Differentially Private Empirical Risk Minimization for AUC Maximization

Puyu Wang[1,2], Zhenhuan Yang[1], Yunwen Lei[3], Yiming Ying[1,*] and Hai Zhang[2]
[1]Department of Mathematics and Statistics, State University of New York at Albany, Albany, NY 12222, USA
[2]School of Mathematics, Northwest University, Xi'an 710127, China
[3]School of Computer Science, University of Birmingham, Birmingham B15 2TT, UK

Abstract

Area under ROC curve (AUC) is a widely used performance measure for imbalanced classification. Oftentimes, the ubiquitous imbalanced data such as financial records from fraud detection or genomic data from cancer diagnosis contains sensitive information, and therefore it is of practical and theoretical importance to develop privacy-preserving AUC maximization algorithms. In this paper, we propose differentially private empirical risk minimization (ERM) for AUC maximization, and systematically study their privacy and utility guarantees. In particular, we establish guarantees on the generalization (utility) performance of the proposed algorithms with fast rates. The technical novelty contains fast rates for the regularized ERM in AUC maximization, which is established using the peeling techniques for Rademacher averages [1] and properties of U-Statistics [2, 3] to handle statistically non-independent pairs of examples in the objective function, and a new error decomposition to handle strongly smooth losses (e.g. least square loss). In addition, we revisit the private ERM with pointwise loss [4, 5] and show optimal rates can be obtained using the uniform convergence approach.

Keywords: Differential privacy, Imbalanced classification, AUC maximization, Empirical risk minimization

1. Introduction

Many learning tasks involve imbalanced classification in which the size of one class is much larger than others. Imbalanced data is abundant in important application domains such as medical diagnosis, network intrusion detection, and fraud detection. In such cases, the quality of classification is often measured by the area under the ROC curve (AUC) [6, 7, 8, 9]. Recently, there are considerable work [10, 11, 12, 13, 14, 15, 16, 17, 18] on *AUC maximization* which have been proven to be effective for handling imbalanced data. Oftentimes, the ubiquitously generated imbalanced data such as financial records from fraud detection or genomic data from cancer diagnosis contains very sensitive information. This has raised serious concerns that the adversaries may be able to infer private information from trained AUC maximization models. As such, it is of practical and theoretical importance to develop privacy-preserving AUC maximization algorithms.

Differential privacy (DP) [19] is a de facto concept for designing algorithm with privacy guarantees. It ensures that the output of a learning algorithm is insensitive to any change of an individual in the dataset. Many studies [20, 21, 4, 22, 23, 24, 5, 25, 26] have focused on developing efficient differentially private learning algorithms while preserving their statistical effectiveness. In particular, the work [4] studied differential privacy for the fundamental supervised learning framework, i.e. empirical risk minimization (ERM). Assuming the loss is convex with Lipschitz gradient and differentiable, the authors investigated both output and objective perturbations with random noise added to the output of the ERM minimizer and the objective function, respectively. Privacy and utility guarantees (generalization performance) are established there. The work [5] improved the analysis in [4] by providing an improved $\Omega(\sqrt{d})$ dependence (d is the data dimension) in the utility guarantees and extended differential privacy results to the case of non-smooth objective functions. In [21], gradient perturbation and exponential sampling were proposed.

However, all the above work applies to the classification and regression problems of *pointwise learning*, i.e. the loss function $\ell(\mathbf{w}, \mathbf{z})$ depends on one single data point $\mathbf{z} = (\mathbf{x}, y)$. The loss function for AUC maximization involves a *pairwise* loss $\ell(\mathbf{w}, \mathbf{z}, \mathbf{z}')$ requiring more delicate techniques for developing privacy-preserving algorithms, as pairs of examples are not statistically independent of each other (e.g. two pairs may share one common example). In this paper, we leverage the previous work [21, 4, 19, 27, 5, 28] to systematically develop and analyze the differentially private ERM framework for AUC maximization. Our main contributions can be summarized as follows.

- We systematically study the output and objective perturbation mechanisms for the regularized ERM in AUC maximization, and provide comprehensive results on their privacy guarantees. The privacy analysis for objective perturbation involves the estimation of two Jacobians differing in a possible rank-n matrix (n is the size of training data) which significantly extends the analysis in [4, 5], where the counterparts differ only in a matrix with rank at most two.
- We provide guarantees on the generalization (utility) per-

formance of the proposed private AUC maximization algorithms with fast rates. In particular, for objective perturbation, we show that the excess population risk can achieve bound $O(\max(\frac{d}{n\epsilon}, \frac{1}{\sqrt{n}}))$ for ϵ-DP and $O(\max(\frac{\sqrt{\log(\frac{1}{\delta})d}}{n\epsilon}, \frac{1}{\sqrt{n}}))$ for (ϵ, δ)-DP. The main technical novelty is the fast rate for the regularized ERM in AUC maximization which extends [29] to the setting of pairwise losses. In contrast to the pointwise learning where the losses involves i.i.d. individual examples, the main challenge to derive such fast rates is that the objective function of AUC maximization involves pairs of examples which are not statistically independent. We overcome this hindrance by sufficiently exploring the properties of U-Statistics [2, 3] to handle this statistical non-independence. We also introduce a new error decomposition which enables us to handle the losses with strong smoothness in unbounded parameter domain (e.g. the least square loss). In addition, we revisit the private ERM with pointwise loss [4, 5] and show optimal rates can be obtained using the uniform convergence approach (see detailed discussion and comparison in Section 5).

- We conduct experimental evaluation of the proposed approaches on various datasets. The results validate the effectiveness of preserving privacy and generalization performance of the proposed private AUC maximization algorithms.

Related Work. Below we review related work on DP and AUC maximization which is by no means extensive. Batch learning algorithms for AUC maximization were studied in [30, 31, 17]. The work of [10, 32, 15, 18] developed online AUC maximization algorithms and established regret bounds. Stochastic gradient-based algorithms were developed for AUC maximization in [12, 14, 16] for the linear case which enjoys cheap per-iteration cost and fast convergence rates. The recent work studied AUC maximization with deep neural networks [13] by casting it into a non-convex concave min-max problem. Nonlinear AUC maximization was also studied in [11]. An appealing stochastic primal-dual algorithm for saddle point problems was developed in [33] which, as a by-product, can be applied to AUC maximization with the least square loss.

Recently, there is a large amount of work on differential privacy for pointwise learning from different perspectives. The private ERM was first studied in [4, 34], although other variants were studied before. The output perturbation was studied in both papers where one releases an output with additive noise. The objective perturbation was introduced in [4] and improved in [35, 23, 5] where the noise was directly added to the ERM objective. The gradient perturbation was studied and analyzed in [21] and further improved in [26, 36]. The work [37, 38] studied regret bounds in online learning. The relation between learnability and stability, and DP was systematically addressed in [39]. Recently, optimal rates for private stochastic convex optimization were investigated in [40, 20, 22]. The work of [41] and [42] considered the DP for rank aggregation which combines multiple ranked lists into a single ranking. [43] proposed differential pairwise privacy for secure metric learning but utility (generalization) analysis is not given. The work [44] studied privacy-preserving pairwise learning algorithms with output perturbation when the parameter domain is bounded and the loss function is Lipschitz and smooth. While we studied the DP with output and objective perturbations in unbounded parameter domain for the most baseline ERM framework. Especially, we proposed a new error decomposition to handle the losses without Lipschitz property. Further, our method provides the better utility bound than theirs: $O(\frac{\sqrt{d}}{\epsilon\sqrt{n}})$ in [44] versus $O(\max(\frac{\sqrt{d}}{\epsilon n}, \frac{1}{\sqrt{n}}))$ given by our Theorem 5.

Organization of the Paper. The paper is organized as follows. In Section 2, we introduce the formulation of AUC maximization and the definition of differential privacy. The proposed private algorithms and privacy guarantees are given in Section 3. In Section 4, we establish guarantees on the generalization (utility) performance of the proposed algorithms. In Section 5, we show, for private ERM algorithms with pointwise losses, that optimal rates can be achieved by using the uniform convergence approach, and discuss the comparison with related work. Examples are given in Section 6. Section 7 concludes the paper. Detailed technical proofs are postponed to Appendix. To facilitate the presentation, Table 1 summarizes the main notations.

Symbol	Meaning
λ	ℓ_2-regularization parameter
$D_{\mathcal{X}}$	diameter of the input space
ϵ, δ	privacy parameters
L	Lipschitz constant
β	smoothness
$\mathcal{R}(\mathbf{w})$	population risk
$\mathcal{R}^{\lambda}(\mathbf{w})$	regularized population risk
$\mathcal{R}_S(\mathbf{w})$	empirical risk
$\mathcal{R}_S^{\lambda}(\mathbf{w})$	regularized empirical risk
$\mathbf{w}^*$	$\arg\inf_{\mathbf{w}\in\mathbb{R}^d} \mathcal{R}(\mathbf{w})$
$\mathbf{w}_{\lambda}$	$\arg\inf_{\mathbf{w}\in\mathbb{R}^d} \mathcal{R}^{\lambda}(\mathbf{w})$
$\widehat{\mathbf{w}}$	$\arg\min_{\mathbf{w}\in\mathbb{R}^d} \mathcal{R}_S^{\lambda}(\mathbf{w})$
$\|\cdot\|$	Euclidean norm unless special remark

Table 1: Summary of Main Notations.

2. Problem Formulation and Notations

Let the input space $\mathcal{X} \subseteq \mathbb{R}^d$, output space $\mathcal{Y} = \{\pm 1\}$, and denote the joint sample space by $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$. Denote the training data by $S = \{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$ with $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, which is assumed to be i.i.d from an unknown distribution P on $\mathcal{X} \times \mathcal{Y}$. Let $S' = \{(\mathbf{x}_i', y_i') : i = 1, \ldots, n\}$ be a *neighboring* data to S, i.e. the datasets S and S' differ only in one single datum. Throughout this paper, we assume that the input space $\mathcal{X}$ is a bounded domain and denote its diameter by $D_{\mathcal{X}} = \sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} \|\mathbf{x} - \mathbf{x}'\|$. The AUC score [2, 6, 9] of a prescribed parameter $\mathbf{w}$ on the

data S is given by

$$\mathrm{AUC}(\mathbf{w}) = \frac{1}{n_+ n_-} \sum_{i,j=1}^{n} \mathbb{I}_{[\mathbf{w}^T\mathbf{x}_i > \mathbf{w}^T\mathbf{x}_j]} \mathbb{I}_{[y_i=1 \wedge y_j=-1]}, \tag{1}$$

where n_+ and n_- denote the numbers of instances in the positive and negative classes, respectively, and $\mathbb{I}_{[\cdot]}$ is the indicator function which returns 1 for the true event and 0 otherwise. Maximizing the AUC score in (1) w.r.t. $\mathbf{w}$ is equivalent to minimizing the quantity $1 - \mathrm{AUC}(\mathbf{w}) = \frac{1}{n_+ n_-} \sum_{i,j=1}^{n} \mathbb{I}_{[\mathbf{w}^T(\mathbf{x}_i - \mathbf{x}_j) \le 0]} \mathbb{I}_{[y_i=1 \wedge y_j=-1]}$. Since the indicator function $\mathbb{I}_{\mathbf{w}^T\mathbf{x}_i \le \mathbf{w}^T\mathbf{x}_j}$ is discontinuous, one often replaces it by a convex surrogate loss $\ell : \mathbb{R} \to [0, \infty)$ such that $\mathbb{I}_{[t \le 0]} \le \ell(t)$. Such losses can be the least square loss $\ell(t) = (1-t)^2$ and logistic regression loss $\ell(t) = \log_2(1 + e^{-t})$.

Now the regularized ERM for AUC maximization (`AUC-ERM`) can be formulated as

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \Big\{ \frac{1}{n_+ n_-} \sum_{i,j=1}^{n} \ell(\mathbf{w}^T(\mathbf{x}_i - \mathbf{x}_j)) \mathbb{I}_{[y_i=1 \wedge y_j=-1]} + \frac{\lambda}{2}\|\mathbf{w}\|^2 \Big\}. \tag{2}$$

Let $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W} \subset \mathbb{R}^d$ be a randomized algorithm taking a data set $S \in \mathcal{Z}^n$ as input. Differential privacy was introduced in [19] as a privacy measure for an algorithm $\mathcal{A}$.

