Differentially Private Stochastic Convex Optimization under a Quantile Loss Function

Du Chen[1] Geoffrey A. Chua[1]

Abstract

We study (ε, δ)-differentially private (DP) stochastic convex optimization under an r-th quantile loss function taking the form $c(u) = ru^+ + (1-r)(-u)^+$. The function is nonsmooth, and we propose to approximate it with a smooth function obtained by convolution smoothing, which enjoys both structure and bandwidth flexibility and can address outliers. This leads to a better approximation than those obtained from existing methods such as Moreau Envelope. We then design private algorithms based on DP stochastic gradient descent and objective perturbation, and show that both algorithms achieve (near) optimal excess generalization risk $O(\max\{\frac{1}{\sqrt{n}}, \frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon}\})$. Through objective perturbation, we further derive an upper bound $O(\max\{\sqrt{\frac{d}{n}}, \sqrt{\frac{d\ln(1/\delta)}{n\varepsilon}}\})$ on the parameter estimation error under mild assumptions on data generating processes. Some applications in private quantile regression and private inventory control will be discussed.

1. Introduction

Stochastic convex optimization (SCO) under a linear quantile loss function, $\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta};\mathbb{P}) := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}\left[\ell(\boldsymbol{\theta};\boldsymbol{x},y)\right]$ where $\ell(\boldsymbol{\theta};\boldsymbol{x},y) := c(y - \boldsymbol{\theta}^\top\boldsymbol{x})$ and $c(u) := ru^+ + (1-r)(-u)^+$, is a fundamental problem in machine learning, and has many applications, such as support vector machine (Suthaharan & Suthaharan 2016), quantile regression (Koenker et al. 2017) and inventory control (Ban & Rudin 2019). Compared to symmetric loss functions (for example, squared function $c(u) = u^2$), the r-th quantile loss function allows imposing asymmetric weights on positive and negative values of u, providing insights into distributional relationships between feature $\boldsymbol{x}$ and dependent variable y.

In practice, SCO is closely related to the Empirical Risk Minimization (ERM) problem, $\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \widehat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{D}) := \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta};\boldsymbol{x}_i,y_i)$ on a dataset $\mathcal{D} := \{(\boldsymbol{x}_i,y_i)\}_{i=1}^n$ of n i.i.d. data points drawn from unknown $\mathbb{P}$. The goal is to output a high-quality estimator $\widehat{\boldsymbol{\theta}}$, from solving an ERM, of $\boldsymbol{\theta}^* := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta};\mathbb{P})$. An estimator's quality is usually measured by *excess generalization risk* $\mathcal{L}(\widehat{\boldsymbol{\theta}};\mathbb{P}) - \mathcal{L}(\boldsymbol{\theta}^*;\mathbb{P})$ or *mean absolute error* $\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2\right]$. The former measure plays an important role in optimization, while the latter is more relevant to statistical inference.

Given a private dataset, estimators may reveal critical information and are at risk of being exploited by attackers. We study the problem under the constraint of *differential privacy*, a mathematically rigorous measure of privacy, which guarantees that the distribution over an algorithm's output is insensitive to a slight change in the dataset. In the private setting, extensive studies have been done to investigate the impact of privacy (Kifer et al. 2012, Bassily et al. 2014, Wang et al. 2017, Bassily et al. 2019, Bassily et al. 2020, Feldman et al. 2020, Bassily et al. 2021b, Bassily et al. 2021a, Asi et al. 2021, Kulkarni et al. 2021, Han et al. 2022).

However, the body of work mentioned above mostly assumes that the loss function $\ell(\boldsymbol{\theta};\boldsymbol{x},y)$ is differentiable and smooth for all $\boldsymbol{\theta}$, which is not the case for a quantile loss function. The quantile loss function is essentially a piecewise linear function with a knot at the origin, at which the curvature is infinite, implying non-differentiability and nonsmoothness. The nonsmoothness will lead to an unstable estimator, prevent gradient-based optimization methods from being efficient, and invalidate uniform stability (Shalev-Shwartz & Ben-David 2014), a crucial property for theoretical analysis.

Some recent works endeavour to address nonsmooth loss functions in DP-SCO. For example, Bassily et al. (2019) and Bassily et al. (2020) proposed to adopt the Noisy Gradient Descent framework, i.e. use a noisy gradient instead of a true gradient to update the estimator in each iteration.

[1] Nanyang Business School, Nanyang Technological University, 639798, Singapore. Correspondence to: Du Chen <chen1443@e.ntu.edu.sg>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Feldman et al. (2020) proposed an iterative localization approach, where an ERM with a localized regularization term is solved for updating estimators in each iteration. Asi et al. (2021) and Kulkarni et al. (2021) extended the iterative localization approach to more general situations. A closely related work, Bassily et al. (2021a) studied Generalized Linear Models with a nonsmooth but bounded loss function while we do not assume boundedness. All these methods were developed for general nonsmooth functions. Though some of them can address the nonsmoothness of a quantile loss, they fail to exploit its special structure, engendering significant performance gaps in practice.

To better take advantage of the structure of a quantile loss, Horowitz (1998) proposed to smooth out the quantile loss function by replacing the implicit indicator function in a quantile loss. However, Horowitz's smooth function gains smoothness at the cost of convexity, i.e. it is no longer globally convex unless $r = 1/2$. Another stream of works (Whang 2006, Kaplan & Sun 2017) proposed to use a smoothed estimating equation estimator, which is essentially the solution to smoothed moment conditions. The solution can be obtained by replacing the implicit indicator function with a kernel counterpart. In this paper, we adopt *convolution smoothing* (Hirschman & Widder 2012) to directly smooth out the entire quantile loss function, resulting in a much smoother and much less variable function (Fernandes et al. 2021). We noticed that a similar convolution smoothing idea was used by Feldman et al. (2018) and Kulkarni et al. (2021) in a DP setting. However, the former only considered Gaussian kernel, while the latter applied it to a strongly convex function whereas the quantile loss function is not strongly convex.

By considering a quantile loss function, our work is also closely related to DP quantile estimation and regression. There is a long history on DP quantile estimation (Dwork & Lei 2009, Dwork et al. 2010, Chan et al. 2011, Bun et al. 2015, Kaplan et al. 2020, Gillenwater et al. 2021, Kaplan et al. 2022). However, these works do not explicitly take regression into consideration, leaving a huge gap in the literature. Our work attempts to fill the gap by considering DP linear quantile regression.

1.1. Our Contributions

The first contribution is the adoption of convolution smoothing for addressing the nonsmoothness of a quantile loss function under a DP context (Section 3.1). We find that, for DP-SCO under a quantile loss, convolution smoothing is preferred over existing methods such as Moreau Envelope. The insight is that convolution smoothing allows us to properly choose both kernel function (structure) and bandwidth (parameter), while Moreau Envelope only allows the choice of smoothness parameter. Discussions on this insight can be found at the end of Section 3.1. Secondly, we find that with convolution smoothing, both gradient perturbation and objective perturbation can achieve (near) optimal excess generalization risks (Theorem 3.4, Theorem 3.6) under very mild assumptions. The third contribution is that we derive an upper bound on the mean absolute error, that is $O(\max\{\sqrt{\frac{d}{n}}, \sqrt{\frac{d\ln(1/\delta)}{n\varepsilon}}\})$, between the private quantile estimator from objective perturbation and the true optimal $\boldsymbol{\theta}^*$ (Theorem 4.5). Lastly, we discuss some applications in statistics and management, and run simulations to demonstrate the superior performance of our approaches empirically. All proofs are deferred to Appendix A and B.

Though quantile loss is specific, it has many applications, for instance, support vector machine (when $r = 1/2$, Suthaharan & Suthaharan 2016), quantile regression (Koenker et al. 2017), inventory control (Ban & Rudin 2019), and serving as clipping thresholds (Andrew et al. 2021). Our in-depth research may pave the way for a deeper understanding on these topics and guide advanced private algorithm design. More importantly, the insight into the superior performance of convolution smoothing may inspire further exploration on other nonsmooth losses.

Notation: We use $\mathcal{X} \times \mathcal{Y} \subseteq \mathcal{B}(B_x) \times \mathbb{R}$ to denote the data domain, where $\mathcal{B}(B)$ is a Euclidean ball with radius B. The r-th quantile loss function is $c(u) = ru^+ + (1 - r)(-u)^+, \forall u \in \mathbb{R}$ where $u^+ := \max\{0, u\}$, and we denote $\ell(\boldsymbol{\theta}; \boldsymbol{x}, y) := c(y - \boldsymbol{\theta}^\top \boldsymbol{x})$, a loss function that takes a vector $\boldsymbol{\theta} \in \mathbb{R}^d$ and a data point $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$ as inputs and outputs a real value. The empirical risk of any $\boldsymbol{\theta} \in \mathbb{R}^d$ w.r.t. loss ℓ and dataset $\mathcal{D} := \{\boldsymbol{x}_i, y_i\}_{i=1}^n$ is defined as $\widehat{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D}) := \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)$. The generalization risk of $\boldsymbol{\theta}$ w.r.t. loss ℓ and distribution $\mathbb{P}$ is defined as $\mathcal{L}(\boldsymbol{\theta}; \mathbb{P}) := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}[\ell(\boldsymbol{\theta}; \boldsymbol{x}, y)]$. Shorthand $\widehat{\mathcal{L}}(\boldsymbol{\theta})$ and $\mathcal{L}(\boldsymbol{\theta})$ are used for the empirical and generalization risk when the dependence is clear from context.

2. Preliminaries

Definition 2.1 (Differential privacy)**.** A randomized algorithm $\mathcal{M} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{Z}$ is (ε, δ)-differential private if, for any pair of neighboring datasets $\mathcal{D} \sim \mathcal{D}'$ that differ in one data point, and for any subset $\mathcal{S} \subseteq \mathcal{Z}$, we have

$$\Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{S}] \leq e^\varepsilon \cdot \Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{S}] + \delta.$$

Definition 2.2 (L-Lipschitz continuity)**.** Let $L > 0$. A function $\ell : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz with respect to norm $\|\cdot\|_2$ over a set $\mathcal{B}$ if for every $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{B}$, we have

$$|\ell(\boldsymbol{\theta}_1) - \ell(\boldsymbol{\theta}_2)| \leq L \cdot \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 .$$

Definition 2.3 (α-strongly convexity)**.** Let $\alpha > 0$. A function $\ell : \mathbb{R}^d \to \mathbb{R}$ is α-strongly convex over a set $\mathcal{B}$ if for every $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{B}$, we have

$$\ell(\boldsymbol{\theta}_1) \geq \ell(\boldsymbol{\theta}_2) + \langle \nabla \ell(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle + \frac{\alpha}{2} \left\| \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\|_2^2 .$$

Definition 2.4 (β-smoothness)**.** Let $\beta > 0$. A function $\ell : \mathbb{R}^d \to \mathbb{R}$ is β-smooth over a set $\mathcal{B}$ if for every $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{B}$, we have

$$\ell(\boldsymbol{\theta}_1) \leq \ell(\boldsymbol{\theta}_2) + \langle \nabla \ell(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle + \frac{\beta}{2} \left\| \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\|_2^2 .$$

Definition 2.5 (τ-uniform stability)**.** A randomized algorithm $\widehat{\boldsymbol{\theta}} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R}^d$ is τ-uniform stable with respect to function $\ell : \mathbb{R}^d \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ if for any pair of neighboring datasets $\mathcal{D} \sim \mathcal{D}'$ that differ in one data point only, we have

$$\sup_{\boldsymbol{x}, y} \mathbb{E} \left[\ell(\widehat{\boldsymbol{\theta}}(\mathcal{D}); \boldsymbol{x}, y) - \ell(\widehat{\boldsymbol{\theta}}(\mathcal{D}'); \boldsymbol{x}, y) \right] \leq \tau,$$

where the expectation is taken over algorithm's randomness.

Lemma 2.6. *(Bousquet & Elisseeff, 2002) Let $\widehat{\boldsymbol{\theta}} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R}^d$ be a τ-uniformly stable algorithm w.r.t. loss function $\ell : \mathbb{R}^d \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Let $\mathbb{P}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{D} \sim \mathbb{P}^n$ be samples i.i.d. drawn from $\mathbb{P}$. Then, we have*

$$\mathbb{E} \left[\mathcal{L}(\widehat{\boldsymbol{\theta}}(\mathcal{D}); \mathbb{P}) - \widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}(\mathcal{D}); \mathcal{D}) \right] \leq \tau,$$

where the expectation is taken over both data sampling $\mathcal{D} \sim \mathbb{P}^n$ and algorithm's randomness.

Of particular interest is the excess generalization risk of a given differentially private algorithm π:

$$\mathcal{R}(\pi; \mathbb{P}) := \mathbb{E}_{\mathcal{D} \sim \mathbb{P}^n, \pi} \left[\mathcal{L}(\widehat{\boldsymbol{\theta}}^{\pi}) \right] - \mathcal{L}(\boldsymbol{\theta}^*),$$

where $\mathcal{L}(\boldsymbol{\theta}) := \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathbb{P}} \left[\ell(\boldsymbol{\theta}; \boldsymbol{x}, y) \right]$ is the expected loss of a given vector $\boldsymbol{\theta}$, and $\boldsymbol{\theta}^*$ is the optimal vector that an oracle with full information of $\mathbb{P}$ can achieve. Following literature in learning theory, the excess generalization risk can be decomposed into two parts,

$$\begin{aligned} \mathcal{R}(\pi; \mathbb{P}) &= \mathbb{E} \left[\mathcal{L}(\widehat{\boldsymbol{\theta}}^{\pi}) - \widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}^{\pi}) \right] + \mathbb{E} \left[\widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}^{\pi}) - \mathcal{L}(\boldsymbol{\theta}^*) \right] \\ &= \mathbb{E} \left[\mathcal{L}(\widehat{\boldsymbol{\theta}}^{\pi}) - \widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}^{\pi}) \right] + \mathbb{E} \left[\widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}^{\pi}) - \widehat{\mathcal{L}}(\boldsymbol{\theta}^*) \right], \end{aligned} \tag{1}$$

where all expectations are taken over $\mathcal{D}$ and π, and the second equality comes from $\mathcal{L}(\boldsymbol{\theta}^*) = \mathbb{E}_{\mathbb{P}^n} \left[\widehat{\mathcal{L}}(\boldsymbol{\theta}^*) \right]$. The above risk decomposition follows a framework in learning theory that, loosely speaking, uniform stability plus shrinking ERM imply learnability. Our risk analysis will heavily rely on (1).

3. Private Algorithms with Convolution Smoothing

In this section, we first show how convolution smoothing works and its impact. Then, DP stochastic gradient descent and objective perturbation algorithms are applied, along with privacy analysis and excess generalization risk analysis.

3.1. Convolution Smoothing

Convolution smoothing is a powerful tool to approximate a nonsmooth function with a smooth function generated from the convolution between the original function and a properly chosen Kernel Function (also known as Approximate Identity). The idea is also used for Kernel Density Estimation and Fourier analysis.

Condition 1 (Kernel Functions). Let $K : \mathbb{R} \to \mathbb{R}_+$ be a nonnegative function having following properties

- **Integrate to 1**: $\int K = 1$;
- **Symmetry**: $K(u) = K(-u)$ for all $u \in \mathbb{R}$;
- **Monotonicity**: $K(u) \leq K(v)$, $\forall \left| u \right| \geq \left| v \right| \in \mathbb{R}$.
- **Finite absolute moments**: $\kappa_1 := \int_{-\infty}^{\infty} \left| u \right| K(u) \, du < \infty$, $\kappa_2 := \int_{-\infty}^{\infty} u^2 K(u) \, du < \infty$, and $\overline{K} := \sup_u K(u) < \infty$.

Any function that satisfies Condition 1 is suitable for our problems. Commonly used kernels are Gaussian Kernel, Logistic Kernel, Uniform Kernel, and Epanechnikov Kernel, summarized in Table 1 under column "$K(u)$".

Given a bandwidth $h > 0$, and a kernel function K, we follow notations in literature and denote, for any $u \in \mathbb{R}$, $K_h(u) := K(u/h)/h$, $\mathcal{K}(u) := \int_{-\infty}^{u} K(v) \, dv$, $\mathcal{K}_h(u) := \mathcal{K}(u/h)$. Then, against the quantile loss functon $c(u) = ru^+ + (1 - r)(-u)^+ = \left| u \right| / 2 + (r - 1/2)u$, the smoothed loss function obtained via convolution smoothing is

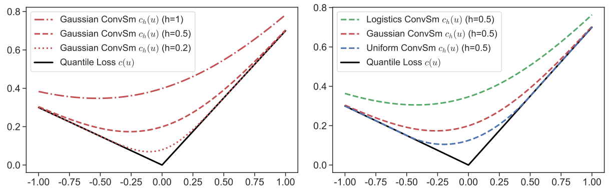
$$\begin{aligned} c_h(u) &:= (c * K_h)(u) \\ &= \int_{-\infty}^{\infty} c(v) K_h(u - v) \, dv \\ &= \frac{h}{2} \int_{-\infty}^{\infty} \left| \frac{u}{h} + v \right| K(v) \, dv + \left(r - \frac{1}{2} \right) u. \end{aligned} \tag{2}$$

Intuitively, the value $c_h(u)$ is a weighted average over u's neighbors, and the weights are given by the adjusted kernel function $K_h(\cdot)$ so that a closer neighbor has a higher weight. With previously mentioned kernel functions, the integral of the first term in (2) has closed forms as shown under the third column in Table 1.

To illustrate the impact of convolution smoothing, we visualize c_h in Figure 1. Mathematically, it can be shown that

Table 1. Kernel Functions

Kernels	$K(u)$	$\int_{-\infty}^{\infty} \lvert u/h+v\rvert\, K(v)\, dv$	$\mathcal{K}_h(u)$
Gaussian	$\frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}}$	$\sqrt{\frac{2}{\pi}} e^{-\frac{(u/h)^2}{2}} + u/h(1-2\Phi(-u/h))$	$\Phi(u/h)$
Logistic	$\frac{e^{-u}}{(1+e^{-u})^2}$	$u/h + 2\ln\left(1+e^{-u/h}\right)$	$(1+e^{-u/h})^{-1}$
Uniform	$\frac{\mathbb{1}\{\lvert u\rvert\le 1\}}{2}$	$\begin{cases} \frac{u^2}{2h^2}+\frac{1}{2}, \text{ if } \lvert u/h\rvert \le 1 \\ \lvert u/h\rvert, \text{ o.w.} \end{cases}$	$\min\{(u/h+1)/2, 1\}\mathbb{1}\{u/h \ge -1\}$
Epanechnikov	$\frac{3}{4}(1-u^2)\mathbb{1}\{\lvert u\rvert \le 1\}$	$\begin{cases} -\frac{u^4}{8h^4}+\frac{3u^2}{4h^2}+\frac{3}{8}, \text{ if } \lvert u/h\rvert \le 1 \\ \lvert u/h\rvert, \text{ o.w.} \end{cases}$	$\begin{cases} -\frac{u^3}{4h^3}+\frac{3u}{4h}+\frac{1}{2}, \text{ if } \lvert u/h\rvert \le 1 \\ \mathbb{1}\{u/h \ge 1\}, \text{ o.w.} \end{cases}$



(a) Bandwidth Impact (b) Structure Impact

Figure 1. Convolution Smoothing. $r = 0.7$

$c_h(\cdot)$ is convex, epi-graphically converges to $c(\cdot)$ as $h \to 0$, and second-order differentiable (Fernandes et al. 2021):

$$c_h'(u) = \mathcal{K}_h(u) + r - 1; \quad c_h''(u) = K_h(u) \ \ge 0.$$

These properties are graphically confirmed by Figure 1(a). Specifically, in Figure 1(a), a smaller bandwidth leads to a better approximation, demonstrating the impact of bandwidth. Moreover, Figure 1(b) highlights the impact of kernel function structure on approximation quality, i.e. a kernel with lighter tails leads to a better approximation.

Similarly, the regression loss function $\ell_h(\boldsymbol{\theta}; \boldsymbol{x}, y) := c_h(y - \boldsymbol{\theta}^\top \boldsymbol{x})$ possesses the same properties, and we provide a summary below for later reference.

