Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain

Yu-Xiang Wang *
University of California, Santa Barbara, CA
Amazon AWS AI, Palo Alto, CA

Abstract

We revisit the problem of linear regression under a differential privacy constraint. By consolidating existing pieces in the literature, we clarify the correct dependence of the feature, label and coefficient domains in the optimization error and estimation error, hence revealing the delicate price of differential privacy in statistical estimation and statistical learning. Moreover, we propose simple modifications of two existing DP algorithms: (a) posterior sampling, (b) sufficient statistics perturbation, and show that they can be *upgraded* into adaptive algorithms that are able to exploit data-dependent quantities and behave nearly optimally *for every instance*. Extensive experiments are conducted on both simulated data and real data, which conclude that both ADAOPS and ADASSP outperform the existing techniques on nearly all 36 data sets that we test on.

1 INTRODUCTION

Linear regression is one of the oldest tools for data analysis (Galton, 1886) and it remains one of the most commonly-used as of today (Draper & Smith, 2014), especially in social sciences (Agresti & Finlay, 1997), econometics (Greene, 2003) and medical research (Armitage et al., 2008). Moreover, many nonlinear models are either intrinsically linear in certain function spaces, e.g., kernels methods, dynamical systems, or can be reduced to solving a sequence of linear regressions, e.g., iterative reweighted least square for generalized Linear models, gradient boosting for additive models and so on (see Friedman et al., 2001, for a detailed review).

In order to apply linear regression to sensitive data such as those in social sciences and medical studies, it is often needed to do so such that the privacy of individuals in the data set is protected. Differential privacy (Dwork et al., 2006b) is a commonly-accepted criterion that provides provable protection against identification and is resilient to arbitrary auxiliary information that might be available to attackers. In this paper, we focus on linear regression with (ϵ, δ)-differentially privacy (Dwork et al., 2006a).

Isn't it a solved problem? It might be a bit surprising why this is still a problem, since several general frameworks of differential privacy have been proposed that cover linear regression. Specifically, in the agnostic setting (without a data model), linear regression is a special case of differentially private empirical risk minimization (ERM), and its theoretical properties have been quite well-understood in a sense that the minimax lower bounds are known (Bassily et al., 2014) and a number of algorithms (Chaudhuri et al., 2011; Kifer et al., 2012) have been shown to match the lower bounds under various assumptions. In the statistical estimation setting where we assume the data is generated from a linear Gaussian model, linear regression is covered by the sufficient statistics perturbation approach for exponential family models (Dwork & Smith, 2010; Foulds et al., 2016), propose-test-release framework (Dwork & Lei, 2009) as well as the the subsample-and-aggregate framework (Smith, 2008), with all three approaches achieving the asymptotic efficiency in the fixed dimension ($d = O(1)$), large sample ($n \to \infty$) regime.

Despite these theoretical advances, very few empirical evaluations of these algorithms were conducted and we are not aware of a commonly-accepted best practice. Practitioners are often left puzzled about which algorithm to use for the specific data set they have. The nature of differential privacy often requires them to set parameters of the algorithm (e.g., how much noise to add) according to the diameter of the parameter domain, as well as properties of a hypothetical worst-case data set, which often leads to

*Corresponding email: `yuxiangw@cs.ucsb.edu` an inefficient use of their valuable data.

The main contribution of this paper is threefold:

1. We consolidated many bits and pieces from the literature and clarified the price of differentially privacy in statistical estimation and statistical learning.
2. We carefully analyzed One Posterior Sample (OPS) and Sufficient Statistics Perturbation (SSP) for linear regression and proposed simple modifications of them into adaptive versions: ADAOPS and ADASSP. Both work near optimally for every problem instance without any hyperparameter tuning.
3. We conducted extensive real data experiments to benchmark existing techniques and concluded that the proposed techniques give rise to the more favorable privacy-utility tradeoff relative to existing methods.

Outline of this paper. In Section 2 we will describe the problem setup and explain differential privacy. In Section 3, we will survey the literature and discuss existing algorithms. Then we will propose and analyze our new method ADASSP and ADAOPS in Section 4 and conclude the paper with experiments in Section 5.

2 NOTATIONS AND SETUP

Throughout the paper we will use $X \in \mathbb{R}^{n\times d}$ and $\boldsymbol{y} \in \mathbb{R}^n$ to denote the design matrix and response vector. These are collections of data points $(x_1, y_1), ..., (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$. We use $\|\cdot\|$ to denote Euclidean norm for vector inputs, ℓ_2-operator norm for matrix inputs. In addition, for set inputs, $\|\cdot\|$ denotes the radius of the smallest Euclidean ball that contains the set. For example, $\|\mathcal{Y}\| = \sup_{y\in\mathcal{Y}} |y|$ and $\|\mathcal{X}\| = \sup_{x\in\mathcal{X}} \|x\|$. Let Θ be the domain of coefficients. Our results do not require Θ to be compact but existing approaches often depend on $\|\Theta\|$. $\lesssim$ and $\gtrsim$ denote greater than or smaller to up to a universal multiplicative constant, which is the same as the big $O(\cdot)$ and the big $\Omega(\cdot)$. $\tilde{O}(\cdot)$ hides at most a logarithmic term. $\prec$ and $\succ$ denote the standard semidefinite ordering of positive semi-definite (psd) matrices. $\cdot \vee \cdot$ and $\cdot \wedge \cdot$ denote the bigger or smaller of the two inputs.

We now define a few data dependent quantities. We use $\lambda_{\min}(X^TX)$ (abbv. $\lambda_{\min}$) to denote the smallest eigenvalue of X^TX, and to make the implicit dependence in d and n clear from this quantity, we define $\alpha := \lambda_{\min}\frac{d}{n\|\mathcal{X}\|^2}$. One can think of α as a normalized smallest eigenvalue of X^TX such that $0 \leq \alpha \leq 1$. Also, $1/\alpha$ is closely related to the condition number of X^TX.

Define the least square solution $\theta^* = (X^TX)^\dagger X^T\boldsymbol{y}$. It is the optimal solution to $\min_\theta \frac{1}{2}\|\boldsymbol{y} - X\theta\|^2 =: F(\theta)$. Similarly, we use $\theta^*_\lambda = (X^TX + \lambda I)^{-1}X^T\boldsymbol{y}$ denotes the optimal solution to the ridge regression objective $F_\lambda(\theta) = F(\theta) + \lambda\|\theta\|^2$.

In addition, we denote the global Lipschitz constant of F as $L^* := \|\mathcal{X}\|^2\|\Theta\| + \|\mathcal{X}\|\|\mathcal{Y}\|$ and data-dependent local Lipschitz constant at θ^* as $L := \|\mathcal{X}\|^2\|\theta^*\| + \|\mathcal{X}\|\|\mathcal{Y}\|$. Note that when $\Theta = \mathbb{R}^d$, $L^* = \infty$, but L will remain finite for every given data set.

Metric of success. We measure the performance of an estimator $\hat{\theta}$ in two ways.

First, we consider the optimization error $F(\hat{\theta}) - F(\theta^*)$ in expectation or with probability $1 - \varrho$. This is related to the prediction accuracy in the distribution-free statistical learning setting.

Second, we consider how well the coefficients can be estimated under the linear Gaussian model:

$$\boldsymbol{y} = X\theta_0 + \mathcal{N}(0, \sigma^2 I_n)$$

in terms of $\mathbb{E}[\|\hat{\theta} - \theta_0\|^2]$ or in some cases $\mathbb{E}[\|\hat{\theta} - \theta_0\|^2|E]$ where E is a high probability event.

