JMLR: Workshop and Conference Proceedings vol 23 (2012) 25.1–25.40 25th Annual Conference on Learning Theory

Private Convex Empirical Risk Minimization and High-dimensional Regression

Daniel Kifer DKIFER@CSE.PSU.EDU
Adam Smith ASMITH@CSE.PSU.EDU
Abhradeep Thakurta AZG161@CSE.PSU.EDU
Department of Computer Science and Engineering
Pennsylvania State University

Editor: Shie Mannor, Nathan Srebro, Robert C. Williamson

Abstract

We consider *differentially private* algorithms for convex empirical risk minimization (ERM). Differential privacy (Dwork et al., 2006b) is a recently introduced notion of privacy which guarantees that an algorithm's output does not depend on the data of any individual in the dataset. This is crucial in fields that handle sensitive data, such as genomics, collaborative filtering, and economics. Our motivation is the design of private algorithms for sparse learning problems, in which one aims to find solutions (e.g., regression parameters) with few non-zero coefficients. To this end:

(a) We significantly extend the analysis of the "objective perturbation" algorithm of Chaudhuri et al. (2011) for convex ERM problems. We show that their method can be modified to use less noise (be more accurate), and to [redacted] We also give a tighter, data-dependent analysis of the additional error introduced by their method.

A key tool in our analysis is a new nontrivial limit theorem for differential privacy which is of independent interest: if a sequence of differentially private algorithms converges, in a *weak* sense, then the limit algorithm is also differentially private.

In particular, our methods give the best known algorithms for differentially private linear regression. These methods work in settings where the number of parameters p is less than the number of samples n.

(b) We give the first two private algorithms for *sparse* regression problems in high-dimensional settings, where p is much larger than n. We analyze their performance for linear regression: under standard assumptions on the data, our algorithms have vanishing empirical risk for $n = poly(s, \log p)$ when there exists a good regression vector with s nonzero coefficients. Our algorithms demonstrate that randomized algorithms for sparse regression problems can be both stable and accurate – a combination which is impossible for deterministic algorithms.

1. Introduction

Problem Setting Given a data set $(d_1, ..., d_n)$ of n individuals, where each observation d_i lies in a fixed domain $\mathcal{T}$, consider the following p-dimensional convex optimization problem:

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{F}} \frac{1}{n}\left(\sum_{i=1}^{n} \ell(\theta; d_i) + r(\theta)\right), \tag{1}$$

where $\ell(\theta; d_i)$ is a real-valued function that is convex in the first parameter $\theta \in \mathbb{R}^p$ for every $d \in \mathcal{T}$, the *regularizer* r is an arbitrary convex function and the *constraint* $\mathbb{F} \subseteq \mathbb{R}^p$ is a closed convex set.

This type of program captures a variety of empirical risk minimization (ERM) problems. For example, when $r = 0$, it can describe the MLE's for linear regression (where $\ell(\theta; d) = (y - \langle x, \theta \rangle)^2$ and $d = (x, y)$)) and logistic regression (where $\ell(\theta; d) = \log(1 + \exp(y\langle x, \theta \rangle))$). In the *Lasso*, widely used for selecting a sparse estimator for linear regression, one adds the regularizer $r(\theta) = \Lambda\|\theta\|_1$ or constrains the solution to $\mathbb{F} = \{\theta : \|\theta\|_1 \leq t\}$; here Λ and t are fixed real numbers.

The regression literature distinguishes two settings depending on the relationship between n (the number of records) and p (the dimension). In the classical *low-dimensional* setting, p is constant or grows polynomially slower than n. In the *high-dimensional* setting, p grows much faster than n. In order for ERM solutions to be meaningful in the high-dimensional setting, one typically has to look for solutions θ with some additional structure, such as sparsity (for vectors) or low rank (for matrices). To make the corresponding optimization problem tractable, the structural constraint is often replaced with a convex regularizer or constraint, such as the ℓ_1 or nuclear norms. This is a prolific area of research; see Negahban et al. (2010) for a brief survey.

Differential privacy Learning algorithms are frequently run on sensitive data (say, genomic data or email transcripts). Although there is substantial social benefit to publishing the results of an analysis over such data, there is a significant risk of inadvertently leaking information about the entries in the data set.

A recent line of work seeks to place private data analysis on rigorous, principled foundations. Our algorithms satisfy *differential privacy* (Dwork et al., 2006b; Dwork, 2006), which emerged from this line of work and is now widely studied in computer science and statistics. See Raskhodnikova and Smith (2010); Roth (2011) for links to papers and surveys. Intuitively, differential privacy requires that datasets differing in only one entry induce similar distributions on the output of a (randomized) algorithm. This implies that an attacker will draw essentially the same conclusions about an individual whether or not that individual's data was used – even if many records are known *a priori* to the attacker. See Dwork (2006); Ganta et al. (2008); Dwork and Naor (2010); Kifer and Machanavajjhala (2011) for further discussion of the implications of differential privacy.

Definition 1 (Differential privacy Dwork et al. (2006b,a)) *A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private if for any two datasets $\mathcal{D}$ and $\mathcal{D}'$ drawn from a domain $\mathcal{T}$ with $|\mathcal{D}\Delta\mathcal{D}'| = 1$ (Δ being the symmetric difference), and for all (Borel) sets $\mathcal{O} \subseteq Range(\mathcal{A})$ the following holds:* $\Pr[\mathcal{A}(\mathcal{D}) \in \mathcal{O}] \leq e^{\epsilon} \Pr[\mathcal{A}(\mathcal{D}') \in \mathcal{O}] + \delta$.

Differentially private algorithms cannot be deterministic and in particular cannot always output the exact minimizer from (1). Our goal is to find algorithms that solve convex ERM with as little additional empirical risk as possible.

Related Work The convex ERM setting considered here was explicitly studied by Chaudhuri et al. (2011); Rubinstein et al. (2009), though variants and special cases had been considered previously. They considered two basic techniques: *output perturbation* (studied by both papers), where one releases the output $\hat{\theta}$ with additive noise, and *objective perturbation* (introduced by Chaudhuri et al. (2011) and further studied by Dwork et al. (2009)), where one releases the (exact) minimizer of a perturbed version of the objective function.

There also exist other techniques for specific convex optimization problems such as order statistics (Nissim et al., 2007; Dwork and Lei, 2009) and linear regression (Dwork and Lei, 2009). The *sample-and-aggregate* framework (Nissim et al., 2007) is a generic technique for designing private

algorithms, which can be instantiated in many different ways. Smith (2011) applied it to a class of statistical problems that includes low-dimensional ERM.

The existing analysis of output perturbation [redacted] than that of objective perturbation. Under the minimal set of assumptions that allow both techniques to apply, the worst-case theoretical guarantees on the two techniques' performance are very similar (Chaudhuri et al., 2011), and are better than the guarantees one gets for the techniques of Dwork and Lei (2009); Smith (2011). However, in experiments objective perturbation performed much better than objective perturbation. This phenomenon was partly explained by Dwork et al. (2009), who showed that in logistic regression, objective perturbation distorts the minimizer much less than output perturbation on "nice" data.

All the previous techniques work in the low-dimensional regime. When $p \gg n$, they fail to provide consistent error estimates.

1.1. Our Contributions

Our two main contributions are improving the objective perturbation technique, and providing the first algorithms for private high-dimensional sparse regression and feature selection.

1.1.1. IMPROVING OBJECTIVE PERTURBATION

With the objective perturbation technique, instead of minimizing the empirical loss $\hat{J}(\theta; \mathcal{D}) = \frac{1}{n}(\sum_i \ell(\theta; d_i) + r(\theta))$, one considers a linear perturbation $J^{\text{priv}}(\theta; \mathcal{D}) = \hat{J}(\theta; \mathcal{D}) + \langle B, \theta \rangle$, where B is a random vector drawn according to a gamma distribution. The output of the algorithm is the minimizer of $J^{\text{priv}}(\cdot; \mathcal{D})$. We improve the treatment of Chaudhuri et al. (2011) in several respects:

More Accurate Objective Perturbation We show that drawing the perturbation B from a Gaussian (instead of gamma) distribution, leads to a $\tilde{\Omega}(\sqrt{p})$ improvement in the utility guarantees of the objective algorithm, at the cost of relaxing the privacy guarantee from $(\epsilon, 0)$- to (ϵ, δ)-differential privacy for negligible δ. When $\delta < 1/n^2$, the relaxed guarantee has very similar semantics to the original (Ganta et al., 2008). This result parallels a similar improvement that is possible for output perturbation (see, *e.g.,* Dwork et al. (2006a)), though the privacy and utility proofs are quite different.

Generalized Privacy Analysis and a Limit Theorem for Differential Privacy We also show that objective perturbation (with either Gaussian or gamma perturbation) continues to be private even when the convex regularizer r is nondifferentiable and when the parameter vector θ is constrained to a closed convex set $\mathbb{F}$. As mentioned above, the privacy proof of CMS required that r be differentiable and θ be unconstrained.

Our analysis greatly extends the range of problems to which objective perturbation applies. For example, it allows one to use objective perturbation for convex programs like the Lasso (where the regularizer r is the L_1 norm) and nuclear norm regularized minimization (Negahban et al., 2010), which was earlier not possible. The extension is also critical for applying the objective perturbation technique to linear regression.

The main tool we use in the above analysis is a *limit theorem* for differential privacy. The theorem states that if a sequence of (ϵ, δ)-differentially private algorithms $\mathcal{A}_1, \mathcal{A}_2, \ldots$ converges in a weak sense, then the limiting algorithm $\mathcal{A} = \lim_{i \to \infty} \mathcal{A}_i$ is also (ϵ, δ)-differentially private. Note

that the probabilistic behavior of $\mathcal{A}$ can be very different from any of the $\mathcal{A}_i$ (see Example 1). We feel this tool is likely to have other applications.

The idea behind our generalized analysis of objective perturbation is to approximate the constrained, nondifferentiable problem in (1) with a sequence of unconstrained, differentiable problems, and apply our limit theorem to the resulting sequence of algorithms. The difficulty is in ensuring that the resulting problems are all convex (so that the previous analysis applies) and converge in an appropriate sense to the original problem.

Data-dependent Utility Analysis Finally, we provide an improved, data-dependent utility analysis. Our approach is inspired by the analysis of Dwork et al. (2009), which was specific to logistic regression. We show that for "nice" data, namely, data sets for which the loss function is *strongly* convex in a neighborhood of its minimizer, objective perturbation has much better error guarantees than in the worst case (roughly, a factor of $\sqrt{p}$ lower or a typical setting of parameters). The assumption of strong convexity is common in the optimization literature (*e.g.,* Nocedal and Wright, 2000; Negahban et al., 2010).

Case Study: Linear Regression We illustrate our results with an application to low-dimensional linear regression. For a typical setting of parameters, we obtain a factor of p improvement in the expected additional risk compared to previous approaches.

1.1.2. SPARSE REGRESSION

The second part of our paper initiates the study of private algorithms for high-dimensional learning with structural constraints. Specifically, we consider algorithms for linear regression that seek a *sparse* vector of regression coefficients. As mentioned above, this is a well-studied problem (without privacy considerations) and a popular approach is to regularize the standard ERM with the ℓ_1 norm of θ (the "Lasso"). The resulting program is convex (making it computationally tractable) and produces sparse solutions with good generalization error in a variety of settings. Roughly, the Lasso performs well when there is an s-sparse vector θ^{sp} that labels the data well and $n = \omega(s \log p)$.

Unfortunately, none of the existing approaches to private convex ERM (including our variant of objective perturbation) perform well on the Lasso when $p \gg n$, never mind when n grows as $\log p$. Nevertheless, we give two algorithms that produce consistent, sparse estimates θ^{sp} when n is a least a polynomial in s and $\log p$. The algorithms are not specific to linear regression, but we analyze them in that setting for convenience. We take a two stage approach: we first privately select a support set of small size and then run the objective perturbation algorithm to select a parameter vector with support on this set. We provide two algorithms for the first stage: 1. *Superpolynomial time, via exponential sampling*: We apply the *exponential mechanism* (McSherry and Talwar, 2007) to sample a good support set of size s. To instantiate the mechanism, we define the "score" of a set of features Γ to be the empirical loss of the best parameter vector with support in Γ. This algorithm is roughly the nonprivate analogue of exhaustive search over all subsets ("L_0 minimization" (Wipf and Rao, 2005)). The algorithm is inefficient but provides a baseline for comparison. 2. *Polynomial time, via sample-and-aggregate*: Following the sample and aggregate framework (Nissim et al., 2007), the algorithm splits the data set into disjoint blocks, selects a support set for each block and then aggregates the results via a novel "voting" aggregation algorithm. The algorithm works under the assumption that many random sub-samples of the data generate the same support set.

In the context of linear regression (and under typical assumptions), the algorithms produce consistent estimates of the support of θ^{sp} when $n = \omega(s^3 \log p)$ and $n = \omega(s^2 \log^2 p)$, respectively.

Designing algorithms that match the performance of the best nonprivate algorithms, even asymptotically, remains an interesting open problem.

Structure of this paper Section 2 details our results on objective perturbation, while Section 3 discusses sparse regression. A summary of notation is included in Appendix A for convenience. The remaining appendices provide omitted details and proofs.

2. Differentially Private Convex Optimization

2.1. Tool: A limit theorem for differentially private algorithms

Establishing that an algorithm $\mathcal{A}$ satisfies differential privacy is often a difficult task. In this section we present a new proof technique for deriving the privacy properties of $\mathcal{A}$ from a sequence of differentially private algorithms $\mathcal{A}_i$. The power of this technique is that we only require a very weak form of convergence; in fact, the limiting probabilistic behavior of the A_i can be quite different from the behavior of $\mathcal{A}$. Our results are summarized in the following theorem (see Appendix B for proof).

Theorem 1 (Successive Approximation) *Let b be a $\mathbb{R}^p$-valued random variable. Let $\mathcal{A}$ be a randomized algorithm induced by the random variable b and some deterministic function ϕ – that is, $\mathcal{A}(\mathcal{D}) \equiv \phi(\mathcal{D}, b)$. Let $\mathcal{A}_1, \mathcal{A}_2, \ldots$ be a sequence of randomized algorithms, where each $\mathcal{A}_i$ is induced by b and some deterministic function ϕ^i (i.e. $\mathcal{A}_i(\mathcal{D}) \equiv \phi^i(\mathcal{D}, b)$). If $\mathcal{A}_1, \mathcal{A}_2, \ldots$ are all (ϵ, δ)-differentially private and $\lim_{i\to\infty} \phi^i(\mathcal{D}, b) = \phi(\mathcal{D}, b)$ (i.e. pointwise convergence for all $\mathcal{D}$ and realized values of b), then $\mathcal{A}$ is also (ϵ, δ)-differentially private.*

It is important to note that differential privacy is a condition on $\Pr[\phi^i(\mathcal{D}, b) \in \mathcal{O}]$ (which is the same as $P(\mathcal{A}_i(\mathcal{D}) \in \mathcal{O})$) yet the pointwise convergence $\lim_{i\to\infty} \phi^i(\mathcal{D}, b) \to \phi(\mathcal{D}, b)$ required by Theorem 1 is too weak to guarantee that $\Pr[\phi^i(\mathcal{D}, b) \in \mathcal{O}] \to \Pr[\phi(\mathcal{D}, b) \in \mathcal{O}]$. In fact, the limiting probabilistic behavior (if it exists) of $\phi^i(\mathcal{D}, b)$ can be quite different from the probabilistic behavior of $\phi(\mathcal{D}, b)$. Nevertheless, Theorem 1 establishes that $\mathcal{A}$ still inherits differential privacy properties from the $\mathcal{A}_i$'s. Consider the following example:

Example 1 *Let $\theta \in \mathbb{R}^p$ be a parameter vector and let $\hat{\mathcal{L}}(\theta; \mathcal{D})$ be a strongly convex, twice continuously differentiable loss function. Let $\phi(\mathcal{D}, b) \equiv \operatorname{argmin}_\theta \hat{\mathcal{L}}(\theta; \mathcal{D}) + \frac{1}{n}(b^T\theta + \|\theta\|_1)$, which is an L_1-regularized minimization problem (with random perturbation $b^T\theta$). We can approximate it (see Appendix C.2) with a sequence $\phi^i(\mathcal{D}, b) \equiv \operatorname{argmin}_\theta \hat{\mathcal{L}}(\theta; \mathcal{D}) + \frac{1}{n}(b^T\theta + r_i(\theta))$ where r_i is an infinitely differentiable regularizer. If b has a continuous probability distribution, then for each fixed $\mathcal{D}$ the distribution of $\phi^i(\mathcal{D}, b)$ has a density (Chaudhuri et al. (2011)) but $\phi(\mathcal{D}, b)$ does not. In fact, the subdifferentials of the L_1 regularizer ensure that for each $\mathcal{D}$, $\phi(\mathcal{D}, b)$ can take values in a lower-dimensional submanifold of $\mathbb{R}^p$ with positive probability (which is not possible for the $\phi^i(\mathcal{D}, b)$ because of their densities).*

2.2. Application: Private Constrained Optimization

We use Theorem 1 to extend the applicability of the differentially private empirical risk minimization framework of Chaudhuri et al. (2011) to allow hard convex constraints and non-differentiable regularizers. Consider the convex program: $\operatorname{argmin}_{\theta\in\mathbb{F}} \hat{\mathcal{L}}(\theta; \mathcal{D}) + \frac{1}{n} r(\theta)$, where $\mathbb{F} \subseteq \mathbb{R}^p$ is a closed convex set, $\mathcal{D} = \{d_1, \ldots, d_n\}$ is a dataset, $\hat{\mathcal{L}}$ is a twice-continuously differentiable convex loss

Algorithm 1 Generalized Objective Perturbation Mechanism (Obj-Pert)

Require: dataset $\mathcal{D} = \{d_1, \ldots, d_n\}$, privacy parameters ϵ and δ ($\delta = 0$ for ϵ-differential privacy), convex regularizer r, a convex domain $\mathbb{F} \subseteq \mathbb{R}^p$, convex loss function $\hat{\mathcal{L}}(\theta; \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; d_i)$ with continuous Hessian, $\| \bigtriangledown \ell(\theta; d)\|_2 \leq \zeta$ (for all $d \in \mathcal{P}$ and $\theta \in \mathbb{F}$), and upper bound λ on the eigenvalues of $\bigtriangledown^2 \ell(\theta; d)$ (for all d and for all $\theta \in \mathbb{F}$).

1: Set $\Delta \geq \frac{2\lambda}{\epsilon}$.
2: **if** require ϵ-differential privacy **then**
3: sample $b \in \mathbb{R}^p$ from the Gamma distribution with density $\nu_1(b; \epsilon, \zeta) \propto e^{-\epsilon \frac{\|b\|_2}{2\zeta}}$
4: **else if** require (ϵ, δ)-differential privacy **then**
5: sample $b \in \mathbb{R}^p$ from $\nu_2(b; \epsilon, \delta, \zeta) = \mathcal{N}\left(0, \frac{\zeta^2(8 \log \frac{2}{\delta} + 4\epsilon)}{\epsilon^2} I_{p\times p}\right)$.
6: **end if**
7: **return** $\theta^{\text{priv}} \equiv \arg\min_{\theta \in \mathbb{F}} \hat{\mathcal{L}}(\theta; \mathcal{D}) + \frac{1}{n} r(\theta) + \frac{\Delta}{2n}\|\theta\|_2^2 + \frac{b^T\theta}{n}$.

function of the form $\hat{\mathcal{L}}(\theta; \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; d_i)$ and r is any (possibly non-differentiable) convex regularizer. When the objective function is γ/n-strongly convex ($\gamma \geq 0$) for all datasets of size n, one adds a quadratic term $\frac{(\Delta-\gamma)^+}{2n}\|\theta\|_2^2$, where Δ depends on the largest possible eigenvalue of the Hessian of $\ell(\theta, d_i)$. This ensures that the objective function is Δ/n-strongly convex and reduces the influence of any single data point. For privacy, a random linear perturbation term $\frac{b^T\theta}{n}$ is then added to the objective function. The full mechanism is described in Algorithm 1. Note that to simplify the discussion, we can w.l.o.g. assume $\gamma = 0$ (*i.e.,* the initial objective function is not strongly convex).

Theorem 2 (Private Convex Optimization via Objective Perturbation) *Let $\mathbb{F}$ be a closed convex subset of $\mathbb{R}^p$. Let $\mathcal{D} = \{d_1, \ldots, d_n\}$ be a dataset, let $\hat{\mathcal{L}}(\theta; \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; d_i)$ be a convex loss function with continuous Hessian, let ζ be the upper bound on $\| \bigtriangledown \ell(\theta; d)\|_2$ and let λ be an upper bound on the eigenvalues of $\bigtriangledown^2 \ell(\theta; d)$ (for all d and for all $\theta \in \mathbb{F}$), and let r be a convex function. Assume that for all $\theta \in \mathbb{F}$ and for all d the rank of $\bigtriangledown^2 \ell(\theta; d)$ is at most one.*

Then Algorithm 1 is $(\epsilon, 0)$-differentially private when b has gamma density ν_1 and (ϵ, δ)-differentially private when b has Gaussian density ν_2.

