization (ERM), where $F(w, X) = \frac{1}{n}\sum_{i=1}^n f(w, x_i)$, algorithms for maintaining differential privacy through observing $\tilde{w}(X)$ is well studied (Chaudhuri et al. 2011; Bassily et al. 2014; Zhang et al. 2017; Wang et al. 2017). Here (and throughout) $X = \{x_i\}_{i=1}^n$ is a data set with observations in some set $\mathcal{X} \subseteq \mathbb{R}^q$, and the weights $w \in \mathbb{R}^d$. More recently, several works have also considered private stochastic convex optimization (SCO), where the goal is to minimize the expected population loss $F(w, X) = \mathbb{E}_{x \sim \mathcal{D}}[f(w, x)]$, given access to n i.i.d

samples $X = \{x_i\}_{i=1}^n$ (Bassily et al. 2019; Feldman et al. 2020; Arora et al. 2020). However, the algorithms in these works are only differentially private for $\delta > 0$, which, as discussed earlier, provides substantially weaker privacy guarantees. Therefore, providing efficient, practical algorithms for $(\epsilon, 0)$-differentially private SCO is an important gap that we fill in the present work.

Our second main contribution is differentially private convex optimization for general (non-ERM, non-SCO) loss functions. While ERM and SCO are useful if average performance is the goal, there are situations where another objective besides average performance is desirable. For example, one may want to train a machine learning model that ensures some subsets of the population are treated fairly (see e.g. (Datta et al. 2015)), or one that is robust to corrupted data or adversarial attacks (Goodfellow et al. 2015), or one that has lower variance to allow for potentially better generalization. One may also want to diminish the effect of outliers or increase sensitivity to outliers. In these cases, it may be more fruitful to consider an alternative loss function that is not of ERM form. For example, the max-loss function

$$F(w, X) = \max\{f(w, x_1), ...f(w, x_n)\}$$

provides a model that has good "worst-case" performance and such F is clearly not of ERM form. The recently proposed "tilted ERM" (TERM) framework (Li et al. 2020) aims to address these shortcomings of standard ERM and encompasses the max-loss mentioned above. As another example, one may want a ML model that offers "accuracy at the top" (AATP) for applications such as recommendation systems, since many users will only browse the first ("top") few suggestions that are returned (Boyd et al. 2012). Maximizing accuracy at the top can be formulated as an optimization problem that is not ERM form (Boyd et al. 2012). Existing differentially privacy utility and runtime results have all been derived specifically for standard ERM or SCO and therefore would not apply to objectives such as TERM and AATP, which do not fall into either of these two (ERM or SCO) categories. Beyond machine learning, non-ERM losses appear in other engineering applications, such as power grid scheduling and sensor networks (Fioretto et al. 2019; Wang et al. 2018).

Our last main contribution is to specialize our theory and framework to DP TERM (discussed above) and adversarial training. In particular, for smooth strongly convex TERM, we derive excess risk bounds that (nearly) extend the optimal differentially private ERM bounds of (Bassily et al. 2014), since the TERM objective encompasses ERM in the limit. In adversarial training, the goal is to train a model that has robust predictions to an adversary's perturbations (with respect to some perturbation set S) of the feature data. This problem has gained a lot of attention in recent years, since it was first observed that neural nets can often be fooled by tiny, human-imperceptible perturbations into misclassifying images (Goodfellow et al. 2015). However, the challenging task of ensuring such adversarial training is executed in a differentially private manner has received much less attention by researchers. Indeed, we are not aware of any prior works that have shown how to keep the adversarial training procedure differentially private and provided adversarial risk and runtime bounds. Perhaps the closest step in this direction is the work (Phan et al. 2020), which provides a differentially private algorithm for training a classifier that is "certifiably robust," in the sense that with high probability, the classifier's predicted label is stable under small perturbations. However, our measure of adversarial robustness is different: we look at excess adversarial risk and provide tight bounds that depend expicitly on the privacy parameters. This allows for an interpretation of the tradeoffs between robustness, privacy, and runtime. Furthermore the algorithm of (Phan et al. 2020), a noisy stochastic batch gradient descent, is quite complicated to implement and does not come with runtime bounds.

Our theory is built on the idea of *output perturbation*. Conceptually, the output perturbation mechanism outputs

$$w_{\mathcal{A}}(X) := \Pi_{B(0,R)}\left[\tilde{w}(X) + b\right],$$

where $\tilde{w}$ is the output of some non-private algorithm and $b \in \mathbb{R}^d$ is some suitably chosen random noise vector. Here $\Pi_{B(0,R)}(z) := \arg\min_{w \in B(0,R)} \|z - w\|_2$ is the projection onto the closed euclidean ball (centered at 0) of radius R, $B(0, R) \subset \mathbb{R}^d$, which is large enough to contain $w^*(X)$; see "Notation" paragraph below for more detail. Technically, projection is necessary for our excess risk analysis in the non-smooth case, since we require Lipschitzness at $w_{\mathcal{A}}(X)$ (recall that strongly convex function cannot be Lipschitz on the whole space). Output perturbation has been studied in the differential privacy literature for many years (Dwork et al. 2006; Chaudhuri et al. 2011; Zhang et al. 2017). In early works, which culminated in (Dwork et al. 2006), the method was introduced and proven to be differentially private. In (Chaudhuri et al. 2011), high probability excess risk and population loss bounds for linear classifiers with strongly convex regularizers in the ERM and SCO settings are given for output perturbation (Chaudhuri et al. 2011, Thm 15, Lemma 16). However, no practical implementation is provided. As a first step in the practical direction, (Zhang et al. 2017) shows how to implement output perturbation with gradient descent in the smooth ERM setting, providing excess empirical risk and runtime bounds. Their privacy analysis is tied to the particular non-private optimization method they use, which hinders their runtime potential and makes their analysis less transparent.

Notation. Recall that a function $h : \mathcal{W} \to \mathbb{R}$ on some domain $\mathcal{W} \subseteq \mathbb{R}^d$ is *convex* if $h(\lambda w + (1-\lambda)w') \leq \lambda h(w) + (1-\lambda)h(w')$ for all $\lambda \in [0, 1]$ and all $w, w' \in \mathcal{W}$. We say h is μ-*strongly convex* if $h(w) - \frac{\mu}{2}\|w\|_2^2$ is convex. Also, h is L-*Lipschitz* if $\|h(w) - h(w')\| \leq L\|w - w'\|_2$ for all $w, w' \in \mathcal{W}$ and β-*smooth* if h is differentiable and $\nabla h(w)$ is β-Lipschitz. We always assume that $F : \mathbb{R}^d \times \mathcal{X}^n :\to \mathbb{R}$ is such that for all $X \in \mathcal{X}^n$, $F(\cdot, X)$ is convex on $\mathbb{R}^d$ and L-Lipschitz on $B(0, R)$, where

$$R := \sup_{X \in \mathcal{X}^n} \inf_{w^*(X)} \|w^*(X)\|_2 + 1,$$

and the infimum is over $w^*(X) \in \operatorname{argmin}_{w \in \mathbb{R}^d} F(w, X)$. We denote the family of such convex, L-Lipschitz (on $B(0, R)$) functions by $\mathcal{G}_{L,R}$. Thus, if $F \in \mathcal{G}_{L,R}$,

then $B(0,R)$ contains at least one global minimizer of $F(\cdot, X)$ in its interior for all $X \in \mathcal{X}^n$. The data universe $\mathcal{X}$ can be be any set, and we often denote $X = (x_1, \cdots, x_n) \in \mathcal{X}^n$. We will also work with the following families of functions, which are each subsets of $\mathcal{G}_{L,R}$, throughout: $\mathcal{F}_{\mu,L,R} := \{F : \mathbb{R}^d \times \mathcal{X}^n \to \mathbb{R} \mid F(\cdot, X)$ is μ-strongly convex $\forall\, X\}$; $\mathcal{H}_{\beta,\mu,L,R} := \{F \in \mathcal{F}_{\mu,L,R} : F(\cdot, X)$ is β-smooth $\forall X\}$; and $\mathcal{J}_{\beta,L,R} := \{F \in \mathcal{G}_{L,R} : F(\cdot, X)$ is β-smooth $\forall\, X \in \mathcal{X}^n\}$. For $F \in \mathcal{H}_{\beta,\mu,L,R}$, denote the condition number by $\kappa = \frac{\beta}{\mu}$. Also, $\mathcal{F}^{ERM}_{\mu,L,R}, \mathcal{H}^{ERM}_{\beta,\mu,L,R}, \ldots$ are defined to be the subset of functions in $\mathcal{F}_{\mu,L,R}, \mathcal{H}_{\beta,\mu,L,R}, \ldots$ that are of ERM form, i.e. $F(w, X) = \frac{1}{n}\sum_{i=1}^n g(w, x_i)$ for all $w \in \mathbb{R}^d$, $X \in \mathcal{X}^n$ for some convex $g : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$. Finally, define the following constant that is used in our choice of Gaussian noise vector: $c_\delta := \sqrt{\log\left(\frac{2}{\sqrt{16\delta+1}-1}\right)}$ (Zhao et al. 2019).

2 Output Perturbation for General Differentially Private Convex Optimization

In this section, we implement the classical output perturbation algorithm and provide a tight analysis for different classes of loss functions:

Strongly Convex Loss

Our framework when $F(w, X)$ in (1) is strongly convex in w is described in Algorithm 1. Note that it accepts as input any non-private optimization algorithm and transforms it into a differentially private one, by adding noise to the approximate minimizer w_T. As a result, we are able to obtain runtimes as fast as non-private optimization method.

Algorithm 1 Black Box Output Perturbation Algorithm for $\mathcal{F}_{\mu,L,R}$ and $\mathcal{H}_{\beta,\mu,L,R}$

Require: non-private (possibly randomized) optimization method $\mathcal{M}$, $n, d \in \mathbb{N}$, privacy parameters $\epsilon > 0, \delta \geq 0$, data set $X \in \mathcal{X}^n$, function $F(w, X) \in \mathcal{F}_{\mu,L,R}$, accuracy parameter $\alpha > 0$ with corresponding iteration number $T = T(\alpha) \in \mathbb{N}$ (such that $\mathbb{E}[F(w_T(X), X) - F(w^*(X), X)] \leq \alpha$).

1: Run $\mathcal{M}$ for $T = T(\alpha)$ iterations to ensure $\mathbb{E}F(w_T(X), X) - F(w^*, X) \leq \alpha$.

2: Add noise to ensure privacy: $w_{\mathcal{A}} := \Pi_{B(0,R)}(w_T + \widehat{z})$, where the pdf $p(\widehat{z})$ of $\widehat{z}$ is proportional to

$$\begin{cases} \exp\left(-\frac{\epsilon\|\widehat{z}\|_2}{\Delta_F + 2\sqrt{\frac{2\alpha}{\mu}}}\right) & \text{if } \delta = 0 \\ \exp\left(-\frac{\epsilon^2\|\widehat{z}\|_2^2}{(\Delta_F + 2\sqrt{\frac{2\alpha}{\mu}})^2(c_\delta + \sqrt{c_\delta^2+\epsilon})^2}\right) & \text{if } \delta > 0 \end{cases},$$

where $\Delta_F := \begin{cases} \frac{2L}{\mu n} & \text{if } F \in \mathcal{F}^{ERM}_{\mu,L,R} \\ \frac{2L}{\mu} & \text{if } F \in \mathcal{F}_{\mu,L,R} \setminus \mathcal{F}^{ERM}_{\mu,L,R} \end{cases}$

3: **return** $w_{\mathcal{A}}$.

First, let us establish the privacy guarantee for this algorithm.

Proposition 2.1 *Algorithm 1 is (ϵ, δ)-differentially private.*

For $F \in \mathcal{F}_{\mu,L,R}$, instantiating Algorithm 1 with the subgradient method (Nesterov 2014) results in the following.

Theorem 2.1 *Take $\mathcal{M}$ to be the Subgradient Method with step sizes $\eta_t = \frac{2}{\mu(t+1)}$ in Algorithm 1. Let $F \in \mathcal{F}_{\mu,L,R}$.*

a) Let $\delta = 0$, $\frac{d}{\epsilon} \leq 1$.

Setting $\alpha = \frac{L^2}{\mu}\frac{d}{\epsilon}$ gives $\mathbb{E}_{\mathcal{A}}F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 9\frac{L^2}{\mu}\frac{d}{\epsilon}$ in $T = \frac{2\epsilon}{d}$ gradient evaluations and runtime 2ϵ.

b) Let $\delta > 0$, $\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon} \leq 1$.

Setting $\alpha = \frac{L^2}{\mu}\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon}$ gives $\mathbb{E}_{\mathcal{A}}F(w^\delta_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 9\frac{L^2}{\mu}\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon}$ in $T = \frac{2\epsilon}{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})} \leq 2\sqrt{\frac{\epsilon}{d}}$ gradient evaluations and runtime $2\sqrt{\epsilon d}$.

If F is additionally β-smooth ($F \in \mathcal{H}_{\beta,\mu,L,R}$), our excess risk bounds and runtime can be improved using Nesterov's Accelerated Gradient Descent (AGD) (Nesterov 2014):

Theorem 2.2 *Let $F \in \mathcal{H}_{\beta,\mu,L,R}$. Take $\mathcal{M}$ to be Nesterov's Accelerated Gradient Descent (AGD) (Nesterov 2014) in Algorithm 1.*

a) Let $\delta = 0$, $\frac{d}{\epsilon} \leq 1$. Then setting $\alpha = \frac{L^2}{\mu}\min\left\{\kappa\left(\frac{d}{\epsilon}\right)^2, 1\right\}$ and $T = \left\lceil\sqrt{\kappa}\log\left(\frac{\beta R^2\mu}{L^2}\max\left\{1, \left(\frac{\epsilon}{d}\right)^2\frac{1}{\kappa}\right\}\right)\right\rceil$ gives $\mathbb{E}_{\mathcal{A}}F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 26\kappa\frac{L^2}{\mu}\left(\frac{d}{\epsilon}\right)^2$ in runtime $\widetilde{O}(\sqrt{\kappa}d)$.

b) Let $\delta \in (0, \frac{1}{2})$, $\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon} \leq 1$. Then setting $\alpha = \frac{L^2}{\mu}\min\left\{\kappa\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon}\right)^2, 1\right\}$ and $T = \left\lceil\sqrt{\kappa}\log\left(\frac{\beta R^2\mu}{L^2}\max\{1, \left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon}\right)^2\frac{1}{\kappa}\}\right)\right\rceil$ gives $\mathbb{E}_{\mathcal{A}}F(w^\delta_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 13.5\kappa\frac{L^2}{\mu}\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon}\right)^2$ in runtime $\widetilde{O}(\sqrt{\kappa}d)$

Convex Loss

Our Black Box implementation procedure for general convex loss (but in the absence of strong convexity) is described in Algorithm 2. Our first result is that this method is differentially private:

Proposition 2.2 *Algorithm 2 is (ϵ, δ)-differentially private.*

Next, we establish excess risk and runtime bounds for Algorithm 2 combined with the subgradient method:

Theorem 2.3 *Let $F \in \mathcal{G}_{L,R}$. Put $\mathcal{M}$ to be the subgradient method in the Black Box Algorithm 2 with α and T as prescribed below. Then there exist choices of $\lambda > 0$ such that the following results hold: a) If $\delta = 0, \frac{d}{\epsilon} \leq 1$,*

Algorithm 2 Black Box Output Perturbation Algorithm with Regularization for $\mathcal{G}_{L,R}$ and $\mathcal{J}_{\beta,L,R}$

Require: Number of data points $n \in \mathbb{N}$, dimension $d \in \mathbb{N}$ of data, non-private (possibly randomized) optimization method $\mathcal{M}$, privacy parameters $\epsilon > 0, \delta \geq 0$, data universe $\mathcal{X}$, data set $X \in \mathcal{X}^n$, function $F(w, X) \in \mathcal{G}_{L,R}$, accuracy and regularization parameters $\alpha > 0, \lambda > 0$ with corresponding iteration number $T = T(\alpha, \lambda) \in \mathbb{N}$ (such that $\mathbb{E}[F_\lambda(w_T(X), X) - F_\lambda(w^*(X), X)] \leq \alpha$).