Lemma 3.1 (Properties of ℓ_h)**.** *Suppose kernel function K satisfies Condition 1. And define $\ell_h(\boldsymbol{\theta}; \boldsymbol{x}, y) := c_h(y - \boldsymbol{\theta}^\top \boldsymbol{x})$, then function $\ell_h : \mathbb{R}^d \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ satisfies the following properties. For any $\boldsymbol{\theta}, \boldsymbol{x}, y$,*

1. *function ℓ_h is second-order differentiable in $\boldsymbol{\theta}$ with*

$$\nabla \ell_h(\boldsymbol{\theta}; \boldsymbol{x}, y) = \left[\mathcal{K}_h\left(\boldsymbol{\theta}^\top \boldsymbol{x} - y\right) - r\right] \cdot \boldsymbol{x};$$
$$\nabla^2 \ell_h(\boldsymbol{\theta}; \boldsymbol{x}, y) = K_h\left(y - \boldsymbol{\theta}^\top \boldsymbol{x}\right) \cdot \boldsymbol{x}\boldsymbol{x}^\top \ \succcurlyeq \boldsymbol{0};$$

2. *function ℓ_h is $\bar{r}B_x =: L$-Lipschitz continuous and $\overline{K}B_x^2/h =: \beta$-smooth in $\boldsymbol{\theta}$ with respect to $\|\cdot\|_2$;*

3. *function ℓ_h is upper and lower bounded by affine functions of ℓ,*

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}, y) \le \ell_h(\boldsymbol{\theta}; \boldsymbol{x}, y) \le \ell(\boldsymbol{\theta}; \boldsymbol{x}, y) + \frac{1}{2} h\kappa_1.$$

Comparison to Moreau Envelope: Moreau Envelope (Parikh et al. 2014) is a powerful tool to address nonsmoothness and is widely utilized in DP context (Bassily et al. 2019, Feldman et al. 2020, Bassily et al. 2021a). It approximates an original nonsmooth function with a smooth function obtained from *infimal convolution* $c_\beta(u) := (f \square g_\beta)(u) := \inf_x\{f(x) + g_\beta(u - x)\}$ by fixing g_β as $\frac{\beta}{2}\|\cdot\|_2^2$. As shown in Figure 2, Moreau approximates a quantile function from below. But when u is away from 0, the approximation gap is large. In contrast, convolution smoothing approximates a quantile function from above and can tolerate extreme values of u better. Furthermore, Moreau leaves the parameter β freely chosen while convolution smoothing, recall the definition $c_h(u) := \int f(x) \cdot g_h(u - x)\, dx$, allows us to choose the g function in addition to bandwidth parameter h, providing both structure flexibility and parameter flexibility. Lastly, since convolution smoothing approximates from above, it naturally provides an upper bound that leads to tighter risk bounds. This idea is explicitly employed in the proof of Theorem 3.6 part 2.

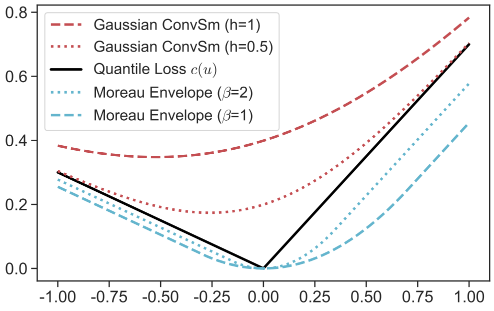


Figure 2. Convolution Smoothing v.s. Moreau Envelope. $r = 0.7$

More General Nonsmooth Functions: Convolution smoothing can also be applied to other nonsmooth functions, such as piecewise linear functions with more than two pieces, as shown in Figure 3. One can expect that the smoothed function should possess similar properties. However, obtaining analytical results will require more effort because of multiple pieces.

With the smooth approximation function from convolution,

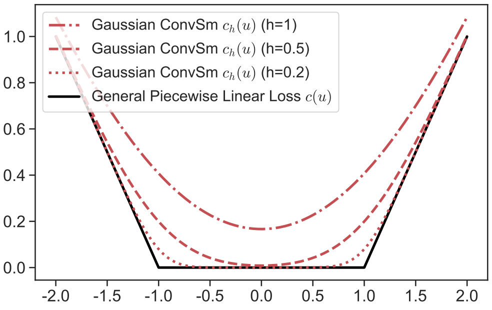


Figure 3. Smoothing Piecewise Linear Function

we next show that some standard differentially private algorithms are suitable for DP-SCO under a quantile loss, and can achieve (near) optimal excess generalizaiton risks.

3.2. DP-Stochastic Gradient Descent

The DP-Stochastic Gradient Descent (DP-SGD) algorithm (Bassily et al. 2014) is based on classic stochastic gradient descent by injecting a carefully calibrated Gaussian noise into the gradient in each iteration. We apply DP-SGD to smoothed quantile loss function ℓ_h rather than ℓ. The nuance appears in Step 5 of Algorithm DP-SGD formally given below.

Algorithm 1 DP-Stochastic Gradient Descent (DP-SGD)

Input: Private dataset $\mathcal{D}$, privacy parameters $\varepsilon \leq 1$, $\delta \geq 0$, kernel function K with bandwidth $h > 0$, Lipschitz parameter $L = \bar{r}B_x$, smoothness parameter $\beta = \overline{K}B_x^2/h$, noise variance $\sigma^2 = 8L^2 \ln(1/\delta)/\varepsilon^2$, step size $\eta > 0$.

```
Set initial point $\widehat{\boldsymbol{\theta}}_{h,1} = \mathbf{0}$
for $t = 1$ to $n^2 - 1$ do
    Uniformly sample a record $(\boldsymbol{x}_{(t)}, y_{(t)})$ from $\mathcal{D}$ with replacement
    Sample a noise vector $\boldsymbol{w}_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{d\times d})$
    $\widehat{\boldsymbol{\theta}}_{h,t+1} \leftarrow \widehat{\boldsymbol{\theta}}_{h,t} - \eta \cdot (\nabla \ell_h(\widehat{\boldsymbol{\theta}}_{h,t}; \boldsymbol{x}_{(t)}, y_{(t)}) + \boldsymbol{w}_t)$
end for
Return $\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGD}} \leftarrow \frac{1}{n^2}\sum_{t=1}^{n^2} \widehat{\boldsymbol{\theta}}_{h,t}$
```

We now state privacy and excess generalization risk guarantees.

Theorem 3.2. *Algorithm DP-SGD is (ε, δ)-differentially private.*

The privacy guarantee follows from Bassily et al. (2020, Theorem 5.1).

Now, we come to analyze the excess generalization risk. Applying part 3 in Lemma 3.1 to the second term in (1), we have:

$$\mathcal{R}(\text{DP-SGD}; \mathbb{P}) \leq \mathbb{E}\left[\mathcal{L}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGD}}) - \widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGD}})\right] + \mathbb{E}\left[\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGD}}) - \widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*)\right] + \frac{1}{2}h\kappa_1.$$

It suffices to show upper bounds for the three terms separately. The key step is to show the uniform stability of Algorithm DP-SGD w.r.t. ℓ. Since ℓ is L-Lipschitz continuous, it remains to control the expected distance between two returned vectors $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}'$ trained on a pair of neighboring datasets. The distance can be controlled if loss function is smooth, and step size is not too large. We formally state the intermediate result in the following lemma.

Lemma 3.3. *In Algorithm DP-SGD, suppose that $\eta \leq 2/\beta$, where β is the smoothness parameter of ℓ_h. Then Algorithm DP-SGD is $\left(\frac{L^2\eta\cdot(1+n^2)}{n}\right)$-uniformly stable with respect to $\ell(\cdot; \boldsymbol{x}, y)$ for any $\boldsymbol{x}, y$.*

The proof utilizes the non-expansiveness property of gradient update rules (Hardt et al., 2016). Combining Lemma 2.6 and Lemma 3.3, we successfully controlled the first term in the excess generalization risk. The second term can be bounded by a classic risk analysis of gradient descent with a fixed step size. Moreover, bounding the third term amounts to properly choosing bandwidth h that adapts to sample size n. Combining these three parts, we arrive at another main contribution of our work.

Theorem 3.4. *Suppose that $\boldsymbol{\theta}^* \in \mathcal{B}(B_\theta)$. Let step size $\eta = B_\theta / \sqrt{2L^2n^3 + (d\sigma^2 + L^2 + \kappa_1 \overline{K} B_x^2/2)n^2 + 2L^2 n}$, bandwidth $h = \eta \overline{K} B_x^2/2$ for Algorithm DP-SGD, then, for any distribution $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$, we have*

$$\mathcal{R}(DP\text{-}SGD; \mathbb{P}) \leq O\left(\max\left\{\frac{1}{\sqrt{n}}, \frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon}\right\}\right).$$

This theorem confirms asymptotic optimality of Algorithm DP-SGD with convolution smoothing. More importantly, we will show in Section 3.4 that the rate is (near) optimal, indicating that the approximation error by convolution smoothing is controllable and does not harm convergence rates.

3.3. Objective Perturbation

We further propose a differentially private algorithm based on objective perturbation (Kifer et al. 2012), which is formally described below. The algorithm boils down to solving a regularized convex optimization problem $\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \widehat{\mathcal{L}}_h^{\mathsf{OP}}(\boldsymbol{\theta}; \mathcal{D}) := \widehat{\mathcal{L}}_h(\boldsymbol{\theta}; \mathcal{D}) + \lambda \|\boldsymbol{\theta}\|_2^2 + \frac{\boldsymbol{b}^\top \boldsymbol{\theta}}{n}$ with a perturbed objective function by a noise vector $\boldsymbol{b}$. Note that the objective perturbation algorithm is applied to the smoothed function $\widehat{\mathcal{L}}_h$ instead of the original function $\widehat{\mathcal{L}}$. As

Algorithm 2 Objective Perturbation (OP)

```
Input: Private dataset $\mathcal{D}$, privacy parameters $\varepsilon > 0, \delta \geq 0$, kernel function $K$ with bandwidth $h > 0$, Lipschitz parameter $L = \bar{r}B_x$, smoothness parameter $\beta = \overline{K}B_x^2/h$, variance $\sigma^2 = L^2\left(8\ln(2/\delta) + 4\varepsilon\right)/\varepsilon^2$
Set any $\lambda \geq \frac{\beta}{n\varepsilon}$
Sample a noise vector $\boldsymbol{b} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{d\times d})$
$\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\boldsymbol{b}) \leftarrow \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \widehat{\mathcal{L}}_h(\boldsymbol{\theta};\mathcal{D}) + \lambda\|\boldsymbol{\theta}\|_2^2 + \frac{\boldsymbol{b}^\top\boldsymbol{\theta}}{n}$
Return: $\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\boldsymbol{b})$
```

a result, the algorithm preserves DP according to Kifer et al. (2012, Theorem 2).

Theorem 3.5. *Algorithm* OP *is (ε, δ)-differentially private.*

Before we characterize the excess generalization risk of OP, we introduce an additional condition on kernel functions.

Condition 2. The kernel function $K : \mathbb{R} \to \mathbb{R}_+$ has a light tail, i.e., there exists a value $v > 0$ such that $K(u) \leq 1/u^3, \forall\, |u| \geq v$.

Theorem 3.6. *Assume $\boldsymbol{\theta}^* \in \mathcal{B}(B_\theta)$. In Algorithm* OP*,*

1. *if we set $\lambda = \frac{1}{B_\theta}\sqrt{\frac{2L^2}{n} + \frac{d\sigma^2}{n^2} + \frac{\kappa_1 \overline{K}B_x^2}{2n\varepsilon}}$, $h = \frac{\overline{K}B_x^2}{\lambda n\varepsilon}$, then, for any distribution $\mathbb{P}$ over $\mathcal{X}\times\mathcal{Y}$, its excess generalization risk satisfies*

$$\mathcal{R}(\mathsf{OP};\mathbb{P}) \leq O\left(\max\left\{\frac{1}{\sqrt{n\varepsilon}}, \frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon}\right\}\right).$$

2. *if kernel function K further satisfies Condition 2 and the residue $u := y - \boldsymbol{\theta}^{*\top}\boldsymbol{x}$ has a finite expected reciprocal $\mathbb{E}_u\left[1/|u|\right] < \infty$. Then, set $\lambda = \frac{1}{B_\theta}\sqrt{\frac{2L^2}{n} + \frac{d\sigma^2}{n^2}}$, $h = \frac{\overline{K}B_x^2}{\lambda n\varepsilon}$, we will have*

$$\mathcal{R}(\mathsf{OP};\mathbb{P}) \leq O\left(\max\left\{\frac{1}{\sqrt{n}}, \frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon}\right\}\right),$$

when $\varepsilon^4 + d\ln(1/\delta)\varepsilon^2 \geq \Omega(\frac{1}{n})$.

The proof follows the same proof idea for Theorem 3.4. Specifically, uniform stability is a straightforward result of the regularizer $\|\cdot\|_2^2$; and shrinking ERM can be derived from empirical risk analysis for regularized optimization problems.

It is remarkable that OP can achieve optimal rates under mild assumptions as shown in part 2 of Theorem 3.6, which is not observed in nonsmooth DP-SCO literature. The additional condition on kernel functions further highlights the importance of structure flexibility. In fact, the condition is not restrictive since all kernels in Table 1 are qualified. The optimality of OP naturally motivates a deeper analysis on OP because the estimators obtained from OP are private M-estimators satisfying first order conditions (FOCs). These FOCs will play a critical role in private quantile regression that will be discussed in Section 4.1

3.4. Lower Bounds

In this subsection, we prove that the convergence rate that our proposed algorithm can achieve is (near) optimal up to logarithmic factors $\ln d$ and $\ln 1/\delta$. Though previous studies (Bassily et al. 2019, Asi et al. 2021) have explored lower bounds for DP-SCO, they use mean estimation to prove the lower bounds, which may not necessarily lead to a lower bound for DP-SCO with quantile functions. To fill the gap, we provide a detailed proof in the Appendix to show the optimality of our algorithms. The optimality claim is formally stated in the following theorem.

Theorem 3.7. *The minimax risk of DP-SCO under a quantile loss function is given as*

$$\inf_{\pi\in\mathcal{F}_{\varepsilon,\delta}} \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \mathcal{R}(\pi;\mathbb{P}) \geq \widetilde{\Omega}\left(\max\left\{\frac{1}{\sqrt{n}}, \frac{\sqrt{d}}{n\varepsilon}\right\}\right),$$

where $\widetilde{\Omega}$ hides the term $\ln d$ in the denominator of the second term in $\max$ operator; $\mathcal{F}_{\varepsilon,\delta}$ is the set of all (ε,δ)-DP mappings from dataset space to estimator space; and $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ is the set of all distributions supported on $\mathcal{X}\times\mathcal{Y}$.

The proof follows a bootstrapping idea in Bassily et al. (2019) but we customize dataset construction. This result confirms that our algorithms with convolution smoothing are (near) optimal. The lower bound in Theorem 3.7 is state-of-the-art in nonsmooth DP-SCO literature, and how to close the gap remains an interesting open question.

4. Applications

4.1. The Private Quantile Regression

Linear quantile regression is one of the most basic and important methods to understand the heterogeneous effect of $\boldsymbol{x}$ on y for a prefixed quantile level r. We assume the underlying true data generating process follows a linear model taking the form $y = \langle\boldsymbol{\theta}^\star, \boldsymbol{x}\rangle + \epsilon^\star(\boldsymbol{x})$ with a prefixed finite $\boldsymbol{\theta}^\star$ but without any restrictions on the endogenous error term $\epsilon^\star(\boldsymbol{x})$. Thus, y could be unbounded because of the error term, even when $\boldsymbol{x}$ is bounded. Under the linearity assumption, the data generating process can be reformulated as $y = \langle\boldsymbol{\theta}^*, \boldsymbol{x}\rangle + \epsilon(\boldsymbol{x})$ with $F_{\epsilon|\boldsymbol{x}}^{-1}(r) = 0$ (Koenker et al. 2017), where $F_{\epsilon|\boldsymbol{x}}$ is the conditional cumulative distribution function of $\epsilon(\boldsymbol{x})$ conditional on $\boldsymbol{x}$; and $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta};\mathbb{P}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}\left[c(y - \boldsymbol{\theta}^\top\boldsymbol{x})\right]$. Using the reformulated form is common in quantile regression literature (Chen et al. 2020, He et al. 2021), and we follow this convention.

Of particular interest in private quantile regression is to find a private estimator to $\boldsymbol{\theta}^*$. Note that the formulation is exactly the same as SCO we considered in the previous section. Thus, we can employ algorithms in Section 3 to generate DP quantile estimators. Since Algorithm OP is optimal and always returns a private minimizer, we restrict our attention to OP only. The measurement of interest is the mean absolute error $\mathbb{E}_{\mathsf{OP}}\left[\left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}-\boldsymbol{\theta}^*\right\|_2\right]$. Denote $\widehat{\boldsymbol{\theta}}_h^{\#}=\arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d}\widehat{\mathcal{L}}_h^{\#}(\boldsymbol{\theta}):=\widehat{\mathcal{L}}_h(\boldsymbol{\theta})+\lambda\left\|\boldsymbol{\theta}\right\|_2^2$, and $\boldsymbol{\theta}_h^*=\arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d}\mathcal{L}_h(\boldsymbol{\theta})$, Then, the estimation error can be decomposed into three parts,

$$\begin{aligned}\mathbb{E}_{\mathsf{OP}}\left[\left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}-\boldsymbol{\theta}^*\right\|_2\right]\leq\ &\mathbb{E}_{\mathsf{OP}}\left[\left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}-\widehat{\boldsymbol{\theta}}_h^{\#}\right\|_2\right]\\&+\left\|\widehat{\boldsymbol{\theta}}_h^{\#}-\boldsymbol{\theta}_h^*\right\|_2+\left\|\boldsymbol{\theta}_h^*-\boldsymbol{\theta}^*\right\|_2,\forall\mathcal{D},\end{aligned}$$

where the three parts are privacy-induced error, estimation error due to sampling, and approximation error due to convolution smoothing, respectively.

Before proceeding, we make some mild assumptions about the true data generating process.

Assumption 4.1. The unknown underlying joint distribution $\mathbb{P}$ of $(\boldsymbol{x},\epsilon(\boldsymbol{x}))$ satisfies

- (**Standardized $\boldsymbol{x}$**): $x_1\equiv 1$, and $\mathbb{E}[x_i]=0,\forall i=2,\ldots,d$.
- (**Lipschitz continuous pdf**): For any $\boldsymbol{x}\in\mathcal{X}$, the conditional pdf $f_{\epsilon|\boldsymbol{x}}$ of $\epsilon(\boldsymbol{x})$ exists, and is l_1-Lipschitz continuous.
- (**Locally strictly positive slope**): There exists a constant $\underline{f}>0$ such that $f_{\epsilon|\boldsymbol{x}}(0)\geq\underline{f}$, for all $\boldsymbol{x}\in\mathcal{X}$.
- (**Widely spread $\boldsymbol{x}$**): Matrix $\boldsymbol{\Sigma}:=\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]\succ\mathbf{0}$ is positive definite, and has a minimal eigenvalue $\rho_1>0$.

The first assumption incorporates an intercept term and does not lose generality. The second assumption is mild as it at least admits Gaussian noise ϵ that is independent of $\boldsymbol{x}$. The third assumption ensures the uniqueness of $\boldsymbol{\theta}^*$. The last assumption is common in literature to help improve estimation accuracy as one can expect to observe $\boldsymbol{x}$ along all directions with reasonable probabilities. Now, we are ready to control the three terms.

Lemma 4.2. *Suppose $\boldsymbol{\theta}^*\in\mathcal{B}(B_\theta)$ and Assumption 4.1 holds. Let bandwidth $h>0$ be small enough such that $\underline{f}-hl_1(\kappa_1+\sqrt{B_x^3\kappa_2}/\rho_1)>0$. Then, we have*

$$\left\|\boldsymbol{\theta}_h^*-\boldsymbol{\theta}^*\right\|_2\leq\frac{h^2l_1\kappa_2}{\rho_1\left(\underline{f}-hl_1\kappa_1\right)}.$$

This lemma confirms that convolution smoothing does not significantly affect estimation, and the error reduces to 0 at a quadratic rate in h. The proof exploits the special structure of the quantile loss function.

Lemma 4.3. *Suppose $\boldsymbol{\theta}^*\in\mathcal{B}(B_\theta)$. For any dataset $\mathcal{D}\sim\mathbb{P}^n$ and $\mathbb{P}$ that satisfies Assumption 4.1, if $\lambda\asymp\frac{1}{B_\theta}\cdot\sqrt{\frac{\kappa_1\overline{K}B_x^2}{n\varepsilon}+\frac{d\sigma^2}{n^2}}$ is set as the same value in Theorem 3.6 part 1, then the estimator $\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}$ given by Algorithm OP satisfies*

$$\mathbb{E}_{\mathsf{OP}}\left[\left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}-\widehat{\boldsymbol{\theta}}_h^{\#}\right\|_2\right]\lesssim\frac{LB_\theta}{B_x\sqrt{\kappa_1\overline{K}}}\cdot\sqrt{\frac{d\ln(1/\delta)}{n\varepsilon}},$$

where κ_1 and $\overline{K}$ are values induced from kernel function.

