(Near) Dimension Independent Risk Bounds for Differentially Private Learning

Prateek Jain PRAJAIN@MICROSOFT.COM
Microsoft Research

Abhradeep Thakurta B-ABHRAG@MICROSOFT.COM
Stanford University and Microsoft Research

Abstract

In this paper, we study the problem of differentially private risk minimization where the goal is to provide differentially private algorithms that have small excess risk. In particular we address the following open problem: *Is it possible to design computationally efficient differentially private risk minimizers with excess risk bounds that do not explicitly depend on dimensionality (p) and do not require structural assumptions like restricted strong convexity?*

In this paper, we answer the question in the affirmative for a variant of the well-known *output* and *objective* perturbation algorithms (Chaudhuri et al., 2011). In particular, we show that under certain assumptions, variants of both output and objective perturbation algorithms have no *explicit* dependence on p; the excess risk depends only on the L_2-norm of the true risk minimizer and that of training points.

Next, we present a novel privacy preserving algorithm for risk minimization over simplex in the generalized linear model, where the loss function is a doubly differentiable convex function. Assuming that the training points have bounded L_∞-norm, our algorithm provides risk bound that has only *logarithmic* dependence on p. We also apply our technique to the online learning setting and obtain a regret bound with similar logarithmic dependence on p. In contrast, the existing differentially private online learning methods incur $O(\sqrt{p})$ dependence.

1. Introduction

Recently, there have been growing concerns regarding potential privacy violation of individual users'/customers' data by modern systems that employ learning and statistical analysis methods. Motivated by such concerns, several recent works have proposed and analyzed privacy preserving learning algorithms (Chaudhuri et al., 2011; Pathak et al., 2010; Kifer et al., 2012; Jain et al., 2012; Duchi et al., 2012; Smith & Thakurta, 2013). All these works use differential privacy (Dwork et al., 2006b) as the notion to define privacy of each individual training data point. Furthermore, they show that not only their methods are differentially private, but they also have bounded excess risk or bounded regret that improves with larger number of training instances.

Most of these existing methods use the standard technique of adding noise to either the learned model or some intermediate construct of the algorithm (Chaudhuri et al., 2011; Kifer et al., 2012; Jain et al., 2012). Subsequently, they provide excess risk/generalization error bounds that scale as a *polynomial* of the dimensionality (p) of the input data.

One of the achievement of machine learning has been that for several learning problems that can be modeled using *generalized linear model*, the excess generalization risk can be shown to be *independent* of the dimensionality of input data points (p), provided that the input data points and the output model are constrained to have bounded L_2-norm. Similarly, for several other classes of problems, the bounds are known to grow only *logarithmically* with p (Shalev-Shwartz et al., 2009; Negahban et al., 2009).

Hence, a long-standing open problem in the domain of differential privacy has been: *Can computationally efficient differentially private learning algorithms be designed which have excess risk/generalization error that is either* independent *or only* logarithmically *dependent on the dimensionality of the problem?* Recently, Kifer et al. (2012); Smith & Thakurta (2013) obtained risk bounds for the sparse regression problem that scale logarithmically with p. However, their method needs additional restrictive assumptions like restricted strong convexity (RSC).

In this paper, we provide the *first* dimension independent risk bounds for generalized linear model based learning methods. In particular, we show that while the "distance" (such as Euclidean distance) between the differentially private learned model and the optimal non-private learned

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

model can depend polynomially on p, the excess risk can still be independent of the dimensionality of the input data, as long as the feature vectors and the underlying parameter vector are bounded in the L_2-norm. Furthermore, we show that if the feature vectors are bounded in the L_∞-norm and the underlying parameter vector is bounded in the L_1-norm, we can get risk bounds which only depend logarithmically on the dimensions.

We propose to use Gaussian noise based perturbation with the *output* and *objective perturbation* algorithms by Chaudhuri et al. (2011). While the privacy of such perturbation was analyzed by Chaudhuri et al. (2011); Kifer et al. (2012), their excess risk analysis for empirical risk minimization (ERM) in the generalized linear models (GLM) were loose and led to polynomial dependence on p. We show that their analysis can be tightened to remove explicit dependence on p and have dependence only on the L_2 norm of the input data points and the output parameter vector.

Our results hold for all L_2 regularized unconstrained ERMs with 1-Lipschitz smooth convex function. We would like to stress that this class of ERM is a popular class and includes important problems such as logistic regression. Furthermore, for analysis of output perturbation based algorithms we don't need smoothness assumption as well and hence can provide excess risk bounds for private-SVM as well.

We then study a class of problems where the parameter vector is bounded to be on a scaled simplex. Recently, such problems have gained a lot of importance as they provide excess risk bounds that scale only logarithmically in dimensions and linearly with the L_1-norm of parameter vector, hence are well-suited for high-dimensional learning.

For these problems, we propose a novel privacy preserving algorithm that draws multiple samples from a distribution defined by the non-private optimal parameter vector and outputs their average. We show that the excess risk for this algorithm scale as $O(\log p/n^{1/3})$, where n is the number of training examples. In comparison, existing approaches (Chaudhuri et al., 2011; Kifer et al., 2012) incurs a $poly(p)$ dependence on the error. We would like to stress that our algorithm fundamentally deviates from existing approaches for private ERM as the existing techniques require $O(p)$ randomized operations to output the parameter vector while we perform only sub-linear (in n) number of randomized operations to give the output parameter vector. Hence, our method requires significantly lesser randomness for $p \gg n$.

Furthermore, most of the existing differential private algorithms either add explicit perturbation (Chaudhuri et al., 2011; Kifer et al., 2012) or uses a well-known exponential mechanism (McSherry & Talwar, 2007). In contrast, our algorithm uses a novel sampling approach which might in itself be of interest for designing novel differentially private algorithms. Our algorithm also ensures that the output is in fact a sparse vector, hence it not only provides privacy but also enables efficient computation.

As a direct application of our private algorithm over the simplex, we provide a privacy preserving variant of the *Follow the Regularized Leader* (FTRL) algorithm commonly used in online learning (Hazan et al., 2007; Shalev-Shwartz, 2011). We show that if the cost functions (in the online learning setting) are linear and the optimization is performed over the simplex, then our proposed algorithm achieves the optimal regret bound of $O(\sqrt{T \log p})$. Here, T refers to the time horizon of the online learning algorithm and p-refers to the dimensionality of the problem. A similar result was obtained by Dwork et al. (2010b) that guarantees $\sqrt{T \log p}$ regret, however their analysis holds only for a *weaker model of privacy* where the adversary cannot distinguish between the *presence or absence of one coordinate in the linear cost function*.