Definition 1. A randomized algorithm $\mathcal{A}$ provides (ϵ, δ)-differential privacy (DP) if, for any two neighboring data sets S and S' differing in one single datum and any set $E \in \mathrm{Range}(\mathcal{A})$, there holds

$$\Pr(\mathcal{A}(S) \in E) \le e^{\epsilon} \Pr(\mathcal{A}(S') \in E) + \delta. \tag{3}$$

In particular, if $\delta = 0$, we call it ϵ-DP.

A basic paradigm to achieve ϵ-DP is to use the L_2-sensitivity of $\mathcal{A}$, which is defined as follows.

Definition 2. The ℓ_2-sensitivity of algorithm $\mathcal{A}$ is defined as $\Delta(\mathcal{A}) = \sup_{S,S'} \|\mathcal{A}(S) - \mathcal{A}(S')\|$, where data sets S and S' differ in one single datum.

Throughout the paper, we always assume that the loss $\ell : \mathbb{R} \to [0, \infty)$ is convex with $\ell(0) = 1$. For any $R \ge 0$, define $B(R) = \sup_{|s| \le R} |\ell'(s)|$. We say that the loss ℓ is L-Lipschitz if, for any $s, t \in \mathbb{R}$, $|\ell(s) - \ell(t)| \le L|s - t|$, and β-strongly smooth if its derivative is β-Lipschitz. In particular, we have the following sensitivity result for `AUC-ERM` defined by (2).

Lemma 1. *`AUC-ERM` (2) has L_2-sensitivity with* $\Delta(\mathcal{A}) = \frac{2D_{\mathcal{X}} B(\sqrt{2/\lambda} D_{\mathcal{X}})}{\lambda} (\frac{1}{n_+} + \frac{1}{n_-})$.

The proof of Lemma 1 is standard which is provided in Appendix C.1 for completeness. The following lemma states that adding Gaussian noise to the output of a randomized algorithm $\mathcal{A}$ can guarantee (ϵ, δ)-DP.

Lemma 2. *([27]) Given any function $\mathcal{A} : \mathcal{Z}^n \to \mathbb{R}^d$ with L_2 sensitivity $\Delta(\mathcal{A})$ and assume that $\sigma \ge \frac{\sqrt{2\ln(1.25/\delta)}\Delta(\mathcal{A})}{\epsilon}$, the following Gaussian mechanism yields (ϵ, δ)-DP:*

$$\mathcal{M}(\mathcal{A}, S, \epsilon) = \mathcal{A}(S) + \mathbf{b}, \quad \mathbf{b} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}).$$

Algorithm 1 Output Perturbation for `AUC-ERM` (`Output-Pert-AUC`)

1: **Inputs:** Data $S = \{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$ and parameters $\lambda, \epsilon, \delta$
2: **Compute:** $B(\sqrt{2/\lambda} D_{\mathcal{X}}) = \sup\{|\ell'(t)| : |t| \le \sqrt{2/\lambda} D_{\mathcal{X}}\}$, $n_+ = \sum_{i=1}^{n} \mathbb{I}_{[y_i=1]}$ and $n_- = \sum_{i=1}^{n} \mathbb{I}_{[y_i=-1]}$
3: **if** require ϵ-differential privacy **then**
4: compute $\gamma = \frac{2D_{\mathcal{X}} B(\sqrt{2/\lambda} D_{\mathcal{X}})}{\epsilon\lambda} (\frac{1}{n_+} + \frac{1}{n_-})$, and sample $\mathbf{b}$ from $\nu_1(\mathbf{b}; \gamma, \epsilon) \propto e^{-\frac{\|\mathbf{b}\|}{\gamma}}$
5: **else if** require (ϵ, δ)-differential privacy **then**
6: computer $\sigma = \frac{2\sqrt{2\log(1.25/\delta)} D_{\mathcal{X}} B(\sqrt{2/\lambda} D_{\mathcal{X}})}{\epsilon\lambda} (\frac{1}{n_+} + \frac{1}{n_-})$, and sample $\mathbf{b}$ from $\nu_2(\mathbf{b}; \epsilon, \delta, \sigma) = \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$.
7: **end if**
8: **return:** $\mathbf{w}_{\mathrm{priv}} = \mathbf{b} + \arg\min_{\mathbf{w} \in \mathbb{R}^d} \mathcal{R}_S^{\lambda}(\mathbf{w})$

3. Privacy-Preserving ERM for AUC Maximization

In this section, we present privacy-preserving algorithms for AUC maximization using output and objective perturbations, and provide a systematical study on its privacy guarantees.

Output Perturbation. In analogy to [4, 19], differential privacy can be achieved by following:

$$\mathbf{w}_{\mathrm{priv}}(S) = \widehat{\mathbf{w}}(S) + \mathbf{b}, \tag{4}$$

where $\mathbf{b}$ is a random noise and $\widehat{\mathbf{w}}(S)$ is the ERM minimizer of `AUC-ERM` (2). In particular, if $\mathbf{b}$ is from distribution with density $\frac{1}{\alpha} \exp(-\frac{\|\mathbf{b}\|}{\gamma})$, where α is a normalizing constant, then it achieves ϵ-DP. If the random noise $\mathbf{b}$ is from the Gaussian distribution $\mathbf{b} \propto \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, it achieves (ϵ, δ)-DP. For simplicity, denote the empirical risk by $\mathcal{R}_S(\mathbf{w}) = \frac{1}{n_+ n_-} \sum_{i,j=1}^{n} \ell(\mathbf{w}^T(\mathbf{x}_i - \mathbf{x}_j)) \mathbb{I}_{[y_i=1 \wedge y_j=-1]}$, and its regularized empirical risk by $\mathcal{R}_S^{\lambda}(\mathbf{w}) = \mathcal{R}_S(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2$. The pseudo-code of the output perturbation for `AUC-ERM` (`Output-Pert-AUC`) is given by `Algorithm` 1. It is worthy of mentioning that `Algorithm` 1 applies to any convex and differentiable loss ℓ as long as $B(\sqrt{2/\lambda} D_{\mathcal{X}}) < \infty$. Its privacy guarantees are stated in the following theorem.

Theorem 1. *Assume that the loss $\ell(\cdot)$ is convex and differentiable. Then, `Algorithm 1 (Out-Pert-AUC)` is ϵ-DP when $\mathbf{b}$ has density ν_1 and (ϵ, δ)-DP when $\mathbf{b}$ is from density ν_2.*

Proof. Consider the output perturbation (4). For ϵ-DP, let $\gamma = \frac{2D_{\mathcal{X}} B(\sqrt{2/\lambda} D_{\mathcal{X}})}{\epsilon\lambda} (\frac{1}{n_+} + \frac{1}{n_-})$. For any S, S' differing in one datum, and any $E \in \mathbb{R}^d$, $\Pr(\mathbf{w}_{\mathrm{priv}}(S) \in E) = \int_E \frac{1}{\alpha} e^{-\|\xi - \widehat{\mathbf{w}}(S)\|/\gamma} d\xi \le \exp(\frac{\sup_{S,S'} \|\widehat{\mathbf{w}}(S) - \widehat{\mathbf{w}}(S')\|}{\gamma}) \int_E \frac{1}{\alpha} e^{-\|\xi - \widehat{\mathbf{w}}(S')\|/\gamma} d\xi \le e^{\epsilon} \Pr(\mathbf{w}_{\mathrm{priv}}(S') \in E)$, where the last inequality used Lemma 1. This completes the proof of the theorem. For (ϵ, δ)-DP, the proof directly follows from Lemma 2. □

Objective Perturbation. An alternative approach to achieve differential privacy is to use the objective perturbation [4]. That is, $\mathbf{w}_{\mathrm{priv}} = \arg\min_{\mathbf{w}} \mathcal{R}_S^{\lambda}(\mathbf{w}) + \mathbf{b}^T\mathbf{w}$, where $\mathbf{b}$ is a random noise generated from distribution with density $\frac{1}{\alpha} \exp(-\frac{\|\mathbf{b}\|}{\gamma})$ or Gaussian distribution. `Algorithm` 2 lists the pseudo-code for the ob-

Algorithm 2 Objective Perturbation for AUC-ERM (Obj-Pert-AUC)

Inputs: Data $S = \{(\mathbf{x}_i, y_i) : i = 1, \dots, n\}$ and parameters $\lambda, \epsilon, \delta, L, \beta$
Compute: $n_+ = \sum_{i=1}^n \mathbb{I}_{[y_i=1]}$ and $n_- = \sum_{i=1}^n \mathbb{I}_{[y_i=-1]}$
if require ϵ-differential privacy **then**
  if $n\log(1 + \frac{\beta D_X^2}{n_+n_-\lambda}) < \epsilon$ **then**
    let $\Delta = 0$, $\gamma = 2nLD_X[n_+n_-(\epsilon - n\log(1 + \frac{\beta D_X^2}{n_+n_-\lambda})]^{-1}$.
  else if $n\log(1 + \frac{\beta D_X^2}{n_+n_-\lambda}) \geq \epsilon$ **then**
    let $\gamma = \frac{4nLD_X}{n_+n_-\epsilon}$ and $\Delta = \frac{\beta D_X^2}{n_+n_-(e^{\frac{\epsilon}{2n}}-1)} - \lambda$
  end if
  sample $\mathbf{b}$ from $\nu_1(\mathbf{b}; \gamma, \epsilon) \propto e^{-\frac{\|\mathbf{b}\|}{\gamma}}$
else if require (ϵ, δ)-differential privacy **then**
  if $n\log(1 + \frac{\beta D_X^2}{n_+n_-\lambda}) < \epsilon$ **then**
    let $\Delta = 0$, $\epsilon' = \epsilon - n\log(1 + \frac{\beta D_X^2}{n_+n_-\lambda})$, and $\sigma = (2\sqrt{2\log(\frac{1}{\delta})} + \sqrt{2\epsilon'})nLD_X/(n_+n_-\epsilon')$
  else if $n\log(1 + \frac{\beta D_X^2}{n_+n_-\lambda}) \geq \epsilon$ **then**
    choose $\epsilon' = \frac{\epsilon}{2}$, $\Delta = \frac{\beta D_X^2}{n_+n_-(e^{\frac{\epsilon}{2n}}-1)} - \lambda$, and $\sigma = (2\sqrt{2\log(\frac{1}{\delta})} + \sqrt{2\epsilon'})nLD_X/(n_+n_-\epsilon')$
  end if
  sample $\mathbf{b}$ from $\nu_2(\mathbf{b}; \epsilon, \delta, \sigma) = \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.
end if
return: $\mathbf{w}_{\text{priv}} = \arg\min\{\mathcal{R}_S^\lambda(\mathbf{w}) + \frac{\Delta}{2}\|\mathbf{w}\|^2 + \mathbf{b}^T\mathbf{w}\}$

jective perturbation for AUC-ERM. Compared to the output perturbation given by Algorithm 1, the loss function has much stronger assumptions, i.e. it is twice-differentiable, Lipschitz and strongly smooth. In particular, we can show the following privacy guarantees for Algorithm 2 (Obj-Pert-AUC).

Theorem 2. *Assume that ℓ is convex and twice-differentiable, L-Lipschitz and β-strongly smooth. Then, Algorithm 2 (Obj-Pert-AUC) achieves ϵ-DP when $\mathbf{b}$ is generated by distribution ν_1 and (ϵ, δ)-DP when $\mathbf{b}$ has Gaussian distribution ν_2.*

Proof. Without loss of generality, assume S and S' differ in the first datum, i.e. $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_1', y_1')$. Now consider an output $\mathbf{w}_{\text{priv}}$ from Algorithm 2. As the objective function is differentiable and strongly convex, the map between $\mathbf{b}$ and $\widehat{\mathbf{w}}_{\text{priv}}$ is bijective, i.e. perfect one-to-one correspondence between $\mathbf{b}$ and $\widehat{\mathbf{w}}_{\text{priv}}$. And we have

$$\mathbf{b} = -\Big(\frac{1}{n_+n_-}\sum_{i,j=1}^n \ell'(\mathbf{w}_{\text{priv}}^T(\mathbf{x}_i - \mathbf{x}_j))(\mathbf{x}_i - \mathbf{x}_j)\mathbb{I}_{[y_i=1\wedge y_j=-1]} + (\lambda + \Delta)\mathbf{w}_{\text{priv}}\Big). \tag{5}$$

Let $\text{pdf}(\mathbf{w}_{\text{priv}}|S)$ and $\text{pdf}(\mathbf{w}_{\text{priv}}|S')$ be the densities of $\mathbf{w}_{\text{priv}}$ given by S and S' respectively. To show the differential privacy, it suffices to estimate the density ratio of $\frac{\text{pdf}(\mathbf{w}_{\text{priv}}|S)}{\text{pdf}(\mathbf{w}_{\text{priv}}|S')}$. To this end, we denote by $\text{pdf}(\mathbf{b}|S)$ and $\text{pdf}(\mathbf{b}|S')$ the densities of the given $\mathbf{w}_{\text{priv}}$, when the datasets are S and S' respectively. Denote by $\mathbf{J}(\mathbf{w}_{\text{priv}} \to \mathbf{b}|S)$ and $\mathbf{J}(\mathbf{w}_{\text{priv}} \to \mathbf{b}|S')$ the Jacobians of the mappings from $\mathbf{w}_{\text{priv}}$ to $\mathbf{b}$ with given S and S', respectively.