See Appendix C for the proof. The main idea is to use Theorem 1 twice. We first consider unconstrained optimization and convolve the regularizer r with a sequence $K_1, K_2, \ldots$ of infinitely differentiable kernels. This results in a sequence of smooth optimization problems that can be solved differentially privately by the results of Chaudhuri et al. (2011). We prove pointwise convergence of their differentially private solutions and then invoke Theorem 1. For constrained optimization, we replace the hard constraint $\theta \in \mathbb{F}$ with a sequence of soft constraints by adding penalties for $\theta \notin \mathbb{F}$ that depend on the distance from θ to $\mathbb{F}$. We again show pointwise convergence and invoke Theorem 1.

2.2.1. UTILITY ANALYSIS

The following lemma bounds the empirical risk (*i.e.,* $\hat{J}(\theta^{\text{priv}}; \mathcal{D}) - \hat{J}(\hat{\theta}; \mathcal{D})$) of Algorithm 1 (Algorithm Obj-Pert).

Lemma 3 (Empirical risk) *Let $\hat{\theta}$ be the minimizer of the empirical objective function $\hat{J}(\theta; \mathcal{D})$ over the closed convex set $\mathbb{F}$ and let θ^{priv} be the output of Algorithm 1. We have $\hat{J}(\theta^{priv}; \mathcal{D}) - \hat{J}(\hat{\theta}; \mathcal{D}) \leq \frac{2\|b\|_2^2}{\Delta n} + \frac{\Delta}{2n}\|\hat{\theta}\|_2^2$.*

The proof of this lemma can be found in Appendix D.1.1. Using the tail bounds for the noise distributions used in Algorithm Obj-Pert, we obtain the following theorem as a corollary of the above lemma. A detailed proof of this theorem is given Appendix D.1.2.

Theorem 4 (Theorem 26, special case) *Assume that $\| \bigtriangledown \ell(\theta; d)\|_2 \leq \zeta$ (for all $\theta \in \mathbb{F}$ and for all $d \in \mathcal{P}$). Let λ be the maximum eigenvalue bound on $\bigtriangledown^2 \ell$.*

1. *(Chaudhuri et al., 2011) With Gamma density ν_1, setting $\Delta = \Theta\left(\frac{\zeta p \log p}{\epsilon \|\hat{\theta}\|_2}\right)$ and assuming $\Delta \geq \frac{\lambda}{2\epsilon}$, we have $\mathbb{E}\left[\hat{J}(\theta^{priv}; \mathcal{D}) - \hat{J}(\hat{\theta}; \mathcal{D})\right] = O\left(\frac{\zeta \|\hat{\theta}\|_2 p \log p}{\epsilon n}\right)$.*
2. *(This paper) With Gaussian density ν_2, setting $\Delta = \Theta\left(\frac{\sqrt{\zeta^2 p \log(1/\delta)}}{\epsilon \|\hat{\theta}\|_2}\right)$ and assuming $\Delta \geq \frac{\lambda}{2\epsilon}$, we have $\mathbb{E}\left[\hat{J}(\theta^{priv}; \mathcal{D}) - \hat{J}(\hat{\theta}; \mathcal{D})\right] = O\left(\frac{\zeta \|\hat{\theta}\|_2 \sqrt{p \log(1/\delta)}}{\epsilon n}\right)$.*

Note that the empirical risk bounds in Theorem 4 are for the ideal choices of Δ. Optimal Δ depends on the L_2 norm of the true minimizer ($\hat{\theta}$) of $\hat{J}$. In practice, if the exact bound on $\|\hat{\theta}\|_2$ is not known, then one can replace it with a loose upper bound, *e.g.,* a bound on the diameter of the convex set $\mathbb{F}$.

The main takeaway from Theorem 4 is that, ignoring the privacy parameters (ϵ, δ), the empirical risk bound for the Gamma distribution (ν_1) is at least $\sqrt{p}$ times higher than for Gaussian distribution (ν_2). Intuitively, this gap arises from the fact that the vectors drawn from ν_2 are more tightly concentrated around the mean as compared to ν_1. For an application of the above theorem to linear regression, see Section 2.2.3.

For Generalized Linear Models (GLM), using a generic conversion theorem from empirical risk to generalization error (Shalev-Shwartz et al., 2009, Theorem 2), one can directly obtain a bound on the generalization error ($\bar{J}(\theta^{\text{priv}}; \mathcal{P}) - \bar{J}(\bar{\theta}; \mathcal{P})$). (See Appendix D.2).

In all utility guarantees in this paper, using the *Gamma* noise distribution results in an $\sqrt{p}$ increase in the error. So in the rest of our discussion, we will only concentrate on *Gaussian* noise distribution and hence guarantee (ϵ, δ)-differential privacy with $\delta > 0$.

2.2.2. REFINED UTILITY GUARANTEES UNDER STRONGER ASSUMPTIONS

In this section, we provide refined utility guarantees for Algorithm Obj-Pert (Algorithm 1) based on stronger assumptions on the underlying dataset. Our analysis is inspired by the work of Dwork et al. (2009) which specifically analyzes logistic regression under such a setting.

For the simplicity of exposition, assume the empirical objective function $\hat{J}(\theta; \mathcal{D})$ equals the empirical loss function $\hat{\mathcal{L}}(\theta; \mathcal{D})$, *i.e.,* the regularizer $r(\theta)$ is set to zero. Suppose the empirical loss function $\hat{\mathcal{L}}$ is (η/n)-strongly convex (for some constant η) within a ball of radius ψ (which will be fixed later) around $\hat{\theta}$, where $\hat{\theta}$ is the minimizer of $\hat{\mathcal{L}}$.

Theorem 5 bounds the empirical risk based on this stronger assumption on the loss function . In order to make the result more informative, we state a special case of Theorem 31 (Appendix E.2) below. In the following theorem we have assumed the following.

Assumption 1 *Assume: i)* $\eta = \Omega(\lambda n/p)$, *ii)* $\psi \geq \frac{p^{3/2}\zeta\sqrt{\log(1/\delta)}}{\lambda n\epsilon^2} + \sqrt{\frac{p}{n}}\|\hat{\theta}\|_2$, *iii)* $n \geq p^2$.

Intuitively, the assumption on η makes sense because if each of $\bigtriangledown^2\ell(\theta; d_i)$ is a rank-one matrix with an eigenvalue $\lambda > 0$ and the eigenvalues of $\bigtriangledown^2\hat{\mathcal{L}}$ are spread out across all dimensions, then we would expect $\sum_{i=1}^n \ell(\theta; d_i)$ to have minimum eigenvalue of Σ to be $\Omega(\lambda n/p)$ (since there are p dimensions).

Theorem 5 (Theorem 31, special case) *Let* $\Delta = 2\lambda/\epsilon$ *(where* λ *is the bound on the maximum eigenvalue of* $\bigtriangledown^2\ell$*). Under Assumption 1, using Gaussian density* ν_2*, we have* $\mathbb{E}\left[\hat{\mathcal{L}}(\theta^{priv}; \mathcal{D}) - \hat{\mathcal{L}}(\hat{\theta}; \mathcal{D})\right] = O\left(\frac{1}{n\epsilon}\left(\frac{p^2\zeta^2\log(1/\delta)}{\lambda n\epsilon} + \lambda\|\hat{\theta}\|_2^2\right)\right)$.

The proof is given in Appendix E.2. We now apply this theorem to linear regression to reduce the error bound by a factor of $\sqrt{p}$.

2.2.3. CASE STUDY: LINEAR REGRESSION

Consider the linear regression problem $y = X\theta^* + w$, where the design matrix X is in $\mathbb{R}^{n\times p}$, output vector y is in $\mathbb{R}^{n\times 1}$, parameter vector θ^* is in $\mathbb{R}^p$, and $w \in \mathbb{R}^{n\times 1}$ is a noise vector. We define the loss function for any given θ as $\hat{\mathcal{L}}(\theta; \mathcal{D}) = \frac{1}{2n}\sum_{i=1}^n (y_i - \langle X_i, \theta\rangle)^2$, where y_i is the i-th entry in the vector y and X_i is the i-th row of the matrix X. The setting we are interested in is where each row of the design matrix X has L_2 norm at most $\sqrt{p}$ and the parameter vector θ^* has L_2 norm at most $\sqrt{p}$. Under this setting we obtain the following empirical risk bounds (Table 1). (For a detailed discussion on the setting of parameters that lead to the following bounds, see Section H.)

Section	Theorem	Empirical risk (ignoring privacy parameters)
Section 2.2.1	Theorem 4 (Part 1)	$\tilde{O}(p^3/n)$
Section 2.2.1	Theorem 4 (Part 2)	$\tilde{O}(p^{5/2}/n)$
Section 2.2.2	Theorem 5	$\tilde{O}(p^2/n)$

Table 1: Empirical risk bounds for linear regression in the "small p, large n" regime.

3. Privacy preserving sparse regression

In sparse regression, we try to estimate $\theta^{\text{sp}} \in \arg\min_{\theta\in\mathbb{F}, \|\theta\|_0 \leq s} \hat{\mathcal{L}}(\theta; \mathcal{D})$, where $\mathbb{F}$ is a convex set and s is the sparsity parameter. We are typically interested in the setting where $s < n$ and $n \ll p$. In order to obtain a private estimate for θ^{sp}, we use the following two stage approach (Algorithm 2):

Algorithm 2 Meta-algorithm for sparse linear regression

1: Output $\hat{\Gamma}$, an estimate of the support for the parameter vector θ^{sp}.
2: Privately minimize the loss function $\hat{\mathcal{L}}(\theta; \mathcal{D})$ over the convex set $\mathbb{F}$ restricted to support $\hat{\Gamma}$ using Algorithm Obj-Pert (Algorithm 1).

We propose two algorithms which allow us to obtain good estimate for the support of the parameter vector θ^{sp}. The first algorithm (Algorithm Exp-mech) is based on the Exponential Mechanism

by McSherry and Talwar (2007). The second algorithm (Algorithm Samp-Agg) is based on the Sample and Aggregate Framework by Nissim et al. (2007).

These algorithms work under incomparable sets of assumptions. Roughly, Algorithm Exp-mech requires a bounded loss function, while Algorithm Samp-Agg works assumes that most random sub-samples of the dataset will be correctly labeled by parameter vectors that share a common small support.

To provide a comparison, we analyze the performance of these two algorithms on a class of widely studied linear regression problems (described in Section 3.1), which satisfy all these sets of assumptions. The algorithms and their performance guarantees are given in Sections 3.2 and 3.3. We often mention restricting the convex set $\mathbb{F}$ to some support Γ (which we represent by $\mathbb{F}_\Gamma$). By restriction we mean the set $\theta \in \mathbb{F}$ whose support lie in Γ: $\{\theta \in \mathbb{F} : \text{supp}(\theta) \subseteq \Gamma\}$.

3.1. Case Study: Sparse Linear Regression

To compare the performance of our two algorithms, we consider their performance on a class of "well-behaved" linear regression instances. This class is very similar to those used in the literature to analyze the LASSO and related non-private approaches to the sparse regression (see, e.g., Negahban et al. (2010)).

We look at the following linear system: $y = X\theta^* + w$, where the design matrix X is in $\mathbb{R}^{n\times p}$, output vector y is in $\mathbb{R}^{n\times 1}$, s-sparse parameter vector θ^* is in $\mathbb{R}^p$, and $w \in \mathbb{R}^{n\times 1}$ is a noise vector. We define the loss function for any given θ as $\hat{\mathcal{L}}(\theta; \mathcal{D}) = \frac{1}{2n}\|y - X\theta\|_2^2$.

In the following, we define what it means to be a "well-behaved" dataset. We use this definition to precisely state the assumptions on the problem.

Definition 2 ((s, σ, Ψ)-well behaved) *A pair (M, w), where M is a $n' \times p$ design matrix and w is a n'-dimensional vector, is* (s, σ, Ψ)-well behaved *if:*

1. $\forall i, \| M_i|_s \|_2 \leq \sqrt{s}$, *where $M_i|_s$ denotes the largest s entries of the i-th row of M.*
2. $\forall j, \|c_j\|_2 \leq \sqrt{n'}$, *where c_j is the j-th column of M.*
3. $\|M^T w\|_\infty \leq 2\sigma\sqrt{n' \log p}$.
4. ***Restricted Strong Convexity (RSC):*** *Given a set of indices $\Gamma \subset [p]$, let $C(\Gamma) = \{\theta \in \mathbb{R}^p : \|\theta_{\Gamma^c}\|_1 \leq 3\|\theta_\Gamma\|_1\}$. Here θ_Γ (respectively, θ_{Γ^c}) denotes θ restricted to entries in Γ (respectively, $\Gamma^c = [p] \setminus \Gamma$). We require that for all Γ of size $|\Gamma| = s$ and for all $\theta \in C(\Gamma)$: $\|M\theta\|_2^2 \geq n'\Psi\|\theta\|_2^2$.*

Remark: If w is i.i.d. sub-gaussian with mean zero and variance σ^2 and if the design matrix $X_{n\times p}$ is generated by sampling the rows i.i.d. from a Gaussian ensemble $\mathcal{N}(0, \Sigma)$, then under reasonable assumptions on n, s, p and Σ, the tuple (X, w) is (s, σ, Ψ)-well-behaved with high probability. (Roughly, a sufficient condition is that the eigenvalues of Σ lie strictly between 0 and 1 and that n grows at least as fast as $s \log p$.) See Negahban et al. (2010) for discussion and references.

The analysis of both our algorithms require the following assumption on the instance (X, y) of the linear regression problem:

Assumption 2 (Sparse-Linear) *We can write $y = M\theta^* + w$ where*

1. $\|\theta^*\|_\infty \leq 1$ *and* $\|\theta^*\|_0 \leq s$.

2. *All nonzero entries of θ^* have absolute value at least Φ.*
3. *The response vector $y \in [-s, s]^n$.*
4. *(X, w) is (s, σ, Ψ)-well behaved.*

The analysis of our second, efficient algorithm requires a slightly stronger requirement on the design matrix and noise. Specifically, our algorithm will partition the dataset into (roughly) $\sqrt{n}$ subsets of $\sqrt{n}$ points. We require that the design matrices and noise vectors for each of these sub-instances be well-behaved. Specifically:

Assumption 3 (Sparse-Linear') *In addition to Assumption 2 (Assumption* Sparse-Linear*),*

3'. *All pairs $(X_1, w_1), ..., (X_{\sqrt{n}}, w_{\sqrt{n}})$ are (s, σ, Ψ)-well-behaved, where (X_i, w_i) are formed by partitioning the rows of (X, w) into $\lceil \sqrt{n} \rceil$ disjoint blocks of $\sqrt{n} \pm 1$ points.*

Note that the assumptions on θ^* are identical in Assumptions Sparse-Linear and Sparse-Linear', and that the Assumption Sparse-Linear' is strictly stronger than Sparse-Linear.

From the above assumptions one can easily conclude that $|\langle x, \theta^* \rangle| \leq s$, where x is any row of the design matrix X. This means, if we truncate the responses $y_1, \cdots y_n$ (in the dataset $\mathcal{D}$) to have values in $[-s, s]$, then the utility of the algorithm will not worsen. Therefore, w.l.o.g. we assume that $y_1, \cdots, y_n$ lie in $[-s, s]$.

In order to compare the two support estimation algorithms, we compare the bounds on the dataset size n such that there are consistent estimates for the empirical risk $\hat{\mathcal{L}}(\theta^{\text{priv}}; \mathcal{D}) - \hat{\mathcal{L}}(\theta^*; \mathcal{D})$ as $n \to \infty$. These bounds are shown in Table 2.

Section	Algorithm	Bound on n	Running time $poly(s, n, p)$
Section 3.2	Exp-mech	$\omega(s^3 \log p)$	no
Section 3.3	Algorithm Samp-Agg	$\omega(s^2 \log^2 p)$	yes

Table 2: Bound on the dataset size n for consistent estimate of empirical risk.

We chose sparse linear regression as a case study because of the following two reasons. First, it demonstrates the use of our successive approximation tool which allows us to guarantee privacy for constrained optimization. In our privacy analysis we assumed that the convex set $\mathbb{F}$ is bounded in order to show that the minimizer of the regression problem does not change by much due to addition or removal of one data entry. Second, sparse linear regression demonstrates the effectiveness of our tighter utility analysis (Section 2.2.2). Using the tighter analysis, we obtained an $\sqrt{s}$ improvement in the utility guarantees.

3.2. Inefficient Feature Selection via Exponential Sampling

At a high level, the algorithm (Algorithm Exp-mech in Appendix F.1) finds a support $\hat{\Gamma}$ of size s such that restricted to this support, the minimum (non-private) loss is close the empirical loss incurred by the true minimizer θ^*. In order to find $\hat{\Gamma}$ privately, Algorithm Exp-mech uses the *exponential mechanism* by McSherry and Talwar (2007). Broadly speaking, the exponential mechanism first defines a score (or quality) function q for all possible outputs of the algorithm in the range space. (Algorithm Exp-mech defines the score function for any support Γ of size s as

$q(\Gamma; \mathcal{D}) = \min_{\theta \in \mathbb{F}_\Gamma} \sum_{i=1}^{n} \ell(\theta; d_i)$ and the range space as all possible supports of size s.) It then picks a support Γ of size s with probability proportional to $\exp(-\epsilon q(\Gamma; \mathcal{D})\alpha)$, where α is an upper bound on $|\ell(\theta; d)|$ for all d in domain $\mathcal{T}$ and for all $\theta \in \mathbb{F}$ restricted to a support of size at most s. It is important to realize that Algorithm Exp-mech may not be computationally efficient.

From the privacy analysis of exponential mechanism, it follows that Algorithm Exp-mech is ϵ-differentially private. The main step in the utility analysis of Algorithm Exp-mech is that a "good" support has high weight in the exponential sampling. Also the utility guarantee relies on the parameter α which essentially bounds the change in the score function for any support Γ when one entry is added or removed from the dataset $\mathcal{D}$. The following utility guarantee is proven in Appendix F.2.

Theorem 6 (Theorem 34, special case) *Assume that $|\ell(\theta; d)| \leq \alpha$ (for all $\theta \in \mathbb{F}_\Gamma$, for all $d \in \mathcal{T}$ and for all support Γ of size at most s). We have $\mathbb{E}\left[\hat{\mathcal{L}}(\phi; \mathcal{D}) - \hat{\mathcal{L}}(\theta^{sp}; \mathcal{D})\right] = O\left(\frac{\alpha s \log p}{\epsilon n}\right)$. Here $\phi = \arg\min_{\theta \in \mathbb{F}_{\hat{\Gamma}}} \hat{\mathcal{L}}(\theta; \mathcal{D})$ and $\hat{\Gamma}$ is the output of Algorithm* Exp-mech.

For linear regression, if we instantiate the first step of Algorithm 2 (Algorithm Meta-Alg) with the exponential sampling described above, for outputting support $\hat{\Gamma}$ while preserving $\epsilon/2$-differential privacy, and execute Algorithm Obj-Pert (Algorithm 1) in the second step with privacy parameters $(\epsilon/2, \delta)$, then we obtain an (ϵ, δ)-differentially private algorithm.

From Theorems 4 and 6, we directly obtain the utility guarantee for the current instantiation of Meta Algorithm 2. See Appendix F.3 for a detailed proof.

Theorem 7 (Theorem 35, special case) *Under Assumption 2 (Assumption* Sparse-Linear*), if we set $\Delta = \Theta(s/\epsilon)$, then we have $\mathbb{E}\left[\hat{\mathcal{L}}(\theta^{priv}; \mathcal{D}) - \hat{\mathcal{L}}(\theta^*; \mathcal{D})\right] = O\left(\frac{1}{n\epsilon}\left(\frac{s \log(1/\delta)}{n\epsilon\Psi} + s^3 \log p\right)\right)$.*

Assuming ϵ, δ and Ψ to be constants, empirical risk goes to zero as $n \to \infty$ as long as $n = \omega(s^3 \log p)$.

3.3. Efficient Feature Selection via Sample and Aggregate Framework

The efficient version of the feature selection algorithm (Algorithm Samp-Agg (Algorithm 5)) uses the *Sample and Aggregate framework* (SAF) by Nissim et al. (2007). At a high-level, in stage one SAF partitions the dataset into blocks $\mathcal{D}_1, \cdots, \mathcal{D}_k$ and executes some non-private algorithm $\mathcal{B}$ on each $\mathcal{D}_i$. In the second stage, it uses a private aggregation function to combine the output of all the k executions of the algorithm $\mathcal{B}$.

