Private Empirical Risk Minimization Beyond the Worst Case: The Effect of the Constraint Set Geometry

Kunal Talwar∗ Abhradeep Thakurta† Li Zhang‡

November 21, 2014

Abstract

Empirical Risk Minimization (ERM) is a standard technique in machine learning, where a model is selected by minimizing a loss function over constraint set. When the training dataset consists of private information, it is natural to use a differentially private ERM algorithm, and this problem has been the subject of a long line of work \cite{CM08, KST12, JKT12, ST13a, DJW13, JT14, BST14, Ull14}. A private ERM algorithm outputs an approximate minimizer of the loss function and its error can be measured as the difference from the optimal value of the loss function. When the constraint set is arbitrary, the required error bounds are fairly well understood \cite{BST14}. In this work, we show that the geometric properties of the constraint set can be used to derive significantly better results. Specifically, we show that a differentially private version of Mirror Descent leads to error bounds of the form $\tilde{O}\left(\frac{G_C}{n}\right)$ for a lipschitz loss function, improving on the $\tilde{O}\left(\sqrt{\frac{p}{n}}\right)$ bounds in \cite{BST14}. Here $p$ is the dimensionality of the problem, $n$ is the number of data points in the training set, and $G_C$ denotes the Gaussian width of the constraint set that we optimize over. We show similar improvements for strongly convex functions, and for smooth functions. In addition, we show that when the loss function is Lipschitz with respect to the $\ell_1$ norm and $C$ is $\ell_1$-bounded, a differentially private version of the Frank-Wolfe algorithm gives error bounds of the form $\tilde{O}(n^{-2/3})$. This captures the important and common case of sparse linear regression (LASSO), when the data $x_i$ satisfies $|x_i|_\infty \leq 1$ and we optimize over the $\ell_1$ ball. We show new lower bounds for this setting, that together with known bounds, imply that all our upper bounds are tight.

∗Email: kunal@kunaltalwar.org. (Part of this research was performed at the now defunct Microsoft Research Silicon Valley.)
†Yahoo Labs, Sunnyvale. abhradeep@yahoo-inc.com
‡Email: liqzhang@yahoo.com. (Part of this research was performed at the now defunct Microsoft Research Silicon Valley.)
1 Introduction

A common task in supervised learning is to select the model that best fits the data. This is frequently achieved by selecting a loss function that associates a real-valued loss with each datapoint \( d \) and model \( \theta \) and then selecting from a class of admissible models, the model \( \theta \) that minimizes the average loss over all data points in the training set. This procedure is commonly referred to as Empirical Risk Minimization (ERM).

The availability of large datasets containing sensitive information from individuals has motivated the study of learning algorithms that guarantee the privacy of individuals contributing to the database. A rigorous and by-now standard privacy guarantee is via the notion of differential privacy. In this work, we will study the design of differentially private algorithms for Empirical Risk Minimization, continuing a long line of work initiated by Chaudhuri and Monteleoni [CM08] and continued in [CMS11, KST12, JKT12, ST13a, DJW13, JT14, BST14, Ull14].

As an example, suppose that the training dataset \( D \) consists of \( n \) pairs of data \( d_i = (x_i, y_i) \) where \( x_i \in \mathbb{R}^p \), usually called the feature vector, and \( y_i \in \mathbb{R} \), the prediction. The goal is to find a “reasonable model” \( \theta \in \mathbb{R}^p \) such that \( y_i \) can be predicted from the model \( \theta \) and the feature vector \( x_i \). The quality of approximation is usually measured by a loss function \( \mathcal{L}(\theta; d_i) \), and the empirical loss is defined as \( \mathcal{L}(\theta; D) = \frac{1}{n} \sum_i \mathcal{L}(\theta; d_i) \).

For example, in the linear model with squared loss, \( \mathcal{L}(\theta; d_i) = (\langle \theta, x_i \rangle - y_i)^2 \). To avoid overfitting or to lend stability to the learning algorithm, it is common to add a regularization term on \( \theta \) or equivalently require \( \theta \) to come from a (typically convex) constraint set \( C \). This leads to the constrained optimization problem of computing \( \arg\min_{\theta \in C} \mathcal{L}(\theta; D) \). For example, in the classical sparse linear regression problem, we set \( C \) to be the \( \ell_1 \) ball. We formalize the problem as follows, and require differential privacy with respect to changes in a single \( d_i \).

**Problem definition:** Given a data set \( D = \{d_1, \cdots, d_n\} \) of \( n \) samples from a domain \( \mathcal{D} \), a convex set \( C \subseteq \mathbb{R}^p \), and a convex loss function \( \mathcal{L} : C \times \mathcal{D} \to \mathbb{R} \), the objective is to output a model \( \theta^{\text{priv}} \in C \), while preserving \((\epsilon, \delta)\)-differential privacy (Definition 2.1). We will define the privacy risk of a mechanism \( A \) as \( R(A) = \max_D \mathbb{E}[R(\theta^{\text{priv}})] \), where

\[
R(\theta) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \mathcal{L}(\theta; d_i) - \min_{\theta \in C} \frac{1}{n} \sum_{i=1}^n \mathcal{L}(\theta; d_i). \tag{1}
\]

We will design mechanisms that make the expected privacy risk as small as possible for any dataset \( D \) consisting of data records in \( \mathcal{D} \). Here the expectation is taken with respect to the internal randomness of the private mechanism.

Previous work on Private ERM has studied this problem under fairly general conditions. For convex loss functions \( \mathcal{L}(\theta; d_i) \) that for every \( d_i \) are 1-Lipschitz as functions from \( (\mathbb{R}^p, \ell_2) \) to \( \mathbb{R} \) (i.e. are Lipschitz in the first parameter with respect to the \( \ell_2 \) norm), and for \( C \) contained in the unit \( \ell_2 \) ball, Bassily, Smith and Thakurta [BST14] showed that the privacy risk is at most \( \tilde{O}(\sqrt{p}/n) \). They also showed that this bound cannot be improved in general, even for the squared loss function. Similarly they gave tight bounds under stronger assumptions on the loss functions (more details below).

In this work, we go beyond these worst-case bounds by exploiting properties of the constraint set \( C \). In the setting of the previous paragraph, we show that the \( \sqrt{p} \) term in the privacy risk can be replaced by the Gaussian Width of \( C \), defined as \( \bar{G}_C = \mathbb{E}_{g \in N(0,1)^p} [\sup_{\theta \in C} \langle \theta, p \rangle] \). Gaussian width is a well studied quantity in Convex Geometry for capturing the global geometry of \( C \). For a \( C \) contained in the the \( \ell_2 \) ball it is never larger than \( O(\sqrt{p}) \) and can be significantly smaller. For example, for the \( \ell_1 \) ball, the Gaussian width is only \( \Theta(\sqrt{\log p}) \). Similarly, we give improved bounds for other assumptions on the loss functions. These bounds are proved by analyzing a noisy version of the mirror descent algorithm [NY83, BT03].

\(^{1}\)Throughout the paper, we use \( \tilde{O}, \tilde{\Omega} \) to hide the polynomial factors in \( 1/\epsilon, \log(1/\delta), \log n, \) and \( \log p \).
In the simplest setting, when the loss function $\mathcal{L}(\cdot, d)$ is convex, and $L_2$-Lipschitz with respect to the $\ell_2$ norm on the parameter space, we get the following result. The precise bounds require a potential function that is tailored to the convex set $\mathcal{C}$. In the following, let $\|\mathcal{C}\|_2$ denote the $\ell_2$ radius of $\mathcal{C}$, and $G_{\mathcal{C}}$ denote the Gaussian width of $\mathcal{C}$.

**Theorem 1.1** (Informal version). There exists an $(\epsilon, \delta)$-differentially private algorithm $A$ such that

$$R(A) = O\left(\frac{L_2 G_{\mathcal{C}} \log(n/\delta)}{\epsilon n}\right).$$

In particular, $R(A) = O\left(\frac{L_2 \|\mathcal{C}\|_2 \sqrt{\log(n/\delta)}}{\epsilon n}\right)$, and if $\mathcal{C}$ is a polytope with $k$ vertices, $R(A) = O\left(\frac{L_2 \|\mathcal{C}\|_2 \sqrt{\log n/\delta}}{\epsilon n}\right)$.

Similar improvements can be shown (Section 3.2) for other constraint sets, such as those bounding the grouped $\ell_1$ norm, interpolation norms, or the nuclear norm when the vector is viewed as a matrix. When one additionally assumes that the loss functions satisfy a strong convexity definition (Appendix A), we can get further improved bounds. Moreover, for smooth loss functions (Section 4), we can show that a simpler objective perturbation algorithm $[CMS11, KST12]$ gives Gaussian-width-dependent bounds similar to the ones above. Our work also implies Gaussian-width-dependent convergence bounds for the noisy (stochastic) mirror descent algorithm, which may be of independent interest.

The bounds based on mirror descent have a dependence on the $\ell_2$ Lipschitz constant. This constant might be too large for some problems. For example, for the popular sparse linear regression problem, one often assumes $x_i$ to have bounded $\ell_\infty$ norm, i.e. each entry of $x_i$, instead of $\|x_i\|_2$, is bounded. The $\ell_2$ Lipschitz constant is then polynomial in $p$ and leads to a loose bound. In these cases, it would be more beneficial to have a dependence on the $\ell_1$ Lipschitz constant. Our next contribution is to address this issue.

We show that when $\mathcal{C}$ is the $\ell_1$ ball, one can get significantly better bounds using a differentially private version of the Frank-Wolfe algorithm. Let $\|\mathcal{C}\|_1$ denote the maximum $\ell_1$ radius of $\mathcal{C}$, and $\Gamma_{\mathcal{C}}$ the curvature constant for $\mathcal{L}$ (precise definition in Section 5).

**Theorem 1.2.** If $\mathcal{C}$ is a polytope with $k$ vertices, then there exists an $(\epsilon, \delta)$-differentially private algorithm $A$ such that

$$R(A) = O\left(\frac{\Gamma_{\mathcal{C}}^{1/3} (L_1 \|\mathcal{C}\|_1)^{2/3} \log(nk) \sqrt{\log(1/\delta)}}{(n\epsilon)^{2/3}}\right).$$

In particular, for the sparse linear regression problem where each $\|x_i\|_\infty \leq 1$, we have that

$$R(A) = O(\log(np/\delta)/(n\epsilon)^{2/3}).$$

Finally, we use the fingerprinting code lower bound technique developed in [BUV14] to show that the upper bound for the sparse linear regression problem, and hence the above result, is nearly tight.

**Theorem 1.3.** For the sparse linear regression problem where $\|x_i\|_\infty \leq 1$, for $\epsilon = 0.1$ and $\delta = 1/n$, any $(\epsilon, \delta)$-differentially private algorithm $A$ must have

$$R(A) \geq \Omega(1/(n \log n)^{2/3}).$$

In Table 1 we summarize our upper and lower bounds. Combining our results with that of [BST14], in particular we show that all the bounds in this paper are essentially tight. The lower bound for the $\ell_1$-norm case does not follow from [BST14], and we provide a new lower bound argument.