The proof exploits strong convexity of $\widehat{\mathcal{L}}_h^{\mathsf{OP}}$.

Lemma 4.4. *Suppose $\boldsymbol{\theta}^*\in\mathcal{B}(B_\theta)$. Under Assumption 4.1, if we set $h\leq o(1)$ and $\lambda\leq o(1)$ such that $B_\theta^2\lambda^2\gtrsim h^2l_1\kappa_2$, then with probability at least $1-\gamma,\forall\gamma\in(0,1)$ over random draw of samples, we have*

$$\left\|\widehat{\boldsymbol{\theta}}_h^{\#}-\boldsymbol{\theta}_h^*\right\|_2\lesssim\frac{1}{\rho_1\underline{f}}\cdot\left(L\sqrt{\frac{d+\ln(1/\gamma)}{n}}+B_\theta\lambda\right).$$

The proof sketch is as follows. We first notice that proving the lemma is equivalent to proving $\widehat{D}_h^{\#}(\boldsymbol{\delta},\boldsymbol{\theta}_h^*):=\widehat{\mathcal{L}}_h^{\#}(\boldsymbol{\theta}_h^*+\boldsymbol{\delta})-\widehat{\mathcal{L}}_h^{\#}(\boldsymbol{\theta}_h^*)>0,\forall\boldsymbol{\delta}\in\partial\mathcal{B}(r_0):=\{\boldsymbol{\delta}\in\mathbb{R}^d:\left\|\boldsymbol{\delta}\right\|_2=r_0\}$ with a radius r_0 chosen as the value on the r.h.s. of the inequality in Lemma 4.4 (Wainwright 2019, Lemma 9.21). We then obtain a lower bound on $\widehat{D}_h^{\#}(\boldsymbol{\delta},\boldsymbol{\theta}_h^*)$, which is a positive term minus some terms. And these subtrahend terms can be further upper bounded through Rademacher Complexity and covering arguments. With vanishing h and λ, the overall lower bound is finally found to be strictly positive with high probability.

Combining the preceding three lemmas, we can derive the following upper bound for the overall estimation error.

Theorem 4.5. *Let K be a kernel function satisfying Condition 1, and assume privacy parameter $\delta\asymp n^{-w}$ for some $w>0$. In Algorithm OP, if $\lambda\asymp\frac{1}{B_\theta}\cdot\sqrt{\frac{\kappa_1\overline{K}B_x^2}{n\varepsilon}+\frac{d\sigma^2}{n^2}}$ and $h=\overline{K}B_x^2/(\lambda n\varepsilon)$ as the same values in Theorem 3.6 part 1, then for any distribution $\mathbb{P}$ that satisfies Assumption 4.1, with probability at least $1-\gamma,\forall\gamma\in(0,1)$ over the random draw of dataset $\mathcal{D}$, the private estimator $\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}$ obtained from Algorithm OP satisfies*

$$\mathbb{E}_{\mathsf{OP}}\left[\left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}-\boldsymbol{\theta}^*\right\|_2\right]\lesssim\frac{1}{\rho_1\underline{f}}\cdot\max\left\{\sqrt{\frac{d}{n}},\sqrt{\frac{d\ln(1/\delta)}{n\varepsilon}}\right\},$$

where we omit an additive term $\sqrt{\ln(1/\gamma)/n}$ in the first term of the max *operator.*

The Theorem gives an upper bound on estimation error w.r.t. $\|\cdot\|_2$.The requirement on δ ensures $B_\theta^2\lambda^2 \gtrsim h^2 l_1 \kappa_2$ so that we can employ Lemma 4.4. Dependencies on other constants are omitted. It is well known that, by Bayes risk arguments, the statistical estimation error is at order $\sqrt{d/n}$. Our derived bound introduces an additional term involving privacy parameters. To our best knowledge, our work is the first to characterize the estimation error of a private quantile estimator in DP context.

4.2. The Private Newsvendor Problem

The Newsvendor problem is the most fundamental problem in inventory control. A newsvendor must decide an order quantity q before daily demand y is realized. After the demand is realized, if $q \leq y$, an underage cost $c_u \cdot (y - q)$ will be incurred; if $q > y$, an overage cost $c_o \cdot (q - y)$ will be incurred, where $c_u, c_o > 0$ are the unit underage cost and the unit overage cost, respectively. The newsvendor aims to minimize the expected cost $\min_{q\in\mathbb{R}} \mathbb{E}_y \left[C(y - q)\right]$ with $C(u) := c_u \cdot u^+ + c_o \cdot (-u)^+ = (c_u + c_o)\cdot c(u)$, where $c(u)$ is the $\frac{c_u}{c_u+c_o}$-th quantile loss function. The private newsvendor problem further requires the order quantity q to be differentially private. When demand y follows the same linear model in Section 4.1, we are safe to restrict the ordering rule to a linear form $q = \widehat{\boldsymbol{\theta}}^\top \boldsymbol{x}$ (Ban & Rudin 2019). Thus, solving a private newsvendor problem amounts to finding a private estimator $\widehat{\boldsymbol{\theta}}$ to minimize the excess expected cost

$$\mathcal{C}(\widehat{\boldsymbol{\theta}}; \mathbb{P}) := (c_u + c_o) \cdot \Big[\mathbb{E}_{\widehat{\boldsymbol{\theta}},(\boldsymbol{x},y)\sim\mathbb{P}} \left[c(y - \widehat{\boldsymbol{\theta}}^\top \boldsymbol{x})\right] - \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}} \left[c(y - \boldsymbol{\theta}^{*\top} \boldsymbol{x})\right]\Big],$$

Note that the excess expected cost is the same as that of a DP-SCO under a quantile loss function, up to a constant factor $c_u + c_o$. Hence, all proposed algorithms in Section 3 are suitable for the private newsvendor problem. Therefore, we have the following conclusions.

Theorem 4.6. *1. Running Algorithm DP-SGD with the same settings as in Theorem 3.4 for the private newsvendor problem yields*

$$\mathcal{C}(\mathsf{SGD}; \mathbb{P}) \leq c_{u+o} \cdot O\left(\max\left\{\frac{1}{\sqrt{n}}, \frac{\sqrt{d \ln(1/\delta)}}{n\varepsilon}\right\}\right),$$

where $c_{u+o} := c_u + c_o$.

2. Running Algorithm OP *with the same settings as in part 2 of Theorem 3.6 for the private newsvendor problem yields*

$$\mathcal{C}(\mathsf{OP}; \mathbb{P}) \leq c_{u+o} \cdot O\left(\max\left\{\frac{1}{\sqrt{n}}, \frac{\sqrt{d \ln(1/\delta)}}{n\varepsilon}\right\}\right),$$

when corresponding assumptions in Theorem 3.6 part 2 are satisfied, where $c_{u+o} := c_u + c_o$.

These statements are corollaries of results in Section 3, and no proof is needed. Moreover, the differential privacy of order quantity q is ensured by DP's post-processing lemma.

5. Experiments

We run simulations on synthetic datasets to demonstrate our theoretical findings empirically. Seven algorithms are implemented, including a non-private regularized model $\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \widehat{\mathcal{L}}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_2^2$, our proposed private algorithms OP and DP-SGD, an empirical (DP-)SGD without smoothing, Moreau envelope (Bassily et al. 2019, Theorem 4.4, Algorithm 1), Phased-ERM (Feldman et al. 2020, Theorem 4.8, Algorithm 3), and Phased-SGD (Bassily et al. 2021a, Theorem 5, Algorithm 4). We examine the excess generalization risk of an estimator $\widehat{\boldsymbol{\theta}}$ relative to the risk of true optimal $\boldsymbol{\theta}^*$, i.e., $\frac{\mathcal{L}(\widehat{\boldsymbol{\theta}})-\mathcal{L}(\boldsymbol{\theta}^*)}{\mathcal{L}(\boldsymbol{\theta}^*)}$, and the relative estimation error $\frac{\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2}{\|\boldsymbol{\theta}^*\|_2}$ in Figure 4, 5 (exogenous error term) and Figure 6, 7 (endogenous error term). Shadow areas represent standard deviations.

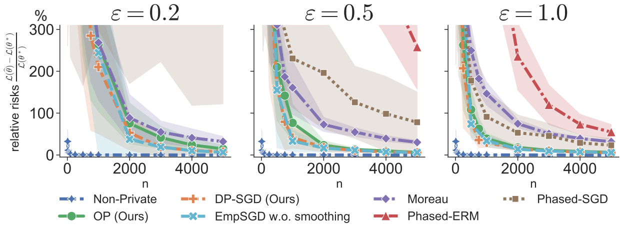


Figure 4. Relative excess generalization risks. Data generating process follows $y = 10 + 5x_1 - 2x_2 + \epsilon$, where $x_1 \sim \mathcal{N}(0, 2^2)$, $x_2 \sim \mathcal{N}(0, 3^2)$ with $(x_1, x_2) \in \mathcal{B}(B_x)$, $B_x = 10$, and $\epsilon \sim \mathcal{N}(0, 3^2)$. Quantile level $r = 0.7$, and in this case $\boldsymbol{\theta}^* = (11.41, 5, -2)$. We set $B_\theta = 2\|\boldsymbol{\theta}^*\|_2$. Privacy parameters ε is set accordingly, and $\delta = 10^{-2}$. Logistic kernel is used. Simulations are repeatedly run 50 times with gradually increasing sample size n.

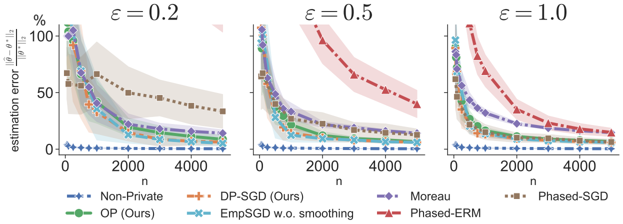


Figure 5. Relative estimation errors. Settings are the same as in Figure 4.

We can see from Figure 4 that under all privacy levels, our convolution smoothing based algorithms OP and DP-SGD outperform existing private methods, and are as good as empirical SGD without smoothing. This is because our approaches exploit the special structure of a quantile loss function explicitly, while others do not. Nevertheless, the authors still respect the universal applicability of other existing methods in tackling nonsmoothness, since our convolution

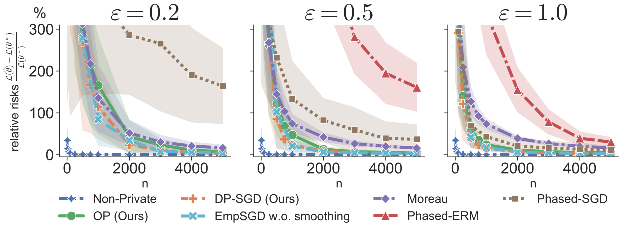


Figure 6. Relative excess generalization risks. Settings are the same as in Figure 4, except $\epsilon(\boldsymbol{x}) = \mathcal{N}(0, 3^2) + x_1 - x_2 + |x_1 x_2|$.

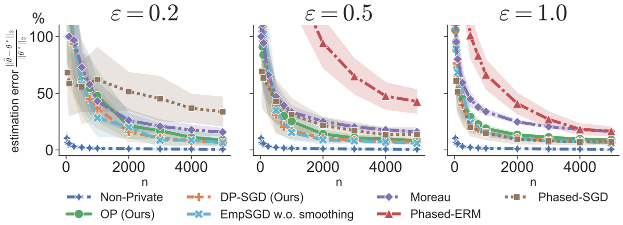


Figure 7. Relative estimation errors. Settings are the same as in Figure 6.

smoothing is exclusively designed for a quantile loss function. In Figure 5, we show the relative estimation errors, and our algorithms still perform better. When the error term is endogenous (Figure 6 and 7), trends remain the same. We defer other large-scale simulation results to Appendix C.

For completeness, we report computational time for solving models in Table 2. Remarkably, the solving time of Algorithm OP does not increase massively compared to its non-private counterpart, while other methods' solving time increases significantly.

6. Conclusions and Future Directions

This work examined differentially private stochastic convex optimization under a quantile loss function. To deal with the nonsmoothness of a quantile function, we proposed to approximate it with a smooth function obtained by convolution smoothing. Convolution smoothing enjoys both structure and parameter flexibility, resulting in a better approximation over existing methods. Based on the smoothed function, we applied DP-SGD and OP, and studied their performances theoretically and empirically. We found that DP-SGD and OP can both achieve (near) optimal excess generalization risks, and are practically appealing. We also derived an estimation error in parameters.

Following our idea, it would be interesting to apply convolution smoothing to more general nonsmooth losses such as general piecewise linear functions, and design private algorithms and derive analytical results accordingly.

Table 2. Solving time (mean±std, in seconds)

	Model	n=100	n=1000	n=5000
d=3	Non-Private	0.0±0.0	0.1±0.0	0.4±0.1
	OP (Ours)	0.0±0.0	0.3±0.0	2.8±0.2
	DP-SGD (Ours)	0.7±0.0	62.4±0.5	1575.1±27.7
	EmpSGD	0.6±0.0	59.3±1.9	1431.8±15.4
	Moreau	0.6±0.0	19.4±0.3	317.7±3.8
	Phased-ERM	0.2±0.0	0.2±0.0	0.3±0.0
	Phased-SGD	0.1±0.0	0.1±0.0	0.1±0.0
d=51	Non-Private	0.1±0.1	0.2±0.1	1.6±0.2
	OP (Ours)	0.1±0.0	0.8±0.3	6.1±0.6
	DP-SGD (Ours)	1.1±0.3	69.5±1.1	1600.0±26.1
	EmpSGD	1.0±0.4	62.2±1.3	1528.3±18.8
	Moreau	0.6±0.1	24.9±0.5	323.6±4.8
	Phased-ERM	1.0±0.1	4.2±0.8	8.4±0.9
	Phased-SGD	0.3±0.1	0.3±0.1	0.3±0.1

References

Andrew, G., Thakkar, O., McMahan, B., and Ramaswamy, S. Differentially private learning with adaptive clipping. *Advances in Neural Information Processing Systems*, 34: 17455–17466, 2021.

Asi, H., Feldman, V., Koren, T., and Talwar, K. Private stochastic convex optimization: Optimal rates in l1 geometry. In *International Conference on Machine Learning*, pp. 393–403. PMLR, 2021.

Ban, G.-Y. and Rudin, C. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1):90–108, 2019.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pp. 464–473. IEEE, 2014.

Bassily, R., Feldman, V., Talwar, K., and Guha Thakurta, A. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019.

Bassily, R., Feldman, V., Guzmán, C., and Talwar, K. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391, 2020.

Bassily, R., Guzmán, C., and Menart, M. Differentially private stochastic optimization: New results in convex and non-convex settings. *Advances in Neural Information Processing Systems*, 34:9317–9329, 2021a.

Bassily, R., Guzmán, C., and Nandi, A. Non-euclidean differentially private stochastic convex optimization. In *Conference on Learning Theory*, pp. 474–499. PMLR, 2021b.

Bousquet, O. and Elisseeff, A. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Bun, M., Nissim, K., Stemmer, U., and Vadhan, S. Differentially private release and learning of threshold functions.

In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pp. 634–649. IEEE, 2015.

Chan, T.-H. H., Shi, E., and Song, D. Private and continual release of statistics. *ACM Transactions on Information and System Security (TISSEC)*, 14(3):1–24, 2011.

Chen, X., Liu, W., Mao, X., and Yang, Z. Distributed high-dimensional regression under a quantile loss function. *Journal of Machine Learning Research*, 21, 2020.

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

Dwork, C., Naor, M., Pitassi, T., and Rothblum, G. N. Differential privacy under continual observation. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 715–724, 2010.

Feldman, V., Mironov, I., Talwar, K., and Thakurta, A. Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 521–532. IEEE, 2018.

Feldman, V., Koren, T., and Talwar, K. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 439–449, 2020.

Fernandes, M., Guerre, E., and Horta, E. Smoothing quantile regressions. *Journal of Business & Economic Statistics*, 39(1):338–357, 2021.

Gillenwater, J., Joseph, M., and Kulesza, A. Differentially private quantiles. In *International Conference on Machine Learning*, pp. 3713–3722. PMLR, 2021.

Han, Y., Liang, Z., Liang, Z., Wang, Y., Yao, Y., and Zhang, J. Private streaming sco in ℓ_p geometry with applications in high dimensional online decision making. In *International Conference on Machine Learning*, pp. 8249–8279. PMLR, 2022.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

He, X., Pan, X., Tan, K. M., and Zhou, W.-X. Smoothed quantile regression with large-scale inference. *Journal of Econometrics*, 2021.

Hirschman, I. I. and Widder, D. V. *The convolution transform*. Courier Corporation, 2012.

Horowitz, J. L. Bootstrap methods for median regression models. *Econometrica*, pp. 1327–1351, 1998.

Kaplan, D. M. and Sun, Y. Smoothed estimating equations for instrumental variables quantile regression. *Econometric Theory*, 33(1):105–157, 2017.

Kaplan, H., Ligett, K., Mansour, Y., Naor, M., and Stemmer, U. Privately learning thresholds: Closing the exponential gap. In *Conference on Learning Theory*, pp. 2263–2285. PMLR, 2020.

Kaplan, H., Schnapp, S., and Stemmer, U. Differentially private approximate quantiles. In *International Conference on Machine Learning*, pp. 10751–10761. PMLR, 2022.

Kifer, D., Smith, A., and Thakurta, A. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pp. 25–1. JMLR Workshop and Conference Proceedings, 2012.

Koenker, R., Chernozhukov, V., He, X., and Peng, L. Handbook of quantile regression. 2017.

Kulkarni, J., Lee, Y. T., and Liu, D. Private non-smooth erm and sco in subquadratic steps. *Advances in Neural Information Processing Systems*, 34:4053–4064, 2021.

Parikh, N., Boyd, S., et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Suthaharan, S. and Suthaharan, S. Support vector machine. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pp. 207–235, 2016.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wang, D., Ye, M., and Xu, J. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30, 2017.

Whang, Y.-J. Smoothed empirical likelihood methods for quantile regression models. *Econometric Theory*, 22(2): 173–205, 2006.

A. Proofs for Section 3

A.1. Proof of Lemma 3.1

Proof. Gradients and Hessian matrix in part 1 of this Lemma can be easily calculated from function $c_h(\cdot)$.

As for part 2, the L-Lipschitz continuity is by noticing $\sup_{\boldsymbol{\theta},\boldsymbol{x},y}\left\|\nabla\ell_h(\boldsymbol{\theta};\boldsymbol{x},y)\right\|_2 = \bar{r}B_x$; the β-smoothness is by noticing the supremum of maximum eigenvalue of Hessian matrix $\sup_{\boldsymbol{\theta};\boldsymbol{x},y}\lambda_{\max}\left(\nabla^2\ell_h(\boldsymbol{\theta};\boldsymbol{x},y)\right) = \sup_{\boldsymbol{\theta};\boldsymbol{x},y}\lambda_{\max}\left(K_h\left(y-\boldsymbol{\theta}^\top\boldsymbol{x}\right)\cdot\boldsymbol{x}\boldsymbol{x}^\top\right) \le \overline{K}B_x^2/h$.