In the context of current discussion, SAF works as follows. First, the dataset $\mathcal{D}$ is partitioned into k blocks $\mathcal{D}_1, \cdots, \mathcal{D}_k$. Then a feature selection algorithm $\mathcal{A}_{\text{supp}}$ is run on each data block $\mathcal{D}_i$. Each execution of Algorithm $\mathcal{A}_{\text{supp}}$ is guaranteed to produce a vector V_i in $\{0, 1\}^p$ (where ones in the vector represent the elements in the support). In the second part, given $V_1, \cdots, V_k$, the aggregation function of SAF picks the top-s coordinates in terms of the average number of votes they received from $V_1, \cdots, V_k$. Since, this choice of top s coordinates cannot be private, controlled amount of noise is added to the average number of votes for each coordinate before selecting the top s (call this set $\hat{\Gamma}$).

The resulting algorithm (Samp-Agg) is ϵ-differentially private (see Appendix G.2).

Theorem 8 *Assume that all k executions of $\mathcal{A}_{supp}$ identify an underlying correct support $\hat{\Gamma}^*$ for each data block $\mathcal{D}_i$. With probability $\geq 1 - p\exp(-\frac{\epsilon k}{4s})$, the output set $\hat{\Gamma}$ equals $\hat{\Gamma}^*$.*

A proof of can be found in Appendix G.3. In the context of sparse linear regression, Algorithm Samp-Agg yields Algorithm 3 as an instantiation of Algorithm Meta-Alg (Algorithm 2). See Appendix G.4 for the detailed algorithm.

Algorithm 3 Sparse linear regression via Sample and Aggregate framework

1: Let $\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{2n}\|y - X\theta\|_2^2 + \frac{\Lambda}{n}\|\theta\|_1$. Define Algorithm $\mathcal{A}_{\text{supp}}$ as the algorithm which returns the top s coordinates of $\hat{\theta}$ (based on absolute value).
2: Run Algorithm Samp-Agg with privacy parameter $\epsilon/2$ and number of data blocks $k = \sqrt{n}$ to return support $\hat{\Gamma}$.
3: Normalize the design matrix $X_{n\times p}$ such that a vector of any s elements from each row has norm at most $\sqrt{s}$ and make sure the response vector y is in $[-s, s]^n$.
4: Set $\mathbb{F} = \{\theta \in \mathbb{R}^p : \|\theta\|_\infty \leq 1\}$. Using Algorithm Obj-Pert (Algorithm 1), minimize the loss function $\hat{\mathcal{L}}(\theta; \mathcal{D})$ over the convex set $\mathbb{F}_{\hat{\Gamma}}$ with privacy parameters $(\epsilon/2, \delta)$.

The above instantiation of Algorithm Meta-Alg is (ϵ, δ)-differentially private (see Theorem 38 in Appendix G.4). The proof of this follows directly from the privacy guarantees of Algorithms Obj-Pert and Samp-Agg . For utility, we get the following:

Theorem 9 (Theorem 44, special case) *Let* $\Lambda = \Theta\left(\sigma n^{1/4}\sqrt{\log p}\right)$ *and* $\Delta = \Theta\left(s/\epsilon\right)$. *Under Assumption 3 (Assumption* Sparse-Linear' *), if* $n \geq (\frac{16\sigma}{\Psi\Phi})^4 s^2 \log^2 p$, *then with probability* $\geq 1 - \left(p \exp\left(-\frac{\epsilon\sqrt{n}}{8s}\right)\right)$ *over the randomization of the support selection step,*

$$\mathbb{E}_b\left[\hat{\mathcal{L}}(\theta^{priv}; \mathcal{D}) - \hat{\mathcal{L}}(\theta^*; \mathcal{D})\right] = O\left(\frac{s^2}{n\epsilon}\left(\frac{s^2 \log(1/\delta)}{n\epsilon\Psi} + 1\right)\right)$$

Here b *is the noise vector in Algorithm* Obj-Pert *(Algorithm 1).*

It is interesting to note that the convergence rate does not have any dependence on the dimensionality p. From the failure probability it can be seen that one needs $n = \omega(s^2 \log^2 p)$ to obtain failure probability that goes down to zero as $n \to \infty$. Hence, it suffices to have $n = \omega(s^2 \log^2 p)$ for consistent empirical risk.

References

Raghav Bhaskar, Srivatsan Laxman, Adam Smith, and Abhradeep Thakurta. Discovering frequent patterns in sensitive data. In *KDD*, 2010.

Patrick Billingsley. *Probability and Measure*. John Wiley and Sons, New York, NY, USA, 1995. ISBN 9780471007104.

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

Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *J. Mach. Learn. Res.*, 8:203–226, December 2007.

Cynthia Dwork. Differential privacy. In *ICALP*, 2006.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *STOC*, 2009.

Cynthia Dwork and Moni Naor. On the difficulties of disclosure prevention, or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1), 2010.

Cynthia Dwork, Krishnaram Kenthapadi, Frank Mcsherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006b.

Cynthia Dwork, Parikshit Gopalan, Huijia Lin, Toniann Pitassi, Guy Rothblum, Adam Smith, and Sergey Yekhanin. An analysis of the Chaudhuri and Monteleoni algorithm. *Innovations in Computer Science (poster)*, 2009. Available as Dwork et al. (2012).

Cynthia Dwork, Parikshit Gopalan, Huijia Lin, Toniann Pitassi, Guy Rothblum, Adam Smith, and Sergey Yekhanin. An analysis of the Chaudhuri and Monteleoni algorithm. Technical Report NAS-TR-0156-2012, Network and Security Research Center, Pennsylvania State University, USA, February 2012.

Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. In *KDD*, 2008.

Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *SIGMOD*, 2011.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *FOCS*, 2007.

Sahand Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *CoRR*, abs/1010.2731, 2010.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, 2007.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2000. ISBN 0387987932.

Sofya Raskhodnikova and Adam Smith. Cse 598a, spring 2010: Algorithmic challenges in data privacy. www.cse.psu.edu/~asmith/privacy598/, 2010.

Aaron Roth. Cis 800/002, fall 2011: The algorithmic foundations of data privacy. http://www.cis.upenn.edu/~aaroth/courses/privacyF11.html, 2011.

Benjamin I. P. Rubinstein, Peter L. Bartlett, Ling Huang, and Nina Taft. Learning in a large function space: Privacy-preserving mechanisms for svm learning. *CoRR*, abs/0911.5708, 2009.

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

Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *STOC*, 2011.

David Wipf and Bhaskar Rao. L_0-norm minimization for basis selection. In *NIPS*, 2005.

A. H. Zemanian. *Distribution theory and transform analysis: an introduction to generalized functions, with applications*. Dover Publications, Inc., New York, NY, USA, 1987.

Appendix A. Notation

Let $\mathcal{D} = \langle d_1, \cdots, d_n \rangle$ be a dataset drawn from a domain of tuples $\mathcal{T}$. Let $\mathcal{P}$ be a distribution over the domain $\mathcal{T}$. Let $\hat{\mathcal{L}}$ be an empirical loss defined as $\hat{\mathcal{L}}(\theta; \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; d_i)$, where $\ell(\theta; d_i)$ is a positive real valued function which is convex in the first parameter $\theta \in \mathbb{R}^p$. We define the stochastic loss for a parameter vector θ over the distribution $\mathcal{P}$ as follows: $\bar{\mathcal{L}}(\theta; \mathcal{P}) = \mathbb{E}_{d\sim\mathcal{P}}[\ell(\theta; d)]$. Let $r : \mathbb{R}^p \to \mathbb{R}^+$ be a convex regularizer. We define the empirical objective function as $\hat{J}(\theta; \mathcal{D}) = \hat{\mathcal{L}}(\theta; \mathcal{D}) + \frac{1}{n} r(\theta)$. Similarly, we define the stochastic objective function as $\bar{J}(\theta; \mathcal{P}) = \bar{\mathcal{L}}(\theta; \mathcal{P}) + \frac{1}{n} r(\theta)$. To obtain differential privacy, we add a "noisy" term $\frac{b^T\theta}{n}$ (where b is a noise vector drawn from some appropriate distribution) and an L_2 penalty $\frac{\Delta}{2n}\|\theta\|_2^2$ to the objective function $\hat{J}(\theta; \mathcal{D})$ (see Algorithm 1).

We denote such an objective function as $J^{\text{priv}}(\theta, b; \mathcal{D}) = \hat{J}(\theta; \mathcal{D}) + \frac{\Delta}{2n}\|\theta\|_2^2 + \frac{1}{n} b^T\theta$. Since the term $\frac{\Delta}{2n}\|\theta\|_2^2$ becomes useful in our utility analysis too, we define $J^{\#}(\theta; \mathcal{D}) = \hat{J}(\theta; \mathcal{D}) + \frac{\Delta}{2n}\|\theta\|_2^2$ to segregate the noise term. See Table 3 for a summary.

Description	Objective Function	Minimizer
Sparse minimizer for $\theta \in \mathbb{F}$ and $\|\theta\|_0 \leq s$	$\hat{\mathcal{L}}(\theta; \mathcal{D})$	θ^{sp}
Empirical loss + regularizer $\frac{1}{n}r$	$\hat{J}(\theta; \mathcal{D})$	$\hat{\theta}$
Expected empirical loss + regularizer $\frac{1}{n}r$	$\bar{J}(\theta; \mathcal{P})$	$\bar{\theta}$
Private objective function ($\hat{J}(\theta; \mathcal{D}) + \frac{\Delta}{2n}\|\theta\|_2^2 + \frac{b^T\theta}{n}$)	$J^{\text{priv}}(\theta, b; \mathcal{D})$	θ^{priv}
$\hat{J}(\theta; \mathcal{D}) + \frac{\Delta}{2n}\|\theta\|_2^2$	$J^{\#}(\theta; \mathcal{D})$	$\theta^{\#}$

Table 3: Various objective functions and their corresponding minimizers over $\mathbb{F}$

Appendix B. Differential Privacy via Successive Approximation

Theorem 1 directly follows from the following lemma. In the following, we use the notation $\phi_{\mathcal{D}}$ and $\phi^i_{\mathcal{D}}$ in place of $\phi(\mathcal{D}, \cdot)$ and $\phi^i(\mathcal{D}, \cdot)$, respectively.

Lemma 10 *Let $\mathcal{D}$ and $\mathcal{D}'$ be two datasets. Let b be a random variable. Let $\phi^1_{\mathcal{D}}, \phi^2_{\mathcal{D}}, \ldots$ be a sequence of functions that converge pointwise to some function $\phi_{\mathcal{D}}$ (*i.e., $\lim_{i\to\infty} \phi^i_{\mathcal{D}}(b) = \phi_{\mathcal{D}}(b)$ *for all values of b). Similarly, let $\phi^1_{\mathcal{D}'}, \phi^2_{\mathcal{D}'}, \ldots$ be a sequence of functions that converge pointwise to some function $\phi_{\mathcal{D}'}$. Let $\mu^i_{\mathcal{D}}$ be the probability measure defined as $\mu^i_{\mathcal{D}}(E) \equiv \Pr[\phi^i_{\mathcal{D}}(b) \in E]$ and define $\mu^i_{\mathcal{D}'}$, $\mu_{\mathcal{D}}$, and $\mu_{\mathcal{D}'}$ similarly. Fix $\epsilon, \delta \geq 0$. If $\mu^i_{\mathcal{D}}(E) \leq e^{\epsilon}\mu^i_{\mathcal{D}'}(E) + \delta$ for all i, then $\mu_{\mathcal{D}}(E) \leq e^{\epsilon}\mu_{\mathcal{D}'}(E) + \delta$.*

Lemma 10 follows immediately from Claims 11, 12, and 13.

Claim 11 *Consider the σ-algebra of Borel subsets of $\mathbb{R}^p$. For any Borel set E, probability measure μ, and $\xi > 0$, there exists an open set A and a closed set B such that: $B \subseteq E \subseteq A$ and*

- $\mu(E) \leq \mu(A) \leq \mu(E) + \xi$
- $\mu(B) \leq \mu(E) \leq \mu(B) + \xi$

Proof We first prove the first condition relating E and A for the cases when E is closed and then when E is a Borel set (the case when E is open is trivial). Then we prove the condition relating E and B by reducing it to previous results.

Part 1: Closed sets E.
Suppose E is a closed subset of $\mathbb{R}^p$. For each $i = 1, 2, \ldots$, define $A^{(i)} = \{y : \inf_{x\in E} \|x-y\|_2 < 1/i\}$. Each $A^{(i)}$ is open since it is the union of open balls of radius $1/i$ around each point of E. Clearly, $E \subseteq A^{(i)}$ and for all i and $A^{(1)} \supseteq A^{(2)} \supseteq \ldots$. Also $E = \bigcap_{i=1}^{\infty} A^{(i)}$ because if a point $x \notin E$ then, since E is closed, the distance between x and E is non-zero and so one of the $A^{(i)}$ does not contain x. The downward continuity property (Billingsley, 1995) of probability measures now ensures that $\lim_{i\to\infty} \mu(A^{(i)}) = \mu(E)$. Thus given $\xi > 0$, there exists an i such that $\mu(A^{(i)}) \leq \mu(E) + \xi$ and $\mu(A^{(i)}) \geq \mu(E)$ because $E \subseteq A^{(i)}$.

Part 2: Borel sets E. Consider the algebra $\mathcal{G}$ consisting of all subsets of $\mathbb{R}^p$ that are (1) open, or (2) closed, or (3) the intersection of an open and a closed set, or (4) the union of an open and closed set. Note that $\mathbb{R}^p \in \mathcal{G}$ and that $\mathcal{G}$ is closed under complementation, finite union, and finite intersection. Given the values of $\mu(C)$ for all $C \in \mathcal{G}$, we can define the outer measure (Billingsley, 1995) μ^* on all subsets $F \subseteq \mathbb{R}^p$ as follows:

$$\mu^*(F) = \inf_{\substack{\{C_1, C_2, \ldots\} \subseteq \mathcal{G} \\ F \subseteq \bigcup C_i}} \sum_i \mu(C_i)$$

where the infimum is taken over all finite and countable collections of sets from $\mathcal{G}$ whose union contains F. Caratheodory's Extension Theorem (Billingsley, 1995) guarantees that $\mu(E) = \mu^*(E)$

for all Borel sets E. Thus for any $\xi > 0$, there exists a finite or countable collection $C_1, C_2, \ldots$ of sets in $\mathcal{G}$ such that $E \subseteq \bigcup_i C_i$ and $\mu(E) \leq \sum_i \mu(C_i) \leq \mu(E) + \xi/2$.

We now replace the C_i with slightly bigger open sets A_i. If C_i is open, then set $A_i = C_i$. If C_i is closed, then use the previous result to find an open set $A_i \supset C_i$ such that $\mu(A_i) \leq \mu(C_i) + \xi/2^{i+1}$. If C_i is the intersection of an open set $\mathcal{O}$ and a closed set H, then replace H with an open set $H' \supset H$ such that $\mu(H') \leq \mu(H) + \xi/2^{i+1}$ and set $A_i = \mathcal{O} \cap H'$. Note that $C_i \subset A_i$ and $\mu(A_i) \leq \mu(C_i \cup (H' \setminus H)) \leq \mu(C_i) + \xi/2^{i+1}$. Finally, if C_i is the union of an open set $\mathcal{O}$ and a close set H, then replace H with an open set $H' \supset H$ such that $\mu(H') \leq \mu(H) + \xi/2^{i+1}$ and set $A_i = \mathcal{O} \cup H'$. Note that $C_i \subset A_i$ and $\mu(A_i) \leq \mu(C_i \cup (H' \setminus H)) \leq \mu(C_i) + \xi/2^{i+1}$.

Set $A = \bigcup_i A_i$. Note that A is open. Then, since $E \subseteq A$,

$$\begin{aligned}
\mu(E) &\leq \mu(A) \leq \sum_i \mu(A_i) \leq \sum_i (\mu(C_i) + \xi 2^{-i-1}) \\
&\leq \xi/2 + \sum_i \mu(C_i) \leq \xi/2 + \mu(E) + \xi/2 \\
&= \mu(E) + \xi
\end{aligned}$$

Part 3: Approximating E from below.
To prove the second part of the theorem, pick an $\xi > 0$ and choose an open set $A \supseteq E^c$ (the complement of E) such that $\mu(E^c) \leq \mu(A) \leq \mu(E^c) + \xi$. Set $B = A^c$. Then B is closed, $B \subseteq E$, and

$$\begin{aligned}
\mu(E^c) &\leq \mu(B^c) \leq \mu(E^c) + \xi \\
&\Rightarrow 1 - \mu(E) \leq 1 - \mu(B) \leq 1 - \mu(E) + \xi \\
&\Rightarrow \mu(B) \leq \mu(E) \leq \mu(B) + \xi
\end{aligned}$$

■

The next result shows that pointwise convergence of the $\phi^i_{\mathcal{D}}$ allows us to upper bound $\Pr[\phi_{\mathcal{D}}(b) \in \mathcal{O}]$ when $\mathcal{O}$ is open and lower bound it when $\mathcal{O}$ is closed.

Claim 12 *Under the assumptions of Lemma 10, for every open set $A \subseteq \mathbb{R}^p$,*

$$\mu_{\mathcal{D}}(A) \leq \liminf_{i \to \infty} \mu^i_{\mathcal{D}}(A)\,.$$

For every closed set $B \subseteq \mathbb{R}^p$,

$$\mu_{\mathcal{D}}(B) \geq \limsup_{i \to \infty} \mu^i_{\mathcal{D}}(B)$$

Proof For any set C, we use the notation $1_{\{\phi_{\mathcal{D}}(b) \in C\}}(b)$ to be the indicator function that is 1 when $\phi_{\mathcal{D}}(b) \in C$ and 0 otherwise, and similarly for $1_{\{\phi^i_{\mathcal{D}}(b) \in C\}}(b)$. Let A be any open set. For any b such that $\phi_{\mathcal{D}}(b) \in A$, there is a bounded open set $\mathcal{O}$ so that $\phi_{\mathcal{D}}(b) \in \mathcal{O} \subseteq A$. Since $\phi^i_{\mathcal{D}}(b)$ converges to $\phi_D(b)$, this means that eventually $\phi^i_{\mathcal{D}}(b) \in \mathcal{O}$ and so $\phi^i_{\mathcal{D}}(b) \in A$. This means that for any b such that $\phi_{\mathcal{D}}(b) \in A$, the indicators $1_{\{\phi^i_{\mathcal{D}}(b) \in A\}}(b)$ converge to $1_{\{\phi_{\mathcal{D}}(b) \in A\}}$ as $i \to \infty$. For b such

that $\phi_{\mathcal{D}}(b) \notin A$, $1_{\{\phi^i_{\mathcal{D}}(b)\in A\}}(b) \geq 0 = 1_{\{\phi_{\mathcal{D}}(b)\in A\}}(b)$. Thus for all b, $\liminf_{i\to\infty} 1_{\{\phi^i_{\mathcal{D}}(b)\in A\}}(b) \geq 1_{\{\phi_{\mathcal{D}}(b)\in A\}}(b)$. By Fatou's Lemma (Billingsley, 1995),

$$\begin{aligned}
\mu_{\mathcal{D}}(A) &= \int 1_{\{\phi_{\mathcal{D}}(b)\in A\}}(b)\, d\mu(b) \\
&\leq \int \liminf_{i\to\infty} 1_{\{\phi^i_{\mathcal{D}}(b)\in A\}}(b)\, d\mu(b) \\
&\leq \liminf_{i\to\infty} \int 1_{\{\phi^i_{\mathcal{D}}(b)\in A\}}(b)\, d\mu(b) \\
&= \liminf_{i\to\infty} \mu^i_{\mathcal{D}}(A)
\end{aligned}$$

To show the second part, let B be a closed set. Consider its complement B^c, which is an open set. Using the previous result,

$$\begin{aligned}
\mu_{\mathcal{D}}(B^c) &\leq \liminf_{i\to\infty} \mu^i_{\mathcal{D}}(B^c) \\
&\Rightarrow 1 - \mu_{\mathcal{D}}(B) \leq \liminf_{i\to\infty} (1 - \mu^i_{\mathcal{D}}(B)) \\
&\Rightarrow \mu_{\mathcal{D}}(B) \geq -\liminf_{i\to\infty} -\mu^i_{\mathcal{D}}(B) \\
&\Rightarrow \mu_{\mathcal{D}}(B) \geq \limsup_{i\to\infty} \mu^i_{\mathcal{D}}(B)
\end{aligned}$$

■

The final result states that the upper bound and lower bound results of Claim 12 are all that we need.

