Noname manuscript No.
(will be inserted by the editor)

Distributionally-Robust Machine Learning Using Locally Differentially-Private Data

Farhad Farokhi

the date of receipt and acceptance should be inserted later

Abstract We consider machine learning, particularly regression, using locally-differentially private datasets. The Wasserstein distance is used to define an ambiguity set centered at the empirical distribution of the dataset corrupted by local differential privacy noise. The radius of the ambiguity set is selected based on privacy budget, spread of data, and size of the problem. Machine learning with private dataset is rewritten as a distributionally-robust optimization. For general distributions, the distributionally-robust optimization problem can be relaxed as a regularized machine learning problem with the Lipschitz constant of the machine learning model as a regularizer. For Gaussian data, the distributionally-robust optimization problem can be solved exactly to find an optimal regularizer. Training with this regularizer can be posed as a semi-definite program.

Keywords Local differential privacy · Machine learning · Distributionally-robust optimization · Regularization · Wasserstein distance

1 Introduction

Advances in machine learning have opened new possibilities for data analytic to address important societal challenges. However, these achievements can be stifled by privacy concerns. Local differential privacy, a variant of the popular differential privacy framework [1, 2], has been touted as an approach for providing privacy guarantees in the presence of an untrusted aggregators or analysts [3–5]. This is because, with local differential privacy, the data can be freely shared. Even, commercial entities have started using local differential privacy to deploy privacy-preserving data aggregation mechanisms [6–8].

F. Farokhi
Department of Electrical and Electronic Engineering, University of Melbourne, Australia.
E-mail: farhad.farokhi@unimelb.edu.au

The additive noise in local differential privacy can degrade the performance of machine learning models. Several studies have looked into providing bounds for the performance degradation caused by local differential privacy noise as a function of the privacy budget and the dataset size [5, 9–12]. These studies however do not use recent advances in distributionally-robust optimization and machine learning (see, e.g., [13, 14]) to compute robust machine learning models with out-of-sample performance guarantees in the presence of local differential privacy noise. They focus on understanding the effect of privacy-preserving noise on established machine learning algorithms.

Distributionally-robust optimization considers uncertain stochastic programs [15] with the ambiguity set for the distribution modeled by discrete distributions [16], moment constraints [17], Kullback-Leibler divergence [18], and the Wasserstein distance [13]. Distributionally-robust optimization has shown significant promises in adversarial machine learning [19] or machine learning with outlier data [20]. It has been used to devise robust classifiers, such as support vector machines [21] and logistic regression [22]. However, it has not been used for training robust machine learning models based on privatized data. Furthermore, the optimal regularizer in this paper has not been previously presented for linear regression with Gaussian data.

In this paper, we use the Wasserstein distance to define an ambiguity set centered at the empirical distribution of the training dataset that is corrupted with local differential privacy noise. This ambiguity set is shown to contain the probability distribution of unperturbed data. The radius of the ambiguity set is a function of the privacy budget, spread of the data, and the size of the problem (i.e., number of inputs and outputs of the machine learning model). Armed with this description of the ambiguity set, we can cast the problem of learning with locally-differentially private data as a distributionally-robust optimization problem. We show that, for general distributions, an upper bound for the worst-case expected loss in the distributionally-robust optimization problem is the empirical sampled-averaged loss plus the Lipschitz-constant of the loss function. Using this, we can relax the distributionally-robust optimization problem as a regularized machine learning problem with the Lipschitz constant as a regularizer. For Gaussian data, the distributionally-robust optimization problem can be solved exactly to find an optimal regularizer for the problem. This approach results in an entirely new regularizer for linear regression.

It should be noted that the main difference of this paper comparing the existing literature in distributionally-robust machine learning is the modeling of locally differentially-private data through the ambiguity set. This is the novelty of the paper and the rest of the results follow from this derivation using existing results in distributionally-robust machine learning [13, 14, 23].

The rest of the paper is organized as follows. In Section 2, we show that machine learning with private dataset can rewritten as a distributionally-robust optimization, which can be relaxed as a regularized machine learning problem with the Lipschitz constant of the machine learning model as a regularizer. In Section 3, we focus on Gaussian data and show that the distributionally-robust optimization problem can be solved exactly to find an optimal regular-

izer. Finally, we present the experiments in Section 4 and conclude the paper in Section 5.

2 Distributionally-Robust Machine Learning with Private Data

We consider supervised learning using training dataset $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathcal{X} \subseteq \mathbb{R}^{p_x}$ is the input/feature vector and $y_i \in \mathcal{Y} \subseteq \mathbb{R}^{p_y}$ is the output/label. The training dataset is composed of independently and identically distributed (i.i.d.) samples from distribution $\mathbb{P}$.

Training machine learning models refers to extracting a model $\mathfrak{M} : \mathbb{R}^{p_x} \times \mathbb{R}^{p_\theta} \to \mathbb{R}^{p_y}$ to describe the relationship between inputs and outputs distributed according to $\mathbb{P}$. This can be done by solving the stochastic program

$$J^* := \min_{\theta \in \Theta} \mathbb{E}^{\mathbb{P}}\{\ell(\mathfrak{M}(x;\theta), y)\}, \tag{1}$$

where θ is the machine learning model parameter, $\Theta \subseteq \mathbb{R}^{p_\theta}$ is the set of feasible parameters, and $\ell : \mathbb{R}^{p_y} \times \mathbb{R}^{p_y} \to \mathbb{R}$ is the loss function.

In the absence of the knowledge of $\mathbb{P}$, the training dataset $\{(x_i, y_i)\}_{i=1}^n$, i.e., its samples, can be used to solve the sample-averaged approximation problem

$$\hat{J} := \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathfrak{M}(x_i;\theta), y_i). \tag{2}$$

The approximation in (2) is often the starting point of machine learning. This is because (2) can be a good proxy for (1), when n is large enough, in the sense of probably approximately correct (PAC) learnability [24]. We make the following standing assumption regarding the distribution of the training data.