Finally, we provide empirical evaluation of our proposed methods and compare them against the objective/output perturbation methods of Chaudhuri et al. (2011) over benchmark data sets. We show that the methods of Chaudhuri et al. (2011) indeed incur test error that grows with p, while our method is able to obtain accurate predictions even for high-dimensional data sets. Similarly, we also evaluate our proposed sampling based method for privacy preserving learning over simplex by simulations over a benchmark data set.

Contributions:

1. We show that by sampling the perturbation from a Gaussian distribution, instead of Gamma distribution as proposed by Chaudhuri et al. (2011), we can obtain dimension independent excess risk for the well-known output and objective perturbation algorithms (Chaudhuri et al., 2011) when applied to the maximum-margin based problems.

2. We provide a sampling based differentially private algorithm for solving a large class of ERMs over scaled simplex and show that the obtained risk bound scales *logarithmically* in p. However, our excess generalization error rate has a worse dependence of $1/n^{1/3}$ on the size of the data set (n), as compared to the optimal rate of $1/\sqrt{n}$.

3. We provide a differentially private version of the Follow The Regularized Leader algorithm for online learning, whose regret scales as $O(\sqrt{T \log p})$ when optimizing over the simplex. Our regret bound matches the non-private regret bound (under similar setting) up to factors depending only on differential privacy parameters ϵ and $\log(1/\delta)$.

4. We provide empirical evaluation of our methods on benchmark data sets. Our evaluation clearly shows that the proposed techniques not only provides significantly tighter error bounds but also provide significantly more accurate predictions on benchmark data sets.

2. Related Works

The problem of differentially private ERM has been studied extensively in the literature. Starting with (Chaudhuri

& Monteleoni, 2008; Chaudhuri et al., 2011; Rubinstein et al., 2009), there has been extensive work both in the *low-dimensional setting* as well as the *high-dimensional setting*. In the low-dimensional setting, where the dimensionality of the problem (p) is smaller than the size of the training data set (n), most of the existing methods provide an excess error bound that has a polynomial dependence on p (Chaudhuri & Monteleoni, 2008; Chaudhuri et al., 2011; Rubinstein et al., 2009; Kifer et al., 2012; Jain et al., 2012). In the *high-dimensional setting*, where the dimensionality of the problem exceeds the size of the data set, the existing methods provide logarithmic dependence on p, but require strong statistical assumptions like Restricted Strong Convexity (Kifer et al., 2012; Smith & Thakurta, 2013). An exception to the above is the work by Jain & Thakurta (2013) that provides a differentially private method for learning with kernels, however their methods assume a stronger model where a few samples from the unlabeled test data set are also available.

In contrast, we show that for generalized linear models with bounded feature space (either in the L_2-norm or L_∞-norm), we can use differentially private regularized ERM that also guarantees excess risk bound which is independent of p or depends logarithmically on the dimensions (p).

Chaudhuri & Hsu (2011) provided a differentially private classifier where the excess generalization error is dependent only the doubling dimension of the hypothesis space. They also provided a matching lower bound. However, their method is defined only for $0-1$ loss function and in general can take exponential time in the number of training points. We also note that the doubling dimension of regularized linear learning models can be shown to be either independent or logarithmically dependent on p (Zhang, 2002). Hence our results do not contradict the lower-bound result of Chaudhuri & Hsu (2011).

3. Background and Problem Formulation

Risk Minimization and Excess Generalization Error: Given a data domain $\mathcal{X}$, an unknown but fixed distribution $Dist$ over $\mathcal{X}$, a fixed convex set $\mathcal{C} \subseteq \mathbb{R}^p$, and a risk (loss) function $\ell : \mathcal{C} \times \mathcal{X} \to \mathbb{R}$, the objective is solve the following *stochastic minimization problem*: $\arg\min_{\boldsymbol{\theta}\in\mathcal{C}} \mathbb{E}_{d\sim Dist}[\ell(\boldsymbol{\theta}; d)]$. For a given vector $\boldsymbol{\theta} \in \mathcal{C}$, the *excess generalization error (or risk)* is defined as an upper bound on:

$$\mathop{\mathbb{E}}_{d\sim Dist}[\ell(\boldsymbol{\theta}; d)] - \min_{\boldsymbol{\theta}\in\mathcal{C}} \mathop{\mathbb{E}}_{d\sim Dist}[\ell(\boldsymbol{\theta}; d)].$$

To minimize the excess risk, we use the standard regularized Empirical Risk Minimization (ERM) method:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta}; d_i) + \frac{r(\boldsymbol{\theta})}{n}, \tag{1}$$

where, the *training data* ($\mathcal{D}$) is drawn i.i.d. from the distribution $Dist$. Also, $r : \mathcal{C} \to \mathbb{R}$ is a *twice differentiable* convex regularizer.

Now the goal of this work is to design differentially private ERM with small *excess risk* bound. We focus on the *generalized linear models* (GLMs), where each data point d is of the form $(\boldsymbol{x}, y)$ with $\boldsymbol{x} \in \mathbb{R}^p$ and $y \in \mathbb{R}$, and the loss function $\ell(\boldsymbol{\theta}; d)$ is given by: $\ell(\boldsymbol{\theta}, d) = \ell(\langle \boldsymbol{x}, \boldsymbol{\theta}\rangle; y)$. Logistic regression, linear regression and support vector machines are some of the classic examples of GLM.

Differential Privacy: Differential privacy (Dwork et al., 2006b;a) ensures that the amount of information an adversary can obtain about an individual from the output of an algorithm $\mathcal{A}$ running on the data set $\mathcal{D}$ is roughly the same irrespective of that individual's presence or absence in the data set $\mathcal{D}$. Formally,

Definition 1 (Differential privacy (Dwork et al., 2006b;a))**.** *A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private if for any two data sets $\mathcal{D}$ and $\mathcal{D}'$ of size n drawn from the domain $\mathcal{X}^n$ with $d_H(\mathcal{D}, \mathcal{D}') = 1$ (d_H being the hamming distance), and for all (Borel) sets $\mathcal{O} \subseteq Range(\mathcal{A})$ the following holds:* $\Pr[\mathcal{A}(\mathcal{D}) \in \mathcal{O}] \leq e^\epsilon \Pr[\mathcal{A}(\mathcal{D}') \in \mathcal{O}] + \delta$.

Choice of ϵ, δ privacy parameters: Smaller values of ϵ and δ imply stronger privacy guarantees. Typically ϵ is set to be a small constant (say 0.1) and δ should be $o(1/n^2)$, where n is the number of records in the data set. See (Kasiviswanathan & Smith, 2008) (Lemma 3.3) for a rigorous justification of the above choices of these parameters. Kasiviswanathan & Smith (2008) also show that the semantic notion privacy (see Definition 2.3 in (Kasiviswanathan & Smith, 2008)) is invariant to the size of the hypothesis space and hence, to provide a fixed level of privacy to a data point, δ is not required to depend on p.

