Statistica Sinica Preprint No: SS-2022-0282	
Title	Differentially Private Regularized Stochastic Convex Optimization with Heavy-Tailed Data
Manuscript ID	SS-2022-0282
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0282
Complete List of Authors	Haihan Xie, Matthew Pietrosanu, Yi Liu, Wei Tu, Bei Jiang and Linglong Kong
Corresponding Authors	Linglong Kong
E-mails	lkong@ualberta.ca

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

Differentially Private Regularized Stochastic Convex Optimization with Heavy-Tailed Data

Haihan Xie[1] Matthew Pietrosanu[1] Yi Liu[1]

Wei Tu[2] Bei Jiang[1] Linglong Kong[1]

[1] *Department of Mathematical and Statistical Sciences, University of Alberta*

[2] *Department of Public Health Sciences, Queen's University*

Abstract: Existing privacy guarantees for convex optimization algorithms do not apply to heavy-tailed data with regularized estimation. This is a notable gap in the differential privacy (DP) literature, given the broad prevalence of non-Gaussian data and penalized optimization problems. In this work, we propose three (ϵ, δ)-DP methods for regularized convex optimization and derive bounds on their population excess risks in a framework that accommodates heavy-tailed data with fewer assumptions (relative to previous works). This work is the first to address DP in generic regularized convex optimization problems with heavy-tailed responses. Two of our methods augment a basic (ϵ, δ)-DP algorithm with robust procedures for privately estimating minibatch gradients. Our numerical analyses highlight the performance of our methods relative to data dimensionality, batch size, and privacy budget, and suggest settings where each approach is favorable.

Key words and phrases: Privacy protection, randomized mechanism, non-smooth regularization, error bound.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

1. Introduction

Progress towards data privacy, especially in the analysis of sensitive financial or medical data (**???**), is undermined by the ubiquity of convex optimization. As general approaches to convex problems offer no inherent privacy guarantees, the development of privacy-aware optimization techniques has the potential to impact a broad range of analytic methods and applications. Differential privacy (**?**) offers a framework to objectively assess privacy guarantees and can be easily incorporated into analytic techniques and algorithms. Differential privacy has consequently grown in popularity in fields where privacy protection is a concern (**?**).

While privacy-preserving convex optimization algorithms have received substantial attention in the literature, most focus has been placed on stochastic gradient descent (SGD) due to its simplicity and computational efficiency (**?**Abadi et al., 2016; **?**; **?**). However, the convergence rates of these differentially private (DP) versions of SGD suffer in most regularized optimization problems, particularly when working with non-smooth regularizers, such as the widely used ℓ_1 penalty for managing sparse, high-dimensional settings (**?**). Recently, DP variants of mirror descent and Frank–Wolfe algorithms have been developed to address this problem (**????**). Additionally, **?** proposed a modification to the exponential mechanism, which allows for

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

the optimal empirical and population risk in private solving of non-smooth objective functions.

Despite these and other advances in DP convex optimization, the current literature relies on one critical assumption to achieve differential privacy: Lipschitz continuity of the loss function. The consequent assumption of a bounded gradient is not valid for many real-world datasets. Settings with heavy-tailed data, as a focal point for the present work, can hardly be called fringe: file sizes, the flood levels of rivers, and major insurance claim amounts, among many other examples, fall into this category. Consider the squared loss function $\ell(\beta, x, y) = (y - x^\top \beta)^2$ as an example, which has an unbounded gradient with respect to β when either covariates x or the response variable y has heavy-tailed distributions. As a result, the privacy guarantees offered by the aforementioned methods may fail to hold in this setting.

More recent works noting this issue are discussed in Section 2 (**???**). In summary, researchers are recognizing the limitations of assuming Lipschitz continuity in the loss function and are developing privacy algorithms to address these limitations for SCO problems. Nevertheless, there has been insufficient attention paid to regularized convex optimization problems in this direction, such as lasso regression, group lasso, and regularized logistic

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

regression, despite their widespread use in statistics and machine learning. Regularization techniques can provide various critical functions such as shrinking estimates, performing variable selection, and reducing model overfitting. The current gap in the literature is surprising, given the ubiquity of regularized problems and also heavy-tailed data.

Throughout this work, we consider a dataset $D = \{z_1, \cdots, z_n\}$ of $z_i = (x_i, y_i) \in \mathcal{Z}$ which independently and identically follow some distribution $\mathcal{D}$. Here, $x_i \in \mathbb{R}^p$ is a feature vector, and $y_i \in \mathbb{R}$ is a heavy-tailed response. Our goal is to minimize a population risk of the form $F_{\mathcal{D}}(\beta) = L_{\mathcal{D}}(\beta) + g(\beta)$ under this setting, namely,

$$\min_{\beta \in \mathbb{R}^p} \big\{ L_{\mathcal{D}}(\beta) + g(\beta) \big\}, \tag{1.1}$$

where $L_{\mathcal{D}}(\beta) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(\beta, z)]$ for some smooth, convex loss function ℓ and where g is a relatively simple, convex function that may be non-differentiable. Unless otherwise specified, we let $\|\cdot\|$ denote the usual ℓ_2 norm on Euclidean space. Let $\nabla L_{\mathcal{D}}$ and $\nabla \ell$ denote the gradient of $L_{\mathcal{D}}(\beta)$ and $\ell(\beta)$, respectively.

In this work, we focus on developing algorithms that can privately, efficiently, and robustly handle regularized stochastic convex optimization (SCO) problems with heavy-tailed responses. It is worth noting that the term "robust" is used to signify our objective of achieving rigorous theoretical derivation and stable numerical outcomes, even in the presence of heavy-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

tailed responses. In this situation, the utilization of first-order algorithms may lead to inaccuracies due to the existence of anomalous gradients. To overcome this problem, we employ robust mean estimators proposed in the literature, which help to provide reliable estimates of gradients. Specifically, our contributions in this paper are three-fold.

(1) We first propose a vanilla (ϵ, δ)-DP algorithm (Algorithm 1) for the regularized SCO problem in (1.1). This approach requires minimal assumptions and is applicable to a wide range of data distributions. In this setting, we establish an upper bound on the expected population excess risk that generally depends on a gradient-clipping mechanism.

(2) Next, we develop an (ϵ, δ)-DP algorithm (Algorithm 2) that utilizes the mean estimator in Kamath et al. (2020) to robustly estimate $\nabla L_{\mathcal{D}}$. We show that, with high probability and for some $k \geq 2$, the output of Algorithm 2 achieves an upper bound of $\tilde{O}[p^{3/2}\{\sqrt{pT}/(n\epsilon)\}^{(k-1)/k} + T^{-1/2}]$ on the population excess risk, where T denotes the total number of iterations. We introduce one additional assumption to achieve this refinement of Algorithm 1: the mean and the order-k central moment of each coordinate of $\nabla \ell$ are bounded.