1: Run $\mathcal{M}$ on $F_\lambda(w, X) = F(w, X) + \frac{\lambda}{2}\|w\|_2^2$ for $T = T(\alpha)$ iterations to ensure $\mathbb{E}F_\lambda(w_T(X), X) - F_\lambda(w_\lambda^*, X) \leq \alpha$.

2: Add noise to ensure privacy: $w_{\mathcal{A}} := \Pi_{B(0,R)}\left[w_T + \hat{z_\lambda}\right]$, where the density $p(\hat{z_\lambda})$ of $\hat{z_\lambda}$ is proportional to $\begin{cases} \exp\{-\frac{\epsilon\|\hat{z_\lambda}\|_2}{\Delta_\lambda + 2\sqrt{\frac{2\alpha}{\lambda}}}\} & \text{if } \delta = 0 \\ \exp\{-\frac{2\epsilon^2\|\hat{z_\lambda}\|_2^2}{(\Delta_\lambda + 2\sqrt{\frac{2\alpha}{\lambda}})^2(c_\delta + \sqrt{c_\delta^2 + \epsilon})^2}\} & \text{if } \delta > 0, \end{cases}$, where $\Delta_\lambda := \begin{cases} \frac{2(L+\lambda R)}{\lambda n} & \text{if } F \in \mathcal{G}_{L,R}^{ERM} \\ \frac{2(L+\lambda R)}{\lambda} & \text{if } F \in \mathcal{G}_{L,R} \setminus \mathcal{G}_{L,R}^{ERM} \end{cases}$ is an upper bound on the L_2 sensitivity of F_λ.

3: **return** $w_{\mathcal{A}}$.

then setting $\alpha = LR\left(\frac{d}{\epsilon}\right)^{3/2}$, $T = \left\lceil 12\left(\frac{\epsilon}{d}\right)^2\right\rceil$ *ensures* $\mathbb{E}_{\mathcal{A}}F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 49LR\sqrt{\frac{d}{\epsilon}}$ *in runtime* $O(\frac{\epsilon^2}{d})$.

b) Let $\delta \in \left(0, \frac{1}{2}\right), \frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon} \leq 1$. *Then setting* $\alpha = LR\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon}\right)^{3/2}$, $T = \left\lceil 12\frac{\epsilon^2}{d(c_\delta + \sqrt{c_\delta^2 + \epsilon})^2}\right\rceil$ *implies* $\mathbb{E}_{\mathcal{A}}F(w_{\mathcal{A}}^\delta(X), X) - F(w^*(X), X) \leq 25LR\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon}\right)^{1/2}$ *in runtime* $O(\epsilon)$.

As in the strongly convex, assuming smoothness allows us to improve our excess risk and runtime bounds:

Theorem 2.4 *Let* $F \in \mathcal{J}_{\beta,L,R}$ *(or* $\mathcal{J}_{\beta,L,R}^{ERM}$*). Take* $\mathcal{M}$ *to be AGD in Algorithm 2.*

a) Let $\delta = 0, \left(\frac{d}{\epsilon}\right)^2 \leq \frac{L}{R\beta}$.

Then setting $\alpha = \frac{L^{4/3}R^{2/3}}{\beta^{1/3}}(\frac{\epsilon}{d})^{2/3}$ *and* $T = \left\lceil \sqrt{2}(\frac{\beta R}{L}\frac{\epsilon}{d})^{1/3}\log\left(2\left(\frac{\beta R}{L}\right)^{4/3}\left(\frac{d}{\epsilon}\right)^{2/3}\right)\right\rceil$ *implies* $\mathbb{E}_{\mathcal{A}}F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 146\beta^{1/3}L^{2/3}R^{4/3}\left(\frac{d}{\epsilon}\right)^{2/3}$ *in runtime* $O\left(d^{2/3}\epsilon^{1/3}\left(\frac{\beta R}{L}\right)^{1/3}\log\left(\left(\frac{\beta R}{L}\right)^{4/3}\left(\frac{d}{\epsilon}\right)^{2/3}\right)\right)$ *for some choice of* $\lambda > 0$.

b) Let $\delta \in \left(0, \frac{1}{2}\right), \left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon}\right)^2 \leq \frac{L}{R\beta}$. *Then setting* $\alpha = \frac{L^{4/3}R^{2/3}}{\beta^{1/3}}\frac{\epsilon}{\sqrt{d(c_\delta + \sqrt{c_\delta^2 + \epsilon})}}\min\{1, \frac{1}{\beta}\}$ *and* $T = \widetilde{O}\left(\sqrt{2}\left(\frac{\beta R}{L}\frac{\epsilon}{\sqrt{d(c_\delta + \sqrt{c_\delta^2 + \epsilon})}}\right)^{1/3}\right)$ *implies* $\mathbb{E}_{\mathcal{A}}F(w_{\mathcal{A}}^\delta(X), X) - F(w^*(X), X) \leq 27\beta^{1/3}L^{2/3}R^{4/3}\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon}\right)^{2/3}$ *in runtime* $\widetilde{O}\left(d^{5/6}\epsilon^{1/6}\left(\frac{\beta R}{L}\right)^{1/3}\right)$ *for some choice of* $\lambda > 0$.

3 DP SCO: Population Loss and Runtime Bounds

Recent works (Bassily et al. 2014, 2019; Feldman et al. 2020) have established tight bounds on the expected population loss of (ϵ, δ)-private stochastic optimization algorithms for $\delta > 0$. However, the important case of $\delta = 0$, which provides the strongest privacy guarantee, has largely been overlooked, except for (Bassily et al. 2014). We show that our simple output perturbation algorithm results in excess population loss bounds that improve the excess population loss bounds in (Bassily et al. 2014) and can be executed in substantially less runtime. Our method also works for $\delta > 0$, where it serves as a simple, flexible, efficient alternative which, unlike other algorithms, allows for arbitrary $\epsilon > 0$.

To proceed, we change some notation for further clarity and introduce some new definitions. Let the data have some (unknown) distribution $x \sim \mathcal{D}$. We are given a sample dataset $X \sim \mathcal{D}^n$. We denote the empirical loss $\widehat{F}(w, X) := \frac{1}{n}\sum_{i=1}^n f(w, x_i)$ and the **population loss** $F(w, \mathcal{D}) := \mathbb{E}_{x\sim\mathcal{D}}f(w, x)$. Our goal is to understand perturbation framework in privately minimizing the **excess population loss**, defined at a point $w \in \mathbb{R}^d$ by $F(w, \mathcal{D}) - \min_{w\in\mathbb{R}^d}(w, \mathcal{D}) = F(w, \mathcal{D}) - F(w^*(\mathcal{D}), \mathcal{D})$, where we denote the parameter that minimizes the population loss by $w^*(\mathcal{D})$. To avoid any ambiguity, we will denote the minimizer of the empirical loss for a given data set X by $\widehat{w}(X)$.

Strongly Convex, Lipschitz Loss

We begin with our results for strongly convex (and possibly non-smooth) loss. By using stochastic subgradient descent (SGD) to approximately minimize the empirical objective $\widehat{F}$ in Algorithm 1, we can attain tight excess population loss bounds efficiently.

Theorem 3.1 *Let* $f(w, x)$ *be* μ*-strongly convex and* L*-Lipschitz in* w *for all* $x \in \mathcal{X}$. *Run* $\mathcal{A}$ *= Algorithm 1 on* $\widehat{F}$ *with* $\mathcal{M}$ *as the stochastic subgradient method with step sizes* $\eta_t = \frac{2}{\mu(t+1)}$ *and* T, α *as prescribed below.*

1. Suppose $\delta = 0$, $\frac{d}{\epsilon n} \leq 1$. *Then putting* $\alpha = \frac{L^2}{\mu n}\min\{\frac{1}{n}, \frac{d}{\epsilon}\}$ *and* $T = \lceil 2\max\{n^2, \frac{n\epsilon}{d}\}\rceil$ *implies* $\mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}}F(w_{\mathcal{A}}(X), \mathcal{D}) - F(w^*(\mathcal{D}), \mathcal{D}) \leq \frac{L^2}{\mu}\left(\frac{5}{n} + 9\frac{d}{\epsilon n}\right)$. *The runtime of this method is* $O(\max\{dn^2, \epsilon n\})$.

2. Suppose $\delta \in (0, \frac{1}{2})$, $\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon n} \leq 1$. *Then putting* $\alpha = \frac{L^2}{\mu n}\min\left\{\frac{1}{n}, \frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon}\right\}$

and $T = \lceil 2n \max\left\{n, \frac{\epsilon}{\sqrt{d}(2c_\delta+\sqrt{\epsilon})}\right\}\rceil$ *implies* $\mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}} F(w^\delta_{\mathcal{A}}(X),\mathcal{D}) - F(w^*(\mathcal{D}),\mathcal{D}) \leq \frac{L^2}{\mu}\left(\frac{5}{n} + 6\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)$. *The resulting runtime is* $O\left(\max\left\{dn^2, n\sqrt{d\epsilon}\right\}\right)$.

The optimal non-private excess population loss for μ-strongly convex, L-Lipschitz f is $O(\frac{L^2}{\mu n})$ (Hazan and Kale 2014). So if $\delta = 0$ and $d \lesssim \epsilon$, then we match the optimal non-private rate and in effect get "privacy for free." Likewise, for $\delta > 0$, if $\sqrt{d}c_\delta + \sqrt{\epsilon} \lesssim \epsilon$, we get privacy for free.

As noted earlier, neither (Bassily et al. 2019), nor (Feldman et al. 2020) provide excess population loss bounds for $\delta = 0$. The only population loss bound with $\delta = 0$ is $O\left(\frac{L^2}{\mu}\left(\frac{d}{\epsilon n}\sqrt{\log(n)} + \frac{1}{n}\right)\right)$ (Bassily et al. 2014, Theorem F.2), obtained by exponential sampling + localization. Thus, we outperform their bounds by a logarithmic factor. Moreover, as noted above, our runtime is $O\left(\max\left\{n^2d, \epsilon n\right\}\right)$ for $\delta = 0$. By contrast, the worst-case runtime of the exponential sampling + localization method is much larger: $\widetilde{O}\left(\frac{L^2}{\mu^2}d^9\frac{n}{\epsilon^2}\max\{1, \frac{L}{\mu}\}\right)$ (Bassily et al. 2014).

Smooth, Strongly Convex, Lipschitz Loss

With the smoothness assumption, we can also use Katyusha (Allen-Zhu 2018), an accelerated stochastic method, instead of SGD for implementation and obtain faster runtimes:

Theorem 3.2 *Let $f(w,x)$ be μ-strongly convex, L-Lipschitz, β-smooth in w for all $x \in \mathcal{X}$, and let $X \in \mathcal{X}^n$. Run $\mathcal{A}$ = Algorithm 1 on $\widehat{F}$ with $\mathcal{M}$ as Katyusha with $T = O\left((n+\sqrt{n\kappa})\log(\frac{LR}{\alpha})\right)$ and α as prescribed below.*

1. Suppose $\delta = 0$, $\frac{d}{\epsilon n} \leq 1$. Then putting $\alpha = \frac{L^2}{\mu n^2}\min\left\{\kappa\left(\frac{d}{\epsilon}\right)^2, 1\right\}$ results in a point $w_{\mathcal{A}}(X)$ such that $\mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}} F(w_{\mathcal{A}}(X),\mathcal{D}) - F(w^*(\mathcal{D}),\mathcal{D}) \leq \frac{L^2}{\mu}\left(\frac{5}{n} + 26\kappa\left(\frac{d}{\epsilon n}\right)^2\right)$ in $T = \widetilde{O}(n+\sqrt{n\kappa})$ stochastic gradient evaluations. The resulting runtime is $\widetilde{O}(d(n+\sqrt{n\kappa}))$.

2. Suppose $\delta \in (0, \frac{1}{2})$, $\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon n} \leq 1$. Then putting $\alpha = \frac{L^2}{\mu n^2}\min\left\{\kappa\left(\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon}\right)^2, 1\right\}$ results in a point $w_{\mathcal{A}}(X)$ such that $\mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}} F(w_{\mathcal{A}}(X),\mathcal{D}) - F(w^*(\mathcal{D}),\mathcal{D}) \leq \frac{L^2}{\mu}\left(\frac{5}{n} + 13.5\kappa\left(\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)^2\right)$ in $T = \widetilde{O}(n+\sqrt{n\kappa})$ stochastic gradient evaluations. The resulting runtime is $\widetilde{O}(d(n+\sqrt{n\kappa}))$.

The non-private statistically optimal population loss is $O(\frac{L^2}{\mu n})$, as for non-smooth. Hence, if $\delta = 0$, we get "privacy for free" whenever $\kappa\left(\frac{d}{\epsilon}\right)^2 \lesssim n$ or $d \lesssim \epsilon$. For $\delta > 0$, we get "privacy for free" whenever $\frac{\kappa p c_\delta^2}{\epsilon} \lesssim n$ or $\sqrt{d}c_\delta + \sqrt{\epsilon} \lesssim \epsilon$. There are no private population loss bounds we are aware of for the smooth, strongly convex, Lipschitz class with $\delta = 0$, other than the one from exponential sampling + localization mentioned above, which does not benefit additionally from smoothness (Bassily et al. 2014). Therefore, our method provides the tightest $(\epsilon, 0)$-differentially private excesss population loss bounds that we are aware of and runs in less time than any competing algorithm.

Convex, Lipschitz Loss

For convex loss $f(w,x)$, applying the regularized output perturbation Algorithm 2 with SGD yields the following excess population loss and runtime bounds:

Theorem 3.3 *Let $f(w,x)$ be μ-strongly convex and L-Lipschitz in w for all $x \in \mathcal{X}$. Run $\mathcal{A}$ = Algorithm 2 on $\widehat{F_\lambda}$ with $\mathcal{M}$ as the stochastic subgradient method with step sizes $\eta_t = \frac{2}{\lambda(t+1)}$ and T, α as prescribed below. There exist choices of $\lambda > 0$ such that the following hold:*

1. Suppose $\delta = 0$, $\frac{d}{\epsilon n} \leq 1$. Set $\alpha = \frac{LR\left(\frac{d}{\epsilon n}\right)^{3/2}}{(1+\frac{d}{\epsilon})^2}$, $T = 2n^2\max\{1, (\frac{\epsilon}{d})^2\}$. Then $\mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}} F(w_{\mathcal{A}}(X),\mathcal{D}) - F(w^*(\mathcal{D}),\mathcal{D}) \leq 32LR\left(\left(\frac{d}{\epsilon n}\right)^{1/2} + \frac{1}{\sqrt{n}}\right)$ in runtime $O(dn^2\max\{1, (\frac{\epsilon}{d})^2\})$.

2. Suppose $\delta \in (0, \frac{1}{2})$. Set $\alpha = LR\left(\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)^{3/2}$, $T = 2\frac{n^2\epsilon}{d}$. Then $\mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}} F(w^\delta_{\mathcal{A}}(X),\mathcal{D}) - F(w^*(\mathcal{D}),\mathcal{D}) \leq 32LR\left(\left(\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)^{1/2} + \frac{1}{\sqrt{n}}\right)$ in runtime $O(n^2\epsilon)$.