Therefore,

$$\frac{\text{pdf}(\mathbf{w}_{\text{priv}}|S)}{\text{pdf}(\mathbf{w}_{\text{priv}}|S')} = \frac{\text{pdf}(\mathbf{b}|S)}{\text{pdf}(\mathbf{b}'|S')} \cdot \frac{|\det(\mathbf{J}(\mathbf{w}_{\text{priv}} \to \mathbf{b}'|S'))|}{|\det(\mathbf{J}(\mathbf{w}_{\text{priv}} \to \mathbf{b}|S))|}. \tag{6}$$

We will estimate (6) in two steps.

Step 1: Firstly, we estimate the ratio between two Jacobians in (6). To this end, denote $X_{ij} = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$ and $X'_{1i} = (\mathbf{x}_1' - \mathbf{x}_i)(\mathbf{x}_1' - \mathbf{x}_i)^T$ and $X'_{j1} = (\mathbf{x}_j - \mathbf{x}_1')(\mathbf{x}_j - \mathbf{x}_1')^T$. In addition, denote $E_{i1} = \frac{1}{n_+n_-}\ell''(\mathbf{w}_{\text{priv}}^T(\mathbf{x}_i - \mathbf{x}_1))X_{i1}\mathbb{I}_{[y_i=1\wedge y_1=-1]}$, $E_{1j} = \frac{1}{n_+n_-}\ell''(\mathbf{w}_{\text{priv}}^T(\mathbf{x}_1 - \mathbf{x}_j))X_{1j}\mathbb{I}_{[y_1=1\wedge y_j=-1]}$, and likewise, $E'_{i1} = \frac{1}{n_+n_-}\ell''(\mathbf{w}_{\text{priv}}^T(\mathbf{x}_i - \mathbf{x}_1'))X'_{i1}\mathbb{I}_{[y_i=1\wedge y_1'=-1]}$, and $E'_{1j} = \frac{1}{n_+n_-}\ell''(\mathbf{w}_{\text{priv}}^T(\mathbf{x}_1' - \mathbf{x}_j))X'_{1j}\mathbb{I}_{[y_1'=1\wedge y_j=-1]}$. Let $A = (\lambda + \Delta)\mathbb{I} + \frac{1}{n_+n_-}\sum_{i,j=2}^n \ell''(\mathbf{w}_{\text{priv}}(\mathbf{x}_i - \mathbf{x}_j))X_{ij}\mathbb{I}_{[y_i=1\wedge y_j=-1]}$ be the common matrix shared by the above two Jacobians. Notice that

$$\begin{aligned}
&\mathbf{J}(\mathbf{w}_{\text{priv}} \to \mathbf{b}|S) \\
&= -\Big(\frac{1}{n_+n_-}\sum_{i,j=1}^n \ell''(\mathbf{w}_{\text{priv}}^T(\mathbf{x}_i - \mathbf{x}_j))(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T\mathbb{I}_{[y_i=1\wedge y_j=-1]} \\
&\qquad + (\lambda + \Delta)\mathbb{I}\Big) \\
&= -(A + \sum_{i=2}^n E_{i1} + \sum_{j=2}^n E_{1j})
\end{aligned}$$

and similarly, $\mathbf{J}(\mathbf{w}_{\text{priv}} \to \mathbf{b}|S') = -(A + \sum_{i=2}^n E'_{i1} + \sum_{j=2}^n E'_{1j})$. Therefore,

$$\begin{aligned}
\frac{|\det(\mathbf{J}(\mathbf{w}_{\text{priv}} \to \mathbf{b}|S'))|}{|\det(\mathbf{J}(\mathbf{w}_{\text{priv}} \to \mathbf{b}|S))|} &= \frac{\det(A + \sum_{i=2}^n E'_{i1} + \sum_{j=2}^n E'_{1j})}{\det(A + \sum_{i=2}^n E_{i1} + \sum_{j=2}^n E_{1j})} \\
&= \frac{\det(A + \sum_{i=2}^n E'_{i1} + \sum_{j=2}^n E'_{1j})}{\det(A)} \cdot \frac{\det(A)}{\det(A + \sum_{i=2}^n E_{i1} + \sum_{j=2}^n E_{1j})} \\
&\leq \frac{\det(A + \sum_{i=2}^n E'_{i1} + \sum_{j=2}^n E'_{1j})}{\det(A)},
\end{aligned} \tag{7}$$

where the last inequality follows from the positive semi-definite (PSD) of $\sum_{i=2}^n E_{i1} + \sum_{j=2}^n E_{1j}$ as the loss ℓ is convex.

Therefore, the estimation of two Jacobians is reduced to the estimation of the ratio $\det(A + \sum_{i=2}^n E'_{i1} + \sum_{j=2}^n E'_{1j})/\det(A)$. Notice, for any $\ell''(z)\mathbf{v}\mathbf{v}^T$ with $z \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^d$, and any PSD matrix $B \geq (\Delta + \lambda)\mathbb{I}$, that

$$\begin{aligned}
\frac{\det(B + \ell''(z)\mathbf{v}\mathbf{v}^T)}{\det(B)} &= \det(\mathbb{I} + \ell''(z)B^{-1/2}\mathbf{v}\mathbf{v}^TB^{-1/2}) \\
&= 1 + \ell''(z)\|B^{-1/2}\mathbf{v}\|^2 \leq 1 + \frac{\beta\|\mathbf{v}\|^2}{\lambda + \Delta},
\end{aligned} \tag{8}$$

where the last inequality follows from the β-smoothness of ℓ and $B \geq (\lambda + \Delta)\mathbb{I}$. We now rewrite

$$\begin{aligned}
&\frac{\det(A + \sum_{i=2}^n E'_{i1} + \sum_{j=2}^n E'_{1j})}{\det(A)} \\
&= \frac{\det(A + \sum_{i=2}^n E'_{i1} + \sum_{j=2}^n E'_{1j})}{\det(A + \sum_{i=2}^n E'_{i1})} \cdot \frac{\det(A + \sum_{i=2}^n E'_{i1})}{\det(A)}.
\end{aligned} \tag{9}$$

For the first term on the right hand side of the above equality, applying (8) recursively implies that

$$\begin{aligned}
&\frac{\det(A+\sum_{i=2}^n E'_{i1}+\sum_{j=2}^n E'_{1j})}{\det(A+\sum_{i=2}^n E'_{i1})}\\
&=\prod_{k=2}^n\Big[\frac{\det(A+\sum_{i=2}^n E'_{i1}+\sum_{j=2}^k E'_{1j})}{\det(A+\sum_{i=2}^n E'_{i1}+\sum_{j=2}^{k-1} E'_{1j})}\Big]\\
&\le\Big(1+\frac{1}{(n_+n_-)}\frac{\beta D_X^2}{(\lambda+\Delta)}\Big)^{n_-}, \qquad (10)
\end{aligned}$$

where we used the fact that there are at most n_- non-zero terms in $\sum_{j=2}^n E_{1j}$, and $\|\mathbf{x}'_1-\mathbf{x}'_j\|\le D_X$. Likewise, we can have

$$\frac{\det(A+\sum_{i=2}^n E'_{i1})}{\det(A)}\le\Big(1+\frac{1}{(n_+n_-)}\frac{\beta D_X^2}{(\lambda+\Delta)}\Big)^{n_+}. \qquad (11)$$

Putting (7), (9), (10) and (11) together implies that

$$\frac{|\det(\mathbf{J}(\mathbf{w}_{\text{priv}}\to\mathbf{b}|S'))|}{|\det(\mathbf{J}(\mathbf{w}_{\text{priv}}\to\mathbf{b}|S))|}\le\Big(1+\frac{1}{(n_+n_-)}\frac{\beta D_X^2}{(\lambda+\Delta)}\Big)^{n}. \qquad (12)$$

Step 2: Now we estimate $\text{pdf}(\mathbf{b}|S)/\text{pdf}(\mathbf{b}'|S')$ when $\mathbf{b}$ is from distribution with density ν_1 or ν_2. Let us consider the case when the noise $\mathbf{b}$ is generated from distribution with density ν_1. In this case, for any given $\mathbf{w}_{\text{priv}}$, (5) implies that $\|\mathbf{b}-\mathbf{b}'\|\le\frac{2nLD_X}{n_+n_-}$. We can write that

$$\frac{\text{pdf}(\mathbf{b}|S)}{\text{pdf}(\mathbf{b}'|S')}=\frac{e^{\frac{-\|\mathbf{b}\|}{\gamma}}}{e^{\frac{-\|\mathbf{b}'\|}{\gamma}}}\le\exp(\frac{2nLD_X}{n_+n_-\gamma}) \qquad (13)$$

Putting (12) and (13) together implies that

$$\frac{\text{pdf}(\mathbf{w}_{\text{priv}}|S)}{\text{pdf}(\mathbf{w}_{\text{priv}}|S')}\le\exp\Big(n\log\Big(1+\frac{\beta D_X^2}{n_+n_-(\lambda+\Delta)}\Big)+\frac{2nLD_X}{n_+n_-\gamma}\Big). \qquad (14)$$

If $n\log(1+\frac{\beta D_X^2}{n_+n_-\lambda})<\epsilon$ then letting $\Delta=0$ and $\gamma=2nLD_X[n_+n_-(\epsilon-n\log(1+\frac{\beta D_X^2}{n_+n_-\lambda})]^{-1}$. From (14), there holds $\frac{\text{pdf}(\mathbf{w}_{\text{priv}}|S)}{\text{pdf}(\mathbf{w}_{\text{priv}}|S')}\le e^{\epsilon}$. Otherwise, letting $\gamma=\frac{4nLD_X}{n_+n_-\epsilon}$ and $\Delta=\frac{\beta D_X^2}{n_+n_-(e^{\frac{\epsilon}{2n}}-1)}-\lambda$, from (14) again, we have $\frac{\text{pdf}(\mathbf{w}_{\text{priv}}|S)}{\text{pdf}(\mathbf{w}_{\text{priv}}|S')}\le e^{\epsilon}$. This completes the proof of the theorem.

Now consider the case that the noise is drawn from Gaussian distribution. Then, assume $\Gamma=\mathbf{b}-\mathbf{b}'$,

$$\begin{aligned}
\frac{\text{pdf}(\mathbf{b}|S)}{\text{pdf}(\mathbf{b}'|S')}&=\frac{e^{\frac{-\|\mathbf{b}\|^2}{2\sigma^2}}}{e^{\frac{-\|\mathbf{b}'\|^2}{2\sigma^2}}}=\exp\Big(\frac{1}{2\sigma^2}(\|\Gamma\|^2-2\langle\mathbf{b},\Gamma\rangle)\Big)\\
&\le\exp\Big(\frac{1}{2\sigma^2}(\|\Gamma\|^2+2|\langle\mathbf{b},\Gamma\rangle|)\Big). \qquad (15)
\end{aligned}$$

Notice that $\|\Gamma\|\le\frac{2nLD_X}{n_+n_-}$, and let the event $\mathcal{E}=\{\mathbf{b}\in\mathbb{R}^d:|\langle\mathbf{b},\Gamma\rangle|\ge\frac{2nLD_X\sigma t}{n_+n_-}\}$. Notice, for one-dimensional Gaussian random variable $Z\sim\mathcal{N}(0,1)$, that for any $t\ge0$, $\Pr(|Z|\ge t)=\frac{2}{\sqrt{2\pi}}\int_t^\infty e^{\frac{-s^2}{2}}ds=\frac{2}{\sqrt{2\pi}}\int_0^\infty e^{-\frac{(s+t)^2}{2}}ds\le\frac{2e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\int_0^\infty e^{-\frac{s^2}{2}}ds=e^{-\frac{t^2}{2}}$. Therefore, since $\langle\mathbf{b},\Gamma\rangle\sim\mathcal{N}(0,\|\Gamma\|^2\sigma^2)$, there holds $\Pr(\mathcal{E})=\Pr(|\langle\mathbf{b},\Gamma\rangle|\ge\frac{2nLD_X\sigma t}{n_+n_-})\le e^{-\frac{t^2}{2}}$. This means, choosing $t=\sqrt{2\log\frac{1}{\delta}}$, we have $\Pr(\mathcal{E})\le\delta$. Hence, given any $\epsilon'>0$, from (15) we have that $\frac{\text{pdf}(\mathbf{b}|S)}{\text{pdf}(\mathbf{b}'|S')}\le e^{\frac{2(nLD_X)^2}{(n_+n_-)^2\sigma^2}+\frac{2nLD_X}{n_+n_-\sigma}\sqrt{2\log(\frac{1}{\delta})}}\le e^{\epsilon'}$ if $\sigma\ge(2\sqrt{2\log(\frac{1}{\delta})}+\sqrt{2\epsilon'})nLD_X/(n_+n_-\epsilon')$ on the event $\mathcal{E}^c$ with its probability at least $1-\delta$.