(3) We propose a second robust (ϵ, δ)-DP algorithm (Algorithm 4) that

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

instead incorporates the robust mean estimator in Holland (2019). It improves the upper bound of Algorithm 2 to $\tilde{O}(pT^{1/4}/\sqrt{n\epsilon} + T^{-1/2})$ in a high probability setting. Compared with Algorithm 2, Algorithm 4 relies on a relatively weaker assumption that only bounds the second moment of each coordinate of $\nabla\ell$.

Due to the space limit, all the proofs and technical lemmas are relegated to the supplementary material.

2. Related Work

As previously mentioned, there are only a few works that address DP SCO for heavy-tailed data. Wang et al. (2020) proposed Algorithm 4 as the first approach to consider this setting. This algorithm combined projected gradient descent with the robust mean estimator from Holland et al. (2019). Our Algorithm 4 draws inspiration from this approach, and we can improve the algorithm's $\tilde{O}\{p/(\epsilon^2 n)^{1/3}\}$ bound on population excess risk in this work. **?** derived new algorithms for SCO with heavy-tailed distributions under concentrated differential privacy. However, this work relies on the smoothness of the objective function and only provides the bounds on expected population excess risk. **?** was the first to study DP SCO for high-dimensional, heavy-tailed data, but did not do so in full generality.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

The authors considered specific convex problems over polytope (i.e., ℓ_1-norm), sparsity (ℓ_0-norm) constraints under ϵ-DP and (ϵ, δ)-DP frameworks. Since only the methods in **?** can handle non-smooth loss function in the presence of heavy-tailed data, we compare them with our approaches in our numerical study.

3. Preliminaries

3.1 Stochastic proximal gradient descent

Proximal gradient descent is a standard approach to the problem in (1.1) due to the efficiency and flexibility of the proximal operator (**??**). Given an initial point $\beta_0 \in \mathbb{R}^p$ and a sequence $(\gamma_t)_{t\in\mathbb{N}}$ of positive step sizes, the update rule for proximal gradient descent is $\beta_{t+1} = \text{prox}_{\gamma_t, g}\left(\beta_t - \gamma_t \nabla L_{\mathcal{D}}\left(\beta_t\right)\right)$, where

$$\text{prox}_{\gamma, g}(\beta) = \underset{b\in\mathbb{R}^p}{\arg\min}\left\{g(b) + (2\gamma)^{-1}\|b - \beta\|^2\right\}$$

is the proximal map associated with g and the step size γ. If g is a proper, convex, and lower semicontinuous function, then the proximal map is unique for any $\beta \in \mathbb{R}^p$ and any $\gamma > 0$ (**?**).

When n is large, the proximal gradient update requires computing n gradients and may be expensive. Stochastic proximal gradient descent (SPGD), or more accurately, mini-batch SPGD, is a stochastic variant

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

of proximal gradient descent that uses a small, size-m subset B_t of the observations to estimate the gradient. The updating procedure is

$$\beta_{t+1} = \text{prox}_{\gamma_t, g}\left(\beta_t - \gamma_t m^{-1} \sum_{z_i \in B_t} \nabla \ell(\beta_t, z_i)\right).$$

If the subset B_t is sampled uniformly from the full dataset, then the stochastic gradient estimate is unbiased, i.e., $\mathbb{E}[m^{-1} \sum_{z_i \in B_t} \nabla \ell(\beta_t, z_i)] = \nabla L_{\mathcal{D}}(\beta_t)$.

3.2 Differential privacy

Loosely speaking, a privacy-preserving algorithm should give similar outputs when applied to neighboring datasets, i.e., datasets that differ in only one record. It should be difficult to determine the contribution of a single individual to the output. Differential privacy provides a formal mathematical framework that quantifies this notion and controls the amount of privacy protection using a prespecified privacy budget ϵ. However, "pure" ϵ-differential privacy is typically too restrictive and can been relaxed to (ϵ, δ)-differential privacy, a framework that accommodates a constant probability δ of failure.

Definition 1. A randomized mechanism $\mathcal{M}$ is called (ϵ, δ)-differentially private if, for all measurable $\mathcal{O}$ in the output space of $\mathcal{M}$ and for any two adjacent datasets $d, d' \in \mathcal{Z}^n$, $\mathbb{P}[\mathcal{M}(d) \in \mathcal{O}] \leq e^{\epsilon} \mathbb{P}[\mathcal{M}(d') \in \mathcal{O}] + \delta$. We write that $\mathcal{M}$ is (ϵ, δ)-DP.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

Approaches for attaining (ϵ, δ)-differential privacy are well established in the literature (**?**). For a mechanism $f : \mathcal{Z}^n \to \mathbb{R}^p$, the Gaussian mechanism is defined as $\mathcal{M}(d, f, \epsilon, \delta) = f(d) + \left(X_1^*, \ldots, X_p^*\right)^\top$, where the X_i^*s are independently drawn from $\mathcal{N}(0, \sigma^2\Delta_2^2)$. Here, $\Delta_2 = \max_{\{d,d' \in \mathcal{Z}^n\}} \|f(d) - f(d')\|$ (d, d' are two adjacent datasets) is the l_2 sensitivity of f and σ is a variance parameter determined by ϵ and δ. It has been shown that the Gaussian mechanism is (ϵ, δ)-DP.

3.3 Assumptions

To facilitate our theoretical investigations, we consider the following assumptions.

(A1) The function g is non-negative, convex, not identically $+\infty$, and lower semicontinuous. The function ℓ is convex and continuously differentiable over the parameter space.

(A2) The parameter space $\mathcal{B} = \{\beta \in \mathbb{R}^p \mid g(\beta) < \infty\}$ is closed and has a finite diameter $\Delta = \sup\{\|\beta - \beta'\| : \beta, \beta' \in \mathcal{B}\}$.

(A3) The function $L_{\mathcal{D}}$ has a K_1-Lipschitz-continuous gradient on $\mathcal{B}$: for any $\beta, \beta' \in \mathcal{B}$, $\|\nabla L_{\mathcal{D}}(\beta) - \nabla L_{\mathcal{D}}(\beta')\| \leq K_1\|\beta - \beta'\|$.

(A4) The function ℓ has a K_2-Lipschitz-continuous gradient on $\mathcal{B}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

(A5) For any $\beta \in \mathcal{B}$ and each coordinate $j \in [p]$, $\mathbb{E}[|\langle \nabla \ell(\beta, z) - \mu, e_j \rangle|^k] \leq 1$ for some $k \geq 2$, where e_j is the jth standard basis vector and $\mu = \mathbb{E}[\nabla \ell(\beta, z)]$ is assumed to be bounded.