Assumption 1 $\mathbb{E}^{\mathbb{P}}\{\exp(\|\xi\|^a)\} < \infty$ *for some* $a > 1$.

This assumption implies that $\mathbb{P}$ is light-tailed. All probability distributions with compact support are light-tailed; however, unbounded noises, such as Gaussian or Laplace, are also light-tailed. This is often an implicit assumption in the machine learning as, for heavy-tailed distributions, the sample average of the loss might not generally converge to the expected loss [25].

Due to privacy concerns, the training dataset is sometimes replaced with a noisy dataset $\{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^n$ in which

$$\tilde{x}_i := x_i + w_i, \tag{3a}$$

$$\tilde{y}_i := y_i, \tag{3b}$$

where $(w_i)_{i=1}^n$ are i.i.d. samples from distribution $\mathbb{W}$. Local differential privacy is a useful and versatile notion of privacy.

Definition 1 (Local Differential Privacy) The reporting mechanism with additive noise in (3) is (ϵ, δ)-locally differentially private if, for all $(x, y), (x', y') \in \mathcal{X} \times \mathcal{Y}$ and any Lebesgue-measurable set $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{Y}$, $\mathbb{P}\{(\tilde{x}_i, \tilde{y}_i) \in \mathcal{A} | x_i = x, y_i = y\} \leq \exp(\epsilon)\mathbb{P}\{(\tilde{x}_i, \tilde{y}_i) \in \mathcal{A} | x_i = x', y_i = y'\} + \delta$.

Assumption 2 $\mathcal{X} \subseteq [\underline{x}, \overline{x}]^{p_x}$.

The box constraint nature of Assumption 2 is not strictly-speaking necessary; however, it simplifies the closed-form expression of the results. The results can be readily extended to any compact sets by using diameter of the set. As demonstrated in the next theorem, we can ensure differential privacy using Laplace and Gaussian additive noises.

Theorem 1 *The following statements hold:*

1. *For $\epsilon > 0$, (3) is ϵ-locally differentially private if w_i is a vector of zero-mean i.i.d. Laplace noise with scale Δ/ϵ, where $\Delta := (\overline{x} - \underline{x})p_x$;*
2. *For $\epsilon, \delta > 0$, (3) is (ϵ, δ)-locally differentially private if w_i is a vector of zero-mean i.i.d. Gaussian noise with standard deviation $\sigma := \sqrt{2\log(1.25/\delta)}\Delta/\epsilon$.*

Proof The proof for the Laplace mechanism follows from [2, Theorem 3.6] while noting that ℓ_1-sensitivity of the query (which is equal to the identity function here) is given by Δ. The proof for the Gaussian noise follows from [2, Theorem A.1]. Note that ℓ_1-sensitivity is an upper bound for ℓ_2-sensitivity.

The privacy-preserving records in the training dataset $(\tilde{x}_i, \tilde{y}_i)_{i=1}^n$ are i.i.d. samples from $\mathbb{D}$, which can be characterized by the convolution of $\mathbb{P}$ and $\mathbb{W}$ [26]. We can define the empirical probability distribution

$$\widehat{\mathbb{D}}_n := \frac{1}{n}\sum_{i=1}^{n} \mathfrak{d}_{(\tilde{x}_i, \tilde{y}_i)},$$

where $\mathfrak{d}_\xi$ is the Dirac distribution function. We can rewrite (2) as

$$\hat{J} := \min_{\theta \in \Theta} \mathbb{E}^{\widehat{\mathbb{D}}_n}\{\ell(\mathfrak{M}(x;\theta), y)\}. \tag{4}$$

The empirical distribution $\widehat{\mathbb{D}}_n$ is in a vicinity (defined using the Wasserstein distance) of the original probability distribution $\mathbb{P}$ with a high probability. The Wasserstein distance is $\mathfrak{W}_q(\mathbb{P}_1, \mathbb{P}_2) := \inf_{\Pi} \left[\int_{\Xi^2} \|\xi_1 - \xi_2\|^q \Pi(\mathrm{d}\xi_1, \mathrm{d}\xi_2)\right]^{1/q}$, where Π is a joint disribution on ξ_1 and ξ_2 with marginals $\mathbb{P}_1$ and $\mathbb{P}_2$, respectively.

Theorem 2 *Assume that $\mathbb{W}$ is the distribution in Theorem 1. There exist constants $c_1, c_2 > 0$ such that $\mathbb{D}^n\{\mathfrak{W}_1(\widehat{\mathbb{D}}_n, \mathbb{P}) \leq \zeta(\gamma) + \sqrt{2p}\Delta/\epsilon\} \geq 1-\gamma$, for the Laplace mechanism and $\mathbb{D}^n\{\mathfrak{W}_1(\widehat{\mathbb{D}}_n, \mathbb{P}) \leq \zeta(\gamma) + \sqrt{2\log(1.25/\delta)p}\Delta/\epsilon\} \geq 1-\gamma$, for the Gaussian mechanism, where*

$$\zeta(\gamma) := \begin{cases} \left(\dfrac{\log(c_1/\gamma)}{c_2 n}\right)^{1/\max\{p,2\}}, & n \geq \dfrac{\log(c_1/\gamma)}{c_2}, \\ \left(\dfrac{\log(c_1/\gamma)}{c_2 n}\right)^{1/a}, & n < \dfrac{\log(c_1/\gamma)}{c_2}, \end{cases}$$

for all $n \geq 1$, $p = p_x + p_y \neq 2$, and $\gamma > 0$.