To prove part 3, it is equivalent to show $c(u) \le c_h(u) \le c(u) + \frac{1}{2}h\kappa_1, \forall u \in \mathbb{R}$. Recall that $c(u) = |u|/2 + (r-1/2)u$. Let $g_h(u) := c_h(u) - c(u)$, we have $g_h'(u) = \mathcal{K}_h(u) \ge 0, \forall u < 0$ and $g_h'(u) = \mathcal{K}_h(u) - 1 \le 0, \forall u > 0$. Since $g_h(\cdot)$ is continuous on $\mathbb{R}$, it takes maximum value at $u = 0$ with $g_h(0) = c_h(0) - c(0) = \frac{h}{2}\int_{-\infty}^{\infty}|v|\,K(v)\,dv$, which proves the right-hand-side in part 3. Moreover, this upper bound is tight by our argument. It remains to show, for any $h > 0$, $\lim_{u\to\infty} g_h(u) \ge 0$ and $\lim_{u\to-\infty} g_h(u) \ge 0$. To prove this result, we can first calculate $g_h(u)$ explicitly:

$$2g_h(u) = \int_{-\infty}^{\infty}|u+vh|\,K(v)\,dv - |u| = 2h\int_{|u|/h}^{\infty}vK(v)\,dv + u\cdot\int_{-u/h}^{u/h}K(v)\,dv - |u|\,, \quad \forall u \in \mathbb{R}.$$

We notice that, when u tends to positive or negative infinity, both limits of $g_h(u)$ exist and equal 0, which completes the proof. □

A.2. Proof of Lemma 3.3

Proof. Because of L-Lipschitz continuity of ℓ, we have

$$\mathbb{E}_{\mathsf{SGD}}\left[\ell(\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGD}}(\mathcal{D});\boldsymbol{x},y) - \ell(\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGD}}(\mathcal{D}');\boldsymbol{x},y)\right] \le L\cdot\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGD}}(\mathcal{D}) - \widehat{\boldsymbol{\theta}}_h^{\mathsf{SGD}}(\mathcal{D}')\right\|_2\right], \quad \forall \boldsymbol{x}, y, \mathcal{D}\sim\mathcal{D}';$$

therefore, it suffices to control the expected deviation between returned vectors trained on two neighboring datasets. For notation brevity, the superscribe $^{\mathsf{SGD}}$ and dependences on $\mathcal{D}$ are omitted throughout this proof, and we use $\widehat{\boldsymbol{\theta}}_h := \widehat{\boldsymbol{\theta}}_h^{\mathsf{SGD}}(\mathcal{D})$ and $\widehat{\boldsymbol{\theta}}_h' := \widehat{\boldsymbol{\theta}}_h^{\mathsf{SGD}}(\mathcal{D}')$. We follow the same idea in Bassily et al. (2019, Lemma 3.4) to complete our proof. We use $\widehat{\boldsymbol{\theta}}_{h,t}$ and $\widehat{\boldsymbol{\theta}}_{h,t}'$ to represent vectors in t-th iteration trained on $\mathcal{D}$ and $\mathcal{D}'$, respectively. Firstly, it can be shown that, when step size $\eta \le 2/\beta$, we have

$$\mathbb{E}_{\mathsf{SGD}}\left[\left\|\widehat{\boldsymbol{\theta}}_{h,t} - \widehat{\boldsymbol{\theta}}_{h,t}'\right\|_2\right] \le \frac{2L\eta t}{n}, \quad \forall t = 1,\ldots,n^2. \tag{3}$$

This can be proved by induction. When $t = 1$, by the setting of initial points $\widehat{\boldsymbol{\theta}}_{h,1} = \widehat{\boldsymbol{\theta}}_{h,1}' = \mathbf{0}$, obviously it is true. Then suppose that it is true for t-th iteration, it remains to check $(t+1)$-th iteration. Let us fix a sequence of noise vector $\{\boldsymbol{w}_t\}_{t=1}^{n^2}$, then,

$$\begin{aligned}\left\|\widehat{\boldsymbol{\theta}}_{h,t+1} - \widehat{\boldsymbol{\theta}}_{h,t+1}'\right\|_2 &= \left\|\left(\widehat{\boldsymbol{\theta}}_{h,t} - \eta(\nabla\ell_h(\widehat{\boldsymbol{\theta}}_{h,t};\boldsymbol{x}_{(t)},y_{(t)}) + \boldsymbol{w}_t)\right) - \left(\widehat{\boldsymbol{\theta}}_{h,t}' - \eta(\nabla\ell_h(\widehat{\boldsymbol{\theta}}_{h,t}';\boldsymbol{x}_{(t)}',y_{(t)}') + \boldsymbol{w}_t)\right)\right\|_2 \\ &= \left\|\left(\widehat{\boldsymbol{\theta}}_{h,t} - \eta\cdot\nabla\ell_h(\widehat{\boldsymbol{\theta}}_{h,t};\boldsymbol{x}_{(t)},y_{(t)})\right) - \left(\widehat{\boldsymbol{\theta}}_{h,t}' - \eta\cdot\nabla\ell_h(\widehat{\boldsymbol{\theta}}_{h,t}';\boldsymbol{x}_{(t)}',y_{(t)}')\right)\right\|_2\end{aligned} \tag{4}$$

Notice that $(\boldsymbol{x}_{(t)},y_{(t)})$ and $(\boldsymbol{x}_{(t)}',y_{(t)}')$ are uniformly drawn from $\mathcal{D}$ and $\mathcal{D}'$, it implies that, with probability $1/n$, we have $(\boldsymbol{x}_{(t)},y_{(t)}) \ne (\boldsymbol{x}_{(t)}',y_{(t)}')$; and with probability $1-1/n$, we have $(\boldsymbol{x}_{(t)},y_{(t)}) = (\boldsymbol{x}_{(t)}',y_{(t)}')$. When not equal, (4) $\le \left\|\widehat{\boldsymbol{\theta}}_{h,t} - \widehat{\boldsymbol{\theta}}_{h,t}'\right\|_2 + 2\eta L$ because of triangular inequality. When equal, gradient update rule is 1-expansive, i.e., (4) $\le \left\|\widehat{\boldsymbol{\theta}}_{h,t} - \widehat{\boldsymbol{\theta}}_{h,t}'\right\|_2$ (Hardt et al. 2016, Lemma 3.6). Consequently, taking expectation over the randomness of algorithm, we have

$$\begin{aligned}\mathbb{E}_{\mathsf{SGD}}\left[\left\|\widehat{\boldsymbol{\theta}}_{h,t+1} - \widehat{\boldsymbol{\theta}}_{h,t+1}'\right\|_2\right] &\le \left(1-\frac{1}{n}\right)\mathbb{E}_{\mathsf{SGD}}\left[\left\|\widehat{\boldsymbol{\theta}}_{h,t} - \widehat{\boldsymbol{\theta}}_{h,t}'\right\|_2\right] + \frac{1}{n}\mathbb{E}_{\mathsf{SGD}}\left[\left\|\widehat{\boldsymbol{\theta}}_{h,t} - \widehat{\boldsymbol{\theta}}_{h,t}'\right\|_2 + 2\eta L\right] \\ &\le \mathbb{E}_{\mathsf{SGD}}\left[\left\|\widehat{\boldsymbol{\theta}}_{h,t} - \widehat{\boldsymbol{\theta}}_{h,t}'\right\|_2\right] + \frac{2\eta L}{n} \\ &\le \frac{2\eta L(t+1)}{n}.\end{aligned}$$

Therefore, by induction, our argument is true. Recall that $\widehat{\boldsymbol{\theta}}_h = \frac{1}{n^2}\sum_{t=1}^{n^2}\widehat{\boldsymbol{\theta}}_{h,t}$ is the averaged vector over all iterations, thus,

$$\mathbb{E}_{\mathsf{SGD}}\left[\left\|\widehat{\boldsymbol{\theta}}_h - \widehat{\boldsymbol{\theta}}'_h\right\|_2\right] \leq \frac{1}{n^2}\sum_{t=1}^{n^2}\mathbb{E}_{\mathsf{SGD}}\left[\left\|\widehat{\boldsymbol{\theta}}_{h,t} - \widehat{\boldsymbol{\theta}}'_{h,t}\right\|_2\right] \leq \frac{1}{n^2}\sum_{t=1}^{n^2}\frac{2\eta Lt}{n} = \frac{(1+n^2)\eta L}{n}$$

□

A.3. Proof of Theorem 3.4

Proof. The excess generalization risk is

$$\mathcal{R}(\mathsf{SGD};\mathbb{P}) \leq \mathbb{E}_{\mathcal{D},\mathsf{SGD}}\left[\mathcal{L}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGD}}) - \widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGD}})\right] + \mathbb{E}_{\mathcal{D},\mathsf{SGD}}\left[\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGD}}) - \widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*)\right] + \frac{1}{2}h\kappa_1.$$

By Lemma 3.3, we know that $\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGD}}$ is uniformly stable w.r.t. ℓ, and therefore according to Lemma 2.6, the first term is upper bounded by $\frac{L^2\eta\cdot(1+n^2)}{n}$. Now, we fix a dataset $\mathcal{D}$ and come to bound the second term:

$$\begin{aligned}
\mathbb{E}_{\mathsf{SGD}}\left[\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_h^{\mathsf{SGD}}) - \widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*)\right] &\leq \frac{1}{n^2}\cdot\mathbb{E}_{\mathsf{SGD}}\left[\sum_{t=1}^{n^2}\left(\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_{h,t}) - \widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*)\right)\right]\\
&\leq \frac{1}{n^2}\cdot\mathbb{E}_{\mathsf{SGD}}\left[\sum_{t=1}^{n^2}\left\langle\widehat{\boldsymbol{\theta}}_{h,t} - \boldsymbol{\theta}^*, \nabla\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_{h,t})\right\rangle\right].\\
&= \frac{1}{n^2}\cdot\mathbb{E}_{\mathsf{SGD}}\left[\sum_{t=1}^{n^2}\left\langle\widehat{\boldsymbol{\theta}}_{h,t} - \boldsymbol{\theta}^*, \nabla\ell_h(\widehat{\boldsymbol{\theta}}_{h,t}, \boldsymbol{x}_{(t)}, y_{(t)}) + \boldsymbol{w}_t\right\rangle\right]\\
&\leq \frac{1}{n^2}\cdot\mathbb{E}_{\mathsf{SGD}}\left[\frac{\|\boldsymbol{\theta}^*\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{n^2}\left\|\nabla\ell_h(\widehat{\boldsymbol{\theta}}_{h,t}, \boldsymbol{x}_{(t)}, y_{(t)}) + \boldsymbol{w}_t\right\|_2^2\right],\\
&\leq \frac{\|\boldsymbol{\theta}^*\|_2^2}{2n^2\eta} + \frac{\eta L^2}{2} + \frac{\eta d\sigma^2}{2},
\end{aligned}$$

where first two lines are due to convexity of $\widehat{\mathcal{L}}_h$; the third line comes from the fact that $\boldsymbol{w}_t$ is drawn from a zero-mean Gaussian distribution and is independent of $\widehat{\boldsymbol{\theta}}_{h,t}$, and $\nabla\ell_h(\widehat{\boldsymbol{\theta}}_{h,t}; \boldsymbol{x}_{(t)}, y_{(t)})$ is an unbiased gradient to $\nabla\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_{h,t})$; the forth line follows from a classic gradient descent analysis (Shalev-Shwartz & Ben-David 2014, Lemma 14.1) and from the gradient descent update rule in our algorithm; the last line is due to L-Lipschitz continuity of ℓ_h and Gaussian vector's upper bounds.

Plugging back into the risk expression, letting $h = \eta\overline{K}B_x^2/2$ (which ensures $\eta\beta = 2$, making Lemma 3.3 hold), and replacing $\|\boldsymbol{\theta}^*\|_2$ with B_θ, we obtain

$$\mathcal{R}(\mathsf{SGD};\mathbb{P}) \leq \frac{L^2\eta\cdot(1+n^2)}{n} + \frac{B_\theta^2}{2n^2\eta} + \frac{\eta L^2}{2} + \frac{\eta d\sigma^2}{2} + \frac{\eta\kappa_1\overline{K}B_x^2}{4}, \quad \forall\eta\geq 0.$$

It is easy to find $\eta^* = B_\theta/\sqrt{2L^2n^3 + (d\sigma^2 + L^2 + \kappa_1\overline{K}B_x^2/2)n^2 + 2L^2n}$ minimizes the r.h.s, and gives

$$\mathcal{R}(\mathsf{SGD};\mathbb{P}) \leq B_\theta L\sqrt{\frac{2}{n} + \frac{1 + 8d\ln(1/\delta)/\varepsilon^2 + \kappa_1\overline{K}B_x^2/2}{n^2} + \frac{2}{n^3}} \lesssim \max\left\{\frac{1}{\sqrt{n}}, \frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon}\right\}.$$

Lastly, we note that the above reasoning can be applied to any distribution $\mathbb{P}$ over $\mathcal{X}\times\mathcal{Y}$, which completes the proof. □

A.4. Proof of Theorem 3.5

Proof. The Theorem directly follows from Kifer et al. (2012, Theorem 2). According to Kifer et al. (2012, Theorem 2), the minimizer $\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\boldsymbol{b})$ will satisfy (ε,δ)-DP if[1] (1) Hessian matrix $\nabla^2\widehat{\mathcal{L}}_h$ is continuous and at most rank-1 (2) loss function ℓ_h

[1] conditions are rephrased to be consistent with our context

is L-Liptschiz continuous and β-smooth (3) noise vector $\boldsymbol{b}$ is sampled from a multivariate Gaussian $\mathcal{N}(0, \sigma^2 \mathbf{I}_{d\times d})$ with variance $\sigma^2 \geq L^2 \left(8\ln(2/\delta) + 4\varepsilon\right)/\varepsilon^2$; (4) and the regularization coefficient $\lambda \geq \beta/(n\varepsilon)$. It is easy to check that all four conditions are satisfied in our settings. □

A.5. Proof of Theorem 3.6

Before proceeding, we summarize all notations for analyzing objective perturbation algorithm in Table 3.

Table 3. Notations for Objective Perturbation

Type	Description	Abbr.	Functions	Minimizers
Empirical	ERM	$\widehat{\mathcal{L}}(\boldsymbol{\theta})$	$\frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)$	$\widehat{\boldsymbol{\theta}}$
	Regularized ERM	$\widehat{\mathcal{L}}^{\#}(\boldsymbol{\theta})$	$\frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) + \lambda \left\|\boldsymbol{\theta}\right\|_2^2$	$\widehat{\boldsymbol{\theta}}^{\#}$
	Private regularized ERM	$\widehat{\mathcal{L}}^{\mathsf{OP}}(\boldsymbol{\theta}; \boldsymbol{b})$	$\frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) + \lambda \left\|\boldsymbol{\theta}\right\|_2^2 + \frac{\boldsymbol{b}^\top \boldsymbol{\theta}}{n}$	$\widehat{\boldsymbol{\theta}}^{\mathsf{OP}}$
Smoothed	ERM	$\widehat{\mathcal{L}}_h(\boldsymbol{\theta})$	$\frac{1}{n}\sum_{i=1}^n \ell_h(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)$	$\widehat{\boldsymbol{\theta}}_h$
	Regularized ERM	$\widehat{\mathcal{L}}_h^{\#}(\boldsymbol{\theta})$	$\frac{1}{n}\sum_{i=1}^n \ell_h(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) + \lambda \left\|\boldsymbol{\theta}\right\|_2^2$	$\widehat{\boldsymbol{\theta}}_h^{\#}$
	Private regularized ERM	$\widehat{\mathcal{L}}_h^{\mathsf{OP}}(\boldsymbol{\theta}; \boldsymbol{b})$	$\frac{1}{n}\sum_{i=1}^n \ell_h(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) + \lambda \left\|\boldsymbol{\theta}\right\|_2^2 + \frac{\boldsymbol{b}^\top \boldsymbol{\theta}}{n}$	$\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}$

Proof. • We first prove part 1 in the Theorem. It suffices to separately bound excess generalization risk's three terms:

$$\mathcal{R}(\mathsf{OP}; \mathbb{P}) \leq \mathbb{E}_{\mathcal{D},\mathsf{OP}}\left[\mathcal{L}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}})\right] + \mathbb{E}_{\mathcal{D},\mathsf{OP}}\left[\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*)\right] + \frac{1}{2}h\kappa_1. \tag{5}$$

In following analysis, we notationally suppress the dependences unless explicitly manipulating $\mathcal{D}$ and $\boldsymbol{b}$.

1. Because L-Lipschitz continuity of ℓ implies $\ell(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D})) - \ell(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D}')) \leq L\left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D}) - \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D}')\right\|_2$, it suffices to control the distance between minimizers trained on two neighboring datasets $\mathcal{D} \sim \mathcal{D}'$. By classic stability analysis for strongly convex optimization problem with a regularizer $\left\|\cdot\right\|_2^2$ (Bousquet & Elisseeff, 2002), we have $\left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D}) - \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\mathcal{D}')\right\|_2 \leq \frac{2L}{n\lambda}$, $\forall \mathcal{D} \sim \mathcal{D}', \boldsymbol{b}$. Therefore, Algorithm OP is $\frac{2L^2}{n\lambda}$-uniform stable w.r.t. ℓ. By Lemma 2.6, we know that
$$\mathbb{E}_{\mathcal{D},\mathsf{OP}}\left[\mathcal{L}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}})\right] \leq \frac{2L^2}{n\lambda}. \tag{6}$$

2. We now fix a dataset $\mathcal{D}$ and a noise vector $\boldsymbol{b}$, and come to bound $\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*)$. By strong convexity of $\widehat{\mathcal{L}}_h^{\mathsf{OP}}$ and Cauchy's inequality, we have
$$\lambda\left\|\widehat{\boldsymbol{\theta}}_h^{\#} - \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}\right\|_2^2 \leq \widehat{\mathcal{L}}_h^{\mathsf{OP}}(\widehat{\boldsymbol{\theta}}_h^{\#}) - \widehat{\mathcal{L}}_h^{\mathsf{OP}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) = \widehat{\mathcal{L}}_h^{\#}(\widehat{\boldsymbol{\theta}}_h^{\#}) - \widehat{\mathcal{L}}_h^{\#}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) + \frac{\boldsymbol{b}^\top \widehat{\boldsymbol{\theta}}_h^{\#}}{n} - \frac{\boldsymbol{b}^\top \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}}{n} \leq \frac{\left\|\boldsymbol{b}\right\|_2 \left\|\widehat{\boldsymbol{\theta}}_h^{\#} - \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}\right\|_2}{n},$$
which gives $\left\|\widehat{\boldsymbol{\theta}}_h^{\#} - \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}\right\|_2 \leq \frac{\left\|\boldsymbol{b}\right\|_2}{n\lambda}$. We further notice that
$$\begin{aligned}
\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*) &\leq \widehat{\mathcal{L}}_h^{\#}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}_h^{\#}(\widehat{\boldsymbol{\theta}}_h^{\#}) + \lambda\left(\left\|\boldsymbol{\theta}^*\right\|_2^2 - \left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}\right\|_2^2\right) \\
&\leq \left[\widehat{\mathcal{L}}_h^{\mathsf{OP}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \frac{\boldsymbol{b}^\top \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}}{n}\right] - \left[\widehat{\mathcal{L}}_h^{\mathsf{OP}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \frac{\boldsymbol{b}^\top \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}}{n}\right] + \lambda\left\|\boldsymbol{\theta}^*\right\|_2^2 \\
&\leq \frac{\left\|\boldsymbol{b}\right\|_2 \cdot \left\|\widehat{\boldsymbol{\theta}}_h^{\#} - \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\boldsymbol{b})\right\|_2}{n} + \lambda\left\|\boldsymbol{\theta}^*\right\|_2^2 \\
&\leq \frac{\left\|\boldsymbol{b}\right\|_2^2}{n^2\lambda} + \lambda\left\|\boldsymbol{\theta}^*\right\|_2^2,
\end{aligned}$$

where the first inequality is because $\widehat{\boldsymbol{\theta}}_h^{\#}$ is the minimizer to $\widehat{\mathcal{L}}_h^{\#}$; the second inequality is by plugging in minimizer $\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}$ into the second bracket, and dropping a negative term $-\lambda \left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}\right\|_2^2$. The reasoning here holds for any $\mathcal{D}$ and $\boldsymbol{b}$, and therefore we have

$$\mathbb{E}_{\mathcal{D},\mathsf{OP}}\left[\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*)\right] \leq \frac{\mathbb{E}_{\boldsymbol{b}}\left[\|\boldsymbol{b}\|_2^2\right]}{n^2\lambda} + \lambda \|\boldsymbol{\theta}^*\|_2^2 \leq \frac{d\sigma^2}{n^2\lambda} + \lambda B_\theta^2. \tag{7}$$

Plugging (6) and (7) back into (5), we obtain

$$\mathcal{R}(\mathsf{OP};\mathbb{P}) \leq \frac{2L^2}{n\lambda} + \frac{d\sigma^2}{n^2\lambda} + \lambda B_\theta^2 + \frac{1}{2}h\kappa_2, \quad \forall \lambda > 0.$$

Lastly, setting $h = \frac{\overline{K}B_x^2}{\lambda n\varepsilon}$, and optimizing over λ, we get the optimal $\lambda^* = \frac{1}{B_\theta}\sqrt{\frac{2L^2}{n} + \frac{d\sigma^2}{n^2} + \frac{\kappa_1\overline{K}B_x^2}{2n\varepsilon}}$, and

$$\mathcal{R}(\mathsf{OP};\mathbb{P}) \leq 2B_\theta\sqrt{\frac{2L^2}{n} + \frac{dL^2(8\ln(2/\delta)+4\varepsilon)}{n^2\varepsilon^2} + \frac{\kappa_1\overline{K}B_x^2}{2n\varepsilon}} \lesssim \max\left\{\frac{1}{\sqrt{n\varepsilon}}, \frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon}\right\},$$

which completes the proof for part 1.