Our results enlarge the set of problems for which privacy comes “for free”. Given $n$ samples from a distribution, suppose that $\theta^*$ is the empirical risk minimizer and $\theta^{\text{priv}}$ is the differentially private approximate minimizer. Then the non-private ERM algorithm outputs $\theta^*$ and inures expected (on the distribution) loss.
| Assumption | Previous work | This work |
|------------|--------------|-----------|
| 1-Lipschitz w.r.t $L_2$-norm and $\|C\|_2 = 1$ | $\frac{\sqrt{p}}{cn}$ [BST14] | $\Omega\left(\frac{\sqrt{p}}{n}\right)$ [BST14] |
| ... and $\lambda$-smooth | $\frac{\sqrt{p+\lambda}}{cn}$ [CMS11] | $\Omega\left(\frac{\sqrt{p}}{n}\right)$ [BST14] (for $\lambda = O(p)$) |
| 1-Lipschitz w.r.t $L_1$-norm, $\|C\|_1 = 1$, and curvature $\Gamma$ | | |

Table 1: Upper and lower bounds for $(\epsilon, \delta)$-differentially private ERM. $k$ denotes the number of corners in the convex set $C$. (In general the dependence is on the Gaussian width of $C$, generalizing $\sqrt{p}$ or $\sqrt{\log k}$.) The curvature parameter is a weaker condition than smoothness, and is in particular bounded by the smoothness. Bounds ignore multiplicative dependence of $\log(1/\delta)$ and in the lower bounds, $\epsilon$ is considered as a constant. The lower bounds of [BST14] have the form $\Omega(\min\{n, \cdots\})$.

equal to the loss($\theta^*$, training − set) + generalization − error, where the generalization error term depends on the loss function, $C$ and on the number of samples $n$. The differentially private algorithm incurs an additional loss of the privacy risk. If the privacy risk is asymptotically smaller than the generalization error, we can think of privacy as coming for free, since under the assumption of $n$ being large enough to make the generalization error small, we are also making $n$ large enough to make the privacy risk small. For many of the problems, by our work we get privacy risk bounds that are close to the best known generalization bounds for those settings. More concretely, in the case when the $\|C\|_2 \leq 1$ and the loss function is 1-Lipschitz in the $\ell_2$-norm, the known generalization error bounds strictly dominates the privacy risk when $n = \omega(G^2_C)$ [SSSSS09, Theorem 7]. In the case when $C$ is the $\ell_1$-ball, and the loss function is the squared loss with $\|x\|_\infty \leq 1$ and $|y| \leq 1$, the generalization error dominates the privacy risk when $n = \omega(\log^3 p)$ [BM03, Theorem 18].

1.1 Related work

In the following we distinguish between the two settings: i) the convex set is bounded in the $\ell_2$-norm and the the loss function is 1-Lipschitz in the $\ell_2$-norm (call it the $(\ell_2/\ell_2)$-setting for brevity), and ii) the convex set is bounded in the $\ell_1$-norm and the the loss function is 1-Lipschitz in the $\ell_1$-norm (call it the $(\ell_1/\ell_\infty)$-setting).

The $(\ell_2/\ell_2)$-setting: In all the works on private convex optimization that we are aware of, either the excess risk guarantees depend polynomially on the dimensionality of the problem ($p$), or assumes special structure to the loss (e.g., generalized linear model [JT14] or linear losses [DNPR10, ST13b]). Similar dependence is also present in the online version of the problem [JKT12, ST13c]. [BST14] recently show that in the private ERM setting, in general this polynomial dependence on $p$ is unavoidable. In our work we show that one can replace this dependence on $p$ with the Gaussian width of the constraint set $C$, which can be much smaller. We use the mirror descent algorithm of Beck and Teboulle [BT03] as our building block.

The $(\ell_1/\ell_\infty)$-setting: The only two results in this setting that we are aware of are by Kifer, Smith and
They achieve excess empirical risk guarantees that depend logarithmically in the dimensions $p$, and thus allowing the guarantees to be meaningful even when $p \gg n$. In fact under certain assumptions on the data (restricted strong convexity (RSC) and mutual incoherence), they obtain privacy risk guarantees that scale as $O(\text{polylog } p/n)$, which is better than our asymptotic dependence of $O(\text{polylog } p/n^{2/3})$. However, arguably these assumptions on the data are strong and may not hold in practice [Was12]. In our work, we provide guarantees which do not require any such data dependent assumptions. We should mention here that both [KST12] and [ST13a] address a harder problem of support recovery, i.e., to figure out a small set of features, optimizing over which will result in low privacy risk. For the specific problem of support recovery, [KST12] provides a computationally inefficient algorithm using the exponential mechanism of McSherry and Talwar [MT07] which does not depend on RSC or mutual incoherence.

**Effect of Gaussian width in risk minimization:** For linear losses, the notions of Rademacher complexities and Gaussian complexities are closely related to the concept of Gaussian width, i.e., when the loss function are of the form $\langle \theta, d \rangle$. One of the initial works that formalized this connection was by Bartlett and Mendelson [BM03]. They in particular bound the excess generalization error by the Gaussian complexity of the constraint set $C$, which is very similar to Gaussian width in the context of linear functions. Recently Chandrashekharan et al. [CRPW12] show that the Gaussian complexity of a constraint set $C$ is very closely related to the number of generic linear measurements one needs to perform to recover an underlying model $\theta^* \in C$.

Shamir and Zhang [SZ13] analyzed the problem of noisy stochastic gradient descent (SGD) for general convex loss functions. Their empirical risk guarantees depend polynomially on the $\ell_2$-norm of the noise vector that gets added during the gradient computation in the SGD algorithm. As a corollary of our results we show that if the noise vector is sub-Gaussian (not necessarily spherical), the polynomial dependence on the $\ell_2$-norm of the noise can be replaced by the Gaussian width of the constraint set.

We should note here that the notion of Gaussian width has been used by Nikolov, Talwar and Zhang [NTZ13], and Dwork, Nikolov and Talwar [DNT13] in the context of differentially private query release mechanisms but in the very different context of answering multiple linear queries over a database.

## 2 Background

### 2.1 Differential Privacy

The notion of differential privacy (Definition 2.1) is by now a defacto standard for statistical data privacy [Dwo06, Dwo08, Dwo09]. One of the reasons for which differential privacy has become so popular is because it provides meaningful guarantees even in the presence of arbitrary auxiliary information. At a semantic level, the privacy guarantee ensures that an adversary learns almost the same thing about an individual independent of his presence or absence in the data set. The parameters ($\epsilon, \delta$) quantify the amount of information leakage. For reasons beyond the scope of this work, $\epsilon \approx 0.1$ and $\delta = 1/n^{\omega(1)}$ are a good choice of parameters. Here $n$ refers to the number of samples in the data set.

**Definition 2.1.** A randomized algorithm $A$ is $(\epsilon, \delta)$-differentially private ([DMNS06, DKS+06]) if, for all neighboring data sets $D$ and $D'$ (i.e., they differ in one record, or equivalently, $d_H(D, D') = 1$) and for all events $S$ in the output space of $A$, we have

$$\Pr(A(D) \in S) \leq e^\epsilon \Pr(A(D') \in S) + \delta.$$

Here $d_H(D, D')$ refers to the Hamming distance.
2.2 Bregman Divergence, Convexity, Norms, and Gaussian Width

In this section we review some of the concepts commonly used in convex optimization, which will be useful in the exposition of our algorithms. In all the definitions below we assume that the set \( C \subseteq \mathbb{R}^p \) is closed and convex.

\( \ell_q \)-norm, \( q \geq 1 \): For \( q \geq 1 \), the \( \ell_q \)-norm for any vector \( v \in \mathbb{R}^p \) is defined as \( \left( \sum_{i=1}^{p} v(i)^q \right)^{1/q} \), where \( v(i) \) is the \( i \)-th coordinate of the vector \( v \).

**Minkowski norm w.r.t a set** \( C \subseteq \mathbb{R}^p \): For any vector \( v \in \mathbb{R}^p \), the Minkowski norm (denoted by \( \|v\|_C \) w.r.t. a centrally symmetric convex set \( C \subseteq \mathbb{R}^p \) is defined as follows.

\[
\|v\|_C = \min \{ r \in \mathbb{R} : v \in rC \}.
\]

**L-Lipschitz continuity w.r.t. norm** \( \| \cdot \|_p \): A function \( \Psi : C \rightarrow \mathbb{R} \) is L-Lipschitz within a set \( C \) w.r.t. a norm \( \| \cdot \|_p \) if the following holds.

\[
\forall \theta_1, \theta_2 \in C, |\Psi(\theta_1) - \Psi(\theta_2)| \leq L \cdot \|\theta_1 - \theta_2\|.
\]

**Convexity and \( \Delta \)-strong convexity w.r.t norm** \( \| \cdot \|_p \): A function \( \Psi : C \rightarrow \mathbb{R} \) is convex if

\[
\forall \theta_1, \theta_2 \in C, \alpha \in [0, 1], \Psi(\alpha \theta_1 + (1 - \alpha) \theta_2) \leq \alpha \Psi(\theta_1) + (1 - \alpha) \Psi(\theta_2).
\]

A function is \( \Delta \)-strongly convex within a set \( C \) w.r.t. a norm \( \| \cdot \| \) if

\[
\forall \theta_1, \theta_2 \in C, \alpha \in [0, 1], \Psi(\alpha \theta_1 + (1 - \alpha) \theta_2) \leq \alpha \Psi(\theta_1) + (1 - \alpha) \Psi(\theta_2) - \Delta \cdot \alpha(1 - \alpha) \frac{1}{2} \|\theta_1 - \theta_2\|^2.
\]

**Bregman divergence**: For any convex function \( \Psi : \mathbb{R}^p \rightarrow \mathbb{R} \), the Bregman divergence defined by \( B_{\Psi} : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R} \) is defined as

\[
B_{\Psi}(\theta_1, \theta_2) = \Psi(\theta_1) - \Psi(\theta_2) - \langle \nabla \Psi(\theta_2), \theta_1 - \theta_2 \rangle.
\]

Notice that Bregman divergence is always positive, and convex in the first argument. It need not be symmetric.

**\( \Delta \)-strong convexity w.r.t a function** \( \Psi \): A function \( f : C \rightarrow \mathbb{R} \) is \( \Delta \)-strongly convex within a set \( C \) w.r.t. a differentiable convex function \( \Psi \) if the following holds.

\[
\forall \theta_1, \theta_2 \in C, \alpha \in [0, 1], f(\alpha \theta_1 + (1 - \alpha) \theta_2) \leq \alpha f(\theta_1) + (1 - \alpha) f(\theta_2) - \frac{\Delta \cdot \alpha(1 - \alpha)}{2} B_{\Psi}(\theta_1, \theta_2).
\]

**Duality**: The following duality property (Fact 2.2) of norms will be useful through the rest of this paper. Recall that for any pair of dual norms \( \| \cdot \|_a \) and \( \| \cdot \|_b \), and \( x, y \in \mathbb{R}^p \), Holder’s inequality says that

\[
|\langle x, y \rangle| \leq \|x\|_a \cdot \|y\|_b.
\]

**Fact 2.2.** The dual of \( \ell_q \) norm is \( \ell_{q'} \)-norm such that \( 1/q + 1/q' = 1 \). The dual of \( \| \cdot \|_C \) is \( \| \cdot \|_{C^*} \), where for any vector \( v \in \mathbb{R}^p \), \( \|v\|_{C^*} = \max_{w \in C} |\langle w, v \rangle| \).

**Gaussian width of a set** \( C \): Let \( b \sim \mathcal{N}(0, I_p) \) be a Gaussian random vector in \( \mathbb{R}^p \). The Gaussian width of a set \( C \) is defined as \( G_C \overset{\text{def}}{=} \mathbb{E}_b \left[ \sup_{w \in C} |\langle b, w \rangle| \right] \).
3 Private Mirror Descent and the Geometry of $C$

In this section we introduce the well-established mirror descent algorithm [NY83] in the context of private convex optimization. We notice that since mirror descent is designed to closely follow the geometry of the convex set $C$, we get much tighter bounds than that were known earlier in the literature for a large class of interesting instantiations of the convex set $C$. More precisely, using private mirror descent one can show that the privacy depends on the Gaussian width (see Section 2.2) instead of any explicit dependence on the dimensionality $p$.

3.1 Private Mirror Descent Method

In Algorithm 1 we define our private mirror descent procedure. The algorithm takes as input a potential function $\Psi$ that is chosen based on the constraint set $C$. $B_\Psi$ refers to the Bregman divergence with respect to $\Psi$. (See Section 2.2.) If $L(\theta; d)$ is not differentiable at $\theta$, we use any sub-gradient at $\theta$ instead of $\nabla L(\theta; d)$.

Algorithm 1 $\mathcal{A}_{\text{Noise-MD}}$: Differentially Private Mirror Descent

**Input:** Data set: $D = \{d_1, \ldots, d_n\}$, loss function: $L(\theta; D) = \frac{1}{n} \sum_{i=1}^{n} L(\theta; d_i)$ (with $\ell_2$-Lipschitz constant $L$ for $L$), privacy parameters: $(\epsilon, \delta)$, convex set: $C$, potential function: $\Psi : C \to \mathbb{R}$, and learning rate: $\eta : [T + 1] \to \mathbb{R}$.