Proof Note that, since $\mathbb{P}$ is light-tailed, $\mathbb{D}$ is also light-tailed if we use the privacy-preserving Laplace or Gaussian noises in Theorem 1. Following [13, Theorem 3.4], we know that $\mathbb{D}^n\{\mathfrak{W}_1(\widehat{\mathbb{D}}_n, \mathbb{D}) \leq \zeta(\gamma)\} \leq 1-\gamma$. Using [27, Lemma 8.6] while noting that $\mathfrak{W}_1(\mathbb{P}, \mathbb{P}) = 0$, we get $\mathfrak{W}_1(\mathbb{D}, \mathbb{P}) \leq \mathfrak{W}_1(\mathbb{P}, \mathbb{P}) + \mathfrak{W}_1(\mathfrak{d}_0, \mathbb{W}) = \mathfrak{W}_1(\mathfrak{d}_0, \mathbb{W}) \leq \mathbb{E}^{\mathbb{W}}\{\|w\|\}$. The Jensen's inequality [28, p. 27] implies that $\mathbb{E}^{\mathbb{W}}\{\|w\|\} \leq \sqrt{\mathbb{E}^{\mathbb{W}}\{\|w\|^2\}}$. Furthermore, for the Laplace noise, we get $\mathbb{E}^{\mathbb{W}}\{\|w\|^2\} = \text{trace}(\mathbb{E}^{\mathbb{W}}\{ww^\top\}) = 2p\Delta^2/\epsilon^2$, and, as a result, $\mathfrak{W}_1(\mathbb{D}, \mathbb{P}) \leq \sqrt{2p}\Delta/\epsilon$. Therefore, $\mathfrak{W}_1(\widehat{\mathbb{D}}_n, \mathbb{P}) \leq \zeta(\gamma) + \sqrt{2p}\Delta/\epsilon$ if $\mathfrak{W}_1(\widehat{\mathbb{D}}_n, \mathbb{D}) \leq \zeta(\gamma)$, which implies that $\mathbb{D}^n\{\mathfrak{W}_1(\widehat{\mathbb{D}}_n, \mathbb{P}) \leq \zeta(\gamma) + \sqrt{2p}\mathfrak{d}/\epsilon\} \geq 1-\gamma$. The proof for the Gaussian noise is the same with the exception that $\mathbb{E}^{\mathbb{W}}\{\|w\|^2\} = \text{trace}(\mathbb{E}^{\mathbb{W}}\{ww^\top\}) = 2p\log(1.25/\delta)\Delta^2/\epsilon^2$.

Theorem 2 is the most fundamental result of this paper. Note that the main difference of this paper comparing the existing literature in distributionally-robust machine learning [13, 14, 23] is the modeling of locally differentially-private data through the ambiguity set, which is the essence of Theorem 2.

If we select ρ large enough, the original distribution $\mathbb{P}$ belongs to the ambiguity set $\{\mathbb{G} : \mathfrak{W}_1(\mathbb{G}, \widehat{\mathbb{D}}_n) \leq \rho\}$. This observation motivates training the model by solving the distributionally-robust optimization problem in

$$\hat{J}_n := \min_{\theta \in \Theta} \sup_{\mathbb{G}:\mathfrak{W}_1(\mathbb{G}, \widehat{\mathbb{D}}_n) \leq \rho} \mathbb{E}^{\mathbb{G}}\{\ell(\mathfrak{M}(x;\theta), y)\}, \tag{5}$$

for constant $\rho > 0$. The correct value of ρ is discussed in the next theorem, which follows from [13, Theorem 3.4].

Theorem 3 *Assume that $\mathbb{W}$ is the distribution in Theorem 1. If $\rho = \zeta(\beta) + \sqrt{2p}\Delta/\epsilon$ for the Laplace mechanism or if $\rho = \zeta(\beta) + \sqrt{2p\log(1.25/\delta)}\Delta/\epsilon$ for the Gaussian mechanism, then $\mathbb{D}^n\{J^* \leq \mathbb{E}^{\mathbb{P}}\{\ell(\mathfrak{M}(x;\hat{\theta}_n), y)\} \leq \hat{J}_n\} \geq 1-\beta$, where $\beta \in (0,1)$ is a significance parameter and the trained model parameter $\hat{\theta}_n \in \Theta$ is the minimizer of (5).*

Proof We focus on the Laplace privacy-preserving noise. In this case, Theorem 2 implies that $\mathbb{D}^n\{\mathfrak{W}_1(\widehat{\mathbb{D}}_n, \mathbb{P}) \leq \zeta(\gamma) + \sqrt{2p}\Delta/\epsilon\} \geq 1-\gamma$. Therefore, with probability of at least $1-\gamma$, $\mathbb{P} \in \{\mathbb{G} : \mathfrak{W}_1(\mathbb{G}, \widehat{\mathbb{D}}_n) \leq \rho\}$ if $\rho = \zeta(\gamma) + \sqrt{2p}\Delta/\epsilon$. This implies that, with probability of at least $1-\gamma$, $\mathbb{E}^{\mathbb{P}}\{\ell(\mathfrak{M}(x;\theta), y)\} \leq \sup_{\mathbb{G}:\mathfrak{W}_1(\mathbb{G}, \widehat{\mathbb{D}}_n) \leq \rho} \mathbb{E}^{\mathbb{G}}\{\ell(\mathfrak{M}(x;\theta), y)\}$ for all θ. Therefore, it must also hold for $\theta = \hat{\theta}_n$ and, as a result, $\mathbb{E}^{\mathbb{P}}\{\ell(\mathfrak{M}(x;\hat{\theta}_n), y)\} \leq \hat{J}_n$ with probability of at least $1-\gamma$. Note that, by definition, $\mathbb{E}^{\mathbb{P}}\{\ell(\mathfrak{M}(x;\hat{\theta}_n), y)\} \geq J^*$. The proof for the Gaussian privacy-preserving noise follows an identical line of reasoning while replacing $\rho = \zeta(\gamma) + \sqrt{2p}\Delta/\epsilon$ with $\rho = \zeta(\gamma) + \sqrt{2p\log(1.25/\delta)}\Delta/\epsilon$.

The optimization problem in (5) involves taking a supremum over the probability density function. This is an infinite-dimensional optimization problem and is hence computationally difficult to solve. We can relax this problem by defining the regularized sample-averaged optimization problem in

$$\tilde{J}_n := \min_{\theta \in \Theta} \Big[\mathbb{E}^{\widehat{\mathbb{D}}_n}\{\ell(\mathfrak{M}(x;\theta), y)\} + \rho L(\theta)\Big]. \tag{6}$$

We can still prove a performance guarantee for the optimizer of (6).

