Differentially Private Generalized Linear Models Revisited

Raman Arora
Department of Computer Science
The Johns Hopkins University
arora@cs.jhu.edu

Raef Bassily
Department of Computer Science & Engineering
Translational Data Analytics Institute (TDAI)
The Ohio State University
bassily.1@osu.edu

Cristóbal Guzmán
Inst. for Mathematical and Comput. Eng.
Fac. de Matemáticas and Esc. de Ingeniería
Pontificia Universidad Católica de Chile
c.guzman@utwente.nl

Michael Menart
Department of Computer Science & Engineering
The Ohio State University
menart.2@osu.edu

Enayat Ullah
Department of Computer Science
The Johns Hopkins University
enayat@jhu.edu

Abstract

We study the problem of (ϵ, δ)-differentially private learning of linear predictors with convex losses. We provide results for two subclasses of loss functions. The first case is when the loss is smooth and non-negative but not necessarily Lipschitz (such as the squared loss). For this case, we establish an upper bound on the excess population risk of $\widetilde{O}\left(\frac{\|w^*\|}{\sqrt{n}} + \min\left\{\frac{\|w^*\|^2}{(n\epsilon)^{2/3}}, \frac{\sqrt{d}\|w^*\|^2}{n\epsilon}\right\}\right)$, where n is the number of samples, d is the dimension of the problem, and w^* is the minimizer of the population risk. Apart from the dependence on $\|w^*\|$, our bound is essentially tight in all parameters. In particular, we show a lower bound of $\widetilde{\Omega}\left(\frac{1}{\sqrt{n}} + \min\left\{\frac{\|w^*\|^{4/3}}{(n\epsilon)^{2/3}}, \frac{\sqrt{d}\|w^*\|}{n\epsilon}\right\}\right)$. We also revisit the previously studied case of Lipschitz losses [SSTT20]. For this case, we close the gap in the existing work and show that the optimal rate is (up to log factors) $\Theta\left(\frac{\|w^*\|}{\sqrt{n}} + \min\left\{\frac{\|w^*\|}{\sqrt{n\epsilon}}, \frac{\sqrt{\text{rank}}\|w^*\|}{n\epsilon}\right\}\right)$, where rank is the rank of the design matrix. This improves over existing work in the high privacy regime. Finally, our algorithms involve a private model selection approach that we develop to enable attaining the stated rates without a-priori knowledge of $\|w^*\|$.

1 Introduction

Ensuring privacy of users' data in machine learning algorithms is an important desideratum for multiple domains, such as social and medical sciences. Differential privacy (DP) has become the gold standard of privacy-preserving data analysis as it offers formal and quantitative privacy guarantees and enjoys many attractive properties from an algorithmic design perspective [DR+14]. However, for various basic machine learning tasks, the known risk guarantees are potentially sub-optimal and pessimistic.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

		H-Smooth, Non-negative				G-Lipschitz	
		$\sqrt{H}\,\Vert w^*\Vert\,\Vert\mathcal{X}\Vert \leq \Vert\mathcal{Y}\Vert$		$\sqrt{H}\,\Vert w^*\Vert\,\Vert\mathcal{X}\Vert > \Vert\mathcal{Y}\Vert$			
		$d \leq \left(\frac{\Vert w^*\Vert\sqrt{H}\Vert\mathcal{X}\Vert n\epsilon}{\Vert\mathcal{Y}\Vert}\right)^{2/3}$	$d > \left(\frac{\Vert w^*\Vert\sqrt{H}\Vert\mathcal{X}\Vert n\epsilon}{\Vert\mathcal{Y}\Vert}\right)^{2/3}$	$d \leq (n\epsilon)^{2/3}$	$d > (n\epsilon)^{2/3}$	rank $\leq n\epsilon$	rank $> n\epsilon$
DP	UB	$\frac{\sqrt{H}\Vert w^*\Vert\Vert\mathcal{X}\Vert\Vert\mathcal{Y}\Vert\sqrt{d}}{n\epsilon}$ Theorem 1	$\frac{\left(\sqrt{H}\Vert w^*\Vert\Vert\mathcal{X}\Vert\right)^{4/3}\Vert\mathcal{Y}\Vert^{2/3}}{(n\epsilon)^{2/3}}$ Theorem 2	$\frac{H\Vert w^*\Vert^2\Vert\mathcal{X}\Vert^2\sqrt{d}}{n\epsilon}$ Theorem 1	$\frac{H\Vert w^*\Vert^2\Vert\mathcal{X}\Vert^2}{(n\epsilon)^{2/3}}$ Theorem 2, 3	$\frac{G\Vert w^*\Vert\Vert\mathcal{X}\Vert\sqrt{\text{rank}}}{n\epsilon}$ [SSTT20]	$\frac{G\Vert w^*\Vert\Vert\mathcal{X}\Vert}{\sqrt{n\epsilon}}$ Theorem 10, 5
	LB	Tight Theorem 4	Tight Theorem 4	$\frac{\sqrt{H}\Vert w^*\Vert\Vert\mathcal{X}\Vert\Vert\mathcal{Y}\Vert\sqrt{d}}{n\epsilon}$ Theorem 4	$\frac{\left(\sqrt{H}\Vert w^*\Vert\Vert\mathcal{X}\Vert\right)^{4/3}\Vert\mathcal{Y}\Vert^{2/3}}{(n\epsilon)^{2/3}}$ Theorem 4	Tight Theorem 6	Tight Theorem 6
Non-private	UB	$\frac{\sqrt{H}\Vert\mathcal{X}\Vert\Vert\mathcal{Y}\Vert\Vert w^*\Vert}{\sqrt{n}}$ [SST10]				$\frac{G\Vert w^*\Vert\Vert\mathcal{X}\Vert}{\sqrt{n}}$ [NY83]	
	LB	$\frac{\min\{\Vert\mathcal{Y}\Vert,\Vert w^*\Vert\Vert\mathcal{X}\Vert\}}{\sqrt{n}} + \min\left\{\Vert\mathcal{Y}\Vert^2, \frac{H\Vert w^*\Vert^2\Vert\mathcal{X}\Vert^2+d\Vert\mathcal{Y}\Vert^2}{n}, \frac{\sqrt{H}\Vert w^*\Vert\Vert\mathcal{Y}\Vert\Vert\mathcal{X}\Vert}{\sqrt{n}}\right\}$ [SST10, Sha15]				Tight [NY83]	

Table 1: Summary of Rates. Parameters: d: dimension, n: sample size, H: smoothness parameter, w^*: minimum norm population risk minimizer, $\|\mathcal{X}\|$: bound on feature vectors, $\|\mathcal{Y}\|$: bound on loss at zero, G: Lipschitzness parameter, rank: expected rank of the design matrix, ϵ: privacy parameter (δ factors omitted). The actual private excess risk bounds are the sum of the expressions shown in the DP rows and their non-private counterparts. Details on non-private lower bounds in Appendix E.

In this work, we make progress towards resolving one of the most basic machine learning problems under differential privacy: learning linear predictors with convex losses.

1.1 Related Work

Differentially private machine learning has been thoroughly studied for over a decade. In the Lipschitz-convex setting, tight rates are known for both the empirical and population risk [BST14, BFTGT19]. Specifically, it was shown that in the constrained setting, dependence on the dimension in the form of $\Omega\big(\frac{\sqrt{d}}{n\epsilon}\big)$ is unavoidable even for generalized linear models (GLM) (see Section 2 for a formal definition). By contrast, in the unconstrained setting, it has been shown that dimension independent rates are possible for GLMs [JT14]. In this setting, assuming prior knowledge of $\|w^*\|$, the best known rate is $O\left(\frac{\|w^*\|}{\sqrt{n}} + \frac{\|w^*\|\sqrt{\text{rank}}}{n\epsilon}\right)$ [SSTT20], where rank is the expected rank of the design matrix. However, without prior knowledge of $\|w^*\|$, these methods exhibit quadratic dependence on $\|w^*\|$. Furthermore, these results crucially rely on the assumption that the loss is Lipschitz to bound the sensitivity. Although gradient clipping has been proposed to remedy this problem [SSTT20, CWH20], it is known that the solution obtained by clipping may not coincide with the one of the original model.

Without Lipschitzness, work on differentially private GLMs has largely been limited to linear regression [Wan18, CWZ21]. Here, dimension independent rates have only been obtained under certain sparsity assumptions.

More generally, smooth non-negative losses have been studied in the non-private setting by [SST10], where it was shown such functions can obtain risk guarantees with linear dependence on the minimizer norm (as in the Lipschitz case). This work also established a lower bound of $\Omega\big(\frac{1}{\sqrt{n}}\big)$ on the excess population risk for this class of loss functions. [Sha15] additionally establishes a lower bound of $\Omega\left\{\min\left\{\frac{\|w^*\|^2+d}{n}, \frac{\|w^*\|}{\sqrt{n}}\right\}\right\}$ on the excess population risk by way of linear regression[1].

1.2 Our Contributions

Smooth nonegative GLMs. Our primary contribution is a new and nearly optimal rate for the problem of differentially private learning of smooth GLMs. In this setting, we focus on characterizing the excess risk in terms of n, d, ϵ and $\|w^*\|$. Specifically, we show that it is possible to achieve a rate of $\widetilde{O}\left(\frac{\|w^*\|}{\sqrt{n}} + \min\left\{\frac{\|w^*\|^2}{(n\epsilon)^{2/3}}, \frac{\sqrt{d}\|w^*\|^2}{n\epsilon}\right\}\right)$ on the excess population risk. Our new rates exhibit an interesting low/high dimensional transition at $d \approx (\|w^*\|\, n\epsilon)^{2/3}$. First, in the low dimensional regime, we develop a novel analysis of noisy gradient decent (GD) inspired by techniques from [SST10]. In particular, we show that Noisy GD gives an improved rate for non-negative smooth functions (not necessarily GLMs). This is based on an average stability analysis of Noisy GD. As we elaborate in

[1] The [SST10] bound assumes $\|\mathcal{Y}\|, H, \|\mathcal{X}\| = \Omega(1)$. The bounds of [Sha15] were originally stated for the constrained setting, but can easily be converted. More details in Appendix E.

Section 3.1, a straightforward application of uniform stability leads to sub-optimal bounds and hence a new analysis is required. We note in passing that this upper bound works for (unconstrained) DP-SCO with smooth (non-Lipschitz) losses, which is of independent interest. For the high dimensional regime, we perform random projections of the data (specifically, the Johnson-Lindenstrauss transform) for dimensionality reduction, roughly reducing the problem to its low dimensional counterpart. We also develop a lower bound for the excess risk under DP of $\widetilde{\Omega}\left(\min\left\{\frac{\|w^*\|^{4/3}}{(n\epsilon)^{2/3}}, \frac{\sqrt{d}\|w^*\|}{n\epsilon}\right\}\right)$. We note that non-privately a lower bound of $\widetilde{\Omega}\left(\frac{1}{\sqrt{n}} + \min\left\{\frac{d+\|w^*\|^2}{n}, \frac{\|w^*\|}{\sqrt{n}}\right\}\right)$ is known on the excess population risk [SST10, Sha15]. We note that these private and non-private lower bounds imply that our bound is optimal up to factors of $\|w^*\|$ (see Table 1).

Lipschitz GLMs. For the Lipschitz case, we close a subtle but important gap in existing rates. In this setting, it has been shown that one can characterize the excess risk in terms of the expected rank of the design matrix, rank, instead of d [SSTT20]. In this setting, the best known rate was $\widetilde{O}\left(\frac{\|w^*\|}{\sqrt{n}} + \frac{\sqrt{\text{rank}}\|w^*\|}{n\epsilon}\right)$. We show an improved rate of $\widetilde{O}\left(\frac{\|w^*\|}{\sqrt{n}} + \min\left\{\frac{\|w^*\|}{\sqrt{n\epsilon}}, \frac{\sqrt{\text{rank}}\|w^*\|}{n\epsilon}\right\}\right)$. This improves in the high privacy regime where $\epsilon \leq \frac{\text{rank}}{n}$. In fact, the upper bound $O\left(\frac{\|w^*\|}{\sqrt{n\epsilon}}\right)$ for this rate can be obtained with only minor adjustments to the regularization method of [JT14]. Our second contribution in this setting is extending the lower bound of [SSTT20] to hold for all values of $\|w^*\| > 0$ and rank $\in [n]$. This is in contrast to the original lower bound which only holds for problem instances where $\|w^*\|^2 = \text{rank}$ and rank $\in [n\epsilon]$.

Model selection. As part of our methods, we develop a differentially private model selection approach which eliminates the need for a-priori knowledge of $\|w^*\|$. Although such methods are well established in the non-private case, (see e.g. [SSBD14]), in the private case no such methods have been established. Our method, as in the non-private case, performs a grid search over estimates of $\|w^*\|$ and picks the best model based on the loss. However, in the private setting we must account for the fact the the loss evaluation must be privatized. This is non-trivial in the non-Lipschitz smooth case as the loss at a point w may grow quadratically with $\|w\|$.

Lower bounds for Non-Euclidean DP-SCO. Our lower bound construction generalizes to Non-Euclidean ℓ_p/ℓ_q variants of DP-SCO with Lipschitz convex losses [BGN21]. Herein, we assume that the loss function is G_q-Lipschitz with respect to ℓ_q norm, and radius of the constraint set is bounded in ℓ_p norm by B_p. For this setting, we give a lower bound of $\Omega\left(G_q B_p \min\left(\frac{1}{(n\epsilon)^{1/p}}, \frac{d^{(p-1)/p}}{n\epsilon}\right)\right)$ on excess empirical/population risk of any (potentially unconstrained) (ϵ, δ)-DP algorithm; see Corollary 4 in Appendix B.4 for a formal statement and proof. For $p = \infty$ and $p \geq 2, d \leq n\epsilon$, this matches the best known upper bounds in [BGN21].

Non-private settings. As by-products, we give the following new results for the non-private setting. For details on the parameters used below we refer to Table 1.

1. We show that gradient descent, when run on convex non-negative $\widetilde{H}$ smooth functions (not necessarily GLMs), it achieves the optimal rate of $O\big(\frac{\sqrt{\widetilde{H}}\|w^*\|\|\mathcal{Y}\|}{\sqrt{n}}\big)$ (see Corollary 1). This is done via an average-stability analysis of gradient descent. This result is interesting as it also shows GD only needs n iterations, which is known not to work for non-smooth SCO [BFGT20, ACKL21, AKL21].
2. In Section D, we give a procedure to boost the confidence of algorithms for risk minimization with convex non-negative $\widetilde{H}$ smooth functions (not necessarily GLMs). The standard boosting analysis based on Hoeffding's inequality does not give a bound with a linear dependence on the parameters $(\|w^*\|, \|\mathcal{X}\|, \|\mathcal{Y}\|)$, and hence a tighter analysis is required.

1.3 Techniques

Upper bounds. We give two algorithms for both the smooth and Lispschitz cases. The first method is simple and has two main steps. First, optimize the regularized empirical risk over the constraint set $\{w : \|w\| \leq B\}$ for some $B \geq \|w^*\|$. Then output a perturbation of the regularized minimizer with Gaussian noise (which is not requried to be in the constraint set). This method is akin to that of [JT14] with the modification that the regularized minimizer is constrained to a ball. We elaborate on this key difference shortly.

The second method is based on dimensionality reduction. We use smaller dimensional data-oblivious embeddings of the feature vectors. A linear JL transform suffices to give embeddings with the required properties. We then run a constrained DP-SCO method (Noisy GD) in the embedded space, and use the transpose of the JL transform to get a d dimensional model. In this method, the embedding dimension required is roughly the threshold on dimension at which the rates switch from dimension dependent to independent bounds. We also remark that [NêUZ20] applied a similar technique to provide dimension independent classification error guarantees for privately learning support vector machines under hard margin conditions.

We note that a crucial part in all of these methods is the use of constrained optimization as a subroutine, where the constraint set is a ball of radius $\|w^*\|$. This is in stark contrast to the Lipschitz case where existing methods such as those presented by [JT14, SSTT20] rely on the fact that projection is not required. In the smooth case however, constrained optimization helps ensure that the norm of the gradient is roughly bounded by the diameter of the constraint set. We note that in the high dimensional regime, the property that the *final* output of the algorithm can have large norm is still crucial to the success of our algorithms.

Lower bounds. For our lower bounds in the smooth case we rely on the connection between stability and privacy. Specifically, we will utilize a lemma from [CH12] which bounds the accuracy of one-dimensional differentially private algorithms. We then combine this with packing arguments to obtain stronger lower bounds for high dimensional problems. For the Lipschitz case, we adapt the method of [SSTT20].

2 Preliminaries

In the following we detail several concepts needed for the presentation of this paper.

Risk minimization. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the domain of features, and $\mathcal{Y} \subseteq \mathbb{R}$ be the domain of responses. A linear predictor is any $w \in \mathbb{R}^d$. Let $\ell : \mathbb{R}^d \times (\mathcal{X} \times \mathcal{Y}) \rightarrow \mathbb{R}$ be a loss function. Given some unknown distribution $\mathcal{D}$ over $(\mathcal{X} \times \mathcal{Y})$, we define the population loss $L(w;\mathcal{D}) = \underset{(x,y)\sim\mathcal{D}}{\mathbb{E}}[\ell(w;(x,y))]$. Given some dataset $S \in (\mathcal{X} \times \mathcal{Y})^n$ drawn i.i.d. from $\mathcal{D}$, the objective is to obtain $\widehat{w} \in \mathbb{R}^d$ which minimizes the excess population risk, $L(\widehat{w};\mathcal{D}) - \min_{w\in\mathbb{R}^d}\{L(w;\mathcal{D})\}$. Given a population risk minimization problem, we will denote w^* to be a *minimum norm* solution to this problem. We define the empirical risk as $\widehat{L}(w;S) = \frac{1}{n}\sum_{(x,y)\in S}\ell(w;(x,y))$. We define the following quantities for notational convenience: $\varepsilon_{\mathsf{risk}}(w) = L(w;\mathcal{D}) - L(w^*;\mathcal{D})$, $\varepsilon_{\mathsf{erm}}(w) = \widehat{L}(w;S) - \widehat{L}(w^*;S)$, and $\varepsilon_{\mathsf{gen}}(w) = L(w;\mathcal{D}) - \widehat{L}(w;S)$. We define $\mathcal{B}_B$ to be the Euclidean ball of radius B on $\mathbb{R}^d$.

Generalized linear models. We will more specifically be interested in the problem of learning generalized linear models, where there exists some function $\phi : \mathbb{R} \times \mathcal{Y} \rightarrow \mathbb{R}$ such that the loss function can be written as $\ell(w;(x,y)) = \phi(\langle w, x\rangle, y)$. We define parameter bounds $\|\mathcal{X}\| = \max_{x\in\mathcal{X}}\|x\|$ and $\|\mathcal{Y}\|^2 = \max_{y\in\mathcal{Y}}|\phi(0,y)|$. Note that for many common loss functions, the latter condition is the moral equivalent of assuming labels bounded by $\|\mathcal{Y}\|$. For ease of notation, we write $\phi(\langle w, x\rangle, y)$ as $\phi_y(\langle w, x\rangle)$ and denote function $\phi_y : z \mapsto \phi(z,y)$. We say that the loss function is G-Lipschitz GLM if all $y \in \mathcal{Y}$, $\phi_y : \mathbb{R} \rightarrow \mathbb{R}$ is G-Lipschitz. We similarly define H-smooth GLM.

Differential privacy [DKM$^+$06]. We restrict our investigation to the class of algorithms which minimize the excess population risk under the constraint of differential privacy. A randomized algorithm $\mathcal{A}$ is said to be (ϵ, δ) differentially private (i.e., (ϵ, δ)-DP) if for any pair of datasets S and S' differing in one point and any event $\mathcal{E}$ in the range of $\mathcal{A}$ it holds that

$$\mathbb{P}[\mathcal{A}(S) \in \mathcal{E}] \leq e^{\epsilon}\mathbb{P}[\mathcal{A}(S') \in \mathcal{E}] + \delta.$$

For our lower bounds, we will make use of the following Lemma from [CH12].

Lemma 1. *Let $\mathcal{Z}$ be a data domain and let S and S' be two datasets each in $\mathcal{Z}^n$ that differ in at most Δ entries, and let $\mathcal{A} : \mathcal{Z}^n \to \mathbb{R}$ be any (ϵ, δ)-DP algorithm. For all $\tau \in \mathbb{R}$, if $\Delta \leq \frac{\log(1/2\gamma)}{\epsilon}$ and $\delta \leq \frac{1}{16}(1 - e^{-\epsilon})$, then $\mathbb{E}\left[|\mathcal{A}(S) - \tau| + |\mathcal{A}(S') - \tau'|\right] \geq \frac{1}{4}|\tau - \tau'|$.*

Finally, we introduce the Johnson-Lindenstrauss (JL) transform to perform random projections.

Definition 1 ((α, β)-JL property)**.** *A distribution over matrices $\mathbb{R}^{k \times d}$ satisfies (α, β)-JL property if for any $u, v \in \mathbb{R}^d$, $\mathbb{P}\left[|\langle \Phi u, \Phi v\rangle - \langle u, v\rangle| > \alpha \|u\| \|v\|\right] \leq \beta$.*

It is well known that several such "data-oblivious" (i.e. independent of u,v) distributions exist with $k = O\left(\frac{\log(1/\beta)}{\alpha^2}\right)$ [Nel11]. We note that the JL property is typically described as approximation of norms (or distances), but it is easy to deduce the above dot product preservation property from it; for completeness we give this as Lemma 12. Finally, we use $\Phi\mathcal{D}$ to denote the push-forward measure of the distribution $\mathcal{D}$ under the map $\Phi : (x, y) \mapsto (\Phi x, y)$. Similarly, given a data set $S = \{(x_i, y_i)\}_i$, we define $\Phi S := \{(\Phi x_i, y_i)_i\}$

3 Smooth Non-negative GLMs

For smooth non-negative GLMs, we present new upper and lower bounds on the excess risk. For our upper bounds, we here assume that the algorithm is given access to some upper bound on $\|w^*\|$, that we denote by B. We later show in Section 5 how to obtain such a rate without prior knowledge of $\|w^*\|$. We also emphasize that the privacy of these algorithms holds regardless of whether or not $B \geq \|w^*\|$.

3.1 Upper Bounds

Before presenting our algorithms, we highlight some key ideas underlying all our methods. A crucial property of non-negative smooth loss functions which allows one to bound sensitivity, and thus ensure privacy, is the self-bounding property (e.g. Sec. 12.1.3 in [SSBD14]), which states that for an $\widetilde{H}$-smooth non-negative function f and $u \in \mathrm{dom(f)}$, $\|\nabla f(u)\| \leq \sqrt{4\widetilde{H}f(u)}$.

This property implies that the gradient grows at most linearly with $\|w\|$. More precisely, we have the following.

Lemma 2. *Let ℓ be an H-smooth non-negative GLM. Then for any $w \in \mathcal{B}_B$ and $(x, y) \in (\mathcal{X} \times \mathcal{Y})$ we have $\|\nabla \ell(w, (x, y))\| \leq 2\|\mathcal{Y}\|\sqrt{H}\|\mathcal{X}\| + 2HB\|\mathcal{X}\|^2$.*

In order to leverage this property, all our algorithms in this setting utilize constrained optimization as a subroutine, where the constraint set is $\mathcal{B}_B$ and the Lipschitz constant is $G = 2\|\mathcal{Y}\|\sqrt{H}\|\mathcal{X}\| + 2HB\|\mathcal{X}\|^2$. This, in conjunction with the self-bounding property ensures reasonable bounds on sensitivity. In turn, this allows us to ensure privacy without excessive levels of noise.