A common approach for designing a differentially private algorithm is via the *global sensitivity framework* defined below. In a lot of the algorithms discussed in this paper, this forms the basic building block. Let $\mathcal{X}^n$ be a domain of data sets (with n data points) and let $f : \mathcal{X}^n \to \mathbb{R}^p$ be a function to be evaluated on a data set $\mathcal{D} \in \mathcal{X}^n$. *Global sensitivity* of the function f is defined as in (2). Here the operator d_H refers to the hamming distance and $\|\cdot\|_q$ refers to the L_q-norm for a specific q.

$$GS(f) = \max_{\mathcal{D},\mathcal{D}'\in\mathcal{X}^n, d_H(\mathcal{D},\mathcal{D}')=1} \|f(\mathcal{D}) - f(\mathcal{D}')\|_q \tag{2}$$

Let $\boldsymbol{b} \in \mathbb{R}^p$ be a random variable sampled from the distribution with density proportional to $e^{-\frac{\epsilon\|\boldsymbol{b}\|_q}{GS(f)}}$. (Dwork et al., 2006b) showed that for a given data set $\mathcal{D}$, an algorithm that outputs $f(\mathcal{D}) + \boldsymbol{b}$ is *ϵ-differentially private*.

Assumptions and notation: Throughout this paper we will assume that the loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is in the generalized linear model, L-Lipschitz continuous in its first parameter, and twice-continuously differentiable.

The feature vectors are bounded in the L_q-norm with the bound being denoted with R_q. We will set $q = 2$ or $q = \infty$ based on the context of the problem. For a fixed distribution $Dist$ over the data domain $\mathcal{X}$, we denote $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\mathcal{C}} \mathbb{E}_{(\boldsymbol{x},y)\sim Dist}[\ell(\langle \boldsymbol{x}, \boldsymbol{\theta}\rangle; y)]$. We denote the diameter of the convex set $\mathcal{C}$ in L_q-norm as $\|\mathcal{C}\|_q$ (when the diameter is finite). For the domain of data points $\mathcal{X}$, R_q denotes an L_q bound on the norm of any $\boldsymbol{x} \in \mathbb{R}^p$ in $\mathcal{X}$.

4. Private Risk Minimization with No Explicit Dependence on Dimensions

In this section, we present two privacy preserving ERM algorithms with dimension independent excess generalization error bounds for L_2-norm bounded data points and L_2-norm bounded parameter space. We also assume that the regularizer $r(\boldsymbol{\theta})$ in (1) is $\frac{\lambda\|\boldsymbol{\theta}\|_2^2}{2}$ (with λ being the regularization parameter) and the convex set $\mathcal{C}$ equals $\mathbb{R}^p$. One way to interpret the assumption on the convex set is in the *improper learning* setting. Although the true risk minimizer lies in a bounded convex set, the algorithm is allowed to produce a hypothesis from $\mathbb{R}^p$. Settings as above are common in several machine learning formulations, especially ones that try to find maximum margin learner such as SVM, L_2 regularized logistic regression etc.

Chaudhuri et al. (2011) proposed two popular algorithms for the problem of privacy preserving ERM, namely *output perturbation* and *objective perturbation*. Roughly, the output perturbation method perturbs the true minimizer of the ERM (1), $\hat{\boldsymbol{\theta}}$, to preserve privacy. While the objective perturbation method perturbs the objective function $J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \ell(\langle \boldsymbol{\theta}, x_i\rangle; y_i) + \frac{\lambda\|\boldsymbol{\theta}\|_2^2}{2n}$ for privacy.

In the following, we provide tighter utility analyses for both output and objective perturbation for GLMs.

4.1. Output Perturbation

The output perturbation method (Chaudhuri et al., 2011) first computes the minimizer $\widehat{\boldsymbol{\theta}}$ of (1) and then adds noise scaled according to the global sensitivity of $\widehat{\boldsymbol{\theta}}$. A formal description of the algorithm is given in (3).

$$\textbf{Output Perturbation: } \boldsymbol{\theta}_{priv}(\mathcal{D}) = \widehat{\boldsymbol{\theta}}(\mathcal{D}) + \boldsymbol{b} \quad (3)$$

Here $\widehat{\boldsymbol{\theta}}(\mathcal{D})$ maps a data set $\mathcal{D}$ to the corresponding minimizer $\widehat{\boldsymbol{\theta}}$ of (1), and $\boldsymbol{b} \in \mathbb{R}^p$ is a random vector whose L_2-norm (v) is distributed according to the Gamma distribution with kernel $e^{-\frac{v\epsilon\lambda}{4LR_2}}$.

Chaudhuri et al. (2011) showed that if the bound on the double derivative of ℓ is c_2 (w.r.t. its first parameter), the excess generalization error scales as $O\left(\frac{{Lc_2}^{1/3} p \log p (R_2\|\boldsymbol{\theta}^*\|_2)^{4/3}}{\epsilon\sqrt{n}}\right)$. Kifer et al. (2012); Thakurta (2013) showed that instead of adding Gamma noise, if Gaussian noise $\mathcal{N}(0, \frac{16(LR_2)^2(\log(1/\delta)+\epsilon)}{\lambda^2\epsilon^2}\mathbb{I}_p)$ is added, then the generalization error improves to $O\left(\frac{{Lc_2}^{1/3}\sqrt{p(\log(1/\delta)+\epsilon)}(R_2\|\boldsymbol{\theta}^*\|_2)^{4/3}}{\epsilon\sqrt{n}}\right)$. Note the $\sqrt{p}$ improved dependence on the dimensionality compared to the bound by Chaudhuri et al. (2011). However, the privacy guarantee is weaker, *i.e.,* the algorithm now satisfies (ϵ, δ)-differential privacy compared to ϵ-differential privacy guarantee provided by Chaudhuri et al. (2011).

In our work we improve the earlier analysis and show that with the same Gaussian noise, one can get generalization error guarantees that are independent of any *explicit* dependence on p. Our result also has improved dependence on parameters L, R_2 and $\|\boldsymbol{\theta}^*\|_2$. We would like to note that our results hold only for the generalized linear model (GLM) and when the regularization function is the squared L_2 norm.