(A6) For any $\beta \in \mathcal{B}$ and each coordinate $j \in [p]$, $\mathbb{E}[\{\nabla_j \ell(\beta, z)\}^2] \leq \tau$, where τ is some known constant.

Assumptions (A1)–(A3) are common in the SPGD literature (Atchadé et al., 2017; **?**). Assumption (A4) pertains to the smoothness of the loss function ℓ, requiring it to be K_2-smooth. This assumption has been taken into consideration in previous literature when dealing with heavy-tailed response data, see **?**Holland (2019); **?**, for example. Assumption (A5) bounds the order-k central moment (for some $k \geq 2$) and the mean of the distribution of $\nabla \ell(\beta, z)$: this is required for the robust estimator in Kamath et al. (2020) which we implement in Algorithm 2. Assumption (A6) is designed for Algorithm 4, which is weaker than Assumption (A5) in the way that it only assumes the gradient of the loss function has a bounded second-order moment.

4. Vanilla DP SPGD

We begin by proposing a basic approach to DP SPGD in Algorithm 1. In the absence of a Lipschitz-continuity assumption on the loss function, we

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

introduce clipping in each iteration to bound the norm of the gradient and achieve a desired level of privacy. Outside of differential privacy, gradient clipping is commonly used to manage exploding gradients and has established practical benefits (**??**). Intuitively, clipping reduces the influence of large gradients created by outlying data. To achieve differential privacy, we apply the Gaussian mechanism after clipping and then perform the proximal update step using the private gradient estimate $\nabla \widehat{L}_{\mathcal{D}}$.

Specifically, the noises introduced in line 6 of Algorithm 1 should be generated from a Gaussian distribution with a variance term of $\sigma^2 C^2 I_p$, where C is a pre-specified clipping threshold, I_p is a $p \times p$ identity matrix, and σ is a noise scale. To ensure Algorithm 1 is (ϵ, δ)-DP, the noise scale needs to satisfy the inequality: $\sigma \geq \sqrt{\log(\delta^{-1})}/(m\epsilon)$. In the algorithm, we set σ equal to the right-hand side of the inequality to meet the minimum requirement on the noise scale. The following proposition establishes a privacy guarantee for the output of Algorithm 1. The proof of this result can be found in the supplementary material.

Proposition 1. *Let $T = O(n^2/m^2)$. There exists a constant c_1 such that Algorithm 1 is (ϵ, δ)-DP for any $0 < \epsilon \leq c_1$ and $0 < \delta < 1$.*

Remark 1. To facilitate practical implementation, Algorithm 1 takes δ as an input, allowing for the adjustment of the noise scale σ based on

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

Algorithm 1 DP SPGD algorithm for solving (1.1)

Input: dataset $D = \{z_i\}_{i=1}^n$, number of iterations T, step-size sequence $(\gamma_t)_{t=0}^{T-1}$, clipping threshold C, batch size m, privacy parameters (ϵ, δ)

initialize β_0

for $t = 0$ to $T-1$ **do**

Sample a batch B_t randomly from D with sampling probability m/n

$\tilde{\ell}_{ti} \leftarrow \nabla \ell(\beta_t, z_i)$ for $i \in B_t$

$\bar{\ell}_{ti} \leftarrow \tilde{\ell}_{ti} / \max\{\|\tilde{\ell}_{ti}\|/C, 1\}$ for $i \in B_t$

$\nabla \widehat{L}_{\mathcal{D}}(\beta_t) \leftarrow m^{-1}(\sum_{i \in B_t} \bar{\ell}_{ti}) + \mathcal{N}(0, \sigma^2 C^2 I_p)$, where $\sigma = \sqrt{\log(\delta^{-1})}/(m\epsilon)$

$\beta_{t+1} \leftarrow \mathrm{prox}_{\gamma_t, g}(\beta_t - \gamma_t \nabla \widehat{L}_{\mathcal{D}}(\beta_t))$

end for

Output: $\bar{\beta}^T = T^{-1} \sum_{t=1}^T \beta_t$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

the desired privacy level. As a result, the privacy guarantee mentioned above holds for all $0 < \delta < 1$. If the noise scale in the algorithm is fixed, the lower bound of δ will depend on ϵ, σ, and m. Specifically, we have $\exp\{-(m\epsilon\sigma)^2\} \leq \delta < 1$.

The theorem below offers a utility guarantee for Algorithm 1 based on the expected population risk, where the randomness of the algorithm is taken into account. According to the result, it is not possible to eliminate the bias introduced by clipping without a more specific clipping mechanism or additional assumptions on the gradients, see **?** for example. Further discussion of this clipping bias is beyond the scope of this paper.

Theorem 1. *Let $\bar{\beta}^T = T^{-1}\sum_{t=1}^{T}\beta_t$ be the output of Algorithm 1. Assume the non-increasing step sizes satisfy $\gamma_t \in (0, K_1^{-1}]$ for all $t \leq T-1$ and let β^* denote an arbitrary minimizer of (1.1). Under Assumptions (A1)–(A3),*

$$\mathbb{E}\left[F_{\mathcal{D}}(\bar{\beta}^T) - F_{\mathcal{D}}(\beta^*)\right] \leq \frac{\Delta^2}{2T\gamma_{T-1}} + T^{-1}\sum_{t=1}^{T}\Delta\mathrm{Bias}_{\|\cdot\|}(\nabla\widehat{L}_{t-1}),$$

where $\mathrm{Bias}_{\|\cdot\|}(\nabla\widehat{L}_{t-1}) = \mathbb{E}[\|\nabla\widehat{L}_{\mathcal{D}}(\beta_{t-1}) - \nabla L_{\mathcal{D}}(\beta_{t-1})\|]$ is the bias of $\nabla\widehat{L}_{\mathcal{D}}(\beta_{t-1})$ with respect to the norm $\|\cdot\|$.

Remark 2. The clipping parameter C in Algorithm 1 plays an important role in practice. It is unclear how to choose a value of C that suitably balances the signal of the clipped gradient and the noise introduced for the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

sake of privacy. **?** showed that standard gradient clipping cannot induce label noise robustness in classification tasks: while the setting of **?** and the present paper differ, this result encourages us to seek better alternatives to simple clipping mechanisms. We explore these alternatives in the following section with two robust mean estimators to better reduce the effect of anomalous gradients caused by heavy-tailed data.