Theorem 4 *Assume that $\mathbb{W}$ is the distribution in Theorem 1 and $\ell(\mathfrak{M}(x;\theta),y)$ is $L(\theta)$-Lipschitz continuous in (x,y) for $\theta \in \Theta$. If $\rho = \zeta(\beta) + \sqrt{2p}\Delta/\epsilon$ for the Laplace mechanism or if $\rho = \zeta(\beta) + \sqrt{2p\log(1.25/\delta)}\Delta/\epsilon$ for the Gaussian mechanism, then $\mathbb{D}^n\{J^* \leq \mathbb{E}^{\mathbb{P}}\{\ell(\mathfrak{M}(x;\tilde{\theta}_n),y)\} \leq \tilde{J}_n\} \geq 1-\beta$, where the trained model parameter $\tilde{\theta}_n \in \Theta$ is the minimizer of* (6).

Proof Again, we focus on the Laplace privacy-preserving noise. As shown in the proof of Theorem 3, $\mathbb{E}^{\mathbb{P}}\{\ell(\mathfrak{M}(x;\theta),y)\} \leq \sup_{\mathbb{G}:\mathfrak{W}_1(\mathbb{G},\widehat{\mathbb{D}}_n)\leq\rho} \mathbb{E}^{\mathbb{G}}\{\ell(\mathfrak{M}(x;\theta),y)\}$ for all θ with probability of at least $1-\gamma$. The duality theorem of Kantorovich and Rubinstein [29] implies that $\sup_{\mathbb{G}:\mathfrak{W}_1(\mathbb{G},\widehat{\mathbb{D}}_n)\leq\rho} \mathbb{E}^{\mathbb{G}}\{\ell(\mathfrak{M}(x;\theta),y)\} \leq \mathbb{E}^{\widehat{\mathbb{D}}_n}\{\ell(\mathfrak{M}(x;\theta),y)\} + L(\theta)\rho$ with probability of at least $1-\gamma$; see the proof of [30, Lemma 2] for more detailed derivations. Both these inequalities must also hold for $\theta = \tilde{\theta}_n$ and, as a result, $\mathbb{E}^{\mathbb{P}}\{\ell(\mathfrak{M}(x;\hat{\theta}_n),y)\} \leq \tilde{J}_n$ with probability of at least $1-\gamma$. Note that, again by definition, $\mathbb{E}^{\mathbb{P}}\{\ell(\mathfrak{M}(x;\hat{\theta}_n),y)\} \geq J^*$. The proof for the Gaussian privacy-preserving noise follows an identical line of reasoning while replacing $\rho = \zeta(\gamma)+\sqrt{2p}\Delta/\epsilon$ with $\rho = \zeta(\gamma)+\sqrt{2p\log(1.25/\delta)}\Delta/\epsilon$.

Theorem 4 shows that by regularizing the sample-averaged cost function, we can train a machine learning model that performs better in the presence of locally differentially private noise. This is an interesting observation demonstrating the value of regularization in machine learning with private data. Following Theorem 4, $\rho = \mathcal{O}(\epsilon^{-1})$ because $\zeta(n) \approx 0$ for large n. Therefore, the regularization weight ρ should increase when reducing ϵ.

Remark 1 Let $p_y = 1$; otherwise, each output can be treated independently. For linear regression, $\mathfrak{M}(x;\theta) = \theta^\top[x^\top\ 1]^\top$ and $\ell(\mathfrak{M}(x;\theta),y) = (\mathfrak{M}(x;\theta)-y)^2/2$. Assume (x,y) belongs to compact set $\mathcal{X}\times\mathcal{Y} \subseteq \mathbb{R}^{p_x}\times\mathbb{R}$. We have $L(\theta) = (X+1+Y)\|\theta\|_*^2$, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$, and $X = \max_{x\in\mathcal{X}}\|x\|$ and $Y = \max_{y\in\mathcal{Y}}|y|$.

3 Linear Regression with Gaussian Data

In this section, we consider the specific case that $(x_i,y_i)_{i=1}^n$ is Gaussian distributed with mean μ and covariance Σ. Furthermore, we assume that $\mathfrak{M}(x;\theta) = Ax+B$ with A,B modeling the machine learning model parameters (instead of θ) and $\ell(\mathfrak{M}(x;A,B),y) = \|\mathfrak{M}(x;A,B)-y\|^2$. Therefore, by using the Gaussian mechanism in Theorem 1, $(\tilde{x}_i,\tilde{y}_i)_{i=1}^n$ is also Gaussian distributed. Note that Assumption 2 no longer holds in this section (as the Gaussian process behind the data has an infinite support). However, with high probability, the data belongs to a bounded set. Therefore, we can adopt local randomized differential privacy instead of local differential privacy [31].

In this case, we can redefine the empirical probability distribution $\widehat{\mathbb{D}}_n$ to be Gaussian with mean $\widehat{\mu}_n$ and covariance $\widehat{\Sigma}_n$, where

$$\widehat{\mu}_n := \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix}\tilde{x}_i\\ \tilde{y}_i\end{bmatrix}, \quad \widehat{\Sigma}_n := \frac{1}{n-1}\sum_{i=1}^{n}\left(\begin{bmatrix}\tilde{x}_i\\ \tilde{y}_i\end{bmatrix} - \widehat{\mu}_n\right)\left(\begin{bmatrix}\tilde{x}_i\\ \tilde{y}_i\end{bmatrix} - \widehat{\mu}_n\right)^\top.$$