The optimal error in either case will depend on the specific design matrix X, optimal solution θ^*, the data domain $\mathcal{X}, \mathcal{Y}$, the parameter domain Θ as well as θ_0, σ^2 in the statistical estimation setting.

Differential privacy. We will focus on estimators that are differential private, as defined below.

Definition 1 (Differential privacy (Dwork et al., 2006b)). *We say a randomized algorithm $\mathcal{A}$ satisfies (ϵ, δ)-DP if for all fixed data set $(X, \boldsymbol{y})$ and data set $(X', \boldsymbol{y}')$ that can be constructed by adding or removing one row (x, y) from $(X, \boldsymbol{y})$, and for any measurable set $\mathcal{S}$ over the probability of the algorithm*

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

Parameter ϵ represents the amount of privacy loss from running the algorithm and δ denotes a small probability of failure. These are user-specified targets to achieve and the differential privacy guarantee is considered meaningful if $\epsilon \leq 1$ and $\delta \ll 1/n$ (see, e.g., Section 2.3.3 of Dwork et al., 2014a, for a comprehensive review).

The pursuit for adaptive estimators. Another important design feature that we will mention repeatedly in this paper is *adaptivity*. We call an estimator $\hat{\theta}$ *adaptive* if it behaves optimally simultaneously for a wide range of parameter choices. Being adaptive is of great practical relevance because we do not need to specify the class of problems or worry about whether our specification is wrong (see examples of adaptive estimators in e.g., Donoho, 1995; Birgé & Massart, 2001). *Adaptivity* is particularly important for differentially private data analysis because often we need to decide the amount of noise to add by the size of the domain. For example, an adaptive algorithm will not rely on conservative upper bounds of θ_0, or a worst case $\lambda_{\min}$ (which would be 0 on any $\mathcal{X}$), and it can take advantage of favorable properties when they exist in the data set. We want to design an estimator that does not take these parameters as inputs and behave nearly optimally for every fixed data set $X \in \mathcal{X}^n, \boldsymbol{y} \in \mathcal{Y}$ under a variety of configuration of $\|\mathcal{X}\|, \|\mathcal{Y}\|, \|\Theta\|$.

3 A SURVEY OF PRIOR WORK

In this section, we summarize existing theoretical results in linear regression with and without differential privacy constraints. We will start with lower bounds.

3.1 Information-theoretic lower bounds

Lower bounds under linear Gaussian model. Under the statistical assumption of linear Gaussian model $\boldsymbol{y} = X\theta_0 + \mathcal{N}(0, \sigma^2)$, the minimax risk for both estimation and prediction are crisply characterized for each fixed design matrix X:

$$\inf_{\hat{\theta}} \sup_{\theta_0 \in \mathbb{R}^d} \mathbb{E}[F(\hat{\theta}) - F(\theta_0)|X] = \frac{d\sigma^2}{2}, \qquad (1)$$

and if we further assume that $n \geq d$ and X^TX is invertible (for identifiability), then

$$\inf_{\hat{\theta}} \sup_{\theta_0 \in \mathbb{R}^d} \mathbb{E}[\|\hat{\theta} - \theta_0\|_2^2|X] = \sigma^2 \mathrm{tr}[(X^TX)^{-1}]. \qquad (2)$$

In the above setup, $\hat{\theta}$ is any measurable function of $\hat{y}$ (note that X is fixed). These are classic results that can be found in standard statistical decision theory textbooks (See, e.g., Wasserman, 2013, Chapter 13).

Under the same assumptions, the Cramer-Rao lower bound mandates that the covariance matrix of any unbiased estimator $\hat{\theta}$ of θ_0 to obey that

$$\mathrm{Cov}(\hat{\theta}) \succ \sigma^2 (X^TX)^{-1}. \qquad (3)$$

This bound applies to every problem instance separately and also implies a sharp lower bound on the prediction variance on every data point x. More precisely, $\mathrm{Var}(\hat{\theta}^T x) \geq \sigma^2 x^T (X^TX)^{-1} x$ for any x.

Minimax risk (1), (2) and the Cramer-Rao lower bound (3) are simultaneously attained by θ^*.

Statistical learning lower bounds. Perhaps much less well-known, linear regression is also thoroughly studied in the distribution-free statistical learning setting, where the only assumption is that the data are drawn iid from some unknown distribution $\mathcal{P}$ defined on some compact domain $\mathcal{X} \times \mathcal{Y}$. Specifically, let the risk ($\mathbb{E}$[loss]) be

$$R(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{P}}[\tfrac{1}{2}(x^T\theta - y)^2] = \tfrac{1}{n}\mathbb{E}_{(X,\boldsymbol{y})\sim\mathcal{P}^n}[F(\theta)].$$

Shamir (2015) showed that when Θ, $\mathcal{X}$ are $\mathcal{Y}$ are Euclidean balls,

$$\inf_{\hat{\theta}} \sup_{\mathcal{P}} \left[\mathbb{E}[n \cdot R(\hat{\theta})] - \inf_{\theta \in \Theta}[n \cdot R(\theta)]\right] \gtrsim \min\{n\|\mathcal{Y}\|^2, \|\Theta\|^2\|\mathcal{X}\|^2 + d\|\mathcal{Y}\|^2, \sqrt{n}\|\Theta\|\|\mathcal{X}\|\|\mathcal{Y}\|\}. \qquad (4)$$

where $\hat{\theta}$ be any measurable function of the data set $X, \boldsymbol{y}$ to Θ and the expectation is taken over the data generating distribution $X, \boldsymbol{y} \sim \mathcal{P}^n$. Note that to be compatible to other bounds that appear in this paper, we multiplied the $R(\cdot)$ by a factor of n. Informally, one can think of $\|\mathcal{Y}\|$ as σ in (1) so both terms depend on $d\sigma^2$ (or $d\|\mathcal{Y}\|^2$), but the dependence on $\|\Theta\|\|\mathcal{X}\|$ is new for the distribution-free setting.

Koren & Levy (2015) later showed that this lower bound is matched up to a constant by Ridge Regression with $\lambda = 1$ and both Koren & Levy (2015) and Shamir (2015) conjecture that ERM without additional regularization should attain the lower bound (4). If the conjecture is true, then the unconstrained OLS is simultaneously optimal for all distributions supported on the smallest ball that contains all data points in $X, \boldsymbol{y}$ for any Θ being an ℓ_2 ball with radius larger than $\|\theta^*\|$.

Lower bounds with (ϵ, δ)-privacy constraints. Suppose that we further require $\hat{\theta}$ to be (ϵ, δ)-differentially private, then there is an additional price to pay in terms of how accurately we can approximate the ERM solution. Specifically, the lower bounds for the *empirical* excess risk for differentially private ERM problem in (Bassily et al., 2014) implies that for $\delta < 1/n$ and sufficiently large n:

1. There exists a triplet of $(\mathcal{X}, \mathcal{Y}, \Theta) \subset \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d$, such that

$$\inf_{\hat{\theta} \text{ is } (\epsilon,\delta)\text{-DP}} \sup_{X \in \mathcal{X}^n, \boldsymbol{y} \in \mathcal{Y}^n} \left[F(\hat{\theta}) - \inf_{\theta \in \Theta} F(\theta)\right] \gtrsim \min\{n\|\mathcal{Y}\|^2, \frac{\sqrt{d}(\|\mathcal{X}\|^2\|\Theta\|^2 + \|\mathcal{X}\|\|\Theta\|\|\mathcal{Y}\|)}{\epsilon}\}. \qquad (5)$$

2. Consider the class of data set $\mathcal{S}$ where all data sets $X \in \mathcal{S} \subset \mathcal{X}^n$ obeys that the inverse condition number $\alpha \geq \alpha^* \geq \frac{d^{1.5}(\|\mathcal{X}\|\|\Theta\|+\|\mathcal{Y}\|)}{n\|\mathcal{X}\|\|\Theta\|\epsilon}$ [1]. There exists a

[1] This requires $\lambda_{\min} \geq \sqrt{d}L/\epsilon$ for all data sets X.

triplet of $(\mathcal{X}, \mathcal{Y}, \Theta) \subset \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d$ such that

$$\inf_{\hat{\theta} \text{ is } (\epsilon,\delta)\text{-DP}} \sup_{X \in \mathcal{S}, \boldsymbol{y} \in \mathcal{Y}^n} \left[F(\hat{\theta}) - \inf_{\theta \in \Theta} F(\theta) \right] \tag{6}$$
$$\gtrsim \min\{n\|\mathcal{Y}\|^2, \frac{d^2(\|\mathcal{X}\|\|\Theta\| + \|\mathcal{Y}\|)^2}{n\alpha^* \epsilon^2}\}.$$

These bounds are attained by a number of algorithms, which we will go over in Section 3.2.