5. Improved DP SPGD via Robust Mean Estimation

5.1 DP SPGD with the Kamath estimator

Our first improved approach is motivated by the robust mean estimator in Kamath et al. (2020). The core idea of this estimator is illustrated in lines 1–9 of Algorithm 3 and formalized in Lemma 2 in the supplementary material. Informally speaking, if the gradient distribution is truncated within a large interval that is centered close to its mean, then the mean of the distribution will not change substantially with truncation. In DP SPGD settings, this result can guarantee the theoretical performance of the clipping mechanism (e.g., in Algorithm 1). Consequently, we adopt the Kamath estimator to obtain a private and robust approximation of the gradient in Algorithm 2. The DP robust mean estimation procedure DPRME on line 4 is provided in Algorithm 3.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

Algorithm 2 DP SPGD (Kamath) for solving (1.1)

Input: dataset $D = \{z_i\}_{i=1}^n$, number of iterations T, step-size sequence $(\gamma_t)_{t=0}^{T-1}$, batch size m, failure probability ζ, scale parameter ϱ, privacy parameters (ϵ, δ)

initialize β_0

for $t = 0$ to $T-1$ **do**

  Sample a batch B_t randomly from D with sampling probability m/n

  $\nabla \widehat{L}_{\mathcal{D}}(\beta_t) \leftarrow \mathrm{DPRME}(\{\nabla \ell(\beta_t, z_i)\}_{i \in B_t}, \zeta, \varrho, \epsilon, \delta)$

  $\beta_{t+1} \leftarrow \mathrm{prox}_{\gamma_t, g}(\beta_t - \gamma_t \nabla \widehat{L}_{\mathcal{D}}(\beta_t))$

end for

Output: $\bar{\beta}^T = T^{-1} \sum_{t=1}^T \beta_t$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

The following theorem establishes the accuracy of the DP robust mean estimator in Algorithm 3.

Theorem 2. *Let $\mathcal{P}$ be a distribution over $\mathbb{R}^p$ with bounded mean μ, and let $X = \{x_1, \ldots, x_n\}$ be independent and identically distributed samples from $\mathcal{P}$. Assume that $\mathbb{E}[|\langle X - \mu, e_j\rangle|^k] \le 1$ for some $k \ge 2$ and all $j \in [p]$, where e_j is the jth standard basis vector. If ϱ in Algorithm 3 is set to be $\varrho \ge 4\|\mu\|_\infty$, with probability at least $1 - \zeta$, the output $\hat{\mu}$ satisfies*

$$\|\hat{\mu} - \mu\| \le O\Big[\frac{\varrho \log(p\zeta^{-1})\sqrt{p\log(\delta^{-1})}}{n\epsilon}\big\{\sqrt{p} + \sqrt{\log(\zeta^{-1})}\big\} + \sqrt{p}\Big\{\sqrt{\frac{\log(p\zeta^{-1})}{n}} + \Big(\frac{4}{\varrho}\Big)^{k-1}\Big\}\Big].$$

We defer the proof of Theorem 2 to Section S2.2 in the supplementary material, whose spirit is similar to that of Theorem 4.1 in **?**. The difference, however, is that some parameters (e.g., the number of splits q and the truncation interval) are modified for better experimental performance. Furthermore, we carefully control the Gaussian noise added for DP guarantee in order to apply the moments-accountant technique in Abadi et al. (2016), which leads to a stronger composition of DP than the basic composition used in **?**.

Note that due to the sampling procedure in Algorithm 2, batches are not independent across iterations. An uniform bound on the gradient estimation

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

Algorithm 3 DP robust mean estimator (DPRME)

Input: samples $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^p$, failure probability ζ, scale parameter ϱ, privacy parameters (ϵ, δ)

$q \leftarrow 3\log(2p/\zeta)$

Set the truncation interval: $I \leftarrow [-\varrho/2, \varrho/2]$

for $j = 1$ to p **do**

  for $i = 1$ to q **do**

    Define the truncated set $Z_j^i \leftarrow \{\text{clip}(x, I) : x \in (X_{(i-1)(n/q)+1}(j), \ldots, X_{in/q}(j))\}$

    $\hat{\mu}_j^i \leftarrow (q/n)\sum_{x \in Z_j^i} x$

  end for

  $\hat{\mu}_j \leftarrow \text{Median}\{\hat{\mu}_j^1, \ldots, \hat{\mu}_j^q\}$

end for

$\hat{\mu} \leftarrow (\hat{\mu}_1, \ldots, \hat{\mu}_p) + \mathcal{N}(0, \sigma^2 I_p)$, where $\sigma = \varrho q\sqrt{p\log(\delta^{-1})}/(n\epsilon)$

Output: $\hat{\mu}$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

error $\|\nabla\widehat{L}_{\mathcal{D}}(\beta_t) - \nabla L_{\mathcal{D}}(\beta)\|$ over $\beta \in \mathcal{B}$ is required. Lemma 3 provides the uniform bound, which is presented in the supplementary material. This bound is used in Corollary 1 to achieve a high probability bound on the population excess risk. Additionally, the corollary below provides the DP guarantee for Algorithm 2.

Corollary 1. *Set $T = O(n^2/m^2)$. Algorithm 2 achieves (ϵ, δ)-DP for any $0 < \epsilon \leq c_2$ and $0 < \delta < 1$, where c_2 is a constant. Under Assumptions (A1)–(A5), If $\varrho = \{\epsilon m/\sqrt{p\log(\delta^{-1})}\}^{1/k} > 4\|\mu\|_\infty$ and $\gamma_t = K_1/\sqrt{t}$ for $t \leq T-1$, then with probability at least $1-\zeta$, Algorithm 2 satisfies*

$$F_{\mathcal{D}}(\bar{\beta}^T) - F_{\mathcal{D}}(\beta^*) \leq \tilde{O}\Big[T^{-\frac{1}{2}} + p^{\frac{3}{2}}\log(\zeta^{-1})\big\{\sqrt{p\log(\delta^{-1})T}/n\epsilon\big\}^{\frac{k-1}{k}}\Big]$$

for any failure probability ζ, where the notation $\tilde{O}$ omits some logarithmic factors and the term of Δ, K_1.

To provide intuition, we can express the bound from Corollary 1 as:

$$F_{\mathcal{D}}(\bar{\beta}^T) - F_{\mathcal{D}}(\beta^*) \leq \tilde{O}\big(T^{-\frac{1}{2}} + p^{3/2}\sigma\big),$$

where σ is defined in Algorithm 3. It is well-known that the lower bound for convergence rate in non-smooth stochastic convex optimization is $\Omega(1/\sqrt{T})$ (**?**). Therefore, our upper bound can be decomposed into two parts: the non-private convergence rate that cannot be improved for non-smooth stochastic

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

convex optimization and the additional excess population error caused by privacy. It is worth noting that the non-private rate is suboptimal when the regularization term is smooth or absent. Even though the regularization term is allowed to be smooth, the literature on SPGD primarily focuses on the use of non-differentiable regularization functions. For instance, **?** and **?** have established the $O(1/\sqrt{T})$ convergence rate for solving (1.1) via SPGD. In this work, we mainly address the challenge of handling non-smooth regularization terms using the proposed approach. Deriving an optimal non-private rate that covers both smooth and non-smooth cases is left as future work. More discussions are given in Section 7.