Combining (12) and the above estimation, we know that, choosing $\sigma=(2\sqrt{2\log(\frac{1}{\delta})}+\sqrt{2\epsilon'})nLD_X/(n_+n_-\epsilon')$, on the event $\mathcal{E}^c$, that

$$\frac{\text{pdf}(\mathbf{w}_{\text{priv}}|S)}{\text{pdf}(\mathbf{w}_{\text{priv}}|S')}\le\exp\Big(n\log\Big(1+\frac{\beta D_X^2}{n_+n_-(\lambda+\Delta)}\Big)+\epsilon'\Big). \qquad (16)$$

Now if $n\log(1+\frac{\beta D_X^2}{n_+n_-\lambda})<\epsilon$ then letting $\Delta=0$. We choose $\epsilon'=\epsilon-n\log(1+\frac{\beta D_X^2}{n_+n_-\lambda})$ and let $\sigma=(2\sqrt{2\log(\frac{1}{\delta})}+\sqrt{2\epsilon'})nLD_X/(n_+n_-\epsilon')$ which indicates that $\frac{\text{pdf}(\mathbf{w}_{\text{priv}}|S)}{\text{pdf}(\mathbf{w}_{\text{priv}}|S')}\le e^{\epsilon}$ on the event $\mathcal{E}^c$. If $n\log(1+\frac{\beta D_X^2}{n_+n_-\lambda})\ge\epsilon$, choose $\epsilon'=\frac{\epsilon}{2}$, $\sigma=(2\sqrt{2\log(\frac{1}{\delta})}+\sqrt{2\epsilon'})nLD_X/(n_+n_-\epsilon')$ and $\Delta=\frac{\beta D_X^2}{n_+n_-(e^{\frac{\epsilon}{2n}}-1)}-\lambda$, on the event $\mathcal{E}^c$.

Therefore, for any set $E\subseteq\mathbb{R}^d$

$$\begin{aligned}
\Pr(\mathbf{w}_{\text{priv}}(S)\in E)&=\Pr(\mathbf{w}_{\text{priv}}(S)\in E\cap\mathcal{E})+\Pr(\mathbf{w}_{\text{priv}}(S)\in E\cap\mathcal{E}^c)\\
&\le\Pr(\mathcal{E})+\Pr(\mathbf{w}_{\text{priv}}(S)\in E\cap\mathcal{E}^c)\\
&\le\delta+\int_{E\cap\mathcal{E}^c}\text{pdf}(\mathbf{w}_{\text{priv}}=\alpha|S))d\alpha\\
&\le\delta+e^{\epsilon}\int_{E\cap\mathcal{E}^c}\text{pdf}(\mathbf{w}_{\text{priv}}=\alpha|S'))d\alpha\\
&\le\delta+e^{\epsilon}\Pr(\mathbf{w}_{\text{priv}}(S')\in E).
\end{aligned}$$

This completes the proof of the theorem. □

4. Generalization Performance

In this section, we systematically study the generalization (utility) guarantees for `Algorithm 1` and `Algorithm 2`. To this end, let the population (true) risk for AUC maximization be defined by $\mathcal{R}(\mathbf{w})=\mathbb{E}[\ell(\mathbf{w}^T(\mathbf{x}-\mathbf{x}'))|y=1,y'=-1]$ which is identical to $\frac{1}{\Pr(y=1)\Pr(y'=-1))}\mathbb{E}[\ell(\mathbf{w}^T(\mathbf{x}-\mathbf{x}'))\mathbb{I}_{[y=1\wedge y'=-1]}]$. Its regularized population risk is defined by $\mathcal{R}^\lambda(\mathbf{w})=\mathcal{R}(\mathbf{w})+\frac{\lambda}{2}\|\mathbf{w}\|^2$, and let $\mathbf{w}_\lambda=\arg\inf_{\mathbf{w}\in\mathbb{R}^d}\mathcal{R}^\lambda(\mathbf{w})$, and $\mathbf{w}^*=\arg\inf_{\mathbf{w}\in\mathbb{R}^d}\mathcal{R}(\mathbf{w})$. The generalization analysis aims to examine the *excess population risk* which is the difference between the risks of the private estimator $\mathbf{w}_{\text{priv}}$ and the best possible one, i.e. $\mathcal{R}(\mathbf{w}_{\text{priv}})-\inf_{\mathbf{w}\in\mathbb{R}^d}\mathcal{R}(\mathbf{w})$.

The generalization analysis for the proposed algorithms critically rely on the following novel fast rates for strongly convex objectives in AUC maximization.

Lemma 3. *Let* $\mathcal{B}=\{\mathbf{w}\in\mathbb{R}^d:\sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}}|\ell'(\mathbf{w}^T(\mathbf{x}-\mathbf{x}'))|\le B\}$, *where* $B>0$ *is a constant. For any* $0<\tau<1$, *with probability*

at least $1-\delta$*, we have, for any* $\mathbf{w}\in\mathcal{B}$*, that*

$$\mathcal{R}^\lambda(\mathbf{w})-\mathcal{R}^\lambda(\mathbf{w}_\lambda)\le\frac{1}{\tau}(\mathcal{R}_S^\lambda(\mathbf{w})-\mathcal{R}_S^\lambda(\mathbf{w}_\lambda))+O\Big(\frac{B^2D_{\mathcal{X}}^2\log(\frac{1}{\delta})n^3}{\lambda\tau(1-\tau)(n_+n_-)^2}\Big)$$
$$\le\frac{1}{\tau}(\mathcal{R}_S^\lambda(\mathbf{w})-\mathcal{R}_S^\lambda(\widehat{\mathbf{w}}))+O\Big(\frac{B^2D_{\mathcal{X}}^2\log(\frac{1}{\delta})n^3}{\lambda\tau(1-\tau)(n_+n_-)^2}\Big).$$

The above lemma is inspired by [29] for the case of pointwise learning using the peeling techniques for Rademacher averages [1]. The novelty in the proof for Lemma 3 is to use the decoupling techniques of U-Statistics [2, 3] to handle the pairwise loss in AUC maximization. The detailed proof of Lemma 3 can be found in Appendix C.2.

4.1. Utility Analysis for Output Perturbation

In this subsection, we present the generalization analysis for the output perturbation given by `Algorithm 1` which applies to any convex loss such as the logistic loss, least square loss and Huber loss. In particular, we will first start with L-Lipschitz and β-smooth losses (e.g. logistic loss and Huber loss), and then consider the smooth losses such as the least square loss.

Firstly, we consider the case when the loss ℓ is Lipschitz and strongly smooth. To this end, let $\mathbb{E}_{\mathbf{b}}[\cdot]$ denote the expectation w.r.t. the noise $\mathbf{b}$. Using the error decomposition often used in the literature (e.g. [4]) to examine the quantity $\mathcal{R}(\mathbf{w}_{\text{priv}})-\inf_{\mathbf{w}}\mathcal{R}(\mathbf{w})$, i.e.

$$\begin{aligned}&\mathcal{R}(\mathbf{w}_{\text{priv}})-\inf_{\mathbf{w}\in\mathbb{R}^d}\mathcal{R}(\mathbf{w})\\&\le[\mathcal{R}^\lambda(\mathbf{w}_{\text{priv}})-\mathcal{R}^\lambda(\mathbf{w}_\lambda)]+[\mathcal{R}^\lambda(\mathbf{w}_\lambda)-\mathcal{R}^\lambda(\mathbf{w}^*)]+\frac{\lambda}{2}\|\mathbf{w}^*\|^2\\&\le[\mathcal{R}^\lambda(\mathbf{w}_{\text{priv}})-\mathcal{R}^\lambda(\mathbf{w}_\lambda)]+\frac{\lambda}{2}\|\mathbf{w}^*\|^2.\end{aligned}\tag{17}$$

where the last inequality used the fact $\mathcal{R}^\lambda(\mathbf{w}_\lambda)\le\mathcal{R}^\lambda(\mathbf{w}^*)$ from the definition of $\mathbf{w}_\lambda$. Therefore,

$$\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})]-\inf_{\mathbf{w}\in\mathbb{R}^d}\mathcal{R}(\mathbf{w})\le\mathbb{E}_{\mathbf{b}}[\mathcal{R}^\lambda(\mathbf{w}_{\text{priv}})-\mathcal{R}^\lambda(\mathbf{w}_\lambda)]+\frac{\lambda}{2}\|\mathbf{w}^*\|^2.\tag{18}$$

Here we assume class 1 is the minority class while class -1 is the majority class, and let the *imbalanced ratio* $\rho=\frac{n_+}{n}$ which means that $\rho\le\frac{1}{2}$ and $n_+=\rho n$ and $n_-=(1-\rho)n$. We have the following generalization bound.

Theorem 3. *If the loss function* ℓ *is* L*-Lipschitz and* β*-strongly smooth, then the output* $\mathbf{w}_{priv}$ *of* `Algorithm 1 (Output-Pert-AUC)` *has the following properties.*

(a) For ϵ*-differential privacy, choosing* $\lambda=\min\{\frac{\beta^{\frac{1}{3}}(LD_{\mathcal{X}}nd)^{\frac{2}{3}}}{(\|\mathbf{w}^*\|\epsilon n_+n_-)^{2/3}},\frac{LD_{\mathcal{X}}\sqrt{\log(\frac{1}{\xi})}n^{\frac{3}{2}}}{\|\mathbf{w}^*\|n_+n_-}\}$ *implies, with probability at least* $1-\xi$*, that*

$$\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{priv})]-\inf_{\mathbf{w}\in\mathbb{R}^d}\mathcal{R}(\mathbf{w})$$
$$=O(\max\{\frac{\beta^{\frac{1}{3}}(LD_{\mathcal{X}}d)^{\frac{2}{3}}\|\mathbf{w}^*\|^{\frac{4}{3}}}{(\rho(1-\rho)\epsilon n)^{2/3}},\frac{LD_{\mathcal{X}}\sqrt{\log(\frac{1}{\xi})}\|\mathbf{w}^*\|}{\rho(1-\rho)\sqrt{n}}\}).$$

(b) For (ϵ,δ)*-differential privacy, choosing* $\lambda=\min\Big\{\frac{(\log(\frac{1}{\delta})\beta)^{\frac{1}{3}}(LD_{\mathcal{X}}n)^{\frac{2}{3}}d^{\frac{1}{3}}}{(\|\mathbf{w}^*\|\epsilon(n_+n_-))^{2/3}},\frac{LD_{\mathcal{X}}\sqrt{\log(\frac{1}{\xi})}n^{\frac{3}{2}}}{\|\mathbf{w}^*\|(n_+n_-)}\Big\}$ *yields, with probability at least* $1-\xi$*, that*

$$\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{priv})]-\inf_{\mathbf{w}\in\mathbb{R}^d}\mathcal{R}(\mathbf{w})$$
$$=O(\max\{\frac{(\log(\frac{1}{\delta})\beta)^{\frac{1}{3}}(LD_{\mathcal{X}})^{\frac{2}{3}}d^{\frac{1}{3}}\|\mathbf{w}^*\|^{\frac{4}{3}}}{(\rho(1-\rho)\epsilon n)^{2/3}},\frac{LD_{\mathcal{X}}\sqrt{\log(\frac{1}{\xi})}\|\mathbf{w}^*\|}{\rho(1-\rho)\sqrt{n}}\}).$$

The proof of Theorem 3 can be found in Appendix C.3.