Finally, we note that our upper bounds for smooth GLMs distinguish low and high dimensional regimes, transitioning at $d = \min\left(\left(\frac{B\sqrt{H}\|\mathcal{X}\|}{\|\mathcal{Y}\|}\right)^{\frac{2}{3}}, 1\right)(n\epsilon)^{\frac{2}{3}}$.

3.1.1 Low dimensional regime

We start with the low dimensional setting where we use techniques developed for constrained DP-SCO for Lipschitz losses (not necessarily GLMs). Existing private algorithms for DP-SCO (e.g., [BFTGT19, FKT20]) lead to excess risk bounds that scale with $\frac{HB^2}{\sqrt{n}}$. On the other hand, the optimal non-private rate [SST10] scales with $\frac{\sqrt{H}B}{\sqrt{n}}$, which may indicate that the private rate implied by the known methods is sub-optimal. We show that this gap can be closed by a novel analysis of private GD.

A standard proof of excess risk of (noisy) gradient descent for smooth convex functions is based on uniform stability [HRS15, BFTGT19]. However, this still leads to sub-optimal rates. Hence we turn to an average stability based analysis of GD, yielding the following result.

Theorem 1. *Let ℓ be a non-negative, convex, $\widetilde{H}$-smooth loss function, bounded at zero by $\|\mathcal{Y}\|^2$. Let $B, n > 0$, $n_0 = \frac{\widetilde{H}B^2}{\|\mathcal{Y}\|^2}$, $G = 2\|\mathcal{Y}\|\sqrt{\widetilde{H}} + 2\widetilde{H}B$. Then, for any $\epsilon, \delta > 0$, Algorithm 1 invoked with $\mathcal{W} = \mathcal{B}_B$, $T = n$, $\sigma^2 = \frac{8G^2 T \log(1/\delta)}{n^2\epsilon^2}$, $\eta = \min\left(\frac{B}{\sqrt{T}\max\left(\sqrt{\widetilde{H}}\|\mathcal{Y}\|, \sigma\sqrt{d}\right)}, \frac{1}{4\widetilde{H}}\right)$ is (ϵ,δ)-differentially private. Further, given a dataset S of $n \geq n_0$ i.i.d samples from an unknown distribution $\mathcal{D}$, the excess risk of output of Algorithm 1 is bounded as,*

$$\mathbb{E}[\varepsilon_{\mathsf{risk}}(\widehat{w})] \leq O\left(\frac{\sqrt{\widetilde{H}}B\|\mathcal{Y}\|}{\sqrt{n}} + \frac{GB\sqrt{d\log(1/\delta)}}{n\epsilon}\right).$$

We note that since H-smooth GLMs satisfy the Theorem condition with parameter $\widetilde{H} \leq H\|\mathcal{X}\|^2$, we obtain results for GLMs as a direct corollary.

Non-private risk bound: As a corollary (see Corollary 1 in Appendix A.3.1), with no privacy constraint, the above result (setting $\epsilon \to \infty$ and $\delta = 1$) shows that gradient descent achieves the optimal excess risk bound, previously shown to be achievable by regularized ERM and one-pass SGD [SST10]. The lower bound, n_0, simply means that the trivial solution "zero" has larger excess risk.

Proof sketch of Theorem 1. The privacy proof simply follows from [BST14] since the loss function is G-Lipschitz in the constraint set. For utility, we first introduce two concepts used in the proof. Let S be a dataset of n i.i.d samples $\{(x_i, y_i)\}_{i=1}^n$, $S^{(i)}$ be the dataset where the i-th data point is replaced by an i.i.d. point (x', y'). Let $\mathcal{A}$ be an algorithm which takes a dataset as input and outputs $\mathcal{A}(S)$. The **average argument stability** of $\mathcal{A}$, denoted as $\varepsilon_{\mathsf{av-stab}}(\mathcal{A})$ is defined as

$$\varepsilon_{\mathsf{av-stab}}(\mathcal{A})^2 = \mathbb{E}_{S,i,(x',y')}\|\mathcal{A}(S) - \mathcal{A}(S^{(i)})\|^2.$$

The **average regret** of gradient descent (Algorithm 1) with iterates $\{w_t\}_{t=1}^T$ is

$$\varepsilon_{\mathsf{reg}}(\mathcal{A}; w^*) = \tfrac{1}{T}\textstyle\sum_{j=1}^T \mathbb{E}[\widehat{L}(w_j; S) - \widehat{L}(w^*; S)].$$

The key arguments are as follows: we first bound the generalization error, or on-average stability, in terms of average argument stability and excess empirical risk (Lemma 4). We then bound average argument stability in terms of average regret (Lemma 5). Finally, in Lemma 6, we provide bounds on excess empirical risk and average regret of noisy gradient descent. Substituting these in the excess risk decomposition gives the claimed bound. The full proof with the above lemmas is deferred to Appendix A.3.1.

Algorithm 1 Noisy GD

```
Input: Dataset $S$, loss function $\ell$, constraint set $\mathcal{W}$, steps $T$, learning rate $\eta$, noise scale $\sigma^2$
$w_0 = 0$
for $t = 1$ to $T - 1$ do
    $\xi \sim \mathcal{N}(0, \sigma^2\mathbb{I})$
    $w_{t+1} = \Pi_{\mathcal{W}}\left(w_t - \eta\left(\widehat{L}(w_t; S) + \xi\right)\right)$
    where $\Pi_{\mathcal{W}}$ is the Euclidean projection on to $\mathcal{W}$
end for
Output: $\widehat{w} = \frac{1}{T}\sum_{t=1}^T w_j$
```

3.1.2 High dimensional regime

In the high dimensional setting, we present two techniques.

JL method. In the JL method, Algorithm 2, we use a data-oblivious JL map Φ to embed all feature vectors in dataset S to $k < d$ dimensions. Let dataset $\widetilde{S} = \{(\Phi x_i, y_i)\}_{i=1}^n$. We then run projected Noisy GD method (Algorithm 1) on the loss with dataset $\widetilde{S}$ and the diameter of the constraint set as $2B$. Finally, we map the returned output $\widetilde{w}$ back to d dimensions using $\Phi^\top$ to get $\widehat{w} = \Phi^\top\widetilde{w}$. We note that no projection is performed on $\widehat{w}$ and thus the output may have large norm due to re-scaling induced by $\Phi^\top$.

Algorithm 2 JL Method

```
Input: Dataset $S = \{(x_1, y_1), ..., (x_n, y_n)\}$, loss function $\ell$, $k$, $B, \eta, T$, $\sigma^2$
  1: Sample JL matrix $\Phi \in \mathbb{R}^{k \times d}$
  2: $\widetilde{S} = \{(\Phi x_1, y_1), (\Phi x_2, y_2), \ldots, (\Phi x_n, y_n)\}$
  3: $\widetilde{w} = \text{NoisyGD}(\widetilde{S}, \ell, \mathcal{B}_{2B}, T, \eta, \sigma^2)$
Output: $\widehat{w} = \Phi^\top \widetilde{w}$
```

Algorithm 3 Regularized Output Perturbation

```
Input: Dataset $S = \{(x_1, y_1), ..., (x_n, y_n)\}$, loss function $\ell$, $\lambda$, $B, \sigma^2$
  1: $\widetilde{w} = \arg\min_{w \in \mathcal{B}_B} \left\{ L(w; S) + \frac{\lambda}{2} \|w\|^2 \right\}$
  2: $\xi \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$
Output: $\widehat{w} = \widetilde{w} + \xi$
```

Theorem 2. *Let* $k = O\left(\frac{B\sqrt{H}\|\mathcal{X}\| \log(2n/\delta) n\epsilon}{\|\mathcal{Y}\|\|\mathcal{X}\| + \sqrt{H}B\|\mathcal{X}\|^2}\right)^{2/3}$, $\mathcal{W} = \mathcal{B}_B$, $T = n$, $\sigma^2 = \frac{8G^2 T \log(1/\delta)}{n^2\epsilon^2}$, $\eta = \min\left(\frac{B}{\sqrt{T}\max\left(\sqrt{H}\|\mathcal{X}\|\|\mathcal{Y}\|, \sigma\sqrt{d}\right)}, \frac{1}{4H\|\mathcal{X}\|^2}\right)$ *and* $n_0 = \frac{HB^2\|\mathcal{X}\|^2}{\|\mathcal{Y}\|^2}$. *Algorithm 2 satisfies* (ϵ, δ)*-differential privacy. Given a dataset* S *of* $n \geq n_0$ *i.i.d samples, of the output* $\widehat{w}$ *is bounded by*

$$\mathbb{E}[\varepsilon_{\mathsf{risk}}(\widehat{w})] \leq \widetilde{O}\left(\frac{\sqrt{H}B\|\mathcal{X}\|\|\mathcal{Y}\|}{\sqrt{n}} + \frac{\left(\sqrt{H}B\|\mathcal{X}\|\sqrt{\|\mathcal{Y}\|}\right)^{\frac{4}{3}} + \left(\sqrt{H}B\|\mathcal{X}\|\right)^2}{(n\epsilon)^{\frac{2}{3}}}\right).$$

Proof sketch of Theorem 2. From the JL property, with $k = O(\log(n/\delta)/\alpha^2)$, w.h.p. norms of all feature vectors and w^*, as well as inner products between are preserved upto an α tolerance (see Definition 1). The preservation of norms of feature vectors implies the gradient norms are preserved, and thus privacy guarantee of sub-routine, Algorithm 1, suffices to establish DP. For the utility proof, from our analysis of Noisy GD (Theorem 1) and using the JL property, the excess risk of $\widehat{w}$ under $\mathcal{D}$ w.r.t. the risk of Φw^* under $\Phi\mathcal{D}$ is bounded as,

$$\mathbb{E}[L(\widehat{w}; \mathcal{D}) - L(\Phi w^*; \Phi\mathcal{D})] \leq O\left(\frac{\sqrt{H}\|\mathcal{X}\|B\|\mathcal{Y}\|}{\sqrt{n}} + \frac{\left(\sqrt{H}\|\mathcal{X}\|\|\mathcal{Y}\| + H\|\mathcal{X}\|^2 B^2\right) B\sqrt{k\log(1/\delta)}}{n\epsilon}\right).$$

From smoothness and JL property, the, "JL error" is:

$$\mathbb{E}[L(\Phi w^*; \Phi\mathcal{D})] - L(w^*; \mathcal{D})] \leq \widetilde{O}\left(\frac{HB^2\|\mathcal{X}\|^2}{k}\right).$$

The above is optimized for the value of k prescribed in Theorem 2, substituting which gives the claimed bound.

Constrained regularized ERM + output perturbation. Our second technique is *constrained* regularized ERM with output perturbation (Algorithm 3). A similar technique for the Lipschitz case was seen in [JT14], however we note that the addition of the constraint set $\mathcal{B}_B$ is crucial in bounding the sensitivity in the smooth case.

Theorem 3. *Let* $n_0 = \frac{HB^2\|\mathcal{X}\|^2}{\|\mathcal{Y}\|^2}$. *Then Algorithm 3 run with* $\sigma^2 = O\left(\frac{\left(\|\mathcal{Y}\|^2 + H^2B^2\|\mathcal{X}\|^4\right)\log(1/\delta)}{\lambda^2 n^2 \epsilon^2}\right)$ *and* $\lambda = \left(\frac{\left(\|\mathcal{Y}\| + HB\|\mathcal{X}\|^2\right)\sqrt{H}\|\mathcal{X}\|}{Bn\epsilon}\right)^{2/3} (\log(1/\delta))^{1/3}$ *satisfies* (ϵ, δ)*-differential privacy. Given a dataset* S *of* $n \geq n_0$ *i.i.d samples, the excess risk of its output* $\widehat{w}$ *is bounded as*

$$\mathbb{E}[\varepsilon_{\mathsf{risk}}(\widehat{w})] \leq \widetilde{O}\left(\frac{\sqrt{H}B\|\mathcal{X}\|\|\mathcal{Y}\| + \|\mathcal{Y}\|^2}{\sqrt{n}} + \frac{\left(\sqrt{H}B\|\mathcal{X}\|\right)^{4/3}\|\mathcal{Y}\|^{2/3} + \left(\sqrt{H}B\|\mathcal{X}\|\right)^2}{(n\epsilon)^{2/3}}\right).$$

We note that we can use the same technique in the low dimensional setting too, yielding a rate of $\frac{\sqrt{H}B\|\mathcal{X}\|\|\mathcal{Y}\|+\|\mathcal{Y}\|^2}{\sqrt{n}} + \frac{GB\sqrt{d}}{n\epsilon}$. However, in contrast to Theorem 1 and 2, these results have an additional $\frac{\|\mathcal{Y}\|^2}{\sqrt{n}}$ term (in both regimes). Thus, in the regime when $\|\mathcal{Y}\| \leq \sqrt{H}B\|\mathcal{X}\|$, the two upper bounds are of the same order.

Proof sketch of Theorem 3. The privacy proof follows the Gaussian mechanism guarantee together with the fact that the ℓ_2-sensitivity of constrained regularized ERM is $O\left(\frac{G}{\lambda n}\right)$ [BE02]. For utility, we use the Rademacher complexity based result of [SST10] to bound the generalization error of $\widetilde{w}$. The other term, error from noise, $\mathbb{E}[L(\widehat{w};\mathcal{D}) - L(\widetilde{w};\mathcal{D})] \leq O(H\sigma^2\|\mathcal{X}\|^2)$ from smoothness of GLM. Combining these two and optimizing for λ gives the claimed result.

3.2 Lower Bounds

The proof technique for our lower bound relies on the connection between differential privacy and sensitivity shown in [CH12]. The underlying mechanisms behind the proof is also similar in nature to the method seen in [SU17]. We note that although we present the following theorem for empirical risk, a reduction found in [BFTGT19] implies the following result holds for population risk as well (up to log factors).

Theorem 4. *Let $\epsilon \in [1/n, 1]$, $\delta \leq \frac{1}{16}(1 - e^{-\epsilon})$. For any (ϵ, δ)-DP algorithm, $\mathcal{A}$, there exists a dataset S with empirical minimizer of norm B such that the excess empirical risk of $\mathcal{A}$ is lower bounded by* $\Omega\left\{\min\left\{\|\mathcal{Y}\|^2, \frac{(B\|\mathcal{X}\|)^{4/3}(H\|\mathcal{Y}\|)^{2/3}}{(n\epsilon)^{2/3}}, \frac{\sqrt{d}B\|\mathcal{X}\|\|\mathcal{Y}\|\sqrt{H}}{n\epsilon}\right\}\right\}$.

The problem instance used in the proof is the squared loss function, and thus this result holds additionally for the more specific case of linear regression. We also note this lower bound implies our upper bound is optimal whenever $B\|\mathcal{X}\| \leq \|\mathcal{Y}\|$, which is a commonly studied regime in linear regression [SST10, KL15]. We here provide a proof sketch and defer the full proof to Appendix A.7.

Proof sketch of Theorem 4 Define $F(w;S) = \frac{1}{n}\sum_{(x,y)\in S}(\langle w, x\rangle - y)^2$. Let $d' < \min\{n, d\}$ and $b, p \in [0,1]$ be parameters to be chosen later. For any $\boldsymbol{\sigma} \in \{\pm 1\}^{d'}$, define the dataset $S_{\boldsymbol{\sigma}}$ which consists of the union of d' subdatasets, $S_1, ..., S_{d'}$ given as follows. Set $\frac{pn}{d'}$ of the feature vectors in S_j as $\|\mathcal{X}\|e_j$ (the rescaled j'th standard basis vector) and the rest as the zero vector. Set $\frac{pn}{2d}(1+b)$ of the labels as $\boldsymbol{\sigma}_j\|\mathcal{Y}\|$ and $\frac{pn}{2d}(1-b)$ labels as $-\boldsymbol{\sigma}_j\|\mathcal{Y}\|$. Let $w^{\boldsymbol{\sigma}} = \arg\min_{w\in\mathbb{R}^d}\{F(w;S_{\boldsymbol{\sigma}})\}$ be the ERM minimizer of $F(\cdot;S_{\boldsymbol{\sigma}})$. Following from Lemma 2 of [Sha15] we have that for any $\bar{w} \in \mathbb{R}^d$ that

$$F(\bar{w};S_{\boldsymbol{\sigma}}) - F(w^{\boldsymbol{\sigma}};S_{\boldsymbol{\sigma}}) \geq \frac{p\|\mathcal{X}\|^2}{2d'}\sum_{j=1}^{d'}(\bar{w}_j - w_j^{\boldsymbol{\sigma}})^2. \tag{1}$$

We will now show lower bounds on the per-coordinate error. Consider any $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ which differ only at index j for some $j \in [d']$. Note that the datasets $S_{\boldsymbol{\sigma}}$ and $S_{\boldsymbol{\sigma}'}$ differ in $\Delta = \frac{pn}{2d'}[(1+b)-(1-b)] = \frac{pbn}{d'}$ points. Let $\tau = w_j^{\boldsymbol{\sigma}} = \frac{\|\mathcal{Y}\|b}{\|\mathcal{X}\|}$ and $\tau' = w_j^{\boldsymbol{\sigma}'} = -\frac{\|\mathcal{Y}\|b}{\|\mathcal{X}\|}$ (i.e. the j components of the empirical minimizers for S and S_j' respectively). Note that $|w_j^{\boldsymbol{\sigma}} - w_j^{\boldsymbol{\sigma}'}| = \frac{2\|\mathcal{Y}\|b}{\|\mathcal{X}\|}$. Setting $d' = \left(\frac{p\|\mathcal{X}\|Bn\epsilon}{\|\mathcal{Y}\|}\right)^{2/3}$ and $b = \left(\frac{\|\mathcal{X}\|B}{\|\mathcal{Y}\|\sqrt{pn\epsilon}}\right)^{2/3}$ ensures $\Delta \leq \frac{1}{\epsilon}$, and thus Lemma 1 can be used to obtain

$$\mathbb{E}\left[|\mathcal{A}(S_{\boldsymbol{\sigma}})_j - w_j^{\boldsymbol{\sigma}}|^2 + |\mathcal{A}(S_{\boldsymbol{\sigma}'})_j - w_j^{\boldsymbol{\sigma}'}|^2\right] \geq \frac{1}{32}\frac{\|\mathcal{Y}\|^2 b^2}{\|\mathcal{X}\|^2} = \frac{B^{4/3}\|\mathcal{Y}\|^{2/3}}{32(\|\mathcal{X}\|pn\epsilon)^{2/3}}.$$

One can now show via a packing argument that

$$\sup_{\boldsymbol{\sigma}\in\{\pm 1\}^{d'}}\left\{\mathbb{E}\left[F(\mathcal{A}(S_{\boldsymbol{\sigma}})) - F(w^{\boldsymbol{\sigma}})\right]\right\} \geq \frac{(\|\mathcal{X}\|B)^{4/3}\|\mathcal{Y}\|^{2/3}p^{1/3}}{128(n\epsilon)^{2/3}},$$

The result then follows from setting $p = \min\left\{1, \frac{d^{3/2}\|\mathcal{Y}\|}{B\|\mathcal{X}\|n\epsilon}\right\}$.

4 Lipschitz GLMs

In the Lipschitz case, we close a subtle gap in existing rates. We recall that in this setting a more precise characterization in terms of the expected rank of the design matrix is possible (as opposed to using d). The best known upper bound is $\widetilde{O}\left(\frac{\|w^*\|\sqrt{\text{rank}}}{n\epsilon}\right)$ assuming knowledge of $\|w^*\|$. This bound was shown to be optimal when $\epsilon \geq \text{rank}/n$ and $\|w^*\| = \sqrt{\text{rank}}$ [SSTT20].

We first show that in the high privacy regime where $\epsilon \leq \text{rank}/n$, an improved rate is possible. Specifically, we show that in this regime constrained regularized ERM with output perturbation and achieves the optimal rate. In fact, we note that the method of [JT14] (i.e. *unconstrained* regularized ERM with output perturbation), can obtain this rate when $\epsilon = O(1)$ if the regularization parameter is set differently. We present the constrained version in order to leverage Rademacher complexity arguments and provide a slightly cleaner bound that holds for all $\epsilon > 0$.

Theorem 5. *Algorithm 3 run with parameters* $\sigma^2 = \frac{4G^2\|\mathcal{X}\|^2\log(1/\delta)}{\lambda^2 n^2\epsilon^2}$ *and* $\lambda = \frac{G\|\mathcal{X}\|(\log(1/\delta))^{1/4}}{B\sqrt{n\epsilon}}$ *satisfies* (ϵ,δ)*-DP. Given a dataset of* n *i.i.d. samples from* $\mathcal{D}$*, its output has excess risk* $\mathbb{E}\left[\varepsilon_{\mathsf{risk}}(\widehat{w})\right] = \widetilde{O}\left(\frac{GB\|\mathcal{X}\|}{\sqrt{n}} + \frac{GB\|\mathcal{X}\|}{\sqrt{n\epsilon}}\right)$.

We also state and prove a similar bound using the JL technique in Appendix B.

Next, we generalize the lower bound of [SSTT20] to show this new bound is optimal for all settings of B, rank, and ϵ. We now show that a modification of the lower bound present in [SSTT20] shows our upper bound is tight. We note that their lower bound only held for problem instances where $\text{rank} = \|w^*\|^2$ and $\epsilon \leq \text{rank}/n$. By contrast, the upper bound $O(\frac{\sqrt{\text{rank}}\|w^*\|}{n\epsilon})$ holds for any values of rank and $\|w^*\|$.

Theorem 6. *Let* $G, \|\mathcal{Y}\|, \|\mathcal{X}\|, B > 0$, $\epsilon \leq 1.2$ *and* $\delta \leq \epsilon$*. For any* (ϵ,δ)*-DP algorithm* $\mathcal{A}$*, there exists a* G*-Lipschitz GLM loss bounded at zero by* $\|\mathcal{Y}\|$ *and a distribution* $\mathcal{D}$ *with* $\|w^*\| \leq B$ *such that the output of* $\mathcal{A}$ *on* $S \sim \mathcal{D}^n$ *satisfies* $\mathbb{E}[\varepsilon_{\mathsf{risk}}(\mathcal{A}(S))] = \Omega\Big(GB\|\mathcal{X}\|\min\Big(1, \frac{1}{\sqrt{n\epsilon}}, \frac{\sqrt{rank}}{n\epsilon}\Big)\Big)$.

All proofs for this section are deferred to Appendix B.

5 Adapting to $\|w^*\|$

Our method for privately adapting to $\|w^*\|$ is given in Algorithm 4. We start by giving a high level overview and defining some necessary preliminaries. The algorithm works in the following manner. First we define a number of “guesses” K for $\|w^*\|$, $B_1, ..., B_K$ where $B_j = 2^j : \forall j \in [K]$. Then given black box access to a DP optimization algorithm, $\mathcal{A}$, Algorithm 4 generates K candidate vectors $w_1, ..., w_K$ using $\mathcal{A}$, training set $S_1 \in (\mathcal{X}\times\mathcal{Y})^{n/2}$, and the guesses $B_1, ..., B_K$. We assume $\mathcal{A}$ satisfies the following accuracy assumption for some confidence parameter $\beta > 0$.