Claim 13 *Let E be a Borel set and let $\mu_{\mathcal{D}}$, $\mu_{\mathcal{D}'}$, $\mu^i_{\mathcal{D}}$, and $\mu^i_{\mathcal{D}'}$ (for all i) be probability measures such that:*

1. *$\mu_{\mathcal{D}}(A) \leq \lim_{i\to\infty} \inf \mu^i_{\mathcal{D}}(A)$ for all open sets $A \subseteq \mathbb{R}^p$ and $\mu_{\mathcal{D}}(B) \geq \lim_{i\to\infty} \sup \mu^i_{\mathcal{D}}(B)$ for all closed sets $B \subseteq \mathbb{R}^p$.*
2. *$\mu_{\mathcal{D}'}(A) \leq \lim_{i\to\infty} \inf \mu^i_{\mathcal{D}'}(A)$ for all open sets $A \subseteq \mathbb{R}^p$ and $\mu_{\mathcal{D}'}(B) \geq \lim_{i\to\infty} \sup \mu^i_{\mathcal{D}'}(B)$ for all closed sets $B \subseteq \mathbb{R}^p$.*

For all $\epsilon, \delta \geq 0$, if $\mu^i_{\mathcal{D}}(E) \leq e^{\epsilon}\mu^i_{\mathcal{D}'}(E) + \delta$ for all i, then $\mu_{\mathcal{D}}(E) \leq e^{\epsilon}\mu_{\mathcal{D}'}(E) + \delta$.

Proof

Part 1: Reduction to open sets.
Let E be a Borel set and let $\mathcal{D}$ and $\mathcal{D}'$ be two datasets that differ by the addition or deletion of one tuple. Assume, by way of contradiction, that the (ϵ, δ)-differential privacy conditions do not hold so that $\mu_{\mathcal{D}}(E) \geq e^{\epsilon}\mu_{\mathcal{D}'}(E) + \delta + \alpha$ for some $\alpha > 0$. Using Claim 11, choose an open set $\mathcal{O} \supseteq E$ such that:

$$\mu_{\mathcal{D}'}(E) \leq \mu_{\mathcal{D}'}(\mathcal{O}) \leq \mu_{\mathcal{D}'}(E) + \frac{\alpha}{2e^{\epsilon}}$$

Therefore

$$\begin{aligned}\mu_{\mathcal{D}}(\mathcal{O}) &\geq \mu_{\mathcal{D}}(E) \geq e^{\epsilon}\mu_{D'}(E) + \delta + \alpha \\ &\geq e^{\epsilon}\left(\mu_{\mathcal{D}'}(\mathcal{O}) - \frac{\alpha}{2e^{\epsilon}}\right) + \delta + \alpha \\ &= e^{\epsilon}\mu_{D'}(\mathcal{O}) + \delta + \alpha/2\end{aligned}$$

and so $\mathcal{O}$ also violates the (ϵ, δ)-differential privacy constraints. Therefore, without loss of generality, we can assume that E is actually an open set.

Part 2: Proof for open sets. Let E be an open set that violates the (ϵ, δ)-differential privacy conditions such that $\mu_{\mathcal{D}}(E) \geq e^{\epsilon}\mu_{\mathcal{D}'}(E) + \delta + \alpha$ for some $\alpha > 0$. We will approximate E from below using both open sets and closed sets as follows. First, note that $E \neq \emptyset$ because the set $\emptyset$ can never violate the differential privacy conditions. Consider the open sets A_i and closed set B_i defined as follows:

$$\begin{aligned}A_i &= \{\theta' : \inf_{\theta \in E^c} \|\theta - \theta'\|_2 < 1/i\} \\ B_i &= \{\theta' : \inf_{\theta \in E^c} \|\theta - \theta'\|_2 \leq 1/i\}\end{aligned}$$

Note that $B_i = \overline{A_i}$ (B_i is the closure of A_i) and E^c is a subset of A_i and B_i for all $i \geq 1$. Now define the open set $\mathcal{O}_i \equiv B_i^c$ and note that $\overline{\mathcal{O}_i} = A_i^c$ and that $\mathcal{O}_i$ and $\overline{\mathcal{O}_i}$ are subsets of E. Finally, note that $\mathcal{O}_1 \subseteq \mathcal{O}_2 \subseteq \ldots$ and $\overline{\mathcal{O}_1} \subseteq \overline{\mathcal{O}_2} \subseteq \ldots$ and

$$\bigcup_{i=1}^{\infty} \mathcal{O}_i = E = \bigcup_{i=1}^{\infty} \overline{\mathcal{O}_i}$$

Now, by the upward continuity property of probability measures (Billingsley, 1995), there exists an i_0 such that for all $i \geq i_o$

$$\begin{aligned}\mu_{\mathcal{D}}(\mathcal{O}_i) &\leq \mu_{\mathcal{D}}(E) \leq \mu_{\mathcal{D}}(\mathcal{O}_i) + \frac{\alpha}{3} \\ \mu_{\mathcal{D}'}(\overline{\mathcal{O}_i}) &\leq \mu_{\mathcal{D}'}(E)\end{aligned}$$

Thus

$$\begin{aligned}\mu_{\mathcal{D}}(E) &\geq e^{\epsilon}\mu_{\mathcal{D}'}(E) + \delta + \alpha \\ &\Rightarrow \mu_{\mathcal{D}}(\mathcal{O}_i) + \frac{\alpha}{3} \geq e^{\epsilon}\mu_{\mathcal{D}'}(\overline{\mathcal{O}_i}) + \delta + \alpha \\ &\Rightarrow \mu_{\mathcal{D}}(\mathcal{O}_i) \geq e^{\epsilon}\mu_{\mathcal{D}'}(\overline{\mathcal{O}_i}) + \delta + \frac{2\alpha}{3}\end{aligned}$$

Then, using the $\liminf$ conditions on open sets and $\limsup$ conditions on closed sets,

$$\begin{aligned}\lim_{j\to\infty} \inf \mu_{\mathcal{D}}^{j}(\mathcal{O}_i) &\geq \mu_{\mathcal{D}}(\mathcal{O}_i) \\ &\geq e^{\epsilon}\mu_{\mathcal{D}'}(\overline{\mathcal{O}_i}) + \delta + \frac{2\alpha}{3} \\ &\geq \lim_{j\to\infty} \sup e^{\epsilon}\mu_{\mathcal{D}'}^{j}(\overline{\mathcal{O}_i}) + \delta + \frac{2\alpha}{3}\end{aligned}$$

Now, since $\mu^j_{\mathcal{D}}(\mathcal{O}_i) \leq \mu^j_{\mathcal{D}}(\overline{\mathcal{O}_i})$:

$$\lim_{j\to\infty} \inf \mu^j_{\mathcal{D}}(\overline{\mathcal{O}_i}) \quad \geq \quad \lim_{j\to\infty} \sup e^{\epsilon} \mu^j_{\mathcal{D}'}(\overline{\mathcal{O}_i}) + \delta + \frac{2\alpha}{3}$$

and so for some j

$$\mu^j_{\mathcal{D}}(\overline{\mathcal{O}_i}) \quad \geq \quad e^{\epsilon} \mu^j_{\mathcal{D}'}(\overline{\mathcal{O}_i}) + \delta + \frac{\alpha}{3}$$

However, this contradicts the fact that the pair of measures $\mu^j_{\mathcal{D}}$, $\mu^j_{\mathcal{D}'}$ satisfy the (ϵ, δ)-differential privacy conditions ($\mu^j_{\mathcal{D}}(\overline{\mathcal{O}_i}) \leq e^{\epsilon}\mu^j_{\mathcal{D}'}(\overline{\mathcal{O}_i}) + \delta$). Therefore E cannot violate the (ϵ, δ)-differential privacy conditions for the measures $\mu_{\mathcal{D}}$ and $\mu_{\mathcal{D}'}$.

■

Appendix C. Appendix: Differential Privacy and Convex Optimization

In order to prove Theorem 2, the starting point is Lemma 14 (Section C.1) which proves differential privacy for the special cases of Algorithm 1 where the regularizer r is twice continuously differentiable and the convex set $\mathbb{F}$ over which we optimize is the entire real space $\mathbb{R}^p$. Afterwards, we will use our *successive approximation* technique to remove these assumptions one-by-one (Sections C.2 and C.3).

C.1. Private Smooth Unconstrained Optimization

Lemma 14 (Differentially Private Smooth Unconstrained Objective Perturbation) *Under the conditions of Theorem 2, if we assume that the convex regularizer r is twice-continuously differentiable, the convex set $\mathbb{F}$ is the entire real space $\mathbb{R}^p$, then*

1. *(Chaudhuri et al., 2011)with Gamma density ν_1 in Algorithm 1 (Algorithm* Obj-Pert *) guarantees ϵ-differential privacy.*
2. *(This paper) with Gaussian density ν_2 in Algorithm 1 (Algorithm* Obj-Pert *) guarantees (ϵ, δ)-differential privacy.*

Proof The first part of Lemma 14 follows directly from Chaudhuri et al. (2011) and hence omitted here. The proof of the second part of Lemma 14 is as follows.

If we want to prove that Algorithm 1 satisfies (ϵ, δ)-privacy, it suffices to show that for all $\alpha \in \mathbb{R}^p$ the following is true.

$$e^{-\epsilon}(\text{pdf}(\theta^{\text{priv}} = \alpha; \mathcal{D}') - \delta) \leq \text{pdf}(\theta^{\text{priv}} = \alpha; \mathcal{D}) \leq e^{\epsilon}\text{pdf}(\theta^{\text{priv}} = \alpha; \mathcal{D}') + \delta \tag{2}$$

First consider an $\alpha \in \mathbb{R}^p$. If we have $\theta^{\text{priv}} = \alpha$, then it means that $\alpha = \arg\min_{\theta\in\mathbb{R}^p} n\hat{\mathcal{L}}(\theta; \mathcal{D}) + r(\theta) + \frac{\Delta}{2}\|\theta\|_2^2 + b^T\theta$. Setting the gradient of the objective function to zero we get the following.

$$b(\alpha; \mathcal{D}) = -\left(n \bigtriangledown \hat{\mathcal{L}}(\alpha; \mathcal{D}) + \bigtriangledown r(\alpha) + \Delta\alpha\right) \tag{3}$$

We have $\frac{\text{pdf}_{\mathcal{D}}(\theta^{\text{priv}}=\alpha)}{\text{pdf}_{\mathcal{D}'}(\theta^{\text{priv}}=\alpha)} = \frac{\nu_2(b(\alpha;\mathcal{D});\epsilon,\delta,\zeta)}{\nu_2(b(\alpha;\mathcal{D}');\epsilon,\delta,\zeta)} \frac{|\det(\bigtriangledown b(\alpha;\mathcal{D}'))|}{|\det(\bigtriangledown b(\alpha;\mathcal{D}))|}$. We bound the ratios of the densities ν_2 and the determinants separately.

First, we show that for all $\alpha \in \mathbb{R}^p$, $e^{-\epsilon} \leq \frac{|\det(\bigtriangledown b(\alpha;\mathcal{D}'))|}{|\det(\bigtriangledown b(\alpha;\mathcal{D}))|} \leq e^{\epsilon}$. The following lemma would be helpful in bounding the ratio.

Lemma 15 (Chaudhuri et al. (2011)) *If A is a full-rank matrix and if E is matrix with rank at most 2, then,*

$$\frac{det(A+E) - \det(A)}{\det(A)} = \lambda_1(A^{-1}E) + \lambda_2(A^{-1}E) + \lambda_1(A^{-1}E)\lambda_2(A^{-1}E)$$

where $\lambda_i(Z)$ is the i-th highest eigenvalue of matrix Z.

Let $A = \bigtriangledown b(\alpha;\mathcal{D}) = -(n \bigtriangledown^2 \hat{\mathcal{L}}(\alpha;\mathcal{D}) + \bigtriangledown^2 r(\alpha) + \Delta \mathbb{I}_p)$, where $\mathbb{I}_p$ is an identity matrix of $p \times p$ dimensions. W.l.o.g. assume that $\mathcal{D}'$ has one entry more as compared to $\mathcal{D}$, and $\mathcal{D}$ has n entries. Let $E = \bigtriangledown^2 \ell(\alpha; d_{n+1})$. Therefore, $|\det(\bigtriangledown b(\alpha;\mathcal{D}'))| = \det(A+E)$. Since $n \bigtriangledown^2 \hat{\mathcal{L}}(\alpha;\mathcal{D}) + \bigtriangledown^2 r(\alpha)$ is positive semi-definite (as both $\hat{\mathcal{L}}$ and r are convex), the smallest eigenvalue of A is Δ. Since E is a positive semi-definite matrix of rank at most one, $A^{-1}E$ has at most one non-zero eigenvalue. Additionally, it follows that $\lambda_1(A^{-1}E) \leq \frac{\lambda_1(E)}{\Delta}$. Applying Lemma 15, we have $\frac{det(A+E)}{det(A)} \leq 1 + \frac{\psi}{\Delta}$, since $\lambda_1(E) \leq \psi$ by assumption. Replacing the value of Δ we get $\frac{|\det(\bigtriangledown b(\alpha;\mathcal{D}'))|}{|\det(\bigtriangledown b(\alpha;\mathcal{D}))|} \leq e^{\frac{\epsilon}{2}}$.

To bound $\frac{\nu_2(b(\alpha;\mathcal{D});\epsilon,\delta,\zeta)}{\nu_2(b(\alpha;\mathcal{D}');\epsilon,\delta,\zeta)}$, recall that the noise vector b is drawn from the Gaussian distribution $\mathcal{N}(0, \beta^2 \mathbb{I}_p)$, where $\beta = \frac{\zeta\sqrt{8\log\frac{2}{\delta}+4\epsilon}}{\epsilon}$ is the standard deviation. Let us assume $\Gamma = b(\alpha;\mathcal{D}) - b(\alpha;\mathcal{D}')$. With this we have the following:

$$\begin{aligned}
\frac{\nu_2(b(\alpha;\mathcal{D});\epsilon,\delta,\zeta)}{\nu_2(b(\alpha;\mathcal{D}');\epsilon,\delta,\zeta)} &= \frac{e^{-\frac{\|b(\alpha;\mathcal{D})\|_2^2}{2\beta^2}}}{e^{-\frac{\|b(\alpha;\mathcal{D}')\|_2^2}{2\beta^2}}} \\
&= e^{\frac{1}{2\beta^2}\left|\|b(\alpha;\mathcal{D})\|_2^2 - \|b(\alpha;\mathcal{D}')\|_2^2\right|} \\
&= e^{\frac{1}{2\beta^2}\left|\|b(\alpha;\mathcal{D})\|_2^2 - \|b(\alpha;\mathcal{D}) - \Gamma\|_2^2\right|} \\
&= e^{\frac{1}{2\beta^2}\left|2\langle b(\alpha;\mathcal{D}),\Gamma\rangle - \|\Gamma\|_2^2\right|}
\end{aligned}$$

Since $\| \bigtriangledown \ell(\theta;\dot{}) \|_2 \leq \zeta$ for all $\theta \in \mathbb{R}^p$ and for all $d \in \mathcal{T}$, therefore $\|\Gamma\|_2 \leq \zeta$. Hence the following is true.

$$e^{\frac{1}{2\beta^2}\left|2\langle b(\alpha;\mathcal{D}),\Gamma\rangle - \|\Gamma\|_2^2\right|} \leq e^{\frac{1}{2\beta^2}\left(|2\langle b(\alpha;\mathcal{D}),\Gamma\rangle| + \|\Gamma\|_2^2\right)} \leq e^{\frac{1}{2\beta^2}\left(|2\langle b(\alpha;\mathcal{D}),\Gamma\rangle| + \zeta^2\right)} \tag{4}$$

The following two lemmas will be useful in bounding $|\langle b(\alpha;\mathcal{D}),\Gamma\rangle|$. Both of them follow from basic probability theory and hence we skip their proofs.

Lemma 16 *Let $Z \sim \mathcal{N}(0, \mathbb{I}_p)$ and $v \in \mathbb{R}^p$ be a fixed vector. Then*

$$\langle Z, v\rangle \sim \mathcal{N}(0, \|v\|_2^2)$$

Note that Γ is independent of the noise vector. Therefore using Lemma 16, we get $\langle b(\alpha;\mathcal{D}),\Gamma\rangle \sim \mathcal{N}(,\|\Gamma\|_2^2\beta^2)$. The following lemma provides a tail bound for normal distribution which we use to bound the probability that the noise vector $b(\alpha;\mathcal{D})$ is not in the set GOOD .

Lemma 17 *Let $Z \sim \mathcal{N}(0,1)$, then for all $t > 1$, we have*

$$\Pr[|Z| > t] \leq e^{-t^2/2}$$

Using this lemma and the fact that $\|\Gamma\|_2 \leq \zeta$, we get $\Pr[|\langle b(\alpha;\mathcal{D}),\Gamma\rangle| \geq \zeta\beta t] \leq e^{-\frac{t^2}{2}}$, where $t > 1$. Let GOOD be the set $\{a \in \mathbb{R}^p | \langle a,\Gamma\rangle| \geq \zeta\beta t\}$. We want the noise vector $b(\alpha;\mathcal{D})$ to be in the set GOOD w.p. at least $1-\delta$. Setting $t = \sqrt{2\log\frac{2}{\delta}}$ implies that $2e^{-\frac{t^2}{2}} = \delta$. To make sure $t \geq 1$, we need to have $\delta \leq \frac{2}{\sqrt{e}}$. This always true for any non trivial δ. Replacing $t = \sqrt{2\log\frac{2}{\delta}}$ in $\zeta\beta t$, we get from Equation 4 that $\frac{\nu_2(b(\alpha;\mathcal{D});\epsilon,\delta,\zeta)}{\nu_2(b(\alpha;\mathcal{D}');\epsilon,\delta,\zeta)} \leq e^{\frac{1}{2\beta^2}\left(\beta\zeta\sqrt{8\log\frac{2}{\delta}}+\zeta^2\right)}$. Solving for β we get $\beta \geq \frac{\zeta\sqrt{8\log\frac{2}{\delta}+4\epsilon}}{\epsilon}$. To complete the argument, we show the following:

$$\begin{aligned}
&\text{pdf}(\theta^{\text{priv}} = \alpha;\mathcal{D}) = \Pr[b \in \text{GOOD}\,]\text{pdf}(\theta^{\text{priv}} = \alpha | b \in \text{GOOD}\,;\mathcal{D}) \\
&+ \Pr[b \in \overline{\text{GOOD}}\,]\text{pdf}(\theta^{\text{priv}} = \alpha | b \in \overline{\text{GOOD}}\,;\mathcal{D}) \\
&\leq \Pr[b \in \text{GOOD}\,]\text{pdf}(\theta^{\text{priv}} = \alpha | b \in \text{GOOD}\,;\mathcal{D}) + \delta \\
&\leq e^{\epsilon}\Pr[b \in \text{GOOD}\,]\text{pdf}(\theta^{\text{priv}} = \alpha | b \in \text{GOOD}\,;\mathcal{D}') + \delta \\
&\leq e^{\epsilon}\text{pdf}(\theta^{\text{priv}} = \alpha;\mathcal{D}') + \delta
\end{aligned}$$

where b is the noise vector in Algorithm 1. This concludes the proof of Lemma 14. ■

C.2. Extension to non-differentiable regularizers via Successive Approximation

Our first goal is to remove the differentiability assumptions on the regularizer r. To do this, we use the *bump function* (Zemanian, 1987) $\Psi(x) : \mathbb{R} \to \mathbb{R}$ and a sequence of kernel functions $K_i(\theta) : \mathbb{R}^p \to \mathbb{R}$ (for $i = 1, 2, \dots$) defined as:

$$\begin{aligned}
\Psi(x) &= \begin{cases} \exp(-\frac{1}{1-x^2}) & \text{if } |x| < 1 \\ 0 & \text{if } |x| \geq 1 \end{cases} \\
K_i(\theta) &= \frac{\Psi(i\|\theta\|_2^2)}{\int\limits_{\theta' \in \mathbb{R}^p} \Psi(i\|\theta'\|_2^2)\, d\theta'}
\end{aligned} \tag{5}$$

The bump function Ψ is infinitely differentiable and all of its derivatives vanish outside the interval $(-1,1)$ (Zemanian, 1987). Therefore the kernels K_i are also infinitely differentiable and their support (and that of their derivatives) is $\left\{\theta \,:\, \|\theta\|_2^2 < 1/i\right\}$. Now, if r is a convex regularizer

(but not necessarily differentiable), then consider the regularizer r_i defined as the convolution of r and K_i:

$$r_i(\theta) = [r * K_i](\theta) \equiv \int_{y \in \mathbb{R}^p} r(\theta - y) K_i(y) dy$$

By the elementary properties of convolution and the smoothness of K_i, the regularizer r_i is infinitely differentiable. Since convolution with K_i is the same as an (infinite) positive linear combination of translations of r, the regularizer r_i is also convex. Thus we will approximate the objective function $J^{\text{priv}}(\theta, b; \mathcal{D}) = \hat{\mathcal{L}}(\theta; \mathcal{D}) + \frac{\Delta}{2n}\|\theta\|_2^2 + \frac{1}{n}(b^T\theta + r(\theta))$ with $J^{\text{priv}^i}(\theta, b; \mathcal{D}) = \hat{\mathcal{L}}(\theta; \mathcal{D}) + \frac{\Delta}{2n}\|\theta\|_2^2 + \frac{1}{n}(b^T\theta + r_i(\theta))$, which has a smooth regularizer and to which Lemma 14 can be applied. In the next lemma we show that the minimizers of J^{priv} and J^{priv^i} converge pointwise. This will enable us to invoke the successive approximations proof technique for guaranteeing privacy via Lemma 10.