Theorem 1. *Let $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$ be i.i.d. samples drawn from a fixed distribution $Dist$ over the domain $\mathcal{X}$. Also, let $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathbb{E}_{(\boldsymbol{x},y)\sim Dist}[\ell(\langle \boldsymbol{\theta}, x\rangle; y)]$. If $\lambda = \frac{LR_2\sqrt{n}}{\|\boldsymbol{\theta}^*\|_2}$, then with probability at least $2/3$ over the randomness of the training data set $\mathcal{D}$ and the randomness of the noise vector $\boldsymbol{b}$, the following is true:*

$$\mathbb{E}_{(\boldsymbol{x},y)\sim Dist}[\ell(\langle \boldsymbol{\theta}_{priv}, x\rangle; y) - \ell(\langle \boldsymbol{\theta}^*, x\rangle; y)] = O\left(\frac{LR_2\|\boldsymbol{\theta}^*\|_2\sqrt{\log\left(\frac{1}{\delta}\right)+\epsilon}}{\epsilon\sqrt{n}}\right),$$

where $\boldsymbol{\theta}_{priv}$ is the output of the output perturbation method (3) with $\boldsymbol{b} \sim \mathcal{N}(0, \frac{16(LR_2)^2(\log(1/\delta)+\epsilon)}{\lambda^2\epsilon^2}\mathbb{I}_p)$.

Proof Idea: The main idea in our proof is that since the learning model is a GLM, the prediction for a feature vector $\boldsymbol{x} \in \mathbb{R}^p$ with a parameter vector $\boldsymbol{\theta}$ depends only on $\langle \boldsymbol{x}, \boldsymbol{\theta}\rangle$. Now for the two parameter vectors $\widehat{\boldsymbol{\theta}}$ (from (1)) and $\boldsymbol{\theta}_{priv}$ (the output of the output perturbation algorithm), and any data point $(\boldsymbol{x}, y)$, the difference in the loss is bounded by the following.

$$|\ell(\langle \boldsymbol{\theta}_{priv}, x\rangle; y) - \ell(\langle \widehat{\boldsymbol{\theta}}, x\rangle; y)| \leq L|\langle \boldsymbol{b}, x\rangle|$$

Since we are interested in bounding the excess generalization error, we only need to bound the right hand side in expectation over $\boldsymbol{b}$ (and then use Markov's inequality). Recall that our noise vector $\boldsymbol{b}$ is a symmetric Gaussian random vector. So, we have $\mathbb{E}_{\boldsymbol{b}}[|\langle \boldsymbol{b}, x\rangle|] \leq \sigma\|x\|_2$, where σ is the standard deviation of $\boldsymbol{b}$. Notice that a naive Cauchy-Schwarz argument would result in a bound of $\sqrt{p}\sigma\|x\|_2$.

The fact that we can bound the expectation with a quantity that does not have any explicit dependence on the dimensionality allows us to get our desired result in Theorem 1. See Appendix A for a detailed proof.

4.2. Objective Perturbation

In this section we discuss the objective perturbation algorithm that perturbs the objective function in (1) by a random linear term to guarantee differential privacy. That is,

$$\textbf{Objective Perturbation: } \boldsymbol{\theta}_{priv} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \ell(\langle\boldsymbol{\theta}, x_i\rangle; y_i) + \frac{\lambda\|\boldsymbol{\theta}\|_2^2}{2n} + \frac{\langle \boldsymbol{b}, \boldsymbol{\theta}\rangle}{n}, \quad (4)$$

where λ is a parameter to the algorithm and $\boldsymbol{b}$ is a zero-mean perturbation. This algorithm was first proposed by Chaudhuri et al. (2011) and was subsequently improved by Kifer et al. (2012). Chaudhuri et al. (2011) showed that the algorithm is ϵ-differentially private when the noise vector $\boldsymbol{b}$ is drawn from the Gamma distribution with kernel $e^{-\frac{\epsilon\|\boldsymbol{b}\|_2}{2LR_2}}$ and $\lambda = \frac{2c_2R_2^2}{\epsilon}$, where $|\ell''(\langle\boldsymbol{\theta}, \boldsymbol{x}\rangle; y)| \leq c_2,\ \forall(\boldsymbol{x}, y) \in \mathcal{X},\ \forall\boldsymbol{\theta} \in \mathbb{R}^p$. (The double derivative is w.r.t. the first parameter of ℓ.) While, when the noise vector $\boldsymbol{b}$ is drawn from $\mathcal{N}\left(0, \frac{4(LR_2)^2(\log\frac{1}{\delta}+\epsilon)}{\epsilon^2}\mathbb{I}_p\right)$, Kifer et al. (2012) showed that the algorithm is (ϵ, δ)-differentially private.

In terms of excess risk bounds, Chaudhuri et al. (2011) showed that under suitable choice of parameter λ, the generalization error of the private algorithm with Gamma noise scales as:

$$O\left(\frac{(LR_2)\|\boldsymbol{\theta}^*\|_2 p\log p}{\epsilon\sqrt{n}}\right).$$

Later, Kifer et al. (2012) improved the excess risk bound to:

$$O\left(\frac{(LR_2)\|\boldsymbol{\theta}^*\|_2\sqrt{p\log(1/\delta)}}{\epsilon\sqrt{n}}\right),$$

where the perturbation vector $\boldsymbol{b}$ is sampled from the Gaussian distribution mentioned above and where λ parameter is selected appropriately.

We now present our dimension independent excess risk bound of the objective perturbation algorithm with Gaussian noise. The dimension independent analysis for objective perturbation is significantly trickier than output perturbation, since unlike output perturbation the exact distribution of $\boldsymbol{\theta}_{priv}$ is hard to evaluate. Our analysis uses a novel reduction of the analysis of objective perturbation algorithm to the analysis of output perturbation as given in Section 4.1.

Theorem 2. *Let $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$ be i.i.d. samples drawn from a fixed distribution $Dist$ over the domain $\mathcal{X}$. Also, let $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim Dist}[\ell(\langle\boldsymbol{\theta}, x\rangle; y)]$. If the regularization coefficient $\lambda = \frac{LR_2\sqrt{n}}{\sqrt{\|\boldsymbol{\theta}^*\|_2^2+1}}$ in (4), then the following holds with probability at least $7/10$ over the randomness of the training data set $\mathcal{D}$ and the randomness of the noise vector $\boldsymbol{b}$:*

$$\mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim Dist}[\ell(\langle\boldsymbol{\theta}_{priv}, \boldsymbol{x}\rangle; y) - \ell(\langle\boldsymbol{\theta}^*, \boldsymbol{x}\rangle; y)] = O\left(\frac{(\log^2 n)(LR_2)^2\|\boldsymbol{\theta}^*\|_2\sqrt{\log\left(\frac{1}{\delta}\right)+\epsilon}}{\epsilon\sqrt{n}}\right),$$

where $\boldsymbol{\theta}_{priv}$ is the output of the objective perturbation method (4) where $\boldsymbol{b} \sim \mathcal{N}\left(0, \sigma^2\mathbb{I}_p\right)$ and $\sigma^2 = \frac{4(LR_2)^2(\log\frac{1}{\delta}+\epsilon)}{\epsilon^2}$.