If we ignore the effect of polylog factors (which is common in the DP SCO literature, such as in **?**, **?**, and **?**), we see that with high probability guarantee, the upper bound of the excess population risk is $\tilde{O}\big[p^{3/2}\{\sqrt{pT}/(n\epsilon)\}^{(k-1)/k} + T^{-1/2}\big]$ for some $k \geq 2$. The selection of batch size m is subject to a trade-off: a smaller value of m facilitates faster convergence of the non-private part $\tilde{O}(T^{-1/2})$, but results in larger levels of noise added to ensure privacy. Optimizing the upper bound in m yields a bound of $\tilde{O}\{p^{(4k-1)/2}/(\epsilon^{k-1}n^{k-1})\}^{1/(2k-1)}$, where $m = (n^k p^{(4k-1)/2}/\epsilon^{k-1})^{1/(2k-1)}$. The bound may vary depending on the order of the bounded moment of $\nabla\ell$. For example, when $k = 2$, our bound takes the form of $\tilde{O}\{p^{7/6}/(n\epsilon)^{1/3}\}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

Moreover, as k approaches infinity, the bound converges to $\tilde{O}(p/\sqrt{n\epsilon})$. However, to achieve this optimal rate, a restrictive constraint on the ratio of dimensionality p and sample size n is required, namely, $p \leq (n\epsilon)^{(2k-2)/(4k-1)}$. This limitation significantly reduces the practicality of the approach. We further investigate the impact of m on the convergence results through simulation studies, as discussed in Section 6.1.

Remark 3. If we simply set $m = n^{1/2}$, the upper bound from Corollary 1 is given by $p^2/(\sqrt{n}\epsilon)(\sqrt{n}\epsilon/\sqrt{p})^{\frac{1}{k}}$, which is looser compared with the bound $p/\sqrt{n} + p^2/(n\epsilon)(n\epsilon/p^{3/2})^{\frac{1}{k}}$ in **?**. Nevertheless, **?** requires the smoothness assumption on the overall loss function and their bound is for expected population excess risk. For heavy-tailed data, an expected population excess risk can not be directly transformed to a useful population excess risk with high probability, whereas the latter one is more commonly used in the literature on robust statistics (**???**).

5.2 DP SPGD with the Holland estimator

As can be seen in Corollary 1, the upper bound of the excess population risk is slightly high, which impels us to explore other potential methods. Holland (2019) provides another approach to robust mean estimation. As noted in Section 2, this estimator has been previously implemented to solve DP SCO

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

problems for heavy-tailed data (**?**). To keep the present work self-contained, we briefly review the Holland estimator. We start with a one-dimensional example. Let $x_1, \ldots, x_n$ be an i.i.d. realizations of a random variable X. Our goal is to robustly estimate $\mathbb{E}[X]$. The estimation procedure in Holland (2019) consists of three steps.

Firstly, define one step as "scaling and truncation". In this step, we scale each sample x_i by a scale parameter s (to be specified later), truncate the scaled samples with a soft truncation function ϕ, and transform the truncated arithmetic mean back to the original scale:

$$\mathbb{E}[X] \approx \frac{s}{n} \sum_{i=1}^{n} \phi\left(\frac{x_i}{s}\right),$$

where

$$\phi(x) = \begin{cases} x - \frac{x^3}{6}, & -\sqrt{2} \le x \le \sqrt{2} \\ \frac{2\sqrt{2}}{3}, & x > \sqrt{2} \\ -\frac{2\sqrt{2}}{3}, & x < -\sqrt{2} \end{cases} \tag{5.2}$$

is as given in Catoni and Giulini (2017). Then, let $\eta_1, \ldots, \eta_n$ be i.i.d. random noise from a zero-mean distribution χ. Multiply each sample x_i by the random noise $1 + \eta_i$ and apply the "scaling and truncation" step to the result. That is,

$$\tilde{x}(\eta) = \frac{s}{n} \sum_{i=1}^{n} \phi\left(\frac{x_i + \eta_i x_i}{s}\right).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

Lastly, the multiplicative noise is smoothed out by taking the expectation with respect to the noise distribution χ. This leads to the final robust estimator

$$\hat{x} = \frac{s}{n} \sum_{i=1}^{n} \int \phi\left(\frac{x_i + \eta_i x_i}{s}\right) d\chi\left(\eta_i\right). \tag{5.3}$$

The computation of each integral in (5.3) relys on the noise distribution χ and the truncation function ϕ. Luckily, Catoni and Giulini (2017) provides an efficient way to evaluate the integral. If $\chi \sim \mathcal{N}(0, 1/\nu)$ (with ν to be specified later), then for any a and b, $\mathbb{E}_\eta[\phi(a + b\sqrt{\nu}\eta)] = a\left(1 - b^2/2\right) - a^3/6 + C(a, b)$, where $C(a, b)$, whose explicit form is given in Section S2.4 in the supplementary material, is a correction factor that is easy to implement. Due to the nature of the truncation function ϕ, the integral in the final step is bounded by $2\sqrt{2}/3$, so the ℓ_2-norm sensitivity of the estimator $\hat{x}$ is $4\sqrt{2}s/(3n)$. Therefore, a mechanism to achieve (ϵ, δ)-DP is as follows:

$$\mathcal{A}(D) = \hat{x} + \mathcal{N}\left(0, 32s^2 \log(\delta^{-1})/(9n^2\epsilon^2)\right), \tag{5.4}$$

whose error bound is given in Lemma 4 (See Section S2.3 in the supplementary material).