The non-private optimal excess population loss for convex, L-Lipschitz functions is $O(\frac{LR}{n})$ (Hazan and Kale 2014). Thus, we have "privacy for free" when $d \lesssim \epsilon, \delta = 0$, or when $d(c_\delta + \sqrt{c_\delta^2+\epsilon}) \lesssim \epsilon^2, \delta \in (0, \frac{1}{2})$. For $\delta = 0$, the only competing excess population loss bound for this class that we are aware of is $\widetilde{O}\left((LR)^2\left(\sqrt{\frac{d}{\epsilon n}} + \frac{1}{\sqrt{n}}\right)\right)$, obtained by the exponential mechanism (Bassily et al. 2014), which is larger than our bound by a factor of LR. Its runtime $\widetilde{O}(R^2d^6n^3\max\{d, \epsilon nR\})$, is also generally much larger than ours.

Smooth, Convex, Lispchitz Loss

If we assume additionally that f is β-smooth, then with Katyusha as the non-private input algorithm, our excess population loss and runtime bounds improve:

Theorem 3.4 *Let $f(w,x)$ be β-smooth, μ-strongly convex, and L-Lipschitz in w for all $x \in \mathcal{X}$. Run $\mathcal{A}$ = Algorithm 2 on $\widehat{F_\lambda}$ with $\mathcal{M}$ as Katyusha with α, T as prescribed below. There exists $\lambda > 0$ such that the following hold:*

1. Suppose $\delta = 0$, $\left(\frac{d}{\epsilon n}\right)^2 \leq \frac{LR}{\beta}$. Set $\alpha = \frac{L^{4/3}R^{2/3}}{\beta^{1/3}}$, and $T = O\left(n + n^{5/6}\left(\frac{\epsilon}{d}\right)^{1/3}\left(\frac{\beta R}{L}\right)^{1/3}\log\left(\left(\frac{\beta R}{L}\right)^{1/3}\left(\frac{d}{\epsilon}\right)^{2/3}n^{4/3}\right)\right)$. Then

$\mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}}F(w_{\mathcal{A}}(X),\mathcal{D}) - F(w^*(\mathcal{D}),\mathcal{D}) \leq 64\beta^{1/3}L^{2/3}R^{4/3}(\frac{1}{\sqrt{n}} + \left(\frac{d}{\epsilon n}\right)^{2/3})$ *in runtime* $O\left(nd + n^{5/6}d^{2/3}\epsilon^{1/3}\left(\frac{\beta R}{L}\right)^{1/3}\right)$.

2. Suppose $\delta \in (0, \frac{1}{2})$, $\left(\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)^2 \leq \frac{LR}{\beta}$. *Set* $\alpha = \frac{L^{4/3}R^{2/3}}{\beta^{1/3}}(\frac{\epsilon n}{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})})^{2/3}\frac{1}{n^2}$, *and* $T = \widetilde{O}\left(n + n^{5/6}\left(\frac{\epsilon}{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}\right)^{1/3}\left(\frac{\beta R}{L}\right)^{1/3}\right)$. *Then* $\mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}}F(w_{\mathcal{A}}^\delta(X),\mathcal{D}) - F(w^*(\mathcal{D}),\mathcal{D}) \leq 64\beta^{1/3}L^{2/3}R^{4/3}\left(\frac{1}{\sqrt{n}} + \left(\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)^{2/3}\right)$ *in runtime* $O\left(nd + n^{5/6}d^{5/6}\epsilon^{1/6}\left(\frac{\beta R}{L}\right)^{1/3}\right)$.

4 Specialization to TERM and Adversarial Training

Differentially Private Tilted ERM (TERM)

Consider for $\tau > 0$ the Tilted ERM (TERM) objective

$$F_\tau(w, X) := \frac{1}{\tau}\log\left(\frac{1}{n}\sum_{i=1}^n e^{\tau f(w,x_i)}\right)$$

(see e.g. (Kort and Bertsekas 1972; Pee and Royset 2011; Cohen and Shashua 2014; Cohen et al. 2016; Katharopoulos and Fleuret 2017; Li et al. 2020) and the references therein for the applications of this loss). It is easy to show that as $\tau \to 0, F_\tau(w,x) \to \frac{1}{n}\sum_{i=1}^n f(w,x_i)$, so this extends the classical ERM framework in the limit. It also encompasses the max loss ($F_{max}(w,X) = \max\{f(w,x_1),...,f(w,x_n)\}$) for instance, by letting $\tau \to \infty$. More generally, a benefit of the TERM framework is that it allows for a continuum of solutions between average and max loss, which, for example, can be calibrated to promote a desired level of fairness in the the machine learning model (Li et al. 2020). Another interpretation of TERM is that as τ increases, the variance of the model decreases, while the bias increases. Thus, τ can also be tuned to improve the generalization of the model via the bias/variance tradeoff. In what follows, we specialize our developed theory to the TERM objective function.

Strongly convex, Lipschitz, twice differentiable $f(\cdot,x)$ Next we show that if f is "nice enough," then $F_\tau \in \mathcal{F}_{\mu,L,R}$, so that the excess risk bounds proved earlier hold; we will also refine these results to show how excess risk depends on the parameter $\tau > 0$.

Lemma 4.1 *((Li et al. 2020, Lemmas 1 and 3)) Assume $f(\cdot,x)$ is a μ-strongly convex, twice differentiable, L-Lipschitz (on $B(0,R)$) loss function for all $x \in \mathcal{X}$. Then $F_\tau(\cdot,X)$ is μ-strongly convex, twice differentiable, L-Lipschitz (on $B(0,R)$) for all $X \in \mathcal{X}^n$.*

By Lemma 4.1, if we put the subgradient method as $\mathcal{M}$ in Algorithm 1, then we get $\mathbb{E}_{\mathcal{A}}F(w_{\mathcal{A}}(X),X) - F(w^*(X),X) \leq 9\frac{L^2}{\mu}\frac{d}{\epsilon}$ in runtime $O(n\epsilon)$ for $\delta = 0$ by invoking Theorem 2.1, with $\alpha = \frac{L^2}{\mu}\frac{d}{\epsilon}, T = \frac{\epsilon}{d}$. Similarly, for $\delta > 0$, we get $\mathbb{E}_{\mathcal{A}}F(w_{\mathcal{A}}^\delta(X),X) - F(w^*(X),X) \leq 9\frac{L^2}{\mu}\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon}$ in runtime $O(n\sqrt{\epsilon d})$. These results hold even as $\tau \to \infty$. If, in addition, f is bounded on $B(0,R)\times\mathcal{X}$, then we get the following more refined excess risk bounds.

Proposition 4.1 *Let $\tau > 0$. Suppose $f(\cdot,x)$ is L-Lipschitz on $B(0,R)$ (where $\|w^*(X)\| \leq R$), twice differentiable, and μ-strongly convex for all $x \in \mathcal{X}$. Moreover, assume $a_R \leq f(w,x) \leq A_R$ for all $w \in B(0,R)$ and all $x \in \mathcal{X}$. Denote $C_\tau = e^{\tau(A_R-a_r)}$. Then running Algorithm 1 on F_τ with $\mathcal{M}$ as the subgradient method yields $\mathbb{E}_{\mathcal{A}}F_\tau(w_{\mathcal{A}}(X),X) - F_\tau(w^*(X),X) \leq 9\frac{L^2C_\tau}{\mu}\frac{d}{\epsilon n}$ in runtime $O\left(\frac{n^2d}{C_\tau}\max\left\{n, \frac{\epsilon}{C_\tau d}\right\}\right)$ for $\delta = 0$. For $\delta \in (0,\frac{1}{2})$, we get $\mathbb{E}_{\mathcal{A}}F_\tau(w_{\mathcal{A}}^\delta(X),X) - F_\tau(w^*(X),X) \leq 9\frac{L^2C_\tau}{\mu}\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon n}$ in runtime $O\left(\frac{nd}{C_\tau}\max\left\{\frac{n^2}{C_\tau}, \frac{\epsilon n}{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}\right\}\right)$.*

Note that the boundedness condition is not very restrictive; indeed, it is automatic if $\mathcal{X}$ is compact (e.g. if data is normalized) and f is continuous, by the extreme value theorem.

Smooth, strongly convex, Lipschitz $f(\cdot,x)$ If we assume additionally that $f(\cdot,X)$ is β-smooth, then the below Lemma implies that $F_\tau \in \mathcal{H}_{\beta,\mu,L,R}$:

Lemma 4.2 *Assume $f(\cdot,x_i)$ is β-smooth and L-Lipschitz for all $i \in [n]$. Then for any $\tau > 0$, the TERM objective $F_\tau(\cdot,X)$ is β_τ-smooth for $X = (x_1,\cdots,x_n)$, where $\beta_\tau := \beta + L^2\tau$.*

Then denoting $\kappa_\tau = \frac{\beta+L^2\tau}{\mu}$ and appealing to Theorem 2.2 gives

$$\mathbb{E}_{\mathcal{A}}F_\tau(w_{\mathcal{A}}(X),X) - F_\tau(w^*(X),X) \leq 4\kappa_\tau\frac{L^2}{\mu}\left(\frac{d}{\epsilon}\right)^2,$$

if $\delta = 0$. For $\delta > 0$, we have $\mathbb{E}_{\mathcal{A}}F_\tau(w_{\mathcal{A}}^\delta(X),X) - F_\tau(w^*(X),X) \leq 4\kappa_\tau\frac{L^2}{\mu}\left(\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon}\right)^2$.

In both cases, the runtime is $\widetilde{O}(nd\sqrt{\kappa_\tau})$, where $\kappa_\tau = \frac{\beta+L^2\tau}{\mu}$. These bounds also hold as $\tau \to \infty$.

If we assume again that f is also bounded on $B(0,R)\times\mathcal{X}$, then we can obtain the following more refined bounds:

Proposition 4.2 *Assume $f(\cdot,x_i)$ is L-Lipschitz on $B(0,R)$, β-smooth, twice differentiable, and μ-strongly convex. Assume further that $a_R \leq f(w,x) \leq A_R$ for all $w \in B(0,R)$ and all $x \in \mathcal{X}$. Denote $C_\tau = e^{\tau(A_R-a_r)}$. Then running Algorithm 1 with $\mathcal{M}$ as AGD yields $\mathbb{E}_{\mathcal{A}}F_\tau(w_{\mathcal{A}}(X),X) - F_\tau(w^*(X),X) \leq 26\kappa_\tau\frac{L^2C_\tau}{\mu}\left(\frac{d}{\epsilon n}\right)^2$ if $\delta = 0$, and $\mathbb{E}_{\mathcal{A}}F_\tau(w_{\mathcal{A}}^\delta(X),X) - F_\tau(w^*(X),X) \leq 13.5\kappa_\tau\frac{L^2C_\tau}{\mu}\left(\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)^2$ if $\delta \in (0,\frac{1}{2})$. Both of these bounds are realized in runtime $\widetilde{O}(nd\sqrt{\kappa_\tau})$.*

Note that we nearly (up to a factor of κ) recover the optimal ERM excess risk bounds of (Bassily et al. 2014) as $\tau \to 0$.

Differentially Private Adversarial Training

In recent years, adversarial attacks on neural networks have raised significant concerns on reliability of these methods in critical applications. Adversarial attacks are input examples to a machine learning model crafted by making small perturbations to legitimate inputs to mislead the network. These adversarial examples lead to highly confident, but incorrect outputs; see e.g. (Szegedy et al. 2014; Goodfellow et al. 2015; Papernot et al. 2016; Eykholt et al. 2018; Moosavi-Dezfooli et al. 2016) and the references therein. A natural approach to training a model that is robust to such adversarial attacks is to solve the following adversarial training problem (Madry et al. 2018; Zhang et al. 2019; Nouiehed et al. 2019):

$$\min_{w\in\mathbb{R}^d} \max_{\mathbf{v}\in S^n} F(w, X+\mathbf{v}, \mathbf{y}). \tag{2}$$

Here $X = (x_1, \cdots, x_n) \in \mathcal{X}^n$ contains the feature data, $\mathbf{y} \in \mathcal{Y}^n$ is the corresponding label/target vector (e.g. $y_i \in \{0, 1\}$ for a binary classification task), and S is a set of permissible adversarial perturbations. As discussed in the Introduction, solving Eq. (2) with a practical differentially private algorithm to ensure privacy, robustness, and computational speed simultaneously is an important open problem. Indeed, we are not aware of any works that provide privacy, robustness ("excess adversarial risk"), and runtime guarantees for solving Eq. (2). In this section, we illustrate how our methods and results developed so far can be easily applied to establish excess adversarial risk bounds (robustness guarantees) and runtime bounds via output perturbation.

Notation and preliminaries Let $D = (X, \mathbf{y}) = ((x_1, y_1), \cdots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ be a given training data set. Assume that the adversarial perturbation set S *is convex and compact with* L_2 *diameter* ρ and that the adversary chooses $\mathbf{v} = (v_1, \cdots, v_n) \in S^n$ corresponding to the training examples $((x_1, y_1), \cdots, (x_n, y_n))$. The following definition is for notational convenience:

Definition 1 *For a loss function* $F : \mathbb{R}^d \times (\mathcal{X} + S)^n \times \mathcal{Y}^n$, *and dataset* $D = (X, \mathbf{y}) \in (\mathcal{X} \times \mathcal{Y})^n$, *denote the function of the weights and perturbations corresponding to* F *by*

$$H_D : \mathbb{R}^d \times S^n \to \mathbb{R},\ H_D(w, \mathbf{v}) := F(w, X + \mathbf{v}, \mathbf{y}).$$

We make the following additional assumptions:

Assumption 1 $H_D(\cdot, \mathbf{v}) \in \mathcal{F}_{\mu,L,R}$ *for all* $\mathbf{v} \in S^n$ *and all* $D \in (\mathcal{X} \times \mathcal{Y})^n$.

Assumption 2 $H_D(w, \cdot)$ *is continuous and concave for all* $w \in B(0, R)$ *and all* $D \in (\mathcal{X} \times \mathcal{Y})^n$.

Note that Assumption 2 is a standard assumption in the min-max literature (Nouiehed et al. 2019; Lin et al. 2020). Together with Assumption 1, it ensures the existence of a saddle point (see definition below) and enables us to (approximately) find the saddle point and implement our output perturbation method efficiently. Next, we recall a basic notion from min-max optimization:

Definition 2 *For* $\alpha \geq 0$, *say a point* $(\widehat{w}, \widehat{\mathbf{v}}) \in \mathbb{R}^d \times S^n$ *is an* α***-saddle point*** *of a convex (in* w*)-concave (in* $\mathbf{v}$*) function* $H(w, \mathbf{v})$ *if*

$$\max_{\mathbf{v}\in S^n} H(\widehat{w}, \mathbf{v}) - \min_{w\in\mathbb{R}^d} H(w, \widehat{\mathbf{v}}) \leq \alpha.$$

Observe that by Assumption 1, Assumption 2, and convexity and compactness of $B(0, R)$ and S, there exists at least one α-saddle point point $(\widehat{w}, \widehat{\mathbf{v}}) \in B(0, R) \times S^n$ of H_D for any $\alpha \geq 0$. Also, if we denote

$$G_D(w) := \max_{\mathbf{v}\in S^n} H_D(w, \mathbf{v}) = \max_{\mathbf{v}\in S^n} F(w, X + \mathbf{v}, \mathbf{y}),$$

which is in $\mathcal{F}_{\mu,L,R}$ (by Assumption 1 and compactness of S^n), and

$$w^*(D) := \operatorname*{argmin}_{w\in\mathbb{R}^d} G_D(w) = \operatorname*{argmin}_{w\in\mathbb{R}^d} \max_{\mathbf{v}\in S} H_D(w, \mathbf{v}) \in B(0, R),$$

then for any α-saddle point $(\widehat{w}, \widehat{\mathbf{v}})$ of H_D, we have:

$$G_D(\widehat{w}) - G_D(w^*(D)) \leq \alpha,$$

by Sion's minimax theorem (Sion 1958).