Lemma 18 (Unconstrained Pointwise Convergence) *Let $\hat{\mathcal{L}}$ be a convex function and r a convex regularizer. Define the kernel function K_i as in Equation 5 and let $r_i(\theta) = [r * K_i](\theta)$ be the convolution between r and K_i. Define the objective function $J^{priv}(\theta, b; \mathcal{D}) = \hat{\mathcal{L}}(\theta; \mathcal{D}) + \frac{\Delta}{2n}\|\theta\|_2^2 + \frac{1}{n}(b^T\theta + r(\theta))$ and $J^{priv^i}(\theta, b; \mathcal{D}) = \hat{\mathcal{L}}(\theta; \mathcal{D}) + \frac{\Delta}{2n}\|\theta\|_2^2 + \frac{1}{n}(b^T\theta + r^i(\theta))$. Define the unconstrained minimizers, for each b, as $\phi_{\mathcal{D}}(b) = \operatorname{argmin}_{\theta \in \mathbb{R}^p} J^{priv}(\theta, b; \mathcal{D})$ and $\phi^i_{\mathcal{D}}(b) = \operatorname{argmin}_{\theta \in \mathbb{R}^p} J^{priv^i}(\theta, b; \mathcal{D})$. Then for every $b \in \mathbb{R}^p$, $\lim_{i \to \infty} \phi^i_{\mathcal{D}}(b) = \phi_{\mathcal{D}}(b)$.*

Proof In order to prove the pointwise convergence of the sequence of functions $\phi^i_{\mathcal{D}}$ to $\phi_{\mathcal{D}}$, we first prove the following claim.

Claim 19 *Let $\mathcal{I} \subseteq \mathbb{R}^p$ be a bounded set and let $B \subseteq \mathbb{R}^p$ be any set. The functions $n \cdot J^{priv^i}$ converge uniformly to $n \cdot J^{priv}$ on $\mathcal{I} \times B$ as $i \to \infty$.*

Proof Choose a $\xi > 0$. Let $\mathcal{I}' = \{y : \inf_{\boldsymbol{x} \in \mathcal{I}} ||y - \boldsymbol{x}||_2 \leq 1\}$ be the set of all points whose distance to $\mathcal{I}$ is at most 1. Note that $\mathcal{I}'$ is closed, bounded, and hence compact. Since $r(\theta)$ is a continuous function defined over the compact set $\mathcal{I}'$, it is then also uniformly continuous on $\mathcal{I}'$. This means that there exists an η (depending only on ξ) such that $|r(\theta_1) - r(\theta_2)| \leq \xi$ whenever $\theta_1, \theta_2 \in \mathcal{I}'$ and $||\theta_1 - \theta_2||_2 \leq \eta$. Now, for any $i > 1/\xi$ and any $\theta \in \mathcal{I}$ and $b \in B$,

$$\begin{aligned}
n|J^{\text{priv}^i}(\theta, b; \mathcal{D}) - J^{\text{priv}}(\theta, b; \mathcal{D})| &= |r_i(\theta) - r(\theta)| \\
&= \left| \int r(\theta - y) K_i(y)\, dy - r(\theta) \right| \\
&= \left| \int [r(\theta - y) - r(\theta)] K_i(y)\, dy \right| \quad \text{(Since the integral of } K_i \text{ is 1)} \\
&\leq \int \left| r(\theta - y) - r(\theta) \right| K_i(y)\, dy
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\{y\,:\,||y||_2^2 \le 1/i\}} \Big| r(\theta - y) - r(\theta) \Big| K_i(y)\, dy \quad \text{(The support of } K_i\text{)} \\
&\le \int_{\{y\,:\,||y||_2^2 \le 1/i\}} \xi K_i(y)\, dy \quad \text{(Since } \theta \in \mathcal{I},\ \theta - y \in \mathcal{I}',\ \text{and } ||y||_2 \le 1/i \le \xi\text{)} \\
&= \xi \quad \text{(Integral of } K_i \text{ over its support is 1)}
\end{aligned}
$$

Thus $n \cdot {J^{\text{priv}}}^i$ converge uniformly to $n \cdot J^{\text{priv}}$ on $\mathcal{I} \times B$. ■

Now with the uniform conververgence of the objective function in hand we use the following steps to complete the proof for Lemma 18.

Step 1: Properties of J^{priv}

In order to prove pointwise convergence, we first establish some simple properties of $J^{\text{priv}}(\theta, b; \mathcal{D})$, which is $\frac{\Delta}{n}$-strongly convex in θ for each fixed b. Recall that for each b, $\phi_{\mathcal{D}}(b)$ returns the unique θ that minimizes $J^{\text{priv}}(\theta, b; \mathcal{D})$ over $\mathbb{R}^p$ (uniqueness is guaranteed by strong convexity). By definition of Δ-strong convexity,

$$
J^{\text{priv}}(t\theta_1 + (1-t)\theta_2, b; \mathcal{D}) \quad \le \quad tJ^{\text{priv}}(\theta_1, b; \mathcal{D}) + (1-t)J^{\text{priv}}(\theta_2, b; \mathcal{D}) - \frac{\Delta}{2n} t(1-t)\|\theta_1 - \theta_2\|_2^2
$$

So for all θ and $t \in (0,1)$, recalling that $\phi_{\mathcal{D}}(b)$ is the minimizer of $J^{\text{priv}}(\cdot, b; \mathcal{D})$ over $\mathbb{R}^p$,

$$
\begin{aligned}
J^{\text{priv}}(\phi_{\mathcal{D}}(b), b; \mathcal{D}) &\le J^{\text{priv}}(t\phi_{\mathcal{D}}(b) + (1-t)\theta, b; \mathcal{D}) \\
&\le tJ^{\text{priv}}(\phi_{\mathcal{D}}(b), b; \mathcal{D}) + (1-t)J^{\text{priv}}(\theta, b; \mathcal{D}) - \frac{\Delta}{2n} t(1-t)\|\phi_{\mathcal{D}}(b) - \theta\|_2^2 \\
(1-t)J^{\text{priv}}(\phi_{\mathcal{D}}(b), b; \mathcal{D}) &\le (1-t)J^{\text{priv}}(\theta, b; \mathcal{D}) - \frac{\Delta}{2n} t(1-t)\|\phi_{\mathcal{D}}(b) - \theta\|_2^2 \\
\frac{\Delta}{2n} t\|\phi_{\mathcal{D}}(b) - \theta\|_2^2 &\le J^{\text{priv}}(\theta, b; \mathcal{D}) - J^{\text{priv}}(\phi_{\mathcal{D}}(b), b; \mathcal{D}) \\
\frac{\Delta}{2n}\|\phi_{\mathcal{D}}(b) - \theta\|_2^2 &\le J^{\text{priv}}(\theta, b; \mathcal{D}) - J^{\text{priv}}(\phi_{\mathcal{D}}(b), b; \mathcal{D}) \qquad (6)
\end{aligned}
$$

Where the last inequality follows by taking limits as $t \to 1$.

Step 2: Choosing Parameters

Choose any b. Now choose a small ξ such that $\Delta/2 > \xi > 0$. Define $\mathcal{I} = \{\theta\ :\ \|\theta - \phi_{\mathcal{D}}(b)\|_2 \le 1\}$ and the corresponding set $B = \{b'\ :\ \phi_{\mathcal{D}}(b') \in \mathcal{I}\}$. Since $\mathcal{I}$ is bounded, we can use the uniform convergence of the $n \cdot {J^{\text{priv}}}^i$ to $n \cdot J^{\text{priv}}$ over $\mathcal{I} \times B$ (from Claim 19). Choose an i_ξ depending only on ξ such that for all $i \ge i_\xi$, $\theta \in \mathcal{I}$, and $b' \in B$ the inequality $n|J^{\text{priv}}(\theta, b'; \mathcal{D}) - {J^{\text{priv}}}^i(\theta, b'; \mathcal{D})| \le \frac{\xi}{3}$ holds.

Step 3: Pointwise Convergence

We now show that $\|\phi^i_{\mathcal{D}}(b) - \phi_{\mathcal{D}}(b)\|_2 \le \sqrt{4\xi/\Delta}$ for all $i \ge i_\xi$. Assume, by way of contradiction, that $\|\phi^i_{\mathcal{D}}(b) - \phi_{\mathcal{D}}(b)\|_2 > \sqrt{4\xi/\Delta}$ for some $i \ge i_\xi$ and $b \in B$. Then, by the strong convexity of ${J^{\text{priv}}}^i$ (in terms of the parameter θ), there is a θ' along the line from $\phi^i_{\mathcal{D}}(b)$ to $\phi_{\mathcal{D}}(b)$ such that

$$
{J^{\text{priv}}}^i(\phi^i_{\mathcal{D}}(b), b; \mathcal{D}) < {J^{\text{priv}}}^i(\theta', b; \mathcal{D}) < {J^{\text{priv}}}^i(\phi_{\mathcal{D}}(b), b; \mathcal{D}) \qquad (7)
$$

$$
\text{and} \quad \|\theta' - \phi_{\mathcal{D}}(b)\|_2 = \sqrt{2\xi/\Delta} < 1 \quad \text{(since we chose } \xi < \Delta/2\text{)} \qquad (8)
$$

Since $\theta' \in \mathbb{R}^p$, so by Equation 8, $\theta' \in \mathcal{I}$. Now, by Equation 6,

$$\begin{aligned}
\xi &= \frac{\Delta}{2}\|\phi_{\mathcal{D}}(b) - \theta'\|_2^2 \leq n|{J^{\text{priv}}}_{\mathcal{D}}(\theta', b) - {J^{\text{priv}}}_{\mathcal{D}}(\phi_{\mathcal{D}}(b), b)| \\
&\leq n \cdot {J^{\text{priv}}}^i_{\mathcal{D}}(\theta', b) + \frac{\xi}{3} - n \cdot {J^{\text{priv}}}^i_{\mathcal{D}}(\phi_{\mathcal{D}}(b), b) + \frac{\xi}{3} \quad \text{(By uniform convergence on } \mathcal{I} \times B\text{)} \\
&\Rightarrow n \cdot {J^{\text{priv}}}^i(\theta', b; \mathcal{D}) \geq n \cdot {J^{\text{priv}}}^i(\phi_{\mathcal{D}}(b), b; \mathcal{D}) + \frac{\xi}{3}
\end{aligned}$$

This contradicts the fact that θ' was chosen to satisfy ${J^{\text{priv}}}^i(\theta', b; \mathcal{D}) < {J^{\text{priv}}}^i(\phi_{\mathcal{D}}(b), b; \mathcal{D})$. Thus $\|\phi^i_{\mathcal{D}}(b) - \phi_{\mathcal{D}}(b)\|_2 \leq \sqrt{4\xi/\Delta}$ for all $i \geq i_\xi$ and therefore $\phi^i_{\mathcal{D}}(b) \to \phi_{\mathcal{D}}(b)$ as $i \to \infty$. ■

Now invoking Lemmas 10 and 14 we directly get the following.

Lemma 20 (Differentially Private Unconstrained Objective Perturbation) *Under the conditions of Theorem 2, if we assume that the convex set $\mathbb{F}$ is the entire real space $\mathbb{R}^p$, then*

- *using Gamma density ν_1, Algorithm 1 (Algorithm* Obj-Pert *) guarantees ϵ-differential privacy.*
- *using Gaussian density ν_2, Algorithm 1 (Algorithm* Obj-Pert *) guarantees (ϵ, δ)-differential privacy.*

C.3. Extension to Hard Convex Constraints via Successive Approximation.

In order to extend Lemma 20 to Theorem 2, we need to show the same (as in Lemma 20) when $\mathbb{F}$ is a closed convex subset of $\mathbb{R}^p$. To show this we will again invoke our successive approximations technique.

Consider the function $f(\theta) = \min_{y \in \mathbb{F}} \|\theta - y\|_2$. This function is zero if $\theta \in \mathbb{F}$ and is increasing as θ goes farther away from $\mathbb{F}$. Also notice that f is a convex function. Now, consider the following unconstrained optimization problem.

$$\phi^i_{\mathcal{D}}(b) = \arg\min_{\theta \in \mathbb{R}^p} {J^{\text{priv}}}^i(\theta, b; \mathcal{D}) = \hat{\mathcal{L}}(\theta; \mathcal{D}) + \frac{\Delta}{2n}\|\theta\|_2^2 + \frac{1}{n}(r(\theta) + b^T\theta + if(\theta)) \tag{9}$$

Correspondingly consider the following optimization problem whose privacy we care about.

$$\phi_{\mathcal{D}}(b) = \arg\min_{\theta \in \mathbb{F}} J^{\text{priv}}(\theta, b; \mathcal{D}) = \hat{\mathcal{L}}(\theta; \mathcal{D}) + \frac{\Delta}{2n}\|\theta\|_2^2 + \frac{1}{n}(r(\theta) + b^T\theta) \tag{10}$$

Similar to Lemma 18, the following Lemma shows the pointwise convergence of $\phi^i_{\mathcal{D}}$ and $\phi_{\mathcal{D}}$.

Lemma 21 (Constrained Pointwise Convergence) *Let $\hat{\mathcal{L}}$ be a convex function and r a convex regularizer. For a given closed convex set $\mathbb{F} \subset \mathbb{R}^p$, define the function $f(\theta) = \min_{y \in \mathbb{F}} \|\theta - y\|_2$. Define the objective function $J^{priv}(\theta, b; \mathcal{D})$ and ${J^{priv}}^i(\theta, b; \mathcal{D})$ as in Equations 9 and 10. Define the minimizers, for each b, as $\phi_{\mathcal{D}}(b) = \operatorname{argmin}_{\theta \in \mathbb{F}} J^{priv}(\theta, b; \mathcal{D})$ and $\phi^i_{\mathcal{D}}(b) = \operatorname{argmin}_{\theta \in \mathbb{R}^p} {J^{priv}}^i(\theta, b; \mathcal{D})$. Then for every $b \in \mathbb{R}^p$, $\lim_{i \to \infty} \phi^i_{\mathcal{D}}(b) = \phi_{\mathcal{D}}(b)$.*

Proof Before we prove Lemma 21, we will state some simple properties about strongly convex functions. These properties will be needed in the argument of the proof of Lemma 21.

Claim 22 *Let g be a Δ-strongly convex function and let $\hat{\theta}$ be the minimizer of g over a convex set M. Then $g(\theta) - g(\hat{\theta}) \geq \frac{\Delta}{2}\|\theta - \hat{\theta}\|_2$ for all $\theta \in M$.*

Claim 23 *Let g be a convex function and let $\theta_1, \theta_2 \in \mathbb{R}^p$ and let $s \geq 1$. Then*

$$\frac{g(\theta_1 + s\theta_2) - g(\theta_1)}{\|s\theta_2\|_2} \geq \frac{g(\theta_1 + \theta_2) - g(\theta_1)}{\|\theta_2\|_2}$$

With these two claims in hand we complete the proof of Lemma 21.

In the following set of arguments we will show that for any $b \in \mathbb{R}^p$, there exists an i_0 s.t. $\forall i > i_0$, $\phi^i_{\mathcal{D}}(b) = \phi_{\mathcal{D}}(b)$. This will then directly imply that ϕ^i converges pointwise to $\phi_{\mathcal{D}}$. Consider any $b \in \mathbb{R}^p$. First note that $\phi^0_{\mathcal{D}}(b)$ is the unconstrained minimizer. By strong convexity, the unconstrained minimizer $\phi^0_{\mathcal{D}}(b)$ exists and the constrained minimizer $\phi_{\mathcal{D}}(b)$ also exists since $\mathbb{F}$ is closed and convex. If $\phi^0_{\mathcal{D}}(b) = \phi_{\mathcal{D}}(b)$, then we are done. If $\phi^0_{\mathcal{D}}(b) \in \mathbb{F}$, then $\phi^0_{\mathcal{D}}(b) = \phi_{\mathcal{D}}(b)$ by strong convexity (since $\phi_{\mathcal{D}}(b)$ is a minimizer over $\mathbb{F}$) and we are also done. Thus, we may assume $\phi^0_{\mathcal{D}}(b) \neq \phi_{\mathcal{D}}(b)$ and $\phi^0_{\mathcal{D}}(b) \notin \mathbb{F}$.

For any θ, let $\theta_{\mathbb{F}}$ be the point in $\mathbb{F}$ that is closest to θ (existence is guaranteed because $\mathbb{F}$ is closed and uniqueness is guaranteed because $\mathbb{F}$ is convex) so that $\|\theta - \theta_{\mathbb{F}}\|_2 = f(\theta)$.

Now consider the set of points $H \equiv \{\theta \ : \ J(\theta, b; \mathcal{D}) \leq J^{\text{priv}}(\phi_{\mathcal{D}}(b), b; \mathcal{D})\}$. Clearly $\phi^i_{\mathcal{D}}(b) \in H$ for all i. Let $d = \sqrt{n \cdot J^{\text{priv}}(\phi_{\mathcal{D}}(b), b; \mathcal{D}) - n \cdot J^{\text{priv}}(\phi^0_{\mathcal{D}}(b), b; \mathcal{D})}$. By Claim 22, H lies in the closed ball B with center $\phi^0_{\mathcal{D}}(b)$ and radius d. Since $\phi_{\mathcal{D}}(b) \in H \subseteq B$, the farthest distance from any point $\theta \in B$ to $\mathbb{F}$ is $\|\theta - \theta_F\|_2 \leq \|\theta - \phi_{\mathcal{D}}(b)\|_2 \leq 2d$.

Consider the function $\kappa : B \times \{v \in \mathbb{R}^p \ : \ \|v\|_2 = 2d\}$ defined as $\kappa(\theta, v) = \frac{n \cdot J^{\text{priv}}(\theta + v, b; \mathcal{D}) - n \cdot J^{\text{priv}}(\theta, b; \mathcal{D})}{\|v\|_2}$. Let m be the supremum of κ (it is finite since κ is a continuous function over a compact set). Then for any $\theta \in H$, using Claim 23 (with $\theta_1 \equiv \theta$, $\theta_2 \equiv \theta_{\mathbb{F}} - \theta$, and $s \equiv \frac{2d}{\|\theta_{\mathbb{F}} - \theta\|_2}$), we have $\frac{n \cdot J^{\text{priv}}(\theta_{\mathbb{F}}, b; \mathcal{D}) - f(\theta)}{\|\theta_{\mathbb{F}} - \theta\|_2} \leq m$.

Now set $\alpha = 2m$. Then for any $\theta \in H \subseteq B$ with $\theta \notin \mathbb{F}$,

$$\begin{aligned}
n \cdot J^{\text{priv}}(\theta, b; \mathcal{D}) + if(\theta) &= n \cdot J^{\text{priv}}(\theta_{\mathbb{F}}, b; \mathcal{D}) + if(\theta) - \frac{n \cdot J^{\text{priv}}(\theta_{\mathbb{F}}, b; \mathcal{D}) - n \cdot J^{\text{priv}}(\theta, b; \mathcal{D})}{\|\theta_{\mathbb{F}} - \theta\|_2}\|\theta_{\mathbb{F}} - \theta\|_2 \\
&\geq n \cdot J^{\text{priv}}(\theta_{\mathbb{F}}, b; \mathcal{D}) + if(\theta) - m\|\theta_{\mathbb{F}} - \theta\|_2 \\
&= n \cdot J^{\text{priv}}(\theta_{\mathbb{F}}, b; \mathcal{D}) + 2mf(\theta) - mf(\theta) \\
&\geq n \cdot J^{\text{priv}}(\phi_{\mathcal{D}}(b), b; \mathcal{D}) + mf(\theta) \\
&> n \cdot J^{\text{priv}}(\phi_{\mathcal{D}}(b), b; \mathcal{D}) \quad \text{(since } \theta \notin \mathbb{F} \text{ and } \mathbb{F} \text{ is closed)}
\end{aligned}$$

Since $\phi^i_{\mathcal{D}}(b) \in H$, this means $\phi^i_{\mathcal{D}}(b) = \phi_{\mathcal{D}}(b)$, contradicting the assumption that $\phi^i_{\mathcal{D}}(b) \notin \mathbb{F}$.