Comparing to the non-private minimax rates on prediction accuracy, the bounds look different in several aspects. First, neither rate for prediction error in (1) or (4) depends on whether the design matrix X is well-conditioned or not, while α^* appears explicitly in (6). Secondly, the dependence on $\|\Theta\|\|\mathcal{X}\|, \|\mathcal{Y}\|, d, n$ are different, which makes it hard to tell whether the optimization error lower bound due to privacy requirement is limiting. One may ask the following question:

When is privacy *for free* in statistical learning?

Specifically, what is the smallest ϵ such that an (ϵ, δ)-DP algorithm matches the minimax rate in (4)? The answer really depends on the relative scale of $\|\mathcal{X}\|\|\Theta\|$ and $\|\mathcal{Y}\|$ and that of n, d. When $\|\mathcal{X}\|\|\Theta\| \asymp \|\mathcal{Y}\|$, (5) says that (ϵ, δ)-DP algorithms can achieve the nonconvex minimax rate provided that $\epsilon \gtrsim \min\left\{\frac{1}{\sqrt{d}} \vee \sqrt{\frac{d}{n}}, \sqrt{\frac{d^2}{n^{1.5}\alpha^*}} \vee \sqrt{\frac{d}{n\alpha^*}}\right\}$. On the other hand, if $\|\mathcal{X}\|\|\Theta\| \asymp \sqrt{d}\|\mathcal{Y}\|$ [2] and $n > d$, then we need $\epsilon \gtrsim \min\left\{\sqrt{d} \vee \frac{d^{3/2}}{n}, \frac{d}{\sqrt{n\alpha^*}} \vee \frac{d^{3/2}}{n\sqrt{\alpha^*}}\right\}$.

The regions are illustrated graphically in Figure 1. In the first case, there is a large region upon $n \gtrsim d$, where meaningful differential privacy (with $\epsilon \leq 1$ and $\delta = o(1/n)$) can be achieved without incurring a significant toll relative to (4). In the second case, we need at least $n \gtrsim d^2$ to achieve "privacy-for-free" in the most favorable case where $\alpha^* = 1$. In the case when X could be rank-deficient, then it is infeasible to achieve "privacy for free" no matter how large n is.

It might be tempting to conclude that one should always prefer Case 1 over Case 2. This is unfortunately not true because the artificial restriction of the model class via a bounded $\|\Theta\|$ also weakens our non-private baseline. In other word, the best solution within a small Θ might be significantly worse than the best solution in $\mathbb{R}^d$.

In practice, it is hard to find a Θ with a small radius that fits all purposes[3] and it is unreasonable to assume $\alpha^* > 0$.

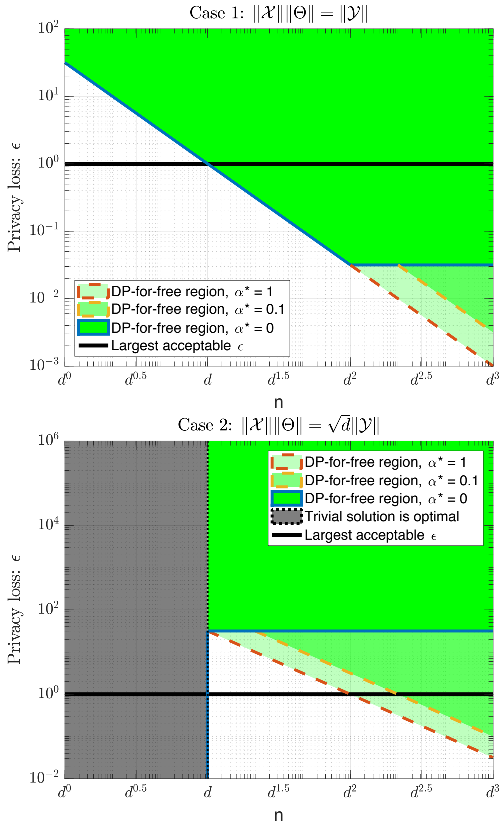


Figure 1: Illustration of the region of ϵ where DP can be obtained without losing minimax rate (4).[Zoom to see!]

This motivates us to go beyond the worst-case and come up with *adaptive* algorithms that work without knowing $\|\theta^*\|$ and α while achieving the minimax rate for the class with $\|\Theta\| = \|\theta^*\|$ and $\alpha^* = \alpha$ (in hindsight).

In Appendix B, we provide an alternative illustration of the lower bounds and highlight the price of differential privacy for different configuration of n, d, α, ϵ.

3.2 Existing algorithms and our contribution

We now survey the following list of five popular algorithms in differentially private learning and highlight the novelty in our proposals [4].

1. Sufficient statistics perturbation (SSP) (Vu & Slavkovic, 2009; Foulds et al., 2016): Release $X^T X$ and $X\boldsymbol{y}$ differential privately and then output $\hat{\theta} = (\widehat{X^T X})^{-1}\widehat{X\boldsymbol{y}}$.
2. Objective perturbation (ObjPert) (Kifer et al.,

[2] This is arguably the more relevant setting. Note that if $x \sim \mathcal{N}(0, I_d)$ and θ is fixed, then $x^T\theta = O_P(d^{-1/2}\|x\|\|\theta\|)$.

[3] If $\|\Theta\| \gg \|\theta^*\|$ then the constraint becomes limiting. If $\|\theta^*\| \ll \|\Theta\|$ instead, then calibrating the noise according to $\|\Theta\|$ will inject more noise than necessary.

[4] While we try to be as comprehensive as possible, the literature has grown massively and the choice of this list is limited by our knowledge and opinions.

2012): $\hat{\theta} = \mathrm{argmin}\, F(\theta) + 0.5\lambda\|\theta\|^2 + Z^T\theta$ with an appropriate λ and Z is an appropriately chosen iid Gaussian random vector.

3. Subsample and Aggregate (Sub-Agg) (Smith, 2008; Dwork & Smith, 2010): Subsample many times, apply debiased MLE to each subset and then randomize the way we aggregate the results.

4. Posterior sampling (OPS) (Mir, 2013; Dimitrakakis et al., 2014; Wang et al., 2015; Minami et al., 2016): Output $\hat{\theta} \sim P(\theta) \propto e^{-\gamma(F(\theta)+0.5\lambda\|\theta\|^2)}$ with parameters γ, λ.

5. NoisySGD (Bassily et al., 2014): Run SGD for a fixed number of iterations with additional Gaussian noise added to the stochastic gradient evaluated on one randomly-chosen data point.