1. Set noise variance $\sigma^2 \leftarrow \frac{32L^2T\log^2(T/\delta)}{(\epsilon n)^2}$.
2. Let $\theta_1$ be an arbitrary point in $C$.
3. **for** $t = 1$ to $T$ **do**
4. \hspace{1em} $\theta_{t+1} = \arg\min_{\theta \in C} \left( \frac{1}{\eta_{t+1}} (\nabla L(\theta_t; D) + b_t) - \nabla \Psi(\theta_t) + \Psi(\theta) \right)$, where $b_t \sim \mathcal{N}(0, \mathbb{I}_p\sigma^2)$.
5. **Output** $\theta_{\text{priv}} \leftarrow \frac{1}{T} \sum_{t=1}^{T} \theta_t$.

**Theorem 3.1** (Privacy guarantee). Algorithm 1 is $(\epsilon, \delta)$-differentially private.

The proof of this theorem is fairly straightforward and follows from by now standard privacy guarantee of Gaussian mechanism [NTZ13], and the strong composition theorem [DRV10]. For a detailed proof, we refer the reader to [BST14, Theorem 2.1]. To establish the utility guarantee in a general form, it will be useful to use a convex body $Q$ on which the potential function $\Psi$ is defined. We will instantiate this theorem with various choices of $Q$ and $\Psi$ depending on $C$ in Section 3.2. In most of the cases, we will set $Q = C$ and the reader may find it convenient to think of that case. Our proof of the theorem below closely follows the analysis of mirror descent from [ST10].

One can obtain stronger guarantees (typically, $O(1/(\eta n)^2)$) under strong convexity assumptions on the loss function. We defer the details of this result to Appendix A.

**Theorem 3.2** (Utility guarantee). Suppose that for any $d \in D$, the loss function $L(\cdot; d)$ is convex and $L$-lipschitz with respect to the $\ell_2$ norm. Let $Q \subseteq \mathbb{R}^p$ be a symmetric convex set with Gaussian width $G_Q$ and $\ell_2$-diameter $\|Q\|_2$, and let $\Psi : Q \to \mathbb{R}$ be $1$-strongly convex w.r.t. $\|\cdot\|_Q$-norm chosen in Algorithm $\mathcal{A}_{\text{Noise-MD}}$ (Algorithm 1). If $T = \frac{\|Q\|_2^2 \mathbb{E}^2 n^2}{L^2 \log^2(n/\delta) G_Q^2}$ and for all $t \in [T + 1]$, $\eta_t = \eta = \frac{1}{L\|Q\|_2 \sqrt{T}}$, then following is true.

$$
\mathbb{E} \left[ L(\theta_{\text{priv}}; D) \right] - \min_{\theta \in C} L(\theta; D) = O \left( \frac{LG_Q \sqrt{\max_{\theta \in C} \Psi(\theta) \log(n/\delta)}}{\epsilon n} \right).
$$
Proof of Theorem 3.2: For the ease of notation we ignore the parameterization of $\mathcal{L}(\theta; D)$ on the data set $D$ and simply refer to as $\mathcal{L}(\theta)$. To begin with, from a direct application of Jensen’s inequality, we have the following.

$$
\mathcal{L}(\theta^{priv}) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta) \leq \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}(\theta_t) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta) \quad (2)
$$

So it suffices to bound the R.H.S. of (2) in order to bound the excess empirical risk. In Claim 3.3, we upper bound the R.H.S. of (2) by a sequence of linear approximations of $\mathcal{L}(\theta)$, thus “linearizing” our analysis.

Claim 3.3. Let $\theta^* = \arg \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta)$. For every $t \in [T]$, let $\gamma_t$ be the sub-gradient of $\mathcal{L}(\theta_t)$ used in iteration $t$ of Algorithm $\mathcal{A}_{Noise-MD}$ (Algorithm 7). Then the convexity of the loss function implies that

$$
\frac{1}{T} \sum_{t=1}^{T} \mathcal{L}(\theta_t) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta) \leq \frac{1}{T} \sum_{t=1}^{T} \langle \gamma_t, \theta_t - \theta^* \rangle. 
$$

Thus it suffices to bound $\frac{1}{T} \sum_{t=1}^{T} \langle \gamma_t, \theta_t - \theta^* \rangle$ in order to bound the privacy risk. By simple algebraic manipulation we have the following. (Recall that $b_t$ is the noise vector used in Algorithm $\mathcal{A}_{Noise-MD}$.)

$$
\eta(\gamma_t + b_t, \theta_t - \theta^*) = A + B + C.
$$

We next upper bound each of the terms $A, B$ and $C$ in (3). By Holder’s inequality, we write

$$
A \leq \|\theta_t - \theta_{t+1}\|_Q \|\eta(\gamma_t + b_t)\|_{Q^*}
$$

$$
\leq \frac{1}{2} \|\theta_t - \theta_{t+1}\|_Q^2 + \frac{\eta^2}{2} \|\gamma_t + b_t\|_Q^2.
$$

$$
\leq \frac{1}{2} \|\theta_t - \theta_{t+1}\|_Q^2 + \frac{\eta^2}{2} (\|\gamma_t\|_{Q^*} + \|b_t\|_{Q^*})^2.
$$

(4)

where we have used the A.M-G.M. inequality in the second step, and the triangle inequality in the third step. Taking expectations over the choice of $b_t$, and using Jensen’s inequality and the definition of Gaussian width, we conclude that

$$
\mathbb{E}_{b_t}[A] \leq \frac{1}{2} \mathbb{E}_{b_t} \left[\|\theta_t - \theta_{t+1}\|_Q^2\right] + \frac{\eta^2}{2} (L\|Q\|_2 + \sigma G_Q)^2. 
$$

(5)

We next proceed to bound the term $B$ in (3). By the definition of $\theta_{t+1}$, it follows that

$$
\langle \eta(\gamma_t + b_t) - \nabla \Psi(\theta_t), \theta_{t+1} \rangle + \Psi(\theta_{t+1}) \leq \langle \eta(\gamma_t + b_t) - \nabla \Psi(\theta_t), \theta^* \rangle + \Psi(\theta^*).
$$
This implies that
\[
B \leq -\Psi(\theta_{t+1}) + \Psi(\theta^*) + \langle \nabla \Psi(\theta_{t+1}), \theta_{t+1} - \theta^* \rangle \\
= -B_{\Psi}(\theta_{t+1}, \theta^*) \leq 0. \tag{6}
\]

One can write the term \( C \) in (3) as follows.
\[
B_{\Psi}(\theta^*, \theta_t) - B_{\Psi}(\theta^*, \theta_{t+1}) - B_{\Psi}(\theta_{t+1}, \theta_t) = \Psi(\theta^*) - \Psi(\theta_t) - \langle \nabla \Psi(\theta_t), \theta^* - \theta_t \rangle \\
- \Psi(\theta^*) + \Psi(\theta_{t+1}) + \langle \nabla \Psi(\theta_{t+1}), \theta^* - \theta_{t+1} \rangle \\
- \Psi(\theta_{t+1}) + \Psi(\theta_t) + \langle \nabla \Psi(\theta_t), \theta_{t+1} - \theta_t \rangle = C \tag{7}
\]

Notice that since \( b_i \) is independent of \( \theta_t, \mathbb{E}[(b_t, \theta_t - \theta^*)] = 0. \) Plugging the bounds (5), (6) and (7) in (3), we have the following.
\[
\eta \mathbb{E}[(\gamma_t, \theta_t - \theta^*)] = \eta \mathbb{E}[(\gamma_t + b_t, \theta_t - \theta^*)] \\
\leq B_{\Psi}(\theta^*, \theta_t) - B_{\Psi}(\theta^*, \theta_{t+1}) + \eta^2 \frac{1}{2}(L\|Q\|_2 + \sigma G_Q)^2 + \frac{1}{2} \|\theta_t - \theta_{t+1}\|_Q^2 - B_{\Psi}(\theta_{t+1}, \theta_t) \tag{8}
\]

In order to bound the term \( D \) in (8), we use the assumption that \( \Psi(\theta) \) is 1-strongly convex with respect to \( \| \cdot \|_Q \). This immediately implies that in (8) \( D \leq 0 \). Using this bound, summing over all \( T \)-rounds, we have
\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[(\gamma_t, \theta_t - \theta^*)] \leq \frac{\max_{\theta \in C} \Psi(\theta)}{\eta T} + \eta \left( L\|Q\|_2 + \sigma G_Q \right)^2 \tag{9}
\]

In the above we used the following property of Bregman divergence: \( B_{\Psi}(\theta^*, \theta_1) \leq \max_{\theta \in C} \Psi(\theta) \). We can prove this fact as follows. Let \( \theta^\dagger = \arg\min_{\theta \in C} \Psi(\theta) \). By the generalized Pythagorean theorem \([Rak09, \text{Chapter 2}]\), it follows that \( B_{\Psi}(\theta^*, \theta_1) \leq B_{\Psi}(\theta^*, \theta^\dagger) \leq B_{\Psi}(\theta^*, \theta^\dagger) \). The last inequality follows from the fact that Bregman divergence is always non-negative. Now since \( \theta^\dagger \) minimizes \( \Psi \) and \( \Psi \) is convex, it follows that \( \langle \nabla \Psi(\theta^\dagger), \theta^* - \theta^\dagger \rangle \geq 0 \). This immediately implies \( B_{\Psi}(\theta^*, \theta^\dagger) \leq \Psi(\theta^*) \leq \max_{\theta \in C} \Psi(\theta) \).

Setting \( T = \frac{\|Q\|_2^2 \eta^2 n^2}{\log^2(n/\delta) G_Q} \) and \( \eta = \frac{\max_{\theta \in C} \Psi(\theta)}{L\|Q\|_2 \sqrt{T}} \), and using (2) and Claim 3.3 we get the required bound. \( \square \)

### 3.2 Instantiation of Private Mirror Descent to Various Settings of \( C \)

In this section we discuss some of the instantiations of Theorem 3.2.

**For arbitrary convex set \( C \subseteq \mathbb{R}^p \) with \( L_2 \)-diameter \( \|C\|_2 \):** Let \( \Psi(\theta) = \frac{1}{2}\|\theta - \theta_0\|_2^2 \) (with some fixed \( \theta_0 \in C \)) and we choose the convex set \( Q \) to be the unit \( L_2 \)-ball in Theorem 3.2. Immediately, we obtain the following as a corollary.
\[
\mathbb{E} [\mathcal{L}(\theta_{\text{priv}}; D)] - \min_{\theta \in C} \mathcal{L}(\theta; D) = O \left( \frac{L \sqrt{p}\|C\|_2 \log(n/\delta)}{\epsilon n} \right). \tag{10}
\]

This is a slight improvement over \([BST14]\).

**For the convex set \( C \subseteq \mathbb{R}^p \) being a polytope:** Let \( C = \text{conv}\{v_1, \ldots, v_k\} \) be a convex polytope formed by the convex combination of vectors \( v_i \in \mathbb{R}^p \) such that for all \( i \in [p], \|v_i\|_2 \leq \|C\|_2 \). Here \( \text{conv}(S) \) corresponds to all the possible convex combinations of the vectors in the set \( S \). Observation 3.4 will be very useful for choosing the correct potential function \( \Psi \) in Algorithm \( A_{\text{Noise-MD}} \) (Algorithm 1).
Fact 3.4 (From [SST11]). For the convex set $C$ defined above, let $Q$ be the convex hull of $C$ and $-C$. The Minkowski norm for any $\theta \in \mathbb{R}^p$ is given by $\|\theta\|_Q = \inf_{\alpha_1, \ldots, \alpha_k, \sum \alpha_i \cdot \alpha_i = \theta} \left( \sum_{i=1}^k |\alpha_i|^q \right)^{1/q}$ be a norm for any $q \in (1, 2]$. Then the function $\Psi(\theta) = \frac{1}{4(q-1)}\|\theta\|_{Q,q}^2$ is $1$-strongly convex w.r.t. $\|\cdot\|_Q$-norm.

In the following we state the following claim which will be useful later.