Remark 1. We note that, in Theorem 3, the choice of λ depends on the norm of $\mathbf{w}^*$. This would require prior knowledge of $\|\mathbf{w}^*\|$, which may be not realistic in practice as the population distribution is unknown. To mitigate this limitation, taking part (a) in Theorem 1 as an example, and one can instead choose $\lambda=\min\{\frac{\beta^{\frac{1}{3}}(LD_{\mathcal{X}}dn)^{\frac{2}{3}}}{(\epsilon n_+n_-)^{2/3}},\frac{LD_{\mathcal{X}}\sqrt{\log(\frac{1}{\xi})}n^{\frac{3}{2}}}{n_+n_-}\}$ which does not require the knowledge of $\mathbf{w}^*$. The resulting rate is of the same order, i.e. $(1+\|\mathbf{w}^*\|)^2\cdot O(\max\{\frac{\beta^{\frac{1}{3}}(LD_{\mathcal{X}}d)^{\frac{2}{3}}}{(\rho(1-\rho)\epsilon n)^{2/3}},\frac{LD_{\mathcal{X}}\sqrt{\log(\frac{1}{\xi})}}{\rho(1-\rho)\sqrt{n}}\})$.

Secondly, we move on to consider the utility guarantee of `Algorithm 1` when the loss ℓ has only strongly smoothness property (e.g. the least square loss). As we see from the error decomposition (17), the critical part is to estimate the term $\mathcal{R}^\lambda(\mathbf{w}_{\text{priv}})-\mathcal{R}^\lambda(\mathbf{w}_\lambda)$ using Lemma 3 which has the requirement that $\sup_{\mathbf{x},\mathbf{x}'}|\ell'(\mathbf{w}_{\text{priv}}^T(\mathbf{x}-\mathbf{x}'))|$ is uniformly bounded. However, if we only know the loss is strongly smooth, $\sup_{\mathbf{x},\mathbf{x}'}|\ell'(\mathbf{w}_{\text{priv}}^T(\mathbf{x}-\mathbf{x}'))|$ can be unbounded as $\mathbf{w}_{\text{priv}}$ is unbounded. As a result, the error decomposition (17) does not apply to this case. To overcome this hindrance, we decompose $\mathcal{R}(\mathbf{w}_{\text{priv}})-\inf_{\mathbf{w}}\mathcal{R}(\mathbf{w})$ as follows,

$$\begin{aligned}&\mathcal{R}(\mathbf{w}_{\text{priv}})-\inf_{\mathbf{w}}\mathcal{R}(\mathbf{w})\\&\le[\mathcal{R}(\mathbf{w}_{\text{priv}})-\mathcal{R}(\widehat{\mathbf{w}})]+[\mathcal{R}^\lambda(\widehat{\mathbf{w}})-\mathcal{R}^\lambda(\mathbf{w}_\lambda)]+[\mathcal{R}^\lambda(\mathbf{w}_\lambda)-\mathcal{R}^\lambda(\mathbf{w}^*)]\\&\quad+\frac{\lambda}{2}\|\mathbf{w}^*\|^2\\&\le[\mathcal{R}(\mathbf{w}_{\text{priv}})-\mathcal{R}(\widehat{\mathbf{w}})]+[\mathcal{R}^\lambda(\widehat{\mathbf{w}})-\mathcal{R}^\lambda(\mathbf{w}_\lambda)]+\frac{\lambda}{2}\|\mathbf{w}^*\|^2,\end{aligned}\tag{19}$$

where the last inequality follows from the fact that $\mathcal{R}^\lambda(\mathbf{w}_\lambda)\le\mathcal{R}^\lambda(\mathbf{w}^*)$. In contrast to (17), we can now show that $\widehat{\mathbf{w}}=\arg\inf_{\mathbf{w}}\mathcal{R}_S^\lambda(\mathbf{w})$ is uniformly bounded, and so does $\sup_{\mathbf{x},\mathbf{x}'}|\ell'(\widehat{\mathbf{w}}^T(\mathbf{x}-\mathbf{x}'))|$. As a result, we can apply Lemma 3 to estimate the term $\mathcal{R}^\lambda(\widehat{\mathbf{w}})-\mathcal{R}^\lambda(\mathbf{w}_\lambda)$ in (19).

The next lemma estimates the upper bound of the term $\mathcal{R}(\mathbf{w}_{\text{priv}})-\mathcal{R}(\widehat{\mathbf{w}})$.

Lemma 4. *If* ℓ *is non-negative and* β*-strongly smooth, the following are true:*

(a) For $\epsilon-$*differential privacy, with probability at least* $1-\xi$*, there holds*

$$\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{priv})-\mathcal{R}(\widehat{\mathbf{w}})]$$
$$=O\Big(\frac{nD_{\mathcal{X}}^4\beta^2d\|\mathbf{w}^*\|}{(n_+n_-)\epsilon\lambda^{\frac{3}{2}}}+\frac{n^{\frac{5}{2}}D_{\mathcal{X}}^5\beta^{\frac{5}{2}}d\sqrt{\log(\frac{1}{\xi})}}{(n_+n_-)^2\epsilon\lambda^{\frac{5}{2}}}+\frac{n^2D_{\mathcal{X}}^4\beta^3(d+d^2)}{(n_+n_-)^2\epsilon^2\lambda^3}\Big).$$

(b) For (ϵ, δ)−differential privacy, with probability at least $1 - \xi$, there holds

$$\begin{aligned}&\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{priv}) - \mathcal{R}(\widehat{\mathbf{w}})] \\ &= O\Big(\frac{nD_{\mathcal{X}}^4\beta^2\|\mathbf{w}^*\|\sqrt{d\log(\frac{1}{\delta})}}{(n_+n_-)\epsilon\lambda^{\frac{3}{2}}} + \frac{n^{\frac{5}{2}}D_{\mathcal{X}}^5\beta^{\frac{5}{2}}\sqrt{d\log(\frac{1}{\delta})\log(\frac{1}{\xi})}}{(n_+n_-)^2\epsilon\lambda^{\frac{5}{2}}} \\ &\quad + \frac{n^2D_{\mathcal{X}}^4\beta^2 d\log(\frac{1}{\delta})}{(n_+n_-)^2\epsilon^2\lambda^3}\Big).\end{aligned}$$

The proof of Lemma 4 is provided in Appendix C.4.

Combining the above lemma and our new error decomposition, we have the following generalization bounds for the output $\mathbf{w}_{\text{priv}}$ of `Algorithm` 1 with strongly smooth losses.

Theorem 4. *If the loss function ℓ is β-strongly smooth, then the output $\mathbf{w}_{priv}$ of `Algorithm 1 (Output-Pert-AUC)` has the following properties.*

(a) For ϵ-DP, setting $\lambda = \frac{D_{\mathcal{X}}^{\frac{8}{5}}\beta^{\frac{4}{5}}(nd)^{\frac{2}{5}}}{(n_+n_-)^{\frac{2}{5}}\epsilon^{\frac{2}{5}}\|\mathbf{w}^*\|^{\frac{2}{5}}}$ implies, with probability at least $1 - 2\xi$, that

$$\begin{aligned}\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{priv})] - \inf_{\mathbf{w}}\mathcal{R}(\mathbf{w}) = O\Big(\max\{&\frac{D_{\mathcal{X}}^{\frac{8}{5}}\beta^{\frac{4}{5}}d^{\frac{2}{5}}\|\mathbf{w}^*\|^{\frac{8}{5}}}{(\rho(1-\rho)\epsilon n)^{\frac{2}{5}}}, \\ &\frac{D_{\mathcal{X}}^{\frac{4}{5}}\beta^{\frac{2}{5}}\log(\frac{1}{\xi})\|\mathbf{w}^*\|^{\frac{4}{5}}}{(\rho(1-\rho))^{\frac{6}{5}}n^{\frac{1}{5}}d^{\frac{4}{5}}}, \frac{D_{\mathcal{X}}\sqrt{\beta\log(\frac{1}{\xi})}\|\mathbf{w}^*\|}{\rho(1-\rho)\sqrt{n}}\}\Big).\end{aligned}$$

(b) For (ϵ, δ)-DP, choosing $\lambda = \frac{D_{\mathcal{X}}^{\frac{4}{3}}\beta^{\frac{2}{3}}(\log(\frac{1}{\xi}))^{\frac{1}{3}}n}{(n_+n_-)^{\frac{2}{3}}\|\mathbf{w}^*\|^{\frac{2}{3}}}$ yields, with probability at least $1 - 2\xi$, that

$$\begin{aligned}&\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{priv})] - \inf_{\mathbf{w}}\mathcal{R}(\mathbf{w}) \\ &= O\Big(\max\{\frac{D_{\mathcal{X}}^2\beta\log(\frac{1}{\delta})\sqrt{d}\|\mathbf{w}^*\|^2}{(\rho(1-\rho))^{\frac{1}{3}}\epsilon^2\sqrt{n}(\log(\frac{1}{\xi}))^{\frac{1}{3}}}, \frac{D_{\mathcal{X}}^{\frac{4}{3}}\beta^{\frac{2}{3}}(\log(\frac{1}{\xi}))^{\frac{1}{3}}\|\mathbf{w}^*\|^{\frac{4}{3}}}{(\rho(1-\rho))^{\frac{2}{3}}n^{\frac{1}{3}}}\}\Big).\end{aligned}$$

Proof. First, we show that $\widehat{\mathbf{w}} \in \mathcal{B}$ with $B = (\beta D_{\mathcal{X}}\sqrt{2/\lambda} + 2\sqrt{\beta})$, where $\mathcal{B}$ is defined in Lemma 3. Using the properties of strongly smoothness, we have $|\ell'(s) - \ell'(0)| \leq \beta|s|$ and $|\ell'(0)| \leq 2\sqrt{\beta\ell(0)}$, thus $|\ell'(s)| \leq 2\sqrt{\beta} + \beta|s|$ as $\ell(0) = 1$. Notice that $\|\widehat{\mathbf{w}}\| \leq \sqrt{2/\lambda}$ and $D_{\mathcal{X}} = \sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}}\|\mathbf{x} - \mathbf{x}'\|$, there holds $|\ell'(\widehat{\mathbf{w}}^T(\mathbf{x} - \mathbf{x}'))| \leq (\beta D_{\mathcal{X}}\sqrt{2/\lambda} + 2\sqrt{\beta})$ for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. Hence, $\widehat{\mathbf{w}} \in \mathcal{B}$. Now, let $\tau = \frac{1}{2}$, and then the error decomposition (19) and Lemma 3 imply, with probability at least $1 - \xi$, that

$$\begin{aligned}&\mathcal{R}(\mathbf{w}_{\text{priv}}) - \inf_{\mathbf{w}}\mathcal{R}(\mathbf{w}) \\ &\leq [\mathcal{R}(\mathbf{w}_{\text{priv}}) - \mathcal{R}(\widehat{\mathbf{w}})] + O\Big(\frac{(\beta D_{\mathcal{X}}\sqrt{2/\lambda} + 2\sqrt{\beta})^2D_{\mathcal{X}}^2\log(\frac{1}{\xi})n^3}{\lambda(n_+n_-)^2}\Big) \\ &\quad + \frac{\lambda}{2}\|\mathbf{w}^*\|^2.\end{aligned}$$

Applying part (a) in Lemma 4 below, and setting $\lambda = \frac{D_{\mathcal{X}}^{\frac{8}{5}}\beta^{\frac{4}{5}}(nd)^{\frac{2}{5}}}{(n_+n_-)^{\frac{2}{5}}\epsilon^{\frac{2}{5}}\|\mathbf{w}^*\|^{\frac{2}{5}}}$, then, for ϵ-DP we have, with probability at least $1 - 2\xi$, that

$$\begin{aligned}\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}}\mathcal{R}(\mathbf{w}) = O\Big(\max\{&\frac{D_{\mathcal{X}}^{\frac{8}{5}}\beta^{\frac{4}{5}}d^{\frac{2}{5}}\|\mathbf{w}^*\|^{\frac{8}{5}}}{(\rho(1-\rho)\epsilon n)^{\frac{2}{5}}}, \\ &\frac{D_{\mathcal{X}}^{\frac{4}{5}}\beta^{\frac{2}{5}}\log(\frac{1}{\xi})\|\mathbf{w}^*\|^{\frac{4}{5}}}{(\rho(1-\rho))^{\frac{6}{5}}n^{\frac{1}{5}}d^{\frac{4}{5}}}, \frac{D_{\mathcal{X}}\sqrt{\beta\log(\frac{1}{\xi})}\|\mathbf{w}^*\|}{\rho(1-\rho)\sqrt{n}}\}\Big).\end{aligned}$$

Applying part (b) in Lemma 4 below, and $\lambda = \frac{D_{\mathcal{X}}^{\frac{4}{3}}\beta^{\frac{2}{3}}(\log(\frac{1}{\xi}))^{\frac{1}{3}}n}{(n_+n_-)^{\frac{2}{3}}\|\mathbf{w}^*\|^{\frac{2}{3}}}$, then, for (ϵ, δ)-DP we have, with probability at least $1 - 2\xi$, that

$$\begin{aligned}&\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}}\mathcal{R}(\mathbf{w}) \\ &= O\Big(\max\{\frac{D_{\mathcal{X}}^2\beta\log(\frac{1}{\delta})\sqrt{d}\|\mathbf{w}^*\|^2}{(\rho(1-\rho))^{\frac{1}{3}}\epsilon^2\sqrt{n}(\log(\frac{1}{\xi}))^{\frac{1}{3}}}, \frac{D_{\mathcal{X}}^{\frac{4}{3}}\beta^{\frac{2}{3}}(\log(\frac{1}{\xi}))^{\frac{1}{3}}\|\mathbf{w}^*\|^{\frac{4}{3}}}{(\rho(1-\rho))^{\frac{2}{3}}n^{\frac{1}{3}}}\}\Big).\end{aligned}$$

This completes the proof of the theorem. □

4.2. Utility Analysis for Objective Perturbation

Now, we turn our attention to the generalization analysis for `Algorithm` 2.