Assumption 1. *There exists a function* $\mathsf{ERR} : \mathbb{R}^+ \mapsto \mathbb{R}^+$ *such that for any* $B \in \mathbb{R}^+$*, whenever* $B \geq \|w^*\|$*, w.p. at least* $1 - \frac{\beta}{4K}$ *under the randomness of* $S_1 \sim \mathcal{D}^{\frac{n}{2}}$ *and* $\mathcal{A}$ *it holds that* $\varepsilon_{\mathsf{risk}}(\mathcal{A}(S_1, B); \mathcal{D}) \leq \mathsf{ERR}(B)$.

After generating the candidate vectors, the goal is to pick guess with the smallest excess population risk in a differentially private manner using a validation set S_2. The following assumption on $\mathcal{A}$ allows us both to ensure the privacy of the model selection algorithm and verify that $\widehat{L}(w_j; S_2)$ provides a tight estimate of $L(w_j; \mathcal{D})$.

Assumption 2. *There exist a function* $\Delta : \mathbb{R}^+ \mapsto \mathbb{R}^+$ *such that for any dataset* $S_2 \in (\mathcal{X}\times\mathcal{Y})^{n/2}$ *and* $B > 0$

$$\underset{\mathcal{A}}{\mathbb{P}}[\exists (x,y) \in S_2 : |\ell(\mathcal{A}(S_1, B); (x,y))| \geq \Delta(B)] \leq \frac{\min\{\delta, \beta\}}{4K}$$

Specifically, our strategy will be to use the Generalized Exponential Mechanism, GenExpMech, of [RS15] in conjunction with a penalized score function. Roughly, this score function penalty ensures the looser guarantees on the population loss estimate when B is large do not interfere with the loss estimates at smaller values of B. We provide the relevant details for GenExpMech in Appendix C.1. We now state our result.

Algorithm 4 Private Grid Search

```
Input: Dataset $\mathcal{S} \in (\mathcal{X} \times \mathcal{Y})^n$, grid parameter $K \in \mathbb{R}$, optimization algorithm: $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^n \times \mathbb{R} \mapsto \mathbb{R}^d$, privacy parameters $(\epsilon, \delta)$
Partition $S$ into two disjoint sets, $S_1$ and $S_2$, of size $\frac{n}{2}$
$w_0 = 0$
for $j \in [K]$ do
    $B_j = 2^j$
    $w_j = \mathcal{A}(S_1, B_j)$
    $\widetilde{L}_j = \widehat{L}(w_j; S_2) + \frac{\Delta(B_j)\log(K/\beta)}{n} + \sqrt{\frac{4\|\mathcal{Y}\|^2 \log(K/\beta)}{n}}$
end for
Set $j^*$ as the output of GenExpMech run with privacy parameter $\frac{\epsilon}{2}$, confidence parameter $\frac{\beta}{4}$, and sensitivity/score pairs $(0, \|\mathcal{Y}\|^2), (\Delta(B_1), \widetilde{L}_1) ..., (\Delta(B_K), \widetilde{L}_K)$,
Output $w_{j^*}$
```

Theorem 7. *Let $\ell : \mathbb{R}^d \times (\mathcal{X} \times \mathcal{Y})$ be a smooth non-negative loss function such that $\ell(0, (x, y)) \leq \|\mathcal{Y}\|^2$ for any $x, y \in (\mathcal{X} \times \mathcal{Y})$. Let $\epsilon, \delta, \beta \in [0, 1]$. Let $K > 0$ satisfy $\mathsf{ERR}(2^K) \geq \|\mathcal{Y}\|^2$. Let $\mathcal{A}$ be an $(\frac{\epsilon}{2K}, \frac{\delta}{2K})$-DP algorithm satisfying Assumption 2. Then Algorithm 4 is (ϵ, δ)-DP. Further, if $\mathcal{A}$ satisfies Assumption 1 and $S_1 \sim \mathcal{D}^{n/2}$ then Algorithm 4 outputs $\bar{w}$ s.t. with probability at least $1 - \beta$,*

$$\varepsilon_{\mathsf{risk}}(\bar{w}; \mathcal{D}) \leq \min\left\{\|\mathcal{Y}\|^2, \mathsf{ERR}(2\max\{\|w^*\|, 1\}) + \sqrt{\frac{4\|\mathcal{Y}\|^2 \log(4K/\beta)}{n}} + \frac{5\Delta(2\max\{\|w^*\|, 1\})}{n\epsilon}\right\}.$$

We note that we develop a generic confidence boosting approach to obtain high probability guarantees from our previously described algorithms in Section 5, and thus obtaining algorithms which satisfy 1 is straightforward. We provide more details on how our algorithms satisfy Assumption 2 in Appendix C.4. The following Theorem details the guarantees implied by this method for output perturbation with boosting (see Theorems 13,15). Full details are in Appendix C.3.

Theorem 8. *Let $K, \epsilon, \delta, \beta > 0$ and $\mathcal{A}$ be the algorithm formed by running 3 with boosting and privacy parameters $\epsilon' = \frac{\epsilon}{K}$, $\delta' = \frac{\delta}{K}$. Then there exists a setting of K such that $K = \Theta\left(\log\left(\max\left\{\frac{\|\mathcal{Y}\|\sqrt{n}}{\|\mathcal{X}\|\sqrt{H}}, \frac{\|\mathcal{Y}\|^2 (n\epsilon)^{2/3}}{\sqrt{H}\|\mathcal{X}\|^2}\right\}\right)\right)$ and Algorithm 4 run with $\mathcal{A}$ and K is (ϵ, δ)-DP and when given $S \sim \mathcal{D}^n$, satisfies the following w.p. at least $1 - \beta$ (letting $B^* = 2\max\{\|w^*\|, 1\}$)*

$$\varepsilon_{\mathsf{risk}}(\widehat{w}) = \widetilde{O}\Bigg(\min\left\{\|\mathcal{Y}\|^2, \frac{\left(\sqrt{H}B^*\|\mathcal{X}\|\right)^{4/3}\|\mathcal{Y}\|^{2/3} + \left(\sqrt{H}B^*\|\mathcal{X}\|\right)^2}{(n\epsilon)^{2/3}} + \frac{\sqrt{H}B^*\|\mathcal{X}\|\max\{\|\mathcal{Y}\|, 1\} + \|\mathcal{Y}\|^2}{\sqrt{n}} + \frac{\|\mathcal{Y}\|^2 + H(B^*\|\mathcal{X}\|)^2}{n\epsilon}\right\}\Bigg).$$

Confidence Boosting: We give an algorithm to boost the confidence of unconstrained, smooth DP-SCO (with possibly non-Lipschitz losses). We split the dataset S into $m + 1$ chunks and run an (ϵ, δ)-DP algorithm over the m chunks to get m models, and then use Report Noisy Max mechanism to select a model with approximately the least empirical risk. We show that this achieves the optimal rate of $\widetilde{O}\left(\frac{\sqrt{H}B\|\mathcal{X}\|\|\mathcal{Y}\|}{\sqrt{n}}\right)$ whereas the previous high probability result of [SST10] had an additional $\widetilde{O}\left(\frac{\|\mathcal{Y}\|^2}{\sqrt{n}}\right)$ term, which was also limited to only GLMs. The key idea is that non-negativity, convexity, smoothness and loss bounded at zero, all together enable strong bounds on the variance of the loss, and consequently give stronger concentration bounds. The details are deferred to Appendix D.

Acknowledgments and Disclosure of Funding

RA and EU are supported, in part, by NSF BIGDATA award IIS-1838139 and NSF CAREER award IIS-1943251. RB's and MM's work on this research is supported by NSF Award AF-1908281 and Google Faculty Research Award. CG's research is partially supported by INRIA Associate Teams project, FONDECYT 1210362 grant, and National Center for Artificial Intelligence CENIA FB210017, Basal ANID. Part of this work was done while CG was at the University of Twente.

References

[ACKL21] Idan Amir, Yair Carmon, Tomer Koren, and Roi Livni. Never go full batch (in stochastic convex optimization). *Advances in Neural Information Processing Systems*, 34, 2021.

[AKL21] Idan Amir, Tomer Koren, and Roi Livni. Sgd generalizes better than gd (and regularization doesn't help). In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 63–92. PMLR, 15–19 Aug 2021.

[BBL03] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Summer school on machine learning*, pages 169–207. Springer, 2003.

[BE02] Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, March 2002.

[BFGT20] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33, 2020.

[BFTGT19] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[BGN21] Raef Bassily, Cristóbal Guzmán, and Anupama Nandi. Non-euclidean differentially private stochastic convex optimization. In *Conference on Learning Theory*, pages 474–499. PMLR, 2021.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[BST14] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

[CH12] Kamalika Chaudhuri and Daniel J. Hsu. Convergence rates for differentially private statistical estimation. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012.

[CWH20] Xiangyi Chen, Steven Z. Wu, and Mingyi Hong. Understanding gradient clipping in private sgd: A geometric perspective. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13773–13782. Curran Associates, Inc., 2020.

[CWZ21] T. Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825 – 2850, 2021.

[DKM^{+}06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, 2006.

[DR^{+}14] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

[FKT20] Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

[HRS15] Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *CoRR*, abs/1509.01240, 2015.

[JT14] Prateek Jain and Abhradeep Guha Thakurta. (near) dimension independent risk bounds for differentially private learning. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 476–484, Bejing, China, 22–24 Jun 2014. PMLR.

[KL15] Tomer Koren and Kfir Y Levy. Fast rates for exp-concave empirical risk minimization. In *NIPS*, pages 1477–1485, 2015.

[Nel11] Jelani Nelson. *Sketching and streaming high-dimensional vectors*. PhD thesis, Massachusetts Institute of Technology, 2011.

[NêUZ20] Huy Lê Nguyễn, Jonathan Ullman, and Lydia Zakynthinou. Efficient Private Algorithms for Learning Large-Margin Halfspaces. In Aryeh Kontorovich and Gergely Neu, editors, *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 704–724. PMLR, 08 Feb–11 Feb 2020.

[NY83] A.S. Nemirovsky and E.R. Yudin. *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience publication. Wiley, 1983.

[RS15] Sofya Raskhodnikova and Adam D. Smith. Efficient lipschitz extensions for high-dimensional graph statistics and node private degree distributions. *CoRR*, abs/1504.07912, 2015.

[Sha15] Ohad Shamir. The sample complexity of learning linear predictors with the squared loss. *J. Mach. Learn. Res.*, 16:3475–3486, 2015.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[SST10] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[SSTT20] Shuang Song, Thomas Steinke, Om Thakkar, and Abhradeep Thakurta. Characterizing private clipped gradient descent on convex generalized linear problems. *CoRR*, abs/2006.06783, 2020.

[SU17] Thomas Steinke and Jonathan Ullman. Tight lower bounds for differentially private selection. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 552–563. IEEE, 2017.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[VGNA20] Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1):e318, 2020.

[Wan18] Yu-Xiang Wang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. In Amir Globerson and Ricardo Silva, editors, *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 93–103. AUAI Press, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work is largely theoretical. Moreover, there are no foreseeable negative societal impacts for our proposed algorithms.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Missing Proofs from Section 3.1 (Smooth GLMs)

A.1 Utility Lemmas

Fact 1. *[SSBD14] For a $\widetilde{H}$-smooth non-negative function f, and for any $u \in \mathrm{dom}(f)$, we have $\|\nabla f(u)\| \leq \sqrt{4\widetilde{H} f(u)}$.*

A.2 Proof of Lemma 2

From the self-bounding property (Fact 1), the $\|\mathcal{Y}\|^2$ bound on loss at zero, and smoothness, we have the following bound on the gradient:

$$\begin{aligned}
\|\nabla \ell(w;(x,y)\| &\leq \|\nabla \ell(0;(x,y))\| + \|\nabla \ell(w;(x,y)) - \nabla \ell(0;(x,y))\| \\
&\leq 2\sqrt{H}\,\|x\|\,\ell(0;(x,y)) + H\,\|x\|^2\,\|w\| \\
&\leq 2\left(\sqrt{H}\,\|\mathcal{Y}\|\,\|x\| + H\,\|x\|^2\,\|w\|\right). \qquad (2)
\end{aligned}$$

Lemma 3. *For any $w \in \mathcal{B}_B$ and any $(x,y) \in (\mathcal{X} \times \mathcal{Y})$ it holds that $\ell(w;(x,y)) \leq 3(\|\mathcal{Y}\|^2 + HB^2\|\mathcal{X}\|^2)$.*

Proof. Using the fact that the loss function is $G = \|\mathcal{Y}\|\sqrt{H}\|\mathcal{X}\| + BH\|\mathcal{X}\|^2$-Lipschitz in the constraint set (Lemma 2) we have

$$\begin{aligned}
\ell(w;(x,y)) &\leq \ell(0;(x,y)) + |\ell(w;(x,y)) - \ell(0;(x,y))| \\
&\leq \|\mathcal{Y}\|^2 + G\,\|w\| \\
&\leq \|\mathcal{Y}\|^2 + \|\mathcal{Y}\|\sqrt{H}B\|\mathcal{X}\| + HB^2\|\mathcal{X}\|^2 \\
&\leq 3\left(\|\mathcal{Y}\|^2 + HB^2\|\mathcal{X}\|^2\right)
\end{aligned}$$

where the last step follow from AM-GM inequality. □

A.3 Low Dimension

Before presenting the proof of Theorem 1, we provide formal statements of its Corollaries.

Corollary 1, stated below, gives an upper bound on excess risk of gradient descent in the non-private setting.

Corollary 1. *Let ℓ be a non-negative convex $\widetilde{H}$ smooth loss function, bounded at zero by $\|\mathcal{Y}\|^2$. Let $n_0 = \frac{\widetilde{H}B^2}{\|\mathcal{Y}\|^2}$, $\mathcal{W} = \mathcal{B}_B$, $T = n$, and $\eta = \min\left(\frac{B}{\sqrt{T}\sqrt{\widetilde{H}}\|\mathcal{Y}\|}, \frac{1}{4\widetilde{H}}\right)$. Given a dataset S of $n \geq n_0$ i.i.d samples from an unknown distribution $\mathcal{D}$, the excess risk of output of Algorithm 1 with $\sigma = 0$ is bounded as,*

$$\mathbb{E}[L(\widehat{w};\mathcal{D}) - L(w^*;\mathcal{D})] \leq O\left(\frac{\sqrt{\widetilde{H}}B\|\mathcal{Y}\|}{\sqrt{n}}\right).$$

Corollary 2 below gives an upper bound on excess risk of noisy gradient descent for non-negative smooth GLMs.

Corollary 2. *Let $n_0 = \frac{H\|\mathcal{X}\|^2B^2}{\|\mathcal{Y}\|^2}$, $\eta = \min\left(\frac{B}{\sqrt{T}\max\left(\sqrt{H}\|\mathcal{X}\|\|\mathcal{Y}\|,\sigma\sqrt{d}\right)}, \frac{1}{4H\|\mathcal{X}\|}\right)$, $\mathcal{W} = \mathcal{B}_B$, $\sigma^2 = \frac{8G^2T\log(1/\delta)}{n^2\epsilon^2}$ and $T = n$. Let ℓ be a non-negative convex H smooth GLM, bounded at zero by $\|\mathcal{Y}\|^2$. Algorithm 1 satisfies (ϵ,δ)-differential privacy. Given a dataset $S \sim \mathcal{D}^n$, $n \geq n_0$, then the excess risk of output of Algorithm 1 is bounded as,*

$$\mathbb{E}[L(\widehat{w};\mathcal{D}) - L(w^*;\mathcal{D})] = O\left(\frac{\sqrt{H}\|\mathcal{X}\|B\|\mathcal{Y}\|}{\sqrt{n}} + \frac{\left(\sqrt{H}\|\mathcal{X}\|\|\mathcal{Y}\| + H\|\mathcal{X}\|^2B^2\right)B\sqrt{d\log(1/\delta)}}{n\epsilon}\right).$$

A.3.1 Proof of Theorem 1

Since Algorithm 1 uses a projection step, the iterates always lie in the constraint set $\{w : \|w\| \leq B\}$. Hence, the function over this constraint set is G-Lipschitz. From the analysis of Noisy (S)GD in [BST14, BFTGT19], we have that the setting of noise variance σ^2 ensures that the algorithm satisfies (ϵ, δ)-DP

We now move to the utility part. We start with the decomposition of excess risk as

$$\mathbb{E}[L(\widehat{w};\mathcal{D}) - L(w^*;\mathcal{D})] = \mathbb{E}[L(\widehat{w};\mathcal{D}) - \widehat{L}(\widehat{w};S)] + \mathbb{E}[\widehat{L}(\widehat{w};S) - \widehat{L}(w^*;S)] \tag{3}$$

The key arguments are as follows: we first bound generalization gap, or on-average stability (first term in the right hand side above), in terms of average argument stability and excess empirical risk (Lemma 4). We then bound average argument stability in terms of average regret (Lemma 5). Finally, in Lemma 6, we provide bounds on excess empirical risk and average regret of gradient descent. Substituting these in the above equation gives the claimed bound. We now fill in the details.

We start with Lemma 4 which gives the following bound on the generalization gap:

$$\mathbb{E}[L(\mathcal{A}(S);\mathcal{D}) - \widehat{L}(\widehat{w};S)] \leq \frac{\sqrt{\widetilde{H}}B}{\sqrt{n}\,\|\mathcal{Y}\|}\left(\mathbb{E}[\widehat{L}(\widehat{w};S) - \widehat{L}(w^*;S)]\right) + \frac{2\sqrt{\widetilde{H}n}\|\mathcal{Y}\|}{B}\varepsilon_{\mathsf{av-stab}}(\mathcal{A})^2 + \frac{\sqrt{\widetilde{H}}B\|\mathcal{Y}\|}{\sqrt{n}} \tag{4}$$

Substituting the bound on $\varepsilon_{\mathsf{av-stab}}(\mathcal{A})$ from Lemma 5, the second term becomes,

$$\begin{aligned}\frac{2\sqrt{\widetilde{H}n}\|\mathcal{Y}\|}{B}\varepsilon_{\mathsf{av-stab}}(\mathcal{A})^2 &\leq \frac{16\widetilde{H}^{3/2}\eta^2T\|\mathcal{Y}\|}{\sqrt{n}B}\frac{1}{n}\sum_{j=1}^{T}\mathbb{E}[\widehat{L}(w_j;S) - \widehat{L}(w^*;S)] + \frac{16\widetilde{H}^{3/2}\eta^2T\|\mathcal{Y}\|}{\sqrt{n}B}\\ &\leq \frac{16\sqrt{\widetilde{H}}B}{\sqrt{n}\|\mathcal{Y}\|}\frac{1}{n}\sum_{j=1}^{T}\mathbb{E}[\widehat{L}(w_j;S) - \widehat{L}(w^*;S)] + \frac{16\sqrt{\widetilde{H}}B}{\sqrt{n}\|\mathcal{Y}\|}\\ &\leq \frac{16}{n}\sum_{j=1}^{T}\mathbb{E}[\widehat{L}(w_j;S) - \widehat{L}(w^*;S)] + \frac{16\sqrt{\widetilde{H}}B}{\sqrt{n}\|\mathcal{Y}\|}.\end{aligned}$$

Substituting the above in Eqn. (4) and using the fact that $\frac{\sqrt{\widetilde{H}}B}{\sqrt{n}\|\mathcal{Y}\|} \leq 1$ from the lower bound on n, we get:

$$\begin{aligned}\mathbb{E}[L(\mathcal{A}(S);\mathcal{D}) - \widehat{L}(\widehat{w};S)] \leq \left(\mathbb{E}[\widehat{L}(\widehat{w};S) - \widehat{L}(w^*;S)]\right) + \frac{16}{n}\left(\sum_{j=1}^{T}\mathbb{E}[\widehat{L}(w_j;S) - \widehat{L}(w^*;S)]\right)\\ + \frac{17\sqrt{\widetilde{H}}B\|\mathcal{Y}\|}{\sqrt{n}}.\end{aligned}$$

From the excess empirical risk guarantee (Lemma 6), the terms excess empirical risk $\mathbb{E}[\widehat{L}(\widehat{w};S) - \widehat{L}(w^*;S)]$ and average regret $\frac{1}{n}\left(\sum_{j=1}^{T}\mathbb{E}[\widehat{L}(w_j;S) - \widehat{L}(w^*;S)]\right)$ are both bounded by the same quantity. Thus, substituting the above in Eqn. (3) and substituting the bound from 6, we have,

$$\begin{aligned}\mathbb{E}[L(\widehat{w};\mathcal{D}) - L(w^*;\mathcal{D})] &\leq \frac{18}{n}\sum_{j=1}^{n}\mathbb{E}[\widehat{L}(w_j;S) - \widehat{L}(w^*;S)] + \frac{17\sqrt{\widetilde{H}}B\|\mathcal{Y}\|}{\sqrt{n}}\\ &\leq O\left(\frac{\sqrt{H}B\|\mathcal{Y}\|}{\sqrt{n}} + \frac{\sqrt{d}GB\log{(1/\delta)}}{n\epsilon}\right).\end{aligned}$$

Substituting the value of G completes the proof.

Lemma 4. *Let $n \geq \frac{\widetilde{H}B^2}{\|\mathcal{Y}\|^2}$. Let ℓ be a non-negative $\widetilde{H}$ smooth convex loss function. Let S be a dataset of n i.i.d. samples from an unknown distribution $\mathcal{D}$, and w^* denote the optimal population risk minimizer. The generalization gap of algorithm $\mathcal{A}$ is bounded as,*

$$\mathbb{E}[L(\mathcal{A}(S);\mathcal{D})-\widehat{L}(\mathcal{A}(S);S)] \leq \frac{\sqrt{\widetilde{H}}B}{\sqrt{n}\,\|\mathcal{Y}\|}\left(\mathbb{E}[\widehat{L}(\mathcal{A}(S);S)-\widehat{L}(w^*;S)]\right)+\frac{2\sqrt{\widetilde{H}n}\|\mathcal{Y}\|}{B}\varepsilon_{\mathsf{av-stab}}(\mathcal{A})^2 + \frac{\sqrt{\widetilde{H}}B\|\mathcal{Y}\|}{\sqrt{n}}.$$

Proof. Let $\widehat{w} := \mathcal{A}(S)$, $S^{(i)}$ be the dataset where the i-th data point is replaced by an i.i.d. point (x', y') and let $\widehat{w}^{(i)}$ be the corresponding output of $\mathcal{A}$.