This completes the proof of Lemma 21. ■

Using the results of Lemmas 20 and 21, and invoking Lemma 10, we complete the proof of Private Convex Optimization theorem, *i.e.,* Theorem 2.

Appendix D. Estimating Empirical Risk and Generalization Error

D.1. Estimating Empirical Risk

To bound the empirical risk mentioned in Section 2.2.1, we need the following helper lemma.

Lemma 24 *Let $\mathcal{D} = \{d_1, \ldots, d_n\}$ be a dataset, and let $\hat{J}(\theta; \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; d_i) + \frac{r(\theta)}{n}$. Let $\theta^{\#} = \arg\min_{\theta \in \mathbb{F}} \hat{J}(\theta; \mathcal{D}) + \frac{\Delta}{2n}\|\theta\|_2^2$ and let θ^{priv} be the output of Algorithm 1 (Algorithm* Obj-Pert *), where $\mathbb{F} \subseteq \mathbb{R}^p$ is a closed convex set. Then*

$$\|\theta^{\#} - \theta^{priv}\|_2 \leq \frac{2\|b\|_2}{\Delta}$$

where b is the noise vector in Algorithm 1.

Proof We have $\theta^{\text{priv}} = \arg\min_{\theta \in \mathbb{F}} \underbrace{J^{\#}(\theta; \mathcal{D}) + \frac{b^T\theta}{n}}_{J^{\text{priv}}(\theta; \mathcal{D})}$, where $J^{\#}(\theta; \mathcal{D}) = \hat{J}(\theta; \mathcal{D}) + \frac{\Delta}{2n}\|\theta\|_2^2$. Similarly, $\theta^{\#} = \arg\min_{\theta \in \mathbb{F}} J^{\#}(\theta; \mathcal{D})$.

Since θ^{priv} is the minimizer of $J^{\text{priv}}(\theta; \mathcal{D})$ and J^{priv} is $\frac{\Delta}{n}$ strongly convex in θ, we have the following (from Claim 22):

$$J^{\text{priv}}(\theta^{\#}; \mathcal{D}) \geq J^{\text{priv}}(\theta^{\text{priv}}; \mathcal{D}) + \frac{\Delta}{2n}\|\theta^{\#} - \theta^{\text{priv}}\|_2^2 \tag{11}$$

$$\Rightarrow J^{\#}(\theta^{\#}; \mathcal{D}) + \frac{b^T\theta^{\#}}{n} \geq J^{\#}(\theta^{\text{priv}}; \mathcal{D}) + \frac{b^T\theta^{\text{priv}}}{n} + \frac{\Delta}{2n}\|\theta^{\#} - \theta^{\text{priv}}\|_2^2 \tag{12}$$

Notice that $J^{\#}(\theta^{\#}; \mathcal{D}) \leq J^{\#}(\theta^{\text{priv}}; \mathcal{D})$, since $\theta^{\#}$ is the minimizer of $J^{\#}(\theta; \mathcal{D})$. Therefore, we have the following:

$$\begin{aligned}
&b^T\theta^{\#} \geq b^T\theta^{\text{priv}} + \frac{\Delta}{2}\|\theta^{\#} - \theta^{\text{priv}}\|_2^2 \\
&\Rightarrow b^T(\theta^{\#} - \theta^{\text{priv}}) \geq \frac{\Delta}{2}\|\theta^{\#} - \theta^{\text{priv}}\|_2^2 \\
&\Rightarrow \|b\|_2\|\theta^{\#} - \theta^{\text{priv}}\|_2 \geq \frac{\Delta}{2}\|\theta^{\#} - \theta^{\text{priv}}\|_2^2 \\
&\Rightarrow \|\theta^{\#} - \theta^{\text{priv}}\|_2 \leq \frac{2\|b\|_2}{\Delta}
\end{aligned}$$

Hence proved. ■

The following corollary bounds the difference in the values of the objective function $J^{\#}$ at θ^{priv} and $\theta^{\#}$. The gap is due to the noise variable b.

Corollary 25 *Let $\theta^{\#} = \arg\min_{\theta \in \mathbb{F}} \hat{J}(\theta; \mathcal{D}) + \frac{\Delta}{2n}\|\theta\|_2^2$ and let θ^{priv} be the output of Algorithm 1 (Algorithm* Obj-Pert *), where $\mathbb{F} \subseteq \mathbb{R}^p$ is a closed convex set. Then*

$$J^{\#}(\theta^{priv}; \mathcal{D}) - J^{\#}(\theta^{\#}; \mathcal{D}) \leq \frac{2\|b\|_2^2}{\Delta n}$$

where b is the noise vector in Algorithm 1.

Proof From Equation 12 of the previous lemma, we have

$$J^{\#}(\theta^{\#};\mathcal{D}) + \frac{b^T\theta^{\#}}{n} \geq J^{\#}(\theta^{\text{priv}};\mathcal{D}) + \frac{b^T\theta^{\text{priv}}}{n} + \frac{\Delta}{2n}\|\theta^{\#} - \theta^{\text{priv}}\|_2^2$$
$$\Rightarrow J^{\#}(\theta^{\text{priv}};\mathcal{D}) - J^{\#}(\theta^{\#};\mathcal{D}) \leq \frac{b^T(\theta^{\#} - \theta^{\text{priv}})}{n} - \frac{\Delta}{2n}\|\theta^{\#} - \theta^{\text{priv}}\|_2^2$$
$$\Rightarrow J^{\#}(\theta^{\text{priv}};\mathcal{D}) - J^{\#}(\theta^{\#};\mathcal{D}) \leq \frac{\|b\|_2\|\theta^{\#} - \theta^{\text{priv}}\|_2}{n}$$
$$\Rightarrow J^{\#}(\theta^{\text{priv}};\mathcal{D}) - J^{\#}(\theta^{\#};\mathcal{D}) \leq \frac{2\|b\|_2^2}{n\Delta}$$

The last inequality follows from Lemma 24. This completes the proof. ■

D.1.1. PROOF OF LEMMA 3

Proof We have

$$\hat{J}(\theta^{\text{priv}};\mathcal{D}) - \hat{J}(\hat{\theta};\mathcal{D}) = (J^{\#}(\theta^{\text{priv}};\mathcal{D}) - J^{\#}(\theta^{\#};\mathcal{D})) + (J^{\#}(\theta^{\#};\mathcal{D}) - J^{\#}(\hat{\theta};\mathcal{D})) + \frac{\Delta}{2n}\|\hat{\theta}\|_2^2 - \frac{\Delta}{2n}\|\theta^{\text{priv}}\|_2^2$$

Notice that $(J^{\#}(\theta^{\#};\mathcal{D}) - J^{\#}(\hat{\theta};\mathcal{D})) \leq 0$. Also from Corollary 25 we have $(J^{\#}(\theta^{\text{priv}};\mathcal{D}) - J^{\#}(\theta^{\#};\mathcal{D})) \leq \frac{2\|b\|_2^2}{n\Delta}$. Hence, we have

$$\hat{J}(\theta^{\text{priv}};\mathcal{D}) - \hat{J}(\hat{\theta};\mathcal{D}) \leq \frac{2\|b\|_2^2}{n\Delta} + \frac{\Delta}{2n}\|\hat{\theta}\|_2^2$$

This completes the proof. ■

D.1.2. PROOF OF THEOREM 4

In order to prove Theorem 4, we prove the following which is a slightly generalized version. Replacing $\Delta = \Theta\left(\frac{\zeta p \log p}{\epsilon\|\hat{\theta}\|_2}\right)$ in the first part of Theorem 26 and $\Delta = \Theta\left(\frac{\sqrt{\zeta^2 p \log\frac{1}{\delta}}}{\epsilon\|\hat{\theta}\|_2}\right)$ in the second part of Theorem 26, we obtain Theorem 4. Since, we are looking at expected error in Theorem 4, we ignore the term γ.

Theorem 26 *Assuming that* $\| \bigtriangledown \ell(\theta; d)\|_2 \leq \zeta$ *(for all* $d \in \mathcal{P}$ *and for all* $\theta \in \mathbb{F}$*), the following are true.*

1. *With Gamma density* ν_1*, w.p.* $\geq 1 - \gamma$

$$\hat{J}(\theta^{priv};\mathcal{D}) - \hat{J}(\hat{\theta};\mathcal{D}) \leq \frac{8\zeta^2 p^2 \log^2\frac{p}{\gamma}}{n\epsilon^2\Delta} + \frac{\Delta}{2n}\|\hat{\theta}\|_2^2$$

2. *With Gaussian density* ν_2*, w.p.* $\geq 1 - \gamma$

$$\hat{J}(\theta^{priv};\mathcal{D}) - \hat{J}(\hat{\theta};\mathcal{D}) \leq \frac{4p\zeta^2(8\log\frac{2}{\delta} + 4\epsilon)\log(1/\gamma)}{n\epsilon^2\Delta} + \frac{\Delta}{2n}\|\hat{\theta}\|_2^2$$

Proof The proof essentially goes via bounding $\|b\|_2$ under the two distributions ν_1 and ν_2 used in Algorithm Obj-Pert (Algorithm 1) and plugging it in Lemma 3.

Recall that distribution $\nu_1(b;\epsilon,\zeta) \propto e^{-\frac{\|b\|_2}{2\zeta}}$. Thus, under the distribution ν_1 for b, we have $\|b\|_2 \sim \Gamma(p, \frac{2\zeta}{\epsilon})$. The following lemma from Chaudhuri et al. (2011) provides a tail bound for *Gamma* distribution.

Lemma 27 (Lemma 4 from Chaudhuri et al. (2011)) *Let X be a random variable drawn from the distribution $\Gamma(p,\theta)$, where p is a positive integer. Then,*

$$\Pr\left[X \geq p\theta \log\frac{p}{\gamma}\right] \leq \gamma$$

Using Lemma 27, w.p. $\geq 1-\gamma$ we have the following:

$$\|b\|_2 \leq \frac{2p\zeta \log\frac{p}{\gamma}}{\epsilon}$$

Plugging in the value of $\|b\|_2$ from above into Lemma 3, we have w.p. $\geq 1-\gamma$

$$\hat{J}(\theta^{\text{priv}};\mathcal{D}) - \hat{J}(\hat{\theta};\mathcal{D}) \leq \frac{8\zeta^2 p^2 \log^2\frac{p}{\gamma}}{n\epsilon^2\Delta} + \frac{\Delta}{2n}\|\hat{\theta}\|_2^2$$

This completes the proof of first part of the theorem.

For the second part, we need to bound $\|b\|_2$ when $b \sim \mathcal{N}\left(0, \mathbb{I}_p \frac{\zeta^2\left(8\log\frac{2}{\delta}+4\epsilon\right)}{\epsilon^2}\right)$. We use the following lemma from Dasgupta and Schulman (2007).

Lemma 28 (Lemma 2 from Dasgupta and Schulman (2007)) *Pick X from the distribution $\mathcal{N}(0,\mathbb{I}_p)$. Then for any $\phi \geq 1$, we have*

$$\Pr[\|X\|_2 \geq \sqrt{\phi p}] \leq e^{-\frac{p}{2}(\phi-1-\log\phi)}$$

In the above lemma, in order to set $e^{-\frac{p}{2}(\phi-1-\log\phi)} \leq \gamma$, we need $1 + \frac{2}{p}\log\frac{1}{\gamma} \leq \frac{\phi}{2}$. Therefore, setting ϕ as above, we have w.p. $\geq 1-\gamma$,

$$\|X\|_2 \leq \sqrt{2 + \frac{2}{p}\log\frac{1}{\gamma}}\sqrt{p}$$
$$\Rightarrow \|X\|_2 \leq \sqrt{2p\log\frac{1}{\gamma}}$$

Using the above bound we have w.p. $\geq 1-\gamma$,

$$\|b\|_2 \leq \sqrt{\frac{2p\zeta^2\left(8\log\frac{2}{\delta}+4\epsilon\right)\log\frac{1}{\gamma}}{\epsilon^2}}$$

Plugging in the value of $\|b\|_2$ from above into Lemma 3, we have w.p. $\geq 1-\gamma$,

$$\hat{J}(\theta^{\text{priv}};\mathcal{D}) - \hat{J}(\hat{\theta};\mathcal{D}) \leq \frac{4p\zeta^2(8\log\frac{2}{\delta}+4\epsilon)\log(1/\gamma)}{n\epsilon^2\Delta} + \frac{\Delta}{2n}\|\hat{\theta}\|_2^2$$

This completes the proof of second part of the theorem. ■

D.2. Estimating Generalization Error

Generalization error : In our presentation of generalization error we restrict ourselves to *Generalized Linear Models* (GLM). In GLM, each data entry d in the dataset is of the form (y, x), where $y \in \mathbb{R}$ and $x \in \mathbb{R}^p$. The loss function $\ell(\theta; d)$ is of the form $\ell_{\text{GLM}}(x^T\theta; y)$, where $d = (y, x)$.

Following is the generalization error bound we obtain as a corollary to Theorem 4 by using (Shalev-Shwartz et al., 2009, Theorem 2) to convert from empirical risk to generalization error. In the rest of the paper, we only concentrate on empirical risk as one can easily convert it to generalization error via the result discussed above. For the simplicity of exposition, in this section we assume that the regularizer $\frac{1}{n}r(\theta)$ (in the loss $\bar{J}(\theta; \mathcal{D})$) is zero for all θ.

Theorem 29 *Consider that for any data entry $d = (y, x)$ (where $y \in \mathbb{R}$ and $x \in \mathbb{R}^p$), $\|x\|_2 \leq R$. Also assume that $|\ell'_{GLM}(u; y)| \leq L$ and $|\ell''_{GLM}(u; y)| \leq (\epsilon\Delta)/(2R^2)$, where $u \in \mathbb{R}$ and $y \in \mathbb{R}$, and the derivatives are w.r.t. u. When using Gaussian density ν_2 for the noise vector b and setting $\Delta = \Theta\left(\frac{\sqrt{(RL)^2 p \log(1/\delta)}}{\epsilon\|\bar{\theta}\|_2}\right)$, we have $\mathbb{E}_b\left[\bar{J}(\theta^{priv}; \mathcal{P}) - \bar{J}(\bar{\theta}; \mathcal{P})\right] = O\left(\frac{(RL)\sqrt{p\log(1/\delta)}\|\bar{\theta}\|_2}{\epsilon n}\right)$.*

The main takeaway from the above theorem is that if we assume $\zeta = RL$ and $\|\hat{\theta}\|_2 \approx \|\bar{\theta}\|_2$ (see Theorem 4), then asymptotically the generalization error is same as the empirical risk. Additionally, comparing the private generalization error to the non-private one, we can show that the private version is worse by a factor of $\sqrt{p}$.

Appendix E. Refined utility guarantees under stronger assumptions

E.1. Parameter Estimation Error bounds

Theorem 30 (Parameter estimation error) *Under the assumption that $\| \bigtriangledown \ell(\theta; d)\|_2 \leq \zeta$ (for all $d \in \mathcal{P}$ and for all $\theta \in \mathbb{F}$), the following are true for Algorithm 1.*

1. *When using the Gaussian density ν_2 and $\psi \geq \frac{\sqrt{32p}\zeta\sqrt{(8\log\frac{2}{\delta}+4\epsilon)\log(1/\gamma)}}{\epsilon(\Delta+\eta)} + 2\sqrt{\frac{\Delta}{(\Delta+\eta)}}\|\hat{\theta}\|_2$, then w.p. $\geq 1-\gamma$ the following are true.*
   - *(a)* $\|\theta^{priv} - \theta^{\#}\|_2 \leq \frac{\sqrt{2p}\zeta\sqrt{(8\log\frac{2}{\delta}+4\epsilon)\log(1/\gamma)}}{\epsilon(\Delta+\eta)}$
   - *(b)* $\|\theta^{priv} - \hat{\theta}\|_2 \leq \frac{\sqrt{2p}\zeta\sqrt{(8\log\frac{2}{\delta}+4\epsilon)\log(1/\gamma)}}{\epsilon(\Delta+\eta)} + \sqrt{\frac{\Delta}{(\Delta+\eta)}}\|\hat{\theta}\|_2$

Here ψ is the radius of the ball around $\hat{\theta}$ where $\hat{\mathcal{L}}$ is $\frac{\eta}{n}$-strongly convex.

Proof By assumption we have $\hat{\mathcal{L}}(\theta; \mathcal{D})$ is η/n-strongly convex in a ball of radius ψ around $\hat{\theta}$, where $\hat{\theta} = \arg\min_{\theta\in\mathbb{F}} \hat{\mathcal{L}}(\theta; \mathcal{D})$. We will fix the value of ψ later.

Recall that

$$\theta^{\text{priv}} = \arg\min_{\theta\in\mathbb{F}} \underbrace{\hat{\mathcal{L}}(\theta; \mathcal{D}) + \frac{\Delta}{2n}\|\theta\|_2^2 + \frac{b^T\theta}{n}}_{J^{\text{priv}}(\theta;\mathcal{D})}$$

and

$$\theta^{\#} = \arg\min_{\theta \in \mathbb{F}} \underbrace{\hat{\mathcal{L}}(\theta; \mathcal{D}) + \frac{\Delta}{2n}\|\theta\|_2^2}_{J^{\#}(\theta;\mathcal{D})}$$

where b is the noise vector used in Algorithm Obj-Pert (Algorithm 1).

Assume for now that $\psi \geq 2\left(\|\theta^{\text{priv}} - \theta^{\#}\|_2 + \|\theta^{\#} - \hat{\theta}\|_2\right)$. We will remove this assumption as we move along. The above assumption implies the following:

1. $\hat{\mathcal{L}}(\theta; \mathcal{D})$ is $\frac{\eta}{n}$-strongly convex in a ball of radius $\|\theta^{\#} - \hat{\theta}\|_2$ around $\theta^{\#}$.
2. $\hat{\mathcal{L}}(\theta; \mathcal{D})$ is $\frac{\eta}{n}$-strongly convex in a ball of radius $\|\theta^{\text{priv}} - \theta^{\#}\|_2$ around θ^{priv}.

In order to bound $\|\theta^{\text{priv}} - \hat{\theta}\|_2$, we first bound $\|\theta^{\text{priv}} - \theta^{\#}\|_2$ and $\|\theta^{\#} - \hat{\theta}\|_2$ individually. Since $\|\theta^{\text{priv}} - \hat{\theta}\|_2 \leq \|\theta^{\text{priv}} - \theta^{\#}\|_2 + \|\theta^{\#} - \hat{\theta}\|_2$, we obtain the required bound.

Since θ^{priv} is the minimizer of $J^{\text{priv}}(\theta; \mathcal{D})$, the following is true from the definition of J^{priv}.

$$J^{\text{priv}}(\theta^{\#}; \mathcal{D}) \geq J^{\text{priv}}(\theta^{\text{priv}}; \mathcal{D}) + \frac{\Delta + \eta}{2n}\|\theta^{\#} - \theta^{\text{priv}}\|_2^2 \tag{13}$$

$$\Rightarrow J^{\#}(\theta^{\#}; \mathcal{D}) + \frac{b^T\theta^{\#}}{n} \geq J^{\#}(\theta^{\text{priv}}; \mathcal{D}) + \frac{b^T\theta^{\text{priv}}}{n} + \frac{\Delta + \eta}{2n}\|\theta^{\#} - \theta^{\text{priv}}\|_2^2 \tag{14}$$

Notice that $J^{\#}(\theta^{\#}; \mathcal{D}) \leq J^{\#}(\theta^{\text{priv}}; \mathcal{D})$, since $\theta^{\#}$ is the minimizer of $J^{\#}(\theta; \mathcal{D})$. Therefore we have the following:

$$b^T\theta^{\#} \geq b^T\theta^{\text{priv}} + \frac{\Delta + \eta}{2}\|\theta^{\#} - \theta^{\text{priv}}\|_2^2 \tag{15}$$

$$\Rightarrow b^T(\theta^{\#} - \theta^{\text{priv}}) \geq \frac{\Delta + \eta}{2}\|\theta^{\#} - \theta^{\text{priv}}\|_2^2 \tag{16}$$

$$\Rightarrow \|b\|_2\|\theta^{\#} - \theta^{\text{priv}}\|_2 \geq \frac{1}{2}\|\theta^{\#} - \theta^{\text{priv}}\|_2^2(\Delta + \eta) \tag{17}$$

$$\Rightarrow \|\theta^{\#} - \theta^{\text{priv}}\|_2 \leq \frac{2\|b\|_2}{(\Delta + \eta)} \tag{18}$$

Here, η is the local strong convexity parameter.