Claim 3.5. If $q = \frac{\log k}{\log k - 1}$, then the following is true for any $\theta \in \mathbb{R}^p$: $\|\theta\|_Q \leq e \cdot \|\theta\|_{Q,q}$.

Proof. First notice that for any vector $v = (v_1, \cdots, v_k)$, $\|v\|_1 \leq k^{1-1/q}\|v\|_q$. This follows from Holder’s inequality. Now setting $q = \log k/(\log k - 1)$, we get $\|v\|_1 \leq e \cdot \|v\|_q$. For any $\theta \in \mathbb{R}^p$, let $a = (\alpha_1, \cdots, \alpha_k)$ be the vector of parameters corresponding to $\|\theta\|_Q$. From the above, we know that $\|a\|_1 \leq e \cdot \|a\|_q$. And by definition, we know that $\|\theta\|_Q \leq \|a\|_1$. This completes the proof. □

Claim 3.5 implies that if $\Psi(\theta) = \frac{1}{2(q-1)}\|\theta\|_{Q,q}^2$ then $\max_{\theta \in C} \Psi(\theta) = O(\log k)$. Additionally due to Fact 3.4 $\Psi(\theta)$ is $O(1)$-strongly convex w.r.t. $\|\cdot\|_Q$. With the above observations, and observing that $G_Q = O(\|C\|_2 \sqrt{\log k})$, setting $Q$ and $\Psi$ as above, we immediately get the following corollary of Theorem 3.2

Notice that the bound does not have any explicit dependence on the dimensionality of the problem.

$$
\mathbb{E} \left[ \mathcal{L}(\theta^{\text{priv}}; D) \right] - \min_{\theta \in C} \mathcal{L}(\theta; D) = O \left( \frac{L\|C\|_2 \sqrt{\log k \log(n/\delta)}}{en} \right). \tag{11}
$$

Notice that this result extends to the standard $p$-dimensional probability simplex: $C = \{ \theta \in \mathbb{R}^p : \sum_{i=1}^p \theta_i = 1, \forall i \in [p], \theta_i \geq 0 \}$. In this case, the only difference is that the term $\log k$ gets replaced by $\log p$ in (11).

For grouped $\ell_1$-norm: For a vector $x \in \mathbb{R}^p$ and a parameter $k$, the grouped $\ell_1$-norm defined as $\|\theta\|_{k,\ell_1,2} = \sum_{i=1}^{[p/k]} \left( \sum_{j=(i-1)k+1}^{ik} |\theta_j|^2 \right)$. If $C$ denotes the convex set centered at zero with radius one with respect to $\|\cdot\|_{(k,\ell_1,2)}$-norm, then it follows from union bound on each of the blocks of coordinates in $[p]$ that $G_C = \sqrt{k \log(p/k)}$. In the following we propose the following choices of $\Psi$ depending on the parameter $k$. (These choices are based on [BTN13, Section 5.2.3].) For a given $M > 1$, divide the coordinates of $\theta$ into $M$ blocks, and denote each block as $\theta^{(j)}$.

$$
\Psi(\theta) = \frac{1}{M \xi} \sum_{j=1}^M \|\theta^{(j)}\|_2^M, M = \begin{cases} 
2, & \text{if } [p/k] \leq 2 \\
1 + 1/(\log(p/k)), & \text{otherwise}
\end{cases}, \xi = \begin{cases} 
1, & \text{if } [p/k] = 1 \\
1/2, & \text{if } [p/k] = 2 \\
1/(e \log(p/k)), & \text{otherwise}
\end{cases}
$$

With this setting of $\Psi(\theta)$ one can show that $\max_{\theta \in C} \Psi(\theta) = O(\sqrt{\log(p/k)})$. Plugging these bounds in Theorem 3.2 we get (13) as a corollary.

$$
\mathbb{E} \left[ \mathcal{L}(\theta^{\text{priv}}; D) \right] - \min_{\theta \in C} \mathcal{L}(\theta; D) = O \left( \frac{L \sqrt{k \log^2(p/k)} \log(n/\delta)}{en} \right). \tag{12}
$$
Similar bounds can be achieved for other forms of interpolation norms, e.g., \( L_1, L_2 \)-interpolation norms:

\[
\|\theta\|_{\alpha, \text{inter}(\ell_1, \ell_2)} = (1 - \alpha)\|\theta\|_1 + \alpha\|\theta\|_2 \quad \text{with} \quad \alpha \in [0, 1].
\]

Notice that since the set \( C = \{ \theta : \|\theta\|_{\alpha, \text{inter}(\ell_1, \ell_2)} \leq 1 \} \) is a subset of \( C_1 + C_2 \), where \( C_1 = \{ (1 - \alpha)\theta : \|\theta\|_1 \leq 1 \} \) and \( C_2 = \{ \alpha\theta : \|\theta\|_2 \leq 1 \} \), it follows that the Gaussian width \( G_C \leq G_{C_1} + G_{C_2} = O((1 - \alpha)\sqrt{\log p} + \alpha\sqrt{p}) \). Additionally from [SST11] it follows that there exists a function \( \Psi(\theta) \) strongly convex w.r.t. \( \|\cdot\|_C \) such that it is at most a constant within the set \( C \).

While using Theorem 3.2 in both of the above settings, we set the convex set \( Q = C \).

**For low-rank matrices:** It is known that the non-private mirror descent extends immediately to matrices [ BTN13]. In the following we show that this is also true for the private mirror descent algorithm in Algorithm 4 (Algorithm \( A_{\text{Noise-MD}} \)). For the matrix setting, we assume \( \theta \in \mathbb{R}^{p \times p} \) and the loss function \( \mathcal{L}(\theta; d) \) is \( L \)-Lipschitz in the Frobenius norm \( \|\cdot\|_F \). From [DTTZ14] it follows that if the noise vector \( b \) in Algorithm \( A_{\text{Noise-MD}} \) is replaced by a matrix \( b \in \mathbb{R}^{p \times p} \) with each entry of \( b \) drawn i.i.d. from \( \mathcal{N}(0, \sigma^2) \) (with the standard deviation \( \sigma \) being the same as in Algorithm \( A_{\text{Noise-MD}} \)), then the \( (\epsilon, \delta) \)-differential privacy guarantee holds. In the following we instantiate Theorem 3.2 for the class of \( m \times m \) real matrices with nuclear norm at most one. Call it the set \( C \). (For a matrix \( \theta, \|\theta\|_{\text{nuc}} \) refers to the sum of the singular values of \( \theta \).) This class is the convex hull of rank one matrices with unit euclidean norm. [CRPW12, Proposition 3.11] shows that the Gaussian width of \( C \) is \( O(\sqrt{m}) \). [ BTN13, Section 5.2.3] showed that the function \( \Psi(\theta) = \frac{4}{\sqrt{e}} \frac{\log(2m)}{2^q(1 + q)} \sum_{i=1}^{m} \sigma_i^{1+q}(\theta) \) with \( q = \frac{1}{2\log(2m)} \) is 1-strongly convex w.r.t. \( \|\cdot\|_{\text{nuc}} \)-norm. Moreover, \( \max_{\theta \in C} \Psi(\theta) = O(\log m) \). Plugging these bounds in Theorem 3.2, we immediately get the following excess empirical risk guarantee.

\[
\mathbb{E} \left[ \mathcal{L}(\theta^{\text{priv}}; D) - \min_{\theta \in C} \mathcal{L}(\theta; D) \right] = O \left( \frac{L\sqrt{m} \log m \log(n/\delta)}{\epsilon n} \right).
\]

### 3.3 Convergence Rate of Noisy Mirror Descent

In this section we analyze the excess empirical risk guarantees of Algorithm 4 (Algorithm \( A_{\text{Noise-MD}} \)) as a purely noisy mirror descent algorithm, and ignoring privacy considerations. Let us assume that the oracle that returns the gradient computation is noisy. In particular each of the \( b_t \) (in Line 4 of Algorithm \( A_{\text{Noise-MD}} \)) is drawn independently from distributions which are mean zero and sub-Gaussian with variance \( \Sigma_{p \times p} \), where \( \Sigma \) is the covariance matrix. For example, this may be achieved by sampling a small number of \( d_i \)'s and averaging \( \nabla \mathcal{L}(\theta_t; d_t) \) over the sampled values. Using the same proof technique of Theorem 3.2 and the observation that \( \mathbb{E}_{b \sim \mathcal{N}(0, I_p)} \max_{\theta \in C} \left| \langle \sqrt{\Sigma} \cdot b, \theta \rangle \right| = O(\sqrt{\lambda_{\max}(\Sigma)G_C}) \), we obtain the following corollary of Theorem 3.2. Here \( \lambda_{\max} \) corresponds to the maximum eigenvalue and we set the convex set \( Q = C \) in Theorem 3.2 for the ease of exposition.

**Corollary 3.6 (Noisy mirror descent guarantee).** Let \( C \subseteq \mathbb{R}^p \) be a symmetric convex set with its \( \ell_2 \) diameter and Gaussian width represented by \( \|C\|_2 \) and \( G_C \) respectively, and let \( \Psi : C \to \mathbb{R} \) be an \( 1 \)-strongly convex function w.r.t. \( \|\cdot\|_C \)-norm chosen in Algorithm \( A_{\text{Noise-MD}} \) (Algorithm 7). For any \( d \in D \), suppose that the loss function \( \mathcal{L}(\theta; d) \) is convex and \( L \)-Lipschitz with respect to the \( \ell_2 \) norm. If for all \( t \in [T+1] \), \( \eta_t = \eta = \frac{1}{(L\|C\|_2 + \sqrt{\lambda_{\max}(\Sigma)})\sqrt{T}} \), then the following is true.

\[
\mathbb{E} \left[ \mathcal{L}(\theta^{alg}; D) - \min_{\theta \in C} \mathcal{L}(\theta; D) \right] = O \left( \sqrt{\max_{\theta \in C} \Psi(\theta)} \left( L\|C\|_2 + \sqrt{\lambda_{\max}(\Sigma)G_C} \right) / \sqrt{T} \right).
\]

Here the expectation is over the randomness of the algorithm.
Corollary [3.6] above improves on the bound obtained in the noisy gradient descent literature [SZ13 Theorem 2] as long as the noise follows the mean zero sub-Gaussian distribution mentioned above and the potential function $\Psi$ exists. In particular it improves on the dependence on dimension by removing any explicit dependence on $p$. For different settings of $\Psi$ depending on the convex set $C$, see Section [3.2]

### 4 Objective Perturbation for Smooth Functions

In this section we show that if the loss function $L$ is twice continuously differentiable, then one can recover similar bounds as in Section 3 using by now well-established objective perturbation algorithm of [CMST11, KST12]. The main contribution in this section is a tighter analysis of objective perturbation using Gaussian width. In the following (Algorithm 2) we first revisit the objective perturbation algorithm from [KST12]. The $(\epsilon, \delta)$-differential privacy guarantee follows from [KST12]. Theorem 4.1 shows privacy risk bounds that are similar to that in Section 3.

**Remark 1.** The smoothness property of the loss function $L$ is used in the privacy analysis. It can be shown that this is in some sense necessary. (See [KST12] for a more detailed discussion.) Standard approaches towards smoothing (like convolving with a smooth function) adversely affects the utility guarantee and results in sub-optimal dependence on the number of data samples ($n$). (See [BST14, Appendix E].)

**Remark 2.** We define the set $Q$ as in Theorem 4.1 because we want to both symmetrize and extend the convex set $C$ to a full-dimensional space. For example, think of the probability simplex in $p$-dimensions as the set $C$, and $Q$ to be the $\ell_1$-ball. Also when there exists a differentiable convex function $\Psi : C \rightarrow \mathbb{R}$ such that $\Psi$ is 1-strongly convex w.r.t. $\|\cdot\|_Q$ and the guarantee in Theorem A.1 holds w.r.t. $\Psi$, then Theorem 4.1 is a special case of Theorem A.1. This in particular captures the following cases: i) $\Psi(\theta) = \frac{1}{2}\|\theta\|_2^2$ (and correspondingly $Q$ being the $\ell_2$-ball), and ii) $\Psi(\theta) = \sum_{i=1}^p \theta_i \log \theta_i$ (and correspondingly $Q$ being the $\ell_1$-ball).