Theorem 5. *Assume that ℓ is convex and twice-differentiable, L-Lipschitz and β-strongly smoothness. We have the following properties for the output $\mathbf{w}_{priv}$ of `Algorithm 2 (Obj-Pert-AUC)`.*

(a) For ϵ-DP, suppose that $\beta \leq \frac{L\sqrt{n\epsilon^2\log(\frac{1}{\xi})+d}}{2D_{\mathcal{X}}\|\mathbf{w}^*\|}$, then choosing $\lambda = \frac{LD_{\mathcal{X}}n\sqrt{d^2+n\log(\frac{1}{\xi})\epsilon^2}}{n_+n_-\epsilon\|\mathbf{w}^*\|}$ implies, with probability at least $1 - \xi$, that

$$\begin{aligned}&\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{priv})] - \inf_{\mathbf{w}\in\mathbb{R}^d}\mathcal{R}(\mathbf{w}) \\ &= O(\max\{\frac{LD_{\mathcal{X}}d\|\mathbf{w}^*\|}{\rho(1-\rho)\epsilon n}, \frac{LD_{\mathcal{X}}\sqrt{\log(\frac{1}{\xi})}\|\mathbf{w}^*\|}{\rho(1-\rho)\sqrt{n}}\}).\end{aligned}$$

(b) For (ϵ, δ)-DP, assume that $\beta \leq \frac{L\left(\sqrt{\log(\frac{1}{\delta})}+\sqrt{\epsilon}\right)^2d+n\log(\frac{1}{\xi})\epsilon^2)^{1/2}}{2D_{\mathcal{X}}\|\mathbf{w}^*\|}$, then selecting $\lambda = \frac{LD_{\mathcal{X}}n\left(\sqrt{\log(\frac{1}{\delta})}+\sqrt{\epsilon}\right)^2d+n\log(\frac{1}{\xi})\epsilon^2)^{1/2}}{\epsilon n_+n_-\|\mathbf{w}^*\|}$ implies, with probability at least $1 - \xi$, that

$$\begin{aligned}&\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{priv})] - \inf_{\mathbf{w}\in\mathbb{R}^d}\mathcal{R}(\mathbf{w}) \\ &= O(\max\{\frac{LD_{\mathcal{X}}(\sqrt{\log(\frac{1}{\delta})} + \sqrt{\epsilon})\sqrt{d}\|\mathbf{w}^*\|}{\rho(1-\rho)\epsilon n}, \frac{LD_{\mathcal{X}}\sqrt{\log(\frac{1}{\xi})}\|\mathbf{w}^*\|}{\rho(1-\rho)\sqrt{n}}\}).\end{aligned}$$

The detailed proof of Theorem 5 can be found in Appendix C.5.

Remark 2. Firstly, from Theorems 3 and 5, one can observe that the bounds for the excess population risk for the objective perturbation are consistently better than the output perturbation. Secondly, similar to the discussion right after Theorem 3 on the choice λ, we can see from the proof of Theorem 5 in Appendix C.5 that one can choose λ independent of

Algorithm 3 Objective Perturbation for ERM (`Obj-Pert`)

Inputs: Data $S = \{(\mathbf{x}_i, y_i) : i = 1, \dots, n\}$ and parameters $\lambda, \epsilon, \delta, L, \beta$

if $\log(1 + \frac{\beta R_2^2}{n\lambda}) < \epsilon$ **then**

  $\Delta = 0$ and $\epsilon' = \epsilon - \log(1 + \frac{\beta R_2^2}{n\lambda})$ and $\sigma = (2\sqrt{2\log(\frac{1}{\delta})} + \sqrt{2\epsilon'})LR_2/\epsilon'$

else if $\log(1 + \frac{\beta R_2^2}{n\lambda}) \geq \epsilon$ **then**

  choose $\epsilon' = \frac{\epsilon}{2}$, $\Delta = \frac{\beta R_2^2}{n(e^{\epsilon/2}-1)} - \lambda$, and $\sigma = (2\sqrt{2\log(\frac{1}{\delta})} + \sqrt{2\epsilon'})LR_2/\epsilon'$

end if

sample $\mathbf{b}$ from $\nu(\mathbf{b}; \epsilon, \delta, \sigma) = \mathcal{N}(\mathbf{0}, \sigma^2\mathbb{I})$.

return: $\mathbf{w}_{\text{priv}} = \arg\min\{J(\mathbf{w}, S) + \frac{\Delta}{2}\|\mathbf{w}\|^2 + \frac{\mathbf{b}^T\mathbf{w}}{n}\}$

the $\|\mathbf{w}^*\|$ and in the corresponding generalization bounds, the term $\|\mathbf{w}^*\|$ is replaced by $(1 + \|\mathbf{w}^*\|)^2$. For instance, in the case of (ϵ, δ)-DP, choosing $\lambda = \frac{LD_{\mathcal{X}}n\left(\sqrt{\log(\frac{1}{\delta})} + \sqrt{\epsilon}\right)^2 d + n\log(\frac{1}{\xi})\epsilon^2\right)^{1/2}}{\epsilon n_+ n_-}$, then, if $\beta \leq \frac{L\left(\sqrt{\log(\frac{1}{\delta})} + \sqrt{\epsilon}\right)^2 d + n\log(\frac{1}{\xi})\epsilon^2\right)^{1/2}}{2D_{\mathcal{X}}}$, with probability $1 - \xi$, $\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}\in\mathbb{R}^d}\mathcal{R}(\mathbf{w}) = (1 + \|\mathbf{w}^*\|)^2 \cdot O(\max\{\frac{LD_{\mathcal{X}}(\sqrt{\log(\frac{1}{\delta})} + \sqrt{\epsilon})\sqrt{d}}{\rho(1-\rho)\epsilon n}, \frac{LD_{\mathcal{X}}\sqrt{\log(\frac{1}{\xi})}}{\rho(1-\rho)\sqrt{n}}\})$.

5. Discussion on Differentially Private ERM with Pointwise Learning

In this section, we firstly revisit the utility bounds (excess population risk) for the objective perturbation in [4, 28, 5] for the standard regularized ERM with pointwise losses. In particular, we show that the optimal rates can be obtained for (ϵ, δ)-DP using the uniform convergence analysis, which was recently established using stabililty analysis in [20]. We will also discuss the relation with the recent results in [20, 21, 22].

5.1. Revisiting the Differentially Private ERM with Pointwise Loss

In this section, we firstly reprove the utility bounds (excess population risk) in [4, 5, 28] for the standard regularized ERM with pointwise losses. In particular, we show that the bound for (ϵ, δ)-DP is indeed optimal using this uniform convergence analysis which was recently established using stabililty analysis in [20]. We should mention that the algorithms are just restatements of those in [4, 5] in our setting, and the proofs are mainly adapted from those papers. We revisited these algorithms and proofs just for completeness.

To illustrate the results clearly, let the population risk and its empirical risk is given by $\mathcal{L}(\mathbf{w}) = \mathbb{E}[\ell(y\mathbf{w}^T\mathbf{x})]$ and $\mathcal{L}(\mathbf{w}, S) = \frac{1}{n}\sum_{i=1}^n \ell(y_i\mathbf{w}^T\mathbf{x}_i)$, respectively. Let $J(\mathbf{w}, S) = \mathcal{L}(\mathbf{w}, S) + \frac{\lambda}{2}\|\mathbf{w}\|^2$, and $J(\mathbf{w}) = \mathcal{L}(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2$. In addition, let $\widehat{\mathbf{w}} = \arg\inf_{\mathbf{w}} J(\mathbf{w}, S)$, $\mathbf{w}^\lambda = \arg\inf_{\mathbf{w}} J(\mathbf{w})$ and $\mathbf{w}^* = \arg\inf_{\mathbf{w}} \mathcal{L}(\mathbf{w})$, and $R_2 = \sup_{\mathbf{x}\in\mathcal{X}}\|\mathbf{x}\|$. For simplicity, we only restrict our attention within the case of the regularizer $\|\mathbf{w}\|^2$, and the loss ℓ is L-Lipschitz, twice differentiable and β-smoothness; but all the results below may hold true for any (possibly non-differentiable) strongly convex regularizer for $\mathbf{w}$ and/or not twice-differentiable losses following the successive approximation argument in [5].

Privacy Guarantees. One can immediately get the following results on privacy guarantees. The proof is essentially from [4, 5] which is given in Appendix C.6 for completeness.

Theorem 6. *[5] Suppose that the loss function ℓ is L-Lipschitz, twice differentiable and β-smooth, then Algorithm 3 is (ϵ, δ)-DP.*

Generalization Performance. We will revisit the objective perturbation in [5] to derive excess population risks. For simplicity, in this section we assume that the pointwise loss ℓ is twice differentiable, L-Lipschitz and β-smooth. The error decomposition (17) in this setting can be restated as

$$\mathcal{L}(\mathbf{w}_{\text{priv}}) - \inf_{\mathbf{w}}\mathcal{L}(\mathbf{w}) \leq [J(\mathbf{w}_{\text{priv}}) - J(\mathbf{w}_\lambda)] + \frac{\lambda}{2}\|\mathbf{w}^*\|^2.$$

Applying the results of [29], we have, with probability $1 - \xi$ over the choice of the data S,

$$J(\mathbf{w}_{\text{priv}}) - J(\mathbf{w}_\lambda) \leq 2[J(\mathbf{w}_{\text{priv}}, S) - J(\widehat{\mathbf{w}}, S)] + O(\frac{L^2R_2^2\log(1/\xi)}{\lambda n}). \tag{20}$$

Consequently,

$$\mathcal{L}(\mathbf{w}_{\text{priv}}) - \inf_{\mathbf{w}}\mathcal{L}(\mathbf{w}) \leq 2[J(\mathbf{w}_{\text{priv}}, S) - J(\widehat{\mathbf{w}}, S)] + O(\frac{L^2R_2^2\log(1/\xi)}{\lambda n}) + \frac{\lambda}{2}\|\mathbf{w}^*\|^2. \tag{21}$$

Now using the above inequality and different estimations for $J(\mathbf{w}_{\text{priv}}, S) - J(\widehat{\mathbf{w}}, S)$ for objective perturbation, we can show the following theorems.