A standard fact (see [SSBD14]) is that generalization gap is equal to on-average stability:

$$\mathbb{E}[L(\widehat{w};\mathcal{D})-\widehat{L}(\widehat{w};S)] = \mathbb{E}[\ell(\widehat{w}^{(i)};(x_i,y_i))-\ell(\widehat{w};(x_i,y_i)].$$

From smoothness and self-bounding property, we have,

$$\begin{aligned}\ell(\widehat{w}^{(i)};(x_i,y_i))-\ell(\widehat{w};(x_i,y_i) &\leq \|\nabla\ell(\widehat{w};(x_i,y_i))\|\left\|\widehat{w}-\widehat{w}^{(i)}\right\|+\frac{\widetilde{H}}{2}\left\|\widehat{w}-\widehat{w}^{(i)}\right\|^2\\ &\leq 2\sqrt{\widetilde{H}\ell((\widehat{w};(x_i,y_i))}\left\|\widehat{w}-\widehat{w}^{(i)}\right\|+\frac{\widetilde{H}}{2}\left\|\widehat{w}-\widehat{w}^{(i)}\right\|^2.\end{aligned}$$

Taking expectation, using Cauchy-Schwarz inequality and substituting average argument stability, we get,

$$\begin{aligned}&\mathbb{E}\left[\ell(\widehat{w}^{(i)};(x_i,y_i))-\ell(\widehat{w};(x_i,y_i)\right]\\ &\leq 2\sqrt{\widetilde{H}\mathbb{E}[\ell((\widehat{w};(x_i,y_i))]}\sqrt{\mathbb{E}[\|\widehat{w}-\widehat{w}^{(i)}\|^2]}+\frac{\widetilde{H}}{2}\mathbb{E}\left[\left\|\widehat{w}-\widehat{w}^{(i)}\right\|^2\right] \qquad (5)\\ &\leq 2\sqrt{\widetilde{H}\mathbb{E}[\widehat{L}(\widehat{w};S)]}\varepsilon_{\mathsf{av-stab}}(\mathcal{A})+\frac{\widetilde{H}\varepsilon_{\mathsf{av-stab}}(\mathcal{A})^2}{2}\\ &\leq \frac{\sqrt{\widetilde{H}}B\mathbb{E}[\widehat{L}(\widehat{w};S)]}{\sqrt{n}\,\|\mathcal{Y}\|}+\frac{\sqrt{\widetilde{H}n}\,\|\mathcal{Y}\|\,\varepsilon_{\mathsf{av-stab}}(\mathcal{A})^2}{B}+\frac{\widetilde{H}\varepsilon_{\mathsf{av-stab}}(\mathcal{A})^2}{2}\\ &\leq \frac{\sqrt{\widetilde{H}}B}{\sqrt{n}\,\|\mathcal{Y}\|}\left(\mathbb{E}[\widehat{L}(\widehat{w};S)-\widehat{L}(w^*;S)]\right)+\frac{\sqrt{\widetilde{H}n}\,\|\mathcal{Y}\|\,\varepsilon_{\mathsf{av-stab}}(\mathcal{A})^2}{B}+\frac{\widetilde{H}\varepsilon_{\mathsf{av-stab}}(\mathcal{A})^2}{2}\\ &+\frac{\sqrt{\widetilde{H}}B}{\sqrt{n}\,\|\mathcal{Y}\|}\mathbb{E}[\widehat{L}(w^*;S)]\\ &\leq \frac{\sqrt{\widetilde{H}}B}{\sqrt{n}\,\|\mathcal{Y}\|}\left(\mathbb{E}[\widehat{L}(\widehat{w};S)-\widehat{L}(w^*;S)]\right)+\left(\frac{\sqrt{\widetilde{H}n}\|\mathcal{Y}\|}{B}+\frac{\widetilde{H}}{2}\right)\varepsilon_{\mathsf{av-stab}}(\mathcal{A})^2\\ &+\frac{\sqrt{\widetilde{H}}B\|\mathcal{Y}\|}{\sqrt{n}} \qquad (6)\end{aligned}$$

where the third inequality follows from AM-GM inequality, and last follows since w^* is the optimal solution: $\mathbb{E}[\widehat{L}(w^*;S)] = L(w^*;\mathcal{D}) \leq L(0;\mathcal{D}) \leq \|\mathcal{Y}\|^2$. Finally, using the lower bound on n, we have $\frac{\widetilde{H}}{2} \leq \frac{\sqrt{\widetilde{H}n}\|\mathcal{Y}\|}{B}$, substituting which gives the claimed bound. □

Lemma 5. *The average argument stability for noisy GD (Algorithm 1) run for T iterations with step size $\eta \leq \frac{4}{\widetilde{H}}$ is bounded as*

$$\varepsilon_{\mathsf{av-stab}}(\mathcal{A})^2 \leq \frac{8\widetilde{H}\eta^2T}{n}\frac{1}{n}\sum_{j=1}^{T}\mathbb{E}[\widehat{L}(w_j;S)-\widehat{L}(w^*;S)]+\frac{8\widetilde{H}\eta^2T\|\mathcal{Y}\|^2}{n}.$$

Proof. The uniform argument stability analysis for (Noisy) (S)GD is limited to the Lipschitz setting [HRS15, BFTGT19, BFGT20] and therefore not directly applicable. We therefore need to modify the arguments to give an average stability analysis in smooth (non-Lipschitz) case.

Let $\widehat{w} := \mathcal{A}(S)$, $S^{(i)}$ be the dataset where the i-th data point is replaced by an i.i.d. point (x', y') and let $\widehat{w}^{(i)}$ be the corresponding output of $\mathcal{A}$. Moroever, let w_i denote the iterate of noisy SGD on dataset S and similarly $\widetilde{w}_i^{(i)}$ for dataset $S^{(i)}$.

We simply couple the Gaussian noise sampled at each iteration to be equal on both datasets. Using the fact the the updates are non-expansive, we have

$$\begin{aligned}\left\|w_{t+1} - w'_{t+1}\right\| &\leq \|w_t - w'_t\| + \frac{\eta\left(\|\nabla \ell(w_t; (x_i, y_i))\| + \|\nabla \ell(w'_t; (x'; y'))\|\right)}{n} \\ &\leq \frac{\eta \sum_{j=1}^{t}\left(\|\nabla \ell(w_j; (x_i, y_i))\| + \left\|\nabla \ell(w'_j; (x'; y'))\right\|\right)}{n}.\end{aligned}$$

From the self-bounding property (Fact 1), $(\|\nabla \ell(w_j; (x_i, y_i))\|) \leq 2\sqrt{\widetilde{H}\ell(w_j; (x_i, y_i))})$. Therefore, we get,

$$\begin{aligned}\mathbb{E}[\|w_t - w'_t\|^2] &\leq \frac{4\widetilde{H}\eta^2 T}{n^2}\sum_{j=1}^{t}\left(\mathbb{E}[\ell(w_j; (x_i, y_i)) + \ell(w'_j; (x', y'))]\right) \\ &= \frac{8\widetilde{H}\eta^2 T}{n^2}\sum_{j=1}^{t}\mathbb{E}[\widehat{L}(w_j; S)].\end{aligned}$$

For the average iterate,

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{w} - \widehat{w}^{(i)}\right\|^2\right] &\leq \frac{1}{T^2}T\sum_{t=1}^{T}\mathbb{E}[\|w_t - w'_t\|^2] \leq \frac{8\widetilde{H}\eta^2 T}{n^2}\sum_{j=1}^{T}\mathbb{E}[\widehat{L}(w_j; S)] \\ &\leq \frac{8\widetilde{H}\eta^2 T}{n}\frac{1}{n}\sum_{j=1}^{T}\mathbb{E}[\widehat{L}(w_j; S) - \widehat{L}(w^*; S)] + \frac{8\widetilde{H}\eta^2 T}{n}\mathbb{E}[\widehat{L}(w^*; S)] \\ &\leq \frac{8\widetilde{H}\eta^2 T}{n}\frac{1}{n}\sum_{j=1}^{T}\mathbb{E}[\widehat{L}(w_j; S) - \widehat{L}(w^*; S)] + \frac{8\widetilde{H}\eta^2 T\|\mathcal{Y}\|^2}{n}\end{aligned}$$

where the last inequality follows since w^* is the population risk minimizer: $\mathbb{E}[\widehat{L}(w^*; S)] = L(w^*; \mathcal{D})] \leq L(0; \mathcal{D}) \leq \|\mathcal{Y}\|^2$. □

Lemma 6. *Let $n \geq \frac{\widetilde{H}B^2}{\|\mathcal{Y}\|^2}$, $\eta = \min\left(\frac{B}{\sqrt{T}\max\left(\sqrt{\widetilde{H}}\|\mathcal{Y}\|, \sigma\sqrt{d}\right)}, \frac{1}{4\widetilde{H}}\right)$, $\sigma^2 = \frac{8G^2 T\log(1/\delta)}{n^2\epsilon^2}$ and $T = n$.*

We have,

$$\mathbb{E}\left[\widehat{L}(\widehat{w}; S) - \widehat{L}(w^*; S)\right] \leq \frac{1}{T}\sum_{j=1}^{T}\mathbb{E}\left[\widehat{L}(w_j; S) - \widehat{L}(w^*; S)\right] \leq O\left(\frac{\sqrt{H}B\|\mathcal{Y}\|}{\sqrt{n}} + \frac{\sqrt{d}G\log\left(1/\delta\right)}{n\epsilon}\right).$$

where w^* is the (minimum norm) population risk minimizer.

Proof. From standard analysis of (S)GD,

$$\begin{aligned}\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] &\leq \mathbb{E}\left[\left\|w_t - \eta\left(\nabla\widehat{L}(w_t; S) + \xi_t\right) - w^*\right\|^2\right] \\ &\leq \mathbb{E}\left[\|w_t - w^*\|^2\right] + \eta^2\mathbb{E}\left[\left\|\nabla\widehat{L}(w_t; S)\right\|^2\right] + \eta^2\sigma^2 d - 2\eta\mathbb{E}\left[\widehat{L}(w_t; S) - \widehat{L}(w^*; S)\right] \\ &\leq \mathbb{E}\left[\|w_t - w^*\|^2\right] + 4\eta^2\widetilde{H}\mathbb{E}[\widehat{L}(w_t; S)] + \eta^2\sigma^2 d - 2\eta\mathbb{E}\left[\widehat{L}(w_t; S) - \widehat{L}(w^*; S)\right]\end{aligned}$$

where the last inequality follows from self-bounding property (Fact 1). Rearranging, and using the fact that $\mathbb{E}[\widehat{L}(w^*;S) = L(w^*;\mathcal{D})] \leq L(0;\mathcal{D}) \leq \|\mathcal{Y}\|^2$ we get,

$$\left(1-2\eta\widetilde{H}\right)\mathbb{E}\left[\widehat{L}(w_t;S)-\widehat{L}(w^*;S)\right] \leq \frac{\mathbb{E}\left[\|w_t-w^*\|^2-\|w_{t+1}-w^*\|^2\right]}{2\eta}+2\eta\left(\widetilde{H}\|\mathcal{Y}\|^2+\sigma^2 d\right)$$

From the choice of η, we have $\left(1-2\eta\widetilde{H}\right) \geq \frac{1}{2}$. Averaging over T iterations, we get that the average regret is,

$$\frac{1}{T}\sum_{j=1}^{T}\mathbb{E}\left[\widehat{L}(w_j;S)-\widehat{L}(w^*;S)\right] \leq \frac{B^2}{\eta T}+4\eta\left(\widetilde{H}\|\mathcal{Y}\|^2+\sigma^2 d\right).$$

Setting $\eta = \min\left(\frac{B}{\sqrt{T}\max\left(\sqrt{\widetilde{H}}\|\mathcal{Y}\|,\sigma\sqrt{d}\right)},\frac{1}{4\widetilde{H}}\right)$, we get

$$\frac{1}{T}\sum_{j=1}^{T}\mathbb{E}\left[\widehat{L}(w_j;S)-\widehat{L}(w^*;S)\right] = O\left(\frac{\sqrt{\widetilde{H}}B\|\mathcal{Y}\|}{\sqrt{T}}+\frac{B\sqrt{d}\sigma}{\sqrt{T}}+\frac{\widetilde{H}B^2}{T}\right).$$

Finally, substituting $\sigma^2 = \frac{8G^2T\log(1/\delta)}{n^2\epsilon^2}$, $T=n$ and using the lower bound on n gives the claimed bound on average regret. Applying convexity to lower bound average regret by excess empirical risk of $\widehat{w}$ gives the same bound on excess empirical risk. □

A.4 High Dimension

Proof of Theorem 2. Let $\alpha \leq 1$ be a parameter to be set later. From the JL property with $k = O\left(\frac{\log(2n/\delta)}{\alpha^2}\right)$, with probability at least $1-\delta/2$, for all data points x_i, $\|\Phi x_i\| \leq (1+\alpha)\|x_i\| \leq 2\|\mathcal{X}\|$, and $\|\Phi w^*\|^2 \leq 2\|w^*\|^2 \leq 2B^2$.

Further, by Lemma 2, for any w in the embedding space with $\|w\|^2 \leq B$, with probability at least $1-\delta/2$, the loss is G Lipschitz where $G = 2\|\mathcal{Y}\|\sqrt{H}\|\mathcal{X}\|+2HB\|\mathcal{X}\|^2$. The privacy guarantee now follows from the privacy of Noisy SGD and post-processing.

For the utility guarantee, let $L(w;\Phi\mathcal{D})$ and $\widehat{L}(w;\Phi S)$ denote population and empirical risk (resp) where test and training feature vectors (resp) are mapped using Φ. The excess risk can be decomposed as:

$$\mathbb{E}\left[L(\Phi^\top\widetilde{w};\mathcal{D})-L(w^*;\mathcal{D})\right] = \mathbb{E}\left[L(\widetilde{w};\Phi\mathcal{D})-L(\Phi w^*;\Phi\mathcal{D})\right]+\mathbb{E}\left[L(\Phi w^*;\Phi\mathcal{D})-L(w^*;\mathcal{D})\right]. \tag{7}$$

The first term in Eqn. (7) is bounded by the utility guarantee of the DP-SCO method. In particular, from Lemma 7 (below), we have

$$\mathbb{E}[L(\widetilde{w};\Phi\mathcal{D})-L(\Phi w^*;\Phi\mathcal{D})] = O\left(\frac{\sqrt{H}\|\mathcal{X}\|B\|\mathcal{Y}\|}{\sqrt{n}}+\frac{\left(\sqrt{H}\|\mathcal{X}\|\|\mathcal{Y}\|+H\|\mathcal{X}\|^2B^2\right)B\sqrt{k\log(1/\delta)}}{n\epsilon}\right).$$

The second term in Eqn. (7) is bounded by the JL property together with smoothness and the fact that w^* is the optimal solution thus $\nabla L(w^*;\mathcal{D})=0$. This gives us

$$\mathbb{E}[L(\Phi w^*;\Phi\mathcal{D})]-L(w^*;\mathcal{D})] \leq \frac{H}{2}\mathbb{E}\left[|\langle\Phi w^*,\Phi x\rangle-\langle w^*,x\rangle|^2\right] \leq \frac{\alpha^2H\|w^*\|^2\|\mathcal{X}\|^2}{2} \leq \widetilde{O}\left(\frac{HB^2\|\mathcal{X}\|^2}{k}\right).$$

Combining, we get,

$$\mathbb{E}\left[L(\Phi^\top\widetilde{w};\mathcal{D})-L(w^*;\mathcal{D})\right] \leq \widetilde{O}\left(\frac{\sqrt{H}\|\mathcal{X}\|B\|\mathcal{Y}\|}{\sqrt{n}}+\frac{\left(\sqrt{H}\|\mathcal{X}\|\|\mathcal{Y}\|+H\|\mathcal{X}\|^2B\right)B\sqrt{k\log(1/\delta)}}{n\epsilon}+\frac{HB^2\|\mathcal{X}\|^2}{k}\right).$$

Setting $k = O\left(\frac{BH\|\mathcal{X}\| \log(2n/\delta) n\epsilon}{G}\right)^{2/3}$ completes the proof.

□

Lemma 7. *Let $\Phi \in \mathbb{R}^{d\times k}$ be a data-oblivious JL matrix. Let $S = \{(x_i, y_i)\}_{i=1}^n$ of n i.i.d data points and let $\Phi S := \{(\Phi x_i, y_i)\}_{i=1}^n$. Let $\widetilde{w} \in \mathbb{R}^k$ be the average iterate returned by Algorithm 1 with $\sigma^2 = O\left(\frac{G^2\|\mathcal{X}\|^2 \log(1/\delta)}{n^2\epsilon^2}\right)$ on dataset ΦS. For $k = \Omega\left(\log\left(n\right)\right)$, the excess risk of $\widetilde{w}$ on $\Phi\mathcal{D}$ is bounded as,*

$$\mathbb{E}[L(\widetilde{w}; \Phi\mathcal{D}) - L(\Phi w^*; \Phi\mathcal{D})] \leq O\left(\frac{\sqrt{H}B\|\mathcal{X}\|\|\mathcal{Y}\|}{\sqrt{n}} + \frac{\left(\sqrt{H}\|\mathcal{X}\|\|\mathcal{Y}\| + H\|\mathcal{X}\|^2 B\right) B\sqrt{k\log\left(1/\delta\right)}}{n\epsilon}\right).$$

Proof. Let $S^{(i)}$ be the dataset where the i-th data point is replaced by an i.i.d. point (x', y') and let $\widetilde{w}^{(i)}$ be the corresponding output of Noisy-SGD on $\Phi S^{(i)}$. Define $\overline{S} := \{S, (x', y')\}$ and let $H(\Phi, \overline{S})$ denote an upper bound on the smoothness parameter of the family of loss function $\{\ell(w; (\Phi x; y))\}_{(x,y)\in\overline{S}}$.

We want to apply Theorem 1, but the theorem requires that $n \geq \frac{2H\|\mathcal{X}\|^2 B^2}{\|\mathcal{Y}\|^2}$. In the proof below, we will use it to bound $H(\Phi, \overline{S}) \leq \frac{n\|\mathcal{Y}\|^2}{B^2}$. We use the JL property to get this. Let $\alpha \leq 1$ be a parameter to be set later. Note that $H(\Phi, \overline{S}) \leq H \sup_{(x,y)\in\overline{S}} \|\Phi x\|^2 \leq H\left(1+\alpha\right)\|\mathcal{X}\|^2 \leq 2H\|\mathcal{X}\|^2$ with probability at least $1-\delta$ for $k = O\left(\frac{\log(2n/\delta)}{\alpha^2}\right)$. Thus, if we assume $n \geq \frac{2H\|\mathcal{X}\|^2 B^2}{\|\mathcal{Y}\|^2}$, then w.h.p. $H(\Phi, \overline{S}) \leq \frac{n\|\mathcal{Y}\|^2}{B^2}$. Also from the JL property, $\|\Phi w^*\| \leq 2B$.

Decomposing excess risk and writing generalization gap as on-average stability, we have

$$\begin{aligned}
\mathbb{E}_{\Phi,S}[L(\widetilde{w}; \Phi\mathcal{D}) - L(\Phi w^*; \Phi\mathcal{D})] &= \mathbb{E}_{\Phi,S}[L(\Phi^\top\widetilde{w}; \mathcal{D}) - \widehat{L}(\Phi^\top\widetilde{w}; S)] + \mathbb{E}_{\Phi,S}[\widehat{L}(\Phi^\top\widetilde{w}; S)] - \widehat{L}(\Phi w^*; \Phi S)] \\
&= \mathbb{E}_{\overline{S},\Phi}[\ell(\widetilde{w}^{(i)}; (\Phi x_i, y_i)) - \ell(\widetilde{w}; (\Phi x_i, y_i)) + \widehat{L}(\Phi^\top\widetilde{w}; S) - \widehat{L}(\Phi w^*; \Phi S)].
\end{aligned}$$

We now fix the randomness of $\overline{S}$ and bound the terms in high probability w.r.t. the random Φ. Let $\text{IAS}(\overline{S}, i) = \left\|\widehat{w} - \widehat{w}^{(i)}\right\|$ denote the instance argument stability. Repeating the analysis in Theorem 1, from Eqn. (5), the first term is bounded as,

$$\begin{aligned}
\ell(\widetilde{w}^{(i)}; (\Phi x_i, y_i)) - \ell(\widetilde{w}; (\Phi x_i, y_i)) &\leq \frac{\sqrt{H(\Phi, \overline{S})B}}{\sqrt{n}Y}\left(\ell(\widehat{w}; (\Phi x_i, y_i)) - \widehat{L}(\Phi w^*; \Phi S)\right) \\
&\quad + \left(\frac{\sqrt{H(\Phi, \overline{S})n}\|\mathcal{Y}\|}{B} + \frac{H(\Phi, \overline{S})}{2}\right)\text{IAS}(\overline{S}, i)^2 + \frac{\sqrt{H(\Phi, \overline{S})B\|\mathcal{Y}\|}}{\sqrt{n}} \\
&\leq \frac{2\sqrt{H}\|\mathcal{X}\|B}{\sqrt{n}Y}\left(\ell(\widehat{w}; (\Phi x_i, y_i)) - \widehat{L}(\Phi w^*; \Phi S)\right) \\
&\quad + \frac{4\sqrt{Hn}\|\mathcal{X}\|\|\mathcal{Y}\|}{B}\text{IAS}(\overline{S}, i)^2 + \frac{2\sqrt{HB\|\mathcal{Y}\|}\|\mathcal{X}\|}{\sqrt{n}} \qquad (8)
\end{aligned}$$

where the last inequality holds from application of JL property: with probability at least $1-\delta$, $H(\Phi, \overline{S}) \leq \frac{n\|\mathcal{Y}\|^2}{B^2}$ (from the lower bound on n) and $H(\Phi, \overline{S}) \leq 2H\|\mathcal{X}\|^2$.

As in the proof of Lemma 5, $\text{IAS}(\overline{S}, i)$ is bounded as,

$$\begin{aligned}
&\text{IAS}(\overline{S}, i)^2 \\
&\leq \frac{8H(\Phi, \overline{S})\eta^2 T}{n}\frac{1}{n}\sum_{j=1}^{T}\mathbb{E}[\ell(w_j; (\Phi x_i, y_i)) + \ell(w_j'; (\Phi x', y')) - 2\widehat{L}(\Phi w^*; \Phi S)] + \frac{8H(\Phi, \overline{S})\eta^2 T\|\mathcal{Y}\|^2}{n} \\
&\leq \frac{16H\|\mathcal{X}\|^2\eta^2 T}{n}\frac{1}{n}\sum_{j=1}^{T}\mathbb{E}[\ell(w_j; (\Phi x_i, y_i)) + \ell(w_j'; (\Phi x', y')) - 2\widehat{L}(\Phi w^*; \Phi S)] + \frac{16H\|\mathcal{X}\|^2\eta^2 T\|\mathcal{Y}\|^2}{n}.
\end{aligned}$$

Substituting the above in equation 8, taking expectation with respect to $\overline{S}$ and from manipulations as in the proof of Theorem 1, we get that with probability at least $1-\delta$, we have

$$\begin{aligned}\mathbb{E}_S[L(\Phi^\top \widetilde{w};\mathcal{D}) - \widehat{L}(\Phi^\top \widetilde{w};S)] &\leq 2\left(\mathbb{E}[\widehat{L}(\widehat{w};\Phi S) - \widehat{L}(\Phi w^*;\Phi S)]\right) \\ &\quad + \frac{32}{n}\left(\sum_{j=1}^{T}\mathbb{E}[\widehat{L}(w_j;\Phi S) - \widehat{L}(w^*;\Phi S)]\right) + \frac{34\sqrt{\widetilde{H}}B\|\mathcal{Y}\|}{\sqrt{n}}.\end{aligned}$$

Let $G(\Phi,S) = G = 2\left\|\mathcal{Y}\right\|\sqrt{H(\Phi,\overline{S})} + 2H(\Phi,\overline{S})\left\|\Phi w^*\right\|$ denote the Lispchitzness parameter of the family of loss functions $\{\ell(w;(\Phi x,y)\}_{(x,y)\in\overline{S}}$. From the analysis in Lemma 6, the average regret and excess empirical risk terms are both bounded by the following quantity with high probability.

$$O\left(\frac{\sqrt{H(\Phi,\overline{S})}B\|\mathcal{Y}\|}{\sqrt{n}} + \frac{\sqrt{k}G(\Phi,S)\log\left(1/\delta\right)}{n\epsilon}\right) \leq O\left(\frac{\sqrt{H}B\|\mathcal{Y}\|}{\sqrt{n}} + \frac{\sqrt{k}G\log\left(1/\delta\right)}{n\epsilon}\right).$$

This gives the following high probability bound,

$$\mathbb{E}[L(\widetilde{w};\Phi\mathcal{D}) - L(\Phi w^*;\Phi\mathcal{D})] \leq O\left(\frac{\sqrt{H}B\|\mathcal{X}\|\|\mathcal{Y}\|}{\sqrt{n}} + \frac{\left(\sqrt{H}\|\mathcal{X}\|\|\mathcal{Y}\| + H\|\mathcal{X}\|^2 B\right)B\sqrt{k\log\left(1/\delta\right)}}{n\epsilon}\right).$$

For the in-expectation bound, note that in the above proof the JL property was used for: with probability at least $1-\delta$, $\sup_{(x,y)\in\overline{S}}\|\Phi x\| \leq \left(1+\frac{\sqrt{\log(n/\delta)}}{k}\right)\|\mathcal{X}\|$ and $\|\Phi w^*\| \leq \left(1+\frac{\sqrt{\log(n/\delta)}}{k}\right)B$. All these quantities appear in the numerator in the above bound. Therefore the excess risk random variable w.r.t Φ has a tail with a poly $\left(\frac{\sqrt{\log(n/\delta)}}{k}\right)$ term. This is a (non-centered) sub-Weibull random variable and from equivalence of tail and moments bounds (e.g. Theorem 3.1 in [VGNA20]) and $\frac{\log(n)}{k} \leq O(1)$, we get the claimed expectation bound.