For a model $w_{\mathcal{A}}$ trained on loss function F (by some randomized algorithm $\mathcal{A}$), the measure of adversarial robustness that we will consider is:

Definition 3 *Let* $w_{\mathcal{A}}$ *be the output of a randomized algorithm* $\mathcal{A}$ *for solving Eq.* (2). *Define the* ***excess adversarial risk*** *of* $\mathcal{A}$ *by*

$$\mathbb{E}_{\mathcal{A}} \max_{\mathbf{v}\in S^n} F(w_{\mathcal{A}}, X + \mathbf{v}, y) - \min_{w\in\mathbb{R}^d} \max_{\mathbf{v}\in S^n} F(w, X + \mathbf{v}, \mathbf{y}) = \mathbb{E}_{\mathcal{A}} G_D(w_{\mathcal{A}}) - G_D(w^*(D)).$$

In what follows, we aim to quantify the tradeoffs between excess adversarial risk, privacy, and runtime for output perturbation. In order to practically implement the output perturbation mechanism, we make the following additional assumption:

Assumption 3 $H_D(w, \cdot)$ *is* β_v*-smooth as a function of* $\mathbf{v} \in S^n$ *for all* $w \in B(0, R)$ *and all* $D = (X, \mathbf{y}) \in (\mathcal{X} \times \mathcal{Y})^n$.

Then Eq. (2) is a smooth (in w and v), strongly convex-concave min-max problem and there are efficient non-private algorithms for solving such problems (Nesterov and Scrimali 2007; Alkousa et al. 2020; Lin et al. 2020). In Algorithm 3, we give a Black Box Algorithm tailored to the special min-max structure of the F that is present in the adversarial training setting. We then instantiate the Algorithm with the near-optimal (in terms of gradient complexity) algorithms of (Lin et al. 2020) and provide upper bounds on the runtime for privately optimizing the adversarially robust model.

Definition 4 *For* $\alpha \geq 0$, *a point* $(\widehat{w}, \widehat{v}) \in B(0, R) \times B_\gamma$ *is an* α*-saddle point of function* $F(w, X + v, y)$ *if*

$$\max_{v\in B_\gamma} F(\widehat{w}, X + v, y) - \min_{w\in B(0,R)} F(w, X + \widehat{v}, y) \leq \alpha.$$

The same arguments used before show that the Algorithm 3 is (ϵ, δ)-differentially private. Furthermore, instantiating Algorithm 3 with the Minimax-AIPP algorithm (Lin et al. 2020) leads to the following guarantees:

Algorithm 3 Black Box Output Perturbation Algorithm for Implementing DP Adversarial Training

Require: Number of data points $n \in \mathbb{N}$, dimension $d \in \mathbb{N}$ of feature data, non-private (possibly randomized) optimization method $\mathcal{M}$, privacy parameters $\epsilon > 0, \delta \geq 0$, feature data universe $\mathcal{X} \subseteq \mathbb{R}^d$ and label universe $\mathcal{Y}$, data set $D = (X, \mathbf{y}) \in (\mathcal{X} \times \mathcal{Y})^n$, convex compact perturbation set $S \subset \mathbb{R}^d$ of diameter $\rho \geq 0$, loss function $F(w, X, \mathbf{y})$ that is L-Lipschitz and μ-strongly convex in w on $B(0, R)$ and concave in $\mathbf{v} \in S^n$, accuracy parameter $\alpha > 0$ with corresponding iteration number $T = T(\alpha) \in \mathbb{N}$ such that (w_T, v_T) is an α-saddle point of $H_D(w, \mathbf{v})$.

1: Run $\mathcal{M}$ for $T = T(\alpha)$ iterations to obtain an α-saddle point (w_T, v_T) of $H_D(w, \mathbf{v}) = F(w, X + \mathbf{v}, \mathbf{y})$.

2: Add noise to ensure differential privacy: $w_{\mathcal{A}}(D) := \Pi_{B(0,R)}(w_T + \widehat{z})$, where the density of $\widehat{z}$ is given as: $p_{\widehat{z}}(t) \propto \begin{cases} \exp\left(-\frac{\epsilon \|t\|_2}{\Delta + 2\sqrt{\frac{2\alpha}{\mu}}}\right) & \text{if } \delta = 0 \\ \exp\left(-\frac{\epsilon^2 \|t\|_2^2}{\left(\Delta + \frac{2\alpha}{\mu}\right)^2 \left(c_\delta + \sqrt{c_\delta^2 + \epsilon}\right)^2}\right) & \text{if } \delta \in \left(0, \frac{1}{2}\right), \end{cases}$ and $\Delta := \begin{cases} \frac{2L}{\mu} & \text{if } F \text{ is not of ERM form} \\ \frac{2L}{\mu n} & \text{if } F \text{ is of ERM form.} \end{cases}$

3: **return** $w_{\mathcal{A}}(D)$.

Theorem 4.1 *Let $D = (X, \mathbf{y}) \in (\mathcal{X} \times \mathcal{Y})^n$ be a dataset and let $\epsilon > 0, \delta \in [0, \frac{1}{2})$. Assume that S is a convex compact set in $\mathbb{R}^d$ of L_2 diameter ρ and that the loss function F is such that Assumption 1, Assumption 2, and Assumption 3 hold (see Definition 1). Furthermore, assume that $H_D(\cdot, v)$ is β-smooth for all $\mathbf{v} \in S^n$ with condition number $\kappa = \beta/\mu$. Run Algorithm 3 with $\mathcal{M} =$ Minimax-AIPP.*

1. Suppose F is not of ERM form.

a) Let $\delta = 0$ and $\frac{d}{\epsilon} \leq 1$. Setting $\alpha = \frac{L^2}{\mu} \min\{\kappa \left(\frac{d}{\epsilon}\right)^2, 1\}$ yields $\mathbb{E}_{\mathcal{A}} G_D(w_{\mathcal{A}}) - \min_{w \in \mathbb{R}^d} G_D(w) \leq 26\kappa \frac{L^2}{\mu} \left(\frac{d}{\epsilon}\right)^2$ in $T = \widetilde{O}\left(\sqrt{\frac{\beta\beta_v}{L^2 \min\left\{\kappa\left(\frac{d}{\epsilon}\right)^2, 1\right\}}} \sqrt{n}\rho\right)$ gradient evaluations and runtime $\widetilde{O}\left(nd\sqrt{\frac{\beta\beta_v}{L^2 \min\left\{\kappa\left(\frac{d}{\epsilon}\right)^2, 1\right\}}} \sqrt{n}\rho\right)$.

b) Let $\delta \in \left(0, \frac{1}{2}\right)$ and $\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon} \lesssim 1$. Setting $\alpha = \frac{L^2}{\mu} \min\left\{\kappa \left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon}\right)^2, 1\right\}$ gives $\mathbb{E}_{\mathcal{A}} G_D(w_{\mathcal{A}}^\delta) - \min_{w \in \mathbb{R}^d} G_D(w) \leq 13.5\kappa \frac{L^2}{\mu} \left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon}\right)^2$ in $T = \widetilde{O}\left(\sqrt{\frac{\beta\beta_v}{L^2 \min\left\{\kappa\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon}\right)^2, 1\right\}}} \sqrt{n}\rho\right)$ gradient evaluations and runtime $\widetilde{O}\left(nd\sqrt{\frac{\beta\beta_v}{L^2 \min\left\{\kappa\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon}\right)^2, 1\right\}}} \sqrt{n}\rho\right)$.

2. Suppose F is of ERM form.

a) Let $\delta = 0$ and $\frac{d}{\epsilon n} \leq 1$. Setting $\alpha = \frac{L^2}{\mu} \frac{1}{n^2} \min\{\kappa \left(\frac{d}{\epsilon}\right)^2, 1\}$ yields $\mathbb{E}_{\mathcal{A}} G_D(w_{\mathcal{A}}) - \min_{w \in \mathbb{R}^d} G_D(w) \leq 26\kappa \frac{L^2}{\mu} \left(\frac{d}{\epsilon n}\right)^2$ in $T = \widetilde{O}\left(n^{3/2} \sqrt{\frac{\beta\beta_v}{L^2 \min\left\{\kappa\left(\frac{d}{\epsilon}\right)^2, 1\right\}}} \rho\right)$ gradient evaluations and runtime $\widetilde{O}\left(n^{5/2} d \sqrt{\frac{\beta\beta_v}{L^2 \min\left\{\kappa\left(\frac{d}{\epsilon}\right)^2, 1\right\}}} \rho\right)$.

b) Let $\delta \in \left(0, \frac{1}{2}\right)$ and $\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon n} \leq 1$. Setting $\alpha = \frac{L^2}{\mu} \frac{1}{n^2} \min\left\{\kappa \left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon}\right)^2, 1\right\}$ gives $\mathbb{E}_{\mathcal{A}} G_D(w_{\mathcal{A}}^\delta) - \min_{w \in \mathbb{R}^d} G_D(w) \leq 13.5\kappa \frac{L^2}{\mu} \left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon n}\right)^2$ in $T = \widetilde{O}\left(n^{3/2} \sqrt{\frac{\beta\beta_v}{L^2 \min\left\{\kappa\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon}\right)^2, 1\right\}}} \rho\right)$ gradient evaluations and runtime $\widetilde{O}\left(n^{5/2} d \sqrt{\frac{\beta\beta_v}{L^2 \min\left\{\kappa\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon}\right)^2, 1\right\}}} \rho\right)$.

If, in addition, $H_D(w, \cdot)$ is μ_v-strongly concave in v for all $w \in B(0, R)$, then the above bounds are all attained and the gradient complexity improves to $T = \widetilde{O}\left(\sqrt{\kappa\kappa_v}\right)$ and runtime improves to $\widetilde{O}\left(nd\sqrt{\kappa\kappa_v}\right)$.

Note also that a similar result for $F \in \mathcal{F}_{\mu,L,R}$ can easily be written down as a consequence of our earlier results. Excess adversarial population loss bounds can also be derived from our DP-SCO results.

5 Conclusion

In this work, we highlighted the importance of differentially private optimization for general non-ERM loss functions, and provided a simple yet practical algorithm for addressing this problem, along with excess risk and runtime bounds. We also used our method to obtain tight population loss and runtime bounds for the differentially private SCO problem, where, unlike prior works, we guarantee (ϵ, δ)-DP for all $\delta \geq 0$. Finally, we applied our results to two practical applications in machine learning: DP Tilted ERM and DP Adversarial Training. An interesting question for future work is whether our excess risk bounds for general non-ERM loss are tight. Currently, the only known DP excess risk lower

bounds are for ERM functions (Bassily et al. 2014).

Acknowledgments

We would like to thank Zeyuan Allen-Zhu, Larry Goldstein, and Adam Smith for helpful comments.

References

Mohammad Alkousa, Darina Dvinskikh, Fedor Stonyakin, Alexander Gasnikov, and Dmitry Kovalev. Accelerated methods for composite non-bilinear saddle point problem, 2020.

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(221):1–51, 2018. URL http://jmlr.org/papers/v18/16-410.html.

Aleksandr Aravkin, James Burke, and Dmitry Drusvyatskiy. *Convex Analysis and Nonsmooth Optimization*. 2017.

Raman Arora, Teodor V. Marinov, and Enayat Ullah. Private stochastic convex optimization: Efficient algorithms for non-smooth objectives. *arXiv:2002.09609v1*, 2020.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Differentially private empirical risk minimization: Efficient algorithms and tight error bounds. *arXiv:1405.7085v2*, 2014.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in Neural Information Processing Systems*, 32:11282–11291, 2019.

Olivier Bousquet and Andre Elisseeff. Stability and generalization. *Journal of machine learning research*, 2002.

Stephen Boyd, Corinna Cortes, Mehryar Mohri, and Ana Radovanovic. Accuracy at the top. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, pages 953–961. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/7fe1f8abaad094e0b5cb1b01d712f708-Paper.pdf.

Sebastian Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8, 2015.

Kamalika Chaudhuri, Claire Monteleoni, and Anand Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

Nadav Cohen and Amnon Shashua. Simnets: A generalization of convolutional networks. *arXiv:1410.0781v3*, 2014.

Nadav Cohen, Or Sharir, and Amnon Shashua. Deep simnets. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4782–4791, 2016. doi: 10.1109/CVPR.2016.517.

Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on privacy enhancing technologies*, 2015(1):92–112, 2015.

Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*. 2014.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology - EUROCRYPT 2006*, volume 4004 of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, 2006.

Kevin Eykholt, Ivan Evtimov, Earlance Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018. doi: 10.1109/CVPR.2018.00175.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in linear time. *arXiv:2005.04763v1*, 2020.

Ferdinando Fioretto, Terrence WK Mak, and Pascal Van Hentenryck. Differential privacy for power grid obfuscation. *IEEE Transactions on Smart Grid*, 11(2):1356–1366, 2019.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1412.6572.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1225–1234, New York, New York, USA, 20–22 Jun 2016. PMLR. URL http://proceedings.mlr.press/v48/hardt16.html.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 2014.

Angelos Katharopoulos and François Fleuret. Biased importance sampling for deep neural network training. *arXiv:1706.00043*, 2017.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 25.1–25.40, Edinburgh, Scotland, 25–27 Jun 2012. PMLR. URL http://proceedings.mlr.press/v23/kifer12.html.

Barry W. Kort and Dimitri P. Bertsekas. A new penalty function method for constrained minimization. In *Proceedings of the 1972 IEEE Conference on Decision and Control and 11th Symposium on Adaptive Processes*, pages 162–166, 1972. doi: 10.1109/CDC.1972.268971.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. *arXiv:2007.01162v1*, 2020.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. *Advances in neural information processing systems*, 28:3384–3392, 2015.

Tianyi Lin, Chi Jin, and Michael I. Jordan. Near-optimal algorithms for minimax optimization. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2738–2779. PMLR, 09–12 Jul 2020. URL http://proceedings.mlr.press/v125/lin20a.html.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016. doi: 10.1109/CVPR.2016.282.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014. ISBN 1461346916.

Yurii Nesterov and Laura Scrimali. Solving strongly monotone variational and quasi-variational inequalities. *Discrete and Continuous Dynamical Systems*, 31, 01 2007. doi: 10.2139/ssrn.970903.

Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems 32*, pages 14934–14942. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9631-solving-a-class-of-non-convex-min-max-games-using-iterative-first-order-methods.pdf.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS P)*, pages 372–387, 2016. doi: 10.1109/EuroSP.2016.36.

E.Y. Pee and J. O. Royset. On solving large-scale finite minimax problems using exponential smoothing. *Journal of Optimization Theory and Applications*, 148(2):390–421, February 2011. URL https://ideas.repec.org/a/spr/joptap/v148y2011i2d10.1007_s10957-010-9759-1.html.

Hai Phan, My T Thai, Han Hu, Ruoming Jin, Tong Sun, and Dejing Dou. Scalable differential privacy with certified robustness in adversarial learning. In *International Conference on Machine Learning*, pages 7683–7694. PMLR, 2020.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, pages 171–176, 1958.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL http://arxiv.org/abs/1312.6199.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30:2722–2731, 2017.