Theorem 7. *[5] Assume that ℓ is convex and twice-differentiable with $|\ell'(t)| \leq L$ and β-smoothness. We have the following properties for the output $\mathbf{w}_{priv}$ of Algorithm 3 (`Obj-Pert`). Assume that $\beta \leq \frac{4L\left(\sqrt{\log(\frac{1}{\delta})} + \sqrt{\epsilon}\right)^2 d + n\log(\frac{1}{\xi})\epsilon^2\right)^{1/2}}{R_2\|\mathbf{w}^*\|}$, then selecting $\lambda = \frac{8LR_2\sqrt{\left(\sqrt{\log(\frac{1}{\delta})} + \sqrt{\epsilon}\right)^2 d + n\log(\frac{1}{\xi})\epsilon^2}}{\epsilon n\|\mathbf{w}^*\|}$ implies, with probability at least $1 - \xi$, that*

$$\mathbb{E}_{\mathbf{b}}[\mathcal{L}(\mathbf{w}_{priv})] - \inf_{\mathbf{w}\in\mathbb{R}^d}\mathcal{L}(\mathbf{w}) = O\Big(\max\Big\{\frac{LR_2(\sqrt{\log(\frac{1}{\delta})} + \sqrt{\epsilon})\sqrt{d}\|\mathbf{w}^*\|}{\epsilon n}, \frac{LR_2\sqrt{\log(\frac{1}{\xi})}\|\mathbf{w}^*\|}{\sqrt{n}}\Big\}\Big).$$

The proof can be found in Appendix C.7.

Remark 3. The choice of λ in Theorem 7 depends on $\|\mathbf{w}^*\|$. Instead, one can choose $\lambda = \frac{8LR_2\sqrt{\left(\sqrt{\log(\frac{1}{\delta})} + \sqrt{\epsilon}\right)^2 d + n\log(\frac{1}{\xi})\epsilon^2}}{\epsilon n}$. Then, if $\beta \leq \frac{\lambda\epsilon n}{2R_2^2} = \frac{4L\sqrt{\left(\sqrt{\log(\frac{1}{\delta})} + \sqrt{\epsilon}\right)^2 d + n\log(\frac{1}{\xi})\epsilon^2}}{R_2}$, we have, with probability at least $1 - \xi$, that

$$\mathbb{E}_{\mathbf{b}}[\mathcal{L}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}}\mathcal{L}(\mathbf{w}) = (1 + \|\mathbf{w}^*\|)^2 \cdot O\Big(\max\Big\{\frac{LR_2(\sqrt{\log(\frac{1}{\delta})} + \sqrt{\epsilon})\sqrt{d}}{\epsilon n}, \frac{LR_2\sqrt{\log(\frac{1}{\xi})}}{\sqrt{n}}\Big\}\Big).$$

5.2. Discussion

Recently there are appealing and important work on considering stochastic optimization algorithms for differential privacy [20, 22] using the approach of algorithmic stability [45, 46]. To be more precise, let $W = \sup_{\mathbf{w}\in\mathcal{W}} \|\mathbf{w}\|$ and the outputs $\mathbf{w}_{\text{priv}}$ of the private algorithms are assumed from $\mathcal{W}$. Then, in these studies, the authors considered $\mathcal{L}(\mathbf{w}_{\text{priv}}) - \inf_{\mathbf{w}\in\mathcal{W}} \mathcal{L}(\mathbf{w})$, i.e. the discrepancy between the risk of the private estimator $\mathbf{w}_{\text{priv}}$ and the best possible in the constrained set $\mathcal{W}$, to which we refer as the *excess population risk in* $\mathcal{W}$. In contrast, we consider here the excess population risks $\mathcal{L}(\mathbf{w}_{\text{priv}}) - \inf_{\mathbf{w}\in\mathbb{R}^d} \mathcal{L}(\mathbf{w})$, i.e. the difference between the risk of the private estimator $\mathbf{w}_{\text{priv}}$ and the best possible. The difference between these two excess errors is the term $\texttt{Approx}(\mathcal{W}) = \inf_{\mathbf{w}\in\mathcal{W}} \mathcal{L}(\mathbf{w}) - \inf_{\mathbf{w}\in\mathbb{R}^d} \mathcal{L}(\mathbf{w})$ which we call it *approximation error* as it is deterministic and measures the difference between the least risk in $\mathcal{W}$ and the least one in $\mathbb{R}^d$. Let $\mathbf{w}^* = \arg\inf_{\mathbf{w}\in\mathbb{R}^d} \mathcal{L}(\mathbf{w})$ be the best possible parameter. By L-Lipschitzness of the loss ℓ, $\texttt{Approx}(\mathcal{W}) = \inf_{\mathbf{w}\in\mathcal{W}} \mathcal{L}(\mathbf{w}) - \inf_{\mathbf{w}\in\mathbb{R}^d} \mathcal{L}(\mathbf{w}) = \inf_{\mathbf{w}\in\mathcal{W}} \mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}^*) \leq LR_2 \inf_{\mathbf{w}\in\mathcal{W}} \|\mathbf{w} - \mathbf{w}^*\|$. Therefore,

$$\mathcal{L}(\mathbf{w}_{\text{priv}}) - \inf_{\mathbf{w}\in\mathbb{R}^d} \mathcal{L}(\mathbf{w}) \leq [\mathcal{L}(\mathbf{w}_{\text{priv}}) - \inf_{\mathbf{w}\in\mathcal{W}} \mathcal{L}(\mathbf{w})] + LR_2 \inf_{\mathbf{w}\in\mathcal{W}} \|\mathbf{w} - \mathbf{w}^*\|. \tag{22}$$

In particular, the very recent work [20] proved the following bounds for both stochastic gradient descent and the exact ERM with pointwise loss using the stability analysis and proved $\mathbb{E}_{\mathbf{b}}[\mathcal{L}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}\in\mathcal{W}} \mathcal{L}(\mathbf{w}) = WL \cdot O(\max\{\frac{1}{\sqrt{n}}, \frac{\sqrt{d\log(\frac{1}{\delta})}}{n\epsilon}\})$. Translating their bounds to our setting, we have that $\mathbb{E}_{\mathbf{b}}[\mathcal{L}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}\in\mathbb{R}^d} \mathcal{L}(\mathbf{w}) = WL \cdot O(\max\{\frac{1}{\sqrt{n}}, \frac{\sqrt{d\log(\frac{1}{\delta})}}{n\epsilon}\}) + LR_2 \inf_{\mathbf{w}\in\mathcal{W}} \|\mathbf{w} - \mathbf{w}^*\|$.To ensure the approximation error is zero (i.e. $\texttt{Approx}(\mathcal{W}) = 0$), one needs to assume the constraint set $\mathcal{W}$ is large enough such that $\mathbf{w}^* \in \arg\inf_{\mathbf{w}\in\mathbb{R}^d} \mathcal{L}(\mathbf{w})$ belongs to $\mathcal{W}$. However, this would require the prior knowledge to guess the norm of $\|\mathbf{w}^*\|$ which may be not realistic due to the unknown population distribution. In contrast, our revisited bound in Theorem 7 is of $O(\max\{\frac{LR_2(\sqrt{\log(\frac{1}{\delta})} + \sqrt{\epsilon})\sqrt{d}}{\epsilon n}, \frac{LR_2\sqrt{\log(\frac{1}{\xi})}}{\sqrt{n}}\})$ which does not involve the approximation error. In addition, our bound does not need prior knowledge of $\mathbf{w}^*$ as discussed above.

6. Experiments

In this section, we present experimental results to verify the effectiveness of our algorithms, i.e. `Algorithm 1` (`Out-Pert-AUC`) and `Algorithm 2` (`Obj-Pert-AUC`). We apply both output perturbation and objective perturbation to AUC maximization (2) with ℓ being either the least square loss or the logistic loss.

6.1. Experimental Setting and Datasets

In particular, for the least square loss, we can analytically calculate the optimal $\hat{\mathbf{w}}_{\text{ls}}$ without building the positive-negative pairs. The details about its analytical solution can be found in Section B. For the logistic regression, we apply mini-batch SGD where at each iteration we build an unbiased gradient estimate based on a randomly selected 100 positive-negative example pairs. We consider step sizes of the form $\eta_t = 0.001/(\lambda t + 1)$. To accelerate the training speed, the SGD algorithm is initialized with $\hat{\mathbf{w}}_{\text{ls}}$ based on the least square loss. The experiments are performed on four benchmark datasets for imbalanced classification including IJCNN, Satimage, Webspam_u and HTTP. The first three datasets are downloaded from the LIBSVM webpage [47]. The HTTP dataset belongs to the KDD Cup'99 dataset, which consists of a wide variety of hand-injected attacks (anomalies) in a closed network [48]. We modify datasets with multiple class labels into datasets with binary class labels by setting the first half of class labels as positive labels, and setting the remaining class labels as negative labels. The statistics (sample size, feature size and imbalance ratio) of these datasets are summarized in Table 2.

datasets	# inst	# feat	n_+/n_-
IJCNN	49990	22	0.1059
Satimage	4435	36	1.26
Webspam_u	350000	254	1.5397
HTTP	567498	3	0.0039

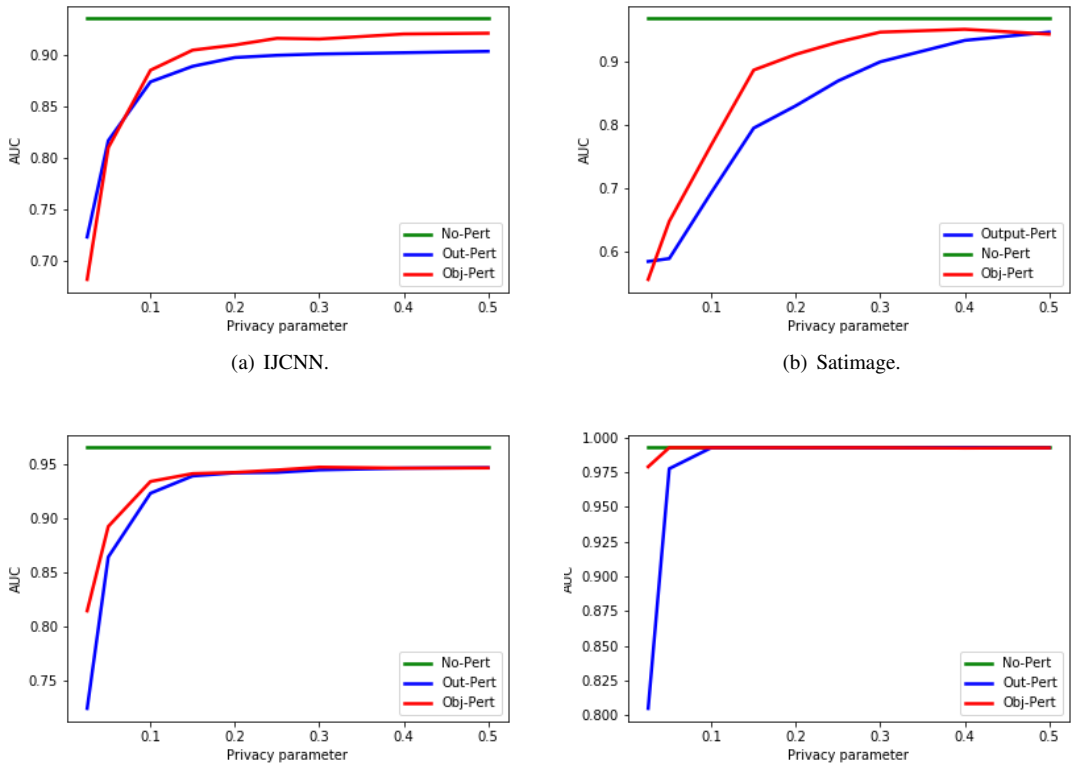
Table 2: Description of the datasets.

We aim to clarify how the prediction behavior would change w.r.t. the privacy parameter ϵ. To this aim, we vary ϵ over the set $\{0.025, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5\}$. Following the same procedure in [4], the regularization parameter λ for `Algorithm 1` and `Algorithm 2` is tuned by 10-fold cross validation over the set $\Lambda = \{10^{-4}, 10^{-3}, \ldots, 1\}$ at the privacy level $\epsilon = 0.1$. We should point out that this procedure may not be ideal and one can find more discussions on tuning the regularization parameter in Section 6.3 below. Due to the randomness of the privacy algorithms, we repeat 60 runs of the randomized training procedure for each parameter setting, and report the average of the AUC scores on the test set as the experimental results.

6.2. Generalization and Privacy

In Figure 1, we compare the behavior of `Algorithm 1` (`Output-Pert-AUC`) and Algorithm 2 (`Obj-Pert-AUC`) for ϵ-DP with the logistic loss. In Figure 2, we compare the behavior of `Output-Pert-AUC` and `Obj-Pert-AUC` for (ϵ, δ)-DP ($\delta = 1/n^2$) with the logistic loss. We include the non-private algorithms in all the figures as a baseline. We can see clearly the trade-off between the testing error and the privacy parameter. The AUC scores on testing data increase as we relax the privacy requirements when ϵ becomes larger.