Note that the generalization error for objective perturbation is almost identical to that of output perturbation except an extra $poly\log n$ factor and an extra factor of LR_2. We conjecture that extra $poly\log n$ factor is an artifact of our analysis and leave further tightening of our bound as topic of future research.

Proof of Theorem 2. Let,

$$J_{priv}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\ell(\langle\boldsymbol{\theta}, \boldsymbol{x}_i\rangle; y_i) + \frac{\lambda\|\boldsymbol{\theta}\|_2^2}{2n} + \frac{\langle\boldsymbol{b}, \boldsymbol{\theta}\rangle}{n} = \frac{1}{n}\sum_{i=1}^{n}\ell(\langle\boldsymbol{\theta}, \boldsymbol{x}_i\rangle; y_i) + \frac{\lambda}{2n}\left\|\boldsymbol{\theta} + \frac{\boldsymbol{b}}{\lambda}\right\|^2 - \frac{\|\boldsymbol{b}\|_2^2}{2\lambda n}.$$

Let $H_{priv}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\ell(\langle\boldsymbol{\theta}, \boldsymbol{x}_i\rangle - \frac{\langle\boldsymbol{b}, \boldsymbol{x}_i\rangle}{\lambda}; y_i) + \frac{\lambda\|\boldsymbol{\theta}\|_2^2}{2n}$. Note that,

$$\left(\arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} H_{priv}(\boldsymbol{\theta})\right) - \frac{\boldsymbol{b}}{\lambda} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} J_{priv}(\boldsymbol{\theta}).$$

Recall that $\boldsymbol{\theta}_{priv} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} J_{priv}(\boldsymbol{\theta})$. Define $\tilde{\boldsymbol{\theta}}_{priv} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} H_{priv}(\boldsymbol{\theta})$. By the observation above, we have $\boldsymbol{\theta}_{priv} = \tilde{\boldsymbol{\theta}}_{priv} - \frac{\boldsymbol{b}}{\lambda}$. With this observation, we can intuitively think that $\tilde{\boldsymbol{\theta}}_{priv}$ is obtained after executing the output perturbation algorithm with the objective function J_{priv}.

We now artificially increase the dimensionality of the problem from p to $p+1$. For every feature vector $\boldsymbol{x} \in \mathbb{R}^p$, define $\boldsymbol{x}^\dagger \in \mathbb{R}^{p+1}$ as the vector $\boldsymbol{x}$ appended with $-\frac{\langle\boldsymbol{b}, \boldsymbol{x}\rangle}{\lambda}$ in the last coordinate. For the vector $\tilde{\boldsymbol{\theta}}_{priv}$, define $\tilde{\boldsymbol{\theta}}^\dagger_{priv} \in \mathbb{R}^{p+1}$ to be the vector $\tilde{\boldsymbol{\theta}}_{priv}$ appended with one in the last coordinate. Using the definition of $\tilde{\boldsymbol{\theta}}_{priv}$, we have,

$$\tilde{\boldsymbol{\theta}}^\dagger_{priv} = \arg\min_{\boldsymbol{\theta}^\dagger\in\mathbb{R}^{p+1}, \boldsymbol{\theta}^\dagger(p+1)=1} G(\boldsymbol{\theta}^\dagger) = \frac{1}{n}\sum_{i=1}^{n}\ell(\langle\boldsymbol{\theta}^\dagger, \boldsymbol{x}_i^\dagger\rangle; y_i) + \frac{\lambda}{2n}\|\boldsymbol{\theta}^\dagger\|_2^2. \quad (5)$$

Now, by using the observation of the previous section, $\tilde{\boldsymbol{\theta}}^\dagger_{priv} = [\boldsymbol{\theta}_{priv} + \frac{\boldsymbol{b}}{\lambda}; 1]$. Also, using the uniform convergence theorem of Shalev-Shwartz et al. (2009) (re-stated in

Theorem 5 of Appendix A) , the following holds with probability $\geq 9/10$ over the randomness of the sampling of the data set $\mathcal{D}$:

$$\mathop{\mathbb{E}}_{\boldsymbol{x},y}[\ell\left(\langle\tilde{\boldsymbol{\theta}}^{\dagger}_{priv},\boldsymbol{x}\rangle;y\right)] - \mathop{\mathbb{E}}_{\boldsymbol{x},y}[\ell\left(\langle\boldsymbol{\theta}^*,\boldsymbol{x}\rangle - \frac{\langle\boldsymbol{b},\boldsymbol{x}\rangle}{\lambda};y\right)] \leq \frac{\lambda}{2n}\|\boldsymbol{\theta}^*\|_2^2 + O\left(\frac{(LR_2{}^{\dagger})^2}{\lambda}\right), \quad (6)$$

where $R_2{}^{\dagger} = \sqrt{\max_{\boldsymbol{x}\in\mathcal{D}}\|\boldsymbol{x}\|_2^2 + \frac{\langle\boldsymbol{b},\boldsymbol{x}\rangle^2}{\lambda^2}}$. By using the above equation and the observation that $\tilde{\boldsymbol{\theta}}^{\dagger}_{priv} = [\boldsymbol{\theta}_{priv} + \frac{\boldsymbol{b}}{\lambda}; 1]$, we get:

$$\mathop{\mathbb{E}}_{\boldsymbol{x},y}[\ell(\langle\boldsymbol{\theta}_{priv},\boldsymbol{x}\rangle - \frac{\langle\boldsymbol{b},\boldsymbol{x}\rangle}{\lambda};y)] \leq \mathop{\mathbb{E}}_{\boldsymbol{x},y}[\ell(\langle\boldsymbol{\theta}^*,\boldsymbol{x}\rangle - \frac{\langle\boldsymbol{b},\boldsymbol{x}\rangle}{\lambda};y)] + \frac{\lambda}{2n}\|\boldsymbol{\theta}^*\|_2^2 + O\left(\frac{(LR_2{}^{\dagger})^2}{\lambda}\right),$$

Now, by using Lipschitz property of ℓ, we get:

$$E_{\boldsymbol{x},y}[\ell(\langle\boldsymbol{\theta}_{priv},\boldsymbol{x}\rangle;y) - \ell(\langle\boldsymbol{\theta}^*,\boldsymbol{x}\rangle;y)] \leq 2L\cdot\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\left|\frac{\langle\boldsymbol{x},\boldsymbol{b}\rangle}{\lambda}\right|\right] + \frac{\lambda}{2n}\|\boldsymbol{\theta}^*\|_2^2 + \frac{O(1)(LR_2{}^{\dagger})^2}{\lambda}. \quad (7)$$

Since, the noise vector b is a vector of i.i.d. Gaussian random variables with standard deviation σ, by the tail probability of Gaussian random vectors we can conclude that with probability at least $9/10$, $R_2{}^{\dagger} = O\left(R_2\sqrt{1+(\sigma\log n/\lambda)^2}\right)$. By a similar argument, with probability at least $9/10$ over the randomness of b, $\mathbb{E}_{\boldsymbol{x}\sim Dist}\left[\left|\frac{\langle\boldsymbol{x},\boldsymbol{b}\rangle}{\lambda}\right|\right] = O\left(\frac{\sigma R_2}{\lambda}\right)$.