We omit detailed operational aspects of these algorithms and focus our discussion on their theoretical guarantees. Interested readers are encouraged to check out each paper separately. These algorithms are proven under different scalings and assumptions. To ensure fair comparison, we make sure that all results are converted to our setting under a subset of the following assumptions.

A.1 $\|\mathcal{X}\|$ is bounded, $\|\mathcal{Y}\|$ is bounded.

A.2 $\|\Theta\|$ is bounded.

A.3 All possible data sets X obey that the smallest eigenvalue $\lambda_{\min}(X^TX)$ is greater than $\frac{n\|\mathcal{X}\|^2}{d}\alpha^*$.

Note that A.3 is a restriction on the domain of the data set, rather than the domain of individual data points in the data set of size n. While it is a little unconventional, it is valid to define differential privacy within such a restricted space of data sets. It is the same assumption that we needed to assume for the lower bound in (6) to be meaningful. As in Koren & Levy (2015), we simplify the expressions of the bound by assuming $\|\mathcal{Y}\| \leq \|\mathcal{X}\|\|\Theta\|$, and in addition, we assume that $\|\mathcal{Y}\| \lesssim \|\mathcal{X}\|\|\theta^*\|$.

Table 1 summarizes the upper bounds of optimization error the aforementioned algorithms in comparison to our two proposals: AdaOPS and AdaSSP. Comparing the rates to the lower bounds in the previous section, it is clear that NoisySGD, ObjPert both achieve the minimax rate in optimization error but their hyperparameter choice depends on the unknown $\|\Theta\|$ and α^*. SSP is adaptive to α and $\|\theta^*\|$ but has a completely different type of issue — it can fail arbitrarily badly for regime covered under (5), and even for well-conditioned problems, its theoretical guarantees only kick in as n gets very large. Our proposed algorithms AdaOPS and AdaSSP are able to simultaneously switch between the two regimes and get the best of both worlds.

Table 2 summarizes the upper bounds for estimation. The second row compares the approximation of θ^* in MSE and the third column summarizes the statistical efficiency of the DP estimators relative to the MLE: θ^* under the linear Gaussian model. All algorithms except OPS are asymptotically efficient. For the interest of (ϵ, δ)-DP, SSP has the fastest convergence rate and does not explicitly depend on the smallest eigenvalue, but again it behaves differently when n is small, while AdaOPS and AdaSSP work optimally (up to a constant) for all n.

3.3 Other related work

The problem of adaptive estimation is closely related to model selection (see, e.g., Birgé & Massart, 2001) and an approach using Bayesian Information Criteria was carefully studied in the differential private setting for the problem of ℓ_1 constrained ridge regression by Lei et al. (2017). Their focus is different to ours in that they care about inferring the correct model, while we take the distribution-free view. Linear regression is also studied in many more specialized setups, e.g., high dimensional linear regression (Kifer et al., 2012; Talwar et al., 2014, 2015), statistical inference (Sheffet, 2017) and so on. For the interest of this paper, we focus on the standard regime of linear regression where $d < n$ and do not use sparsity or ℓ_1 constraint set to achieve the $\log(d)$ dependence. That said, we acknowledge that Sheffet (2017) analyzed SSP under the linear Gaussian model (the third row in Table 2and their techniques of adaptively adding regularization have inspired AdaSSP.

4 MAIN RESULTS

In this section, we present and analyze AdaOPS and AdaSSP that achieve the aforementioned adaptive rate. The pseudo-code of these two algorithms are given in Algorithm 1 and Algorithm 2.

The idea of both algorithms is to release key data-dependent quantities differentially privately and then use a high probability confidence interval of these quantities to calibrate the noise to privacy budget as well as to choose the ridge regression's hyperparameter λ for achieving the smallest prediction error. Specifically, AdaOPS requires us to release both the smallest eigenvalue $\lambda_{\min}$ of X^TX and the local Lipschitz constant $L := \|\mathcal{X}\|(\|\mathcal{X}\|\|\theta^*_\lambda\| + \|\mathcal{Y}\|)$, while AdaSSP only needs the smallest eigenvalue $\lambda_{\min}$.

In both AdaSSP and AdaOPS, we choose λ by minimizing an upper bound of $F(\tilde{\theta}) - F(\theta^*)$ in the form of "variance" and "bias"

$$\tilde{O}(\frac{d\|\mathcal{X}\|^4\|\theta^*\|^2}{\lambda + \lambda_{\min}}) + \lambda\|\theta^*\|^2.$$

Table 1: Summary of optimization error bounds. This table compares the (expected or high probability) additive suboptimality of different differentially private linear regression procedures relative to the (non-private) empirical risk minimizer θ^*. In particular, the results for NoisySGD holds in expectation and everything else with probability $1 - \varrho$ (hiding at most a logarithmic factor in $\sqrt{1/\varrho}$). Constant factors are dropped for readability.

	$F(\hat{\theta}) - F(\theta^*)$	Assumptions	Remarks
NoisySGD	$\frac{\sqrt{d\log(\frac{n}{\delta})}\Vert\mathcal{X}\Vert^2\Vert\Theta\Vert^2}{\epsilon}$	A.1, A.2	Theorem 2.4 (Part 1) of (Bassily et al., 2014).
	$\frac{d^2\log(\frac{n}{\delta})\Vert\Theta\Vert^2}{\alpha^* n\epsilon^2}$	A.1, A.2, A.3	Theorem 2.4 (Part 2) of (Bassily et al., 2014)
ObjPert	$\frac{\sqrt{d\log(\frac{1}{\delta})}\Vert\mathcal{X}\Vert^2\Vert\Theta\Vert\Vert\theta^*\Vert}{\epsilon}$	A.1, A.2	Theorem 4 (Part 2) of (Kifer et al., 2012).
	$\frac{d^2\log(\frac{1}{\delta})\Vert\Theta\Vert^2}{\alpha^* n\epsilon^2}$	A.1, A.2, A.3	Theorem 5 & Appendix E.2 of (Kifer et al., 2012).
OPS	$\frac{d\Vert\mathcal{X}\Vert^2\Vert\Theta\Vert^2}{\epsilon}$	A.1, A.2	Results for ϵ-DP (Wang et al., 2015)
SSP	$\frac{d^2\log(\frac{1}{\delta})\Vert\mathcal{X}\Vert^2\Vert\theta^*\Vert^2}{\alpha n\epsilon^2}$	A.1	Adaptive to $\Vert\theta^*\Vert, X, \alpha$, but requires $n = \Omega(\frac{d^{1.5}\log(4/\delta)}{\alpha\epsilon})$ [5].
AdaOPS & AdaSSP	$\frac{\sqrt{d\log(\frac{1}{\delta})}\Vert\mathcal{X}\Vert^2\Vert\theta^*\Vert^2}{\epsilon} \wedge \frac{d^2\log(\frac{1}{\delta})\Vert\theta^*\Vert^2}{\alpha n\epsilon^2}$	A.1	Adaptive in $\Vert\theta^*\Vert, X, \alpha$.

Table 2: Summary or estimation error bounds under the linear Gaussian model. On the second column we compare the approximation of MLE θ^* in mean square error up to a universal constant. On the third column, we compare the relative efficiency. The relative efficiency bounds are simplified with the assumption of $\alpha = \Omega(1)$, which implies that $\mathrm{tr}[(X^TX)^{-1}] = O(d^2n^{-1}\|\mathcal{X}\|^{-2})$ and $\mathrm{tr}[(X^TX)^{-2}] = O(dn^{-1}\|\mathcal{X}\|^{-2}\mathrm{tr}[(X^TX)^{-1}])$. $\tilde{O}(\cdot)$ hides $\mathrm{polylog}(1/\delta)$ terms.