So far we have only considered robust mean estimation in one dimension. We generalize this procedure to arbitrary dimensions in Algorithm 4: in each iteration, we use a size-m minibatch to estimate each coordinate of

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

Algorithm 4 DP SPGD (Holland) algorithm for solving (1.1)

Input: dataset $D = \{z_i\}_{i=1}^n$, number of iterations T, step-size sequence $(\gamma_t)_{t=0}^{T-1}$, batch size m, failure probability ζ, gradient second-moment bound τ, privacy parameters (ϵ, δ)

initialize β_0

$\nu \leftarrow \sqrt{\log(\zeta^{-1})}$

$s \leftarrow \sqrt{m\epsilon\tau}/\{\log(\zeta^{-1})\log^{1/4}(\delta^{-1})\}$

for $t = 0$ to $T-1$ **do**

Sample a batch B_t randomly from D with sampling probability m/n

$\tilde{\ell}_{ti} \leftarrow \nabla\ell(\beta_t, z_i)$ for $i \in B_t$

Calculate the robust gradient:

$$\tilde{l}_t \leftarrow \frac{1}{m}\sum_{i\in B_t}\Big\{\tilde{l}_{ti}\Big(1-\frac{\tilde{l}_{ti}^2}{2s^2\nu}\Big)-\frac{\tilde{l}_{ti}^3}{6s^2}\Big\}+\frac{s}{m}\sum_{i\in B_t}C\Big(\frac{\tilde{l}_{ti}}{s},\frac{|\tilde{l}_{ti}|}{s\sqrt{\nu}}\Big)$$

$\nabla\widehat{L}_{\mathcal{D}}(\beta_t) \leftarrow \tilde{l}_t + \mathcal{N}(0, \sigma^2 I_p)$, where $\sigma = 4s\sqrt{2p\log(\delta^{-1})}/(3m\epsilon)$

$\beta_{t+1} \leftarrow \mathrm{prox}_{\gamma_t, g}(\beta_t - \gamma_t\nabla\widehat{L}_{\mathcal{D}}(\beta_t))$

end for

Output: $\bar{\beta}^T = T^{-1}\sum_{t=1}^T \beta_t$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

the gradient. Operations on line 7 of Algorithm 4 are made elementwise. To privatize the robust mean estimation procedure, we add Gaussian noise to the robust gradient, whose ℓ_2-norm sensitivity is $4s\sqrt{2p}/3m$. Theorem 3 uniformly bounds the approximation error of $\nabla\widehat{L}_{\mathcal{D}}$ with high probability. Corollary 2 uses this bound to establish a $(1-\zeta)$-probability upper bound on the population excess risk, where ζ is some pre-determined failure probability.

Theorem 3. *Under Assumptions (A2)–(A4) and (A6), with probability at least* $1-\zeta$ *and for any* $\beta \in \mathcal{B}$*, the gradient estimator* $\nabla\widehat{L}_{\mathcal{D}}(\beta)$ *in Algorithm 4 satisfies*

$$\|\nabla\widehat{L}_{\mathcal{D}}(\beta)-\nabla L_{\mathcal{D}}(\beta)\| \leq O\Big[\sqrt{p\tau\log^{\frac{1}{2}}(\delta^{-1})\{\log(p\zeta^{-1})+p\log(\Delta\sqrt{m})\}/(m\epsilon)}\Big],$$

where the batch size $m = O(n/\sqrt{T})$.

Corollary 2. *Set* $T = O(n^2/m^2)$*. Algorithm 4 achieves* (ϵ,δ)*-DP for any* $0 < \epsilon \leq c_3$ *and* $0 < \delta < 1$*, where* c_3 *is a constant. Let* $\gamma_t = K_1/\sqrt{t}$ *for* $t \leq T-1$*. Then for any given failure probability* ζ*, under Assumptions (A1)–(A4) and (A6), Algorithm 4 satisfies*

$$F_{\mathcal{D}}(\bar{\beta}^T) - F_{\mathcal{D}}(\beta^*) \leq \tilde{O}\big\{T^{-\frac{1}{2}} + p\log^{\frac{1}{4}}(\delta^{-1})\log^{\frac{1}{2}}(\zeta^{-1})T^{\frac{1}{4}}/\sqrt{n\epsilon}\big\}$$

with probability at least $1-\zeta$*. The notation* $\tilde{O}$ *omits some logarithmic terms and those that depend only on* Δ*,* τ*, and* K_1*.*

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

Remark 4. It is notable that with a weaker assumption, Corollary 2 improves the upper bound of Corollary 1 to $\tilde{O}(pT^{1/4}/\sqrt{n\epsilon}+T^{-1/2})$. By setting $m=(np)^{2/3}/\epsilon^{1/3}$, we obtain an optimal bound of the form $\tilde{O}\{p^{2/3}/(n\epsilon)^{1/3}\}$, which outperforms the bound of $\tilde{O}\{p^3/(\epsilon^2 n)\}^{1/3}$ in **?** for general smooth convex loss functions. Note that both works have the same requirement on p to achieve optimal rates, specifically $p \leq (n\epsilon)^{1/2}$. Moreover, even if $p > (n\epsilon)^{1/2}$, selecting a batch size close to the sample size (e.g., $m = n^{2/3}$) still results in a superior upper bound when $0 < \epsilon < 1$ compared to the bound in **?**. Our experimental results show that DP SPGD (Holland) algorithm performs better with a relatively large batch size in both regression and classification tasks.

6. Experimental Results

In this section, we examine the performance of the proposed algorithms on synthetic and real-world data. We treat SPGD as the nonprivate benchmark and compare our methods with it under several settings. We write DP-SPGD(K) and DP-SPGD(H) to refer to our methods using the Kamath and Holland estimators, respectively. In addition, we make a comparison with heavy-tailed private LASSO (HTP-LASSO) and heavy-tailed DP Frank Wolfe method (HTDP-FW) proposed in **?** for high dimensional DP SCO with

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

heavy-tailed data. Throughout the following section, results are presented for 20 datasets (for the simulation studies) and across 20 training–test splits (for the real-world analyses).

6.1 Simulation studies

We consider linear and binary logistic regression models with an ℓ_1 penalty. We fix the sample size at $n = 10000$ and consider varying data dimensionalities $p = 20, 40, 60, 80, 100, 150$. We let X denote the $n \times p$ data matrix with each column scaled to a unit norm. The first 10 elements of the effect vector $\beta^* \in \mathbb{R}^p$ are set to be ± 1 while the rest are set to zero. The responses in the linear and logistic models are generated following $Y = X\beta^* + e$ and $Y = \text{sign}([1+\exp\{-(X\beta^*+e)\}]^{-1} - 1/2)$, respectively. The errors e in these models are independently generated from $T(2)$ and the centred $(1/2, 1/2)$ log-logistic distribution.