- As to part 2, we decompose the excess generalization risk of OP in a different way:

$$\begin{aligned}
\mathcal{R}(\mathsf{OP};\mathbb{P}) &= \mathbb{E}_{\mathcal{D},\mathsf{OP}}\left[\mathcal{L}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}})\right] + \mathbb{E}_{\mathcal{D},\mathsf{OP}}\left[\widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}(\boldsymbol{\theta}^*)\right], && \text{(by (1))}\\
&\leq \mathbb{E}_{\mathcal{D},\mathsf{OP}}\left[\mathcal{L}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}})\right] + \mathbb{E}_{\mathcal{D},\mathsf{OP}}\left[\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}(\boldsymbol{\theta}^*)\right], && \text{(since } \ell_h \geq \ell\text{)}\\
&= \mathbb{E}_{\mathcal{D},\mathsf{OP}}\left[\mathcal{L}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}})\right] + \mathbb{E}_{\mathcal{D},\mathsf{OP}}\left[\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*)\right] + \mathbb{E}_{\mathcal{D}}\left[\widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*) - \widehat{\mathcal{L}}(\boldsymbol{\theta}^*)\right]\\
&= \mathbb{E}_{\mathcal{D},\mathsf{OP}}\left[\mathcal{L}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}})\right] + \mathbb{E}_{\mathcal{D},\mathsf{OP}}\left[\widehat{\mathcal{L}}_h(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}) - \widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*)\right] + \mathcal{L}_h(\boldsymbol{\theta}^*) - \mathcal{L}(\boldsymbol{\theta}^*). && (8)
\end{aligned}$$

The first and second terms are well-studied above. We only need to focus on the third term. Note that $\mathcal{L}_h(\boldsymbol{\theta}^*) - \mathcal{L}(\boldsymbol{\theta}^*) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}}\left[c_h(y - \boldsymbol{\theta}^{*\top}\boldsymbol{x}) - c(y - \boldsymbol{\theta}^{*\top}\boldsymbol{x})\right] =: \mathbb{E}_{\boldsymbol{x},\epsilon(\boldsymbol{x})}\left[g_h(\epsilon(\boldsymbol{x}))\right]$. From the proof of part 3 of Lemma 3.1, we know function $g_h(u)$ has a closed-form:

$$\begin{aligned}
g_h(u) &= h\int_{|u|/h}^{\infty} vK(v)\,dv + \frac{1}{2}\left(u \cdot \int_{-u/h}^{u/h} K(v)\,dv - |u|\right), \quad \forall u \in \mathbb{R}\\
&\leq h\int_{|u|/h}^{\infty} vK(v)\,dv && \text{(since negative in parenthese)}\\
&\leq h\int_{|u|/h}^{\infty} \frac{1}{v^2}\,dv && (K(\cdot)\text{ light tail \& } h \text{ small enough})\\
&\leq \frac{h^2}{|u|}.
\end{aligned}$$

As a result, the difference $\mathcal{L}_h(\boldsymbol{\theta}^*) - \mathcal{L}(\boldsymbol{\theta}^*)$ is upper bounded as

$$\mathcal{L}_h(\boldsymbol{\theta}^*) - \mathcal{L}(\boldsymbol{\theta}^*) \leq h^2 \cdot \mathbb{E}_{\boldsymbol{x},\epsilon(\boldsymbol{x})}\left[\frac{1}{|\epsilon(\boldsymbol{x})|}\right] =: h^2 M, \tag{9}$$

which is well-defined by our assumption. Substituting (6), (7), (9) back into (8) and setting $h = \frac{\overline{K}B_x^2}{\lambda n\varepsilon}$, we obtain:

$$\mathcal{R}(\mathsf{OP};\mathbb{P}) \leq \frac{2L^2}{n\lambda} + \frac{d\sigma^2}{n^2\lambda} + \lambda B_\theta^2 + \frac{\overline{K}^2 B_x^4 M}{\lambda^2 n^2\varepsilon^2}, \quad \forall \lambda > 0.$$

Lastly, set $\lambda = \frac{1}{B_\theta}\sqrt{\frac{2L^2}{n} + \frac{d\sigma^2}{n^2}}$, we get

$$\mathcal{R}(\mathsf{OP};\mathbb{P}) \leq 2B_\theta\sqrt{\frac{2L^2}{n} + \frac{d\sigma^2}{n^2}} + \frac{\overline{K}^2 B_x^4 M B_\theta^2}{2L^2 n\varepsilon^2 + d\sigma^2\varepsilon^2}.$$

To ensure the first term on the r.h.s. dominates the second term, it suffices to have $\sqrt{\frac{1}{n}+\frac{d\sigma^2}{n^2}} \geq \Omega\left(\frac{1}{n\varepsilon^2}\right)$. A sufficient condition to make it true is $\varepsilon^4 + d\ln(1/\delta)\varepsilon^2 \geq \Omega(\frac{1}{n})$. Consequently,

$$\mathcal{R}(\mathsf{OP};\mathbb{P}) \leq O\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon}\right),$$

if $\varepsilon^4 + d\ln(1/\delta)\varepsilon^2 \geq \Omega(\frac{1}{n})$. It is easy to check, under these parameters, bandwidth $h \to 0$ as $n \to \infty$, validating (9).

□

A.6. Proof of Theorem 3.7

Proof. The proof includes four steps. In first three steps, we gradually find out better lower bounds for the minimax risk of a (ε,δ)-DP algorithm $\mathsf{A} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{C}$, where $\mathcal{C} := \mathcal{B}(B_\theta)$. These lower bounds finally lead to a minimax risk of a (ε',δ')-DP d-dimensional classification problem. The sample complexity of DP classification problems for achieving a certain accuracy is well studied, from which we can derive the accuracy under a specific sample size. Consequently, the accuracy provides an overall lower bound.

1. **Step 1: relax the feasible region of the $\inf$ problem & restrict the feasible region of the $\sup$ problem**

   We first lower bound the minimax risk by relaxing the feasible region of the $\inf$ problem. Note that, in the minimax risk $\inf_{\mathsf{A}\in\mathcal{F}_{\varepsilon,\delta}} \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \mathcal{R}(\mathsf{A};\mathbb{P})$, the $\inf$ is taken over $\mathcal{F}_{\varepsilon,\delta}$, a set of mappings from $\mathcal{X}^n \times \mathcal{Y}^n$ to an Euclidean ball $\mathcal{C} = \mathcal{B}(B_\theta) := \left\{\boldsymbol{\theta}\in\mathbb{R}^d : \|\boldsymbol{\theta}\|_2 \leq B_\theta\right\}$. If we relax the output space from an Euclidean ball $\mathcal{C}$ to a superset $\mathcal{C}_1 := \left\{\boldsymbol{\theta}\in\mathbb{R}^d : |\theta_j| \leq B_\theta, \forall j = 1\ldots d\right\}$ a box set, the corresponding feasible set of mappings $\mathcal{F}^1_{\varepsilon,\delta}$, which includes all DP mappings from $\mathcal{X}^n \times \mathcal{Y}^n$ to $\mathcal{C}_1$, is therefore much larger. Thus, we get a lower bound by relaxing the feasible region of the $\inf$ problem:

   $$\inf_{\mathsf{A}\in\mathcal{F}_{\varepsilon,\delta}} \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \mathcal{R}(\mathsf{A};\mathbb{P}) \geq \inf_{\mathsf{A}\in\mathcal{F}^1_{\varepsilon,\delta}} \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \mathcal{R}(\mathsf{A};\mathbb{P}) \tag{10}$$

   We further discover a lower bound by restricting the feasible region of the $\sup$ problem. Specifically, we restrict $\mathbb{P}$ to a smaller set of probability measures $\mathbb{P}_\mathcal{S}$ where $\mathcal{S} \in \mathcal{X}^n \times \mathcal{Y}^n$ is a n-samples dataset with a specific structure, and $\mathbb{P}_\mathcal{S}$ is a distribution generated from the given dataset $\mathcal{S}$ by assigning each sample probability $1/n$:

   $$(10) \geq \inf_{\mathsf{A}\in\mathcal{F}^1_{\varepsilon,\delta}} \sup_{\mathbb{P}_\mathcal{S}} \mathcal{R}(\mathsf{A};\mathbb{P}_\mathcal{S})$$

   The inequality holds as $\mathcal{S}$ will be restricted to a specific structure to be elaborated later. As a result of that, $\mathbb{P}_\mathcal{S}$ is optimized over a more restrictive region, leading to a smaller objective value. Therefore, the r.h.s in above inequality is a lower bound.

   The dataset $\mathcal{S}$ is constructed as follows: let contextual matrix $\boldsymbol{X} := \{\boldsymbol{x}_i\}_{i=1}^n$ be an $n \times d$ matrix, where $\boldsymbol{x}_i \in \mathcal{X} := \{\frac{-B_x}{\sqrt{d}}, \frac{B_x}{\sqrt{d}}\}^d$. Denote the average vector $\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$, and its sign vector $sign(\overline{\boldsymbol{x}}) := (sign(\overline{x}_1), \ldots, sign(\overline{x}_d)) \in \{-1,1\}^d$. Then, the dataset $\mathcal{S}$ is constructed as

   $$\mathcal{S} = (\boldsymbol{X}\cdot sign(\overline{\boldsymbol{x}})\cdot B_\theta, \boldsymbol{X}).$$

   By our construction, dataset $\mathcal{S}$ depends only on contextual matrix $\boldsymbol{X}$, therefore, the $\sup_{\mathbb{P}_\mathcal{S}}$ is actually optimizing over $\boldsymbol{X}$. We will use $\mathcal{S}(\boldsymbol{X})$ to indicate such dependence when necessary. We conclude this step by writing out the lower bound explicitly for later reference:

   $$\begin{aligned}\inf_{\mathsf{A}\in\mathcal{F}_{\varepsilon,\delta}} \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \mathcal{R}(\mathsf{A};\mathbb{P}) &\geq \inf_{\mathsf{A}\in\mathcal{F}^1_{\varepsilon,\delta}} \sup_{\mathbb{P}_\mathcal{S}} \mathcal{R}(\mathsf{A};\mathbb{P}_\mathcal{S}) \\ &= \inf_{\mathsf{A}\in\mathcal{F}^1_{\varepsilon,\delta}} \sup_{\mathbb{P}_\mathcal{S}} \mathbb{E}_{\mathsf{A},\mathcal{D}\sim\mathbb{P}^n_\mathcal{S}}\left[\mathcal{L}(\mathsf{A}(\mathcal{D});\mathbb{P}_\mathcal{S})\right] - \min_{\boldsymbol{\theta}\in\mathcal{C}_1} \mathcal{L}(\boldsymbol{\theta};\mathbb{P}_\mathcal{S}),\end{aligned} \tag{11}$$

   where $\mathcal{R}(\mathsf{A};\mathbb{P}_\mathcal{S})$ is the excess generalization risk of A under distribution $\mathbb{P}_\mathcal{S}$.

2. **Step 2: reduce excess generalization risk to excess empirical risk**

   We notice that, since $\mathbb{P}_{\mathcal{S}}$ is an n-valued distribution generated from $\mathcal{S}$, for any $\boldsymbol{\theta}$, the population risk $\mathcal{L}(\boldsymbol{\theta};\mathbb{P}_{\mathcal{S}})$ w.r.t. $\mathbb{P}_{\mathcal{S}}$ is equal to an empirical risk $\widehat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{S})$ w.r.t. $\mathcal{S}$ by definitions of $\mathcal{L}$ and $\widehat{\mathcal{L}}$. Therefore,

$$\begin{aligned}(11) &= \inf_{\mathsf{A}\in\mathcal{F}^1_{\varepsilon,\delta}} \sup_{\mathbb{P}_{\mathcal{S}}} \mathbb{E}_{\mathsf{A},\mathcal{D}\sim\mathbb{P}^n_{\mathcal{S}}}\left[\widehat{\mathcal{L}}(\mathsf{A}(\mathcal{D});\mathcal{S})\right] - \min_{\boldsymbol{\theta}\in\mathcal{C}_1}\widehat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{S}) \\ &= \inf_{\mathsf{A}\in\mathcal{F}^1_{\varepsilon,\delta}} \sup_{\mathbb{P}_{\mathcal{S}}} \mathbb{E}_{\mathsf{A}}\left[\mathbb{E}_{\mathcal{D}\sim\mathbb{P}^n_{\mathcal{S}}}\left[\widehat{\mathcal{L}}(\mathsf{A}(\mathcal{D});\mathcal{S})\right]\right] - \min_{\boldsymbol{\theta}\in\mathcal{C}_1}\widehat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{S}) \qquad (12)\end{aligned}$$

   The inner expectation in the first term is taken over $\mathcal{D}\sim\mathbb{P}^n_{\mathcal{S}}$, that is, we i.i.d. draw n samples with replacement from a given $\mathcal{S}$. Thus, we can treat "$\mathsf{subsampling}$ $\mathcal{D}$ from $\mathcal{S}$, then run $\mathsf{A}(\mathcal{D})$" as a new algorithm B, which takes $\mathcal{S}$ as the input and outputs an estimator $\mathsf{A}(\mathcal{D})$. By a mild revision of notations only, we have

$$(12) = \inf_{\substack{\mathsf{B}:\ \mathsf{subsampling,\ then\ run\ A},\\ \text{where } \mathsf{A}\in\mathcal{F}^1_{\varepsilon,\delta}}} \sup_{\mathcal{S}} \mathbb{E}_{\mathsf{B}}\left[\widehat{\mathcal{L}}(\mathsf{B}(\mathcal{S});\mathcal{S})\right] - \min_{\boldsymbol{\theta}\in\mathcal{C}_1}\widehat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{S}). \qquad (13)$$

   It would be helpful to check whether algorithm B is DP. Following the definition of DP, we consider two neighboring dataset $\mathcal{T} := ((\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_k,y_k),\cdots,(\boldsymbol{x}_n,y_n))$ and $\mathcal{T}' := ((\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}'_k,y'_k),\cdots,(\boldsymbol{x}_n,y_n))$ that differ in k-th sample only. Datasets $\mathcal{T}$ and $\mathcal{T}'$ here not necessarily follow the specific structure in Step 1; instead, they are conceptual here for checking if B is DP only. Denote the set $\mathcal{I}\in\{1,\ldots,n\}^n$ as an index set with indices from i.i.d. $\mathsf{sampling}$ with replacement, and let $\mathcal{T}(\mathcal{I})$ be the resulting dataset with index set $\mathcal{I}$. Denote the number of different samples between $\mathcal{T}(\mathcal{I})$ and $\mathcal{T}'(\mathcal{I})$ as $\Delta(\mathcal{I}) := |\mathcal{T}(\mathcal{I})\backslash\mathcal{T}'(\mathcal{I})|$. As $\mathcal{T}$ and $\mathcal{T}'$ are neighboring and differ in k-th sample only, the value $\Delta(\mathcal{I})$ follows an n-trial Binomial distribution with success probability $1/n$; thus, it should be small with high probability. Specifically, we should have

$$\Pr_{\mathcal{I}}\left[\Delta(\mathcal{I})\geq z+1\right] = \Pr\left[\mathsf{Binomial}(n,1/n)\geq z+1\right] \leq \exp\left(-z^2/3\right), \quad \forall z>0,$$

   where the inequality follows from multiplicative Chernoff upper tail bound. Equivalently, the above implies that $\Delta(\mathcal{I})\geq 3\sqrt{\ln(1/\gamma)}+1 := u$ with probability at most γ. Now, we are ready to check if B is DP by definition: for any subset $\mathcal{U}$ of the output space of $\mathsf{B}(\mathcal{T})$, we have

$$\begin{aligned}\Pr_{\mathsf{B}}\left[\mathsf{B}(\mathcal{T})\in\mathcal{U}\right] &\leq \Pr_{\mathsf{B}|\mathcal{I}}\left[\mathsf{B}(\mathcal{T})\in\mathcal{U}|\Delta(\mathcal{I})\leq u\right]\cdot\Pr_{\mathcal{I}}\left[\Delta(\mathcal{I})\leq u\right]+\gamma \\ &= \Pr_{\mathsf{A}|\mathcal{I}}\left[\mathsf{A}(\mathcal{T}(\mathcal{I}))\in\mathcal{U}|\Delta(\mathcal{I})\leq u\right]\cdot\Pr_{\mathcal{I}}\left[\Delta(\mathcal{I})\leq u\right]+\gamma \\ &\leq \left(e^{\Delta(\mathcal{I})\cdot\varepsilon}\Pr_{\mathsf{A}|\mathcal{I}}\left[\mathsf{A}(\mathcal{T}'(\mathcal{I}))\in\mathcal{U}|\Delta(\mathcal{I})\leq u\right]+\Delta(\mathcal{I})\delta e^{\Delta(\mathcal{I})\varepsilon}\right)\cdot\Pr_{\mathcal{I}}\left[\Delta(\mathcal{I})\leq u\right]+\gamma \\ &\leq e^{u\varepsilon}\Pr_{\mathsf{B}|\mathcal{I}}\left[\mathsf{B}(\mathcal{T}')\in\mathcal{U}|\Delta(\mathcal{I})\leq u\right]\cdot\Pr_{\mathcal{I}}\left[\Delta(\mathcal{I})\leq u\right]+(u\delta e^{u\varepsilon}+\gamma) \\ &\leq e^{\varepsilon'}\Pr_{\mathsf{B}}\left[\mathsf{B}(\mathcal{T}')\in\mathcal{U}\right]+\delta',\end{aligned}$$

   where the third line follows from the fact that A is (ε,δ)-DP and Group Privacy Lemma (Vadhan 2017, Lemma 2.2). Therefore, the algorithm B is (ε',δ')-DP with $\varepsilon' := u\varepsilon$ and $\delta' := u\delta e^{u\varepsilon}+\gamma$. However, algorithm B is very restrictive as it must follow a "$\mathsf{subsampling}$, then run A" framework. If we remove the framework requirement and only require $\mathsf{B}\in\mathcal{F}^1_{\varepsilon',\delta'}$ a set of (ε',δ')-DP mappings from $\mathcal{X}^n\times\mathcal{Y}^n$ to $\mathcal{C}_1$, we will get a lower bound to (13):

$$\begin{aligned}(13) &= \inf_{\substack{\mathsf{B}\in\mathcal{F}^1_{\varepsilon',\delta'}:\ \mathsf{subsampling,\ then\ run\ A},\\ \text{where } \mathsf{A}\in\mathcal{F}^1_{\varepsilon,\delta}}} \sup_{\mathcal{S}} \mathbb{E}_{\mathsf{B}}\left[\widehat{\mathcal{L}}(\mathsf{B}(\mathcal{S});\mathcal{S})\right] - \min_{\boldsymbol{\theta}\in\mathcal{C}_1}\widehat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{S}) \\ &\geq \inf_{\mathsf{B}\in\mathcal{F}^1_{\varepsilon',\delta'}} \sup_{\mathcal{S}} \mathbb{E}_{\mathsf{B}}\left[\widehat{\mathcal{L}}(\mathsf{B}(\mathcal{S});\mathcal{S})\right] - \min_{\boldsymbol{\theta}\in\mathcal{C}_1}\widehat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{S}). \qquad (14)\end{aligned}$$

3. **step 3: convert excess empirical risk to DP binary classification error**

   Now, we start to analyze the excess empirical risk $\mathbb{E}_{\mathsf{B}}\left[\widehat{\mathcal{L}}(\mathsf{B}(\mathcal{S});\mathcal{S})\right] - \min_{\boldsymbol{\theta}\in\mathcal{C}_1}\widehat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{S})$ for any given dataset $\mathcal{S}$. Recall that the dataset $\mathcal{S}$ is constructed as $(\boldsymbol{X}\cdot sign(\overline{\boldsymbol{x}})\cdot B_\theta,\boldsymbol{X})$ in Step 1. Hence, the empirical minimizer is