	Approxi. MLE: $\mathbb{E}\Vert\hat{\theta} - \theta^*\Vert^2$	Rel. efficiency: $\frac{\mathbb{E}\Vert\hat{\theta}-\theta_0\Vert^2}{\mathbb{E}\Vert\theta^*-\theta_0\Vert^2}$	Remarks
Sub-Agg	$O\left(\frac{\mathrm{poly}(d,\Vert\Theta\Vert,\Vert\mathcal{X}\Vert,\alpha^{-1})}{\epsilon^{6/5}n^{6/5}}\right)$	$1 + \tilde{O}(\frac{\mathrm{poly}(d,\Vert\Theta\Vert,\Vert\mathcal{X}\Vert)}{n^{1/5}\epsilon^{6/5}})$	ϵ-DP, suboptimal in n, possibly also in d(Dwork & Smith, 2010).
OPS	$O(\frac{\Vert\mathcal{X}\Vert^2\Vert\Theta\Vert^2}{\epsilon})\mathrm{tr}[(X^TX)^{-1}]$	$\tilde{O}(\frac{\Vert\mathcal{X}\Vert^2\Vert\Theta\Vert^2}{\epsilon\sigma^2})$	ϵ-DP, adaptive in X, but not asymptotically efficient (Wang et al., 2015).
SSP	$O\left(\frac{\log(\frac{1}{\delta})\Vert\mathcal{X}\Vert^4\Vert\theta^*\Vert^2}{\epsilon^2}\mathrm{tr}[(X^TX)^{-2}]\right)$	$1 + \tilde{O}(\frac{d\Vert\mathcal{X}\Vert^2\Vert\theta_0\Vert^2}{n\epsilon^2\sigma^2} + \frac{d^3}{n^2\epsilon^2})$	Adaptive in $\Vert\theta^*\Vert, X$, no explicit dependence on α, but requires large n. (Sheffet, 2017, Theorem 5.1)
AdaOPS & AdaSSP	$O\left(\frac{d\log(\frac{1}{\delta})\Vert\mathcal{X}\Vert^2\Vert\theta^*\Vert^2}{\alpha n\epsilon^2}\mathrm{tr}[(X^TX)^{-1}]\right)$	$1 + \tilde{O}(\frac{d\Vert\mathcal{X}\Vert^2\Vert\theta_0\Vert^2}{n\epsilon^2\sigma^2} + \frac{d^3}{n^2\epsilon^2})$	Adaptive in $\Vert\theta^*\Vert, X, \alpha$.

Algorithm 1 ADAOPS: One-Posterior Sample estimator with adaptive regularization

```
input Data $X$, $\boldsymbol{y}$. Privacy budget: $\epsilon$, $\delta$, Bounds: $\|\mathcal{X}\|, \|\mathcal{Y}\|$.
  1. Calculate the minimum eigenvalue $\lambda_{\min}(X^TX)$.
  2. Sample $Z \sim \mathcal{N}(0,1)$ and privately release
     $\tilde{\lambda}_{\min} = \max\left\{\lambda_{\min} + \frac{\sqrt{\log(6/\delta)}}{\epsilon/4}Z - \frac{\log(6/\delta)}{\epsilon/4}, 0\right\}$.
  3. Set $\bar{\epsilon}$ as the positive solution of the quadratic equation
     $\bar{\epsilon}^2/(2\log(6/\delta)) + \bar{\epsilon} - \epsilon/4 = 0.$
  4. Set $\varrho = 0.05$, $C_1 = \big(d/2 + \sqrt{d\log(1/\varrho)} + \log(1/\varrho)\big)\log(6/\delta)/\bar{\epsilon}^2$, $C_2 = \log(6/\delta)/(\epsilon/4)$, $t_{\min} = \max\{\frac{\|\mathcal{X}\|^2(1+\log(6/\delta))}{2\epsilon} - \tilde{\lambda}_{\min}, 0\}$ and solve
     $\lambda = \operatorname*{argmin}_{t \geq t_{\min}} \frac{\|\mathcal{X}\|^4 C_1[1 + \|\mathcal{X}\|^2/(t+\tilde{\lambda}_{\min})]^{2C_2}}{t + \tilde{\lambda}_{\min}} + t. \quad (7)$
     which has a unique solution.
  5. Calculate $\hat{\theta} = (X^TX + \lambda I)^{-1}X^T\boldsymbol{y}$.
  6. Sample $Z \sim \mathcal{N}(0,1)$ and privately release
     $\Delta = \log(\|\mathcal{Y}\| + \|\mathcal{X}\|\|\hat{\theta}\|) + \frac{\log(1+\|\mathcal{X}\|^2/(\lambda+\tilde{\lambda}_{\min}))}{\epsilon/(4\sqrt{\log(6/\delta)})}Z + \frac{\log(1+\|\mathcal{X}\|^2/(\lambda+\tilde{\lambda}_{\min}))}{\epsilon/(4\log(6/\delta))}$. Set $\tilde{L} := \|\mathcal{X}\|e^{\Delta}$.
  7. Calibrate noise by choosing $\tilde{\epsilon}$ as the positive solution of the quadratic equation
     $\frac{\tilde{\epsilon}^2}{2}\left[\frac{1}{\log(6/\delta)}\frac{1+\log(6/\delta)}{\log(6/\delta)}\right] + \tilde{\epsilon} - \epsilon/2 = 0. \quad (8)$
     and then set $\gamma = \frac{(\tilde{\lambda}_{\min}+\lambda)\tilde{\epsilon}^2}{\log(6/\delta)\tilde{L}^2}$.
output $\tilde{\theta} \sim p(\theta|X,\boldsymbol{y}) \propto e^{-\frac{\gamma}{2}(\|\boldsymbol{y}-X\theta\|^2+\lambda\|\theta\|^2)}$.
```

Algorithm 2 ADASSP: Sufficient statistics perturbation with adaptive damping

```
input Data $X$, $\boldsymbol{y}$. Privacy budget: $\epsilon$, $\delta$, Bounds: $\|\mathcal{X}\|, \|\mathcal{Y}\|$.
  1. Calculate the minimum eigenvalue $\lambda_{\min}(X^TX)$.
  2. Privately release $\tilde{\lambda}_{\min} = \max\left\{\lambda_{\min} + \frac{\sqrt{\log(6/\delta)}}{\epsilon/3}\|\mathcal{X}\|^2 Z - \frac{\log(6/\delta)}{\epsilon/3}\|\mathcal{X}\|^2, 0\right\}$,
     where $Z \sim \mathcal{N}(0,1)$.
  3. Set $\lambda = \max\{0, \frac{\sqrt{d\log(6/\delta)\log(2d^2/\rho)}\|\mathcal{X}\|^2}{\epsilon/3} - \tilde{\lambda}_{\min}\}$
  4. Privately release $\widehat{X^TX} = X^TX + \frac{\sqrt{\log(6/\delta)}\|\mathcal{X}\|^2}{\epsilon/3}Z$
     for $Z \in \mathbb{R}^{d\times d}$ is a symmetric matrix and every element from the upper triangular matrix is sampled from $\mathcal{N}(0,1)$.
  5. Privately release $\widehat{X\boldsymbol{y}} = X\boldsymbol{y} + \frac{\sqrt{\log(6/\delta)}\|\mathcal{X}\|\|\mathcal{Y}\|}{\epsilon/3}Z$
     for $Z \sim \mathcal{N}(0, I_d)$.
output $\tilde{\theta} = (\widehat{X^TX} + \lambda I)^{-1}\widehat{X\boldsymbol{y}}$
```

Note that while $\|\theta^*\|^2$ cannot be privately released in general due to unbounded sensitivity, it appears in both terms and do not enter the decision process of finding the optimal λ that minimizes the bound. This convenient feature follows from our assumption that $\|\mathcal{Y}\| \lesssim \|\mathcal{X}\|\|\theta^*\|$. Dealing with the general case involving an arbitrary $\|\mathcal{Y}\|$ is an intriguing open problem.