**Algorithm 2 Objective Perturbation [KST12]**

**Input:** Data set: $D = \{d_1, \ldots, d_n\}$, loss function: $L(\theta; D) = \frac{1}{n} \sum_{i=1}^n L(\theta; d_i)$ (with $\ell_2$-Lipschitz constant $L$ for $C$), privacy parameters: $(\epsilon, \delta)$, convex set: $C$ (denote the diameter in $\ell_2$-norm by $\|C\|_2$), upper and lower bounds $\lambda_{\max}, \lambda_{\min}$ on the eigenvalues of $\nabla^2 L(\theta; d)$ (for all $d$ and for all $\theta \in C$).

1. Set $\zeta = \max \left\{ \frac{2\lambda_{\max}}{n\epsilon} - \min_{\theta \in C, d \in D} \lambda_{\min}(\nabla^2 L(\theta; d)), 0 \right\}$.

2. Output $\theta^{\text{priv}} \leftarrow \arg \min_{\theta \in C} L(\theta; D) + \frac{\zeta}{2} \|\theta - \theta_0\|_2^2 + \langle b, \theta \rangle$, where $b \sim N(0, \frac{L^2(2\log(1/\delta))}{(n\epsilon)^2} I_{p \times p})$ and $\theta_0 \in C$ is fixed.

**Theorem 4.1 (Utility guarantee).** Suppose that $C \subseteq \mathbb{R}^p$ has diameter $\|C\|_2$ and Gaussian width $G_C$. Further suppose that for all $d \in D$, the loss function $L(\cdot; d)$ is twice continuously differentiable, and for all $\theta \in C$, $\|\nabla^2 L(\theta; d)\|$ has spectral norm at most $\lambda_{\max}$. Then Algorithm 2 satisfies the following guarantees.

1. **Lipschitz case:** Suppose that for any $d \in D$, the loss function $L(\cdot; d)$ is convex and $L$-lipschitz w.r.t. the $\ell_2$ norm. Then

$$E \left[ L(\theta^{\text{priv}}; D) \right] - \min_{\theta \in C} L(\theta; D) = O \left( \frac{L G_C \sqrt{\log(1/\delta)} + \lambda_{\max} \|C\|_2^2}{en} \right).$$


2. **Lipschitz and strongly convex case:** Suppose that for any \( d \in D \), the loss function \( L(\cdot; d) \) is \( L \)-lipschitz in the \( \ell_2 \) norm, and \( \Delta \)-strongly convex with respect to \( \| \cdot \|_Q \), where \( Q \) is the symmetric convex hull of \( C \). If \( \Delta \geq \frac{2\|C\|_2^2\lambda_{\max}}{\eta n} \), then the following is true.

\[
E \left[ L(\theta^{priv}; D) \right] - \min_{\theta \in C} L(\theta; D) = O \left( \frac{(LG_c)^2 \log(1/\delta)}{\Delta(n \epsilon)^2} \right).
\]

**Proof:** For the ease of notation, we will drop the dependence on the data set \( D \), and represent the loss function \( L(\theta; D) \) as \( L(\theta) \). Let \( J(\theta) = L(\theta) + \frac{\Delta}{2} \| \theta \|_2^2 \) and let \( J^{priv}(\theta) = J(\theta) + \langle b, \theta \rangle \). Also let \( \hat{\theta} = \arg \min_{\theta \in C} J(\theta) \). We denote the variance of the noise in Algorithm 2 by \( \sigma^2 = \frac{L^2(2 \log(1/\delta))}{(n \epsilon)^2} \).

**Case 1 (Loss function \( L \) is Lipschitz):** By the optimality of \( \theta^{priv} \), the following is true.

\[
J^{priv}(\hat{\theta}) \geq J^{priv}(\theta^{priv}) \\
\iff J(\hat{\theta}) + \langle b, \hat{\theta} \rangle \geq J(\theta^{priv}) + \langle b, \theta^{priv} \rangle \\
\iff J(\theta^{priv}) - J(\hat{\theta}) \leq \langle b, \hat{\theta} - \theta^{priv} \rangle \\
\Rightarrow E \left[ J(\theta^{priv}) - J(\hat{\theta}) \right] = O \left( \frac{LG_c \sqrt{\log(1/\delta)}}{\epsilon n} \right).
\]

The last equality follows from the definition of Gaussian width and the variance of the noise vector \( b \). Let \( \theta^* = \arg \min_{\theta \in C} J(\theta) \). From (14) the definition of \( J(\theta) \), and that \( \theta \) minimizing \( J(\theta) \), the following is true.

\[
E \left[ L(\theta^{priv}) - L(\theta^*) \right] = E \left[ J(\theta^{priv}) - J(\theta^*) \right] + \frac{\zeta}{2} \| \theta^* - \theta_0 \|_2^2 - \frac{\zeta}{2} \| \theta^{priv} - \theta_0 \|_2^2 \\
\leq E \left[ J(\theta^{priv}) - J(\theta) \right] + \frac{\zeta}{2} \| \theta^* - \theta_0 \|_2^2 \\
= O \left( \frac{LG_c \sqrt{\log(1/\delta)} + \lambda_{\max} \| C \|_2^2}{\epsilon n} \right). \tag{15}
\]

**Case 2 (Loss function \( L \) is Lipschitz and strongly convex):** First notice that by the definition of Minkowski norm, for any vector \( v \in C \), \( \| v \|_Q \geq \| v \|_2/\| C \|_2 \). This implies that if \( L \) is \( \Delta \)-strongly convex w.r.t. \( \| \cdot \|_Q \)-norm, then it is \( \Delta/\| C \|_2^2 \) strongly convex w.r.t. \( \| \cdot \|_2 \)-norm. Hence with the lower bound on \( \Delta \)-satisfied, \( \zeta \) in Algorithm 2 is always zero.

By the definition of strong convexity of \( L \), the following is true.

\[
L(\theta^*) \geq L(\theta^{priv}) + \frac{\Delta}{2} \| \theta^{priv} - \theta^* \|_Q^2 \\
\iff L(\theta^*) + \langle b, \theta^* \rangle - \langle b, \theta^{priv} \rangle \geq L(\theta^{priv}) + \langle b, \theta^{priv} \rangle - \langle b, \theta^{priv} \rangle + \frac{\Delta}{2} \| \theta^{priv} - \theta^* \|_Q^2 \\
\Rightarrow \langle b, \theta^{priv} - \theta^* \rangle \geq \frac{\Delta}{2} \| \theta^{priv} - \theta^* \|_Q^2 \\
\Rightarrow \langle b, \| \theta^{priv} - \theta^* \|_Q \rangle \geq \frac{\Delta}{2} \| \theta^{priv} - \theta^* \|_Q \\
\Rightarrow \max_{v \in Q} \langle b, v \rangle \geq \frac{\Delta}{2} \| \theta^{priv} - \theta^* \|_Q \\
\Rightarrow \| \theta^{priv} - \theta^* \|_Q \leq \frac{2 \max_{v \in Q} \langle b, v \rangle}{\Delta} = \frac{2 \| b \|_Q^*}{\Delta}. \tag{16}
\]
In the above we have used the fact $\mathcal{L}(\theta^*) + \langle b, \theta^* \rangle \leq \mathcal{L}(\theta^{priv}) + \langle b, \theta^{priv} \rangle$ (due to the optimality condition). Using (16) we get the following.

$$\mathcal{L}(\theta^*) + \langle b, \theta^* \rangle \geq \mathcal{L}(\theta^{priv}) + \langle b, \theta^{priv} \rangle \Rightarrow \mathcal{L}(\theta^{priv}) - \mathcal{L}(\theta^*) \leq \|b\|_Q \cdot \|\theta^{priv} - \theta^*\|_Q$$

$$\Rightarrow \mathcal{L}(\theta^{priv}) - \mathcal{L}(\theta^*) \leq \frac{2\|b\|_Q}{\Delta}$$

$$\Rightarrow \mathbb{E} \left[ \mathcal{L}(\theta^{priv}) \right] - \mathcal{L}(\theta^*) = O \left( \frac{\sigma^2 G_Q^2}{\Delta} \right) \Rightarrow O \left( \frac{(L_G)^2 \log(1/\delta)}{\Delta(n\epsilon)^2} \right).$$

This completes the proof. In the last step we used the fact that $G_Q = \Theta(G_C)$. \qed

5 Private Convex Optimization by Frank-Wolfe algorithm

The algorithms in the previous section work best when the objective function is Lipschitz with respect to $\ell_2$ norm. But in many machine learning tasks, especially those with sparsity constraint, the objective function is often Lipschitz with respect to $\ell_1$ norm. For example, in high-dimensional linear regression setting e.g. the classical LASSO algorithm [Tib96], we would like to compute $\arg\min_{\theta \in C} \frac{1}{2n} \|X\theta - y\|_2^2$. Suppose that $|x_{ij}|, |y_j| = O(1)$. Then $\mathcal{L}(\theta) = \frac{1}{2n} \|X\theta - y\|_2^2$ is $O(1)$-Lipschitz with respect to $\ell_1$-norm but is $O(p)$-Lipschitz with respect to $\ell_2$-norm. So applying the private mirror-descent would result in a fairly loose bound. In this section, we will show that in these cases it is more effective to use the private version of the classical Frank-Wolfe algorithm.

5.1 Frank-Wolfe algorithm

The Frank-Wolfe algorithm [FW56] can be regarded as a “greedy” algorithm which always moves towards the optimum solution in the first order approximation (see Algorithm 4 for an exact convergence rate). (Due to the optimality condition). How fast Frank-Wolfe algorithm converges depends how much curvature $\mathcal{L}$ has. Define $\theta^* = \arg\min_{\theta \in C} \mathcal{L}(\theta)$. The following definition is from [Cla10 Jag13].

**Definition 5.1 (Curvature constant).** For $\mathcal{L} : C \rightarrow \mathbb{R}$, define $\Gamma_C$ as below.

$$\Gamma_C := \sup_{\theta_1, \theta_2, \in C, \gamma \in [0,1], \theta_3 = \theta_1 + \gamma (\theta_2 - \theta_1)} \frac{2}{\gamma^2} \left( \mathcal{L}(\theta_3) - \mathcal{L}(\theta_1) - \langle \theta_3 - \theta_1, \nabla \mathcal{L}(\theta_1) \rangle \right).$$

**Remark 3.** One can show ([Cla10 Jag13]) that for any $q, r \geq 1$ such that $q^{-1} + r^{-1} = 1$, $\Gamma_C$ is upper bounded by $\lambda \|\mathcal{L}\|_q^2$, where $\lambda = \max_{\theta \in C, \|v\|_y = 1} \|\nabla^2 \mathcal{L}(\theta) \cdot v\|_r$.

**Remark 4.** One useful bound is for the quadratic programming $\mathcal{L}(\theta) = \theta^T X^T X \theta + b^T \theta$. In this case, by [Cla10], $\Gamma_C \leq \max_{a, b \in X \cdot C} \|a - b\|_2^2$. When $\mathcal{L}$ is centrally symmetric, we have the bound $\Gamma_C \leq 4 \max_{\theta \in C} \|X\theta\|_2^2$.

The following theorem shows the convergence of Frank-Wolfe algorithm

**Theorem 5.2 (Cla10 Jag13).** In Algorithm 3 if we set $\mu = 1/T$, then $\mathcal{L}(\theta^*) - \mathcal{L}(\theta_T) = O(\Gamma_C / T)$. 

13
### Algorithm 3 Frank-Wolfe algorithm

**Input:** \( C \subseteq \mathbb{R}^p, L : C \rightarrow \mathbb{R}, \mu \)

1: Choose an arbitrary \( \theta_1 \) from \( C \);
2: for \( t = 1 \) to \( T - 1 \) do
3: \( \tilde{\theta}_t = \arg\min_{\theta \in C} \langle \nabla L(\theta_t), (\theta - \theta_t) \rangle \);
4: \( \theta_{t+1} = \theta_t + \mu (\tilde{\theta}_t - \theta_t) \);
5: return \( \theta_T \).