$\widehat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\mathcal{C}_1} \widehat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{S}) = sign(\bar{\boldsymbol{x}})\cdot B_\theta \in \mathcal{C}_1$, and the empirical risk $\widehat{\mathcal{L}}(\widehat{\boldsymbol{\theta}};\mathcal{S}) = 0$. We therefore only need to focus our attention on $\mathbb{E}_{\mathsf{B}}\left[\widehat{\mathcal{L}}(\mathsf{B}(\mathcal{S});\mathcal{S})\right]$. It is straightforward to show

$$\begin{aligned}
\mathbb{E}_{\mathsf{B}}\left[\widehat{\mathcal{L}}(\mathsf{B}(\mathcal{S});\mathcal{S})\right] &= \mathbb{E}_{\mathsf{B}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(r\cdot(y_i - \mathsf{B}(\mathcal{S})^\top \boldsymbol{x}_i)^+ + (1-r)\cdot(\mathsf{B}(\mathcal{S})^\top\boldsymbol{x}_i - y_i)^+\right)\right]\\
&\geq \min\{r,1-r\}\mathbb{E}_{\mathsf{B}}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\left(sign(\bar{\boldsymbol{x}})\cdot B_\theta - \mathsf{B}(\mathcal{S})\right)^\top \boldsymbol{x}_i\right|\right]\\
&\geq \min\{r,1-r\}\mathbb{E}_{\mathsf{B}}\left[\left|(sign(\bar{\boldsymbol{x}})\cdot B_\theta - \mathsf{B}(\mathcal{S}))^\top\bar{\boldsymbol{x}}\right|\right] && \text{(by triangular inequality)}
\end{aligned}$$

By simple algebra, the absolute value in the expectation operator is

$$\begin{aligned}
\left|(sign(\bar{\boldsymbol{x}})\cdot B_\theta - \mathsf{B}(\mathcal{S}))^\top\bar{\boldsymbol{x}}\right| &= \left|\sum_{j=1}^{d}\left[sign(\bar{\boldsymbol{x}})_j\cdot B_\theta - \mathsf{B}(\mathcal{S})_j\right]\cdot sign(\bar{\boldsymbol{x}})_j\,|\bar{x}_j|\right|\\
&= \left|\sum_{j=1}^{d}|\bar{x}_j|\left(B_\theta - \mathsf{B}(\mathcal{S})_j\cdot sign(\bar{\boldsymbol{x}})_j\right)\right|\\
&= \sum_{j=1}^{d}|\bar{x}_j|\left(B_\theta - \mathsf{B}(\mathcal{S})_j\cdot sign(\bar{\boldsymbol{x}})_j\right) && \text{(Since } |\mathsf{B}(\mathcal{S})_j| \leq B_\theta\text{)}\\
&\geq B_\theta\cdot\sum_{j=1}^{d}|\bar{x}_j|\,\mathbb{1}\left\{sign(\bar{\boldsymbol{x}})_j \neq sign(\mathsf{B}(\mathcal{S})_j)\right\}.
\end{aligned}$$

Putting above analysis together, we further obtain a lower bound:

$$\begin{aligned}
(14) &\geq \min\{r,1-r\}B_\theta\cdot \inf_{\mathsf{B}\in\mathcal{F}^1_{\varepsilon',\delta'}}\sup_{\mathcal{S}}\mathbb{E}_{\mathsf{B}}\left[\sum_{j=1}^{d}|\bar{x}_j|\,\mathbb{1}\left\{sign(\bar{\boldsymbol{x}})_j \neq sign(\mathsf{B}(\mathcal{S})_j)\right\}\right]\\
&= \min\{r,1-r\}B_\theta\cdot \inf_{\mathsf{B}\in\mathcal{F}^1_{\varepsilon',\delta'}}\sup_{\boldsymbol{X}\in\mathcal{X}^n}\mathbb{E}_{\mathsf{B}}\left[\sum_{j=1}^{d}|\bar{x}_j|\,\mathbb{1}\left\{sign(\bar{\boldsymbol{x}})_j \neq sign(\mathsf{B}(\mathcal{S}(\boldsymbol{X}))_j)\right\}\right]\\
&\geq \min\{r,1-r\}B_\theta\cdot \inf_{\substack{\mathsf{C}:\mathcal{X}^n\to\{-1,1\}^d;\\ \mathsf{C}\text{ is }(\varepsilon',\delta')\text{-DP}}}\sup_{\boldsymbol{X}\in\mathcal{X}^n}\mathbb{E}_{\mathsf{C}}\left[\sum_{j=1}^{d}|\bar{x}_j|\,\mathbb{1}\left\{sign(\bar{\boldsymbol{x}})_j \neq \mathsf{C}(\boldsymbol{X})_j\right\}\right], \qquad (15)
\end{aligned}$$

where the second line is due to the dependence between $\mathcal{S}$ and $\boldsymbol{X}$ that $\mathcal{S}$ does not provide more information than $\mathcal{X}$; the third line follows a similar idea to (14), and $\mathsf{C}:\mathcal{X}^n\to\{-1,1\}^d$ is a DP binary classification algorithm. The expectation term in (15) can be thought of as the worst-case expected error of estimating the $sign$ vector of the average vector of $\boldsymbol{X}$ when using a DP binary classification algorithm C.

4. **step 4: bounding the classification error with sample complexity**

To facilitate further analysis, we denote $\mathsf{Error}(\mathsf{C},\boldsymbol{X}) := \mathbb{E}_{\mathsf{C}}\left[\sum_{j=1}^{d}|\bar{x}_j|\,\mathbb{1}\left\{sign(\bar{\boldsymbol{x}})_j \neq \mathsf{C}(\boldsymbol{X})_j\right\}\right]$ as the expected error of a DP binary classification algorithm C for a given $n\times d$ matrix $\boldsymbol{X}$. Denote the smallest number of samples to achieve a certain minimax risk $\alpha > 0$ as the sample complexity:

$$\mathsf{S}(\alpha,\varepsilon,\delta) := \min\{n : \inf_{\mathsf{C}\in\mathcal{F}_{\varepsilon,\delta},\text{classifier }\mathsf{C}}\ \sup_{\boldsymbol{X}\in\mathcal{X}^n}\mathsf{Error}(\mathsf{C},\boldsymbol{X}) \leq \alpha\}$$

By Proposition 1 and Lemma D.2 in (Asi et al., 2021), for ε',δ' defined in step 3, the sample complexity satisfies

$$\mathsf{S}(\alpha,\varepsilon',\delta') \geq \Omega\left(\frac{\sqrt{d}}{\alpha\varepsilon'\ln d}\right).$$

The lower bound on sample complexity implies a lower bound on the minimax risk:

$$\inf_{\mathsf{C}\in\mathcal{F}_{\varepsilon',\delta'},\text{classifier }\mathsf{C}}\ \sup_{\boldsymbol{X}\in\mathcal{X}^n} \mathsf{Error}(\mathsf{C},X) \geq \Omega\left(\frac{\sqrt{d}}{n\varepsilon'\ln d}\right) = \Omega\left(\frac{\sqrt{d}}{n\varepsilon\ln d}\right).$$

Combing preceding four steps, we obtain

$$\inf_{\mathsf{A}\in\mathcal{F}_{\varepsilon,\delta}}\ \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \mathcal{R}(\mathsf{A};\mathbb{P}) \geq \widetilde{\Omega}\left(\frac{\sqrt{d}}{n\varepsilon}\right), \tag{16}$$

where a logarithmic factor $\ln d$ is hidden. Moreover, since the oracle complexity of stochastic convex optimization is $\Omega(\frac{1}{\sqrt{n}})$, the minimax risk is therefore further lower bounded by the maximum between oracle complexity and (16):

$$\inf_{\mathsf{A}\in\mathcal{F}_{\varepsilon,\delta}}\ \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \mathcal{R}(\mathsf{A};\mathbb{P}) \geq \widetilde{\Omega}\left(\max\left(\frac{1}{\sqrt{n}},\frac{\sqrt{d}}{n\varepsilon}\right)\right),$$

which is the desired bound.

□

B. Proofs for Section 4

B.1. Proof of Lemma 4.2

Proof. We first compute the following expected value conditional on $\boldsymbol{x}$,

$$\begin{aligned}
\mathbb{E}_{\epsilon|\boldsymbol{x}}\left[\mathcal{K}(-\epsilon/h)\middle|\boldsymbol{x}\right] &= \int_{-\infty}^{\infty}\mathcal{K}(-t/h)\,dF_{\epsilon|\boldsymbol{x}}(t) \\
&= \frac{1}{h}\int_{-\infty}^{\infty}F_{\epsilon|\boldsymbol{x}}(t)K(-t/h)\,dt && \text{(integration by parts)} \\
&= \int_{-\infty}^{\infty}F_{\epsilon|\boldsymbol{x}}(-uh)K(u)\,du && (\text{let } u=-t/h) \\
&= r+\int_{-\infty}^{\infty}\left[F_{\epsilon|\boldsymbol{x}}(-uh)-F_{\epsilon|\boldsymbol{x}}(0)\right]K(u)\,du && \left(F_{\epsilon|\boldsymbol{x}}(0)=r \text{ and } \int K=1\right) \\
&= r+\int_{-\infty}^{\infty}K(u)\left[\int_0^{-uh}f_{\epsilon|\boldsymbol{x}}(t)-f_{\epsilon|\boldsymbol{x}}(0)\,dt\right]du \\
&\leq r+\int_{-\infty}^{\infty}K(u)\int_0^{|-uh|}l_1 t\,dt\,du && (f_{\epsilon|\boldsymbol{x}} \text{ is } l_1\text{-Lipschitz}) \\
&= r+\frac{1}{2}h^2 l_1\underbrace{\int_{-\infty}^{\infty}u^2K(u)\,du}_{=:\kappa_2}.
\end{aligned} \tag{17}$$

By above calculation, we can upper bound $\nabla\mathcal{L}_h(\boldsymbol{\theta}^*)$ as below:

$$\begin{aligned}
\nabla\mathcal{L}_h(\boldsymbol{\theta}^*) &= \mathbb{E}_{y,\boldsymbol{x}}\left[\left[\mathcal{K}_h(\boldsymbol{\theta}^{*\top}\boldsymbol{x}-y)-r\right]\boldsymbol{x}\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\epsilon|\boldsymbol{x}}\left[\mathcal{K}(-\epsilon/h)-r\middle|\boldsymbol{x}\right]\cdot\boldsymbol{x}\right] \\
&\leq \frac{1}{2}h^2 l_1\kappa_2\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}]. && \text{(by (17))}
\end{aligned} \tag{18}$$

Since we assume $\mathbb{E}[\boldsymbol{x}]=[1,0,\ldots,0]^\top$, naturally we have

$$\|\nabla\mathcal{L}_h(\boldsymbol{\theta}^*)\|_2 \leq \frac{1}{2}h^2 l_1\kappa_2. \tag{19}$$

Thus, we can upper bound the following value

$$\begin{aligned}\langle \nabla\mathcal{L}_h(\boldsymbol{\theta}_h^*) - \nabla\mathcal{L}_h(\boldsymbol{\theta}^*), \boldsymbol{\theta}_h^* - \boldsymbol{\theta}^*\rangle &\le \|\nabla\mathcal{L}_h(\boldsymbol{\theta}^*)\|_2 \cdot \|\boldsymbol{\theta}_h^* - \boldsymbol{\theta}^*\|_2 && (\text{since } \nabla\mathcal{L}_h(\boldsymbol{\theta}_h^*) = \mathbf{0})\\ &\le \frac{1}{2}h^2 l_1\kappa_2 \cdot \|\boldsymbol{\theta}_h^* - \boldsymbol{\theta}^*\|_2 . && (20)\end{aligned}$$

In addition, we can lower bound the left-hand-side of (20), but before showing that, we first lower bound the Hessian matrix $\nabla^2\mathcal{L}_h(\boldsymbol{\theta}), \forall\boldsymbol{\theta}$.

It can be shown that, for a given $\boldsymbol{\theta}$,

$$\begin{aligned}\mathbb{E}_{y|\boldsymbol{x}}\left[K_h(y - \boldsymbol{\theta}^\top\boldsymbol{x})\middle|\boldsymbol{x}\right] &= \int_{-\infty}^{\infty}\frac{1}{h}K\left(\frac{\boldsymbol{\theta}^{*\top}\boldsymbol{x} - \boldsymbol{\theta}^\top\boldsymbol{x} - t}{h}\right)dF_{\epsilon|\boldsymbol{x}}(t)\\ &= \int_{-\infty}^{\infty}K(u)f_{\epsilon|\boldsymbol{x}}(-\boldsymbol{\delta}^\top\boldsymbol{x} - uh)\,du && (\boldsymbol{\delta} := \boldsymbol{\theta} - \boldsymbol{\theta}^*;\ \text{let } u = \frac{-\boldsymbol{\delta}^\top\boldsymbol{x} - t}{h})\\ &\ge \int_{-\infty}^{\infty}K(u)\left(f_{\epsilon|\boldsymbol{x}}(0) - l_1\left|\boldsymbol{\delta}^\top\boldsymbol{x} + uh\right|\right)du && (f_{\epsilon|\boldsymbol{x}} \text{ is } l_1\text{-Lipschitz})\\ &\ge \underline{f} - hl_1\underbrace{\int_{-\infty}^{\infty}|u|\,K(u)\,du}_{=:\kappa_1} - l_1\left|\boldsymbol{\delta}^\top\boldsymbol{x}\right| .\end{aligned}$$

Therefore, the Hessian matrix of function $\mathcal{L}_h$ at point $\boldsymbol{\theta} \in \mathbb{R}^d$ satisfies

$$\nabla^2\mathcal{L}_h(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{y|\boldsymbol{x}}\left[K_h(y - \boldsymbol{\theta}^\top\boldsymbol{x})\middle|\boldsymbol{x}\right]\cdot\boldsymbol{x}\boldsymbol{x}^\top\right] \succcurlyeq \mathbb{E}_{\boldsymbol{x}}\left[\left(\underline{f} - hl_1\kappa_1 - l_1\left|\boldsymbol{\delta}^\top\boldsymbol{x}\right|\right)\cdot\boldsymbol{x}\boldsymbol{x}^\top\right],$$

where $\boldsymbol{\delta} := \boldsymbol{\theta} - \boldsymbol{\theta}^*$ is a vector that originates from $\boldsymbol{\theta}^*$ to $\boldsymbol{\theta}$. By mean value theorem for vector-valued functions, we know that ,

$$\nabla\mathcal{L}_h(\boldsymbol{\theta}) - \nabla\mathcal{L}_h(\boldsymbol{\theta}^*) = \int_0^1 \nabla^2\mathcal{L}_h(\boldsymbol{\theta}^* + \lambda\boldsymbol{\delta})\,d\lambda\cdot\boldsymbol{\delta}, \quad \forall\boldsymbol{\theta}.$$

By plugging the expression of Hessian matrix $\nabla^2\mathcal{L}_h$, we are able to show that, for any finite $\boldsymbol{\delta} := \boldsymbol{\theta} - \boldsymbol{\theta}^*$,

$$\begin{aligned}\langle\nabla\mathcal{L}_h(\boldsymbol{\theta}) - \nabla\mathcal{L}_h(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle &= \boldsymbol{\delta}^\top\cdot\int_0^1\nabla^2\mathcal{L}_h(\boldsymbol{\theta}^* + \lambda\boldsymbol{\delta})\,d\lambda\cdot\boldsymbol{\delta}\\ &\ge \int_0^1\mathbb{E}_{\boldsymbol{x}}\left[\left(\underline{f} - hl_1\kappa_1 - l_1\left|\lambda\boldsymbol{\delta}^\top\boldsymbol{x}\right|\right)\cdot\left(\boldsymbol{\delta}^\top\boldsymbol{x}\right)^2\right]d\lambda\\ &= \mathbb{E}_{\boldsymbol{x}}\left[\int_0^1\left(\underline{f} - hl_1\kappa_1 - l_1\left|\lambda\boldsymbol{\delta}^\top\boldsymbol{x}\right|\right)d\lambda\cdot\left(\boldsymbol{\delta}^\top\boldsymbol{x}\right)^2\right]\\ &\ge \rho_1(\underline{f} - hl_1\kappa_1)\,\|\boldsymbol{\delta}\|_2^2 - \frac{1}{2}l_1B_x^3\,\|\boldsymbol{\delta}\|_2^3, && (21)\end{aligned}$$

where the switch of integrals in third line holds from Fubini's Theorem. Thus, if we consider $\boldsymbol{\delta}_h^* := \boldsymbol{\theta}_h^* - \boldsymbol{\theta}^*$, we get

$$\langle\nabla\mathcal{L}_h(\boldsymbol{\theta}_h^*) - \nabla\mathcal{L}_h(\boldsymbol{\theta}^*), \boldsymbol{\theta}_h^* - \boldsymbol{\theta}^*\rangle \ge \rho_1(\underline{f} - hl_1\kappa_1)\,\|\boldsymbol{\delta}_h^*\|_2^2 - \frac{1}{2}l_1B_x^3\,\|\boldsymbol{\delta}_h^*\|_2^3 . \tag{22}$$

Combining upper bound (20) and lower bound (22) gives

$$\underbrace{\frac{1}{2}l_1B_x^3}_{=:a>0}\|\boldsymbol{\delta}_h^*\|_2^3 - \underbrace{\rho_1(\underline{f} - hl_1\kappa_1)}_{=:b>0}\|\boldsymbol{\delta}_h^*\|_2^2 + \underbrace{\frac{1}{2}h^2l_1\kappa_2}_{=:c>0}\|\boldsymbol{\delta}_h^*\|_2 \ge 0,$$

solving which results in three solutions,

$$\|\boldsymbol{\delta}_h^*\|_2 \ge 0; \quad \text{or } \|\boldsymbol{\delta}_h^*\|_2 \le \frac{2c}{b + \Delta^{1/2}}; \quad \text{or } \|\boldsymbol{\delta}_h^*\|_2 \ge \frac{b + \Delta^{1/2}}{2a},$$

provided that $\Delta := b^2 - 4ac > 0$. We can rule out the third solution by contradiction. Suppose that the third solution is true, then $\|\boldsymbol{\delta}_h^*\|_2 > b/(2a) > \sqrt{c/a}$. Since $\|\boldsymbol{\delta}_h^*\|_2 > \sqrt{c/a}$, there exists $\alpha \in (0,1)$ such that $\alpha \cdot \|\boldsymbol{\delta}_h^*\|_2 = \sqrt{c/a}$. With this α, let us denote $\boldsymbol{\theta}_\alpha := (1-\alpha)\boldsymbol{\theta}^* + \alpha\boldsymbol{\theta}_h^*$. We notice that, by (20),

$$\langle -\nabla\mathcal{L}_h(\boldsymbol{\theta}^*), \boldsymbol{\theta}_\alpha - \boldsymbol{\theta}^*\rangle = \langle \nabla\mathcal{L}_h(\boldsymbol{\theta}^*), \alpha\boldsymbol{\delta}_h^*\rangle \le c\alpha \left\|\boldsymbol{\delta}_h^*\right\|_2 . \tag{23}$$

Moreover, since $\langle \nabla\mathcal{L}_h(\boldsymbol{\theta}_\alpha), \boldsymbol{\theta}_\alpha - \boldsymbol{\theta}^*\rangle \le 0$, we also have

$$\begin{aligned}\langle -\nabla\mathcal{L}_h(\boldsymbol{\theta}^*), \boldsymbol{\theta}_\alpha - \boldsymbol{\theta}^*\rangle &\ge \langle \nabla\mathcal{L}_h(\boldsymbol{\theta}_\alpha) - \nabla\mathcal{L}_h(\boldsymbol{\theta}^*), \boldsymbol{\theta}_\alpha - \boldsymbol{\theta}^*\rangle \\ &\ge b\left\|\boldsymbol{\theta}_\alpha - \boldsymbol{\theta}^*\right\|_2^2 - a\left\|\boldsymbol{\theta}_\alpha - \boldsymbol{\theta}^*\right\|_2^3 && \text{(by (21))} \\ &= \alpha^2 \left\|\boldsymbol{\delta}_h^*\right\|_2^2 \left(b - a\cdot\sqrt{c/a}\right) && (24)\end{aligned}$$

Combining (23) and (24) together and cancelling out $\alpha\left\|\boldsymbol{\delta}_h^*\right\|_2$ on both sides, we obtain

$$\alpha\left\|\boldsymbol{\delta}_h^*\right\|_2 \le \frac{c}{b - \sqrt{ac}}, \tag{25}$$

which is true without the need of changing inequality symbol as $\Delta > 0$ implying both sides are positive. However, we further have

$$\sqrt{c/a} = \alpha\left\|\boldsymbol{\delta}_h^*\right\|_2 \le \frac{c}{b-\sqrt{ac}} < \frac{c}{\sqrt{ac}} = \sqrt{c/a},$$

a contradiction. Therefore, we are safe to rule out the third solution. As a result,

$$0 \le \left\|\boldsymbol{\delta}_h^*\right\|_2 \le \frac{2c}{b + \Delta^{1/2}} \le \frac{2c}{b},$$

provided that $\Delta > 0$ and $b > 0$, for which a sufficient condition is $\underline{f} - hl_1(\kappa_1 + \sqrt{B_x^3\kappa_2/\rho_1}) > 0$. □