Jun Wang, Rongbo Zhu, Shubo Liu, and Zhaohui Cai. Node location privacy protection based on differentially private grids in industrial wireless sensor networks. *Sensors*, 18(2): 410, 2018.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/zhang19p.html.

Jiaqi Zhang, Kai Zheng, Wenlong Mou, and Liwei Wang. Efficient private erm for smooth objectives. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3922–3928, 2017.

Jun Zhao, Teng Wang, Tao Bai, Kwok-Yan Lam, Zhiying Xu, Shuyu Shi, Xuebin Ren, Xinyu Yang, Yang Liu, and Han Yu. Reviewing and improving the gaussian mechanism for differential privacy. *arXiv:1911.12060v2*, 2019.

Output Perturbation for General Differentially Private Convex Optimization: Supplemental Material

Andrew Lowy[1] **and Meisam Razaviyayn** [2]
[1, 2]University of Southern California
[1] lowya@usc.edu and [2] razaviya@usc.edu

Proofs of Section 2 Results

We will require a definition and a lemma for several proofs.

Definition 1 *Define the L_2* ***sensitivity*** *of a (strongly convex in w) function $F : \mathbb{R}^d \times \mathcal{X}^n \to \mathbb{R}$ as*

$$\Delta_F := \sup_{X, X' \in \mathcal{X}^n, |X \Delta X'| \leq 2} \|w^*(X) - w^*(X')\|_2,$$

where $w^*(X) = \arg\min_{w \in \mathbb{R}^d} F(w, X)$.

The following result generalizes Corollary 8 in (Chaudhuri et al. 2011) (beyond smooth linear ERM classifiers with bounded label space).

Lemma 0.1 *For any $F \in \mathcal{F}_{\mu,L,R}$, $\Delta_F \leq \frac{2L}{\mu}$. For $F \in \mathcal{F}_{\mu,L}^{ERM}$, $\Delta_F \leq \frac{2L}{\mu n}$.*

The proof relies on the following generalization (to non-differentiable functions on a possibly constrained ($\mathcal{W} \neq \mathbb{R}^d$) domain) of Lemma 7 from (Chaudhuri et al. 2011).

Lemma 0.2 *Let $G(w), g(w)$ be continuous convex functions on some convex closed set $\mathcal{W} \subseteq \mathbb{R}^p$ and suppose that $G(w)$ and $G(w) + g(w)$ are μ-strongly convex. Assume further that g is L_g-Lipschitz.*
Define $w_1 = \arg\min_{w \in \mathcal{W}} G(w)$ and $w_2 = \arg\min_{w \in \mathcal{W}} [G(w) + g(w)]$. Then $\|w_1 - w_2\|_2 \leq \frac{L_g}{\mu}$.

Proof: At a point $w \in \mathcal{W}$, let $h(w)$ and $H(w)$ denote subgradients of g and G, respectively. By first-order optimality conditions, for all $w \in \mathcal{W}$, we have

$$\langle H(w_1), (w - w_1) \rangle \geq 0$$

and

$$\langle H(w_2), (w - w_2) \rangle + \langle h(w_2), (w - w_2) \rangle \geq 0.$$

Plugging w_2 for w in the first inequality and w_1 for w in the second inequality, and then subtracting gives:

$$\langle H(w_1) - H(w_2), w_1 - w_2 \rangle \leq \langle h(w_2), w_1 - w_2 \rangle. \quad (1)$$

Now, by strong convexity of G, we have

$$\mu \|w_1 - w_2\|_2^2 \leq \langle H(w_1) - H(w_2), w_1 - w_2 \rangle.$$

Combining this with Eq. (1) and using Cauchy-Schwartz yields:

$$\begin{aligned}\mu \|w_1 - w_2\|_2^2 &\leq \langle H(w_1) - H(w_2), w_1 - w_2 \\ &\leq \langle h(w_2), w_1 - w_2 \rangle \leq \|h(w_2)\|_2 \|w_1 - w_2\|_2.\end{aligned}$$

Finally, using L_g-Lipschitzness of g and dividing the above inequality by $\mu \|w_1 - w_2\|_2$ gives the Lemma.

Now we can prove Lemma 0.1: Let $X, X' \in \mathcal{X}^n$ such that (WLOG) $x_n \neq x'_n$, but all other data points are the same. Apply Lemma 0.2 to $G(w) = F(w, X), g(w) = F(w, X') - F(w, X)$ and note that g is $2L$-Lipschitz by the triangle inequality. For F of ERM form, we have $g(w) = \frac{1}{n}[f(w, x'_n) - f(w, x_n)]$, which is $2\frac{L}{n}$-Lipschitz. This completes the proof.

Proof of Proposition 2.1

Define $\Delta_T := \sup_{|X \Delta X'| \leq 2} \|w_T(X) - w_T(X')\|_2$. First, we show that $\Delta_T \leq \Delta_F + 2\sqrt{\frac{2\alpha}{\mu}}$. Now, note that since $F(w_T(X), X) - F(w^*, X) \leq \alpha$ by the choice of $T = T(\alpha)$, and since $F(\cdot, X)$ is μ-strongly convex, we have

$$\begin{aligned}\|w_T - w^*\|_2^2 \leq \frac{2}{\mu}[F(w_T, X) - F(w^*, X) - \\ \langle \nabla F(w^*, X), w_T - w^* \rangle] \leq \frac{2}{\mu}\alpha.\end{aligned}$$

Hence for any data sets $X, X' \in \mathcal{X}^n$ such that $|X \Delta X'| \leq 2$, we have

$$\begin{aligned}\|w_T(X) - w_T(X')\| &\leq \|w_T(X) - w^*(X)\| \\ &\quad + \|w^*(X) - w^*(X')\| \\ &\quad + \|w^*(X') - w_T(X')\| \\ &\leq \sqrt{\frac{2}{\mu}\alpha} + \Delta_F + \sqrt{\frac{2}{\mu}\alpha}.\end{aligned}$$

Now recall the well-known post-processing property of differential privacy, which states that any function (e.g. projecting onto $\mathcal{W}$) of an (ϵ, δ)- differentially private method is itself (ϵ, δ)- differentially private (Dwork and Roth 2014). By this fact, it suffices to show that the algorithm $w_{\mathcal{A}'}(X) := w_T(X) + \hat{z}$ (without projection) is differentially private. Assume first $\delta = 0$. Then by the definition of differential privacy, it suffices to show that for any

$s \in \text{range}(w_{\mathcal{A}'})$ (for which $p'(s) \neq 0$) and any $X, X' \in \mathcal{X}^n$ such that $|X \Delta X'| \leq 2$,

$$\frac{p(s)}{p'(s)} \leq e^{\epsilon},$$

where p and p' are the probability density functions (pdfs) of $\mathcal{A}'(X)$ and $\mathcal{A}'(X')$ respectively. Now note that $p(s) = p_z(s - w_T(X))$ and $p'(s) = p_z(s - w_T(X'))$, where p_z is the pdf of the noise. Then

$$\begin{aligned}\frac{p(s)}{p'(s)} &= \frac{p_z(s - w_T(X))}{p_z(s - w_T(X'))} \\ &= \exp\left(\frac{-\epsilon\|s - w_T(X)\|_2 + \epsilon\|s - w_T(X')\|_2}{\Delta_T}\right) \\ &\leq \exp\left(\frac{\epsilon\|w_T(X) - w_T(X')\|_2}{\Delta_T}\right) \leq \exp(\epsilon),\end{aligned}$$

where the second to last line uses the reverse triangle inequality and the last line uses the definition of Δ_T. Now, by the post-processing property (Dwork and Roth 2014), we conclude that the algorithm $\mathcal{A}(X) = w_{\mathcal{A}}(X) = \Pi_{\mathcal{W}}(\mathcal{A}'(X))$ is $(\epsilon, 0)$- differentially private.

For $\delta > 0$, the proof follows similarly, using the following result of (Zhao et al. 2019):

Theorem 0.1 *(Theorem 5 in (Zhao et al. 2019)) For $\delta \in (0, .5)$, (ϵ, δ)- differentially privacy can be achieved by adding Gaussian noise with mean 0 and standard deviation $\sigma = \frac{(c+\sqrt{c^2+\epsilon})\Delta}{\epsilon\sqrt{2}}$ to each dimension of a query with L_2 sensitivity Δ. Here $c = \sqrt{\log\left(\frac{2}{\sqrt{16\delta+1}-1}\right)}$.*

Here, our "query" is $w_T(X)$, which has sensitivity Δ_T, and hence $\mathcal{A}'$ (as defined above) is (ϵ, δ). Applying the post-processing property again shows that $\mathcal{A}$ is (ϵ, δ)-differentially private.

Proof of Theorem 2.1

The result is a corollary of (the first part of) the following:

Theorem 0.2 *Run Algorithm 1 with arbitrary inputs.*

1. *Suppose $F \in \mathcal{F}_{\mu,L,R}$.*
   a) Let $\delta = 0$, $\frac{d}{\epsilon} \leq 1$.
   Then $\mathbb{E}_{\mathcal{A}}F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 2\sqrt{2}\left(\frac{L^2}{\mu} + L\sqrt{\frac{2\alpha}{\mu}}\right)\left(\frac{d}{\epsilon}\right) + \alpha$.
   In particular, setting $\alpha = \frac{L^2}{\mu}\frac{d}{\epsilon}$ gives $\mathbb{E}_{\mathcal{A}}F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 9\frac{L^2}{\mu}\left(\frac{d}{\epsilon n}\right)$.
   b) Let $\delta > 0$, $\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon} \leq 1$.
   Then $\mathbb{E}_{\mathcal{A}}F(w_{\mathcal{A}}^\delta(X), X) - F(w^*(X), X) \leq 2\left(\frac{L^2}{\mu n} + L\sqrt{\frac{2\alpha}{\mu}}\right)\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon}\right) + \alpha$.
   In particular, setting $\alpha = \frac{L^2}{\mu n}\min\left\{\frac{1}{n}, \frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon}\right\}$ gives $\mathbb{E}_{\mathcal{A}}F(w_{\mathcal{A}}^\delta(X), X) - F(w^*(X), X) \leq 6\frac{L^2}{\mu}\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)$.
2. *Suppose $F \in \mathcal{F}_{\mu,L}^{ERM}$.*
   a) Let $\delta = 0$, $\frac{d}{\epsilon n} \leq 1$. Then $\mathbb{E}_{\mathcal{A}}F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 2\sqrt{2}\left(\frac{L^2}{\mu}\left(\frac{1}{n}\right) + L\sqrt{\frac{2\alpha}{\mu}}\right)\left(\frac{d}{\epsilon}\right) + \alpha$.
   In particular, setting $\alpha = \frac{L^2}{\mu n}\min\{\frac{1}{n}, \frac{d}{\epsilon}\}$ gives $\mathbb{E}_{\mathcal{A}}F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 9\frac{L^2}{\mu}\left(\frac{d}{\epsilon n}\right)$.
   b) Let $\delta > 0$, $\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon n} \leq 1$.
   Then $\mathbb{E}_{\mathcal{A}}F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 2\left(\frac{L^2}{\mu n} + L\sqrt{\frac{2\alpha}{\mu}}\right)\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon}\right) + \alpha$.
   In particular, setting $\alpha = \frac{L^2}{\mu n}\min\left\{\frac{1}{n}, \frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon}\right\}$ gives $\mathbb{E}_{\mathcal{A}}F(w_{\mathcal{A}}^\delta(X), X) - F(w^*(X), X) \leq 6\frac{L^2}{\mu}\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)$.

To prove this, we will need to recall the following basic fact about projections onto closed, convex sets:

Lemma 0.3 *Let $\mathcal{W}$ be a closed, convex set in $\mathbb{R}^d$, and let $a \in \mathbb{R}^d$. Then $\pi = \Pi_{\mathcal{W}}(a)$ if and only if $\langle a - \pi, w - \pi\rangle \leq 0$ for all $w \in \mathcal{W}$.*

We include the proof for completeness: Fix any $a \in \mathbb{R}^d$. By definition, $\Pi_{\mathcal{W}}(a) = \arg\min_{w\in\mathcal{W}} \frac{1}{2}\|w - a\|_2^2 := \arg\min_{w\in\mathcal{W}} h(w)$. Then $\nabla h(w) = w - a$ and by first-order optimality conditions, $\pi = \arg\min_{w\in\mathcal{W}} h(w)$ if and only if for all $w \in \mathcal{W}$,

$$\langle \nabla h(\pi), w - \pi\rangle \geq 0 \Leftrightarrow \langle \pi - a, w - \pi\rangle \geq 0 \Leftrightarrow \langle a - \pi, w - \pi\rangle \leq 0.$$

Now we can prove the Theorem: Notice

$$\begin{aligned}&\mathbb{E}_{\mathcal{A}}F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) = \\ &\mathbb{E}F(\Pi_{\mathcal{W}}(w_T(X) + \hat{z}), X) - F(w^*(X), X) \\ &= \mathbb{E}[\Pi_{\mathcal{W}}(w_T(X) + \hat{z}), X) - F(w_T(X), X)] \\ &+ \mathbb{E}[F(w_T(X), X) - F(w^*(X), X)] \\ &\leq L\mathbb{E}\|\hat{z}\|_2 + \alpha,\end{aligned}$$

where we used Lemma 0.3 in the last inequality. Now for $\delta = 0$, $\mathbb{E}_{\mathcal{A}}F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq \sqrt{2}L(\Delta_F + 2\sqrt{\frac{2\alpha}{\mu}})\left(\frac{d}{\epsilon}\right) + \alpha$. Then using Lemma 0.1 to bound Δ_F proves the first statement in each of 1a) and 2a). Similarly, for $\delta > 0$, we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{A}}F(w_{\mathcal{A}}^\delta(X), X) - F(w^*(X), X) \leq \\ &L\left(\Delta_F + 2\sqrt{\frac{2\alpha}{\mu}}\right)\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon}\right) + \alpha.\end{aligned}$$

Appealing again to Lemma 0.1 completes the proof. The verification that the prescribed choices of α achieve the respective fundamental upper bounds is routine, using the assumptions stated in the theorem.

Next, we recall the following convergence guarantee for strongly convex, Lipschitz objective:

Proposition 0.1 *(Aravkin et al. 2017, Proposition 5.3) For $F \in \mathcal{F}_{\mu,L,R}$, running the subgradient method for $T = \frac{2L^2}{\mu\alpha}$ iterations and step sizes $\eta_t = \frac{2}{\mu(t+1)}$ results in a point $\hat{w_T} \in \mathcal{W}$ such that $F(\hat{w_T}, X) - F(w^*(X), X) \leq \alpha$. Here $\hat{w_T} = \operatorname{argmin}_{w_t, t\in[T]} F(w_t, X)$.*

Then Theorem 2.1 is a direct consequence of Proposition 0.1 and Theorem 0.2.

Proof of Theorem 2.2

We begin by proving an analogue of Theorem 0.2 for β-smooth loss functions.

Theorem 0.3 *Run Algorithm 1 on $F \in \mathcal{H}_{\beta,\mu,L,R}$ with arbitrary inputs.*

1. *Suppose $F \in \mathcal{H}_{\beta,\mu,L,R}$.*
   a) Let $\delta = 0$, $\frac{d}{\epsilon} \leq 1$. Then $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 4\beta \left(\frac{L}{\mu} + \sqrt{\frac{2\alpha}{\mu}}\right)^2 \left(\frac{d}{\epsilon}\right)^2 + \alpha$.
   In particular, setting $\alpha = \frac{L^2}{\mu} \min\left\{\kappa \left(\frac{d}{\epsilon}\right)^2, 1\right\}$ gives $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 26\kappa \frac{L^2}{\mu} \left(\frac{d}{\epsilon}\right)^2$.
   b) Let $\delta \in (0, \frac{1}{2})$, $\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon} \leq 1$.
   Then $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}^\delta(X), X) - F(w^*(X), X) \leq 2\beta \left(\frac{L}{\mu} + \sqrt{\frac{2\alpha}{\mu}}\right)^2 \left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon}\right)^2 + \alpha$.
   In particular, setting $\alpha = \frac{L^2}{\mu} \min\left\{\kappa \left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon}\right)^2, 1\right\}$ gives $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}^\delta(X), X) - F(w^*(X), X) \leq 13.5\kappa \frac{L^2}{\mu} \left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon}\right)^2$.