While the Frank-Wolfe algorithm does not necessarily provide faster convergence compared to the gradient-descent based method, it has two major advantages. First, on Line 3, it reduces the problem to solving a minimization of linear function. When \( C \) is defined by small number of vertices, e.g. when \( C \) is an \( \ell_1 \) ball, the minimization can be done by checking \( \langle \nabla L(\theta_t), x \rangle \) for each vertex \( x \) of \( C \). This can be done efficiently. Secondly, each step in Frank-Wolfe takes a convex combination of \( \theta_t \) and \( \tilde{\theta}_t \), which is on the boundary of \( C \). Hence each \( \theta_t \) is inside \( C \) (sometimes called projection free), and the final outcome \( \theta_T \) is the convex combination of up to \( T \) points on the boundary of \( C \) (or vertices of \( C \) when \( C \) is a polytope). Such outcome might be desired, for example when \( C \) is a polytope, as it corresponds to a sparse solution. Due to these reasons Frank-Wolfe algorithm has found many applications in machine learning [SSSZ10, HK12, Cla10].

#### 5.2 Private Frank-Wolfe Algorithm

We now present a private version of the Frank-Wolfe algorithm. We can achieve privacy by replacing Line 3 with its private version in one of two ways. In the first variant, we apply exponential mechanism [MT07] to guarantee privacy; and in the second variant, we apply objective perturbation. The first variant works especially well when \( C \) is a polytope defined by polynomially many vertices. In this case, we show that the error depends on the \( \ell_1 \)-Lipschitz constant, which can be much smaller than the \( \ell_2 \)-Lipschitz constant. In particular, the private Frank-Wolfe algorithm is nearly optimal for the important high-dimensional sparse linear regression (or compressive sensing) problem.

In the following we provide the private Frank-Wolfe algorithm (Algorithm 4) for the polytope case, i.e., when \( C \) is a convex hull of a finite set \( S \) of vertices (or corners). In this case, we know that any linear function is minimized at one point of \( S \) per the following basic fact.

**Fact 5.3.** Let \( C \subseteq \mathbb{R}^p \) be the convex hull of a set \( S \subseteq \mathbb{R}^p \). For any vector \( v \in \mathbb{R}^p \), the minimizer of the optimization problem \( \theta = \arg\min_{\theta \in C} \langle \theta, v \rangle \) is a point in \( S \).

When \( S \) is a finite set, i.e., when \( C \) is a polytope, we can apply the exponential mechanism [MT07] to sample a vertex from \( S \) as \( \tilde{\theta}_t \). This results in \( \log |S| \) dependency of the error. When \( |S| \) is polynomial in \( p \), it leads to a so-called dimension-independent bound. While the exponential mechanism may be applied to the general \( C \) as well, it’s error would depend on the size of a cover of the set of points on the boundary of \( C \) that can be the true optimizers in Line 3. If we used an arbitrary net of \( C \) as a cover, the bound is likely to have polynomial dependence on \( p \). Hence for general convex set \( C \), in \( \mathcal{A}_{Noise-FW(Gen-convex)} \) (Algorithm 5 in Appendix B), we use objective perturbation instead and obtain an error dependent on the Gaussian width of \( C \).

**Theorem 5.4** (Privacy guarantee). Algorithm \( \mathcal{A}_{Noise-FW(polytope)} \) (Algorithm 4) is \((\epsilon, \delta)\)-differentially private.

The proof of privacy follows from a straightforward use of exponential mechanism [MT07, BLST10] (the noisy maximum version from [BLST10, Theorem 5]) and the strong composition theorem [DRV10].
Algorithm 4 $A_{\text{Noise-FW(polytope)}}$: Differentially Private Frank-Wolfe Algorithm (Polytope Case)

Input: Data set: $D = \{d_1, \cdots, d_n\}$, loss function: $\mathcal{L}(\theta; D) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\theta; d_i)$ (with $\ell_1$-Lipschitz constant $L_1$ for $\ell$), privacy parameters: $(\epsilon, \delta)$, convex set: $C = \text{conv}(S)$ with $\|C\|_1$ denoting $\max_{s \in S} \|s\|_1$ and $S$ being the set of corners.

1: Choose an arbitrary $\theta_1$ from $C$;
2: for $t = 1$ to $T - 1$ do
3: \hspace{1em} $\forall s \in S$, $\alpha_s \leftarrow \langle s, \nabla \mathcal{L}(\theta_t; D) \rangle + \text{Lap} \left( \frac{L_1 \|C\|_1 \sqrt{8T \log(1/\delta)}}{n \epsilon} \right)$, where $\text{Lap}(\lambda) \sim \frac{1}{2\lambda} e^{-|x|}/\lambda$.
4: \hspace{1em} $\tilde{\theta}_t \leftarrow \arg \min_{s \in S} \alpha_s$.
5: \hspace{1em} $\theta_{t+1} \leftarrow (1 - \mu) \theta_t + \mu \tilde{\theta}_t$, where $\mu = \frac{1}{T+2}$.
6: Output $\theta_{\text{priv}} = \theta_T$.

Theorem 5.5 we prove the utility guarantee for the private Frank-Wolfe algorithm for the convex polytope case. Define $\Gamma_L = \max_{D \in \mathcal{D}} C_L$ over all the possible data sets in $D$.

**Theorem 5.5 (Utility guarantee).** Let $L_1, S$ and $\|C\|_1$ be defined as in Algorithms \[Algorithm A_{\text{Noise-FW(polytope)}}, \] and $\mathcal{L}(\cdot; d)$ be an upper bound on the curvature constant (defined in Definition 5.1) for the loss function $\mathcal{L}((\cdot); d)$ that holds for all $d \in D$. In Algorithm $A_{\text{Noise-FW(polytope)}}$, if we set $T = \frac{\Gamma_L^{3/4} (n\epsilon)^{3/2}}{(L_1 \|C\|_1)^{2/3}}$, then the privacy risk is as follows.

$$\mathbb{E} \left[ \mathcal{L}(\theta_{\text{priv}}; D) \right] - \min_{\theta \in C} \mathcal{L}(\theta; D) = O \left( \frac{\Gamma_L^{1/3} (L_1 \|C\|_1)^{2/3} \log(n|S|) \sqrt{\log(1/\delta)}}{(n\epsilon)^{3/2}} \right).$$

Here the expectation is over the randomness of the algorithm.

**Proof.** For ease of notation we hide the dependence of $\mathcal{L}$ on the data set $D$ and represent it simply as $\mathcal{L}(\theta)$. In order to prove the utility guarantee we first invoke the utility guarantee of the non-private noisy Frank-Wolfe algorithm from [Jag13, Theorem 1].

**Theorem 5.6 (Non-private utility guarantee [Jag13]).** Assume the conditions in Theorem 5.5 and let $\beta > 0$ be fixed. Recall that $\mu = 1/(T + 2)$ and let $\phi_1 \in C$. Suppose that $\langle s_1, \cdots, s_T \rangle$ is a sequence of vectors from $C$, with $\phi_{t+1} = (1 - \mu) \phi_t + \mu s_t$ such that for all $t \in [T]$,

$$\langle s_t, \nabla \mathcal{L}(\phi_t) \rangle \leq \min_{s \in C} \langle s, \nabla \mathcal{L}(\phi_t) \rangle + \frac{1}{2} \beta \mu \Gamma_L.$$

Then,

$$\mathcal{L}(\phi_T) - \min_{\theta \in C} \mathcal{L}(\theta) \leq \frac{2\Gamma_L}{T + 2} (1 + \beta).$$

Since the convex set $C$ is a polytope with corners in $S$, if $s_t$ in Theorem 5.6 corresponds to $\tilde{\theta}_t$ in Algorithm $A_{\text{Noise-FW(polytope)}}$, and $\phi_t$ corresponds to $\theta_t$ in $A_{\text{Noise-FW(polytope)}}$, then using the tail properties of Laplace distribution and Fact 5.3 one can show that with probability at least $1 - \zeta$, the term $\beta$ in Theorem 5.6 is at most $O \left( \frac{L_1 \|C\|_1 \sqrt{8T \log(1/\delta) \log(|S|/\zeta)}}{\mu n \epsilon} \right)$. Plugging in this bound in Theorem 5.6, we immediately get that with probability at least $1 - \zeta$,

$$\mathcal{L}(\theta_T) - \min_{\theta \in C} \mathcal{L}(\theta) = O \left( \frac{\Gamma_L}{T} + \frac{L_1 \|C\|_1 \sqrt{8T \log(1/\delta) \log(|S|/\zeta)}}{n \epsilon} \right).$$  \hspace{1em} (17)
From, \( \text{(17)} \) we can conclude the following in expectation.

\[
E \left[ \mathcal{L}(\theta_T) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta) \right] = O \left( \frac{\Gamma_{\mathcal{L}}}{T} + \frac{L_1\|C\|_1\sqrt{8T \log(1/\delta)}}{n\epsilon} \log(TL_1\|C\|_1 \cdot |S|) \right).
\]  

Setting \( T = \frac{\Gamma_{\mathcal{L}}^{2/3}(\mu\epsilon)^{2/3}}{(L_1\|C\|_1)^{2/3}} \) results in the claimed utility guarantee.

\[ \square \]

### 5.3 Optimality of private Frank-Wolfe algorithm for sparse linear regression

We now apply the private Frank-Wolfe algorithm \( \mathcal{A}_{\text{Noise--FW(polytope)}} \) to the important case of sparse linear regression (also called the compressive sensing) problem. We show that the private Frank-Wolfe algorithm leads to a nearly tight bound.

**Problem definition:** Given a data set \( \hat{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \) of \( n \)-samples from the domain \( D = \{(x, y) : x \in \mathbb{R}^p, y \in [-1, 1], |x|_\infty \leq 1\} \), and the convex set \( \mathcal{C} \) is set to \( \ell_1 \)-ball in \( p \)-dimensions. For any \( \theta \in \mathcal{C} \), define the loss function as,

\[
\mathcal{L}(\theta; D) = \frac{1}{2n} \sum_{i \in [n]} (\langle x_i, \theta \rangle - y_i)^2.
\]

The objective is to compute \( \theta^{\text{priv}} \in \mathcal{C} \) to minimize \( \mathcal{L}(\theta; D) \) while preserving privacy with respect to any change of individual \( (x_i, y_i) \) pair. The non-private setting of the above problem is a variant of the least squares problem with \( \ell_1 \) regularization, which was started by the work of LASSO [Tib96, T+97] and has attracted intensive study in the past years [HTF01, DJ04, CT05, Don06, CT07, BRT09, BM12, RWY09, Zha13]. One important reason for using \( \ell_1 \) regularization is to induce sparse solutions, i.e., \( \theta \) with small number of non-zero coordinates. This is especially interesting for the so called “high-dimensional” setting where \( p \gg n \). Indeed, via a long line of work [DJ04, CT05, Don06, Wai06, CT07, BRT09], it has been shown that under suitable condition of \( X \), using \( \ell_1 \) regularization can indeed produce a nearly optimal sparse solution, providing theoretical support to the empirical success of LASSO.

Since the \( \ell_1 \) ball is the convex hull of \( 2p \) vertices, we can apply the private Frank-Wolfe algorithm \( \mathcal{A}_{\text{Noise--FW(polytope)}} \). For the above setting, it is easy to check that the \( \ell_1 \)-Lipschitz constant is bounded by \( O(1) \). Further, by applying the bound on quadratic programming Remark 4, we have that \( C_{\mathcal{L}} \leq 4 \max_{\theta \in \mathcal{C}} \|X\theta\|_2^2 = O(1) \) since \( C \) is the unit \( \ell_1 \) ball, and \( |x_i| \leq 1 \). Hence \( \Gamma = O(1) \). Now applying Theorem 5.5 we have that

**Corollary 5.7.** Let \( D = \{(x_1, y_1), \ldots, (x_n, y_n)\} \) of \( n \) samples from the domain \( D = \{(x, y) : \|x\|_\infty \leq 1, |y| \leq 1\} \), and the convex set \( \mathcal{C} \) equal to the \( \ell_1 \)-ball. The output \( \theta^{\text{priv}} \) of Algorithm \( \mathcal{A}_{\text{Noise--FW(polytope)}} \) ensures the following.

\[
E[\mathcal{L}(\theta^{\text{priv}}, D) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D)] = O \left( \frac{\log(np/\delta)}{(n\epsilon)^{2/3}} \right).
\]

**Remark 5.** Compared to the previous work [KST12, ST13], the above upper bound makes no assumption of restricted strong convexity or mutual incoherence, which might be too strong for realistic settings [Was12]. But we note that here we measure the error in terms of predication risk but not on the parameters as in [KST12, ST13].