□

A.5 Constrained Regularized ERM with Output Perturbation

We here state a key result from [SST10]. Let $\mathfrak{R}_n(B,\|\mathcal{X}\|)$ denote the expected Rademacher complexity of linear predictors with norm bound by B, with n datapoints and norm of each point bounded by $\|\mathcal{X}\|$.

Theorem 9. *[[SST10]] Let ℓ be an H-smooth GLM and let R be a bound on the loss function. For a dataset S of n i.i.d samples, with probability at least $1-\beta$, for all w such that $\|w\|\leq B$, we have*

$$\begin{aligned}L(w;\mathcal{D}) \leq \widehat{L}(w;S) + O\Bigg(&\sqrt{\widehat{L}(w;S)}\left(\sqrt{H}\log^{1.5}(n)\mathfrak{R}_n(B,\|\mathcal{X}\|) + \sqrt{\frac{R\log\left(1/\beta\right)}{n}}\right) \\ &+ H\log^3(n)\mathfrak{R}_n^2(B,\|\mathcal{X}\|) + \frac{R\log\left(1/\beta\right)}{n}\Bigg).\end{aligned} \tag{9}$$

Corollary 3. *Let $\widetilde{w} = \arg\min_{w\in\mathcal{B}_B}\left\{\widehat{L}(w;S)\right\}$ and $n \geq \frac{HB^2\|\mathcal{X}\|^2}{\|\mathcal{Y}\|^2}$. Then with probability at least $1-\beta$ under the randomness of S we have*

$$L(\widetilde{w};\mathcal{D}) - \widehat{L}(\widetilde{w};S) = \widetilde{O}\left(\frac{\sqrt{H}B\|\mathcal{X}\|\|\mathcal{Y}\|\sqrt{\log\left(1/\beta\right)}}{\sqrt{n}} + \frac{\|\mathcal{Y}\|^2\log\left(1/\beta\right)}{\sqrt{n}}\right).$$

Further, in expectation under the randomness of S it holds that

$$\mathbb{E}[L(\widetilde{w};\mathcal{D}) - \widehat{L}(\widetilde{w};S)] = \widetilde{O}\left(\frac{\sqrt{H}\|\mathcal{Y}\|B\left\|\mathcal{X}\right\|}{\sqrt{n}} + \frac{\|\mathcal{Y}\|^2}{\sqrt{n}}\right).$$

Proof. In our application, w will be the output of regularized ERM $\widetilde{w}$, thus $\widehat{L}(\widetilde{w};S) \leq \widehat{L}_\lambda(0;S) \leq \|\mathcal{Y}\|^2$.

From Lemma 3 we have that $R \leq 3(\|\mathcal{Y}\|^2 + HB^2\|\mathcal{X}\|^2)$. Also, $\mathfrak{R}_n(B, \|\mathcal{X}\|) \leq O\left(\frac{B\|\mathcal{X}\|}{\sqrt{n}}\right)$ [SSBD14]. We now plug in the quantities into the above theorem to get that with probability at least $1-\beta$,

$$\begin{aligned}
&L(\widetilde{w};\mathcal{D}) - \widehat{L}(\widetilde{w};S) \\
&= \widetilde{O}\Bigg(\frac{\sqrt{H}B\|\mathcal{X}\|}{\sqrt{n}}\left(\|\mathcal{Y}\| + \frac{\sqrt{H}B\|\mathcal{X}\|}{\sqrt{n}}\right) \\
&+ \frac{\left(\|\mathcal{Y}\| + \sqrt{H}B\|\mathcal{X}\|\right)\sqrt{\log\left(1/\beta\right)}}{\sqrt{n}}\left(\|\mathcal{Y}\| + \frac{\left(\|\mathcal{Y}\| + \sqrt{H}B\|\mathcal{X}\|\right)\sqrt{\log\left(1/\beta\right)}}{\sqrt{n}}\right)\Bigg) \\
&= \widetilde{O}\left(\frac{\sqrt{H}B\|\mathcal{X}\|\|\mathcal{Y}\|\sqrt{\log\left(1/\beta\right)}}{\sqrt{n}} + \frac{\|\mathcal{Y}\|^2\log\left(1/\beta\right)}{\sqrt{n}}\right)
\end{aligned}$$

where the last follows by simplifications using the lower bound on n.

For the in expectation result, observe that the above Eqn (9) is a tail bound for a sub-Gamma random variable with variance parameter $O\left(\frac{\sqrt{H}B\|\mathcal{X}\|\|\mathcal{Y}\|}{\sqrt{n}}\right)$ and scale parameter $O\left(\frac{\|\mathcal{Y}\|^2}{\sqrt{n}}\right)$. From the equivalence of tail and moment bounds of sub-Gamma random variables, [BLM13] we get the claimed in-expectation bound. □

A.6 Proof of Theorem 3

Proof. Recall from Lemma 2 that the (un-regularized) loss is G-Lipschitz on the constraint set with $\|w\| \leq B$, where $\widetilde{G} := \left(\sqrt{H}\,\|\mathcal{Y}\| + HB\,\|\mathcal{X}\|\right)\|\mathcal{X}\|$. Note that from a standard analysis [BE02], we get that ℓ_2 sensitivity (or uniform argument stability) is $O\left(\frac{G}{n\lambda}\right)$, hence $\sigma^2 = O\left(\frac{G^2\log(1/\delta)}{\lambda^2 n^2\epsilon^2}\right)$ ensures (ϵ,δ)-DP.

For the utility analysis, the excess risk can be decomposed as follows,

$$\mathbb{E}[L(\widehat{w};\mathcal{D}) - L(w^*;\mathcal{D})] = \mathbb{E}[L(\widetilde{w};\mathcal{D}) - \widehat{L}(\widetilde{w};S)] + \mathbb{E}[\widehat{L}(\widetilde{w};S) - \widehat{L}(w^*;S)] + \mathbb{E}[L(\widetilde{w}+\xi;\mathcal{D}) - L(\widetilde{w};\mathcal{D})].$$

The first term is bounded by the Rademacher complexity result (Corollary 3).

The second term is simply bounded by $\frac{\lambda}{2}\|\widetilde{w}\|^2 \leq \frac{\lambda}{2}B^2$ since $\widetilde{w}$ lies in the constraint set. The third term is bounded using smoothness as follows:

$$\begin{aligned}
\mathbb{E}[L(\widetilde{w}+\xi;\mathcal{D}) - L(\widetilde{w};\mathcal{D})] &= \mathbb{E}\left[\phi_y\left(\langle\widetilde{w}+\xi, x\rangle\right) - \phi_y\left(\langle\widetilde{w}, x\rangle\right)\right] \\
&\leq \mathbb{E}\left[\phi_y'(\langle\widetilde{w}, x\rangle)\langle\xi, x\rangle + \frac{H}{2}\left|\langle\xi, x\rangle\right|^2\right] \\
&\leq \frac{H}{2}\sigma^2\|\mathcal{X}\|^2 = O\left(\frac{HG^2\|\mathcal{X}\|^2\log\left(1/\delta\right)}{\lambda^2 n^2\epsilon^2}\right)
\end{aligned}$$

where the last inequality follows since $\mathbb{E}\xi = 0$ and $\langle\xi, x\rangle \sim \mathcal{N}(0, \sigma^2\|x\|^2)$. We now plug in $\lambda = \left(\frac{G\sqrt{H}\|\mathcal{X}\|}{B}\right)^{2/3}\frac{(\log(1/\delta))^{1/3}}{(n\epsilon)^{2/3}}$. and $G = \left(\sqrt{H}\,\|\mathcal{Y}\| + HB\,\|\mathcal{X}\|\right)\|\mathcal{X}\|$ to get,

$$\mathbb{E}[L(\widehat{w};\mathcal{D}) - L(w^*;\mathcal{D})] \leq \widetilde{O}\left(\frac{\sqrt{H}B\,\|\mathcal{X}\|\,\|\mathcal{Y}\| + \|\mathcal{Y}\|^2}{\sqrt{n}} + \frac{\left(\sqrt{H}B\,\|\mathcal{X}\|\right)^{4/3}\|\mathcal{Y}\|^{2/3} + \left(\sqrt{H}B\,\|\mathcal{X}\|\right)^2}{(n\epsilon)^{2/3}}\right).$$

□

A.7 Proof of Theorem 4

Proof. Define $F(w;S) = \frac{1}{n}\sum_{(x,y)\in S}(\langle w, x\rangle - y)^2$. Let $d' < \min\{n,d\}$ and $b, p \in [0,1]$ be parameters to be chosen later. For any $\boldsymbol{\sigma} \in \{\pm 1\}^{d'}$, define the dataset $S_{\boldsymbol{\sigma}}$ which consists of the union of d' subdatasets, $S_1, ..., S_{d'}$ given as follows. Set $\frac{pn}{d'}$ of the feature vectors in S_j as $\|\mathcal{X}\|e_j$ (the rescaled j'th standard basis vector) and the rest as the zero vector. Set $\frac{pn}{2d}(1+b)$ of the labels as $\boldsymbol{\sigma}_j\|\mathcal{Y}\|$ and $\frac{pn}{2d}(1-b)$ labels as $-\boldsymbol{\sigma}_j\|\mathcal{Y}\|$. Let $w^{\boldsymbol{\sigma}} = \arg\min_{w\in\mathbb{R}^d}\{F(w;S_{\boldsymbol{\sigma}})\}$ be the ERM minimizer of $F(\cdot;S_{\boldsymbol{\sigma}})$. Following from Lemma 2 of [Sha15] we have that for any $\bar{w} \in \mathbb{R}^d$ that

$$F(\bar{w};S_{\boldsymbol{\sigma}}) - F(w^{\boldsymbol{\sigma}};S_{\boldsymbol{\sigma}}) \geq \frac{p\|\mathcal{X}\|^2}{2d'}\sum_{j=1}^{d'}(\bar{w}_j - w_j^{\boldsymbol{\sigma}})^2. \tag{10}$$

We will now show lower bounds on the per-coordinate error. Consider any $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ which differ only at index j for some $j \in [d']$. Note that the datasets $S_{\boldsymbol{\sigma}}$ and $S_{\boldsymbol{\sigma}'}$ differ in $\Delta = \frac{pn}{2d'}[(1+b)-(1-b)] = \frac{pbn}{d'}$ points. Let $\tau = w_j^{\boldsymbol{\sigma}} = \frac{\|\mathcal{Y}\|b}{\|\mathcal{X}\|}$ and $\tau' = w_j^{\boldsymbol{\sigma}'} = -\frac{\|\mathcal{Y}\|b}{\|\mathcal{X}\|}$ (i.e. the j components of the empirical minimizers for S and S'_j respectively). Note that $|w_j^{\boldsymbol{\sigma}} - w_j^{\boldsymbol{\sigma}'}| = \frac{2\|\mathcal{Y}\|b}{\|\mathcal{X}\|}$. We thus have by Lemma 1 that for a certain $b = b(\epsilon, n, d, B, p, \|\mathcal{Y}\|)$, $\mathcal{A}$ must satisfy

$$\mathbb{E}\left[|\mathcal{A}(S_{\boldsymbol{\sigma}})_j - w_j^{\boldsymbol{\sigma}}| + |\mathcal{A}(S_{\boldsymbol{\sigma}'})_j - w_j^{\boldsymbol{\sigma}'}|\right] \geq \frac{1}{4}\frac{\|\mathcal{Y}\|b}{\|\mathcal{X}\|}. \tag{11}$$

Since we need $\Delta \leq \frac{1}{\epsilon}$, we must set $b \leq \frac{d'}{pn\epsilon}$. Furthermore, if we are interested in problems with minimizer norm at most B, we need $b \leq \frac{\|\mathcal{X}\|B}{\|\mathcal{Y}\|\sqrt{d'}}$ to ensure the norm of the minimizer is bounded by B. Balancing these two restrictions, we set $d' = \left(\frac{p\|\mathcal{X}\|Bn\epsilon}{\|\mathcal{Y}\|}\right)^{2/3}$ which yields $b = \left(\frac{\|\mathcal{X}\|B}{\|\mathcal{Y}\|\sqrt{pn\epsilon}}\right)^{2/3}$. Assuming such settings of b and d' are possible (e.g. $b \in [0,1]$) we can apply Jensen's inequality and the fact that $(a+b)^2 \leq 2(a^2+b^2)$ to Eqn. (11) to obtain

$$\mathbb{E}\left[|\mathcal{A}(S_{\boldsymbol{\sigma}})_j - w_j^{\boldsymbol{\sigma}}|^2 + |\mathcal{A}(S_{\boldsymbol{\sigma}'})_j - w_j^{\boldsymbol{\sigma}'}|^2\right] \geq \frac{B^{4/3}\|\mathcal{Y}\|^{2/3}}{32(\|\mathcal{X}\|pn\epsilon)^{2/3}}.$$

We will now show this implies there exists a $\boldsymbol{\sigma} \in \{\pm 1\}^{d'}$ such that $F(\mathcal{A}(S_{\boldsymbol{\sigma}})) - F(w^{\boldsymbol{\sigma}}) = \Omega\left(\frac{(\|\mathcal{X}\|B)^{4/3}\|\mathcal{Y}\|^{2/3}p^{1/3}}{(n\epsilon)^{2/3}}\right)$. To prove this, we have the following analysis. Let $U = \{\pm 1\}^{d'}$ and let $\boldsymbol{\sigma}_{-j}$ denote the vector $\boldsymbol{\sigma}$ with its j'th component negated. We have

$$\begin{aligned}
\sup_{\boldsymbol{\sigma}\in U}\left\{\mathbb{E}\left[F(\mathcal{A}(S_{\boldsymbol{\sigma}})) - F(w^{\boldsymbol{\sigma}})\right]\right\} &\geq \frac{1}{|U|}\sum_{\boldsymbol{\sigma}\in U}\mathbb{E}\left[F(\mathcal{A}(S_{\boldsymbol{\sigma}})) - F(w^{\boldsymbol{\sigma}})\right] \\
&= \frac{p\|\mathcal{X}\|^2}{d'|U|}\sum_{j\in[d']}\sum_{\boldsymbol{\sigma}\in U}\mathbb{E}\left[|\mathcal{A}(S_{\boldsymbol{\sigma}})_j - w_j^{\boldsymbol{\sigma}}|^2\right] \\
&= \frac{p\|\mathcal{X}\|^2}{d'|U|}\sum_{j\in[d']}\sum_{\boldsymbol{\sigma}\in U:\boldsymbol{\sigma}_j=1}\mathbb{E}\left[|\mathcal{A}(S_{\boldsymbol{\sigma}})_j - w_j^{\boldsymbol{\sigma}}|^2 + |\mathcal{A}(S_{\boldsymbol{\sigma}_{-j}})_j - w_j^{\boldsymbol{\sigma}_{-j}}|^2\right] \\
&\geq \frac{(\|\mathcal{X}\|B)^{4/3}\|\mathcal{Y}\|^{2/3}p^{1/3}}{128(n\epsilon)^{2/3}}.
\end{aligned}$$

We recall this bound holds providing the settings of d' and b fall into the range $[1, \min\{n,d\}]$ and $[0,1]$ respectively. First note $b > 0$ always. Furthermore, $d' < \min\{n,d\}$ and $b < 1$ whenever

$$\frac{\|\mathcal{X}\|^2B^2}{\|\mathcal{Y}\|^2n\epsilon} \leq p \leq \min\left\{1, \frac{d^{3/2}\|\mathcal{Y}\|}{\|\mathcal{X}\|Bn\epsilon}\right\}. \tag{12}$$

In the following, assume $B \leq \frac{\|\mathcal{Y}\|\min\{\sqrt{n\epsilon},\sqrt{d}\}}{\|\mathcal{X}\|}$. Note this is no loss of generality as we can obtain problem instances with arbitrarily large B by adding a dummy point with $x = c'e_{d'+1}$ and

$y = \|\mathcal{Y}\|$ where c' is arbitrarily small so that the minimizer norm is B. Under this assumption on B, using the restrictions on p, it can be verified that $d' > 1$ whenever $\epsilon > \frac{1}{n}$. Thus we have $b \in [0,1]$ and $d' \in [1, \min\{n,d\}]$ as required. Furthermore this assumption on B implies $\frac{\|\mathcal{X}\|^2 B^2}{\|\mathcal{Y}\|^2 n\epsilon} \leq \min\left\{1, \frac{d^{3/2}\|\mathcal{Y}\|}{\|\mathcal{X}\| B n\epsilon}\right\}$ and thus a valid setting of p is possible. We now turn to setting p in a way which satisfies (12). We consider two cases, the high and low dimensional regimes.

Case 1: $d \geq \left(\frac{B\|\mathcal{X}\|n\epsilon}{\|\mathcal{Y}\|}\right)^{2/3}$. Setting $p = 1$ gives a lower bound of $\Omega\left(\min\left\{\|\mathcal{Y}\|^2, \frac{B^{4/3}\|\mathcal{X}\|^{4/3}\|\mathcal{Y}\|^{2/3}}{32(n\epsilon)^{2/3}}\right\}\right)$, where the min with $\|\mathcal{Y}\|^2$ from the upper bound on B.

Case 2: $d \leq \left(\frac{B\|\mathcal{X}\|n\epsilon}{\|\mathcal{Y}\|}\right)^{2/3}$. Setting $p = \frac{d^{3/2}\|\mathcal{Y}\|}{B\|\mathcal{X}\|n\epsilon}$ we obtain a bound of $\Omega\left(\min\left\{\|\mathcal{Y}\|^2, \frac{\sqrt{d}B\|\mathcal{X}\|\|\mathcal{Y}\|}{n\epsilon}\right\}\right)$ which we note is no larger than the bound from Case 1 in the low dimensional regime and no smaller than the bound from Case 1 in the high dimensional regime. Thus we can write the total bound as the minimum of these two bounds. The $\|\mathcal{Y}\|^2$ term again comes from the restriction on B.

To obtain results for arbitrary H, we can set $F(w;S) = \frac{H}{2n}\sum_{(x,y)\in S}(\langle w,x\rangle - \frac{2}{\sqrt{H}}y)^2$. This satisfies H-smoothness and loss bounded at zero by $\|\mathcal{Y}\|^2$. Then substituting $\|\mathcal{Y}\|$ in the previous expressions for $2\|\mathcal{Y}\|/\sqrt{H}$ and multiplying through by $H/2$ one obtains the claimed bound. □

B Missing Proofs from Section 4 (Lipschitz GLMs)

B.1 Proof of Theorem 5

Proof. Note that from a standard analysis [BE02], we get that ℓ_2 sensitivity of the regularized minimizer $\widetilde{w}$ is $\left(\frac{2G\|\mathcal{X}\|}{n\lambda}\right)$, hence $\sigma^2 = \frac{4G^2\|\mathcal{X}\|^2\log(1/\delta)}{\lambda^2 n^2\epsilon^2}$ ensures (ϵ,δ)-DP.

For the utility analysis, the excess risk can be decomposed as follows,

$$\mathbb{E}[L(\widehat{w};\mathcal{D}) - L(w^*;\mathcal{D})] = \mathbb{E}[L(\widetilde{w};\mathcal{D}) - \widehat{L}(\widetilde{w};S)] + \mathbb{E}[\widehat{L}(\widetilde{w};S) - \widehat{L}(w^*;S)] + \mathbb{E}[L(\widetilde{w}+\xi;\mathcal{D}) - L(\widetilde{w};\mathcal{D})]$$

The first term is bounded as $O\left(\frac{BG\|\mathcal{X}\|}{\sqrt{n}}\right)$ from Rademacher complexity results on bounded linear predictors [SSBD14].

The second term is simply bounded by $\frac{\lambda}{2}\|\widetilde{w}\|^2 \leq \frac{\lambda}{2}B^2$ since $\widetilde{w}$ is the regularized ERM. The third term is bounded via the following

$$\begin{aligned}
\mathbb{E}[L(\widetilde{w}+\xi;\mathcal{D}) - L(\widetilde{w};\mathcal{D})] &= \mathbb{E}\left[\phi_y\left(\langle \widetilde{w}+\xi, x\rangle\right) - \phi_y\left(\langle \widetilde{w}, x\rangle\right)\right] \\
&\leq \mathbb{E}\left[G\left|\langle \xi, x\rangle\right|\right] \\
&\leq G\sigma\|\mathcal{X}\| \leq \frac{4G^2\|\mathcal{X}\|^2\sqrt{\log(1/\delta)}}{\lambda n\epsilon}
\end{aligned}$$

where the last inequality follows since $\mathbb{E}\xi = 0$ and $\langle \xi, x\rangle \sim \mathcal{N}(0, \sigma^2\|x\|^2)$. Now plugging in $\lambda = \frac{G\|\mathcal{X}\|(\log(1/\delta))^{1/4}}{B\sqrt{n\epsilon}}$ obtains the following result

$$\mathbb{E}[L(\widehat{w};\mathcal{D}) - L(w^*;\mathcal{D})] = O\left(\frac{BG\|\mathcal{X}\|}{\sqrt{n}} + \frac{BG\|\mathcal{X}\|\log(1/\delta)^{1/4}}{\sqrt{n\epsilon}}\right).$$

□

B.2 Upper Bound using JL Method

Theorem 10. *Let $k = O(\log(2n/\delta)n\epsilon)$, $\sigma^2 = \frac{8TG^2\|\mathcal{X}\|^2\log(2/\delta)}{n^2\epsilon^2}$, $\eta = \frac{B}{G\|\mathcal{X}\|\left(1+\frac{\sqrt{k\log(2/\delta)}}{n\epsilon}\right)T^{3/4}}$ and $T = n^2$. Algorithm 2 satisfies (ϵ,δ)-differential privacy. Given a dataset S of n i.i.d samples,* *the excess risk of its output $\widetilde{w}$ is bounded as*

$$\mathbb{E}[\varepsilon_{\mathsf{risk}}(\widehat{w})] = \widetilde{O}\left(\frac{GB\left\|\mathcal{X}\right\|}{\sqrt{n}} + \frac{GB\|\mathcal{X}\|}{\sqrt{n\epsilon}}\right).$$

Proof. Let $\alpha \leq 1$ be a parameter to be set later. From the JL property, with $k = O\left(\frac{\log(2n/\delta)}{\alpha^2}\right)$, with probability at least $1 - \frac{\delta}{2}$, for feature vectors have $\left\|\Phi x_i\right\| \leq (1+\alpha)\left\|x_i\right\| \leq 2\left\|\mathcal{X}\right\|$, and $\left\|\Phi w^*\right\|^2 \leq 2\left\|w^*\right\|^2 \leq 2B^2$. Thus, gradient of loss for data point (x,y) in S at any w is bounded as $\left\|\ell(w;(\Phi x, y)\right\| = \phi_y'(\langle w, \Phi x\rangle)\left\|\Phi x\right\| \leq 2G\left\|\mathcal{X}\right\|$. The privacy guarantee thus follows from the privacy of DP-SCO and post-processing.