We set $T = 5n^2/m^2$ as the total number of iterations and $\zeta = 0.1$ as the failure probability in Algorithms 2 and 4. The privacy parameters are set as $\delta = n^{-1}$ and $\epsilon = 0.5, 1, 3$. For comparison, the performance of HTP-LASSO and HTDP-FW are evaluated in the lasso and logistic regression cases, respectively. Parameters in these algorithms (e.g., the total number of iterations T) are set the same way as in **?**. Because of the use of exponential

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

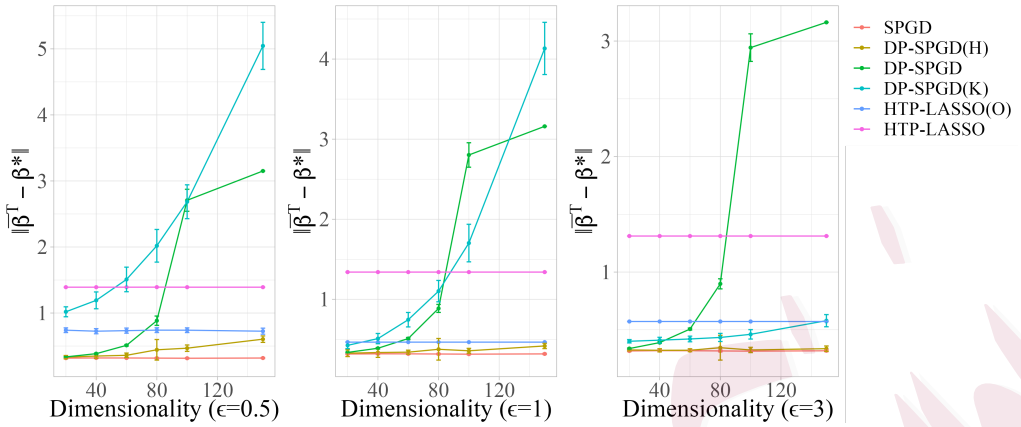

Figure 1: Simulation study results for the lasso model: accuracy vs. data dimensionality p under different privacy budgets ϵ.

mechanism, the performance of HTP-LASSO and HTDP-FW are highly related to the ℓ_1 norm of β^*, which is unknown in practice. Thus, we use (O) to specify algorithms utilizing oracle information (i.e., HTP-LASSO(O) and HTDP-FW(O)). The following numerical experiments examine the influence of batch size m and dimensionality p on algorithm performance, measured as $\|\bar{\beta}^T - \beta^*\|$.

Figures S1 and S2 (See Section S1 in the supplementary material) present results for the lasso and ℓ_1-regularized logistic regression models with respect to batch size: we use these results to determine an optimal batch size for our approaches in each setting. DP-SPGD(H) and DP-SPGD(K) perform better with large batch sizes, especially in the linear case. DP-SPGD typically

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

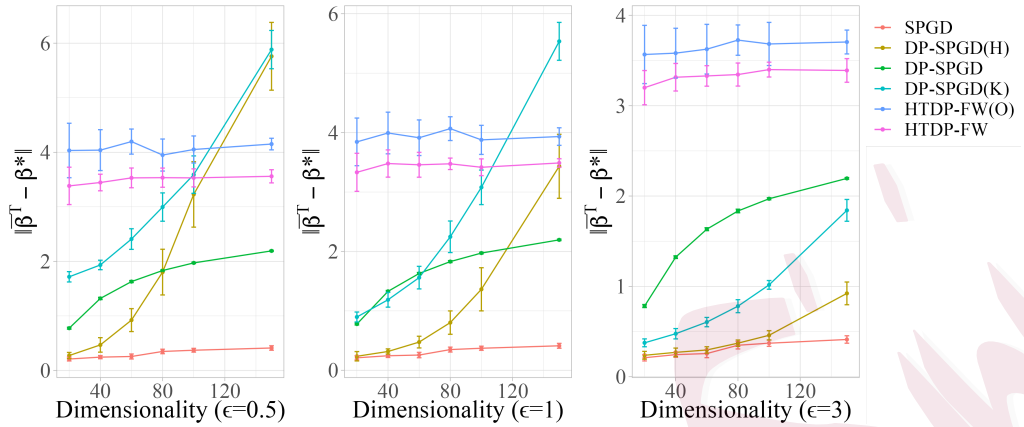

Figure 2: Simulation study results for the ℓ_1-regularized logistic regression model: accuracy vs. data dimensionality p under different privacy budgets ϵ.

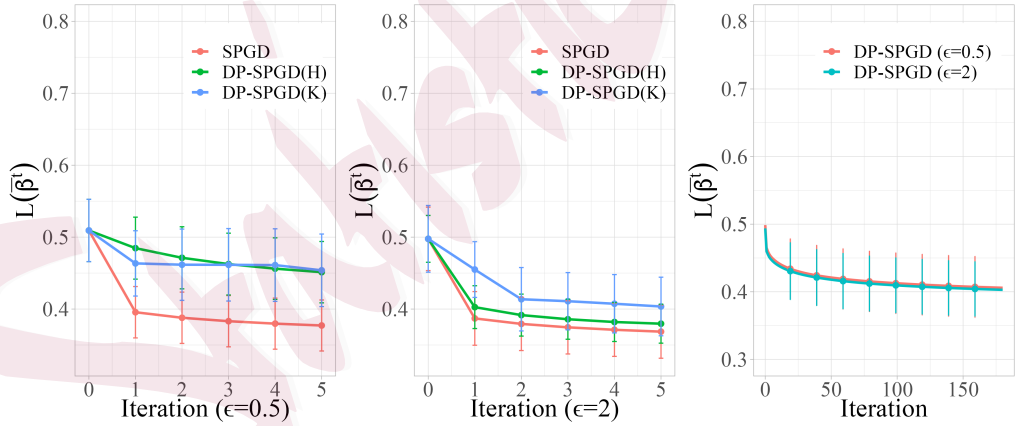

Figure 3: Results for the blog feedback data analysis: empirical risk vs. iteration t under different privacy budgets ϵ.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

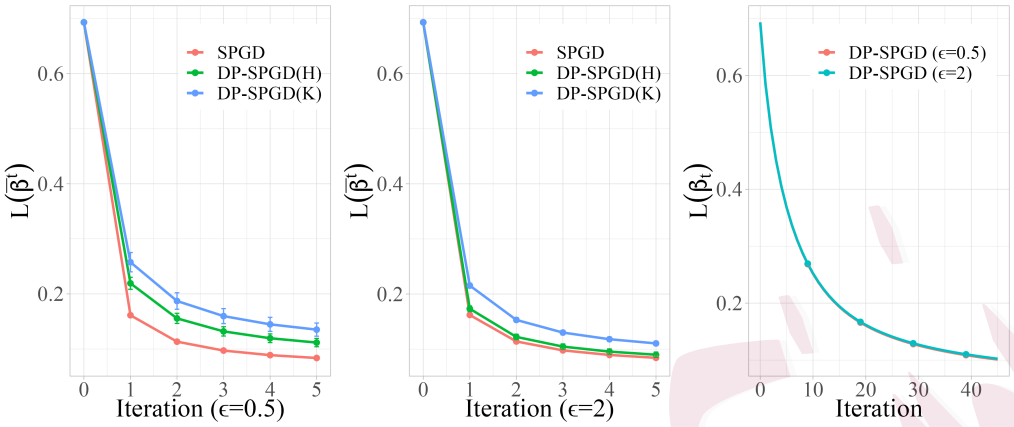

Figure 4: Results for the crop mapping data analysis: empirical risk vs. iteration t under different privacy budgets ϵ. In the right panel, the $\epsilon = 0.5$ and $\epsilon = 2$ series are nearly identical.

requires more iterations (i.e., a larger T and more computation time) to achieve competitive performance.