In the following, we shall show that to ensure privacy, the error bound in Corollary 5.7 is nearly optimal in terms of the dominant factor of \( 1/n^{2/3} \).
Theorem 5.8 (Optimality of private Frank-Wolfe). Let $C$ be the $\ell_1$-ball. For every sufficiently large $n$, for every $(\epsilon, \delta)$-differentially private algorithm $A$, with $\epsilon \leq 0.1$ and $\delta = o(1/n^2)$, there exists a data set $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of $n$ samples from the domain $D = \{(x, y) : \|x\|_\infty \leq 1, |y| \leq 1\}$ such that

$$
\mathbb{E}[\mathcal{L}(A(D); D) - \min_{\theta \in C} \mathcal{L}(\theta; D)] = \Omega \left( \frac{1}{n^{2/3}} \right).
$$

We prove the lower bound by following the fingerprinting codes argument of [BUV14] for lowerbounding the error of $(\epsilon, \delta)$-differentially private algorithms. Similar to [BUV14] and [DTTZ14], we start with the following lemma which is implicit in [BUV14]. The matrix $X$ in Theorem 5.9 is the padded Tardos code used in [DTTZ14] Section 5). For any matrix $X$, denote by $X_{(i)}$ the matrix obtained by removing the $i$-th row of $X$. Call a column of a matrix a concensus column if the entries in the column are either all 1 or all −1. The sign of a concensus column is simply the concensus value of the column. Write $w = n/ \log m$ and $p = 1000m^2$.

Theorem 5.9. [Corollary 16 from [DTTZ14], restated] Let $m$ be a sufficiently large positive integer. There exists a matrix $X \in \{-1, 1\}^{(w+1) \times p}$ with the following guarantee. For each $i \in [1, w + 1)$, there are at least 0.999$p$ concensus columns $W_i$ in each $X_{(i)}$. In addition, for algorithm $A$ on input matrix $X_{(i)}$ where $i \in [1, w + 1)$, if with probability at least $2/3$, $A(X_{(i)})$ produces a $p$-dimensional sign vector which agrees with at least $\frac{3}{4}p$ columns in $W_i$, then $A$ is not $(\epsilon, \delta)$ differentially private with respect to single row change (to some other row in $X$).

Write $\tau = 0.001$. Let $k = \tau wp$. We first form an $k \times p$ matrix $Y$ where the column vectors of $Y$ are mutually orthogonal $\{1, -1\}$ vectors. This is possible as $k \gg p$. Now we construct $w + 1$ databases $D_i$ for $1 \leq i \leq w + 1$ as follows. For all the databases, they contain the common set of pair $(z_j, 0)$ for $1 \leq j \leq k$ where $z_j = (Y_{j1}, \ldots, Y_{jp})$ is the $j$-th row vector of $Y$. In addition, each $D_i$ contains $w$ pairs $(x_j, 1)$ for $x_j = (X_{j1}, \ldots, X_{jk})$ for $j \neq i$. Then $\mathcal{L}(\theta; D_i)$ is defined as follows (for the ease of notation in this proof, we work with the un-normalized loss. This does not affect the generality of the arguments in any way.)

$$
\mathcal{L}(\theta; D_i) = \sum_{j \neq i} (x_j \cdot \theta - 1)^2 + \sum_{j=1}^{k} (y_j \cdot \theta)^2 = \sum_{j \neq i} (x_j \cdot \theta - 1)^2 + k\|\theta\|_2^2.
$$

The last equality is due to the columns of $Y$ are mutually orthogonal $\{-1, 1\}$ vectors. For each $D_i$, consider $\theta^* \in \left\{-\frac{1}{p}, \frac{1}{p}\right\}^p$ such that the sign of the coordinates of $\theta^*$ matches the sign for the consensus columns of $X_{(i)}$. Plugging $\theta^*$ in $\mathcal{L}(\theta^*; \hat{D})$ we have the following,

$$
\mathcal{L}(\theta^*; \hat{D}) \leq \sum_{i=1}^{w} (2\tau)^2 + k/p \quad \text{since the number of consensus columns is at least } (1 - \tau)p
$$

$$
= (\tau + 4\tau^2)w.
$$

We now prove the crucial lemma, which states that if $\theta$ is such that $\|\theta\|_1 \leq 1$ and $\mathcal{L}(\theta; D_i)$ is small, then $\theta$ has to agree with the sign of consensus columns of $X_{(i)}$.

Lemma 5.10. Suppose that $\|\theta\|_1 \leq 1$, and $\mathcal{L}(\theta; D_i) < 1.1\tau w$. For $j \in W_i$, denote by $s_j$ the sign of the consensus column $j$. Then we have

$$
|\{j \in W_i : \text{sign}(\theta_j) = s_j\}| \geq \frac{3}{4}p.
$$
Proof. For any $S \subseteq \{1, \ldots, p\}$, denote by $\theta|_S$ the projection of $\theta$ to the coordinate subset $S$. Consider three subsets $S_1, S_2, S_3$, where

$$S_1 = \{j \in W_i : \text{sign}(\theta_j) = s_j\},$$
$$S_2 = \{j \in W_i : \text{sign}(\theta_j) \neq s_j\},$$
$$S_3 = \{1, \ldots, p\} \setminus W_i.$$

The proof is by contradiction. Assume that $|S_1| < \frac{3}{4}p$.

Further denote $\theta_i = \theta|_{S_i}$ for $i = 1, 2, 3$. Now we will bound $\|\theta_1\|_1$ and $\|\theta_3\|_1$ using the inequality $\|x\|_2 \geq \|x\|_1/\sqrt{d}$ for any $d$-dimensional vector.

$$\|\theta_3\|_2^2 \geq \|\theta_3\|_1^2/|S_3| \geq \|\theta_3\|_2^2/(\tau p).$$

Hence $k\|\theta_3\|_2^2 \geq w\|\theta_3\|_1^2$. But $k\|\theta_3\|_2^2 \leq k\|\theta\|_2^2 \leq 1.1\tau w$, so that $\|\theta_3\|_1 \leq \sqrt{1.1\tau} \leq 0.94$.

Similarly by the assumption of $|S_1| < \frac{3}{4}p$,

$$\|\theta_1\|_2^2 \geq \|\theta_1\|_1^2/|S_1| \geq 4\|\theta_1\|_1^2/(3p).$$

Again using $k\|\theta\|_2^2 < 1.1\tau w$, we have that $\|\theta_1\|_1 \leq \sqrt{1.1 \times 3/4} \leq 0.91$.

Now we have $|\langle x_i, \theta \rangle - 1| = \|\theta_1\|_1 - 2\|\theta_2\|_1 + \beta_i - 1$ where $|\beta_i| \leq \|\theta_3\|_1 \leq 0.04$. By $\|\theta_1\|_1 + \|\theta_2\|_1 + \|\theta_3\|_1 \leq 1$, we have

$$|\langle x_i, \theta \rangle - 1| \geq 1 - \|\theta_1\|_1 - |\beta_i| \geq 1 - 0.94 - 0.04 = 0.05.$$

Hence we have that $\mathcal{L}(\theta; D_i) \geq (0.05)^2 w \geq 1.1\tau w$. This leads to a contradiction. Hence we must have $|S_1| \geq \frac{3}{4}p$.

With Theorem 5.9 and Lemma 5.10, we can now prove Theorem 5.8.

Proof. Suppose that $\mathcal{A}$ is private. And for the datasets we constructed above,

$$\mathbb{E}[\mathcal{L}(\mathcal{A}(D_i); D_i) - \min_{\theta} \mathcal{L}(\theta; D_i)] \leq cw,$$

for sufficiently small constant $c$. By Markov inequality, we have with probability at least $2/3$, $\mathcal{L}(\mathcal{A}(D_i); D_i) - \min_{\theta} \mathcal{L}(\theta; D_i) \leq 3cw$. By (19), we have $\min_{\theta} \mathcal{L}(\theta; D_i) \leq (\tau + 4\tau^2)w$. Hence if we choose $c$ a constant small enough, we have with probability $2/3$,

$$\mathcal{L}(\mathcal{A}(D_i); D_i) < (\tau + 4\tau^2 + 3c)w \leq 1.1\tau w.$$  

(20)

By Lemma 5.10, (20) implies that $\mathcal{A}(D_i)$ agrees with at least $\frac{3}{4}p$ concensus columns in $X(i)$. However by Theorem 5.9 this violates the privacy of $\mathcal{A}$. Hence we have that there exists $i$, such that

$$\mathbb{E}[\mathcal{L}(\mathcal{A}(D_i); D_i) - \min_{\theta} \mathcal{L}(\theta; D_i)] > cw.$$ 

Recall that $w = m/\log m$ and $n = w + wp = O(m^3/\log m)$. Hence we have that

$$\mathbb{E}[\mathcal{L}(\mathcal{A}(D_i); D_i) - \min_{\theta} \mathcal{L}(\theta; D_i)] = \Omega(n^{1/3}/\log^{2/3} n).$$

The proof is completed by converting the above bound to the normalized version of $\Omega(1/(n \log n)^{2/3})$. 

\[ \square \]
References

[BLST10] Raghav Bhaskar, Srivatsan Laxman, Adam Smith, and Abhradeep Thakurta. Discovering frequent patterns in sensitive data. In KDD, New York, NY, USA, 2010.

[BM03] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. The Journal of Machine Learning Research, 3:463–482, 2003.

[BM12] Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. IEEE Transactions on Information Theory, 2012.

[BRT09] Peter J Bickel, Ya’acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. The Annals of Statistics, pages 1705–1732, 2009.

[BST14] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization, revisited. In FOCS, 2014.

[BT03] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. Operations Research Letters, 31(3):167–175, 2003.

[BTN13] Aharon Ben-Tal and Arkadi Nemirovski. Lectures on modern convex optimization. Lecture notes, 2013.

[BUV14] Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In STOC, 2014.

[Cla10] Kenneth L. Clarkson. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. ACM Transactions on Algorithms, 2010.

[CM08] Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. In NIPS, 2008.

[CMS11] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. JMLR, 12:1069–1109, 2011.

[CRPW12] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. Foundations of Computational Mathematics, 12(6):805–849, 2012.

[CT05] Emmanuel Candes and Terrance Tao. Decoding by linear programming. IEEE Transactions on Information Theory, 51, 2005.

[CT07] Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics, pages 2313–2351, 2007.

[DJ04] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. Annals of Statistics, pages 962–994, 2004.

[DJW13] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In FOCS, 2013.

[DKM+06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In EUROCRYPT, pages 486–503, 2006.
[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

[DNPR10] Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N. Rothblum. Differential privacy under continual observation. In *STOC*, 2010.

[DNT13] Cynthia Dwork, Aleksandar Nikolov, and Kunal Talwar. Efficient algorithms for privately releasing marginals via convex relaxations. *arXiv preprint arXiv:1308.1385*, 2013.

[Don06] David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 2006.

[DRV10] Cynthia Dwork, Guy N. Rothblum, and Salil P. Vadhan. Boosting and differential privacy. In *FOCS*, 2010.

[DSSST10] John C Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In *COLT*, 2010.

[DTTZ14] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *STOC*, 2014.

[Dwo06] Cynthia Dwork. Differential privacy. In *ICALP*, LNCS, pages 1–12, 2006.

[Dwo08] Cynthia Dwork. Differential privacy: A survey of results. In *TAMC*, pages 1–19. Springer, 2008.

[Dwo09] Cynthia Dwork. The differential privacy frontier. In *TCC*, pages 496–502. Springer, 2009.

[FW56] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[HK12] Elad Hazan and Satyen Kale. Projection-free online learning. In *ICML*, 2012.

[HTF01] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., 2001.

[Jag13] Martin Jaggi. Revisiting {Frank-Wolfe}: Projection-free sparse convex optimization. In *ICML*, 2013.

[JKT12] Prateek Jain, Pravesh Kothari, and Abhradeep Thakurta. Differentially private online learning. In *COLT*, pages 24.1–24.34, 2012.