In order to bound $\|\theta^{\#} - \hat{\theta}\|_2$, we first notice the following:

$$\hat{\mathcal{L}}(\hat{\theta}) + \frac{\Delta}{2n}\|\hat{\theta}\|_2^2 \geq \hat{\mathcal{L}}(\theta^{\#}) + \frac{\Delta}{2n}\|\theta^{\#}\|_2^2 + \frac{\Delta + \eta}{2n}\|\theta^{\#} - \theta^{\text{priv}}\|_2^2 \tag{19}$$

$$\Rightarrow \Delta\|\hat{\theta}\|_2^2 \geq (\Delta + \eta)\|\hat{\theta} - \theta^{\#}\|_2^2 \tag{20}$$

$$\Rightarrow \|\hat{\theta} - \theta^{\#}\|_2 \leq \sqrt{\frac{\Delta}{\Delta + \eta}}\|\hat{\theta}\|_2 \tag{21}$$

From Equations 18 and 21, it follows that

$$\|\theta^{\text{priv}} - \hat{\theta}\|_2 \leq \frac{2\|b\|_2}{(\Delta + \eta)} + \sqrt{\frac{\Delta}{\Delta + \eta}}\|\hat{\theta}\|_2$$

Equations 18 and 21 also imply a bound on the radius ψ, so that the initial assumption $\psi \geq 2\left(\|\theta^{\text{priv}} - \theta^{\#}\|_2 + \|\theta^{\#} - \hat{\theta}\|_2\right)$ is true. We set $\psi \geq 2\left(\frac{2\|b\|_2}{(\Delta+\eta)} + \sqrt{\frac{\Delta}{\Delta+\eta}}\|\hat{\theta}\|_2\right)$ to satisfy the above assumption.

From the tail bound calculations for $\|b\|_2$ in Appendix D.1.2, w.p. $\geq 1-\gamma$ we have the following.

- In Algorithm Obj-Pert (Algorithm 1) when the noise distribution is ν_2, we have

$$\|b\|_2 \leq \sqrt{\frac{2p\zeta^2\left(8\log\frac{2}{\delta} + 4\epsilon\right)\log\frac{1}{\gamma}}{\epsilon^2}}$$

Plugging in these bounds for $\|b\|_2$, completes the proof. ■

E.2. Proof of Theorem 5

In order to prove Theorem 5, we prove the following slightly generalized version. Substituting the parameters from Assumption 1 in Theorem 31 and setting $\Delta = 2\lambda/\epsilon$, we obtain Theorem 5. Note that in Theorem 5 we ignore the term γ, since there we are dealing with expected error.

Theorem 31 *When using the Gaussian distribution function ν_2 and $\psi \geq \frac{\sqrt{p}\zeta\sqrt{\log(1/\delta)\log(1/\gamma)}}{\epsilon(\Delta+\eta)} + \sqrt{\frac{\Delta}{(\Delta+\eta)}}\|\hat{\theta}\|_2$, then w.p. $\geq 1-\gamma$ the following is true.*

$$\hat{\mathcal{L}}(\theta^{priv}; \mathcal{D}) - \hat{\mathcal{L}}(\hat{\theta}; \mathcal{D}) = O\left(\frac{p\zeta^2\log(1/\delta)\log(1/\gamma)}{n\epsilon^2(\Delta+\eta)} + \frac{\Delta}{n}\|\hat{\theta}\|_2^2\right)$$

Here ψ is the radius of the ball around $\hat{\theta}$ where $\hat{\mathcal{L}}$ is $\frac{\eta}{n}$-strongly convex.

Proof Using Equation 14 from Appendix E.1 we have

$$J^{\#}(\theta^{\#}; \mathcal{D}) + \frac{b^T\theta^{\#}}{n} \geq J^{\#}(\theta^{\text{priv}}; \mathcal{D}) + \frac{b^T\theta^{\text{priv}}}{n} + \frac{\Delta+\eta}{2n}\|\theta^{\#} - \theta^{\text{priv}}\|_2^2 \tag{22}$$

$$\Rightarrow J^{\#}(\theta^{\text{priv}}; \mathcal{D}) - J^{\#}(\theta^{\#}; \mathcal{D}) \leq \frac{b^T(\theta^{\#} - \theta^{\text{priv}})}{n} - \frac{\Delta+\eta}{2n}\|\theta^{\#} - \theta^{\text{priv}}\|_2^2 \tag{23}$$

$$\Rightarrow J^{\#}(\theta^{\text{priv}}; \mathcal{D}) - J^{\#}(\theta^{\#}; \mathcal{D}) \leq \frac{\|b\|_2\|\theta^{\#} - \theta^{\text{priv}}\|_2}{n} \tag{24}$$

The last step follows from Cauchy-Schwarz inequality. Recall that

$$\hat{\mathcal{L}}(\theta^{\text{priv}}; \mathcal{D}) - \hat{\mathcal{L}}(\hat{\theta}; \mathcal{D}) = (J^{\#}(\theta^{\text{priv}}; \mathcal{D}) - J^{\#}(\theta^{\#}; \mathcal{D})) + (J^{\#}(\theta^{\#}; \mathcal{D}) - J^{\#}(\hat{\theta}; \mathcal{D})) + \frac{\Delta}{2n}\|\hat{\theta}\|_2^2 - \frac{\Delta}{2n}\|\theta^{\text{priv}}\|_2^2$$

Notice that $J^{\#}(\theta^{\#}; \mathcal{D}) - J^{\#}(\hat{\theta}; \mathcal{D}) \leq 0$. Using Equation 24 we have the following.

$$\hat{\mathcal{L}}(\theta^{\text{priv}}; \mathcal{D}) - \hat{\mathcal{L}}(\hat{\theta}; \mathcal{D}) \leq \frac{\|b\|_2\|\theta^{\#} - \theta^{\text{priv}}\|_2}{n} + \frac{\Delta}{2n}\|\hat{\theta}\|_2^2$$

The theorem follows from using the tail bounds for $\|b\|_2$ under the distributions ν_1 and ν_2 (see Appendix E.1) and Theorem 30. ■

Appendix F. Exponential Mechanism based High-dimensional Regression

F.1. Details of Algorithm Exp-mech

Algorithm 4 Exponential Mechanism based feature selection (Exp-mech)

Require: dataset: $\mathcal{D} = \{d_1, \ldots, d_n\}$, privacy parameters: (ϵ, δ), loss function: $\hat{\mathcal{L}}(\theta; \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; d_i)$, dimensionality of the problem: p, number of data points: n, L_2 penalization parameter: Δ, support size of θ^*: s, closed convex set: $\mathbb{F}$, α: bound on $|\ell(\theta; d)|$ restricted to any support of size s and for any $d \in \mathcal{T}$

1: For any s-sparse subspace Γ, let score function $q(\Gamma; \mathcal{D}) = \min_{\theta \in \mathbb{F}_\Gamma} \sum_{i=1}^{n} \ell(\theta; d_i)$, where $\mathbb{F}_\Gamma$ refers to the vectors in $\mathbb{F}$ with support in Γ. Pick a subspace $\hat{\Gamma}$ w.p. $\propto e^{-\frac{\epsilon}{2\alpha} q(\Gamma; \mathcal{D})}$.

2: **return** $\hat{\Gamma}$

F.1.1. PRIVACY

Theorem 32 *Algorithm* Exp-mech *(Algorithm 4) is ϵ-differentially private.*

Proof In order to prove the theorem, we bound the *sensitivity* of the score function $q(\Gamma; \mathcal{D})$ (*i.e.,* the maximum absolute change in $q(\theta; \mathcal{D})$ when one entry of $\mathcal{D}$ is modified) via the following lemma.

Lemma 33 *Sensitivity of the score function $q(\Gamma; \mathcal{D}) = \min_{\theta \in \mathbb{F}_\Gamma} \sum_{i=1}^{n} \ell(\theta; \mathcal{D})$ is bounded by $\alpha \geq \max_{\theta \in \mathbb{F}_\Gamma, d \in \mathcal{T}} \ell(\theta; d)$, where Γ is any s-sparse subset and $\mathcal{T}$ is the domain from which the data entries are drawn.*

Proof Let $\mathcal{D}'$ be any dataset which either has one entry more (less) than $\mathcal{D}$. W.l.o.g. we assume that $\mathcal{D}'$ has one entry more as compared to $\mathcal{D}$ (*i.e.,* $\mathcal{D}'$ has entry d_{n+1} which $\mathcal{D}$ does not). To bound the sensitivity of q we need to bound $|q(\Gamma; \mathcal{D}') - q(\Gamma; \mathcal{D})|$ for any database pairs $\mathcal{D}$ and $\mathcal{D}'$, and any subset Γ of size at most s. The bound is as follows.

$$
\begin{aligned}
|q(\Gamma; \mathcal{D}') - q(\Gamma; \mathcal{D})| &= \left| \min_{\theta \in \mathbb{F}_\Gamma} n\hat{\mathcal{L}}(\theta; \mathcal{D}') - \min_{\theta \in \mathbb{F}_\Gamma} n\hat{\mathcal{L}}(\theta; \mathcal{D}) \right| \\
&= \left| \min_{\theta \in \mathbb{F}_\Gamma} \left(n\hat{\mathcal{L}}(\theta; \mathcal{D}) + \ell(\theta; d_{n+1}) \right) - \min_{\theta \in \mathbb{F}_\Gamma} \left(n\hat{\mathcal{L}}(\theta; \mathcal{D}) \right) \right| \\
&\leq \left| \min_{\theta \in \mathbb{F}_\Gamma} n\hat{\mathcal{L}}(\theta; \mathcal{D}) + \max_{\theta \in \mathbb{F}_\Gamma} \ell(\theta; d_{n+1}) - \min_{\theta \in \mathbb{F}_\Gamma} n\hat{\mathcal{L}}(\theta; \mathcal{D}) \right| \\
&= \max_{\Gamma, \theta \in \mathbb{F}_\Gamma, d \in \mathcal{T}} \ell(\theta; d) \leq \alpha
\end{aligned}
$$

With this the bound in the above lemma follows. ■

Now for two datasets $\mathcal{D}$ and $\mathcal{D}'$, the ratio of the probabilities for picking any support of size s is as follows.

$$
\begin{aligned}
\frac{\Pr[\hat{\Gamma}(\mathcal{D}) = \Gamma]}{\Pr[\hat{\Gamma}(\mathcal{D}') = \Gamma]} &\leq \frac{e^{-\frac{\epsilon q(\Gamma; \mathcal{D})}{2\alpha}}}{e^{-\frac{\epsilon q(\Gamma; \mathcal{D}')}{2\alpha}}} \cdot \frac{\sum_\Gamma e^{-\frac{\epsilon q(\Gamma; \mathcal{D}')}{2\alpha}}}{\sum_\Gamma e^{-\frac{\epsilon q(\Gamma; \mathcal{D})}{2\alpha}}} \\
&\leq e^{\epsilon}
\end{aligned}
$$

The lower bound of $e^{-\epsilon}$ also follows symmetrically. With this the proof is complete.

■

F.2. Proof of Theorem 6

In order to prove Theorem 6, we prove a slightly more general version stated below. Since in Theorem 6 we are dealing with expected error, we ignore the term γ.

Theorem 34 *Assume that $|\ell(\theta; d)| \leq \alpha$ (for all $\theta \in \mathbb{F}_\Gamma$, for all $d \in \mathcal{T}$ and for all support Γ of size s). With probability $\geq 1 - \gamma$, we have*

$$\hat{\mathcal{L}}(\phi; \mathcal{D}) - \hat{\mathcal{L}}(\theta^{sp}; \mathcal{D}) = \frac{2\alpha s \log(p/\gamma)}{\epsilon n}$$

where $\phi = \arg\min_{\theta \in \mathbb{F}_{\hat{\Gamma}}} \hat{\mathcal{L}}(\theta; \mathcal{D})$ and $\hat{\Gamma}$ is the support selected by Algorithm Exp-mech *(Algorithm 4).*

Proof Let $\Gamma_{\min}$ be the support of size $\leq s$ which minimizes $\min_{\theta \in \mathbb{F}} \hat{\mathcal{L}}_\Gamma(\theta; \mathcal{D})$ w.r.t. Γ. Recall that $\hat{\Gamma}$ is the support output by exponential sampling. Based on the distribution used for exponential sampling, we have the following for any $\kappa > 0$.

$$\Pr\left[\min_{\theta \in \mathbb{F}_{\hat{\Gamma}}} \hat{\mathcal{L}}(\theta; \mathcal{D}) \geq \min_{\theta \in \mathbb{F}_{\Gamma_{\min}}} \hat{\mathcal{L}}(\theta; \mathcal{D}) + \frac{\kappa}{n}\right] \leq \binom{p}{s} \exp\left(-\frac{\epsilon\kappa}{2\alpha}\right)$$

$$\Rightarrow \Pr\left[\min_{\theta \in \mathbb{F}_{\hat{\Gamma}}} \hat{\mathcal{L}}(\theta; \mathcal{D}) \geq \hat{\mathcal{L}}(\theta^{\text{sp}}; \mathcal{D}) + \frac{\kappa}{n}\right] \leq \binom{p}{s} \exp\left(-\frac{\epsilon\kappa}{2\alpha}\right)$$

The last inequality follows from the fact that $\hat{\mathcal{L}}(\theta^{\text{sp}}; \mathcal{D}) = \min_{\theta \in \mathbb{F}} \hat{\mathcal{L}}_{\Gamma_{\min}}(\theta; \mathcal{D})$. Setting the R.H.S. $\leq \gamma$, we have $\kappa \leq \frac{2\alpha s}{\epsilon} \log \frac{p}{\gamma}$. Thus w.p. $\geq 1 - \gamma$ we have

$$\hat{\mathcal{L}}(\phi; \mathcal{D}) - \hat{\mathcal{L}}(\theta^{\text{sp}}; \mathcal{D}) \leq \frac{2\alpha s}{n\epsilon} \log \frac{p}{\gamma}$$

This completes the proof. ■

F.3. Proof of Theorem 7

In order to prove Theorem 7, we prove a slightly more general version stated below. Setting $\alpha = 4s^2$, $\zeta = 2s^{3/2}$, $\lambda = s$ and $\|\phi\|_2 \leq \sqrt{s}$, and substituting $\Delta = \Theta\left(\frac{s}{\epsilon}\right)$ in Theorem 35 we obtain the required bound for Theorem 7. (See Appendix H for an explanation about the setting of above parameters.) Since in Theorem 7 we are dealing with expected error, we ignore the term γ.

Theorem 35 *Let $\hat{\mathcal{L}}(\theta; \mathcal{D})$ be Ψ-strongly convex for a given dataset $\mathcal{D}$ when the support of $\theta \in \mathbb{F}$ is restricted to any set Γ of size $\leq s$. Assuming that $\|\ell(\theta; d)\|_2 \leq \zeta$, $|\ell(\theta; d)| \leq \alpha$, λ is the bound on the maximum eigenvalue of $\bigtriangledown^2 \ell$ (for all $\theta \in \mathbb{F}_\Gamma$, for all $d \in \mathcal{T}$ and for all support Γ of size s), with probability $\geq 1 - \gamma$, the following is true.*

$$\hat{\mathcal{L}}(\theta^{priv}; \mathcal{D}) - \hat{\mathcal{L}}(\theta^*; \mathcal{D}) \leq \frac{16 s \zeta^2 (8 \log \frac{2}{\delta} + 2\epsilon) \log(2/\gamma)}{n\epsilon^2(\Delta + n\Psi)} + \frac{4\alpha s}{n\epsilon} \log \frac{2p}{\gamma} + \frac{\Delta}{2n} \|\phi\|_2^2$$

where $\phi = \arg\min_{\theta \in \mathbb{F}_{\hat{\Gamma}}} \hat{\mathcal{L}}(\theta; \mathcal{D})$ *and* $\hat{\Gamma}$ *is the support chosen by Algorithm* Exp-mech *(Algorithm 4).*

Proof We bound $\hat{\mathcal{L}}(\theta^{\text{priv}}; \mathcal{D}) - \hat{\mathcal{L}}(\theta^*; \mathcal{D})$ in two parts A and B mentioned below.

$$\hat{\mathcal{L}}(\theta^{\text{priv}}; \mathcal{D}) - \hat{\mathcal{L}}(\theta^*; \mathcal{D}) = \underbrace{\hat{\mathcal{L}}(\phi; \mathcal{D}) - \hat{\mathcal{L}}(\theta^*; \mathcal{D})}_{A} + \underbrace{\hat{\mathcal{L}}(\theta^{\text{priv}}; \mathcal{D}) - \hat{\mathcal{L}}(\phi; \mathcal{D})}_{B}$$

Let us first concentrate on part A. From Theorem 34, w.p. $\geq 1 - \frac{\gamma}{2}$ we have

$$A \leq \frac{4\alpha s}{n\epsilon} \log \frac{2p}{\gamma}$$

Notice that after selecting $\hat{\Gamma}$, the problem has reduced to a s-dimensional subspace. Now invoking Theorem 31 restricted to the support $\hat{\Gamma}$, setting the failure probability to $\gamma/2$ and plugging $\epsilon/2$, w.p. $\geq 1 - \frac{\gamma}{2}$ we have

$$B \leq \frac{16 s \zeta^2 (8 \log \frac{2}{\delta} + 2\epsilon) \log(2/\gamma)}{n\epsilon^2(\Delta + n\Psi)} + \frac{\Delta}{2n} \|\phi\|_2^2$$

Using the bounds for A and B above, Theorem 34 follows. ■

Appendix G. Efficient Feature Selection via Sample and Aggregate Framework

G.1. Details of Algorithm Samp-Agg

Algorithm 5 Samp-Agg : Sample and Aggregate based feature selection

Require: dataset: $\mathcal{D} = \{d_1, \ldots, d_n\}$, privacy parameter: ϵ, algorithm: $\mathcal{A}_{\text{supp}}$, dimensionality of the problem: p, number of data points: n, support size of θ^*: s, and number of blocks: k, convex set $\mathbb{F}$

1: Partition the dataset $\mathcal{D}$ into k blocks of size $\psi = \frac{n}{k}$ each. Call the blocks $\mathcal{D}_1, \cdots, \mathcal{D}_k$.
2: **for** $i = 1$ to k **do**
3: Set $V_i = \mathcal{A}_{\text{supp}}(\mathcal{D}_i, s, \mathbb{F})$ $\{V_i \in \{0,1\}^p$ is an indicator vector for the support. $\mathcal{A}_{\text{supp}}(\cdot, s)$ is guaranteed to produce a support of size at most $s.\}$
4: **end for**
5: Set $G = \frac{1}{k}\sum_{i=1}^{k} V_i + Lap\left(\frac{2s}{k\epsilon}\right)^p$ $\{Lap(\lambda)^p$ is a vector of i.i.d. Laplace r.v. with scaling parameter $\lambda.\}$
6: $\hat{\Gamma} \leftarrow$ indices of the largest s-coordinates in G.
7: **return** $\hat{\Gamma}$

G.2. Privacy Guarantee for Algorithm Samp-Agg (Algorithm 5)

Theorem 36 *Algorithm* Samp-Agg *(Algorithm 5) is ϵ-differentially private.*

Proof To prove the theorem, we notice that in Algorithm Samp-Agg each entry of the dataset $\mathcal{D}$ lies in only one of the data blocks $\mathcal{D}_i$. Now consider the set of indicator vectors $V_i \in \{0,1\}^n$

returned by Algorithm $\mathcal{A}_{\text{supp}}$ for each data block $\mathcal{D}_i$. One can view V_i 's to be votes $\in \{0,1\}$ given to each coordinate by the data block $\mathcal{D}_i$. We define a score function for any coordinate $c \in [p]$ and for any dataset $\mathcal{D}$ as $q(c,\mathcal{D}) = \frac{1}{k}\sum_{i=1}^{k} V_i(c)$, where $V_i(c)$ is the vote for coordinate c in the i-th block and k is number of blocks. Notice that the sensitivity of the score function q is bounded by $\frac{1}{k}$. This means that by removing (adding) one entry from (to) $\mathcal{D}$, one can change $q(c,\mathcal{D})$ by at most $\frac{1}{k}$ for any dataset $\mathcal{D}$ and for any $c \in [p]$. We now invoke the following theorem by Bhaskar et al. (2010).