For the utility guarantee, we decompose excess risk as:

$$\mathbb{E}\left[L(\widehat{w};\mathcal{D}) - L(w^*;\mathcal{D})\right] = \mathbb{E}\left[L(\widetilde{w};\Phi\mathcal{D}) - L(\Phi w^*;\Phi S)\right] + \mathbb{E}\left[L(\Phi w^*;\Phi S) - L(w^*;\mathcal{D})\right]. \quad (13)$$

The first term in Eqn. (13) is bounded by the utility guarantee of the DP-SCO method. In particular, from Lemma 8 (below), we have

$$\mathbb{E}\left[L(\widetilde{w};\Phi\mathcal{D}) - L(\Phi w^*;\Phi S)\right] \leq \frac{GB\left\|\mathcal{X}\right\|}{\sqrt{n}} + O\left(\frac{G\left\|\mathcal{X}\right\|B\sqrt{k}}{n\epsilon}\right).$$

For the second term in Eqn (13), we use JL-property and Lipschitzness of GLM:

$$\begin{aligned}\mathbb{E}\left[L(\Phi w^*;\Phi S) - L(w^*;\mathcal{D})\right] &\leq G\mathbb{E}\left[\left|\langle \Phi x, \Phi w^*\rangle - \langle x, w^*\rangle\right|\right] \\ &\leq \alpha G\left\|\mathcal{X}\right\|\left\|w^*\right\| \\ &= O\left(\frac{G\left\|\mathcal{X}\right\|B\sqrt{\log\left(2n/\delta\right)}}{\sqrt{k}}\right).\end{aligned}$$

Combining, we get,

$$\mathbb{E}\left[L(\widehat{w};\mathcal{D}) - L(w^*;\mathcal{D})\right] \leq \widetilde{O}\left(\frac{G\left\|\mathcal{X}\right\|B}{\sqrt{n}} + \frac{G\left\|\mathcal{X}\right\|B\sqrt{k}}{n\epsilon} + \frac{G\left\|\mathcal{X}\right\|B}{\sqrt{k}}\right).$$

Balancing parameters by setting $k = \widetilde{O}(n\epsilon)$ gives the claimed bound. □

Lemma 8. *Let $\Phi \in \mathbb{R}^{d\times k}$ be a data-oblivious JL matrix. Let $S = \{(x_i, y_i)\}_{i=1}^n$ of n i.i.d data points and let $\Phi S := \{(\Phi x_i, y_i)\}_{i=1}^n$. Let $\widetilde{w} \in \mathbb{R}^k$ be the average iterate returned noisy SGD procedure with Gaussian noise variance $\sigma^2 = O\left(\frac{G^2\left\|\mathcal{X}\right\|^2\log(1/\delta)}{n^2\epsilon^2}\right)$ on dataset ΦS. For $k = \Omega\left(\log\left(n\right)\right)$, the excess risk of $\widetilde{w}$ on $\Phi\mathcal{D}$ is bounded as,*

$$\mathbb{E}_{\Phi,S}[L(\widetilde{w};\Phi\mathcal{D}) - L(\Phi w^*;\Phi\mathcal{D})] \leq O\left(\frac{GB\|\mathcal{X}\|}{\sqrt{n}} + \frac{GB\|\mathcal{X}\|\sqrt{k\log\left(1/\delta\right)}}{n\epsilon}\right). \quad (14)$$

Proof. The proof uses the analysis for excess risk bound for Noisy SGD in [BFGT20] with the JL transform. Let $S^{(i)}$ be the dataset where the i-th data point is replaced by an i.i.d. point (x', y') and let $\widetilde{w}^{(i)}$ be the corresponding output of Noisy-SGD on $\Phi S^{(i)}$. Define $\overline{S} := \{S, (x', y')\}$ and let $G(\Phi, \overline{S})$ denote an upper bound Lipschitzness parameter of the family of loss functions $\{\ell(w;(\Phi x; y))\}_{(x,y)\in\overline{S}}$.

We decompose the excess risk as:

$$\mathbb{E}_{\Phi,S}[L(\widetilde{w};\Phi\mathcal{D}) - L(\Phi w^*;\Phi\mathcal{D})] = \mathbb{E}_{\Phi,S}[L(\Phi^\top\widetilde{w};\mathcal{D}) - \widehat{L}(\Phi^\top\widetilde{w};S)] + \mathbb{E}_{\Phi,S}[\widehat{L}(\Phi^\top\widetilde{w};S)] - \widehat{L}(\Phi w^*;\Phi S)].$$

A well-known fact (see [SSBD14]) is that generalization gap is equal to on-average-stability:

$$\begin{aligned}\mathbb{E}_{\Phi,S}[L(\Phi^\top \widetilde{w};\mathcal{D}) - \widehat{L}(\Phi^\top \widetilde{w};S)] &= \mathbb{E}_{\overline{S},\Phi}[\ell(\Phi^\top \widetilde{w}^{(i)};(x_i,y_i)) - \ell(\Phi^\top \widetilde{w};(x_i,y_i))]\\ &= \mathbb{E}_{\overline{S},\Phi}[\ell(\widetilde{w}^{(i)};(\Phi x_i,y_i)) - \ell(\widetilde{w};(\Phi x_i,y_i))].\end{aligned}$$

From the analysis in Theorem 3.3 from [BFGT20], it follows that

$$\ell(\widetilde{w}^{(i)};(\Phi x_i,y_i)) - \ell(\widetilde{w};(\Phi x_i,y_i)) \leq G(\Phi;\overline{S})\left\|\widetilde{w}^{(i)} - \widetilde{w}\right\| \leq O\left(G(\Phi;\overline{S})^2\left(\eta\sqrt{T} + \frac{\eta T}{n}\right)\right). \tag{15}$$

where the last equality follows from the GLM structure of the loss function.

From analysis of Noisy-SGD [BST14], the other term (excess empirical risk) can be bounded as

$$\widehat{L}(\widetilde{w};\Phi S)] - \widehat{L}(\Phi w^*;\Phi S) \leq O\left(\eta\left(G(\Phi;\overline{S})^2 + \frac{kG(\Phi;\overline{S})^2\log(1/\delta)}{n^2\epsilon^2}\right) + \frac{\|\Phi w^*\|^2}{T\eta}\right) \tag{16}$$

We now take expectation with respect to Φ. Note that $G(\Phi,\overline{S}) = \sup_{(x,y)\in\overline{S}}\sup_w |g_x(\langle w,\Phi x\rangle)|\,\|\Phi x\|$, where g_x is the an element of the sub-differential of ϕ_x at $\langle w,x\rangle$. By the Lipschitzness assumption $|g_x(\langle w,\Phi x\rangle)| \leq G$ and from the JL property, with probability at least $1-\delta$, $\sup_{(x,y)\in\overline{S}}\|\Phi x\| \leq \left(1+\frac{\sqrt{\log(n/\delta)}}{k}\right)\|\mathcal{X}\|$. Similarly, $\left\|\Phi^\top w^*\right\| \leq \left(1+\frac{\sqrt{\log(n/\delta)}}{k}\right)B$. Observe that the tail is that of a (non-centered) sub-Gaussian random variable with variance parameter $O\left(\frac{1}{k}\right)$. Using the equivalence of tail and moment bounds of sub-Gaussian random variables [BLM13] and the fact that $\frac{\log(n)}{k} \leq O(1)$, we get the following bounds for Eqn. (15) and (16):

$$\mathbb{E}_\Phi\left[\ell(\widetilde{w}^{(i)};(\Phi x_i,y_i)) - \ell(\widetilde{w};(\Phi x_i,y_i)\right] \leq O\left(G^2\|\mathcal{X}\|^2\left(\eta\sqrt{T}+\frac{\eta T}{n}\right)\right)$$

$$\mathbb{E}_\Phi\left[\widehat{L}(\widetilde{w};\Phi S)] - \widehat{L}(\Phi w^*;\Phi S)\right] \leq \widetilde{O}\left(\eta\left(G^2\|\mathcal{X}\|^2 + \frac{kG^2\|\mathcal{X}\|^2\log(1/\delta)}{n^2\epsilon^2}\right) + \frac{B^2}{T\eta}\right).$$

Finally, as in [BFGT20], taking expectation with respect $\overline{S}$ in the two inequalities above, setting $\eta = \frac{B}{G\|\mathcal{X}\|\left(1+\frac{\sqrt{k\log(1/\delta)}}{n\epsilon}\right)T^{3/4}}$ and $T = n^2$ and combining gives the claimed bound. □

B.3 Proof of Theorem 6

The proof follows from the more general Theorem 11 stated below. Instantiating Theorem 11 with $p = q = 2$ satisfies all the requirements of our Theorem 6. Finally, let w^* be the population risk minimizer of the hard instance in Theorem 11. We then have,

$$\begin{aligned}\mathbb{E}[L(\mathcal{A}(S);\mathcal{D}) - \min_w L(w;\mathcal{D})] &= \mathbb{E}[L(\mathcal{A}(S);\mathcal{D}) - L(w^*;\mathcal{D})]\\ &\geq \mathbb{E}[L(\mathcal{A}(S);\mathcal{D}) - \min_{w:\|w\|_2\leq B} L(w;\mathcal{D})]\\ &\geq \mathbb{E}[L(\widehat{w};\mathcal{D}) - L(\widetilde{\mathrm{w}};\mathcal{D})]\\ &= \Omega\left(GB\|\mathcal{X}\|\min\left(1,\frac{1}{\sqrt{n\epsilon}},\frac{\sqrt{\text{rank}}}{n\epsilon}\right)\right).\end{aligned}$$

This completes the proof.

B.4 Lower bound for Non-Euclidean DP-GLM

Theorem 11. *Let $G, \|\mathcal{Y}\|, \|\mathcal{X}\|, B > 0$, $\epsilon \leq 1.2, \delta \leq \epsilon$ and $p, q \geq 1$. For any (ϵ, δ)-DP algorithm $\mathcal{A}$, there exists sets $\mathcal{X}$ and $\mathcal{Y}$ such that for any $x \in \mathcal{X}$, $\|x\|_q \leq 1$, a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, a G-Lipschitz GLM loss bounded at zero by $\|\mathcal{Y}\|$ and a $\widetilde{w}$ with $\|\widetilde{w}\|_p \leq B$ such that the output of $\mathcal{A}$ on $S \sim \mathcal{D}^n$ satisfies*

$$\mathbb{E}[L(\mathcal{A}(S);\mathcal{D}) - L(\widetilde{w};\mathcal{D})] = \Omega\left(GB\|\mathcal{X}\| \min\left(1, \frac{1}{(n\epsilon)^{1/p}}, \frac{(rank)^{(p-1)/p}}{n\epsilon}\right)\right).$$

Proof. The construction below is from [SSTT20] with some changes. We provide the complete proof below. Consider the following 1-Lipschitz loss function:

$$\ell(w;(x,y)) = |y - \langle w, x\rangle|.$$

Firstly, we argue that we can assume $\|\mathcal{Y}\| = \infty$. This is because for any arbitrarily large value of y we can translate the above function below so that the the loss is bounded by $\|\mathcal{Y}\|$. However, this translation doesn't change the excess risk. Also, as in [BST14], it suffices to consider $G = 1$, since we can simply scale the 1-Lipschitz loss function and get a factor of the G in the lower bound. Let $d' \leq \min(\text{rank}, n\epsilon)$, $0 \leq \alpha \leq 1$ and $\beta > 0$ be parameters to be set later. Since $d' \leq \text{rank}$, without loss of generality, we will represent the features x as d' dimensional vectors.

Consider a distribution $\mathcal{D}$, where $x = e_0 := \vec{0}$ with probability $1-\alpha$, and with probability α, $x \sim \text{Unif}(\{\|\mathcal{X}\| e_i\}_{i=1}^{d'})$. Note that $\|x\|_q \leq \|\mathcal{X}\|$ and for any $q \geq 1$ for $x \in \text{supp}(\mathcal{D}_x)$ where $\mathcal{D}_x$ denotes the marginal of $\mathcal{D}$ w.r.t. the x variable. This implies the loss $w \mapsto \ell(w;(x,y))$ is $\|\mathcal{X}\|$-Lipschitz in ℓ_q-norm, when $x \in \text{supp}(\mathcal{D}_x)$ and for all y. Let the fingerprinting code $z \in \{0,1\}^{d'}$ be drawn from a product distribution with mean $\mu \in [0,1]^{d'}$ where each co-ordinate $\mu_i \sim \text{Beta}(\beta,\beta)$. Finally we have $y = \frac{B}{(d')^{1/p}}\langle x, z\rangle$. Let $S = \{(x_i,y_i)\}_{i=1}^n$ be n i.i.d. samples drawn from $\mathcal{D}$. Define $\widetilde{w} = \frac{B}{(d')^{1/p}}\mu$; note that $\|\widetilde{w}\|_p \leq B$.

Let $\mathcal{A}$ be any (ϵ,δ)-DP algorithm, which given S outputs $\widehat{w}$. Its excess risk with respect to $\widetilde{w}$ can be lower bounded as,

$$\begin{aligned}\mathbb{E}[L(\widehat{w};\mathcal{D}) - L(\widetilde{w};\mathcal{D})] &= \mathbb{E}[|y - \langle \widehat{w}, x\rangle| - |y - \langle \widetilde{w}, x\rangle|] \\ &\geq \mathbb{E}[|\langle \widehat{w} - \widetilde{w}, x\rangle| - 2|y - \langle \widetilde{w}, x\rangle|]. \end{aligned} \tag{17}$$

The last term above is upper bounded as:

$$\mathbb{E}[|y - \langle \widetilde{w}, x\rangle|] = \frac{B}{(d')^{1/p}}\mathbb{E}[|\langle z, x\rangle - \langle \mu, x\rangle|] = \frac{B\|\mathcal{X}\|}{(d')^{1/p}}\frac{\alpha}{d'}\sum_{i=1}^{d'}\mathbb{E}|z_i - \mu_i|.$$

By direct computation,

$$\mathbb{E}|z_i - \mu_i| = \mathbb{E}[|1-\mu_i|\mu_i + |-\mu_i|(1-\mu_i)] = 2\mathbb{E}\mu_i(1-\mu_i) = \frac{2\beta}{1+2\beta}.$$

This gives us that

$$\mathbb{E}[|y - \langle \widetilde{w}, x\rangle|] \leq \frac{2\alpha\beta B\|\mathcal{X}\|}{(1+2\beta)(d')^{(p+1)/p}}. \tag{18}$$

The first term in the right hand side Inequality (17) is,

$$\mathbb{E}[|\langle \widehat{w} - \widetilde{\mu}, x\rangle| = \frac{B}{(d')^{1/p}}\mathbb{E}\left|\left\langle \frac{(d')^{1/p}}{B}\widehat{w} - \mu, x\right\rangle\right| = \frac{B\|\mathcal{X}\|\alpha}{(d')^{(p+1)/p}}\mathbb{E}\sum_{j=1}^{d'}\left|\frac{(d')^{1/p}}{B}\widehat{w}_j - \mu_j\right|. \tag{19}$$

Define $v \in \mathbb{R}^{d'}$ with $v_i = \frac{(d')^{1/p}\widehat{w}_i}{B}$. Note that,

$$\sum_{j=1}^{d'}\left|\frac{(d')^{1/p}}{B}\widehat{w}_j - \mu_j\right| = \|v - \mu\|_1 \geq \|v - \mu\|_2 \geq \|\widehat{v} - \mu\|_2 \geq \|\widehat{v} - \mu\|_2^2$$

where the first inequality above follows from relationship between ℓ_1 and ℓ_2 norm, the second follows from projection property and $\widehat{v}$ denotes the projection of v onto $[0,1]^d$, and the last inequality follows from boundedness of coordinates of $\widehat{v}-\mu$.

The key step now is application of the fingerprinting lemma (simplified below, see Lemma B.1 in [SSTT20] for complete statement) from [SU17] which roughly speaking, "relates the error to correlation"; for any $\widehat{v} \in [0,1]^{d'}$, we have

$$\mathbb{E}\left\|\widehat{v}-\mu\right\|_2^2 \geq \frac{d'}{4\left(1+2\beta\right)} - \frac{1}{\beta}\sum_{i=1}^{n}\mathbb{E}\left\langle \widehat{v}, z_i-\mu\right\rangle . \tag{20}$$

As in [SSTT20], the correlation simplifies as

$$\mathbb{E}\left\langle \widehat{v}, z_i-\mu\right\rangle = \sum_{j=0}^{d'}\mathbb{E}[\left\langle \widehat{v}, z_i-\mu\right\rangle |x_i=e_j]\mathbb{P}[x_i=\|\mathcal{X}\|e_j] = \frac{\alpha}{d'}\mathbb{E}[\widehat{v}_j\left(z_i-\mu\right)_j|x_i=\|\mathcal{X}\|e_j]$$

where the last equality follows since $\widehat{v}$ only depends on the coordinate of z_i for which $x_i=1$. Now, let $\widehat{v}^{\sim i}$ denotes the solution obtained if z_i were replaced by an independent sample. From boundedness, $\left\|\widehat{v}_j\left(z_i-\mu\right)_j\right\|_\infty \leq 1$. Using differential privacy, we have,

$$\mathbb{E}[\widehat{v}_j\left(z_i-\mu\right)_j|x_i=\|\mathcal{X}\|e_j] \leq \left(e^\epsilon-1\right)\mathbb{E}\left[\left|\widehat{v}_j^{\sim i}\left(z_i-\mu\right)_j\right||x_i=\|\mathcal{X}\|e_j\right]+\delta \leq \left(e^\epsilon-1\right)+\delta \leq 3\epsilon$$

where the last inequality uses the assumption that $\epsilon \leq 1.2$ and $\delta \leq \epsilon$. Plugging this in Eqn. (20), we get

$$\mathbb{E}\left\|\widehat{v}-\mu\right\|_2^2 \geq \frac{d'}{4\left(1+2\beta\right)} - 3\alpha n\epsilon.$$

Plugging the above in Eqn. (19), we get

$$\mathbb{E}[\left|\left\langle \widehat{w}-\widetilde{w}, x\right\rangle\right| \geq \frac{B\left\|\mathcal{X}\right\|\alpha}{d'^{(p+1)/p}}\left(\frac{d'}{4\left(1+2\beta\right)} - 3\alpha n\epsilon\right).$$

Using the above, and the bound in Eqn. (17), we get that,

$$\begin{aligned}\mathbb{E}[L(\widehat{w};\mathcal{D})-L(\widetilde{w};\mathcal{D})] &\geq \frac{B\left\|\mathcal{X}\right\|\alpha}{d'^{(p+1)/p}}\left(\frac{d'}{4\left(1+2\beta\right)} - 3\alpha n\epsilon\right) - \frac{2\alpha\beta B\left\|\mathcal{X}\right\|}{\left(1+2\beta\right)\left(d'\right)^{1/p}} \\ &= \frac{B\left\|\mathcal{X}\right\|\alpha}{\left(1+2\beta\right)\left(d'\right)^{1/p}}\left(\frac{1}{4} - \frac{3\alpha\left(1+2\beta\right)n\epsilon}{d'} - 2\beta\right).\end{aligned}$$

Now, we set $\beta=\frac{1}{16}$. When rank $\leq 48n\epsilon$ we set $d'=$ rank and $\alpha = \min\left(\frac{d'}{48(1+2\beta)n\epsilon}, 1\right)$. The minimum term becomes $\frac{d'}{48(1+2\beta)n\epsilon}$ and obtain an excess risk lower bound of,

$$\mathbb{E}[L(\widehat{w};\mathcal{D})-L(\widetilde{w};\mathcal{D})] \geq \frac{B\left\|\mathcal{X}\right\|\left(d'\right)^{(p-1)/p}}{1024n\epsilon}.$$

On the other hand, when rank $> 48n\epsilon$, set $\alpha=1$ and $d'=\lfloor 48n\epsilon\rfloor$ (so that it is at least 1) to get an excess risk lower bound of,

$$\mathbb{E}[L(\widehat{w};\mathcal{D})-L(\widetilde{w};\mathcal{D})] \geq \min\left(\frac{B\left\|\mathcal{X}\right\|}{16\left(n\epsilon\right)^{1/p}}, B\left\|\mathcal{X}\right\|\right).$$

Finally, note that we can arbitrarily increase the rank of the construction beyond $48n\epsilon$ by adding datapoints with orthogonal feature vectors of small enough magnitude and arbitrary labels. Combining the two lower bounds obtains the claimed bound. □

Corollary 4. *Let $G, B > 0$, $\epsilon \leq 1.2, \delta \leq \epsilon$ and $p, q \geq 1$. Let $\mathcal{W} \subset \mathbb{R}^d$ such that for any $w \in \mathcal{W}$, $\|w\|_p \leq B$. For any (ϵ, δ)-DP algorithm $\mathcal{A}$, there exists a set $\mathcal{Z}$, a distribution $\mathcal{D}$ over $\mathcal{Z}$ and a loss function $w \mapsto \ell(w; z)$, which is convex, G-Lipschitz w.r.t. ℓ_q norm for $w \in \mathcal{W}$, for all $z \in \mathcal{Z}$ such that the output of $\mathcal{A}$ on $S \sim \mathcal{D}^n$ (which may not lie in $\mathcal{W}$), satisfies*

$$\mathbb{E}_{\mathcal{A},S}[\mathbb{E}_{z\sim\mathcal{D}}\ell(\mathcal{A}(S);z) - \min_{w\in\mathcal{W}} \mathbb{E}_{z\sim\mathcal{D}}\ell(w;z)] = \Omega\left(GB\min\left(1, \frac{1}{(n\epsilon)^{1/p}}, \frac{d^{(p-1)/p}}{n\epsilon}\right)\right).$$

Using generalization properties of differential privacy, we get the same bound as above for excess empirical risk; see Corollary B.4 in [SSTT20] for details.