[JT14] Prateek Jain and Abhradeep Thakurta. (near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning (ICML)*, 2014.

[KST12] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *COLT*, pages 25.1–25.40, 2012.

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *FOCS*, pages 94–103. IEEE, 2007.

[NTZ13] Aleksandar Nikolov, Kunal Talwar, and Li Zhang. The geometry of differential privacy: The sparse and approximate cases. In *STOC*, 2013.
[NY83] A. S. Nemirovski and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons, 1983.

[Rak09] Alexander Rakhlin. Lecture notes on online learning. *Lecture notes*, 2009.

[RWY09] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over $\ell_q$-balls. *ArXiv e-prints*, October 2009.

[SSS07] Shai Shalev-Shwartz and Yoram Singer. Logarithmic regret algorithms for strongly convex repeated games. 2007.

[SSSS09] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic Convex Optimization. In *COLT*, 2009.

[SSSZ10] Shai Shalev-Shwartz, Nathan Srebro, and Tong Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 2010.

[SST11] Nati Srebro, Karthik Sridharan, and Ambuj Tewari. On the universality of online mirror descent. In *NIPS*, 2011.

[ST10] Karthik Sridharan and Ambuj Tewari. Convex games in banach spaces. In *COLT*, 2010.

[ST13a] Adam Smith and Abhradeep Thakurta. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *COLT*, 2013.

[ST13b] Adam Smith and Abhradeep Thakurta. Follow the perturbed leader is differentially private with optimal regret guarantees. *Manuscript in preparation*, 2013.

[ST13c] Adam Smith and Abhradeep Thakurta. Nearly optimal algorithms for private online learning in full-information and bandit settings. In *NIPS*, 2013.

[SZ13] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *ICML*, pages 71–79, 2013.

[T+97] Robert Tibshirani et al. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996.

[Ull14] Jonathan Ullman. Private multiplicative weights beyond linear queries. *CoRR*, abs/1407.1571, 2014.

[Wai06] Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity using $\ell_1$-constrained quadratic programs. In *IEEE Transactions on Information Theory*, 2006.

[Was12] Larry Wasserman. Restricted isometry property, rest in peace. *Blog-post*, 2012.

[Zha13] Li Zhang. , nearly optimal minimax estimator for high dimensional sparse linear regression. *Annals of Statistics*, 2013.
A Tighter Guarantees of Mirror Descent for Strongly Convex Functions

In this section we study Algorithm 1 (Algorithm $A_{\text{Noise-MD}}$) in the context of strongly convex functions with the following form: Every loss function $\ell(\theta; d)$ is $L$-Lipschitz in the $L_2$-norm and $\Delta$-strongly convex with respect to some differentiable convex function $\Psi : \mathcal{C} \rightarrow \mathbb{R}$, for any $\theta \in \mathcal{C}$ and $d \in \mathcal{D}$. (See Section 2.4 for a definition.) This setting has previously been studied in [DSSST10, SSS07]. Two common examples are: i) $\Psi(\theta) = \frac{1}{2} \|\theta\|_2^2$ and $\ell(\theta; d)$ is $\Delta$-strongly convex w.r.t. $\|\cdot\|_2$, and ii) for composite loss functions $\ell(\theta; d) = g(\theta; d) + \Delta \Psi(\theta)$ if $\Psi(\theta) = \sum_{i=1}^p \theta_i \log(\theta_i)$, then $\ell(\theta; d)$ is $\Delta$-strongly convex w.r.t. $\Psi(\theta)$ within the probability simplex which is in turn $1$-strongly convex w.r.t. $\|\cdot\|_1$ [DSSST10 Section 5]. In the following we show that one can get a much sharper dependence on $n$ (compared to Theorem 3.2) under strong convexity.

Remark 6. [BST14] analyzed the setting of strong convexity w.r.t. $\ell_2$-norm, and in particular provided tight error guarantees. For this case, Theorem A.1 leads to similarly tight bounds, and thus the lower bounds in [BST14] imply that in general, our guarantee cannot be improved.

**Theorem A.1** (Utility guarantee for strongly convex functions). Let $Q$ be the symmetric convex hull of $\mathcal{C}$. Assume that every loss function $\ell(\theta; d)$ is $L$-Lipschitz in the $\ell_2$-norm and $\Delta$-strongly convex with respect to some differentiable $1$-strongly convex function $\Psi : \mathcal{C} \rightarrow \mathbb{R}$, for any $\theta \in \mathcal{C}$ and $d \in \mathcal{D}$. Let $\|\cdot\|$ be the $\ell_2$-diameter of the set $\mathcal{C}$, and $G_C$ be the Gaussian width. In Algorithm $A_{\text{Noise-MD}}$ (Algorithm 7), if we set $T = \frac{(\|\|\|_2\epsilon n)^2}{G_C^2}$, the potential function to be $\Psi$ and $\eta_t = \frac{2}{\Delta t}$, then following is true.

$$\mathbb{E} [\mathcal{L}(\theta^{\text{priv}}; D)] - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D) = O \left( \frac{(LG_C \log(n/\delta))^2 \log(\|\|\|_2(n\epsilon)/G_C)}{\Delta(n\epsilon)^2} \right).$$

Here the expectation is over the randomness of the algorithm.

**Proof.** For ease of notation, we hide the dependence of $\mathcal{L}(\theta; D)$ on the data set $D$, and simply represent it as $\mathcal{L}(\theta)$. The first part of the proof is fairly standard and exactly same as that of Theorem 3.2 till (2). Following the same notation, it suffices to bound $\frac{1}{T} \sum_{t=1}^T \mathcal{L}(\theta_t) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta)$. Rest of the proof differs from Theorem 3.2 to the extent that we now work with a quadratic approximation (Claim A.2) to the loss function instead of a linear application (Claim 3.3).

**Claim A.2.** Let $\theta^* = \arg \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta)$. For every $t \in [T]$, let $\gamma_t$ be the sub-gradient of $\mathcal{L}(\theta_t; D)$ used in iteration $t$ of Algorithm $A_{\text{Noise-MD}}$ (Algorithm 7). Then, the following is true.

$$\frac{1}{T} \sum_{t=1}^T \mathcal{L}(\theta_t) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta) \leq \frac{1}{T} \sum_{t=1}^T [\langle \gamma_t, \theta_t - \theta^* \rangle - \Delta \cdot B_\Psi(\theta^*, \theta_t)].$$

The proof of this claim is a direct consequence of the definition of strong convexity. Now using (8) from the proof of Theorem 3.2 and summing over the $T$ iterations, we have the following.

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E} [\langle \gamma_t, \theta_t - \theta^* \rangle - \Delta \cdot B_\Psi(\theta^*, \theta_t)] \leq \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[ \frac{B_\Psi(\theta^*, \theta_t) - B_\Psi(\theta^*, \theta_{t+1})}{\eta_{t+1}} - \Delta \cdot B_\Psi(\theta^*, \theta_t) \right]$$

$$+ \left( \frac{2L\|\|_2^2 + \sigma G_C^2}{2T} \right) \sum_{t=1}^T \eta_{t+1}$$

$$= \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[ B_\Psi(\theta^*, \theta_t) \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - \Delta \right) \right] + \left( \frac{2L\|\|_2^2 + \sigma G_C^2}{2T} \right) \sum_{t=1}^T \eta_t$$

(21)
Now setting $\eta_t = \frac{1}{\Delta t}$ and using Claim A.2, we obtain the following.

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} [\mathcal{L}(\theta_t)] - \mathcal{L}(\theta^*) \leq \frac{2(2L\|C\|_2 + \sigma G_C)^2}{\Delta T} \log T$$

$$= O \left( \frac{\log T}{\Delta} \left( \frac{L\|C\|_2}{\sqrt{T}} + \frac{LG_C \log(n/\delta)}{n\epsilon} \right)^2 \right)$$

(22)

Setting $T = \|C\|_2^2(n\epsilon)^2/G_C^2$ in (22), we obtain the required excess risk bound as follows:

$$\mathbb{E} [\mathcal{L}(\theta_{\text{priv}})] - \min_{\theta \in C} \mathcal{L}(\theta) = O \left( \frac{(LG_C \log(n/\delta))^2 \log(\|C\|_2^2(n\epsilon)/G_C^2)}{\Delta(n\epsilon)^2} \right).$$

$\square$

## B Missing Details for Private Frank-Wolfe for the $\ell_2$-bounded Case

In this section we provide the details of the private Frank-Wolfe algorithm for the $\ell_2$-bounded case, along with the privacy and utility guarantees.

Here for a data set $D = \{d_1, \cdots, d_n\}$, define objective function as the empirical loss function $\mathcal{L}(\theta; D) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; d_i)$. We define $L_2$ the $\ell_2$-Lipschitz constant, respectively, of $\mathcal{L}$ over all the possible data sets.

**Algorithm 5 $A_{\text{Noise-FW(General Convex)}}$:** Differentially Private Frank-Wolfe Algorithm (General Convex Case)

**Input:** Data set: $D = \{d_1, \cdots, d_n\}$, loss function: $\mathcal{L}(\theta; D) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; d_i)$ (with $\ell_2$-Lipschitz constant $L_2$ for $\ell$), privacy parameters: $(\epsilon, \delta)$, convex set: $C$ bounded in the $\ell_2$-norm, denoted by $\|C\|_2$.

1. choose an arbitrary $\theta_1$ from $C$;
2. for $t = 1$ to $T - 1$ do
3. $\theta_t = \arg \min_{\theta \in C} (\nabla \mathcal{L}(\theta_t; D) + b_t, \theta)$, where $b_t \sim \mathcal{N}(0, \|\mu\sigma^2\|)$ and $\sigma^2 \leftarrow \frac{32L_2 T \log^2(n/\delta)}{(n\epsilon)^2}$.
4. $\theta_{t+1} \leftarrow (1 - \mu)\theta_t + \mu\bar{\theta}_t$, where $\mu = \frac{1}{T+2}$.
5. Output $\theta_{\text{priv}} = \theta_T$.

**Theorem B.1** (Privacy guarantee). Algorithm $A_{\text{Noise-FW(General Convex)}}$ (Algorithm 5) is $(\epsilon, \delta)$-differentially private.

The proof of privacy is exactly same as the proof of privacy in Theorem 3.2. In the following we provide the utility guarantee for Algorithm $A_{\text{Noise-FW(General Convex)}}$.

**Theorem B.2** (Utility guarantee). Let $L_2$ and $\|C\|_2$ be defined as in Algorithm 5 (Algorithm $A_{\text{Noise-FW(General Convex)}}$). Let $G_C$ the Gaussian width of the convex set $C \subseteq \mathbb{R}^p$, and let $\Gamma_C$ be the curvature constant (defined in Definition 5.1) for the loss function $\ell(\theta; d)$ for all $\theta \in C$ and $d \in D$. In Algorithm $A_{\text{Noise-FW}}$ if we set $T = \frac{\Gamma_C^{-1/3} (L_2 G_C)^{2/3}}{(n\epsilon)^{2/3}}$, then the excess empirical risk is as follows.

$$\mathbb{E} [\mathcal{L}(\theta_{\text{priv}}; D)] - \min_{\theta \in C} \mathcal{L}(\theta; D) = O \left( \frac{\Gamma_C^{-1/3} (L_2 G_C)^{2/3} \log^2(n/\delta)}{(n\epsilon)^{2/3}} \right).$$

Here the expectation is over the randomness of the algorithm and $\Gamma_C$ is the curvature constant.
Proof. Recall $\sigma^2 = \frac{32L_2T\log^2(T/\delta)}{(n\epsilon)^2}$. Using the property of Gaussian width (Section 2.2), and a similar analysis as that of the convex polytope case, we can conclude the following.

$$
\mathbb{E}\left[\mathcal{L}(\theta_T) - \min_{\theta \in C} \mathcal{L}(\theta)\right] = O\left(\frac{\Gamma_c}{T} + \frac{L_2G_C\sqrt{T}\log^2(T/\delta)}{n\epsilon}\right).$$

Setting $T = \frac{\Gamma_c^2/(n\epsilon)^{2/3}}{(L_2G_C)^{2/3}}$, results in the utility guarantee. \qed