2. *Suppose $F \in \mathcal{H}^{ERM}_{\beta,\mu,L,R}$.*
   a) Let $\delta = 0$, $\frac{d}{\epsilon n} \leq 1$. Then $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 4\beta \left(\frac{L}{\mu n} + \sqrt{\frac{2\alpha}{\mu}}\right)^2 \left(\frac{d}{\epsilon}\right)^2 + \alpha$.
   In particular, setting $\alpha = \frac{L^2}{\mu n^2} \min\left\{\kappa \left(\frac{d}{\epsilon}\right)^2, 1\right\}$ gives $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 26 \frac{L^2}{\mu} \kappa \left(\frac{d}{\epsilon n}\right)^2$.
   b) Let $\delta \in (0, \frac{1}{2})$, $\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon n} \leq 1$.
   Then $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}^\delta(X), X) - F(w^*(X), X) \leq 2\beta \left(\frac{L}{\mu n} + \sqrt{\frac{2\alpha}{\mu}}\right)^2 \left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon}\right)^2 + \alpha$.
   In particular, setting $\alpha = \frac{L^2}{\mu n^2} \min\left\{\kappa \left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon}\right)^2, 1\right\}$ gives $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}^\delta(X), X) - F(w^*(X), X) \leq 13.5 \frac{L^2}{\mu} \kappa \left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)^2$.

To prove Theorem 0.3, write

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) = \\
&\mathbb{E} F(\Pi_{\mathcal{W}}(w_T(X) + \hat{z}), X) - F(w^*(X), X) \\
&= \mathbb{E} F(\Pi_{\mathcal{W}}(w_T(X) + \hat{z}), X) - F(w_T(X), X) \\
&+ F(w_T(X), X) - F(w^*(X), X) \\
&\leq \frac{\beta}{2} \mathbb{E} \|\hat{z}\|_2^2 + \alpha,
\end{aligned}$$

where in the last line we used the descent lemma and Lemma 0.3. Now for $\delta = 0$,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq \\
&\beta \left(\Delta_F + 2\sqrt{\frac{2\alpha}{\mu}}\right)^2 \left(\frac{d}{\epsilon}\right)^2 + \alpha.
\end{aligned}$$

Substituting the bounds on Δ_F from Lemma 0.1 proves the first statement in each of 1a) and 2a). Similarly, for $\delta > 0$,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq \\
&\frac{\beta}{2} \left[\frac{d(c_\delta + \sqrt{c_\delta^2 + \epsilon})^2 \left(\Delta_F + 2\sqrt{\frac{2\alpha}{\mu}}\right)^2}{\epsilon^2}\right] + \alpha.
\end{aligned}$$

Again appealing to Lemma 0.1 establishes the first statements in 1b) and 2b). Verification that the prescribed choices of α achieve the claimed bounds is straightforward, completing the proof of Theorem 0.3.

Recall that running $T = \sqrt{\kappa} \log\left(\frac{(\mu+\beta)R^2}{2\alpha}\right) = \widetilde{O}(\sqrt{\kappa})$ iterations (each iteration involves a full gradient evaluation of F) of Nesterov's AGD is sufficient for finding $\hat{w}_T(X)$ such that $F(\hat{w}_T(X), X) - F(w^*(X), X) \leq \alpha$ (Bubeck 2015). Combining this with Theorem 0.3 yields Theorem 2.2.

Proof of Proposition 2.2

By standard arguments, it suffices to show that $\sup_{|X\Delta X'|\leq 2} \|w_T(X) - w_T(X')\|_2 \leq \Delta_\lambda + 2\sqrt{\frac{2\alpha}{\mu}}$. Let $|X \Delta X'| \leq 2$. Then

$$\begin{aligned}
\|w_T(X) - w_T(X')\|_2 \leq &\|w_T(X) - w_\lambda^*(X)\| + \\
&\|w_\lambda^*(X) - w_\lambda^*(X')\| + \\
&\|w_\lambda^*(X') - w_T(X')\| \\
&\leq \sqrt{\frac{2\alpha}{\lambda}} + \Delta_\lambda + \sqrt{\frac{2\alpha}{\lambda}}.
\end{aligned}$$

The last line follows because: λ-strong convexity of F_λ implies the middle term is bounded according to Lemma 0.1 and the other two terms are both bounded by a similar argument as in the proof of Proposition 2.1, again using λ-strong convexity of F_λ.

Proof of Theorem 2.3

As before, we first prove a general result for our black box Algorithm 2, whence Theorem 2.3 will follow as a corollary.

Theorem 0.4 *Run Algorithm 2 with $\lambda > 0$ chosen as follows: If $\delta = 0$, set*

$$\lambda := \begin{cases} \frac{L}{R\sqrt{1+\frac{\epsilon n}{d}}} & \text{if } \in \mathcal{G}_{L,R}^{ERM} \\ \frac{L}{R\sqrt{1+\frac{\epsilon}{d}}} & \text{if } \in \mathcal{G}_{L,R} \setminus \mathcal{G}_{L,R}^{ERM}. \end{cases}$$

If $\delta \in \left(0, \frac{1}{2}\right)$, set

$$\lambda := \begin{cases} \frac{L}{R\sqrt{1+\frac{\epsilon n}{d^{1/2}(c_\delta+\sqrt{c_\delta^2+\epsilon})}}} & \text{if } \in \mathcal{G}_{L,R}^{ERM} \\ \frac{L}{R\sqrt{1+\frac{\epsilon}{d^{1/2}(c_\delta+\sqrt{c_\delta^2+\epsilon})}}} & \text{if } \in \mathcal{G}_{L,R} \setminus \mathcal{G}_{L,R}^{ERM}. \end{cases}$$

1. Suppose $F \in \mathcal{G}_{L,R}$.

a) Let $\delta = 0, \frac{d}{\epsilon} \leq 1$. Then $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 16.5LR\sqrt{\frac{d}{\epsilon}} + 32\sqrt{\alpha}\sqrt{LR}(\frac{\epsilon}{d})^{1/4}$. In particular, setting $\alpha = LR\left(\frac{d}{\epsilon}\right)^{3/2}$ implies $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 49LR\sqrt{\frac{d}{\epsilon}}$.

b) Let $\delta \in (0, 1/2)$, $\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon} \leq 1$. Then $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}^\delta(X), X) - F(w^*(X), X) \leq 13LR\left(\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon}\right)^{1/2} + 12\sqrt{\alpha}\sqrt{LR}(\frac{\epsilon}{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})})^{1/4}$. In particular, setting $\alpha = LR\left(\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon}\right)^{3/2}$ implies $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}^\delta(X), X) - F(w^*(X), X) \leq 25LR\left(\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon}\right)^{1/2}$.

2. Suppose $F \in \mathcal{G}_{L,R}^{ERM}$.

a) Let $\delta = 0, \frac{d}{\epsilon n} \leq 1$. Then $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 16.5LR\sqrt{\frac{d}{\epsilon n}} + 16\sqrt{\alpha}\sqrt{LR}(\frac{\epsilon n}{d})^{1/4}(1 + \frac{d}{\epsilon})$. In particular, setting $\alpha = \frac{LR\left(\frac{d}{\epsilon n}\right)^{3/2}}{(1+\frac{d}{\epsilon})^2}$ implies $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 33LR\sqrt{\frac{d}{\epsilon n}}$.

b) Let $\delta \in (0, \frac{1}{2}), \frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon n} \leq 1$. Then $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}^\delta(X), X) - F(w^*(X), X) \leq 13LR\left(\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)^{1/2} + 12\sqrt{\alpha}\sqrt{LR}(\frac{\epsilon n}{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})})^{1/4}$. In particular, setting $\alpha = LR\left(\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)^{3/2}$ implies $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}^\delta(X), X) - F(w^*(X), X) \leq 25LR\left(\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)^{1/2}$.

To prove Theorem 0.4, decompose excess risk into 3 terms:

$$\begin{aligned} &\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) = \\ &\underbrace{\mathbb{E}_A[F(w_T(X) + z_\lambda, X) - F_\lambda(w_T(X) + z_\lambda, X)]}_{\text{ⓐ}} \\ &+ \underbrace{\mathbb{E}_A F_\lambda(w_T(X) + z_\lambda, X) - F_\lambda(w_\lambda^*(X), X)}_{\text{ⓑ}} \\ &+ \underbrace{F_\lambda(w_\lambda^*(X), X) - F(w^*(X), X)}_{\text{ⓒ}}. \end{aligned}$$

Now bound the terms as follows: ⓐ $= -\frac{\lambda}{2}\mathbb{E}\|w_{\mathcal{A},\lambda}(X)\|_2^2 \leq 0$; ⓑ $\leq (L + \lambda R)\mathbb{E}\|w_T(X) - w_\lambda^*(X) + z_\lambda\| \leq (L + \lambda R)[\sqrt{\frac{2\alpha}{\lambda}} + \mathbb{E}\|z_\lambda\|_2]$; and ⓒ $= [F_\lambda(w_\lambda^*(X), X) - F_\lambda(w^*(X), X)] + [F_\lambda(w^*(X), X) - F(w^*(X), X)] < 0 + \frac{\lambda}{2}\|w^*(X)\|_2^2 \leq \frac{\lambda R^2}{2}$ since $w_\lambda^*(X)$ is the unique minimizer of $F_\lambda(\cdot, X)$ by strong convexity. Hence $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq (L + \lambda R)\left[\sqrt{\frac{2\alpha}{\lambda}} + \mathbb{E}\|z_\lambda\|_2\right] + \frac{\lambda R^2}{2}$.

Assume that $F \in \mathcal{G}_{L,R}^{ERM}$: the proof of the non-ERM case is very similar (but simpler since n disappears). First Suppose $\delta = 0$, so $\mathbb{E}\|z_\lambda\|_2 \leq \frac{\sqrt{2}d(\Delta_\lambda + 2\sqrt{\frac{2\alpha}{\lambda}})}{\epsilon}$. Then $\Delta_\lambda \leq \frac{2(L+\lambda R)}{\lambda n} \leq \frac{4L}{\lambda n} \leq 4\sqrt{2}R\sqrt{\frac{\epsilon}{nd}}$ by our choice of λ and assumption that $\frac{d}{\epsilon n} \leq 1$. Also, $\frac{1}{\lambda} < \sqrt{2}\frac{R}{L}\sqrt{\frac{d}{\epsilon n}}$. Using these estimates and the above estimate for excess risk gives

$$\begin{aligned} &\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq 2L\times \\ &\left[2\sqrt{\alpha}\sqrt{\frac{R}{L}}\left(\frac{\epsilon n}{d}\right)^{1/4} + \sqrt{2}\frac{d}{\epsilon}\left(4\sqrt{2}R\sqrt{\frac{\epsilon}{nd}} + 4\sqrt{2}\sqrt{\alpha}\sqrt{\frac{R}{L}}(\frac{\epsilon n}{d})^{1/4}\right)\right] \\ &+ LR\frac{\sqrt{\frac{d}{\epsilon n}}}{2} \\ &\leq 16\sqrt{\alpha}\sqrt{LR}\left(\frac{\epsilon n}{d}\right)^{1/4}(1 + \frac{d}{\epsilon}) + 16.5LR\sqrt{\frac{d}{\epsilon n}}, \end{aligned}$$

as claimed. Then plug in α.

Next, suppose $\delta \in (0, \frac{1}{2})$. Then $\mathbb{E}\|z_\lambda\|_2 \leq \frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon}\left(\Delta_\lambda + 2\sqrt{\frac{2\alpha}{\lambda}}\right)$. Now $\lambda = \frac{L}{R\sqrt{1+\frac{\epsilon n}{d^{1/2}(c_\delta+\sqrt{c_\delta^2+\epsilon})}}} < \frac{L}{R\sqrt{\frac{\epsilon n}{d^{1/2}(c_\delta+\sqrt{c_\delta^2+\epsilon})}}} \leq \frac{L}{R}\sqrt{\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon n}}$ and $\Delta_\lambda \leq \frac{4L}{\lambda n} \leq 4\sqrt{2}\frac{R}{n}(\frac{\epsilon n}{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})^2})^{1/2}$. Using these estimates and the estimates for each component of excess risk, as above, gives the first statement of part 2b). Using the assumptions and estimates above, it is easy to verify that the prescribed choices of α yield the second statements in each part of the theorem.

We can now prove Theorem 2.3: Recall that the iteration complexity of the subgradient method on F_λ for finding w_T such that $F_\lambda(w_T(X), X)) - F_\lambda(w_\lambda^*(X), X) \leq \alpha$ is $T \leq \frac{2(L+\lambda R)^2}{\lambda\alpha} \leq \frac{8L^2}{\lambda\alpha}$ (Aravkin et al. 2017) by λ-strong

convexity and $(L+\lambda R)$-Lipschitzness of $F_\lambda(\cdot, X)$, as well as our choices of λ (see proof of Theorem 0.4). Then plugging in the exact choices of λ and α gives the risk bound results. For runtime results, multiply the iteration complexity by d since each iteration requires one gradient evaluation of F.