At these optimal batch sizes, we examine the effect of data dimensionality p and the privacy budget ϵ. These results are shown in Figures 1 and 2. We first compare methods proposed in this paper. DP-SPGD(H) outperforms the others except in the logistic case for a small ϵ. As the privacy budget increases, DP-SPGD performance decreases below that of DP-SPGD(K) and DP-SPGD(H). The performance of DP-SPGD depends primarily on clipping parameter and batch size while DP-SPGD(K) and DP-SPGD(H) are more sensitive to changes in the privacy budget. Our empirical and

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

Table 1: Running time for different methods ($p = 50$).

Method	Running time(s)
DP-SPGD	31.2866
DP-SPGD(K)	20.1467
DP-SPGD(H)	28.8785
HTP-LASSO	202.944
HTDP-FW	1.5017

theoretical results are consistent: DP-SPGD(H) is superior to DP-SPGD(K) in all the cases, and the increase in data dimensionality has less effect on the performance of DP-SPGD(H).

In addition, Figure 1 indicates that DP-SPGD(H) dominates other methods, even HTP-LASSO(O), the algorithm utilizing extra information learned from the simulation dataset. HTP-LASSO outperforms DP-SPGD(K) and DP-SPGD when the privacy budget is small or the dimensionality is high. Obviously, HTP-LASSO performs worse without the oracle information. In the logistic case (See Figure 2), the performance of HTDP-FW is not ideal. Weirdly, HTDP-FW doesn't benefit from the oracle information.

To have an idea of how efficient our approaches are, we examine the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

wall-clock running time for each method. Table 1 summarizes the running time of implementing each method per simulation run. We observe that the running time for HTP-LASSO is much larger than other methods. HTDP-FW converges quickly but to an unsatisfactory result as shown in Figure 4. DP-SPGD takes a longer time than DP-SPGD(H) and DP-SPGD(K) because it requires more iterations.

6.2 Real data analysis

We next consider two real-world applications. We fit a lasso and an ℓ_1-penalized logistic regression model to a blog feedback dataset (with $n = 52396$ and $p = 280$) and a crop-mapping dataset (with $n = 20000$ and $p = 173$), respectively. These datasets are both available online on the UCI Machine Learning Repository (**?**). We aim to predict how many comments a post will receive and categorize crop types. The number of comments received follows a heavy-tailed distribution, whose range is from 0 to 1424. In each case, the data is divided into a training and test set following a 70–30 split. Performance is assessed using the test set through metrics such as empirical risk, root mean square error (RMSE), and classification accuracy. The privacy parameters are set to $\delta = n^{-1}$ and $\epsilon = 0.5, 2$. Each algorithm is implemented with parameters that yield the best performance.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

As an example, for the lasso model, we apply DP-SPGD with a batch size of $m = n/6$ and $m = n$ for the other algorithms.

Empirical risk results on test sets are illustrated in Figures 3 and 4. Besides, the supplementary material includes Tables S1 and S2, which present a comparison of the RMSE and classification accuracy, respectively, of different methods. Our analysis aligns with prior simulation studies, indicating that DP-SPGD(H) and DP-SPGD(K) outperform DP-SPGD in scenarios with a large privacy budget. Notably, both methods perform comparably to DP-SPGD but with a shorter execution time due to the efficacy of robust mean estimators. DP-SPGD(H) exhibits superior prediction performance, as evidenced by lower RMSE and higher classification accuracy, when compared to DP-SPGD(K). This advantage is particularly pronounced in scenarios with a large privacy budget.

7. Discussion

In this paper, we proposed three approaches to regularized DP SCO: two of our algorithms incorporate existing robust mean estimators to address the general difficulties associated with gradient clipping. Our work, motivated by the ubiquity of both nonsmooth regularization in statistics and heavy-tailed data in the real world, fills a notable gap in the differential privacy

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

literature on regularized DP SCO methods for heavy-tailed data. For each of our algorithms, we established theoretical bounds on population excess risk under assumptions of varying strength. By comparing these bounds with related results in the general DP SCO literature, we demonstrated that our methods yield superior performance on a more general class of problems and under milder assumptions. Our extensive numerical studies suggested settings where our algorithms perform more- or less favourably.

Our work leaves open numerous doors for further development. First, even though our proposed algorithms can solve a wide class of regularized SCO problems, the generality of our approach to some extent comes at the cost of estimator accuracy. Through a lasso-specific approach, **?** improved the performance bound of (ϵ, δ)-DP lasso estimators to $\tilde{O}\{\log(p)/(n\epsilon)^{2/5}\}$. This convergence rate has logarithmic dependence in dimension, making it more suitable for high-dimensional cases. This example highlights the potential for further improvements to general DP SCO frameworks. Second, as discussed in Section 5.1, the enhancement of the excess population risk is possible, particularly under smooth population risk conditions. This can be achieved by incorporating the scale of the regularization term g into the risk bound, but it necessitates additional assumptions on g, such as Lipschitz continuity. One possible approach is to follow **?** and bound the subgradient

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

of g. Third, to explore the optimality of proposed algorithms, a lower bound for DP regularized SCO with heavy-tailed data can be developed. Lastly, **?** showed that the f-DP framework can provide better privacy guarantees. Consequently, f-DP may provide an alternative approach that can more effectively balance privacy and utility in DP SCO.

Supplementary Materials

We display the additional numerical results and technical details in the supplementary material.

Acknowledgements

The authors thank the Associate Editor and referees for their constructive comments. The work of Wei Tu was supported by funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the catalyst fund from Queen's University. Bei Jiang and Linglong Kong were partially supported by grants from the Canada CIFAR AI Chairs program, the Alberta Machine Intelligence Institute (AMII), and Natural Sciences and Engineering Council of Canada (NSERC), and Linglong Kong was also partially supported by grants from the Canada Research Chair program from NSERC.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0282

Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, T6G 2G1, Canada

E-mail: haihan1@ualberta.ca

Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, T6G 2G1, Canada

E-mail: pietrosa@ualberta.ca

Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, T6G 2G1, Canada

E-mail: yliu16@ualberta.ca

Department of Public Health Sciences and Canadian Cancer Trials Group, Queen's University, Kingston, Canada

E-mail: wei.tu@queensu.ca

Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, T6G 2G1, Canada

E-mail: bei1@ualberta.ca

Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, T6G 2G1, Canada

E-mail: lkong@ualberta.ca