Combining the above observations with (7), we get (w.p. $\geq 9/10$):

$$E_{\boldsymbol{x},y}[\ell(\langle\boldsymbol{\theta}_{priv},\boldsymbol{x}\rangle;y) - \ell(\langle\boldsymbol{\theta}^*,\boldsymbol{x}\rangle;y)] \leq = \frac{O(1)(LR_2)^2}{\lambda}\left(1 + \frac{\sigma^2\log^2 n}{\lambda^2} + \frac{\sigma}{LR_2} + \frac{O(1)\lambda\|\boldsymbol{\theta}^*\|_2^2}{n}\right).$$

Theorem now follows by setting σ according to the theorem statement and by selecting $\lambda = \frac{LR_2\sqrt{n}}{\|\boldsymbol{\theta}^*\|_2}$. □

5. Private ERM Over Simplex with Logarithmic Dependence on Dimensions

In this section we present a differentially private ERM algorithm for L_∞-bounded data points and parameter vectors restricted the simplex, with excess generalization error bound that scales logarithmically in the dimensionality. As a price for improved dependence on the dimensionality, our excess generalization error has worse dependence on the data set size (n). Our error scales as $1/n^{1/3}$, compared to the non-private optimum of $1/\sqrt{n}$.

As a corollary of our result, we derive improved regret bound for differentially private online learning with linear costs considered in (Dwork et al., 2010b). Our regret guarantee depends logarithmically on the dimensionality, compared to the polynomial dependence of (Dwork et al., 2010b). However, our privacy guarantees are weaker, *i.e.,* we guarantee (ϵ,δ)-differential privacy as compared to ϵ-differential privacy of (Dwork et al., 2010b).

5.1. Private ERM with Entropy Regularization

Consider the following ERM problem over simplex $\Delta = \{\boldsymbol{\theta}\in\mathbb{R}^p : \sum_i \theta_i = 1 \text{ and } \forall i, \theta_i > 0\}$:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\theta\in\Delta}\frac{1}{n}\sum_{i=1}^{n}\ell(\langle\boldsymbol{x}_i,\boldsymbol{\theta}\rangle;y_i) + \frac{\lambda}{n}\sum_{j=1}^{p}\theta_j\log(\theta_j), \quad (8)$$

where $(\boldsymbol{x}_i,y_i)\sim Dist, \forall i$. If we choose $\lambda = O(\sqrt{n\log p})$, then using standard Rademacher or covering number arguments one can show that the excess generalization error of $\widehat{\boldsymbol{\theta}}$ is bounded by $O\left(\frac{\sqrt{\log p}R_\infty}{\sqrt{n}}\right)$, where $R_\infty = \max_i\|\boldsymbol{x}_i\|_\infty$ (Kakade et al., 2008; Shalev-Shwartz et al., 2009).

If we use either output or objective perturbation algorithm from Section 4 to obtain a differentially private variant of the above, then the excess generalization error will scale as $O(\sqrt{p}R_\infty/\sqrt{n})$, as opposed to $O\left(\sqrt{\log p}R_\infty/\sqrt{n}\right)$ in the non-private case.

In this section, we present a novel differentially private algorithm for solving ERM over simplex such that the excess risk scales as $O(\log n\log pR_\infty/n^{1/3})$. Our algorithm heavily exploits the fact that the optimization is over a simplex, and involves sampling non-uniformly from probability vectors from the simplex. In our method we first computes the non-private ERM solution $\hat{\boldsymbol{\theta}}$ using (8). Now, we treat $\widehat{\boldsymbol{\theta}}$ as a discrete probability distribution over $\{1,2,\ldots,p\}$ and sample m i.i.d. points $[a_1\ a_2\ \ldots\ a_m], a_j\in[p]$, from $\hat{\boldsymbol{\theta}}$. The private output vector $\boldsymbol{\theta}_{priv}$ is given by:

$$\boldsymbol{\theta}_{priv} = \frac{1}{m}\sum_{j=1}^{m}\boldsymbol{e}_{a_j}, \quad (9)$$

where $\boldsymbol{e}_{a_j}$ denotes the a_j-th canonical basis vector and m is a parameter that we specify later in the theorems for privacy and excess risk.

In the following, we provide formal privacy guarantee (Theorem 3) and excess generalization risk bounds (Theorem 4) for $\boldsymbol{\theta}_{priv}$ (9). See Appendix B for proofs.

Theorem 3 (Privacy guarantee)**.** *Let $\ell:\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ be a differentiable smooth function. Let L_g be the Lipschitz*

constant of ℓ's gradient and let L be its Lipschitz constant. Then:

- $\boldsymbol{\theta}_{priv}$ *is* ϵ*-differentially private when* $m = \frac{\epsilon\lambda}{LR_\infty}\left(4 + \frac{2nR_\infty{}^2}{\lambda}L_g\right)^{-1}$.
- $\boldsymbol{\theta}_{priv}$ *is* (ϵ, δ)*-differentially private when* $m = \left(\frac{\epsilon\lambda}{\log(1/\delta)}\right)^2\left(32 + \frac{16nR_\infty{}^2}{\lambda}L_g\right)^{-2}$.