Proof of Theorem 2.4

We begin with a more general result:

Theorem 0.5 *Run Algorithm 2 with accuracy α as given below and λ given as follows: if $\delta = 0$, set* $\lambda := \begin{cases} \left(\frac{\beta L^2}{R^2}\right)^{1/3} \left(\frac{d}{\epsilon n}\right)^{2/3} & \text{if } \in \mathcal{J}^{ERM}_{\beta,L,R} \\ \left(\frac{\beta L^2}{R^2}\right)^{1/3} \left(\frac{d}{\epsilon}\right)^{2/3} & \text{if } \in \mathcal{J}_{\beta,L,R} \setminus \mathcal{J}^{ERM}_{\beta,L,R}. \end{cases}$
If $\delta \in (0, \frac{1}{2})$, set $\lambda := \begin{cases} \left(\frac{\beta L^2}{R^2}\right)^{1/3} \left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)^{2/3} & \text{if } \in \mathcal{J}^{ERM}_{\beta,L,R} \\ \left(\frac{\beta L^2}{R^2}\right)^{1/3} \left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon}\right)^{2/3} & \text{if } \in \mathcal{J}_{\beta,L,R} \setminus \mathcal{J}^{ERM}_{\beta,L,R}. \end{cases}$
1. Suppose $F \in \mathcal{J}_{\beta,L,R}$.
a) Let $\delta = 0$, $\left(\frac{d}{\epsilon}\right)^2 \le \frac{L}{R\beta}$.
Then $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \le 65\beta^{1/3}L^{2/3}R^{4/3}\left(\frac{d}{\epsilon}\right)^{2/3} + \alpha[16(\frac{R\beta}{L})^{2/3}(\frac{d}{\epsilon})^{4/3} + 1] + 64\sqrt{\alpha}R\sqrt{\beta}(\frac{d}{\epsilon})$.
In particular, setting $\alpha = \min\left\{\frac{L^{4/3}R^{2/3}}{\beta^{1/3}}(\frac{\epsilon}{p})^{2/3}, \beta^{1/3}L^{2/3}R^{4/3}\left(\frac{d}{\epsilon}\right)^{2/3}\right\}$ *implies* $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \le \beta^{1/3}L^{2/3}R^{4/3}146\left(\frac{d}{\epsilon}\right)^{2/3}$.
b) Let $\delta \in (0, \frac{1}{2})$, $\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon}\right)^2 \le \frac{L}{R\beta}$.
Then $\mathbb{E}_{\mathcal{A}} F(w^\delta_{\mathcal{A}}(X), X) - F(w^*(X), X) \le 16.5\beta^{1/3}L^{2/3}R^{4/3}\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon}\right)^{2/3} + \alpha\left[1 + \left(\frac{\beta R}{L}\right)^{2/3}\frac{(d(c_\delta+\sqrt{\epsilon})^2)^{2/3}}{\epsilon^{4/3}}\right] + 8\sqrt{\alpha}R\sqrt{\beta}\frac{\sqrt{d}(c_\delta+\sqrt{\epsilon})}{\epsilon}$.
In particular, setting

$$\alpha = \min\left\{\frac{L^{4/3}R^{2/3}}{\beta^{1/3}}\left(\frac{\epsilon}{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}\right)^{2/3}, \beta^{1/3}L^{2/3}R^{4/3}\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon}\right)^{2/3}\right\}$$

implies $\mathbb{E}_{\mathcal{A}} F(w^\delta_{\mathcal{A}}(X), X) - F(w^*(X), X) \le 27\beta^{1/3}L^{2/3}R^{4/3}\left(\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon}\right)^{2/3}$.
2.Suppose $F \in \mathcal{J}^{ERM}_{\beta,L,R}$.
a) Let $\delta = 0$, $\left(\frac{d}{\epsilon n}\right)^2 \le \frac{L}{R\beta}$.
Then $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \le 65\beta^{1/3}L^{2/3}R^{4/3}\left(\frac{d}{\epsilon n}\right)^{2/3} + \alpha\left[16(\frac{R\beta}{L})^{2/3}\frac{d^{4/3}n^{2/3}}{\epsilon^{4/3}} + 1\right] + 64\sqrt{\alpha}R\beta^{1/2}\frac{d}{\epsilon}$.
In particular, setting $\alpha = \min\left\{\frac{L^{4/3}R^{2/3}}{\beta^{1/3}}(\frac{\epsilon}{p})^{2/3}\frac{1}{n^{4/3}}, \beta^{1/3}L^{2/3}R^{4/3}\left(\frac{d}{\epsilon n}\right)^{2/3}\right\}$ *implies* $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \le 146\beta^{1/3}L^{2/3}R^{4/3}\left(\frac{d}{\epsilon n}\right)^{2/3}$.
b) Let $\delta \in (0, \frac{1}{2})$, $\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)^2 \le \frac{L}{R\beta}$.
Then $\mathbb{E}_{\mathcal{A}} F(w^\delta_{\mathcal{A}}(X), X) - F(w^*(X), X) \le 16.5\beta^{1/3}L^{2/3}R^{4/3}\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)^{2/3} + \alpha\left[1 + n^{2/3}\left(\frac{\beta R}{L}\right)^{2/3}\frac{(d(c_\delta+\sqrt{c_\delta^2+\epsilon})^2)^{2/3}}{\epsilon^{4/3}}\right] + 8\sqrt{\alpha}R\sqrt{\beta}\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon}$.
In particular, setting

$$\alpha = \min\left\{\frac{L^{4/3}R^{2/3}}{\beta^{1/3}}\left(\frac{\epsilon n}{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}\right)^{2/3}\frac{1}{n^2}, \beta^{1/3}L^{2/3}R^{4/3}\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)^{2/3}\right\}$$

implies $\mathbb{E}_{\mathcal{A}} F(w^\delta_{\mathcal{A}}(X), X) - F(w^*(X), X) \le 27\beta^{1/3}L^{2/3}R^{4/3}\left(\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)^{2/3}$.

To prove Theorem 0.5, first decompose excess risk into 3 terms as usual:

$$\begin{aligned} &\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) = \\ &\underbrace{\mathbb{E}_A[F(w_T(X) + z_\lambda, X) - F_\lambda(w_T(X) + z_\lambda, X)]}_{\text{ⓐ}} + \\ &\underbrace{\mathbb{E}_A F_\lambda(w_T(X) + z_\lambda, X) - F_\lambda(w^*_\lambda(X), X)}_{\text{ⓑ}} + \\ &\underbrace{F_\lambda(w^*_\lambda(X), X) - F(w^*(X), X)}_{\text{ⓒ}}. \end{aligned}$$

Now bound the terms as follows: ⓐ $= -\frac{\lambda}{2}\mathbb{E}\|w_{\mathcal{A},\lambda}(X)\|_2^2 \le 0$. By the descent lemma (note $F_\lambda(\cdot, X)$ is $(\beta+\lambda)$-smooth) and independence of the noise z_λ and $\nabla F_\lambda(w, X)$, we have ⓑ =

$$\begin{aligned} &\mathbb{E}[F_\lambda(w_T + z_\lambda, X) - F_\lambda(w_T, X)] + \mathbb{E}[F_\lambda(w_T, X) - F_\lambda(w^*_\lambda, X)] \\ &\le \mathbb{E}[\langle \nabla F_\lambda(w_T(X), X)), z_\lambda\rangle + \frac{\beta+\lambda}{2}\|z_\lambda\|_2^2] + \alpha \\ &= \frac{\beta+\lambda}{2}\mathbb{E}\|z_\lambda\|_2^2 + \alpha. \end{aligned}$$

Lastly, ⓒ $= [F_\lambda(w^*_\lambda(X), X) - F_\lambda(w^*(X), X)] + [F_\lambda(w^*(X), X) - F(w^*(X), X)] < 0 + \frac{\lambda}{2}\|w^*(X)\|_2^2 \le \frac{\lambda R^2}{2}$ since $w^*_\lambda(X)$ is the unique minimizer of $F_\lambda(\cdot, X)$ by strong convexity. Hence $\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \le \frac{\lambda R^2}{2} + (\beta+\lambda)\mathbb{E}\|z_\lambda\|_2^2 + \alpha$.

Assume that $F \in \mathcal{J}^{ERM}_{\beta,L,R}$: the proof of the non-ERM case is very similar, but simpler since n disappears. First Suppose $\delta = 0$, so $\mathbb{E}\|z_\lambda\|_2^2 \leq \frac{2d^2(2\sqrt{\frac{2\alpha}{\lambda}}+\Delta_\lambda)^2}{\epsilon^2}$. Note that for our choice of λ given in Theorem 0.5, $\lambda \leq \beta$ and $(L+\lambda R) \leq 2L$ by the assumption $\left(\frac{d}{\epsilon n}\right)^2 \leq \frac{L}{R\beta}$ and since $L \leq R\beta$ always holds for $F \in \mathcal{J}_{\beta,L,R}$. Also, $\Delta_\lambda \leq \frac{2(L+\lambda R)}{\lambda n} \leq \frac{4L}{\lambda n} \leq 4(\frac{LR^2}{\beta})^{1/3}(\frac{\epsilon n}{d})^{2/3}$. Therefore,

$$\mathbb{E}_{\mathcal{A}} F(w_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq \frac{\lambda R^2}{2} + 16\beta\left(\frac{d}{\epsilon}\right)^2 \times \left[4\frac{L^2}{\lambda^2 n^2} + \frac{\alpha}{\lambda} + 4\sqrt{\frac{\alpha}{\lambda}}\frac{L}{\lambda n}\right] + \alpha.$$

Plugging in $\lambda = \left(\frac{\beta L^2}{R^2}\right)^{1/3}\left(\frac{d}{\epsilon n}\right)^{2/3}$, using the estimates above and the fact that $\frac{d}{\epsilon n} \leq 1$ by assumption gives

$$\lambda R^2, \beta\left(\frac{d}{\epsilon}\right)^2 \frac{L^2}{\lambda^2 n^2} \leq \beta^{1/3} L^{2/3} R^{4/3}\left(\frac{d}{\epsilon n}\right)^{2/3}.$$

Also, by the choice of λ,

$$\beta\left(\frac{d}{\epsilon}\right)^2\left[\frac{\alpha}{\lambda} + 4\sqrt{\frac{\alpha}{\lambda}}\frac{L}{\lambda n}\right] \leq \alpha\left[\left(\frac{R\beta}{L}\right)^{2/3}\frac{d^{4/3}n^{2/3}}{\epsilon^{4/3}}\right] + 4\sqrt{\alpha}R\beta^{1/2}\left(\frac{d}{\epsilon}\right).$$

Putting these pieces together proves the first statement in part 2a). Verifying the second statement is routine, using the assumptions stated in the theorem and the estimates obtained above.

Now Suppose $\alpha \in (0, \frac{1}{2})$. Then $\mathbb{E}\|z_\lambda\|_2^2 \leq \frac{d(c_\delta+\sqrt{c_\delta^2+\epsilon})(2\sqrt{\frac{2\alpha}{\lambda}}+\Delta_\lambda)^2}{\epsilon^2}$. Again, the assumption $\left(\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)^2 \leq \frac{L}{R\beta}$ implies $\lambda \leq \beta$ and $L+\lambda R \leq 2L$. Expanding the square in the numerator implies

$$\mathbb{E}_{\mathcal{A}} F(w^\delta_{\mathcal{A}}(X), X) - F(w^*(X), X) \leq \frac{\lambda R^2}{2} \times \beta\frac{d(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon^2}\left[\Delta_\lambda^2 + 2\Delta_\lambda\sqrt{\frac{\alpha}{\lambda}} + \frac{\alpha}{\lambda}\right] + \alpha$$
$$\leq 16.5\beta^{1/3}L^{2/3}R^{4/3}\left(\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)^{2/3} + \beta\frac{d(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon^2}\left[\frac{8L}{\lambda n}\sqrt{\frac{\alpha}{\lambda}} + \frac{\alpha}{\lambda}\right] + \alpha,$$

where in the last line we plugged in estimates for λ and Δ_λ to bound the first two terms in the sum. Then plugging in $\lambda = \left(\frac{\beta L^2}{R^2}\right)^{1/3}\left(\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right)^{2/3}$ and using the estimates and assumptions completes the proof of the first statement of the theorem. Again, the second statement is easy to verify by plugging in the prescribed α. This proves Theorem 0.5.

Theorem 2.4 then follows from the above theorem and the iteration complexity of AGD (Nesterov 2014).

Proofs of Section 3 Results

We begin with a definition and lemma from (Bassily et al. 2019) that will be very useful for us.

Definition 2 *(Uniform stability) Let $\alpha > 0$. An algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{W}$ is α-uniformly stable (w.r.t loss function f) if for any pair of data sets $X, X' \in \mathcal{X}^n$ differing by at most one point (i.e. $|X \Delta X'| \leq 2$), we have*

$$\sup_{x \in \mathcal{X}} \mathbb{E}_{\mathcal{A}}[f(\mathcal{A}(X), x) - f(\mathcal{A}(X'), x)] \leq \alpha.$$

Lemma 0.4 *Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{W}$ be an α-uniformly stable algorithm w.r.t. loss function f and let $X \sim \mathcal{D}^n$. Then*

$$\mathbb{E}_{X\sim\mathcal{D}^n, \mathcal{A}}[F(\mathcal{A}(X), \mathcal{D}) - \hat{F}(\mathcal{A}(X), X)] \leq \alpha.$$

Combining Lemma 0.4 with our excess empirical risk bounds from before for each ERM function class will enable us to upper bound the population loss.

Proof of Theorem 3.1

Decompose

$$\mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}} F(w_{\mathcal{A}}(X), \mathcal{D}) - F(w^*(\mathcal{D}), \mathcal{D}) = \underbrace{\mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}}[F(w_{\mathcal{A}}(X), \mathcal{D}) - \hat{F}(w_{\mathcal{A}}(X), X)]}_{\text{ⓐ}} + \underbrace{\mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}}[\hat{F}(w_{\mathcal{A}}(X), X) - \hat{F}(\hat{w}(X), X)]}_{\text{ⓑ}} + \underbrace{\mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}}\hat{F}(\hat{w}(X), X) - F(w^*(\mathcal{D}), \mathcal{D})}_{\text{ⓒ}},$$

as before. Observe that $\Delta_T := \sup_{|X\Delta X'|\leq 2}\|w_T(X) - w_T(X')\|_2 \leq \Delta_{\hat{F}} + 2\sqrt{\frac{2\alpha}{\mu}} \leq 2(\frac{L}{\mu n} + \sqrt{\frac{2\alpha}{\mu}})$, and $\mathcal{A}$ is $L\Delta_T$-uniformly stable with respect to f. Hence ⓐ $\leq 2(\frac{L^2}{\mu n} + L\sqrt{\frac{2\alpha}{\mu}})$, by Lemma 0.4. Plugging in the choices of α given in the Corollary shows that ⓐ $\leq 5\frac{L^2}{\mu}$. Next, by Theorem 0.2,

$$\text{ⓑ} = \mathbb{E}_{\mathcal{A}}[\hat{F}(w_{\mathcal{A}}(X), X) - \hat{F}(\hat{w}(X), X)] \leq 9\frac{L^2}{\mu}\frac{d}{\epsilon n}$$

if $\delta = 0$ and $\mathbb{E}_{\mathcal{A}}[\hat{F}(w_{\mathcal{A}}(X), X) - \hat{F}(\hat{w}(X), X)] \leq 6\frac{L^2}{\mu}\frac{\sqrt{d}(c_\delta+\sqrt{c_\delta^2+\epsilon})}{\epsilon n}$ if $\delta \in (0, \frac{1}{2})$ with the given choices of α and T. Finally, ⓒ ≤ 0. Putting these estimates together completes the proof of the excess loss bounds.

Recall that running SGD on $F \in \mathcal{F}^{ERM}_{\mu,L}$ with $\eta_t = \frac{2}{\mu(t+1)}$ produces a point $\hat{w}_T$ such that $\mathbb{E}F(\hat{w}_T, X) - F(w^*(X), X) \leq \alpha$ in $T = \frac{2L^2}{\mu\alpha}$ stochastic gradient evaluations (Bubeck 2015, Theorem 6.2). Since each iteration amounts to just one gradient evaluation of $f(w, x_i)$ (with runtime d), the resulting runtime of the full method is $O(dT)$.

Proof of Theorem 3.2

The proof follows exactly as the proof of Theorem 3.1, but using Theorem 0.3 instead of Theorem 0.2 to bound ⓑ. The runtime bounds follow from applying the following:

Theorem 0.6 *(Allen-Zhu 2018, Theorem 2.1 simplified) Let* $F : \mathbb{R}^d \to \mathbb{R}$, $F(w) = \frac{1}{n}\sum_{i=1}^n f_i(w) = \frac{1}{n}\sum_{i=1}^n g_i(w) + \psi(w)$, *where each* g_i *is convex and* β*-smooth,* ψ *is* μ*-strongly convex, and* $\kappa := \frac{\beta}{\mu}$*. Then running Katyusha for* $T = O\left((n+\sqrt{n\kappa})\log\left(\frac{F(w_0)-F(w^*)}{\alpha}\right)\right)$ *stochastic gradient iterations returns a point* $\widetilde{w}_T \in \mathbb{R}^d$ *such that* $F(\widetilde{w}_T) - F(w^*) \leq \alpha$.