Note that $\mathfrak{W}_2(\widehat{\mathbb{D}}_n, \mathbb{D}) \to 0$ with probability one as $n \to \infty$ [32, Theorem 2.1]. Here, we consider the big data regime ($n \gg 1$) so, without loss of generality, $\mu \approxeq \widehat{\mu}_n$. Therefore, by using [32, Theorem 2.1], the original distribution $\mathbb{P}$ would belong to the ambiguity set $\{\mathbb{G} : \mathbb{G} \text{ is Gaussian} \wedge \mathbb{E}^{\mathbb{G}}\{(x,y)\} = \widehat{\mu}_n \wedge \mathfrak{W}_2(\mathbb{G}, \widehat{\mathbb{D}}_n) \leq \rho\}$ if we select $\rho = \sqrt{2\log(1.25/\delta)p}\Delta/\epsilon$. This observation motivates training the machine learning model by solving the distributionally-robust optimization problem in

$$\hat{J}_n := \min_{A,B} \sup_{\substack{\mathbb{G}\,:\,\mathbb{G} \text{ is Gaussian} \\ \mathbb{E}^{\mathbb{G}}\{(x,y)\} = \widehat{\mu}_n \\ \mathfrak{W}_2(\mathbb{G}, \widehat{\mathbb{D}}_n) \leq \rho}} \mathbb{E}^{\mathbb{G}}\{\ell(\mathfrak{M}(x;A,B),y)\}. \tag{7}$$

Theorem 5 *The optimization problem in* (7) *is equivalent with*

$$\hat{J}_n := \min_{A,B} \mathbb{E}^{\widehat{\mathbb{D}}_n}\{\ell(\mathfrak{M}(x;A,B),y)\} + \lambda(A), \tag{8}$$

where

$$\begin{aligned}\lambda(A) := \inf_{\xi:\xi I \succ \begin{bmatrix}A & -I\end{bmatrix}^\top \begin{bmatrix}A & -I\end{bmatrix}} & \Big[\xi(\rho^2 - \mathit{trace}(\widehat{\Sigma}_n)) \\ & + \xi^2 \mathit{trace}((\xi I - \begin{bmatrix}A & -I\end{bmatrix}^\top \begin{bmatrix}A & -I\end{bmatrix})^{-1}\widehat{\Sigma}_n)\Big] \\ & - \begin{bmatrix}A & -I\end{bmatrix}\widehat{\Sigma}_n\begin{bmatrix}A & -I\end{bmatrix}^\top.\end{aligned}$$

Proof Following [33, Proposition 7], if $\widehat{\mathbb{D}}_n$ and $\mathbb{G}$ are both Gaussian with same mean $\mu \approxeq \widehat{\mu}_n$, we get $\mathfrak{W}_2(\widehat{\mathbb{D}}_n, \mathbb{G})^2 = \mathrm{trace}(\Sigma_G + \widehat{\Sigma}_n - 2(\widehat{\Sigma}_n^{1/2}\Sigma_G\widehat{\Sigma}_n^{1/2})^{1/2})$, where Σ_G denotes the covariance of $\mathbb{G}$. Note that $\mathbb{E}^{\mathbb{G}}\{\ell(\mathfrak{M}(x;A,B),y)\} = \mathrm{trace}(\begin{bmatrix}A & -I\end{bmatrix}(\Sigma_G + \widehat{\mu}_n\widehat{\mu}_n^\top)\begin{bmatrix}A & -I\end{bmatrix}^\top + BB^\top + \begin{bmatrix}A & -I\end{bmatrix}\widehat{\mu}_n B^\top + B\widehat{\mu}_n^\top\begin{bmatrix}A & -I\end{bmatrix}^\top)$. Using [14, Proposition 2.8], we get

$$\begin{aligned}\sup_{\substack{\mathbb{G}\,:\,\mathbb{G} \text{ is Gaussian} \\ \mathbb{E}^{\mathbb{G}}\{(x,y)\} = \widehat{\mu}_n \\ \mathfrak{W}_2(\mathbb{G}, \widehat{\mathbb{D}}_n) \leq \rho}} \mathbb{E}^{\mathbb{G}}\{\ell(\mathfrak{M}(x;A,B),y)\} = & \mathrm{trace}\Big(\begin{bmatrix}A & -I\end{bmatrix}\widehat{\mu}_n\widehat{\mu}_n^\top\begin{bmatrix}A & -I\end{bmatrix}^\top \\ & + BB^\top + \begin{bmatrix}A & -I\end{bmatrix}\widehat{\mu}_n B^\top \\ & + B\widehat{\mu}_n^\top\begin{bmatrix}A & -I\end{bmatrix}^\top\Big) + f(A),\end{aligned}$$

where $f(A) := \inf_{\xi:\xi I \succ \begin{bmatrix}A & -I\end{bmatrix}^\top \begin{bmatrix}A & -I\end{bmatrix}} [\xi(\rho^2 - \mathrm{trace}(\widehat{\Sigma}_n)) + \xi^2\mathrm{trace}((\xi I - \begin{bmatrix}A & -I\end{bmatrix}^\top \begin{bmatrix}A & -I\end{bmatrix})^{-1}\widehat{\Sigma}_n)]$. The rest follows from that $\mathrm{trace}(\begin{bmatrix}A & -I\end{bmatrix}\widehat{\mu}_n\widehat{\mu}_n^\top\begin{bmatrix}A & -I\end{bmatrix}^\top + BB^\top + \begin{bmatrix}A & -I\end{bmatrix}\widehat{\mu}_n B^\top + B\widehat{\mu}_n^\top\begin{bmatrix}A & -I\end{bmatrix}^\top) = \mathbb{E}^{\widehat{\mathbb{D}}_n}\{\ell(\mathfrak{M}(x;A,B),y)\} - \begin{bmatrix}A & -I\end{bmatrix}\widehat{\Sigma}_n\begin{bmatrix}A & -I\end{bmatrix}^\top$.

The regularization in Theorem 5 is completely novel in the context of linear regression. In what follows, we provide a more tractable formulation for (8) in the form of semi-definite program. Note that semi-definite programs are shown to be solvable in polynomial time by interior-point method [34].