Theorem 4 (Utility guarantee)**.** *Let* $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$ *be i.i.d. samples drawn from a fixed distribution* $Dist$ *over the domain* $\mathcal{X}$*. Also, let* $\lambda = \frac{n^{2/3}}{\epsilon^{1/3}\log^{1/3} p}$ *and* $m = \left(\frac{\epsilon\lambda}{\log(1/\delta)}\right)^2\left(32 + \frac{16nR_\infty{}^2}{\lambda}L_g\right)^{-2}$*. Then, the following holds with probability at least* $2/3$ *over the randomness of the training data set* $\mathcal{D}$ *and the randomness of the algorithm:*

$$\mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim Dist}[\ell(\langle\boldsymbol{\theta}_{priv}, \boldsymbol{x}\rangle; y) - \ell(\langle\boldsymbol{\theta}^*, \boldsymbol{x}\rangle; y)] = O\left(\frac{(LR_\infty)^3(1+L_g)\log n \log p \log\left(\frac{1}{\delta}\right)}{(n\epsilon)^{1/3}}\right),$$

where $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Delta} \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim Dist}[\ell(\langle\boldsymbol{\theta}, \boldsymbol{x}\rangle; y)]$ *and* $\boldsymbol{\theta}_{priv}$ *is obtained by* (9) *with* m *as given above.*

Similarly, with a proper choice of λ and m, we can obtain an ϵ-differentially private algorithm whose excess risk scale as $O(\text{poly}(\log n \log p)/(\sqrt{\epsilon}n^{1/4}))$.

5.2. Private Online Learning over Simplex

In this section, we study the problem of private online learning over the simplex. The main goal of this section is to demonstrate that the techniques developed in the previous section can be related to existing methods for private online learning. Furthermore, it leads to improved regret bound for differentially private online learning over simplices as compared to the existing works of Dwork et al. (2010a); Jain et al. (2012). In particular, similar to the offline learning setting, existing privacy preserving methods incur additional $O(\sqrt{p})$ multiplicative factor in the regret. In contrast, our proposed method is able to bring down the multiplicative factor to $\sqrt{\log(p)}$ for online learning with linear costs over the simplex.

We assume that the cost functions $\ell_t(\boldsymbol{\theta})$ provided at each step are linear, i.e., $\ell_t(\boldsymbol{\theta}) = \langle\boldsymbol{x}_t, \boldsymbol{\theta}\rangle$. Also, we assume a "weak" adaptive adversary which cannot see the prediction at the t-th step beforehand. Now, we first consider the popular Follow-the-regularized-leader (FTRL) algorithm with entropy regularization for this problem (Shalev-Shwartz, 2011). Using FTRL, the t-th step parameter vector is given by:

$$\widehat{\boldsymbol{\theta}}_{t+1} = \arg\min_{\boldsymbol{\theta}\in\Delta} \frac{1}{t}\sum_{\tau=1}^{t}\langle\boldsymbol{\theta}, \boldsymbol{x}_\tau\rangle + \frac{\lambda}{t}\sum_i \theta_i\log(\theta_i), \quad (10)$$

where τ-th step loss function is given by: $\ell_\tau(\boldsymbol{\theta}) = \langle\boldsymbol{x}_\tau, \boldsymbol{\theta}\rangle$. Notice that the optimization problem in (10) at every step t, is equivalent to (8). Hence, we can use the same method as given in the previous section for computing the privacy preserving update $\boldsymbol{\theta}^{priv}_{t+1}$. In particular, we select $m = 1$ and sample one index a from the probability distribution $\widehat{\boldsymbol{\theta}}_{t+1}$. Hence,

$$\boldsymbol{\theta}^{priv}_{t+1} = \boldsymbol{e}_a,$$

where $a \in \{1, 2, \ldots, p\}$ is sampled from the discrete distribution $\widehat{\boldsymbol{\theta}}_{t+1}$.

In online learning the objective is to bound the *regret*, given by the following:

$$Regret(T) = \mathbb{E}\left[\sum_{t=1}^{T}\langle\boldsymbol{x}_t, \boldsymbol{\theta}^{priv}_t\rangle\right] - \min_{\boldsymbol{\theta}\in\Delta}\sum_t\langle\boldsymbol{x}_t, \boldsymbol{\theta}\rangle \quad (11)$$

From the standard online learning literature (Corollary 2.14 from (Shalev-Shwartz, 2011)), it follows that (11) is upper bounded by $\lambda\log p + \frac{TR^2_\infty}{\lambda}$. And from the privacy analysis of Theorem 3, it follows that the above algorithm is (ϵ, δ)-differentially private (over all the T-steps) as long as $\lambda \geq \frac{R_\infty\sqrt{8T\log(1/\delta)}}{\epsilon}$. Therefore setting, $\lambda = \frac{R_\infty\sqrt{8T\log p\log(1/\delta)}}{\epsilon}$ in (10) directly gives us an (ϵ, δ)-differentially private FTRL algorithm which has the following regret bound in (12).

$$Regret(T) = O\left(\frac{R_\infty\sqrt{T\log p\log(1/\delta)}}{\epsilon}\right) \quad (12)$$

Remark: The regret bound of our private algorithm matches the non-private optimal regret bound up to factors of $\log(1/\delta)/\epsilon$. Our result directly improves on the result of Dwork et al. (2010b) (adapted to our privacy model), where the regret implied by the analysis of (Dwork et al., 2010b) is $O\left(\sqrt{(Tp)}\text{poly}\log(T)/\epsilon\right)$. It is worth mentioning that we provide a weaker (ϵ, δ)-differential privacy guarantee as opposed to ϵ-differential privacy guarantee of Dwork et al. (2010b). By adapted to our privacy model, we mean the following: Dwork et al. (2010b) guarantees the indistinguishability of the presence or absence of one coordinate of a single linear cost function, where as in this paper we ensure the indistinguishability of one complete linear cost function. In fact in the privacy model considered in (Dwork et al., 2010b), they managed to get a regret guarantee that scales as $O(\sqrt{T\log p}\,\text{poly}\log T)$.

6. Experiments

In this section, we first validate our theoretical analysis of the normal distribution based output and objective-perturbation methods (denoted as **Output-Gauss**, **Objective-Gauss** respectively) for L_2 regularized

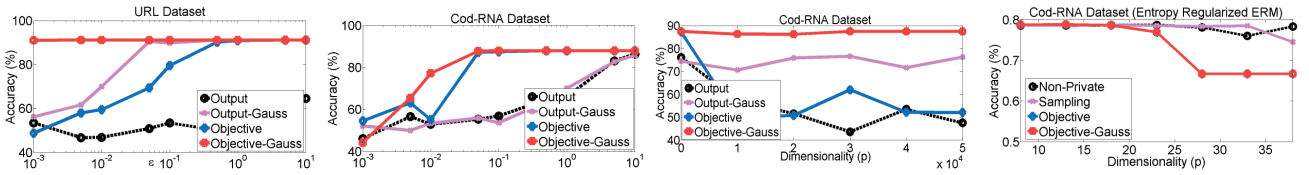

Figure 1. (a), (b) Accuracy achieved by various methods for varying values of ϵ. Objective-Gauss outperforms the other methods significantly. (c) Accuracy achieved by various methods for artificially bloated dimensionality p of the dataset ($\epsilon = 5$, $\delta = 10^{-3}$). Accuracy of both Objective-Gauss and Output-Gauss is nearly constant w.r.t. p, while Objective and Output's accuracy suffers for large p. (d) Accuracy achieved by our sampling based method for artifically bloated dimensionality of the dataset. Our method achieves similar accuracy to the non-private classifier and is slightly better than Objective-Gauss and Objective perturbation methods.