Clearly, given any $F \in \mathcal{H}^{ERM}_{\beta,\mu,L,R}$ and any $X \in \mathcal{X}^n$, we can write $F(w,X) = \frac{1}{n}\sum_{i=1}^n f(w,x_i) = \frac{1}{n}\sum_{i=1}^n g(w,x_i) + \psi(w)$, where $\psi(w) := \frac{\mu}{2}\|w\|_2^2$ is μ-strongly convex, $g(\cdot,x_i) = f(\cdot,x_i) - \psi(\cdot)$ is convex and $(\beta-\mu)$-smooth (hence β-smooth) for all $x_i \in X$.

Proof of Theorem 3.3

Decompose

$$\mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}} F(w_{\mathcal{A}}(X),\mathcal{D}) - F(w^*(\mathcal{D}),\mathcal{D}) =$$
$$\underbrace{\mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}}[F(w_{\mathcal{A}}(X),\mathcal{D}) - \hat{F}(w_{\mathcal{A}}(X),X)]}_{ⓐ} +$$
$$\underbrace{\mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}}[\hat{F}(w_{\mathcal{A}}(X),X) - \hat{F}(\hat{w}(X),X)]}_{ⓑ} +$$
$$\underbrace{\mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}}\hat{F}(\hat{w}(X),X) - F(w^*(\mathcal{D}),\mathcal{D})}_{ⓒ},$$

as before. First, $\mathcal{A}$ is $L\Delta_T$-uniformly stable with respect to f, and $\Delta_T = \sup_{|X\Delta X'|\leq 2}\|w_T(X) - w_T(X')\|_2 \leq \Delta_\lambda + 2\sqrt{\frac{2\alpha}{\lambda}} \leq 2[\frac{L+\lambda R}{\lambda n} + \sqrt{\frac{2\alpha}{\lambda}}]$, using Theorem 2.3 and strong convexity of F_λ. Hence $ⓐ \leq 2L\left[\frac{L+\lambda R}{\lambda n} + \sqrt{\frac{2\alpha}{\lambda}}\right]$ by Lemma 0.4. Now

$$ⓑ \leq (L+\lambda R)\left[\sqrt{\frac{2\alpha}{\lambda}} + \mathbb{E}\|z_\lambda\|_2\right] + \frac{\lambda R^2}{2},$$

by decomposing and bounding as done in the proof of Theorem 0.4. Also, $ⓒ \leq 0$ as usual. Then combining the above with our choices of λ, α, T and using plugging in expected values of the noise completes the proof of the loss bounds. The runtime bounds follow from the SGD runtime bounds stated earlier and the choices of λ prescribed above.

Proof of Theorem 3.4

The proof follows almost exactly as the proof of Theorem 3.3 above, but uses the descent lemma to bound ⓑ with a $\frac{(\beta+\lambda)}{2}\mathbb{E}\|z_\lambda\|^2$ term (instead of $(L+\lambda R)\mathbb{E}\|z_\lambda\|_2$) and uses Theorem 2.4 (Katyusha) instead of Theorem 2.3 (SGD), along with the alternative choices of α and λ, to obtain the stated loss bounds in faster runtime.

Proofs of Section 4 Results

Proof of Proposition 4.1

We first require a tighter estimate of the sensitivity of F_τ :

Lemma 0.5 *Let* $\tau > 0$. *Suppose* $f(\cdot,x)$ *is* L*-Lipschitz on* $B(0,R)$ *(where* $\|w^*(X)\| \leq R$*) and* μ*-strongly convex for all* $x \in \mathcal{X}$. *Moreover, assume* $a_R \leq f(w,x) \leq A_R$ *for all* $w \in \mathcal{W}$ *and all* $x \in \mathcal{X}$. *Then*

$$\Delta_{F_\tau} \leq \frac{2LC_\tau}{\mu n}$$

where $C_\tau := e^{\tau(A_R - a_R)}$.

To prove the lemma, we follow the same approach used in proving **??**. Let $X, X' \in \mathcal{X}^n$ such that $|X\Delta X'| \leq 2$ and assume WLOG that $x_n \neq x'_n$. We apply Lemma 0.2 with $g_\tau(w) := F_\tau(w,X) - F_\tau(w,X')$ and $G_\tau(w) := F_\tau(w,X')$. Denote $v_i(\tau,w) = \frac{e^{\tau f(w,x_i)}}{\sum_{j=1}^n e^{\tau f(w,x_j)}}$ and $v'_i(\tau,w) = \frac{e^{\tau f(w,x'_i)}}{\sum_{j=1}^n e^{\tau f(w,x'_j)}}$. Then observe that

$$\begin{aligned}\nabla g_\tau(w) &= \nabla F_\tau(w,X) - \nabla F_\tau(w,X') \\ &= \sum_{i=1}^n v_i(\tau,w)\nabla f(w,x_i) - v'_i(\tau,w)\nabla f(w,x'_i) \\ &= v_n(\tau,w)\nabla f(w,x_n) - v'_n(\tau,w)\nabla f(w,x'_n).\end{aligned}$$

Now convexity and L-Lipschitzness of f imply

$$\|\nabla g_\tau(w)\| \leq 2L\max\{v_n(\tau,w), v'_n(\tau,w)\} \leq 2LC_\tau.$$

This proof is completed by noticing that G_τ is μ-strongly convex and appealing to Lemma 0.2.

Then to obtain Proposition 4.1, simply plug the estimate for Δ_{F_τ} given above into the proof of Theorem 0.2 and set $\alpha = \frac{L^2 C_\tau}{\mu n}\min\left\{\frac{1}{n}, \frac{C_\tau d}{\epsilon}\right\}$ for $\delta = 0$ and $\alpha = \frac{L^2}{\mu}C_\tau \min\left\{\frac{C_\tau}{n^2}, \frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2+\epsilon})}{\epsilon n}\right\}$ for $\delta \in (0, \frac{1}{2})$. The runtime bounds are proved by recalling that the iteration complexity of the subgradient method for obtaining an α-suboptimal point w_T of F_τ is $T = \left\lceil \frac{2L^2}{\mu\alpha}\right\rceil$ (and noting that each gradient evaluation of F_τ has runtime $O(nd)$).

Proof of Lemma 4.2

First, we have $\nabla F_\tau(w,X) = \sum_{i=1}^n v_i(w,\tau)\nabla f_i(w)$, where $v_i(w,\tau) := \frac{e^{\tau f(w,x_i)}}{\sum_{j=1}^n e^{\tau f(w,x_j)}}$ and we denote $f_i(w) := f(w,x_i)$. Now for any $w_1, w_2 \in \mathcal{W}$, we have $\|\nabla F_\tau(w_1,X) - \nabla F_\tau(w_2,X)\|_2 = \|\sum_{i=1}^n (v_i(w_1,\tau)\nabla f_i(w_1) - v_i(w_2,\tau)f_i(w_2))\|_2 \leq \sum_{i=1}^n (v_i(w_1,\tau)\beta\|w_1 - w_2\|_2 + L_{v_i}\|w_1 - w_2\|_2 L) \leq \beta\|w_1 - w_2\|_2 + L\|w_1 - w_2\|_2\sum_{i=1}^n (L_{v_i})$, where L_{v_i} denotes the Lipschitz constant of $v_i(w,\tau)$ as a function of w. Next, we compute L_{v_i} by bounding $\|\nabla v_i(w,\tau)\|$: $\nabla v_i(w,\tau) = \frac{\tau e^{\tau f_i(w)}\nabla f_i(w)\left(\sum_{j=1}^n e^{\tau f_j(w)}\right) - \tau\sum_{j=1}^n e^{\tau f_j(w)}\nabla f_j(w) e^{\tau f_i(w)}}{\left(\sum_{j=1}^n e^{\tau f_j(w)}\right)^2}$

$= \frac{\tau \sum_{j=1}^n e^{\tau f_j(w)} e^{\tau f_i(w)} (\nabla f_i(w) - \nabla f_j(w))}{\left(\sum_{j=1}^n e^{\tau f_j(w)}\right)^2}$. Taking the norm of both sides and using L-Lipschitzness of $f(\cdot, x_i)$ implies

$$L_{v_i} \leq 2L\tau \frac{e^{\tau f_i(w)}}{\sum_{j=1}^n e^{\tau f_j(w)}},$$

and hence

$$\sum_{i=1}^n L_{v_i} \leq 2L\tau.$$

Therefore, $\|\nabla F_\tau(w_1, X) - \nabla F_\tau(w_2, X)\|_2 \leq \beta\|w_1 - w_2\|_2 + L\|w_1 - w_2\|_2 \sum_{i=1}^n L_{v_i}$
$\leq \left(\beta + 2L^2\tau\right)\|w_1 - w_2\|_2$, which completes the proof.

Proof of Proposition 4.2

Plugging the estimate for Δ_{F_τ} from Lemma 0.5 into the proof of Theorem 0.3 and set

$$\alpha = \frac{L^2}{\mu n^2} \begin{cases} \min\left\{\kappa_\tau \left(\frac{d}{\epsilon}\right)^2, C_\tau^2\right\} & \text{if } \delta = 0 \\ \min\left\{\kappa_\tau \left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon}\right)^2, C_\tau^2\right\} & \text{if } \delta \in (0, \frac{1}{2}) \end{cases}.$$

This yields

$$\leq 26\kappa_\tau \frac{L^2 C_\tau}{\mu}\left(\frac{d}{\epsilon n}\right)^2$$

if $\delta = 0$, and

$$\leq 13.5\kappa_\tau \frac{L^2 C_\tau}{\mu}\left(\frac{\sqrt{d}(c_\delta + \sqrt{c_\delta^2 + \epsilon})}{\epsilon n}\right)^2$$

if $\delta \in (0, \frac{1}{2})$. The runtime claim follows by recalling the $\widetilde{O}(\sqrt{\kappa_\tau})$ gradient complexity of AGD.

Proof of Theorem 4.1

Theorem 4.1 is an immediate consequence of Theorem 0.3 combined with the following:

Proposition 0.2 *((Lin et al. 2020, Thm 5.1/Cor 5.2)) Assume $G(\cdot, v)$ is μ-strongly convex, β-smooth, (with condition number $\kappa = \beta/\mu$) for all $v \in B_\nu$, $G(w, \cdot)$ is β_v-smooth and concave as a function of $v \in B_\nu$ for all $w \in \mathcal{W}$. Then*
1. Minimax-APPA returns an α-saddle point in at most $T = \widetilde{O}(\sqrt{\frac{\kappa\beta_v}{\alpha}}\nu)$ total gradient evaluations.
2. If, in addition, $G(w, \cdot)$ is μ_v-strongly concave, then Minimax-APPA returns an α-saddle point in at most $T = \widetilde{O}(\sqrt{\kappa\kappa_v})$ gradient evaluations, where $\kappa_v = \frac{\beta_v}{\mu_v}$ is the condition number.

References

Mohammad Alkousa, Darina Dvinskikh, Fedor Stonyakin, Alexander Gasnikov, and Dmitry Kovalev. Accelerated methods for composite non-bilinear saddle point problem, 2020.

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(221):1–51, 2018. URL http://jmlr.org/papers/v18/16-410.html.

Aleksandr Aravkin, James Burke, and Dmitry Drusvyatskiy. *Convex Analysis and Nonsmooth Optimization*. 2017.

Raman Arora, Teodor V. Marinov, and Enayat Ullah. Private stochastic convex optimization: Efficient algorithms for non-smooth objectives. *arXiv:2002.09609v1*, 2020.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Differentially private empirical risk minimization: Efficient algorithms and tight error bounds. *arXiv:1405.7085v2*, 2014.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in Neural Information Processing Systems*, 32:11282–11291, 2019.

Olivier Bousquet and Andre Elisseeff. Stability and generalization. *Journal of machine learning research*, 2002.

Stephen Boyd, Corinna Cortes, Mehryar Mohri, and Ana Radovanovic. Accuracy at the top. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, pages 953–961. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/7fe1f8abaad094e0b5cb1b01d712f708-Paper.pdf.

Sebastian Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8, 2015.

Kamalika Chaudhuri, Claire Monteleoni, and Anand Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

Nadav Cohen and Amnon Shashua. Simnets: A generalization of convolutional networks. *arXiv:1410.0781v3*, 2014.

Nadav Cohen, Or Sharir, and Amnon Shashua. Deep simnets. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4782–4791, 2016. doi: 10.1109/CVPR.2016.517.

Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on privacy enhancing technologies*, 2015(1):92–112, 2015.

Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*. 2014.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology - EUROCRYPT 2006*, volume 4004 of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, 2006.

Kevin Eykholt, Ivan Evtimov, Earlance Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018. doi: 10.1109/CVPR.2018.00175.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in linear time. *arXiv:2005.04763v1*, 2020.

Ferdinando Fioretto, Terrence WK Mak, and Pascal Van Hentenryck. Differential privacy for power grid obfuscation. *IEEE Transactions on Smart Grid*, 11(2):1356–1366, 2019.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1412.6572.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1225–1234, New York, New York, USA, 20–22 Jun 2016. PMLR. URL http://proceedings.mlr.press/v48/hardt16.html.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 2014.

Angelos Katharopoulos and François Fleuret. Biased importance sampling for deep neural network training. *arXiv:1706.00043*, 2017.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 25.1–25.40, Edinburgh, Scotland, 25–27 Jun 2012. PMLR. URL http://proceedings.mlr.press/v23/kifer12.html.

Barry W. Kort and Dimitri P. Bertsekas. A new penalty function method for constrained minimization. In *Proceedings of the 1972 IEEE Conference on Decision and Control and 11th Symposium on Adaptive Processes*, pages 162–166, 1972. doi: 10.1109/CDC.1972.268971.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. *arXiv:2007.01162v1*, 2020.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. *Advances in neural information processing systems*, 28:3384–3392, 2015.

Tianyi Lin, Chi Jin, and Michael I. Jordan. Near-optimal algorithms for minimax optimization. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2738–2779. PMLR, 09–12 Jul 2020. URL http://proceedings.mlr.press/v125/lin20a.html.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016. doi: 10.1109/CVPR.2016.282.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014. ISBN 1461346916.

Yurii Nesterov and Laura Scrimali. Solving strongly monotone variational and quasi-variational inequalities. *Discrete and Continuous Dynamical Systems*, 31, 01 2007. doi: 10.2139/ssrn.970903.

Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems 32*, pages 14934–14942. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9631-solving-a-class-of-non-convex-min-max-games-using-iterative-first-order-methods.pdf.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS P)*, pages 372–387, 2016. doi: 10.1109/EuroSP.2016.36.

E.Y. Pee and J. O. Royset. On solving large-scale finite minimax problems using exponential smoothing. *Journal of Optimization Theory and Applications*, 148(2):390–421, February 2011. URL https://ideas.repec.org/a/spr/joptap/v148y2011i2d10.1007_s10957-010-9759-1.html.

Hai Phan, My T Thai, Han Hu, Ruoming Jin, Tong Sun, and Dejing Dou. Scalable differential privacy with certified robustness in adversarial learning. In *International Conference on Machine Learning*, pages 7683–7694. PMLR, 2020.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, pages 171–176, 1958.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL http://arxiv.org/abs/1312.6199.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30:2722–2731, 2017.

Jun Wang, Rongbo Zhu, Shubo Liu, and Zhaohui Cai. Node location privacy protection based on differentially private grids in industrial wireless sensor networks. *Sensors*, 18(2): 410, 2018.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning*

Research, pages 7472–7482. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/zhang19p.html.

Jiaqi Zhang, Kai Zheng, Wenlong Mou, and Liwei Wang. Efficient private erm for smooth objectives. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3922–3928, 2017.

Jun Zhao, Teng Wang, Tao Bai, Kwok-Yan Lam, Zhiying Xu, Shuyu Shi, Xuebin Ren, Xinyu Yang, Yang Liu, and Han Yu. Reviewing and improving the gaussian mechanism for differential privacy. *arXiv:1911.12060v2*, 2019.