Theorem 37 (Modified Theorem 4 from Bhaskar et al. (2010)) *Let $A = \{a_1, \cdots, a_p\}$ be a set of elements and let $\mathcal{D}$ be a dataset which assigns score $q(a,\mathcal{D})$ to each element $a \in A$. Also let Δq be the upper bound on the sensitivity of q,* i.e., *by removing (adding) one entry from (to) $\mathcal{D}$, one can change $q(a,\mathcal{D})$ by at most Δq for any dataset $\mathcal{D}$ and for any $a \in A$. If one picks a set $\hat{\Gamma}$ of s highest entries from A based on the noisy scores defined by $q_{noisy}(a,\mathcal{D}) = q(a,\mathcal{D}) + Lap\left(\frac{2s(\Delta q)}{\epsilon}\right)$, then $\hat{\Gamma}$ is ϵ-differentially private.*

Here $Lap(\kappa)$ denotes the Laplace distribution with scaling parameter κ.

In Theorem 37, setting A to be the set of p-coordinates, $\Delta q = \frac{1}{k}$ and setting privacy parameter to ϵ, we have $\hat{\Gamma}$ (the output of Algorithm Samp-Agg (Algorithm 5)) to be ϵ-differentially private.

This completes the proof. ■

G.3. Proof of Theorem 8

Proof To select a support $\hat{\Gamma}$, Algorithm Samp-Agg (Algorithm 5) does the following. It first finds $A = \frac{1}{k}\sum_{i=1}^{k} V_i$, where $V_i \in \{0,1\}^p$ is an indicator vector indicating whether a particular coordinate $\in [p]$ is in the support for data block $\mathcal{D}_i$ (see Line 5 of Algorithm Samp-Agg (Algorithm 5)). To each coordinate of A, it *independently* adds noise Lap $\left(\frac{2s}{k\epsilon}\right)$. Call this noisy vector A_{noise}. Now $\hat{\Gamma}$ is the set of s coordinates with the highest value in A_{noise}.

By assumption, $\mathcal{A}_{\text{supp}}$ identifies the correct support $\hat{\Gamma}^*$ for all k of the data blocks. This means that the entries in A corresponding to the correct support has value one and all others are set to zero. Let GOOD be the coordinates in $\hat{\Gamma}^*$ and let BAD be the complementary set. Then for any $0 \leq \psi \leq 1$, the following is true.

$$\Pr[a \text{ coordinate} \in \text{GOOD has value} \leq 1-\psi \text{ in } A_{\text{noise}}] \leq \frac{1}{2}e^{-\frac{\psi\epsilon k}{4s}}$$

By union bound, this in turn implies the following.

$$\Pr[\textit{any} \text{ coordinate} \in \text{GOOD has value} \leq 1-\psi \text{ in } A_{\text{noise}}] \leq \frac{s}{2}e^{-\frac{\psi\epsilon k}{2s}}$$

Similarly,

$$\Pr[\textit{any} \text{ coordinate} \in \text{GOOD has value} \geq \psi \text{ in } A_{\text{noise}}] \leq \frac{p}{2}e^{-\frac{\psi\epsilon k}{2s}}$$

Therefore,

$$\Pr[\text{any coordinate} \in \text{GOOD is left out or any coordinate} \in \text{BAD is chosen}] \leq pe^{-\frac{\psi\epsilon k}{2s}}$$

Setting $\psi = \frac{1}{2}$, we have the R.H.S. of the above expression to be $p \exp\left(-\frac{\epsilon k}{4s}\right)$. Hence w.p. $\geq 1 - p\exp\left(-\frac{\epsilon k}{4s}\right)$, we have $\hat{\Gamma}$ as the correct support.

This completes the proof. ■

G.4. Private Sparse Linear Regression via Sample and Aggregate Framework

We look at the following linear system: $y = X\theta^* + w$, where the design matrix $X \in \mathbb{R}^{n\times p}$, output vector $y \in \mathbb{R}^{n\times 1}$, parameter vector $\theta^* \in \mathbb{R}^p$ (which is guaranteed to be s-sparse), and $w \in \mathbb{R}^{n\times 1}$ is a noise vector.

In order to obtain a private estimate of θ^*, we use Algorithm Samp-Agg (Algorithm 5) for support selection and Algorithm Obj-Pert (Algorithm 1) for privately solving the convex optimization problem restricted to the feature set selected. We define the loss function as $\hat{\mathcal{L}}(\theta; \mathcal{D}) = \frac{1}{2n}\|y - X\theta\|_2^2$. Now as discussed in Algorithm Samp-Agg , we need to instantiate the Algorithm $\mathcal{A}_{\text{supp}}$ which identifies the true support. To this end, we solve the following L_1-penalized linear regression (also known as LASSO) on each of the data blocks $\mathcal{D}_i$.

$$\hat{\theta} \in \arg\min_{\theta\in\mathbb{F}} \hat{\mathcal{L}}(\theta; \mathcal{D}_i) + \frac{\Lambda}{n}\|\theta\|_1 \tag{25}$$

One can guarantee that if the dataset follows some statistical conditions (namely, Ψ **Restricted Strong Convexity (RSC)** [Assumption Sparse-Linear' (Assumption 3)]), then $\|\hat{\theta} - \theta^*\|_2$ will be small (Negahban et al., 2010). Since θ^* has s-non-zero entries, picking the top s-coordinates of θ_{temp} (based on absolute value) will provide a good support. An implicit assumption here is that the minimum absolute value of any non-zero coordinate of θ^* is bounded away from zero.

Once the support $\hat{\Gamma}$ is chosen via sample and aggregate framework, the low dimensional problem is solved via Algorithm Obj-Pert (Algorithm 1). The details are provided in Algorithm 6. There are two main features specific to this algorithm. First, the low-dimensional convex optimization (in Line 5) is performed on a closed convex set $\mathbb{F} = \{\theta \in \mathbb{R}^p : \|\theta\|_\infty \leq 1\}$. Our privacy proof needs this bound to guarantee that $\bigtriangledown\hat{\mathcal{L}}(\theta; \mathcal{D})$ does not change by much if one entry is added (removed) to (from) $\mathcal{D}$. But bounding the convex set $\mathbb{F}$ means that we cannot use the results of objective perturbation for unconstrained optimization (originally proposed by Chaudhuri et al. (2011) (see Lemma 14)). This is where our privacy guarantee for constrained optimization (Lemma 21 and Theorem 2) becomes useful. The other feature of Algorithm 6 is that we truncate the vector y to form y_{new}. In the proof of Theorem 43 (utility theorem), we claim that truncating does not degrade the utility.

G.4.1. PRIVACY ANALYSIS

Theorem 38 *Algorithm 6 is (ϵ, δ)-differentially private.*

Proof We prove the privacy in two stages. In the first stage we prove that Line 2 of Algorithm 6 is $\frac{\epsilon}{2}$-differentially private. The proof of this directly follows from Theorem 36. In the second stage, we prove that Line 5 of Algorithm 6 is $(\frac{\epsilon}{2}, \delta)$-differentially private. Using the composition property of differential privacy (Dwork and Lei (2009)), we conclude that Algorithm 6 is (ϵ, δ)-differentially private.

Algorithm 6 Private Sparse Linear Regression

Require: dataset: $\mathcal{D} = (y, X)$, privacy parameters: ϵ and δ, sparsity parameter: Λ, dimensionality of the problem: p, number of data points: n, support size of θ^*: s, L_2 penalization: Δ

1: Define $\mathcal{A}_{\text{supp}}((y, X), s)$ as below:

- $\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{2n}\|y - X\theta\|_2^2 + \frac{\Lambda}{n}\|\theta\|_1$
- Return the top s-coordinates of $\hat{\theta}$ based on absolute value.

2: Call Algorithm 5 with parameters $\mathcal{D}, \frac{\epsilon}{2}, \mathcal{A}_{\text{supp}}, p, n, s, k = \sqrt{n}$. Store the support returned as $\hat{\Gamma}$.

3: Let $y_{new} \in \mathbb{R}^n$ s.t. $\forall i \in [n], y_{new}(i) = \begin{cases} s & \text{if } s < y_i \\ -s & \text{if } y_i < -s \\ y_i & \text{otherwise} \end{cases}$

4: For each row X_i of the matrix X, pick the top s-coordinates (in terms of absolute value) and call it v_i. If any $\|v_i\|_2 \geq \sqrt{s}$, then set $X_i = \frac{\sqrt{s}X_i}{\|X_i\|_2}$.

5: Call Algorithm 1 with the following parameters: i) Dataset $\mathcal{D} = (y_{new}, X)$, ii) Loss function $\hat{\mathcal{L}}(\theta; \mathcal{D}) = \frac{1}{2n}\|y_{new} - \langle X, \theta\rangle\|_2^2$, iii) Sensitivity parameters $\zeta = 2s^{3/2}$ and $\lambda = s$, iv) Privacy parameters $(\frac{\epsilon}{2}, \delta)$, v) Parameters s (dimensionality), n (size of the dataset), vi) Convex set $\mathbb{F} = \{\theta \in \mathbb{R}^p : \|\theta\|_\infty \leq 1, \text{supp}(\theta) \subseteq \hat{\Gamma}\}$, vii) L_2 penalization Δ.

6: **return** The output returned by Algorithm 1.

To prove that Line 5 of Algorithm 6 is $(\frac{\epsilon}{2}, \delta)$-differentially private, we first bound the term ζ which is the upper bound on $\| \bigtriangledown \ell(\theta; d)\|_2$ when $\theta \in \mathbb{F}$ is restricted to any support Γ of size at most s. The following lemma provides this bound.

Lemma 39 *Let $\theta \in \mathbb{F}$ is restricted to any support Γ of size at most s. We have $\| \bigtriangledown \ell(\theta; d)\|_2 < 2s^{3/2}$*

.

Proof Consider an $y \in [-s, s]$ and a vector $v \in \mathbb{R}^p$ s.t. restricted to any support Γ of size s, $\|x|_\Gamma\|_2 \leq \sqrt{s}$. Now consider $d = (y, x)$ to be any data entry of the dataset $\mathcal{D}$. We have the following:

$$\begin{aligned}
&\| \bigtriangledown \ell(\theta; d)\|_2 = \frac{1}{2}\| \bigtriangledown_\theta (y - \langle x|_\Gamma, \theta\rangle)^2\|_2 \\
&= \|(y - \langle x_\Gamma, \theta\rangle)x|_\Gamma{}^T\|_2 \\
&\leq \|y \cdot x|_\Gamma\|_2 + \|x|_\Gamma\|_2\|x|_\Gamma\|_1 \\
&\leq s^{3/2} + s^{3/2} \\
&< 2s^{3/2}
\end{aligned}$$

The last inequality follows from the facts that each entry of y'_{new} is between $[-s, s]$, each row of X restricted to the support Γ has L_2-norm at most $\sqrt{s}$, and $\|\theta\|_\infty \leq 1$. Here restricting any vector $x \in \mathbb{R}^p$ to support Γ means to set all the coordinates of x outside Γ to zero. ■

We now upper bound the maximum eigenvalue of $\bigtriangledown^2 \ell_\Gamma(\theta; \mathcal{D}) = \frac{1}{2}(y - x|_\Gamma^T\theta)^2$, where $y \in [-s, s]$

and $x \in \mathbb{R}^p$ with $\|x|_\Gamma\|_2 \leq \sqrt{s}$. We have $\bigtriangledown^2 \ell_\Gamma(\theta; \mathcal{D}) = x|_\Gamma^T x|_\Gamma$. From Lemma 40 below it follows that the highest eigenvalue of $x|_\Gamma^T x|_\Gamma$ is bounded by s.

Lemma 40 (Chaudhuri et al. (2011)) *Let $A = \sum_j w_j x_j^T x_j$, for $1 \times p$ vectors x_j. Then,*

$$\sum_{i=1}^{p} |\lambda_i(A)| \leq \sum_j |w_j| \cdot \|x_j\|_2^2$$

where λ_i is the i-th eigenvalue of A.

Setting $\zeta = 2s^{3/2}$ and $\lambda = s$, and invoking Theorem 2, it follows that Line 5 of Algorithm 6 is $(\frac{\epsilon}{2}, \delta)$-differentially private. ■

G.4.2. UTILITY ANALYSIS

Lemma 41 *Let $\Lambda = 4\sigma n^{1/4}\sqrt{\log p}$ and $\Delta = \Theta(s/\epsilon)$. Under Assumption* Sparse-Linear' *(Assumption 3) on the design matrices $X_1, \cdots, X_{\sqrt{n}}$ and noise vectors $w_1, \cdots, w_{\sqrt{n}}$ (corresponding to data blocks $\mathcal{D}_1, \cdots, \mathcal{D}_{\sqrt{n}}$), if $n \geq (\frac{16\sigma}{\Psi\Phi})^4 s^2 \log^2 p$, then Algorithm $\mathcal{A}_{supp}$ (in Line 1 of Algorithm 6) outputs a set of size s which contains the correct support of θ^*.*

Lemma 41 follows from (Negahban et al., 2010, Corollary 2). It roughly states that under Assumption 3 (Assumption Sparse-Linear'), with high probability the L_2 distance between $\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{2n}\|y - X\theta\|_2^2 + \frac{\Lambda}{n}\|\theta\|_1$ and θ^* roughly goes down as $\sqrt{\frac{s \log p}{n}}$, where Λ is an appropriately chosen parameter. This means that for sufficiently large n, w.h.p. the support of the top-s coordinates of $\hat{\theta}$ is the support for θ^*. Following is a detailed proof for the above lemma.

Proof In order to prove Lemma 41, we first state the following lemma from Negahban et al. (2010).

Lemma 42 (Corollary 2 from Negahban et al. (2010)) *Consider the optimization problem defined as $\hat{\theta} \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{2n}\|y - X\theta\|_2^2 + \frac{1}{n}\Lambda\|\theta\|_1$, where $y \in \mathbb{R}^n$ is the response vector, $X^{n \times p}$ is the design matrix. Under the assumption that the tuple (X, w) is (s,σ,Ψ)-well behaved, if we set $\Lambda = 4\sigma\sqrt{n \log p}$, then we have*

$$\|\hat{\theta} - \theta^*\|_2^2 \leq \frac{64\sigma^2}{\Psi^2} \frac{s \log p}{n}$$

Here w is the noise in the linear system.

In order to use the above lemma, we observe the following. First, any design matrix (subsampled from the original design matrix X) on which $\mathcal{A}_{\text{supp}}$ (defined in Line 1 of Algorithm 6) executes has $\sqrt{n}$ number of rows. (Here n is the number of rows in the original design matrix X.) Second, note that by Assumption 3 (Assumption Sparse-Linear') each of the $\sqrt{n}$ design matrices on which $\mathcal{A}_{\text{supp}}$ executes in Algorithm 6 follows restricted strong convexity. Thus, from Lemma 42 the following is true for each data block $\mathcal{D}_i$.

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{8\sigma}{\Psi} \frac{\sqrt{s \log p}}{n^{1/4}} \tag{26}$$

Let Φ be the minimum absolute value of any non-zero coordinate of θ^*. Now, setting the R.H.S. of Equation 26 to $\frac{\Phi}{2}$, we have $n \geq (\frac{16\sigma}{\alpha\Phi})^4 s^2 \log^2 p$. This means that when $n \geq (\frac{16\sigma}{\Psi\Phi})^4 s^2 \log^2 p$, for any coordinate $i \in [p]$ where $\theta^*(i) \neq 0$, we have $|\hat{\theta}(i)| \geq \frac{\Phi}{2}$. Similarly, for any coordinate $i \in [p]$ where $\theta^*(i) = 0$, we have $|\hat{\theta}(i)| < \frac{\Phi}{2}$.

Thus, if we pick the top s coordinates (in terms of absolute value), then all the coordinates in the support of θ^* will be chosen. This completes the proof. ■

Theorem 43 (Theorem 44, special case) *Let* $\Lambda = 4\sigma n^{1/4}\sqrt{\log p}$ *and* $\Delta = \Theta\left(s/\epsilon\right)$. *Under Assumption* Sparse-Linear' *(Assumption 3), if* $n \geq (\frac{16\sigma}{\Psi\Phi})^4 s^2 \log^2 p$, *then w.p.* $\geq 1 - \left(p \exp\left(-\frac{\epsilon\sqrt{n}}{8s}\right)\right)$ *we have* $\mathbb{E}_b\left[\hat{\mathcal{L}}(\theta^{priv}; \mathcal{D}) - \hat{\mathcal{L}}(\theta^*; \mathcal{D})\right] = O\left(\frac{1}{n\epsilon}\left(\frac{s^4 \log(1/\delta)}{n\epsilon\Psi} + s^2\right)\right)$. *Here* b *is the noise vector in Algorithm* Obj-Pert *(Algorithm 1).*

The above theorem (Theorem 43) almost follows directly from Lemma 41, and Theorems 4 and 8. In order to prove Theorem 43, we first prove a slightly general version below. Plugging in the value of $\Delta = \Theta\left(s/\epsilon\right)$ yields Theorem 43. Since we are dealing with expectation in Theorem 43, we ignore the term γ.

Theorem 44 (Utility) *Under the conditions of Lemma 41, w.p.* $\geq 1 - \left(p \exp\left(-\frac{\epsilon\sqrt{n}}{8s}\right) + \gamma\right)$ *we have*

$$\hat{\mathcal{L}}(\theta^{priv}; \mathcal{D}) - \hat{\mathcal{L}}(\theta^*; \mathcal{D}) \leq \frac{64 s^4 \left(8 \log \frac{2}{\delta} + 2\epsilon\right) \log \frac{1}{\gamma}}{n\epsilon^2(\Delta + \Psi)} + \frac{\Delta}{2n}\|\theta^{Emp}\|_2^2$$

Here $\theta^{Emp} \in \arg\min_{\theta \in \mathbb{F}_{\hat{\Gamma}}} \hat{\mathcal{L}}(\theta; \mathcal{D})$.

Proof We prove this theorem in two stages. In the first stage, we lower bound the probability with which the correct support of θ^* is chosen in Line 2 of Algorithm 6. In the second stage, we bound the empirical risk for the optimization problem (restricted to the support $\hat{\Gamma}$ chosen). Note that if the correct support of θ^* is chosen, then the empirical risk bound of the second stage corresponds to the empirical risk bound for the actual problem. We conclude the proof by combining the failure probabilities of stages one and two.

Stage one: From Lemma 41, it follows that Algorithm $\mathcal{A}_{\text{supp}}$ (see Line 1 of Algorithm 6) outputs the correct support. Once the correct support is chosen, the problem reduces to an s-dimensional problem.

Stage two: We complete the proof of *stage two* by invoking Theorem 31 with parameters $\zeta = 2s^{3/2}$, $k = \sqrt{n}$ and dimensionality of the problem $= s$.

■

Appendix H. Low-dimensional linear regression

Consider the linear regression problem,

$$y = X\theta^* + w \tag{27}$$

where the design matrix $X \in \mathbb{R}^{n\times p}$, output vector $y \in \mathbb{R}^{n\times 1}$, parameter vector $\theta^* \in \mathbb{R}^p$, and $w \in \mathbb{R}^{n\times 1}$ is a noise vector. We define the loss function for any given θ as $\hat{\mathcal{L}}(\theta; \mathcal{D}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \langle X_i, \theta\rangle)^2$, where y_i is the i-th entry in the vector y and X_i is the i-th row of the matrix X. The setting we are interested in is where each row of the design matrix X has L_2 norm at most $\sqrt{p}$ and the parameter vector θ^* has L_2 norm at most $\sqrt{p}$. Notice that since we assume θ^* and the rows of X have norm at most $\sqrt{p}$, so truncating y into $[-p, p]$ will not hurt utility guarantees. Therefore, w.l.o.g. we assume that $y \in [-p, p]$. Also, since θ^* is assumed to have norm at most $\sqrt{p}$, we assume that the convex set over which the optimization is performed has L_2 norm at most $\sqrt{p}$, *i.e.,* $\mathbb{F} = \left\{\theta \in \mathbb{R}^p : \|\theta\|_2 \leq \sqrt{p}\right\}$.

A natural question is "when does the above parameters setting is meaningful?". One possible setting where it is meaningful is when each entry of the design matrix X is assumed to be constant and each each entry of the parameter vector θ^* is also assumed to be constant.

Under this setting we want to bound the gradient of $\frac{1}{2}(y_i - \langle X_i, \theta\rangle)^2$ by ζ for any $\theta \in \mathbb{F}$. It is easy to see that the gradient is $X_i^T(y_i - \langle X_i, \theta\rangle)$. Therefore, under the choice of parameters in the problem we have $\zeta = 2p^{3/2}$.

Similarly, to bound the maximum eigenvalue of $\bigtriangledown^2 \frac{1}{2}(y_i - \langle X_i, \theta\rangle)^2$ by λ, we first notice that the hessian is $X_i^T X_i$. Since $\|X_i\|_2 \leq \sqrt{p}$, the maximum eigenvalue of the matrix is p. Hence, we can set $\lambda = p$.