B.2. Proof of Lemma 4.3

Proof. By strong convexity of $\widehat{\mathcal{L}}_h^{\mathsf{OP}}$, for any fixed $\mathcal{D}$ and $\boldsymbol{b}$, we have

$$\begin{aligned}\lambda\left\|\widehat{\boldsymbol{\theta}}_h^{\#} - \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\boldsymbol{b})\right\|_2^2 &\le \widehat{\mathcal{L}}_h^{\mathsf{OP}}(\widehat{\boldsymbol{\theta}}_h^{\#}) - \widehat{\mathcal{L}}_h^{\mathsf{OP}}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\boldsymbol{b})) \\ &= \widehat{\mathcal{L}}_h^{\#}(\widehat{\boldsymbol{\theta}}_h^{\#}) - \widehat{\mathcal{L}}_h^{\#}(\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\boldsymbol{b})) + \frac{\boldsymbol{b}^\top\widehat{\boldsymbol{\theta}}_h^{\#}}{n} - \frac{\boldsymbol{b}^\top\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\boldsymbol{b})}{n} \\ &\le \frac{\|\boldsymbol{b}\|_2\left\|\widehat{\boldsymbol{\theta}}_h^{\#} - \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\boldsymbol{b})\right\|_2}{n},\end{aligned}$$

which implies $\left\|\widehat{\boldsymbol{\theta}}_h^{\#} - \widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\boldsymbol{b})\right\|_2 \le \|\boldsymbol{b}\|_2/(n\lambda)$. And further replacing $\lambda \asymp \frac{1}{B_\theta}\cdot\sqrt{\frac{\kappa_1\overline{K}B_x^2}{n\varepsilon} + \frac{d\sigma^2}{n^2}}$ and taking expectation over $\boldsymbol{b}$, we have

$$\mathbb{E}_{\boldsymbol{b}}\left[\left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\boldsymbol{b}) - \widehat{\boldsymbol{\theta}}_h^{\#}\right\|_2\right] \lesssim B_\theta \cdot \frac{\mathbb{E}_{\boldsymbol{b}}\left[\|\boldsymbol{b}\|_2\right]}{n\cdot\sqrt{\frac{\kappa_1\overline{K}B_x^2}{n\varepsilon} + \frac{d\sigma^2}{n^2}}} \lesssim \frac{LB_\theta}{B_x\sqrt{\kappa_1\overline{K}}}\cdot\sqrt{\frac{d\ln(1/\delta)}{n\varepsilon}}.$$

□

B.3. Proof of Lemma 4.4

Lemma B.1 (Lemma 9.21 in Wainwright (2019), rephrased)**.** *Denote* $\boldsymbol{\theta}^* := \arg\min_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$, $\widehat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}}\widehat{\mathcal{L}}(\boldsymbol{\theta})$. *For any given radius* r_0, *if* $\widehat{\mathcal{L}}(\boldsymbol{\theta}^* + \boldsymbol{\delta}) - \widehat{\mathcal{L}}(\boldsymbol{\theta}^*) > 0$, $\forall\boldsymbol{\delta} \in \{\boldsymbol{\delta}\in\mathbb{R}^d : \|\boldsymbol{\delta}\|_2 = r_0\}$, *then* $\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 \le r_0$.

Proof. For any $\boldsymbol{\delta} \in \mathbb{R}^d$ and $\boldsymbol{\theta} \in \mathbb{R}^d$, we define the population-level zero-order Taylor expansion remainder of $\mathcal{L}_h(\boldsymbol{\theta} + \boldsymbol{\delta})$ at point $\boldsymbol{\theta}$ as $D_h(\boldsymbol{\delta}, \boldsymbol{\theta}) = \mathcal{L}_h(\boldsymbol{\theta}+\boldsymbol{\delta}) - \mathcal{L}_h(\boldsymbol{\theta})$, the first-order remainder as $R_h(\boldsymbol{\delta}, \boldsymbol{\theta}) = D_h(\boldsymbol{\delta}, \boldsymbol{\theta}) - \nabla \mathcal{L}_h(\boldsymbol{\theta})^\top \boldsymbol{\delta}$. Correspondingly, the sample-level remainders are defined as $\widehat{D}_h(\boldsymbol{\delta}, \boldsymbol{\theta}) = \widehat{\mathcal{L}}_h(\boldsymbol{\theta} + \boldsymbol{\delta}) - \widehat{\mathcal{L}}_h(\boldsymbol{\theta})$ and $\widehat{R}_h(\boldsymbol{\delta}, \boldsymbol{\theta}) = \widehat{D}_h(\boldsymbol{\delta}, \boldsymbol{\theta}) - \nabla \widehat{\mathcal{L}}_h(\boldsymbol{\theta})^\top \boldsymbol{\delta}$. And similarly for regularized variants, $\widehat{D}^{\#}_h(\boldsymbol{\delta}, \boldsymbol{\theta}) = \widehat{\mathcal{L}}^{\#}_h(\boldsymbol{\theta} + \boldsymbol{\delta}) - \widehat{\mathcal{L}}^{\#}_h(\boldsymbol{\theta})$ and $\widehat{R}^{\#}_h(\boldsymbol{\delta}, \boldsymbol{\theta}) = \widehat{D}^{\#}_h(\boldsymbol{\delta}, \boldsymbol{\theta}) - \nabla \widehat{\mathcal{L}}^{\#}_h(\boldsymbol{\theta})^\top \boldsymbol{\delta}$.

With these notations, for any $\boldsymbol{\delta} \in \mathbb{R}^d$, we have

$$\begin{aligned}
\widehat{D}^{\#}_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*_h) &= \underbrace{R_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) + \nabla \mathcal{L}_h(\boldsymbol{\theta}^*)^\top \boldsymbol{\delta} - D_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*)}_{=0} + \widehat{D}^{\#}_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*_h) \\
&= R_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) + \nabla \mathcal{L}_h(\boldsymbol{\theta}^*)^\top \boldsymbol{\delta} - \left(D_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) - \widehat{D}^{\#}_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) \right) - \left(\widehat{D}^{\#}_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) - \widehat{D}^{\#}_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*_h) \right)
\end{aligned}$$

To find an upper bound r_0 on $\left\| \widehat{\boldsymbol{\theta}}^{\#}_h - \boldsymbol{\theta}^*_h \right\|_2$, by Lemma 9.21 in Wainwright (2019) (i.e. Lemma B.1), it suffices to show that the upper bound r_0 satisfies $\widehat{D}^{\#}_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*_h) > 0, \forall \boldsymbol{\delta} \in \partial \mathcal{B}(r_0) := \left\{ \boldsymbol{\delta} \in \mathbb{R}^d : \|\boldsymbol{\delta}\|_2 = r_0 \right\}$. Our proof strategy is to first lower bound $\widehat{D}^{\#}_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*_h)$ and show this lower bound is strictly greater than 0 for any $\boldsymbol{\delta} \in \partial \mathcal{B}(r_0)$ with our choice of r_0. Note that

$$\widehat{D}^{\#}_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*_h) \geq R_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) - \|\nabla \mathcal{L}_h(\boldsymbol{\theta}^*)\|_2 \|\boldsymbol{\delta}\|_2 - \left(D_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) - \widehat{D}^{\#}_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) \right) - \left(\widehat{D}^{\#}_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) - \widehat{D}^{\#}_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*_h) \right). \tag{26}$$

Hence, it suffices to lower bound first term, and upper bound last three terms. Keep $\boldsymbol{\delta} \in \partial \mathcal{B}(r_0)$ in mind, we address them one by one. For notational brevity, we use same notations as in Appendix B.1 and denote

$$a := \frac{1}{2} l_1 B_x^3, \quad b := \rho_1(\underline{f} - h l_1 \kappa_1), \quad \text{and } c := \frac{1}{2} h^2 l_1 \kappa_2.$$

Note that all three values are positive for finite but large enough n due to the setting $h \leq o(1)$.

1. Following the same argument for (21), we can show that
$$R_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) \geq \rho_1(\underline{f} - h l_1 \kappa_1) \|\boldsymbol{\delta}\|_2^2 - \frac{1}{2} l_1 B_x^3 \|\boldsymbol{\delta}\|_2^3 = -a r_0^3 + b r_0^2 \tag{27}$$

2. By (19), we know that $\|\nabla \mathcal{L}_h(\boldsymbol{\theta}^*)\|_2 \leq \frac{1}{2} h^2 l_1 \kappa_2$; thus,
$$\|\nabla \mathcal{L}_h(\boldsymbol{\theta}^*)\|_2 \|\boldsymbol{\delta}\|_2 \leq c r_0. \tag{28}$$

3. The third term $D_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) - \widehat{D}^{\#}_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*)$ is well-studied in Lemma B.2. Specifically, we know from Lemma B.2 that, for any given $r_0 \in (0, 2\|\boldsymbol{\theta}^*\|_2)$, and $\gamma \in (0,1)$, with high probability at least $1 - \gamma$,
$$\sup_{\boldsymbol{\delta} \in \partial \mathcal{B}(r_0)} \left\{ D_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) - \widehat{D}^{\#}_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) \right\} \leq 3 r_0 L \sqrt{\frac{4d + \ln(1/\gamma)}{n}} + 2\lambda r_0 \|\boldsymbol{\theta}^*\|_2 - \lambda r_0^2. \tag{29}$$

4. By L-Lipschitz continuity of $\widehat{\mathcal{L}}_h$, we can show that, for any $\boldsymbol{\delta} \in \partial \mathcal{B}(r_0)$,
$$\begin{aligned}
\widehat{D}^{\#}_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) - \widehat{D}^{\#}_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*_h) &= \left[\widehat{\mathcal{L}}_h(\boldsymbol{\theta}^* + \boldsymbol{\delta}) - \widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*_h + \boldsymbol{\delta}) \right] + \left[\widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*_h) - \widehat{\mathcal{L}}_h(\boldsymbol{\theta}^*) \right] \\
&\quad + \lambda \|\boldsymbol{\theta}^* + \boldsymbol{\delta}\|_2^2 - \lambda \|\boldsymbol{\theta}^*_h + \boldsymbol{\delta}\|_2^2 + \lambda \|\boldsymbol{\theta}^*_h\|_2^2 - \lambda \|\boldsymbol{\theta}^*\|_2^2 \\
&\leq 2L \cdot \|\boldsymbol{\theta}^*_h - \boldsymbol{\theta}^*\|_2 + 2\lambda (\boldsymbol{\theta}^* - \boldsymbol{\theta}^*_h)^\top \boldsymbol{\delta} \\
&\leq (2L + 2\lambda r_0) \|\boldsymbol{\theta}^* - \boldsymbol{\theta}^*_h\|_2 \\
&\leq (4L + 4\lambda r_0) \cdot c/b
\end{aligned} \tag{30}$$

Plugging (27), (28), (29), and (30) into (26) and replacing $\|\boldsymbol{\theta}^*\|_2$ with its upper bound B_θ, we obtain by rearranging that, for any $\gamma \in (0,1)$, with probability at least $1 - \gamma$, the following inequality holds for any $\boldsymbol{\delta} \in \partial \mathcal{B}(r_0)$,

$$\begin{aligned}
\widehat{D}^{\#}_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*_h) &\geq -a r_0^3 + (b + \lambda) r_0^2 - \left(3L \sqrt{\frac{4d + \ln(1/\gamma)}{n}} + c + 4\lambda c/b + 2\lambda B_\theta \right) r_0 - 4cL/b \\
&\geq -a r_0^3 + b r_0^2 - \left(3L \sqrt{\frac{4d + \ln(1/\gamma)}{n}} + 2\lambda B_\theta \right) r_0 - (4\lambda c r_0/b + c r_0 + 4cL/b).
\end{aligned} \tag{31}$$

Denote $C := 3L\sqrt{\frac{4d+\ln(1/\gamma)}{n}} + 2\lambda B_\theta$. We conjecture that $r_0 \asymp C/b$ is a valid radius. Please keep in mind that we require $h \le o(1)$ and $\lambda \le o(1)$, this implies $r_0 \searrow 0$ when n tends to infinity. Plugging $r_0 \asymp C/b$ into (31), we obtain that, by removing dominated terms r_0^3, $\lambda c r_0$ and $c r_0$,

$$\begin{aligned}(31) &\asymp \frac{C^2}{b} - \frac{cL}{b}\\ &\asymp \frac{1}{\rho_1 \underline{f}} \cdot (C^2 - h^2 l_1 \kappa_2 L)\\ &\asymp \frac{L}{\rho_1 \underline{f}} \cdot \left(L \cdot \frac{d + \ln(1/\gamma)}{n} + B_\theta^2 \lambda^2 - h^2 l_1 \kappa_2 \right) \gtrsim 0,\end{aligned}$$

where the last inequality is from our assumptions that $B_\theta^2 \lambda^2 \gtrsim h^2 l_1 \kappa_2$. Therefore, the radius

$$r_0 \asymp C/b = \frac{1}{\rho_1 \underline{f}} \cdot \left(L\sqrt{\frac{d+\ln(1/\gamma)}{n}} + h^2 l_1 \kappa_1 + \lambda B_\theta \right) \asymp \frac{1}{\rho_1 \underline{f}} \cdot \left(L\sqrt{\frac{d+\ln(1/\gamma)}{n}} + B_\theta \lambda \right)$$

is valid, which completes the proof. □

Lemma B.2. *Given any $n \ge 2$, $\gamma \in (0,1)$ and $r_0 \in (0, 2\|\boldsymbol{\theta}^*\|_2)$, with probability at least $1-\gamma$, the inequality*

$$\sup_{\boldsymbol{\delta} \in \partial\mathcal{B}(r_0)} \left\{ D_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) - \widehat{D}_h^{\#}(\boldsymbol{\delta}, \boldsymbol{\theta}^*) \right\} \le 3\bar{r} r_0 B_x \sqrt{\frac{4d + \ln(1/\gamma)}{n}} + 2\lambda r_0 \|\boldsymbol{\theta}^*\|_2 - \lambda r_0^2,$$

holds, where $D_h(\boldsymbol{\delta}, \boldsymbol{\theta}) := \mathcal{L}_h(\boldsymbol{\theta} + \boldsymbol{\delta}) - \mathcal{L}_h(\boldsymbol{\theta})$, $\widehat{D}_h^{\#}(\boldsymbol{\delta}, \boldsymbol{\theta}) := \widehat{\mathcal{L}}_h^{\#}(\boldsymbol{\theta} + \boldsymbol{\delta}) - \widehat{\mathcal{L}}_h^{\#}(\boldsymbol{\theta})$, $\partial\mathcal{B}(r_0) := \left\{ \boldsymbol{\delta} \in \mathbb{R}^d \middle| \|\boldsymbol{\delta}\|_2 = r_0 \right\}$, and $\bar{r} := \max\{1-r, r\}$.

Proof. We follow the same proof idea for Lemma C.1 in He et al. (2021) to complete our proof. For any given radius $r_0 > 0$, denote $\alpha_\epsilon(r_0) = \frac{n(1-\epsilon)}{2\bar{r} r_0}$, and define a new random variable

$$\Delta_\epsilon(r_0) := \alpha_\epsilon(r_0) \sup_{\boldsymbol{\delta} \in \partial\mathcal{B}(r_0)} \left\{ D_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) - \widehat{D}_h^{\#}(\boldsymbol{\delta}, \boldsymbol{\theta}^*) \right\},$$

parameterized by a constant $\epsilon \in (0,1)$ to be determined later. To investigate the tail performance of $\Delta_\epsilon(r_0)$, we apply Chernoff bound and obtain

$$\mathbb{P}\left[\Delta_\epsilon(r_0) \ge u\right] \le \inf_{k>0} \left\{ \exp\left(\ln^{\mathbb{E}_{\mathcal{D}}\left[e^{k\Delta_\epsilon(r_0)}\right]} - ku \right) \right\}, \ \forall u > 0. \tag{32}$$

Hence, we can first control the moment generating function (MGF) of $\Delta_\epsilon(r_0)$. Define $g(\boldsymbol{\delta}, \boldsymbol{\theta}^*) = \|\boldsymbol{\theta}^* + \boldsymbol{\delta}\|_2^2 - \|\boldsymbol{\theta}^*\|_2^2$, then the moment generating function is, for any $k > 0$,

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[e^{k\Delta_\epsilon(r_0)}\right] &= \mathbb{E}_{\mathcal{D}}\left[\exp\left(\frac{kn(1-\epsilon)}{2\bar{r} r_0} \sup_{\boldsymbol{\delta} \in \partial\mathcal{B}(r_0)} \left\{ D_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) - \widehat{D}_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) - \lambda g(\boldsymbol{\delta}, \boldsymbol{\theta}^*) \right\}\right)\right]\\ &\le \mathbb{E}_{\mathcal{D}}\left[\exp\left(\frac{kn(1-\epsilon)}{2\bar{r} r_0} \sup_{\boldsymbol{\delta} \in \partial\mathcal{B}(r_0)} \left\{ D_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) - \widehat{D}_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) \right\} - k\lambda\alpha_\epsilon(r_0) \inf_{\boldsymbol{\delta} \in \partial\mathcal{B}(r_0)} g(\boldsymbol{\delta}, \boldsymbol{\theta}^*)\right)\right]\\ &= \mathbb{E}_{\mathcal{D}}\left[\exp\left(\frac{kn(1-\epsilon)}{2\bar{r} r_0} \sup_{\boldsymbol{\delta} \in \partial\mathcal{B}(r_0)} \left\{ D_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) - \widehat{D}_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*) \right\}\right)\right] \cdot \exp\left(k\lambda\alpha_\epsilon(r_0)\beta(r_0)\right).\end{aligned} \tag{33}$$

The last line comes from the fact

$$\inf_{\boldsymbol{\delta} \in \partial\mathcal{B}(r_0)} g(\boldsymbol{\delta}, \boldsymbol{\theta}^*) = \inf_{\boldsymbol{\delta} \in \partial\mathcal{B}(r_0)} \|\boldsymbol{\delta}\|_2^2 + 2\boldsymbol{\delta}^\top \boldsymbol{\theta}^* = -(-r_0^2 + 2r_0 \|\boldsymbol{\theta}^*\|_2) \quad =: -\beta(r_0) \le 0,$$

where $-\beta(r_0) \leq 0$ is due to $r_0 \in (0, 2\left\|\boldsymbol{\theta}^*\right\|_2)$. By Rademacher Symmetrization Lemma (see, for example, Proposition 4.11 in Wainwright 2019) and noticing that the exponential function is convex and increasing, we can further upper bound (33) with Rademacher Complexity as shown below,

$$(33) \leq \mathbb{E}_{\mathcal{D}, z_i}\left[\exp\left(\frac{kn(1-\epsilon)}{\bar{r}r_0} \sup_{\boldsymbol{\delta}\in\partial\mathcal{B}(r_0)}\left\{\frac{1}{n}\sum_{i=1}^{n} z_i d_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*; \boldsymbol{x}_i, y_i)\right\}\right)\right] \cdot \exp\left(k\lambda\alpha_\epsilon(r_0)\beta(r_0)\right), \tag{34}$$

where $z_i, \forall i \in \{1, 2, \ldots, n\}$ are independent Rademacher random variables. Note that the function $d_h(\boldsymbol{\delta}, \boldsymbol{\theta}^*; \boldsymbol{x}_i, y_i) = \ell_h\left(\boldsymbol{\delta} + \boldsymbol{\theta}^*; \boldsymbol{x}_i, y_i\right) - \ell_h\left(\boldsymbol{\theta}^*; \boldsymbol{x}_i, y_i\right) = c_h\left(y_i - \boldsymbol{\theta}^{*\top}\boldsymbol{x_i} - \boldsymbol{\delta}^\top\boldsymbol{x_i}\right) - c_h\left(y_i - \boldsymbol{\theta}^{*\top}\boldsymbol{x_i}\right)$ is $\bar{r}$-lipschitz continuous in $\langle\boldsymbol{\delta}, \boldsymbol{x}_i\rangle$, and $d_h(\cdot) = 0$ when $\langle\boldsymbol{\delta}, \boldsymbol{x}_i\rangle = 0$. Therefore, by Ledoux-Talagrand contraction inequality (see Theorem 11.6 in Boucheron et al. 2013), we have

$$\begin{aligned}
(34) &\leq \mathbb{E}_{\mathcal{D}, z_i}\left[\exp\left(\frac{kn(1-\epsilon)}{r_0} \sup_{\boldsymbol{\delta}\in\partial\mathcal{B}(r_0)}\left\{\frac{1}{n}\sum_{i=1}^{n} z_i \left\langle\boldsymbol{\delta}, \boldsymbol{x}_i\right\rangle\right\}\right)\right] \cdot \exp\left(k\lambda\alpha_\epsilon(r_0)\beta(r_0)\right) \\
&= \mathbb{E}_{\mathcal{D}, z_i}\left[\exp\left(k(1-\epsilon)\left\|\sum_{i=1}^{n} z_i \boldsymbol{x}_i\right\|_2\right)\right] \cdot \exp\left(k\lambda\alpha_\epsilon(r_0)\beta(r_0)\right),
\end{aligned} \tag{35}$$

where the second line is obtained by cancelling out n and then plugging in the maximizer $\boldsymbol{\delta}^* = r_0 \cdot \frac{\sum_{i=1}^{n} z_i\boldsymbol{x}_i}{\left\|\sum_{i=1}^{n} z_i\boldsymbol{x_i}\right\|_2}$.