Theorem 6 *The solution to* (8) *is given by*

$$\min_{A,B,\xi,Z} \xi(\rho^2 - \mathit{trace}(\widehat{\Sigma}_n)) + \mathit{trace}(Z) + \mathit{trace}\Big(\begin{bmatrix} A & -I \end{bmatrix} \widehat{\mu}_n \widehat{\mu}_n^\top \begin{bmatrix} A & -I \end{bmatrix}^\top + BB^\top + \begin{bmatrix} A & -I \end{bmatrix} \widehat{\mu}_n B^\top + B \widehat{\mu}_n^\top \begin{bmatrix} A & -I \end{bmatrix}^\top \Big), \tag{9a}$$

$$\text{s.t.} \quad \begin{bmatrix} Z & \widehat{\Sigma}_n^{1/2}\xi & 0 \\ \widehat{\Sigma}_n^{1/2}\xi & \xi I & \begin{bmatrix} A & -I \end{bmatrix}^\top \\ 0 & \begin{bmatrix} A & -I \end{bmatrix} & I \end{bmatrix} \succeq 0. \tag{9b}$$

Proof Let $Z \succeq 0$ be such that $\xi^2 \widehat{\Sigma}_n^{1/2} (\xi I - \begin{bmatrix} A & -I \end{bmatrix}^\top \begin{bmatrix} A & -I \end{bmatrix})^{-1} \widehat{\Sigma}_n^{1/2} \preceq Z$. Using the Schur complement [35], we can transform this inequality into

$$\begin{bmatrix} Z & \widehat{\Sigma}_n^{1/2}\xi & 0 \\ \widehat{\Sigma}_n^{1/2}\xi & \xi I & \begin{bmatrix} A & -I \end{bmatrix}^\top \\ 0 & \begin{bmatrix} A & -I \end{bmatrix} & I \end{bmatrix} \succeq 0. \tag{10}$$

Using the Schur complement, the constraint in computing $f(A)$ becomes

$$\begin{bmatrix} \xi I & \begin{bmatrix} A & -I \end{bmatrix}^\top \\ \begin{bmatrix} A & -I \end{bmatrix} & I \end{bmatrix} \succeq 0. \tag{11}$$

Note that (11) is a subset of (10) and thus need not be added to the constraints.

4 Experimental Result

Here, we demonstrate the performance of distributionally-robust machine learning on two practical datasets. In what follows, we use the mean square error, i.e., $\mathbb{E}^{\mathbb{P}}\{\ell(\mathfrak{M}(x;A,B),y)\} = \mathbb{E}^{\mathbb{P}}\{(Ax+B-y)^2\}$, for training and test. The experiments are performed on a 2019 MacBook Pro with 2.4 GHz Quad-Core Intel Core i5 with 16 GB 2133 MHz LPDDR3 on MATLAB R2020a. In each experiment, we have randomly selected 1000 training datasets and presented the average results along with the variances.

4.1 Loan Dataset

The dataset contains information of roughly 887,000 loans[1]. The inputs contain loan information, e.g., loan size, and borrower information, e.g., age. The

[1] `https://www.kaggle.com/wendykan/lending-club-loan-data`

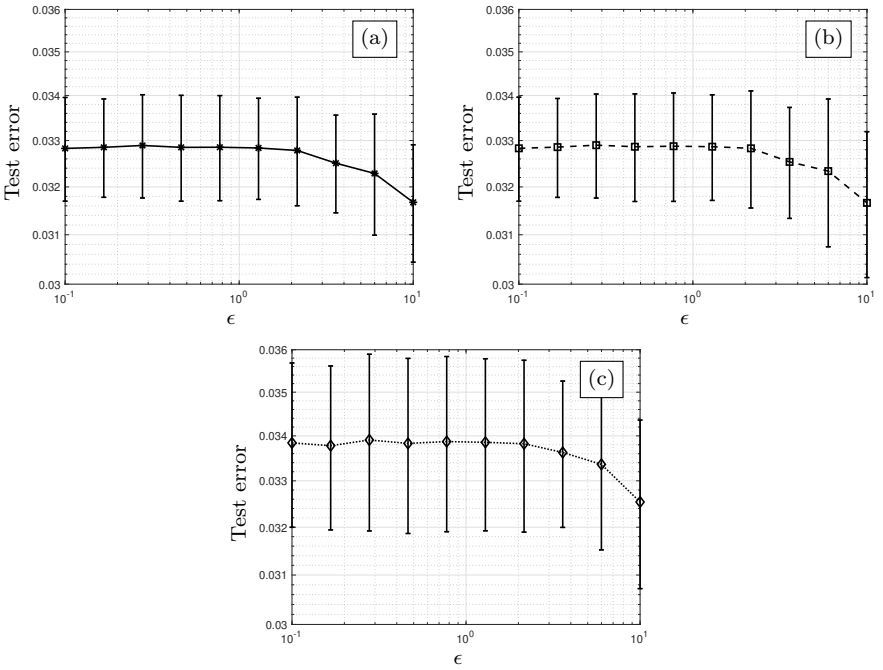

Fig. 1 Performance of the linear regression for the Loan dataset trained on the locally-differential private dataset tested on the original probability distribution using (9) [a], without any regularization [b], and using ℓ_2-norm regularization [c].