C Missing Details for Section 5 (Adapting to $\|w^*\|$)

C.1 Generalized Exponential Mechanism

Theorem 12. *[RS15] Let $K > 0$ and $S \in \mathcal{Z}^n$. Let $q_1, ..., q_K$ be functions s.t. for any adjacent datasets S, S' it holds that $|q_j(S) - q_j(S')| \leq \gamma_j : \forall j \in [K]$. There exists an Algorithm,* GenExpMech, *such that when given sensitivity-score pairs $(\gamma_1, q_1(S)), ..., (\gamma_N, q_N(S))$, privacy parameter $\epsilon > 0$ and confidence parameter $\beta > 0$, outputs $j \in [N]$ such that with probability at least $1 - \beta$ satisfies $q_j(S) \leq \min\limits_{j\in[N]}\left\{q_j(S) + \frac{4\gamma_j \log(N/\beta)}{\epsilon}\right\}$.*

C.2 Proof of Theorem 7

Note that by assumptions on $\mathcal{A}$, the process of generating $w_1, ..., w_K$ is $(\epsilon/2, \delta/2)-DP$. Furthermore, by Assumption 2 with probability at least $\delta/2$ the sensitivity values passed to GenExpMech bound sensitivity. Thus by the privacy guarantees of GenExpMech and composition we have that the entire algorithm is (ϵ, δ)-DP.

We now prove accuracy. In order to do so, we first prove that with high probability every $\widetilde{L}_j$ is an upper bound on the true population loss of w_j. Specifically, define $\tau_j = \frac{\Delta(B_j)\log(4K/\beta)}{n} + \sqrt{\frac{4\|\mathcal{Y}\|^2\log(4K/\beta)}{n}}$ (i.e. the term added to each $L(w_j; S_2)$ in Algorithm 4). Note it suffices to prove

$$\mathbb{P}\left[\exists j \in [K] : |L(w_j; S_2) - L(w_j; \mathcal{D})| \geq \tau_j\right] \leq \frac{\beta}{2}. \tag{21}$$

Fix some $j \in [K]$. Note that the non-negativity of the loss implies that $\ell(w^*_B; (x,y)^2) \geq 0$. The excess risk assumption then implies that $\mathop{\mathbb{E}}\limits_{(x,y)\sim\mathcal{D}}\left[\ell(w_j; (x,y))^2\right] \leq 4\|\mathcal{Y}\|^2$, which in turn bounds the variance. Further, with probability at least $1 - \frac{\beta}{4K}$ it holds that for all $(x,y) \in S_2$ that $\ell(w, (x,y)) \leq \Delta_0 + \Delta(B)$. Thus by Bernstein's inequality we have

$$\mathbb{P}\left[|L(w; S_2) - L(w; \mathcal{D})| \geq t\right] \leq \exp\left(-\frac{t^2n^2}{\Delta(B_j)tn + 4n\|\mathcal{Y}\|^2}\right) + \frac{\beta}{4K}$$

Thus it suffices to set $t = \frac{\Delta(B_j)\log(4K/\beta)}{n} + \sqrt{\frac{4\|\mathcal{Y}\|^2\log(4K/\beta)}{n}}$ to ensure $\mathbb{P}\left[|L(w; S_2) - L(w; \mathcal{D})| \geq t\right] \leq \frac{\beta}{2K}$. Taking a union bound over all $j \in K$ establishes (21). We now condition on this event for the rest of the proof.

Now consider the case where $j^* \neq 0$ and $\|w^*\| \leq 2^K$. Note that the unconstrained minimizer w^* is the constrained minimizer with respect to any $\mathcal{B}_r$ for $r \geq \|w^*\|$. With this in mind, let $j' = \min\limits_{j\in[K]}\{j : w^* \in \mathcal{B}_{2^j}\}$ (i.e. the index of the smallest ball containing w^*). In the following we condition on the event that $\forall j \in [K], j \geq j'$, the parameter vector w_j satisfies excess population risk at most ERR(2^j). We note by Assumption 2 that this (in addition to the event given in (21)) happens with probability at least $1 - \frac{3\beta}{4}$. By the guarantees of GenExpMech, with probability at least $1 - \beta$

we (additionally) have

$$\begin{aligned} L(w_{j^*};\mathcal{D}) \leq L(w_{j^*};S_2) + \tau_{j^*} &\leq \min_{j\in[K]}\left\{L(w_j;S_2) + \tau_j + \frac{4\Delta(B_j)\log\left(4K/\beta\right)}{n\epsilon}\right\} \\ &\leq L(w_{j'};S_2) + \tau_{j'} + \frac{4\Delta(B_{j'})\log\left(4K/\beta\right)}{n\epsilon} \\ &\leq L(w_{j'};\mathcal{D}) + 2\tau_{j'} + \frac{4\Delta(B_{j'})\log\left(4K/\beta\right)}{n\epsilon}. \end{aligned}$$

Since $2^{j'} \leq \max\left\{2\left\|w^*\right\|, 1\right\}$ we have

$$\begin{aligned} L(w_{j^*};\mathcal{D}) - L(w^*;\mathcal{D}) &\leq L(w_{j'};\mathcal{D}) - L(w^*;\mathcal{D}) + 2\tau_{j'} + \frac{4\Delta(B_{j'})\log\left(4K/\beta\right)}{n\epsilon} \\ &\leq \mathsf{ERR}(2\max\left\{\left\|w^*\right\|, 1\right\}) + 2\tau_{j'} + \frac{4\Delta(\max\left\{2\left\|w^*\right\|, 1\right\})\log\left(4K/\beta\right)}{n\epsilon} \\ &\leq \mathsf{ERR}(2\max\left\{\left\|w^*\right\|, 1\right\}) + \sqrt{\frac{4\|\mathcal{Y}\|^2\log\left(4K/\beta\right)}{n}} + \frac{5\Delta(\max\left\{2\left\|w^*\right\|, 1\right\})\log\left(4K/\beta\right)}{n\epsilon} \end{aligned}$$

where the second inequality comes from the fact the assumption that $\|w^*\| \leq \|\mathcal{Y}\|^2$. Now note that by the assumption that $\mathsf{ERR}(2^K) \geq \|\mathcal{Y}\|^2$, whenever $\|w^*\| \geq 2^K$ it holds that $\|\mathcal{Y}\|^2 \leq \mathsf{ERR}(\|w^*\|)$. However since the sensitivity-score pair $(0, \|\mathcal{Y}\|^2)$ is passed to GenExpMech, the excess risk of the output is bounded by at most $\|\mathcal{Y}\|^2$ by the guarantees of GenExpMech).

C.3 Proof of Theorem 8

Let $\widehat{w}$ denote the output of the regularized output perturbation method with boosting and noise and privacy parameters $\epsilon' = \frac{\epsilon}{K}$ and $\delta' = \frac{\delta}{K}$. We have by Theorem 15 that with probability at least $1 - \frac{\beta}{4K}$ that

$$\begin{aligned} L(\widehat{w};\mathcal{D}) - L(w^*;\mathcal{D}) = \widetilde{O}\Bigg(&\frac{\sqrt{H}B\left\|\mathcal{X}\right\|\left\|\mathcal{Y}\right\| + \left\|\mathcal{Y}\right\|^2}{\sqrt{n}} + \frac{\left(\left(\sqrt{H}B\left\|\mathcal{X}\right\|\right)^{4/3}\left\|\mathcal{Y}\right\|^{2/3} + \left(\sqrt{H}B\left\|\mathcal{X}\right\|\right)^2\right)}{(n\epsilon)^{2/3}} \\ &+ \frac{\left(\|\mathcal{Y}\|^2 + HB^2\|\mathcal{X}\|^2\right)}{n\epsilon} + \frac{\left(\|\mathcal{Y}\| + \sqrt{H}B\|\mathcal{X}\|\right)}{\sqrt{n}}\Bigg). \end{aligned}$$

Note that this is no smaller than $\|\mathcal{Y}\|^2$ when $B = \Omega\left(\max\left\{\frac{\|\mathcal{Y}\|\sqrt{n\epsilon}}{\|\mathcal{X}\|\sqrt{H}}, \frac{\|\mathcal{Y}\|^2(n\epsilon)^{2/3}}{\sqrt{H}\|\mathcal{X}\|^2}\right\}\right)$, and thus it suffices to set $K = \Theta\left(\log\left(\max\left\{\frac{\|\mathcal{Y}\|\sqrt{n}}{\|\mathcal{X}\|\sqrt{H}}, \frac{\|\mathcal{Y}\|^2(n\epsilon)^{2/3}}{\sqrt{H}\|\mathcal{X}\|^2}\right\}\right)\right)$ to satisfy the condition of the Theorem statement.

Let σ_j denote the level noise used for when the guess for $\|w^*\|$ is B_j. To establish Assumption 2, by Lemma 10 we have that this assumption is satisfies with $\Delta(B) = \|\mathcal{Y}\|^2 + H\|\mathcal{X}\|^2\sigma_j^2\log\left(K/\min\left\{\beta,\delta\right\}\right)) + HB^2\|\mathcal{X}\|^2$. In particular, we note for the setting of σ_j implied by Theorem 15 and the setting of K we have for all $j \in [K]$ that $H\|\mathcal{X}\|^2\sigma_j^2 = \widetilde{O}(\|\mathcal{Y}\|^2)$. Thus $\Delta(B) = \widetilde{O}\left(\|\mathcal{Y}\|^2 + HB^2\|\mathcal{X}\|^2\right)$. The result then follows from Theorem 7.

C.4 Stability Results for Assumption 2

Lemma 9. *Algorithm 1 run with constraint set $\mathcal{B}_B$ satisfies Assumption 2 with $\Delta(B) = \|\mathcal{Y}\|^2 + HB^2\|\mathcal{X}\|^2$.*

The proof is straightforward using Lemma 3 (provided in the Appendix). For the output perturbation method, we can obtain similar guarantees. Here however, we must account for the fact that the output may not lie in the constraint set. We also remark that the JL-based method (Algorithm 2) can also enjoy this same bound. However, in this case one must apply the norm adaptation method to the intermediate vector $\widetilde{w}$, as $\Phi^\top\widetilde{w}$ may have large norm.

Lemma 10. *Algorithm 3 run with parameter B and σ satisfies Assumption 2 with $\Delta(B) = \|\mathcal{Y}\|^2 + H\|\mathcal{X}\|^2\sigma^2\log{(K/\delta)}) + HB^2\|\mathcal{X}\|^2$*

Proof. Note that since S and S' differ in only one point, it suffices to show that for any (x,y), (x',y') that $\ell(\widehat{w};(x,y)) \leq \|\mathcal{Y}\|^2 + HB^2\|\mathcal{X}\|^2 + H\|\mathcal{X}\|^2\sigma^2\log{(K/\delta)}$ and similarly for $\ell(\widehat{w},(x',y'))$. Let $w \in \mathcal{B}_B$ and let $\widehat{w} = w + b$ where $b \sim \mathcal{N}(0,\mathbb{I}_d\sigma^2)$. We have by previous analysis $\ell(\widehat{w};(x,y)) \leq \|\mathcal{Y}\|^2 + HB^2\|\mathcal{X}\|^2 + H\left\langle b,x\right\rangle^2$. Since $\left\langle b,x\right\rangle$ is distributed as a zero mean Gaussian with variance at most $\|\mathcal{X}\|^2\sigma^2$, we have $\mathbb{P}[\mid\left\langle b,x\right\rangle\mid \geq t] \leq \exp\left(\frac{-t^2}{\|\mathcal{X}\|^2\sigma^2}\right)$. Setting $t = \|\mathcal{X}\|\sigma\log{(K/\delta)}$ we obtain $\mathbb{P}[\mid\left\langle b,x\right\rangle\mid^2 \geq \|\mathcal{X}\|^2\sigma^2\log{(K/\delta)}] \leq \delta/K$. Thus with probability at least $1-\delta/K$ it holds that $\ell(\widehat{w};(x,y)) \leq \|\mathcal{Y}\|^2 + HB^2\|\mathcal{X}\|^2 + H\|\mathcal{X}\|^2\sigma^2\log{(K/\delta)}$. □

D Missing Details for Boosting

Algorithm 5 Boosting

```
Input: Dataset $S$, loss function $\ell$, Algorithm $\mathcal{A}$, $\widetilde{\sigma}$ privacy parameters $\epsilon, \delta$
Split the dataset $S$ into equally sized chunks $\{S_i\}_{i=1}^{m+1}$
For each $i \in [m+1]$, $\widehat{w}_i = \mathcal{A}\left(S_i, \frac{\epsilon}{2}, \delta\right)$
$i^* = \arg\max_{i\in[m]}\left(-\widehat{L}(\widehat{w}_i; S_{m+1}) + \text{Lap}(0,\widetilde{\sigma})\right)$
Output: $\widehat{w}_{i^*}$
```

We state the result of the boosting procedure in a general enough setup so as apply to our proposed algorithms. This leads to additional conditions on the base algorithm since our proposed methods may not produce the output in the constrained set.

Theorem 13. *Let ℓ be a non-negative, $\widetilde{H}$ smooth, convex loss function. Let $\epsilon, \delta > 0$. Let $\mathcal{A} : (S,\epsilon,\delta) \mapsto \mathcal{A}(S,\epsilon,\delta)$ be an algorithm such that*

1. *$\mathcal{A}$ satisfies (ϵ,δ)-DP*
2. *For any fixed S, $\mathcal{A}(S)$ is γ^2 sub-Gaussian [Ver18]:*

$$\sup_{\|u\|=1} \mathbb{E}\left[\exp\left(\left\langle\mathcal{A}(S),u\right\rangle^2/\gamma^2\right)\right] \leq 2$$

3. *For any fixed S, $\mathbb{P}_{(x,y)}[\ell(\mathcal{A}(S);(x,y)) > \Delta\left(\gamma,\beta\right)] < \beta$*
4. *Given a dataset S of n i.i.d. points, $\mathbb{E}[L(\mathcal{A}(S);\mathcal{D}) - \min_{w\in\mathcal{B}_B} L(w;\mathcal{D})] \leq \mathsf{ERR}\left(n,\epsilon,\gamma\right)$*

Let $\widetilde{\sigma}^2 = \frac{4(\|\mathcal{Y}\|^2 + \widetilde{H}\widetilde{\gamma}^2\|\mathcal{X}\|^2)}{n\epsilon}$ and $n_0 = \frac{16\gamma^2\log^8(4/\beta)\widetilde{H}}{\|\mathcal{Y}\|^2}$. Algorithm 5 with Algorithm $\mathcal{A}$ as input satisfies (ϵ,δ)-DP. Given a dataset S of $n \geq n_0$ samples, with probability at least $1-\beta$, the excess risk of its output $\widehat{w}_{i^*}$ is bounded as,

$$L(\widehat{w};\mathcal{D}) - L(w^*;\mathcal{D}) \leq \widetilde{O}\Bigg(\mathsf{ERR}\left(\frac{n}{4\log{(4/\beta)}},\frac{\epsilon}{2},\gamma\right) + \frac{2\Delta(\gamma,\beta/2)}{n\epsilon} + \frac{2\Delta\left(\gamma,\frac{\beta}{2n}\right)}{n} + \frac{32\gamma\sqrt{\widetilde{H}}\|\mathcal{Y}\|}{\sqrt{n}} + \frac{16\|\mathcal{Y}\|}{\sqrt{n}} + \frac{128\widetilde{H}\gamma^2}{n}\Bigg).$$

We first establish the following concentration bound for convex $\widetilde{H}$ smooth non-negative functions.

Lemma 11. *Let ℓ be a convex $\widetilde{H}$ smooth non-negative function. Let S be a dataset of n i.i.d. samples. Let w be a random variable which is γ^2 sub-Gaussian and independent of S and let $\Delta\left(\gamma,\beta\right)$ be such* *that* $\mathbb{P}_{(x,y)}[\ell(w;(x,y)) > \Delta(\gamma,\beta)] \leq \beta$*. Then, with probability at least* $1-\beta$*,*

$$\begin{aligned}\widehat{L}(w;S) &\leq (1+T(n,\beta))\,L(w;\mathcal{D}) + U(n,\beta)\\ L(w;\mathcal{D}) &\leq (1+T(n,\beta))\,\widehat{L}(w;S) + V(n,\beta)\end{aligned}$$

where $T(n,\beta) := \frac{4\gamma\log(4/\beta)\sqrt{\widetilde{H}}}{\|\mathcal{Y}\|\sqrt{n}}$, $U(n,\beta) := \frac{4\gamma\log(4/\beta)\|\mathcal{Y}\|\sqrt{\widetilde{H}}}{\sqrt{n}} + \frac{\|\mathcal{Y}\|\sqrt{\log(2/\beta)}}{\sqrt{n}}$ *and* $V(n,\beta) := \frac{4\gamma\log(4/\beta)\sqrt{\widetilde{H}}\|\mathcal{Y}\|}{\sqrt{n}} + \frac{2\Delta\left(\gamma,\frac{\beta}{4n}\right)\log(2/\beta)}{n} + \frac{\|\mathcal{Y}\|\sqrt{\log(2/\beta)}}{\sqrt{n}} + \frac{48\widetilde{H}\gamma^2\log^2(4/\beta)}{n}$.

Proof. With probability at least $1-\frac{\beta}{4}$, for each $(x,y)\in S$, $\ell(w;(x,y)) \leq \Delta\left(\gamma,\frac{\beta}{4n}\right)$. We condition on this event and apply Bernstein inequality to the random variable $L(w;\mathcal{D}) - \widehat{L}(w;S)$:

$$\mathbb{P}\left[\left|L(w;\mathcal{D}) - \widehat{L}(w;S)\right| > t\right] \leq \exp\left(-\frac{3nt^2}{6n\mathbb{E}[\left(L(w;\mathcal{D}) - \widehat{L}(w;S)\right)^2] + 2\Delta\left(\gamma,\frac{\beta}{4n}\right)t}\right)$$

This gives us that

$$\left|L(w;\mathcal{D}) - \widehat{L}(w;S)\right| \leq \frac{\Delta\left(\gamma,\frac{\beta}{4n}\right)\log(2/\beta)}{n} + \sqrt{\mathbb{E}\left(L(w;\mathcal{D}) - \widehat{L}(w;S)\right)^2\log(2/\beta)} \tag{22}$$

The term $\mathbb{E}[\left(L(w;\mathcal{D}) - \widehat{L}(w;S)\right)^2 = \frac{1}{n}\mathbb{E}[(\ell(w;(x,y)) - \mathbb{E}[\ell(w;(x,y)])^2] \leq \frac{1}{n}\mathbb{E}[(\ell(w;(x,y)))^2]$. Now,

$$\begin{aligned}\mathbb{E}[(\ell(w;(x,y)))^2] &\leq 2\mathbb{E}[(\ell(w;(x,y)) - \ell(0;(x,y))^2] + 2\mathbb{E}[(\ell(0;(x,y)))^2]\\ &\leq 2\mathbb{E}[(\langle\nabla\ell(w;(x,y)),w\rangle)^2] + 2\|\mathcal{Y}\|^2\end{aligned}$$

where the last step follows from convexity. We now use the fact that w is γ^2-sub-Gaussian, therefore $\langle\nabla\ell(w;(x,y)),w\rangle \leq \gamma\sqrt{\log(4/\beta)}\,\|\nabla\ell(w;(x,y))\|$ with probability at least $1-\beta/4$. We now use self-bounding property of non-negative smooth functions to get,

$$\begin{aligned}\mathbb{E}[(\ell(w;(x,y)))^2] &\leq 2\mathbb{E}[\|\nabla\ell(w;(x,y))\|^2\,\gamma^2\log(4/\beta) + 2\|\mathcal{Y}\|^2\\ &\leq 8\widetilde{H}\mathbb{E}[\ell(w;(x,y))]\gamma^2\log(4/\beta) + 2\|\mathcal{Y}\|^2\\ &= 8\widetilde{H}L(w;\mathcal{D})\gamma^2\log(4/\beta) + 2\|\mathcal{Y}\|^2\end{aligned}$$

Plugging the above in Eqn (22) gives us,

$$\begin{aligned}\left|L(w;\mathcal{D}) - \widehat{L}(w;S)\right| &\leq \frac{\Delta\left(\gamma,\frac{\beta}{4n}\right)\log(2/\beta)}{n} + 4\sqrt{\frac{\left(\widetilde{H}L(w;\mathcal{D})\gamma^2\log(4/\beta) + \|\mathcal{Y}\|^2\right)\log(1/\beta)}{n}}\\ &\leq \frac{\Delta\left(\gamma,\frac{\beta}{4n}\right)\log(2/\beta)}{n} + 4\sqrt{\frac{\widetilde{H}L(w;\mathcal{D})}{n}}\gamma\log(4/\beta) + \frac{\|\mathcal{Y}\|\sqrt{\log(2/\beta)}}{\sqrt{n}}.\end{aligned} \tag{23}$$

A simple application of AM-GM inequality gives,

$$\widehat{L}(w;S) \leq \left(1 + \frac{4\gamma\log(4/\beta)\sqrt{\widetilde{H}}}{\|\mathcal{Y}\|\sqrt{n}}\right)L(w;\mathcal{D}) + \frac{4\gamma\log(4/\beta)\|\mathcal{Y}\|\sqrt{\widetilde{H}}}{\sqrt{n}} + \frac{\|\mathcal{Y}\|\sqrt{\log(2/\beta)}}{\sqrt{n}}$$

This proves the first part of the lemma. For the second part, we use the following fact about non-negative numbers A, B, C [BBL03] (see after proof of Theorem 7)

$$A \le B + C\sqrt{A} \implies A \le B + C^2 + \sqrt{B}C$$

Thus, from Eqn. (23),

$$\begin{aligned}
L(w;\mathcal{D}) &\le \widehat{L}(w;S) + \frac{\Delta\left(\gamma, \frac{\beta}{4n}\right)\log(2/\beta)}{n} + \frac{\|\mathcal{Y}\|\sqrt{\log(2/\beta)}}{\sqrt{n}} + \frac{16\widetilde{H}\gamma^2\log^2(4/\beta)}{n} \\
&\quad + \frac{4\gamma\log(4/\beta)\sqrt{\widetilde{H}}}{\sqrt{n}}\left(\sqrt{\widehat{L}(w;S)} + \sqrt{\frac{\Delta\left(\gamma, \frac{\beta}{4n}\right)\log(2/\beta)}{n}} + \sqrt{\frac{\|\mathcal{Y}\|\sqrt{\log(2/\beta)}}{\sqrt{n}}}\right) \\
&\le \widehat{L}(w;S) + \frac{4\gamma\log(4/\beta)\sqrt{\widetilde{H}\widehat{L}(w;S)}}{\sqrt{n}} + \frac{\Delta\left(\gamma, \frac{\beta}{4n}\right)\log(2/\beta)}{n} + \frac{\|\mathcal{Y}\|\sqrt{\log(2/\beta)}}{\sqrt{n}} \\
&\quad + \frac{16\widetilde{H}\gamma^2\log^2(4/\beta)}{n} + \frac{4\gamma\sqrt{\widetilde{H}\Delta\left(\gamma, \frac{\beta}{4n}\right)}\log^{3/2}(4/\beta)}{n} + \frac{4\gamma\sqrt{\widetilde{H}\|\mathcal{Y}\|}\log^{5/4}(2/\beta)}{n^{3/4}} \\
&\le \left(1 + \frac{4\gamma\log(4/\beta)\sqrt{\widetilde{H}}}{\|\mathcal{Y}\|\sqrt{n}}\right)\widehat{L}(w;S) + \frac{4\gamma\log(4/\beta)\|\mathcal{Y}\|\sqrt{\widetilde{H}}}{\sqrt{n}} + \frac{\Delta\left(\gamma, \frac{\beta}{4n}\right)\log(2/\beta)}{n} \\
&\quad + \frac{\|\mathcal{Y}\|\sqrt{\log(2/\beta)}}{\sqrt{n}} + \frac{16\widetilde{H}\gamma^2\log^2(4/\beta)}{n} + \frac{4\gamma\sqrt{\widetilde{H}\Delta\left(\gamma, \frac{\beta}{4n}\right)}\log^{3/2}(4/\beta)}{n} \\
&\quad + \frac{4\gamma\sqrt{\widetilde{H}\|\mathcal{Y}\|}\log^{5/4}(2/\beta)}{n^{3/4}}
\end{aligned}$$

where the last inequality follows from AM-GM inequality. Simplifying the expressions yields the claimed bound. □

Proof of Theorem 13. Since the models $\{\widehat{w}_i\}_{i=1}^m$ are trained on disjoint datasets, by parallel composition $\{\widehat{w}_i\}_{i=1}^m$ satisfies $\left(\frac{\epsilon}{2}, \frac{\delta}{2}\right)$-DP. We know that probability at least $1 - \frac{\delta}{2}$, $\ell(w;(x,y)) \le \Delta\left(\gamma, \frac{\delta}{2}\right)$. Thus conditioning on this event, from the guarantee of the report noisy max procedure, we have that it satisfies $\left(\frac{\epsilon}{2}\right)$-DP. The privacy proof thus follows from absorbing the failure probability into δ part and adaptive composition.