Denote $\boldsymbol{w} := \sum_{i=1}^{n} z_i\boldsymbol{x}_i$. Now, we focus on the ℓ_2-norm term $\left\|\boldsymbol{w}\right\|_2$.Let a set of d-dimensional points $\{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_{N_\epsilon}\}$ with cardinality N_ϵ be the ϵ-*covering* of a unit ball $\mathcal{B}$ with respect to ℓ_2 norm. Then, we know that there must exist a point $\boldsymbol{p}_j \in \mathcal{B}, j \in \{1, 2, \ldots, N_\epsilon\}$ such that

$$(1-\epsilon)\left\|\boldsymbol{w}\right\|_2 \leq \langle\boldsymbol{p}_j, \boldsymbol{w}\rangle;$$

otherwise, the unit ball $\mathcal{B}$ is not covered by the given ϵ-*covering* $\{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_{N_\epsilon}\}$. Therefore, we are able to further upper bound (35) with

$$\begin{aligned}
(35) &\leq \mathbb{E}_{\mathcal{D}, z_i}\left[\exp\left(k \cdot \max_{j\in\{1,\ldots,N_\epsilon\}} \langle\boldsymbol{p}_j, \boldsymbol{w}\rangle\right)\right] \cdot \exp\left(k\lambda\alpha_\epsilon(r_0)\beta(r_0)\right) \\
&\leq \exp\left(k\lambda\alpha_\epsilon(r_0)\beta(r_0)\right) \cdot \sum_{j=1}^{N_\epsilon} \mathbb{E}_{\mathcal{D}, z_i}\left[\exp\left(k \cdot \langle\boldsymbol{p}_j, \boldsymbol{w}\rangle\right)\right]
\end{aligned} \tag{36}$$

For any given $\boldsymbol{p}_j \in \mathcal{B}$, we note that $\mathbb{E}_{\mathcal{D}, z_i}\left[\exp\left(k \cdot \langle\boldsymbol{p}_j, \boldsymbol{w}\rangle\right)\right] = \mathbb{E}_{\mathcal{D}, z_i}\left[\exp\left(k\sum_{i=1}^{n} v_{i,j}\right)\right]$ is the MGF of the sum over n independent zero-mean random variable $v_{i,j} := \langle\boldsymbol{p}_j, z_i\boldsymbol{x}_i\rangle, \forall i = 1, \ldots, n$. Moreover, since we assume $\boldsymbol{x} \in \mathcal{B}(B_x)$, the random variable $v_{i,j}$ is bounded almost surely as $|v_{i,j}| \leq B_x$. Thus, $v_{i,j}$ belongs to $\mathsf{SubGaussian}(\sigma^2)$ with $\sigma^2 = B_x^2$, and

$$\mathbb{E}_{\mathcal{D}, z_i}\left[\exp\left(k \cdot \langle\boldsymbol{p}_j, \boldsymbol{w}\rangle\right)\right] \leq \exp\left(\frac{nk^2\sigma^2}{2}\right), \ \forall\boldsymbol{p_j} \in \mathcal{B}, \forall k > 0, \tag{37}$$

Plugging (37) into (36), and combining (33), (34), (35), and (36) all together, we finally obtain an upper bound for the MGF of $\Delta_\epsilon(r_0)$ as below,

$$\mathbb{E}_{\mathcal{D}}\left[e^{k\Delta_\epsilon(r_0)}\right] \leq \exp\left(k\lambda\alpha_\epsilon(r_0)\beta(r_0)\right) \cdot \sum_{j=1}^{N_\epsilon} \exp\left(\frac{nk^2\sigma^2}{2}\right), \forall\epsilon \in (0,1), \forall r_0 \in (0, 2\left\|\boldsymbol{\theta}^*\right\|_2), \forall k > 0, \tag{38}$$

with $\alpha_\epsilon(r_0) = \frac{n(1-\epsilon)}{2\bar{r}r_0} > 0$ and $\beta(r_0) = -r_0^2 + 2r_0\left\|\boldsymbol{\theta}^*\right\|_2 > 0$.

Replacing the MGF of $\Delta_\epsilon(r_0)$ in (32) with its upper bound (38), we obtain that, for any $u > 0$,

$$\begin{aligned}
\mathbb{P}\left[\Delta_\epsilon(r_0) \geq u\right] &\leq \inf_{k>0}\left\{\exp\left(k\lambda\alpha_\epsilon(r_0)\beta(r_0) + \ln^{N_\epsilon} + \frac{nk^2\sigma^2}{2} - ku\right)\right\} \\
&\leq \inf_{k>0}\left\{\exp\left(k\lambda\alpha_\epsilon(r_0)\beta(r_0) + d\ln\left(1 + \frac{2}{\epsilon}\right) + \frac{nk^2\sigma^2}{2} - ku\right)\right\}
\end{aligned} \tag{39}$$

where the second inequality is by the fact that the covering number N_ϵ is known to satisfy $N_\epsilon \le (1+\frac{2}{\epsilon})^d$. After replace u with $\alpha_\epsilon(r_0)\cdot\left(\frac{2\sqrt{2}\bar{r}r_0\sigma}{1-\epsilon}\sqrt{\frac{v}{n}}+\lambda\beta(r_0)\right)$ for any $v>0$, (39) becomes, for any $v>0$,

$$\begin{aligned}
&\mathbb{P}\left[\sup_{\boldsymbol{\delta}\in\mathbb{B}(r_0)}\left\{D_h(\boldsymbol{\delta},\boldsymbol{\theta}^*)-\widehat{D}_h^{\#}(\boldsymbol{\delta},\boldsymbol{\theta}^*)\right\}\ge\frac{2\sqrt{2}\bar{r}r_0\sigma}{1-\epsilon}\sqrt{\frac{v}{n}}+\lambda\beta(r_0)\right]\\
&\le\inf_{k>0}\left\{\exp\left(d\ln\left(1+\frac{2}{\epsilon}\right)+\frac{nk^2\sigma^2}{2}-k\sigma\sqrt{2vn}\right)\right\}\\
&=\exp\left(d\ln\left(1+\frac{2}{\epsilon}\right)-v\right),
\end{aligned}\tag{40}$$

where the last line is by taking $k^*=\frac{1}{\sigma}\cdot\sqrt{\frac{2v}{n}}$. Lastly, if we set $\epsilon=\frac{2}{e^4-1}$, then $\frac{2\sqrt{2}}{1-\epsilon}\le 3$ and the right-hand-side of (40) becomes $\exp(4d-v)$. Set $\exp(4d-v)=\gamma$ and solve for v, we get the desired lemma. □

B.4. Proof of Theorem 4.5

Proof. This Theorem is a direct conclusion from combing Lemma 4.2, 4.3, and 4.4. Specifically, if we set $\lambda\asymp\frac{1}{B_\theta}\cdot\sqrt{\frac{\kappa_1\overline{K}B_x^2}{n\varepsilon}+\frac{d\sigma^2}{n^2}}$ and $h=\overline{K}B_x^2/(\lambda n\varepsilon)$ as same values in Theorem 3.6, then we have:

1. By Lemma 4.2, we have
$$\|\boldsymbol{\theta}_h^*-\boldsymbol{\theta}^*\|_2\le\frac{h^2l_1\kappa_2}{\rho_1\left(\underline{f}-hl_1\kappa_1\right)}\lesssim\frac{1}{\rho_1\underline{f}}\cdot\frac{l_1\kappa_2\overline{K}B_x^2}{\kappa_1}\cdot\frac{1}{n\varepsilon}.\tag{41}$$

2. By Lemma 4.3, we have
$$\mathbb{E}_{\mathsf{OP}}\left[\left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}-\widehat{\boldsymbol{\theta}}_h^{\#}\right\|_2\right]\lesssim\frac{LB_\theta}{B_x\sqrt{\kappa_1\overline{K}}}\cdot\sqrt{\frac{d\ln(1/\delta)}{n\varepsilon}}.\tag{42}$$

3. Note that if we set λ and h as stated, it implies
$$\lambda\asymp\frac{1}{B_\theta}\cdot\sqrt{\frac{\kappa_1\overline{K}B_x^2}{n\varepsilon}+\frac{d\ln(1/\delta)}{n^2\varepsilon^2}};\quad h\asymp\frac{\overline{K}B_x^2}{\sqrt{n\varepsilon\kappa_1\overline{K}B_x^2+d\ln(1/\delta)}}.$$

   To employ Lemma 4.4, we must ensure $B_\theta^2\lambda^2\gtrsim h^2l_1\kappa_2$, or equivalently,
$$\frac{\kappa_1\overline{K}B_x^2}{n\varepsilon}+\frac{d\ln(1/\delta)}{n^2\varepsilon^2}-\frac{l_1\kappa_2\overline{K}B_x^2}{n\varepsilon\kappa_1+d\ln(1/\delta)/(\overline{K}B_x^2)}\gtrsim 0.$$

   A sufficient condition to make above statement true is $\delta\asymp n^{-w}$ for some $w>0$, since the third term is at the rate $1/(n\varepsilon+dw\ln(n))$, which will be cancelled out by the first term whose rate is $1/(n\varepsilon)$. Therefore, under the condition $\delta\asymp n^{-w}$, by Lemma 4.4, we obtain that, with probability at least $1-\gamma$,
$$\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}_h^{\#}-\boldsymbol{\theta}_h^*\right\|_2&\lesssim\frac{1}{\rho_1\underline{f}}\cdot\left(L\sqrt{\frac{d+\ln(1/\gamma)}{n}}+B_\theta\lambda\right)\\
&\lesssim\frac{1}{\rho_1\underline{f}}\cdot\left(L\sqrt{\frac{d+\ln(1/\gamma)}{n}}+\sqrt{\frac{\kappa_1\overline{K}B_x^2}{n\varepsilon}+\frac{d\ln(1/\delta)}{n^2\varepsilon^2}}\right)
\end{aligned}\tag{43}$$

Combining three parts (41), (42), and (43) together, we obtain

$$\mathbb{E}_{\boldsymbol{b}}\left[\left\|\widehat{\boldsymbol{\theta}}_h^{\mathsf{OP}}(\boldsymbol{b})-\boldsymbol{\theta}^*\right\|_2\right]\lesssim\frac{1}{\rho_1\underline{f}}\cdot\max\left\{\sqrt{\frac{d+\ln(1/\gamma)}{n}},\sqrt{\frac{d\ln(1/\delta)}{n\varepsilon}}\right\}.$$

□

C. More Simulations

In the main body, we have shown simulation results when $d = 3$. Now, we consider a data generating process with larger $d = 51$, and let $y = 10 + \boldsymbol{\theta}^\top \boldsymbol{x} + \epsilon$, where elements of $\boldsymbol{\theta} \in \mathbb{R}^{51}$ take values ascendingly from [-2, 5] with even steps. The feature vector $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{x}}, \Sigma_{\boldsymbol{x}})$ with mean $\boldsymbol{\mu}_{\boldsymbol{x}} = [1, 0, \ldots, 0] \in \mathbb{R}^{51}$ and covariance matrix $\Sigma_{\boldsymbol{x}} = Diag([0, \frac{1}{\sqrt{50}}, \ldots, \frac{1}{\sqrt{50}}])$. We set the error term $\epsilon \sim \mathcal{N}(0, 3^2)$. Quantile level $r = 0.7$. Other settings are the same as in Figure 4. We can see from Figure 8 and 9 that our private algorithms outperform existing algorithms. We note that Phased-ERM does not appear in any pictures because its relative risk (or relative estimation error) is beyond y-axis's range. The conjectured reason is as follows. In i-th iteration of Phased-ERM, the estimator is updated as $\widehat{\boldsymbol{\theta}}_t = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}} \frac{1}{n_i}\sum_{t=1}^{n_i} \ell(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) + \frac{1}{n_i\eta_i}\left\|\boldsymbol{\theta} - (\widehat{\boldsymbol{\theta}}_{t-1} + \boldsymbol{\xi}_{t-1})\right\|_2^2$, where $\boldsymbol{\xi}_{t-1} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$ with $\sigma = 4L(\eta_i/\varepsilon)\sqrt{\ln(1/\delta)}$, $n_i = 2^{-i}n$, and $\eta_i = 4^{-i}\eta$. That implies the coefficient of $\|\cdot\|_2^2$-regularizer $\frac{1}{n_i\eta_i}$ is $\frac{2^{3i}}{n\eta}$, which increases exponentially fast in iteration counter i. The dramatically increasing regularizer's coefficient anchors $\widehat{\boldsymbol{\theta}}_t$ around estimators in first few iterations, but the estimators in first few iterations are seriously affected by injected noises. As a consequence, the final estimator's practical performance is unappealing. Nevertheless, when sample size keeps increasing, the excess generalization risk of Phased-ERM will still converge (see Figure 10)

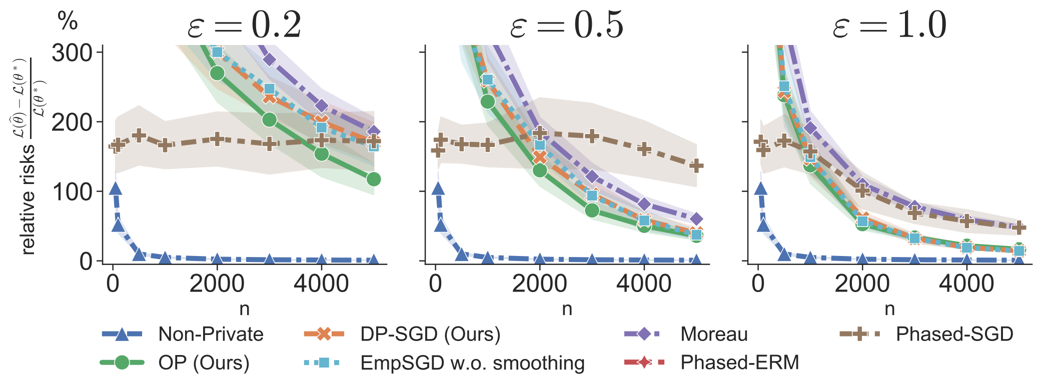


Figure 8. Relative excess generalization risks (d=51, *exogenous* error term)

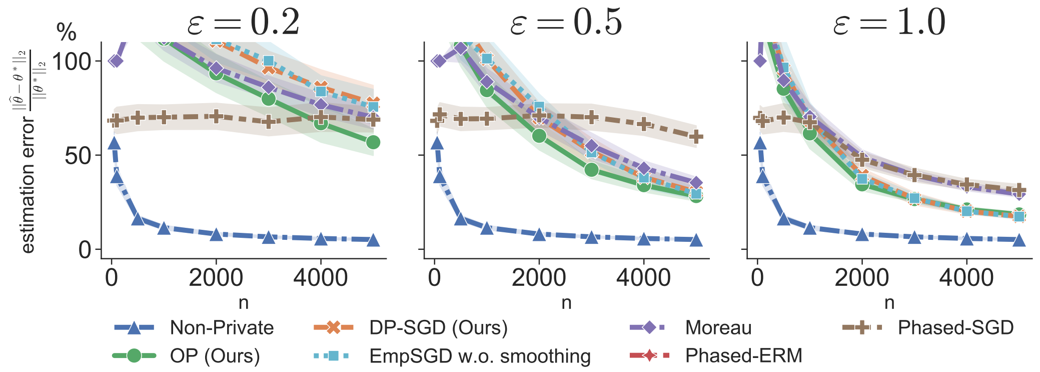


Figure 9. Relative estimation errors (d=51, *exogenous* error term)

Now, we switch to a case with an endogenous error term $\epsilon(\boldsymbol{x}) = \mathcal{N}(0, 3^2) + x_1 - x_2 + |x_1 x_2|$. Other settings remain unchanged. Figure 11 shows the relative excess generalization risks while Figure 12 shows the relative estimation errors. It is clear that our private algorithms still perform better among all private algorithms.

References

Asi, H., Feldman, V., Koren, T., and Talwar, K. Private stochastic convex optimization: Optimal rates in l1 geometry. In *International Conference on Machine Learning*, pp. 393–403. PMLR, 2021.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pp. 464–473. IEEE, 2014.

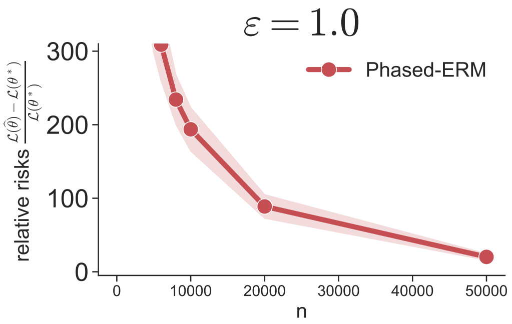


Figure 10. Relative estimation errors of Phased-ERM (d=51, *exogenous* error term)

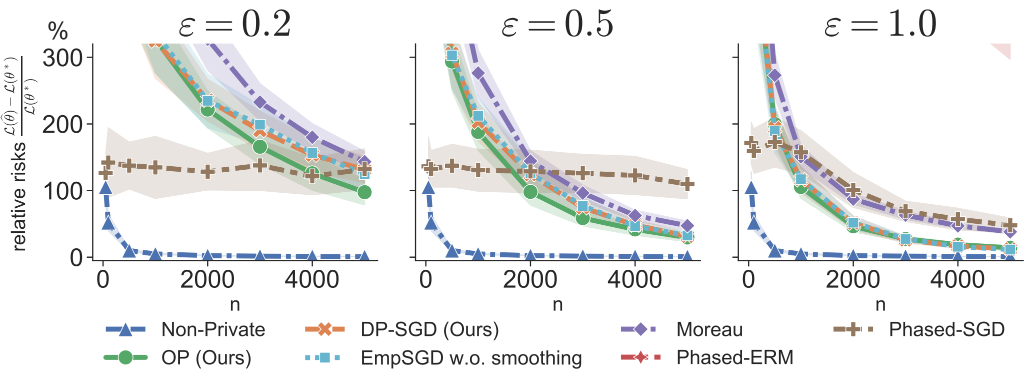


Figure 11. Relative excess generalization risks (d=51, *endogenous* error term)

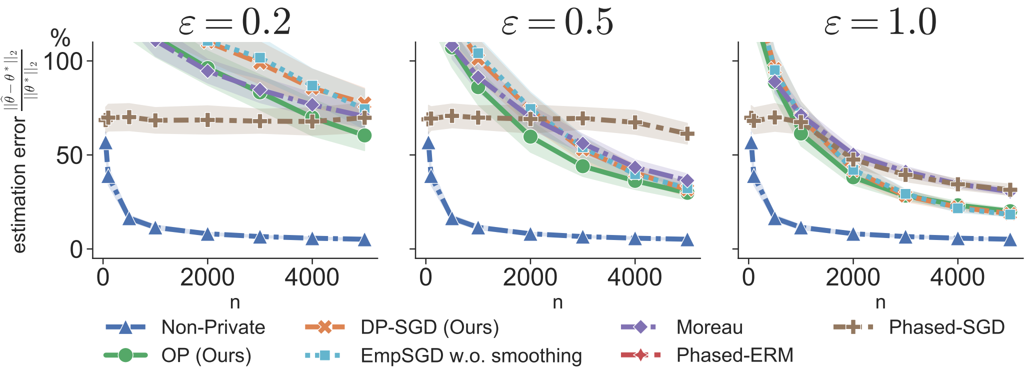


Figure 12. Relative estimation errors (d=51, *endogenous* error term)

Bassily, R., Feldman, V., Talwar, K., and Guha Thakurta, A. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Bousquet, O. and Elisseeff, A. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

He, X., Pan, X., Tan, K. M., and Zhou, W.-X. Smoothed quantile regression with large-scale inference. *Journal of Econometrics*, 2021.

Kifer, D., Smith, A., and Thakurta, A. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pp. 25–1. JMLR Workshop and Conference Proceedings, 2012.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Vadhan, S. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pp. 347–450. Springer, 2017.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.