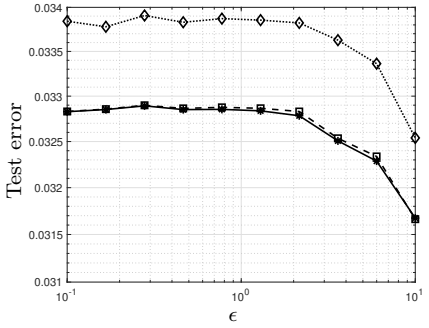

Fig. 2 Performance of the linear regression for the Loan dataset trained on the locally-differential private dataset tested on the original probability distribution using (9) (solid), without any regularization (dashed), and using ℓ_2-norm regularization (dotted).

outputs are the interest rates. Categorical features, e.g., state of residence, are encoded with integer numbers. Unique identifiers, e.g., identity, and irrelevant attributes, e.g., URLs, are removed. We scale all input attributes and outputs to be between zero and one to meet Assumption 2. We consider linear

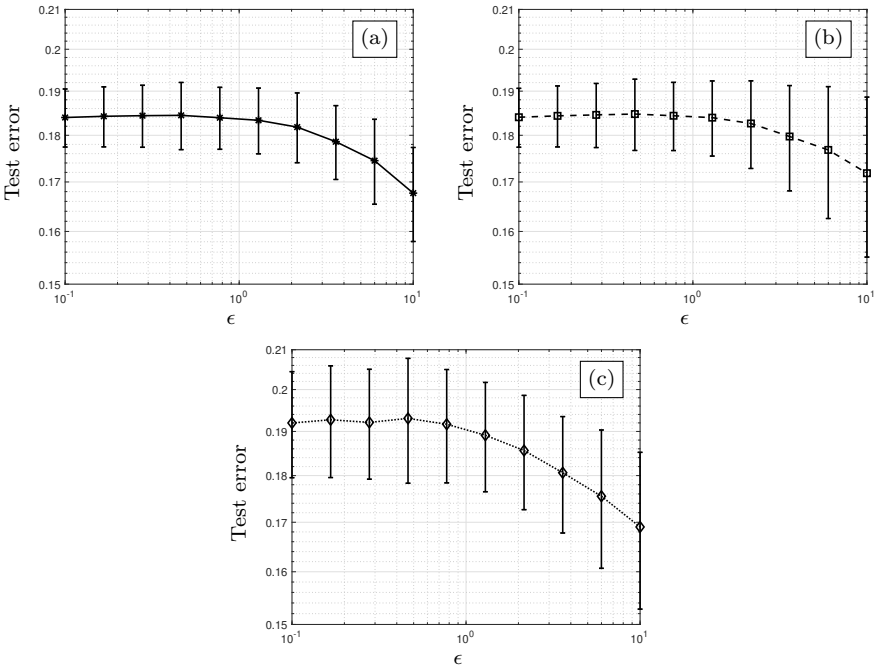

Fig. 3 Performance of the linear regression for the Adult dataset trained on the locally-differential private dataset tested on the original probability distribution using (9) [a], without any regularization [b], and using ℓ_2-norm regularization [c].

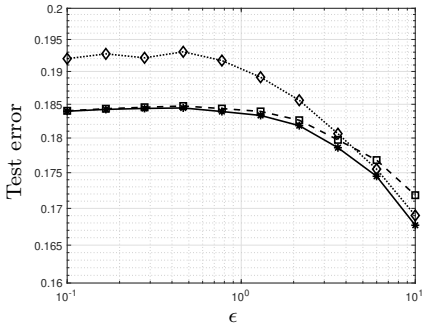

Fig. 4 Performance of the linear regression for the Adult dataset trained on the locally-differential private dataset tested on the original probability distribution using (9) (solid), without any regularization (dashed), and using ℓ_2-norm regularization (dotted).

regression framework in Remark 1 with 2-norm regularization and Section 3 with the novel regularization term. We use the Gaussian mechanism in Theorem 1 to generate locally-differentially private datasets with $\delta = 10^{-2}$ and

$\epsilon \in [10^0, 10^2]/p_x$. We use 50 entries of the dataset for training and the remaining entries for evaluation.

Figure 1 illustrates the performance of the linear regression for the loan dataset trained on the locally-differential private dataset tested on the original probability distribution. Figure 1 (a) shows the performance of the linear regression using the novel regularization term in Section 3, Figure 1 (b) considers linear regression without regularization, and Figure 1 (c) illustrates the performance of the linear regression using the 2-norm regularization in Remark 1 with $\rho = 10^{-2}$. To generate the graphs, we have randomly selected 1000 training datasets and presented the average results along with the variances. In this figure, the curves illustrate the average test error and the vertical lines correspond to standard deviation. To be able to compare these three methods, the average test errors for all three methods are overlaid in the same plot in Figure 2. These plots illustrate that regularization clearly improves the out-of-sample performance of the model. Furthermore, the optimal regularization in Section 3 can be better than the generic regularization based on the Lipschitz constant of the loss function in (6). Note that, in the generic regularization based on the Lipschitz constant of the loss function in (6) and Remark 1, we are optimizing an upper bound of the distributionally-robust optimization in (5). This can result in a more conservative approach. Also note that, intuitively, as ϵ gets smaller, the amount of the additive noise gets larger and the fitness of all the methods degrades similarly (due to the large magnitude of the noise).

4.2 Adult Dataset

Here, we demonstrate the performance of distributionally-robust machine learning on the Adult Dataset[2] containing nearly 49,000 records with features, e.g., age and education, and a binary output indicating whether an individuals earns more than $50,000. We scale all input attributes and outputs to be between zero and one in line with Assumption 2. We consider the linear regression framework in Remark 1 with 2-norm regularization and Section 3 with the novel regularization term. We use the Gaussian mechanism in Theorem 1 to generate locally-differentially private datasets with $\delta = 10^{-2}$ and $\epsilon \in [10^0, 10^2]/p_x$. We use 50 entries of the dataset for training and the remaining entries for evaluation.

Figure 3 illustrates the performance of the linear regression for the loan dataset trained on the locally-differential private dataset tested on the original probability distribution. Figure 3 (a) shows the performance of the linear regression using the novel regularization term in Section 3, Figure 3 (b) considers linear regression without regularization, and Figure 3 (c) illustrates the performance of the linear regression using the 2-norm regularization in Remark 1 with $\rho = 10^{-2}$. Similarly, the average test errors for all three methods

[2] `https://archive.ics.uci.edu/ml/datasets/adult`

are overlaid in the same plot in Figure 4. Similar to the case of the loan dataset, the optimal regularization in Section 3 can be better than the generic regularization based on the Lipschitz constant of the loss function in (6).