ERM. Next, we show that our sampling based method (denoted as **Sampling**) for entropy regularized ERM also achieves similar accuracy to the non-private ERM.

For our first set of experiments, we apply SVM based classifiers on two benchmark datasets: URL and Cod-RNA. We use a subset of the URL dataset which has $100,000$ data points and its dimensionality is around 20M. Cod-RNA has around 60K data points and its dimensionality is 8. We use 70% of the data for training and the remaining 30% for test. All the results presented are averaged over 20 runs and our code uses a modification of the LIBLINEAR method for solving the perturbed SVM problem.

We evaluate the standard Output-perturbation (denoted as **Output**) and Objective-perturbation (denoted as **Objective**) methods by (Chaudhuri et al., 2011) against **Output-Gauss**, **Objective-Gauss** methods. We first show that as shown by our theoretical analysis, the test error of **Output-Gauss** and **Objective-Gauss** is indeed independent of the training data's dimensionality. Moreover, the test error of the gamma-distribution based perturbation used by (Chaudhuri et al., 2011) indeed increases with the dimensionality.

We first apply all the four methods mentioned above to the URL and the Cod-RNA dataset. We set the regularization parameter $\lambda = 0.001$ and $\delta = 10^{-3}$. Figures 1 (a), (b) shows accuracy achieved by different algorithms on the URL, Cod-RNA dataset, with varying privacy parameter ϵ. Clearly, Objective perturbation methods are significantly better than the Output perturbation based methods. Moreover, normal distribution based perturbation is able to obtain significantly higher accuracy.

We now study how the accuracy of various methods vary with the dimensionality of the dataset (Figure 1 (c)). To this end, we use Cod-RNA dataset and artificially blot up its dimensionality by adding zeros to the feature space vectors. Note that this does not effect the baseline classifier and its accuracy. Clearly, as predicted by our analysis, accuracy of the Objective-Gauss and the Output-Gauss does not change with large increments in the ambient dimensionality p. In contrast, accuracy of both Objective and Output perturbation algorithms suffer heavily for larger p.

Above experiments suggest that our theoretical analysis of objective perturbation (Theorem 2) may be loose. One open problem is to investigate the tightness of the current analysis.

Finally, we study our **Sampling** method for entropy regularized ERM. To this end, we first solve a entropy regularized least squares problem. We then threshold predicted values to obtain class values. We conduct experiments on Cod-RNA dataset with $\epsilon = 10, \delta = 10^{-3}$ and by using 70% of the data for training and the remaining for test. Figure 1 (d) shows accuracy achieved by various methods with artificially bloated dimensionality of the data. Clearly, accuracy achieved by our **Sampling** method is similar to the non-private classifier and is significantly better than both **Objective** and **Objective-Gauss** for larger values of p.

7. Discussion

Our dimension independence analysis for objective perturbation holds only if the optimization is over the unconstrained space $\mathbb{R}^p$ (see Theorem 1). In fact, it can be easily shown that both output and objective perturbation approach will fail to provide dimension independent risk bound if the solution of the ERM is contrained to lie in an arbitrary convex set $\mathcal{C} \subset \mathbb{R}^p$. For example, say optimal solution $\boldsymbol{\theta}$ is constrained to lie in the positive orthant and let the optimal solution to the ERM is 0. In this case, it is easy to see that both objective and output perturbation methods will give excess risk that scales as $\sqrt{p}$.

Another limitation of our analysis is that we need to assume that the regularization function of the ERM is either squared L_2 norm or the negative entropy function. Tight excess risk analysis (in terms of dependence on p) for other regularization functions (for example, L_1 norm) is still an open problem and is left as a topic of future research.

Finally, our algorithm for privacy preserving entropy regularized ERM uses a sampling based technique that is significantly different from the existing differential privacy learning techniques. For future research, we want to explore this technique in more detail and possibly, apply the technique to other similar problems.

References

Chaudhuri, Kamalika and Hsu, Daniel. Sample complexity bounds for differentially private learning. *Journal of Machine Learning Research - Proceedings Track*, 19:155–186, 2011.

Chaudhuri, Kamalika and Monteleoni, Claire. Privacy-preserving logistic regression. In *NIPS*, 2008.

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

Duchi, John C., Jordan, Michael I., and Wainwright, Martin J. Privacy aware learning. In *NIPS*, 2012.

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

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

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

Dwork, Cynthia, Naor, Moni, Pitassi, Toniann, and Rothblum, Guy N. Differential privacy under continual observation. In *Proceedings of the 42nd ACM symposium on Theory of computing*, 2010b.

Dwork, Cynthia, Rothblum, Guy N, and Vadhan, Salil. Boosting and differential privacy. In *FOCS*, 2010c.

Hazan, Elad, Agarwal, Amit, and Kale, Satyen. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 2007.

Jain, Prateek and Thakurta, Abhradeep. Differentially private kernel learning. In *ICML*, 2013.

Jain, Prateek, Kothari, Pravesh, and Thakurta, Abhradeep. Differentially private online learning. In *COLT*, 2012.

Kakade, Sham M, Sridharan, Karthik, and Tewari, Ambuj. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NIPS*, 2008.

Kasiviswanathan, Shiva Prasad and Smith, Adam. A note on differential privacy: Defining resistance to arbitrary side information. *CoRR*, arXiv:0803.39461 [cs.CR], 2008.

Kifer, Daniel, Smith, Adam, and Thakurta, Abhradeep. Private convex empirical risk minimization and high-dimensional regression. In *COLT*, 2012.

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

Negahban, Sahand, Ravikumar, Pradeep, Wainwright, Martin J., and Yu, Bin. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *NIPS*, 2009.

Pathak, Manas A., Rane, Shantanu, and Raj, Bhiksha. Multiparty differential privacy via aggregation of locally trained classifiers. In *NIPS*, 2010.

Rubinstein, Benjamin IP, Bartlett, Peter L, Huang, Ling, and Taft, Nina. Learning in a large function space: Privacy-preserving mechanisms for svm learning. *arXiv preprint arXiv:0911.5708*, 2009.

Shalev-Shwartz, Shai. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 2011.

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

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

Thakurta, Abhradeep Guha. Differentially private convex optimization for empirical risk minimization and high-dimensional regression. *PhD Dissertation*, 2013.

Zhang, Tong. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.