We proceed to the utility part. Let $\widetilde{w}$ be the model among $\{\widehat{w}_i\}_{i=1}^m$ with minimum empirical error on the set S_{m+1}. The excess risk of each $\widehat{w}_i$ is bounded as,

$$\mathbb{E}[L(\widehat{w}_i;\mathcal{D})] - L(w^*;\mathcal{D}) \le \mathsf{ERR}\left(\frac{n}{m+1}, \frac{\epsilon}{2}, \gamma\right)$$

From Markov's inequality, with probability at least $3/4$, $L(\widehat{w}_i;\mathcal{D}) \le L(w^*;\mathcal{D}) + 4\mathsf{ERR}\left(\frac{n}{m+1}, \frac{\epsilon}{2}\right)$. From independence of $\{w_i\}_{i=1}^m$, with probability at least $1 - 1/4^m = 1 - \frac{\beta}{4}$, there exists one model, say $\widehat{w}_{i^*}$, such $L(\widehat{w}_{i^*};\mathcal{D}) \le L(w^*;\mathcal{D}) + 4\mathsf{ERR}\left(\frac{n}{m+1}, \frac{\epsilon}{2}\right)$.

Also, from the guarantee of Report-Noisy-Max, we have that with probability at least $1 - \beta/4$

$$L(\widehat{w};S_{m+1}) \le L(\widetilde{w};S_{m+1}) + \frac{\Delta(\gamma,\beta/4)(m+1)\log^2(4m/\delta)}{n\epsilon}$$

Now, we apply Lemma 11. From a union bound, with probability at least $1 - \frac{\beta}{2}$, all $\{w_i\}_{i=1}^m$ satisfy the inequalities in Lemma 11 with β substituted as $\frac{\beta}{2m}$.

Thus, for the output $\widehat{w}$, probability at least $1 - \frac{\beta}{2}$,

$$\begin{aligned}
&L(\widehat{w}; \mathcal{D}) \\
&\leq \left(1 + T\left(\frac{n}{m+1}, \frac{\beta}{2m}\right)\right) L(\widehat{w}; S_{m+1}) + V\left(\frac{n}{(m+1)}, \frac{\beta}{2m}\right) \\
&\leq \left(1 + T\left(\frac{n}{m+1}, \frac{\beta}{2m}\right)\right) L(\widetilde{w}; S_{m+1}) + \left(1 + T\left(\frac{n}{m+1}, \frac{\beta}{2m}\right)\right) \frac{\Delta(\gamma, \beta/4)(m+1)\log^2(4m/\delta)}{n\epsilon} \\
&+ V\left(\frac{n}{(m+1)}, \frac{\beta}{2m}\right) \\
&\leq \left(1 + T\left(\frac{n}{m+1}, \frac{\beta}{2m}\right)\right) L(w_{i^*}; S_{m+1}) + \frac{2\Delta(\gamma, \beta/4)(m+1)\log^2(4m/\delta)}{n\epsilon} + V\left(\frac{n}{(m+1)}, \frac{\beta}{2m}\right)
\end{aligned}$$

where in the above we use that $T\left(\frac{n}{m+1}\right) \leq 1$ given the lower bound on n and the setting of m. Furthermore, the last inequality follows because $\widetilde{w}$ has lowest empirical risk on S_{m+1}. Let $W(n, m, \beta) = \frac{2\Delta(\gamma,\beta/4)(m+1)\log^2(4m/\delta)}{n\epsilon} + V\left(\frac{n}{(m+1)}, \frac{\beta}{2m}\right)$. We now apply the other guarantee in Lemma 11 and the fact that w_{i^*} has small excess risk. With probability at least $1 - \delta/2$,

$$\begin{aligned}
&L(\widehat{w}; \mathcal{D}) \\
&\leq \left(1 + T\left(\frac{n}{m+1}, \frac{\beta}{2m}\right)\right)^2 L(w_{i^*}; \mathcal{D}) + \left(1 + T\left(\frac{n}{m+1}, \frac{\beta}{2m}\right)\right) U\left(\frac{n}{m+1}, \frac{\beta}{2m}\right) + W(n, m, \beta) \\
&\leq L(w^*; \mathcal{D}) + 2T\left(\frac{n}{m+1}, \frac{\beta}{2m}\right) L(w^*; \mathcal{D}) + 4\mathsf{ERR}\left(\frac{n}{m+1}, \frac{\epsilon}{2}\right) \\
&+ 8T\left(\frac{n}{m+1}, \frac{\beta}{2m}\right) 4\mathsf{ERR}\left(\frac{n}{m+1}, \frac{\epsilon}{2}\right) + 2U\left(\frac{n}{m+1}, \frac{\beta}{2m}\right) + W(n, m, \beta)
\end{aligned}$$

Let $X(n, m, \beta) = 4\mathsf{ERR}\left(\frac{n}{m+1}, \frac{\epsilon}{2}\right) + 8T\left(\frac{n}{m+1}, \frac{\beta}{2m}\right) 4\mathsf{ERR}\left(\frac{n}{m+1}, \frac{\epsilon}{2}\right) + 2U\left(\frac{n}{m+1}, \frac{\beta}{2m}\right) + W(n, m, \beta)$. Note that $m = 4\log(4/\beta)$ and $T\left(\frac{n}{m+1}, \frac{\beta}{2m}\right) \leq \frac{16\gamma\log^4(2/\beta)\sqrt{H}}{\|\mathcal{Y}\|\sqrt{n}}$. Substituting this and using the fact that, $L(w^*; \mathcal{D}) \leq L(0; \mathcal{D}) \leq \|\mathcal{Y}\|^2$, we get that with probability at least $1 - \delta$,

$$L(\widehat{w}; \mathcal{D}) \leq L(w^*; \mathcal{D}) + \frac{16\gamma\log^4(2/\beta)\sqrt{H}\|\mathcal{Y}\|}{\sqrt{n}} + X(n, 4\log(2/\beta), \beta)$$

Substituting and simplifying the $X(n, 4\log(4/\beta), \beta)$ we have that

$$\begin{aligned}
&X(n, 4\log(2/\beta), \beta) \\
&\leq 12\mathsf{ERR}\left(\frac{n}{4\log(4/\beta)}, \frac{\epsilon}{2}, \gamma\right) + \frac{2\Delta(\gamma, \beta/4)\log^3(4/\beta)\log(2/\delta)}{n\epsilon} + \frac{2\Delta\left(\gamma, \frac{\beta}{2n}\right)\log^4(8/\beta)}{n} \\
&+ \frac{16\gamma\log^4(8/\beta)\sqrt{\widetilde{H}}\|\mathcal{Y}\|}{\sqrt{n}} + \frac{16\|\mathcal{Y}\|\log^4(8/\beta)}{\sqrt{n}} + \frac{128\widetilde{H}\gamma^2\log^4(8/\beta)}{n}
\end{aligned}$$

Hence, with probability at least $1 - \beta$,

$$\begin{aligned}
&L(\widehat{w}; \mathcal{D}) \\
&\leq L(w^*; \mathcal{D}) + 12\mathsf{ERR}\left(\frac{n}{4\log(4/\beta)}, \frac{\epsilon}{2}, \gamma\right) + \frac{2\Delta(\gamma, \beta/4)\log^3(4/\beta)\log(2/\delta)}{n\epsilon} + \frac{2\Delta\left(\gamma, \frac{\beta}{2n}\right)\log^4(8/\beta)}{n} \\
&+ \frac{32\gamma\log^4(8/\beta)\sqrt{\widetilde{H}}\|\mathcal{Y}\|}{\sqrt{n}} + \frac{16\|\mathcal{Y}\|\log^4(8/\beta)}{\sqrt{n}} + \frac{128\widetilde{H}\gamma^2\log^4(8/\beta)}{n}
\end{aligned}$$

□

D.1 Boosting the JL Method

Theorem 14. *The boosting procedure (Algorithm 5) using the JL method (Algorithm 2) as Algorithm $\mathcal{A}$ satisfies (ϵ,δ)-DP, and with probability at least $1-\beta$, its output $\widehat{w}_{i^*}$ has excess risk,*

$$L(\widehat{w}_{i^*};\mathcal{D}) - L(w^*;\mathcal{D}) \leq \widetilde{O}\Bigg(\frac{\sqrt{H}B\left\|\mathcal{X}\right\|\left\|\mathcal{Y}\right\|}{\sqrt{n}} + \frac{\left(\left(\sqrt{H}B\left\|\mathcal{X}\right\|\right)^{4/3}\left\|\mathcal{Y}\right\|^{2/3} + \left(\sqrt{H}B\left\|\mathcal{X}\right\|\right)^{2}\right)}{(n\epsilon)^{2/3}} + \frac{\left(\|\mathcal{Y}\|^2 + HB^2\|\mathcal{X}\|^2\right)}{n\epsilon} + \frac{\left(\|\mathcal{Y}\| + \sqrt{H}B\|\mathcal{X}\|\right)}{\sqrt{n}}\Bigg)$$

Proof. For the JL method, we consider the boosting procedure in the k dimension space – this is only for the sake of analysis and the algorithm remains the same. In particular, consider the distribution to $\Phi\mathcal{D}$ which, when sampled from, gives the data point $(\Phi x, y)$ where $(x,y)\sim\mathcal{D}$.

Suppose the boosting procedure gives the following k dimensional models: $\widetilde{w}_1,\cdots\widetilde{w}_m$; note that the norm of all these are bounded by $2B$. Let $\widetilde{w}^* \in \arg\min_{\|w\|\leq 2B} L(w;\Phi\mathcal{D})$. Since the outputs satisfy $\|\widetilde{w}_i\|\leq 2B$, the sub-Gaussian parameter $\gamma = O(B)$. We now compute the other parameter $\Delta(\gamma,\beta)$, which is the high probability bound on loss. Note that for a fixed data point $(\Phi x, y)$, an H smooth, non-negative, bounded at zero loss, at any point w s.t. $\|w\|\leq 2B$ is upper bounded by $2\left(\left\|\mathcal{Y}\right\|^2 + 4B^2H\left\|\Phi x\right\|^2\right)$. From the JL guarantee, with the given k, the term $\left\|\Phi x\right\|^2 \leq 2\left\|\mathcal{X}\right\|$ with probability at least $1-\beta/4$. This gives us $\Delta(2B,\beta) = 2\left(\left\|\mathcal{Y}\right\|^2 + 16\left\|\mathcal{X}\right\|^2B^2\right)$.

We now invoke 13 substituting Δ, γ and $\widetilde{H} = H\|\mathcal{X}\|^2$ to get that with probability at least $1-\frac{\beta}{2}$ output satisfies,

$$L(\widetilde{w}_{i^*};\Phi\mathcal{D}) \leq L(\widetilde{w}^*;\Phi\mathcal{D}) + \frac{128B\sqrt{H}\|\mathcal{X}\|\|\mathcal{Y}\|\log^4\left(8/\beta\right)}{\sqrt{n}} + \alpha\left(\frac{n}{4\log\left((8/\beta)\right)},\frac{\epsilon}{2},2B\right) + \frac{128\left(\|\mathcal{Y}\|^2 + HB^2\|\mathcal{X}\|^2\right)\log^4\left(8/\beta\right)}{n\epsilon} + \frac{128\left(\|\mathcal{Y}\| + \sqrt{H}B\|\mathcal{X}\|\right)\log^4\left(8/\beta\right)}{\sqrt{n}}$$

Define $W := \frac{128B\sqrt{H}\|\mathcal{X}\|\|\mathcal{Y}\|\log^4\left(8/\beta\right)}{\sqrt{n}} + \frac{128\left(\|\mathcal{Y}\|^2+HB^2\|\mathcal{X}\|^2\right)\log^4\left(8/\beta\right)}{n\epsilon} + \frac{128\left(\|\mathcal{Y}\|+\sqrt{H}B\|\mathcal{X}\|\right)\log^4\left(8/\beta\right)}{\sqrt{n}}$. The excess risk of the final output $\widehat{w}_{i^*} = \Phi^\top\widetilde{w}_{i^*}$ is bounded as,

$$\begin{aligned}
L(\widehat{w}_{i^*};\mathcal{D}) - L(w^*;\mathcal{D}) &= L(\widetilde{w}_{i^*};\Phi\mathcal{D}) - L(w^*;\mathcal{D}) \\
&\leq L(\widetilde{w}^*;\Phi\mathcal{D}) + \alpha\left(\frac{n}{4\log\left(8/\beta\right)},\frac{\epsilon}{2},2B\right) + W - L(w^*;\mathcal{D}) \\
&\leq L(\Phi w^*;\Phi\mathcal{D}) + \alpha\left(\frac{n}{4\log\left(8/\beta\right)},\frac{\epsilon}{2},2B\right) + W - L(w^*;\mathcal{D}) \\
&\leq \alpha\left(\frac{n}{4\log\left(8/\beta\right)},\frac{\epsilon}{2},2B\right) + W + \frac{H}{2}\left|\langle\Phi x,\Phi w^*\rangle - \langle x,w^*\rangle\right|^2
\end{aligned}$$

where the last inequality follows from smoothness and that $\nabla L(w^*;\mathcal{D}) = 0$. Finally, from the JL property, with probability at least $1-\frac{\beta}{2}$, $\left|\langle\Phi x,\Phi w^*\rangle - \langle x,w^*\rangle\right|^2 \leq \alpha^2\left\|w^*\right\|^2\left\|\mathcal{X}\right\|^2$. Combining and substituting the values of α and $\alpha\left(\frac{n}{4\log(8/\beta)},\frac{\epsilon}{2}\right)$ from Theorem 2 gives the claimed result. □

D.2 Boosting Output Perturbation Method

Theorem 15. *The boosting procedure (Algorithm 5) using the output perturbation method (Algorithm 3) as input Algorithm $\mathcal{A}$ satisfies (ϵ,δ)-DP, and with probability at least $1-\beta$, its output $\widehat{w}_{i^*}$ has* *excess risk,*

$$L(\widehat{w}_{i^*};\mathcal{D}) - L(w^*;\mathcal{D}) \leq \widetilde{O}\Bigg(\frac{\sqrt{H}B\,\|\mathcal{X}\|\,\|\mathcal{Y}\| + \|\mathcal{Y}\|^2}{\sqrt{n}} + \frac{\left(\left(\sqrt{H}B\,\|\mathcal{X}\|\right)^{4/3}\|\mathcal{Y}\|^{2/3} + \left(\sqrt{H}B\,\|\mathcal{X}\|\right)^2\right)}{(n\epsilon)^{2/3}} + \frac{(\|\mathcal{Y}\|^2 + HB^2\|\mathcal{X}\|^2)}{n\epsilon} + \frac{\left(\|\mathcal{Y}\| + \sqrt{H}B\|\mathcal{X}\|\right)}{\sqrt{n}}\Bigg)$$

Proof. Firstly, note that $\widetilde{H} = H\|\mathcal{X}\|^2$. We now compute γ and Δ to invoke Theorem 13. Since the algorithm simply adds Gaussian noise of variance $\sigma^2\mathbb{I}_d$ to a vector $\widetilde{w}$ where $\|\widetilde{w}\| \leq B$, we have that $\gamma^2 \leq B^2 + \sigma^2$. For the bound on loss parameter Δ, by direct computation, $\Delta(\gamma,\beta) \leq 2\left(\|\mathcal{Y}\|^2 + H\langle\widetilde{w}+\xi, x\rangle^2\right) \leq 2\|\mathcal{Y}\|^2 + 2HB^2\|\mathcal{X}\|^2 + 2H\langle\xi,x\rangle^2 \leq 2\|\mathcal{Y}\|^2 + 2HB^2\|\mathcal{X}\|^2 + 2H\sigma^2\log(1/\beta)$ where the last inequality follows since $\langle\xi,x\rangle \sim \mathcal{N}(0,\|x\|^2)$. Plugging these and the value of σ^2 from Theorem 3 into Theorem 13 gives the claimed bound. □

E Non-private Lower Bounds

We first note a simple one-dimensional lower bound.

Theorem 16. *(Implicit in [SST10]) Let $\|\mathcal{X}\| \geq 1$ and $H \geq 2$. For any Algorithm $\mathcal{A}$, there exists a 1-dimensional H-smooth non-negative GLM, $\ell : \mathbb{R} \times (\mathcal{X}\times\mathcal{Y}) \mapsto \mathbb{R}$, and a distribution $\mathcal{D}$ over $(\mathcal{X}\times\mathcal{Y})$ with $\|w^*\| = \Theta(\min\{\|\mathcal{Y}\|, B\}$ such that the excess population risk of the output of $\mathcal{A}$ when run on $S \sim \mathcal{D}^n$ is lower bounded as $\Omega\left(\frac{\min\{\|\mathcal{Y}\|,B\|\mathcal{X}\|\}}{\sqrt{n}}\right)$.*

We remark that this requires a slight modification of the example used in [SST10]. Specifically, therein the loss function is defined as

$$\ell(w,(x,y)) = \begin{cases} (w-y)^2 & |w-y| \leq \frac{1}{2} \\ |w-y| - 1/4 & |w-y| \geq \frac{1}{2} \end{cases}$$

with $y \in \{\pm 1\}$ and $x = 1$. Our statement is obtained by setting the domain of labels as $\{\pm\min\{B, \|\mathcal{Y}\|\}\}$. A reduction in [Sha15] enables lower bounds from problem instances with general $\|\mathcal{X}\|$ and $\|w^*\| = R$ to instances with $\|\mathcal{X}\| = 1$ to $\|w^*\| = R\|\mathcal{X}\|$.

We now show that the lower bounds presented in [Sha15] to the unconstrained setting. We start by stating the original bound from [Sha15] which holds for squared loss.

Theorem 17. *Let $\ell(w,(x,y)) = \frac{1}{2}(\langle w,x\rangle - y)^2$ be the squared loss function and $B > 0$. Then for any algorithm, $\mathcal{A}$, there exists a distribution $\mathcal{D}$ over $(\mathcal{X}\times\mathcal{Y})$ and a constant C such that for $S \sim \mathcal{D}^n$ it holds that $\mathbb{E}\left[L(\mathcal{A}(S);\mathcal{D}) - L(w_B^*;\mathcal{D})\right] \geq C\min\left\{\|\mathcal{Y}\|^2, \frac{B^2\|\mathcal{X}\|^2 + d\|\mathcal{Y}\|^2}{n}, \frac{B\|\mathcal{Y}\|\|\mathcal{X}\|}{\sqrt{n}}\right\}$, where $w_B^* = \underset{w\in\mathcal{B}_B}{\arg\min}\{L(w;\mathcal{D})\}$.*

We now make three observations. First we note that this Theorem holds even for $\mathcal{A}(S) \notin \mathcal{B}_B$. Second we note that the construction of the distribution in the above Theorem is such that $\min\limits_{w^*\in\mathcal{B}_B}\{L(w;\mathcal{D})\} = \min\limits_{w^*\in\mathbb{R}^d}\{L(w;\mathcal{D})\}$. Finally, note that $\frac{H}{2}(\langle w,x\rangle - \frac{y}{\sqrt{H}})^2 = \frac{1}{2}(\sqrt{H}\langle w,x\rangle - y)^2$. This gives the following corollary.

Corollary 5. *Let $B > 0$. Then for any algorithm, $\mathcal{A}$, there exists a distribution $\mathcal{D}$ over $(\mathcal{X}\times\mathcal{Y})$ and an H-smooth non-negative GLM, $\ell : \mathbb{R}^d \times (\mathcal{X}\times\mathcal{Y}) \mapsto \mathbb{R}$, with minimizer $w^* = \underset{w\in\mathbb{R}^d}{\arg\min}\{L(w;\mathcal{D})\}$ such that $\|w^*\| = B$ and for $S \sim \mathcal{D}^n$ it holds that*

$$\mathbb{E}\left[L(\mathcal{A}(S);\mathcal{D}) - L(w^*;\mathcal{D})\right] = \Omega\left\{\min\left\{\|\mathcal{Y}\|^2, \frac{HB^2\|\mathcal{X}\|^2 + d\|\mathcal{Y}\|^2}{n}, \frac{\sqrt{H}B\|\mathcal{Y}\|\|\mathcal{X}\|}{\sqrt{n}}\right\}\right\}.$$

F Additional Results

Lemma 12. *Let* $\Phi \in \mathbb{R}^{d \times k}$ *be a random matrix such that for any* $u \in \mathbb{R}^d$*, with probability at least* $1-\beta$*,* $(1-\alpha)\left\|u\right\|^2 \leq \left\|\Phi u\right\|^2 \leq (1+\alpha)\left\|u\right\|^2$*. Then for any* u, v*, with probability at least* $1-2\beta$*,* $|\langle \Phi u, \Phi v\rangle - \langle u, v\rangle| \leq \alpha \left\|u\right\| \left\|v\right\|$.

Proof. Firstly, note that it suffices to prove the result for unit vectors u and v. From the norm preservation result, with probability at least $1-2\beta$, we have that,

$$(1-\alpha)\left\|u+v\right\|^2 \leq \left\|\Phi(u+v)\right\| \leq (1+\alpha)\left\|u+v\right\|^2$$
$$(1-\alpha)\left\|u-v\right\|^2 \leq \left\|\Phi(u-v)\right\| \leq (1+\alpha)\left\|u-v\right\|^2$$

Therefore, we have

$$\begin{aligned}\langle \Phi u, \Phi v\rangle &= \frac{1}{4}\left(\left\|\Phi\left(u+v\right)\right\|^2 - \left\|\Phi\left(u-v\right)\right\|^2\right)\\ &\leq \frac{1}{4}\left((1+\alpha)\left\|u+v\right\|^2 - (1-\alpha)\left\|u-v\right\|^2\right)\\ &\leq \langle u, v\rangle + \alpha\end{aligned}$$

This gives us that $\langle \Phi u, \Phi v\rangle \leq \langle u, v\rangle + \alpha$. The other inequality follows in the same way. □