5 Conclusions

We considered machine learning, particularly regression, using private datasets. We posed machine learning with private datasets as a distributionally-robust optimization with an ambiguity set parameterized by the Wasserstein distance. For general distributions, the distributionally-robust optimization problem was relaxed as a regularized machine learning problem with the Lipschitz constant of the machine learning model as a regularizer. For Gaussian data, the distributionally-robust optimization problem was solved exactly to find an optimal regularizer.

References

1. C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography* (S. Halevi and T. Rabin, eds.), (Berlin, Heidelberg), pp. 265–284, Springer Berlin Heidelberg, 2006.
2. C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.
3. P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 492–542, 2016.
4. R. Dewri, "Local differential perturbations: Location privacy under approximate knowledge attackers," *IEEE Transactions on Mobile Computing*, vol. 12, no. 12, pp. 2360–2372, 2013.
5. J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438, 2013.
6. X. Ren, C.-M. Yu, W. Yu, S. Yang, X. Yang, J. A. McCann, and S. Y. Philip, "LoPub: High-dimensional crowdsourced data publication with local differential privacy," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 9, pp. 2151–2166, 2018.
7. Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," in *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 1054–1067, 2014.
8. J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang, "Privacy loss in Apple's implementation of differential privacy on MacOS 10.12," *arXiv preprint arXiv:1709.02753*, 2017.
9. A. Smith, A. Thakurta, and J. Upadhyay, "Is interaction necessary for distributed private learning?," in *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 58–77, IEEE, 2017.
10. D. Wang, M. Gaboardi, and J. Xu, "Empirical risk minimization in non-interactive local differential privacy revisited," in *Advances in Neural Information Processing Systems*, pp. 965–974, 2018.
11. K. Zheng, W. Mou, and L. Wang, "Collect at once, use effectively: Making non-interactive locally private learning possible," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 4130–4139, 2017.
12. D. Wang, A. Smith, and J. Xu, "Noninteractive locally private learning of linear models via polynomial approximations," in *Algorithmic Learning Theory*, pp. 898–903, 2019.

13. P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, no. 1-2, pp. 115–166, 2018.
14. V. A. Nguyen, D. Kuhn, and P. M. Esfahani, "Distributionally robust inverse covariance estimation: The wasserstein shrinkage estimator," *arXiv preprint arXiv:1805.07194*, 2018.
15. A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust optimization*, vol. 28. Princeton University Press, 2009.
16. K. Postek, D. den Hertog, and B. Melenberg, "Computationally tractable counterparts of distributionally robust constraints on risk measures," *SIAM Review*, vol. 58, no. 4, pp. 603–650, 2016.
17. E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Operations research*, vol. 58, no. 3, pp. 595–612, 2010.
18. Z. Hu and L. J. Hong, "Kullback-Leibler divergence constrained distributionally robust optimization," *Available at Optimization Online*, 2013.
19. A. Sinha, H. Namkoong, and J. Duchi, "Certifiable distributional robustness with principled adversarial training," in *Proceedings of the Machine Learning and Computer Security Workshop (co-located with Conference on Neural Information Processing Systems 2017)*, vol. 2, 2017.
20. R. Chen and I. C. Paschalidis, "A distributionally robust optimization approach for outlier detection," in *2018 IEEE Conference on Decision and Control (CDC)*, pp. 352–357, IEEE, 2018.
21. C. Lee and S. Mehrotra, "A distributionally-robust approach for finding support vector machines," *Manuscript, available at http://www. optimization-online. org/DB_HTML/2015/06/4965. html*, 2015.
22. S. Shafieezadeh Abadeh, P. M. Mohajerin Esfahani, and D. Kuhn, "Distributionally robust logistic regression," *Advances in Neural Information Processing Systems*, vol. 28, pp. 1576–1584, 2015.
23. D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, "Wasserstein distributionally robust optimization: Theory and applications in machine learning," in *Operations Research & Management Science in the Age of Analytics*, pp. 130–166, INFORMS, 2019.
24. M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 2009.
25. C. Brownlees, E. Joly, G. Lugosi, *et al.*, "Empirical risk minimization for heavy-tailed losses," *The Annals of Statistics*, vol. 43, no. 6, pp. 2507–2536, 2015.
26. F. Farokhi, "Deconvoluting kernel density estimation and regression for locally differentially private data," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.
27. P. J. Bickel, D. A. Freedman, *et al.*, "Some asymptotic theory for the bootstrap," *The annals of statistics*, vol. 9, no. 6, pp. 1196–1217, 1981.
28. T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 2012.
29. L. V. Kantorovich and G. Rubinshtein, "On a space of totally additive functions," *Vestn. Lening. Univ*, vol. 13, pp. 52–59, 1958.
30. F. Farokhi, "Why does regularization help with mitigating poisoning attacks?," *Neural Processing Letters*, 2021.
31. R. Hall, A. Rinaldo, and L. Wasserman, "Random differential privacy," *Journal of Privacy and Confidentiality*, vol. 4, no. 2, pp. 43–59, 2012.
32. T. Rippl, A. Munk, and A. Sturm, "Limit laws of the empirical wasserstein distance: Gaussian distributions," *Journal of Multivariate Analysis*, vol. 151, pp. 90–109, 2016.
33. C. R. Givens and R. M. Shortt, "A class of wasserstein metrics for probability distributions.," *The Michigan Mathematical Journal*, vol. 31, no. 2, pp. 231–240, 1984.
34. L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM review*, vol. 38, no. 1, pp. 49–95, 1996.
35. F. Zhang, *The Schur complement and its applications*, vol. 4. Springer Science & Business Media, 2006.

Author/s:
Farokhi, F

Title:
Distributionally-robust machine learning using locally differentially-private data

Date:
2021-06-10

Citation:
Farokhi, F. (2021). Distributionally-robust machine learning using locally differentially-private data. OPTIMIZATION LETTERS, 16 (4), pp.1167-1179. https://doi.org/10.1007/s11590-021-01765-6.

Persistent Link:
http://hdl.handle.net/11343/276524