A tricky situation for ADAOPS is that the choice of γ depends on λ through $\tilde{L}$, which is the local Lipschitz constant at the ridge regression solution θ^*_λ. But the choice of λ also depends on γ since the "variance" term above is inversely proportional to γ. Our solution is to express $\tilde{L}$ (hence γ) as a function of λ and solve the nonlinear univariate optimization problem (7).

We are now ready to state the main results.

Theorem 2. *Algorithm 1 outputs $\tilde{\theta}$ which obeys that*

(i) It satisfies (ϵ, δ)-DP.

(ii) Assume $\|\mathcal{Y}\| \lesssim \|\mathcal{X}\|\|\theta^*\|$. With probability $1-\varrho$, $F(\tilde{\theta}) - F(\theta^*) \leq$

$$O\left(\frac{\sqrt{d+\log(\frac{1}{\varrho})}\|\mathcal{X}\|^2\|\theta^*\|^2}{\epsilon/\sqrt{\log(\frac{1}{\delta})}} \wedge \frac{d[d+\log(\frac{1}{\varrho})]\|\theta^*\|^2}{\alpha n\epsilon^2/\log(\frac{1}{\delta})}\right).$$

(iii) Assume that $\mathbf{y}|X$ obeys a linear Gaussian model and X is full-rank. Then there is an event E satisfying $\mathbb{P}(E) \geq 1 - \delta/3$ and $E \perp\!\!\!\perp \mathbf{y}|X$, such that $\mathbb{E}[\tilde{\theta}|X,E] = \theta_0$ and

$$\mathrm{Cov}[\tilde{\theta}|X,E] \prec \left(1 + O\left(\frac{\tilde{C}d\log(6/\delta)}{\sigma^2\alpha n\epsilon^2}\right)\right)\sigma^2(X^TX)^{-1}$$

where constant

$$\tilde{C} := \|\mathcal{Y}\|^2 + \|\mathcal{X}\|^2(\|\theta_0\|^2 + \sigma^2\mathrm{tr}[(X^TX)^{-1}]).$$

The proof, deferred to Appendix D, makes use of a fine-grained DP-analysis through the recent per instance DP techniques (Wang, 2017) and then convert the results to DP by releasing data dependent bounds of α and the magnitude of a ridge-regression output θ^*_λ with an adaptively chosen λ. Note that $\|\theta^*_\lambda\|$ does not have a bounded global sensitivity. The method to release it differentially privately (described in Lemma 12) is part of our technical contribution.

The ADASSP algorithm is simpler and enjoys slightly stronger theoretical guarantees.

Theorem 3. *Algorithm 2 outputs $\tilde{\theta}$ which obeys that*

(i) It satisfies (ϵ, δ)-DP.

(ii) Assume $\|\mathcal{Y}\| \lesssim \|\mathcal{X}\|\|\theta^*\|$. With probability $1-\varrho$, $F(\tilde{\theta}) - F(\theta^*) \leq$

$$O\left(\frac{\sqrt{d\log(\frac{d^2}{\varrho})}\|\mathcal{X}\|^2\|\theta^*\|^2}{\epsilon/\sqrt{\log(\frac{6}{\delta})}} \wedge \frac{\|\mathcal{X}\|^4\|\theta^*\|^2\mathrm{tr}[(X^TX)^{-1}]}{\epsilon^2/[\log(\frac{6}{\delta})\log(\frac{d^2}{\varrho})]}\right)$$

(iii) Assume that $\mathbf{y}|X$ *obeys a linear Gaussian model and* X *has a sufficiently large* α*. Then there is an event* E *satisfying* $\mathbb{P}(E) \geq 1 - \delta/3$ *and* $E \perp\!\!\!\perp \mathbf{y}|X$*, such that* $\mathbb{E}[\tilde{\theta}|X, E] = \theta_0$ *and*

$$\mathbb{E}[\|\tilde{\theta} - \theta_0\|^2 | X, E]$$
$$= \sigma^2 \mathrm{tr}[(X^T X)^{-1}] + O\left(\frac{\tilde{C}\|\mathcal{X}\|^2 \mathrm{tr}[(X^T X)^{-2}]}{\epsilon^2/\log(\frac{6}{\delta})}\right),$$

with the same constant $\tilde{C}$ *in Theorem 2 (iii).*

The proof of Statement (1) is straightforward. Note that we release the eigenvalue $\lambda_{\min}(X^T X)$, $X\boldsymbol{y}$ and $X^T X$ differentially privately each with parameter $(\epsilon/3, \delta/3)$. For the first two, we use Gaussian mechanism and for $X^T X$, we use the Analyze-Gauss algorithm (Dwork et al., 2014b) with a symmetric Gaussian random matrix. The result then follows from the composition theorem of differential privacy. The proof of the second and third statements is provided in Appendix C. The main technical challenge is to prove the concentration on the spectrum and the Johnson-Lindenstrauss-like distance preserving properties for symmetric Gaussian random matrices (Lemma 6). We note that while SSP is an old algorithm the analysis of its theoretical properties is new to this paper.

Remarks. Both AdaOPS and AdaSSP match the smaller of the two lower bounds (5) and (6) for each problem instance. They are slightly different in that AdaOPS preserves the shape of the intrinsic geometry while AdaSSP's bounds are slightly stronger as they do not explicitly depend on the smallest eigenvalue.

5 EXPERIMENTS

In this section, we conduct synthetic and real data experiments to benchmark the performance of AdaOPS and AdaSSP relative to existing algorithms we discussed in Section 3. NoisySGD and Sub-Agg are excluded because they are dominated by ObjPert and an (ϵ, δ)-DP version of OPS (see Appendix F for details)[6].

Prediction accuracy in UCI data sets experiments. The first set of experiments is on training linear regression on a number of UCI regression data sets. Standard z-scoring are performed and all data points are normalized to having an Euclidean norm of 1 as a preprocessing step. The results on four of the data sets are presented in Figure 2. As we can see, SSP is unstable for small data. ObjPert suffers from a pre-defined bound $\|\Theta\|$ and does not converge to nonprivate solution even with a large ϵ. OPS performs well but still does not take advantage of the strong convexity that is intrinsic to the data set. AdaOPS and AdaSSP on the other hand are able to nicely interpolate between the trivial solution and the non-private baseline and performed as well as or better than baselines for all ϵ. More detailed quantitative results on all the 36 UCI data sets are presented in Table 3.

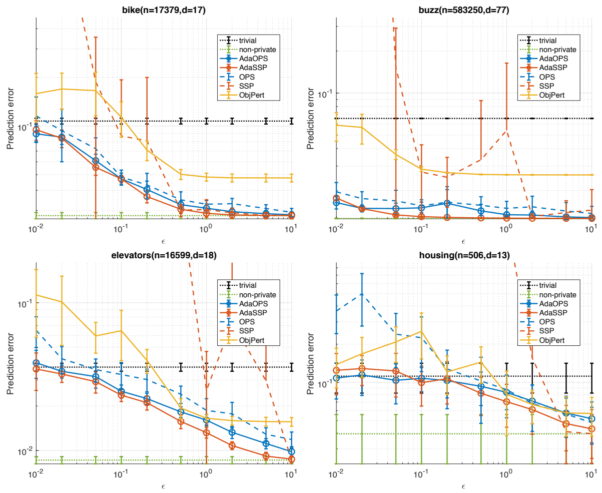
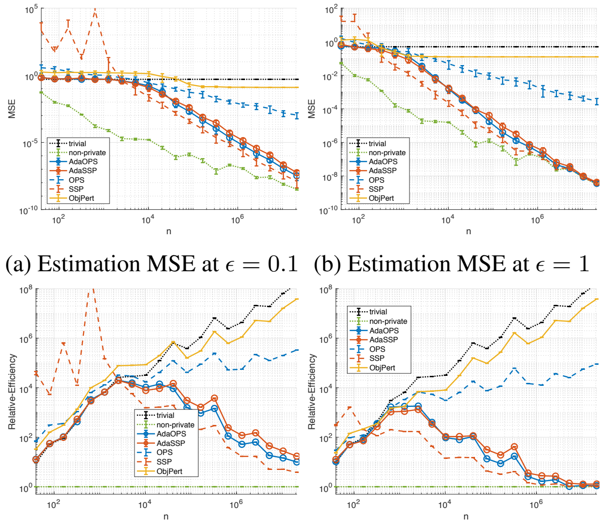


Figure 2: Example of results of differentially private linear regression algorithms on UCI data sets for a sequence of ϵ. Reported on the y-axis is the cross-validation prediction error in MSE and their confidence intervals.



(a) Estimation MSE at $\epsilon = 0.1$ (b) Estimation MSE at $\epsilon = 1$

(c) Rel. efficiency at $\epsilon = 0.1$ (d) Rel. efficiency at $\epsilon = 1$

Figure 3: Example of differentially private linear regression under linear Gaussian model with an increasing data size n. We simulate the data from $d = 10$, θ_0 drawn from a uniform distribution defined on $[0, 1]^d$. We generate $X \in \mathbb{R}^{n \times d}$ as a Gaussian random matrix and then generate $y \sim \mathcal{N}(X\theta_0, I_d)$. We used $\epsilon = 1$ and $\epsilon = 0.1$, both with $\delta = 1/n^2$. The results clearly illustrate the asymptotic efficiency of the proposed approaches.

[6] The code to reproduce all experimental results are available at https://github.com/yuxiangw/optimal_dp_linear_regression.

Table 3: Summary of UCI data experiments at $\epsilon = 0.1, \delta = \min\{1e-6, 1/n^2\}$. The boldface denotes the DP algorithm where the standard deviation is smaller than the error (a positive quantity), and the 95% confidence interval covers the observed best performance among benchmarked DP algorithms.

	Trivial	non-private	ObjPert	OPS	SSP	AdaOPS	AdaSSP
3droad	0.0275±0.00014	0.0265±0.00012	0.0267±0.00013	0.027±0.00026	**0.0265±0.00019**	**0.0265±0.00019**	**0.0265±0.00019**
airfoil	0.103±0.0069	0.0533±0.0074	0.356±0.064	**0.138±0.086**	0.232±0.28	**0.0914±0.015**	**0.0878±0.014**
autompg	0.113±0.011	0.0221±0.0032	**0.143±0.096**	0.242±0.11	5.44±6.1	**0.098±0.03**	**0.115±0.047**
autos	0.13±0.042	0.0274±0.011	**0.17±0.13**	0.308±0.13	1.7e+03±2.5e+03	**0.136±0.066**	**0.132±0.064**
bike	0.107±0.0028	0.0279±0.00078	0.113±0.018	**0.0484±0.005**	0.0869±0.067	**0.0471±0.004**	**0.0471±0.0026**
breastcancer	0.194±0.027	0.139±0.025	**0.212±0.078**	**0.269±0.13**	9.54e+03±1.9e+04	**0.204±0.037**	**0.196±0.051**
buzz	0.0658±0.00015	0.0127±4.6e-05	0.0285±0.00071	0.0156±0.001	0.0272±0.0097	0.0151±0.00095	**0.013±9.7e-05**
challenger	0.141±0.084	0.138±0.088	0.323±0.28	0.338±0.13	3.07±3.9	**0.159±0.13**	**0.146±0.093**
concrete	0.127±0.0043	0.0445±0.0033	0.237±0.076	0.181±0.042	1.94±1.8	**0.12±0.011**	**0.119±0.016**
concreteslump	0.149±0.039	0.0245±0.0071	0.349±0.094	0.549±0.24	3.14±2.5	**0.151±0.064**	**0.165±0.065**
elevators	0.0367±0.0014	0.00861±0.00031	0.0647±0.015	0.0327±0.0042	0.645±0.98	**0.0252±0.0026**	**0.0237±0.0022**
energy	0.235±0.012	0.0232±0.0023	0.332±0.09	**0.161±0.083**	1.7e+03±3.4e+03	**0.167±0.034**	**0.15±0.032**
fertility	0.0977±0.024	0.0863±0.024	0.203±0.04	0.639±0.16	439±8.6e+02	**0.108±0.048**	**0.115±0.032**
forest	0.0564±0.0081	0.0571±0.0086	0.12±0.022	0.177±0.036	41.9±77	**0.0622±0.017**	**0.0675±0.013**
gas	0.112±0.0062	0.0214±0.0028	0.109±0.015	**0.0546±0.012**	0.923±0.63	0.0801±0.0078	0.0875±0.0073
houseelectric	0.122±0.00017	0.0136±1.4e-05	0.0409±0.00027	0.0144±0.00017	**0.0136±2.2e-05**	**0.0136±2.2e-05**	**0.0136±2.2e-05**
housing	0.112±0.019	0.0394±0.01	0.253±0.063	0.225±0.065	2.24±2.3	**0.108±0.023**	**0.0997±0.035**
keggdirected	0.117±0.00095	0.0188±0.0011	0.0637±0.0042	0.0266±0.0019	0.23±0.33	0.0227±0.0015	**0.0212±0.0011**
keggundirected	0.0694±0.00074	0.00475±8.9e-05	0.0365±0.0028	0.0166±0.0033	0.353±0.4	0.0107±0.0012	**0.00912±0.00046**
kin40k	0.0634±0.0012	0.0632±0.0013	0.0871±0.0092	0.0717±0.0026	**0.0633±0.002**	**0.0639±0.0021**	**0.064±0.0021**
machine	0.121±0.013	0.0395±0.0051	0.282±0.14	0.347±0.14	2.27e+03±4.5e+03	**0.105±0.025**	**0.141±0.068**
parkinsons	0.17±0.0026	0.128±0.0024	0.211±0.014	**0.157±0.011**	132±2.6e+02	**0.159±0.0065**	**0.156±0.0064**
pendulum	0.0226±0.0061	0.0181±0.0049	0.118±0.027	0.122±0.041	24.8±45	**0.0276±0.011**	0.0346±0.0069
pol	0.345±0.0028	0.135±0.0023	0.302±0.032	**0.196±0.02**	281±5.3e+02	0.214±0.0056	0.214±0.0061
protein	0.167±0.0011	0.119±0.0014	0.158±0.01	0.137±0.0044	**0.149±0.06**	0.129±0.0015	**0.125±0.0026**
pumadyn32nm	0.0935±0.0039	0.0941±0.0039	0.124±0.0046	0.111±0.005	8.92e+03±1.8e+04	**0.0968±0.0065**	**0.0966±0.0063**
servo	0.184±0.039	0.0752±0.022	0.366±0.077	0.574±0.26	2.03±1.5	**0.195±0.065**	**0.198±0.081**
skillcraft	0.0439±0.0021	0.0203±0.0017	0.0817±0.013	0.0519±0.0099	4.72±4.3	**0.037±0.008**	**0.039±0.0056**
slice	0.196±0.0021	0.0283±0.00051	0.174±0.0053	**0.0924±0.0035**	11.2±9.4	0.0992±0.0021	0.132±0.0015
sml	0.211±0.0089	0.0143±0.00066	0.23±0.03	**0.0955±0.029**	59.9±80	0.134±0.0075	0.147±0.013
solar	0.0118±0.0042	0.0106±0.0038	0.0994±0.023	0.0667±0.017	5.95±9.6	**0.0165±0.0062**	**0.0204±0.0073**
song	0.0917±0.0003	0.0636±0.00033	0.0838±0.0014	0.072±0.00035	**0.0644±0.0005**	0.0685±0.00045	0.0697±0.00029
stock	0.0583±0.0095	0.013±0.0023	0.122±0.026	0.157±0.055	46.8±66	**0.0582±0.023**	**0.0651±0.024**
tamielectric	0.334±0.002	0.334±0.0021	0.341±0.0021	0.343±0.0065	**0.335±0.0033**	**0.337±0.0047**	**0.335±0.0033**
wine	0.0566±0.0028	0.0202±0.00099	0.153±0.028	0.0911±0.016	11.7±17	**0.058±0.011**	**0.0599±0.01**
yacht	0.105±0.017	0.0176±0.0055	0.273±0.076	0.371±0.14	4.92±6.8	**0.0967±0.035**	**0.109±0.03**

Parameter estimation under linear Gaussian model. To illustrate the performance of the algorithms under standard statistical assumptions, we also benchmarked the algorithms on synthetic data generated by a linear Gaussian model. The results, shown in Figure 3 illustrates that as n gets large, AdaOPS and AdaSSP with $\epsilon = 0.1$ and $\epsilon = 1$ converge to the maximum likelihood estimator at a rate faster than the optimal statistical rate that MLE estimates θ^*, therefore at least for large n, differential privacy comes for free. Note that there is a gap in SSP and AdaSSP for large n, this can be thought of as a cost of adaptivity as AdaSSP needs to spend some portion of its privacy budget to release $\lambda_{\min}$, which SSP does not, this can be fixed by using more careful splitting of the privacy budget.

6 CONCLUSION

In this paper, we presented a detailed case-study of the problem of differentially private linear regression. We clarified the relationships between various quantities of the problems as they appear in the private and non-private information-theoretic lower bounds. We also surveyed the existing algorithms and highlighted that the main drawback using these algorithms relative to their non-private counterpart is that they cannot adapt to data-dependent quantities. This is particularly true for linear regression where the ordinary least square algorithm is able to work optimally for a large class of different settings.

We proposed AdaOPS and AdaSSP to address the issue and showed that they both work in unbounded domain. Moreover, they smoothly interpolate the two regimes studied in Bassily et al. (2014) and behave nearly optimally for every instance. We tested the two algorithms on 36 real-life data sets from the UCI machine learning repository and we see significant improvement over popular algorithms for almost all configurations of ϵ.

Acknowledgements

The author thanks the anonymous reviewers for helpful feedbacks and Zichao Yang for sharing the 36 UCI regression data sets as was used in (Yang et al., 2015).

References

Agresti, A., & Finlay, B. (1997). Statistical methods for the social sciences.

Armitage, P., Berry, G., & Matthews, J. N. S. (2008). *Statistical methods in medical research*. John Wiley & Sons.

Bassily, R., Smith, A., & Thakurta, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Foundations of Computer Science (FOCS-14)*, (pp. 464–473). IEEE.

Birgé, L., & Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, *3*(3), 203–268.

Chaudhuri, K., Monteleoni, C., & Sarwate, A. D. (2011). Differentially private empirical risk minimization. *The Journal of Machine Learning Research*, *12*, 1069–1109.

Dimitrakakis, C., Nelson, B., Mitrokotsa, A., & Rubinstein, B. I. (2014). Robust and private Bayesian inference. In *Algorithmic Learning Theory*, (pp. 291–305). Springer.

Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE transactions on information theory*, *41*(3), 613–627.

Draper, N. R., & Smith, H. (2014). *Applied regression analysis*, vol. 326. John Wiley & Sons.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., & Naor, M. (2006a). Our data, ourselves: Privacy via distributed noise generation. In *International Conference on the Theory and Applications of Cryptographic Techniques*, (pp. 486–503). Springer.

Dwork, C., & Lei, J. (2009). Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, (pp. 371–380). ACM.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006b). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography*, (pp. 265–284). Springer.

Dwork, C., Roth, A., et al. (2014a). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, *9*(3–4), 211–407.

Dwork, C., & Smith, A. (2010). Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, *1*(2), 2.

Dwork, C., Talwar, K., Thakurta, A., & Zhang, L. (2014b). Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *ACM symposium on Theory of computing (STOC-14)*, (pp. 11–20). ACM.

Foulds, J., Geumlek, J., Welling, M., & Chaudhuri, K. (2016). On the theory and practice of privacy-preserving Bayesian data analysis. In *Conference on Uncertainty in Artificial Intelligence (UAI-16)*, (pp. 192–201). AUAI Press.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*, vol. 1. Springer series in statistics Springer, Berlin.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Kifer, D., Smith, A., & Thakurta, A. (2012). Private convex empirical risk minimization and high-dimensional regression. *Journal of Machine Learning Research*, *1*, 41.

Koren, T., & Levy, K. (2015). Fast rates for exp-concave empirical risk minimization. In *Advances in Neural Information Processing Systems*, (pp. 1477–1485).

Laurent, B., & Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, (pp. 1302–1338).

Lei, J., Charest, A.-S., Slavkovic, A., Smith, A., & Fienberg, S. (2017). Differentially private model selection with penalized and constrained likelihood. *Journal of the Royal Statistical Society*.

Minami, K., Arai, H., Sato, I., & Nakagawa, H. (2016). Differential privacy without sensitivity. In *Advances in Neural Information Processing Systems*, (pp. 956–964).

Mir, D. J. (2013). *Differential privacy: an exploration of the privacy-utility landscape*. Ph.D. thesis, Rutgers University.

Shamir, O. (2015). The sample complexity of learning linear predictors with the squared loss. *Journal of Machine Learning Research*, *16*, 3475–3486.

Sheffet, O. (2017). Differentially private ordinary least squares. In *International Conference on Machine Learning (ICML-17)*, (pp. 3105–3114).

Smith, A. (2008). Efficient, differentially private point estimators. *arXiv preprint arXiv:0809.4794*.

Stewart, G. W. (1998). Perturbation theory for the singular value decomposition. Tech. rep.

Talwar, K., Thakurta, A., & Zhang, L. (2014). Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry. *arXiv preprint arXiv:1411.5417*.

Talwar, K., Thakurta, A. G., & Zhang, L. (2015). Nearly optimal private lasso. In *Advances in Neural Information Processing Systems*, (pp. 3025–3033).

Vu, D., & Slavkovic, A. (2009). Differential privacy for clinical trial data: Preliminary evaluations. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, (pp. 138–143). IEEE.

Wang, Y.-X. (2017). Per-instance differential privacy and the adaptivity of posterior sampling in linear and ridge regression. *arXiv preprint arXiv:1707.07708*.

Wang, Y.-X., Fienberg, S., & Smola, A. (2015). Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning (ICML-15)*, (pp. 2493–2502).

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

Yang, Z., Wilson, A., Smola, A., & Song, L. (2015). A la carte–learning fast kernels. In *Artificial Intelligence and Statistics (AISTATS-15)*, (pp. 1098–1106).