In particular, for the dataset Webspam_u and HTTP, both `Output-Pert-AUC` and `Obj-Pert-AUC` achieve a high AUC score for a low privacy parameter $\epsilon = 0.1$. It is also clear that objective perturbation outperforms output perturbation in our experiments, which is in line with the observation in the pointwise learning case [4]. The underlying reason could be that the parameters γ and σ for perturbation



(a) IJCNN. (b) Satimage.

(c) Webspam_u. (d) HTTP.

Figure 1: AUC scores versus privacy parameter for ϵ-DP with the logistic loss. No-Pert means non-private algorithm, Output-Pert means `Output-Pert-AUC` for ϵ-DP and Obj-Pert means `Obj-Pert-AUC` for ϵ-DP.

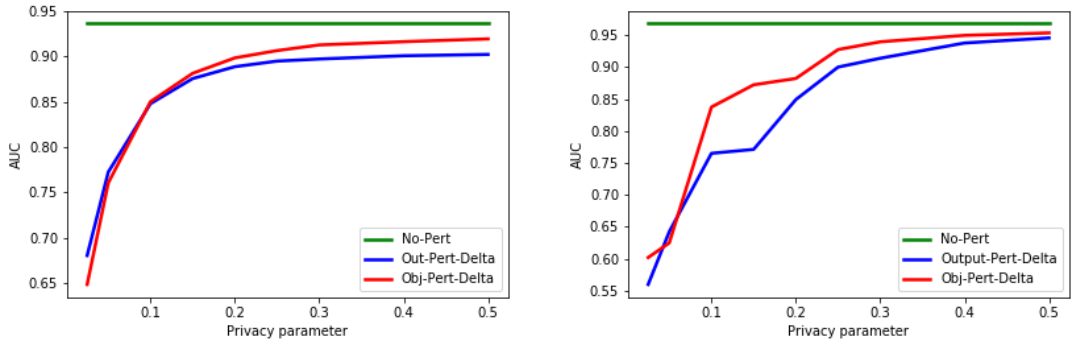
in `Algorithm` 1 (`Output-Pert-AUC`) grow reciprocally w.r.t. the regularization parameter λ. As a comparison, the parameters γ and σ have a milder logarithmic dependency on λ in `Algorithm` 2 (`Obj-Pert-AUC`). This means a heavier perturbation for `Output-Pert-AUC` than `Obj-Pert-AUC` if λ is small. Our theoretical analysis (see Theorem 3 and Theorem 5) also indicates that `Obj-Pert-AUC` enjoys a stronger theoretical guarantee than `Output-Pert-AUC`, which is consistent with the results of pointwise learning.

In Figure 3, we show the behavior of `Output-Pert-AUC` for ϵ-DP and (ϵ, δ)-DP ($\delta = 1/n^2$) when learning with the least square loss. We do not include results for `Obj-Pert-AUC` since the least square loss is not Lipschitz continuous, which is required for `Obj-Pert-AUC`. It is observed that the performance of `Output-Pert-AUC` with the least square loss is not as good as that with the logistic loss. For the dataset IJCNN, we achieve the AUC score 0.9 for $\epsilon < 0.2$ when using the logistic loss, while, for the least square loss, we require $\epsilon = 0.5$ to achieve a similar AUC score. The underlying reason is that $B(R) \leq 1$ for the logistic loss, while $B(R) > 2$ for the least square loss. That is, we impose a heavier perturbation for the least square loss when using the same privacy parameter.

6.3. Parameter Tuning

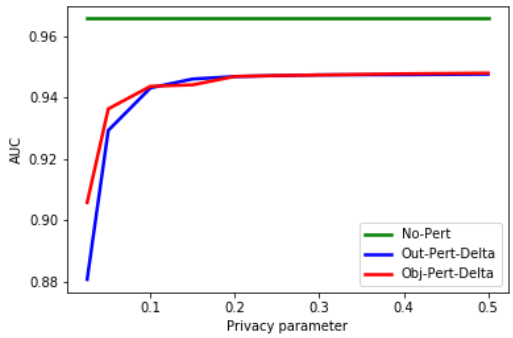
Tuning the regularization parameter λ is a critical issue in machine learning tasks. A widely used approach is to use cross validation: using data held out as the validation set, training classifiers using the remaining data for different values of λ, and selecting the best λ corresponding to the prediction performance on the validation set. The tuning of λ for privacy-preserving algorithms becomes more challenging as the standard cross-validation procedure mentioned above may violate differential privacy since the correct validation set is usually not available.

In many studies [49, 50, 24], the experiments were conducted with a fixed λ. There are a few works on designing parameter-tuning algorithms. For instance, the work [4] presents a differentially private parameter-tuning algorithm and provides its performance guarantee. Specifically, the algorithm first divides the training set into several disjoint sets and trains each λ using the private algorithm on a different set, then scores the performance of predictors on a validation set and chooses the output by exponential mechanism [51]. Another approach is end-to-end differentially private training and validation algorithm [52] which obeys a certain stability condition. The algorithm trains classifiers on the same training set with privacy budget for each parameter, and then uses a differentially private procedure to select the best parameter. Their experiments show that the performance of the stability algorithm is better than the parameter tuning algorithm in [4] and random choice of parameter. However, all the above methods focus on the pointwise learning. It will be a very interesting future direction to apply these approaches to the task of AUC maximization.

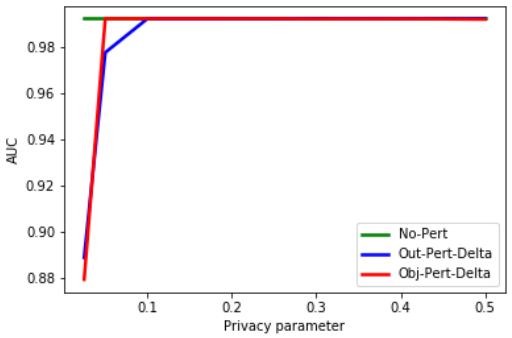


(a) IJCNN.

(b) Satimage.



(c) Webspam_u.



(d) HTTP.

Figure 2: AUC scores versus privacy parameter for (ϵ, δ)-DP with the logistic loss and $\delta = 1/n^2$. No-Pert means non-private algorithm, Output-Pert-Delta means `Output-Pert-AUC` for (ϵ, δ)-DP and Obj-Pert-Delta means `Obj-Pert-AUC` for (ϵ, δ)-DP.

7. Conclusion

In this paper, we proposed differentially private ERM algorithms for the important problem of AUC maximization in imbalanced classification. In particular, we systematically studied the privacy guarantees for output perturbation and objective perturbation with respect to both ϵ-DP and (ϵ, δ)-DP. Furthermore, we established utility guarantees on their generalization performance with fast rates. The main technical difficulty for deriving generalization bounds of the proposed algorithms is that the objective function for AUC maximization involves statistically dependent pairs of examples. To this end, we introduced a new error decomposition and developed fast rates through a novel combination of peeling techniques for Rademacher averages [53, 29] and the properties of U-Statistics [2, 3].

Future work can be the design of private stochastic optimization algorithms based on gradient perturbation for AUC maximization with optimal rates, which may require further exploration the appealing ideas in [20, 22] and the algorithmic stability in the setting of AUC maximization.

Acknowledgement. This work was done while Puyu Wang was a visiting student at SUNY Albany. The corresponding author is Yiming Ying, whose work is supported by NSF IIS-1816227 and IIS-2008532. The work of Hai Zhang is supported by NSFC U1811461.

References

[1] P. L. Bartlett, O. Bousquet, S. Mendelson, Local Rademacher complexities, The Annals of Statistics 33 (2005) 1497–1537.

[2] S. Clémençon, G. Lugosi, N. Vayatis, Ranking and empirical minimization of u-statistics, The Annals of Statistics (2008) 844–874.

[3] V. De la Pena, E. Giné, Decoupling: from dependence to independence, Springer Science & Business Media, 2012.

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

[5] D. Kifer, A. Smith, A. Thakurta, Private convex empirical risk minimization and high-dimensional regression, in: Conference on Learning Theory, 2012, pp. 25–1.

[6] C. Cortes, M. Mohri, Auc optimization vs. error rate minimization, in: Advances in neural information processing systems, 2004, pp. 313–320.

[7] A. P. Bradley, The use of the area under the roc curve in the evaluation of machine learning algorithms, Pattern recognition 30 (1997) 1145–1159.

[8] T. Fawcett, An introduction to roc analysis, Pattern recognition letters 27 (2006) 861–874.

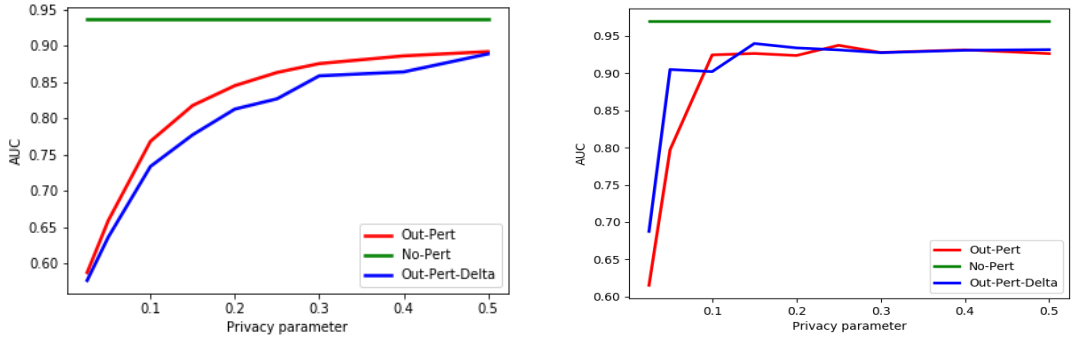
[9] J. A. Hanley, B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (roc) curve., Radiology 143 (1982) 29–36.

[10] W. Gao, R. Jin, S. Zhu, Z.-H. Zhou, One-pass auc optimization, in: International Conference on Machine Learning, 2013, pp. 906–914.

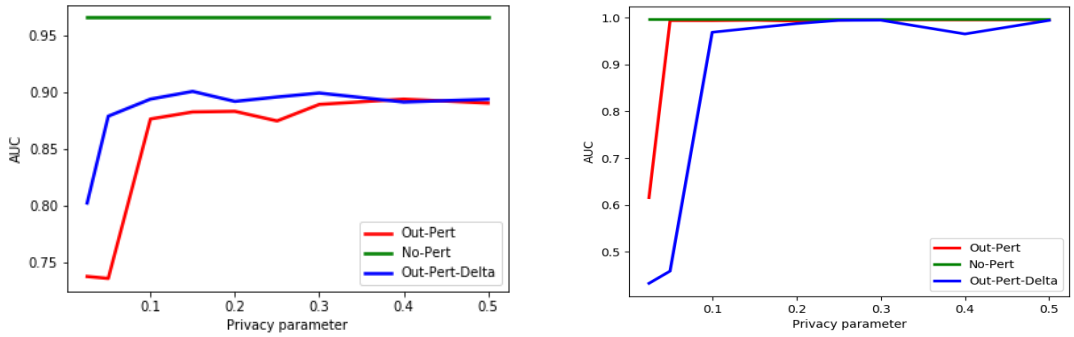
[11] M. Khalid, I. Ray, H. Chitsaz, Scalable nonlinear auc maximization methods, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2018, pp. 292–307.

[12] M. Liu, X. Zhang, Z. Chen, X. Wang, T. Yang, Fast stochastic auc maximization with o (1/n)-convergence rate, in: International Conference on Machine Learning, 2018, pp. 3195–3203.

[13] M. Liu, Z. Yuan, Y. Ying, T. Yang, Stochastic auc maximization with deep neural networks, in: International Conference on Learning Representations, 2020.



(a) IJCNN. (b) Satimage.



(c) Webspam_u. (d) HTTP.

Figure 3: AUC scores versus privacy parameter for output perturbation with the least square loss. No-Pert means non-private algorithm, Output-Pert means `Output-Pert-AUC` for ϵ-DP and Output-Pert-Delta means `Output-Pert-AUC` for (ϵ, δ)-DP ($\delta = 1/n^2$).

[14] M. Natole, Y. Ying, S. Lyu, Stochastic proximal algorithms for auc maximization, in: International Conference on Machine Learning, 2018, pp. 3707–3716.

[15] Y. Wang, R. Khardon, D. Pechyony, R. Jones, Generalization bounds for online learning algorithms with pairwise loss functions, in: Conference on Learning Theory, 2012, pp. 13–1.

[16] Y. Ying, L. Wen, S. Lyu, Stochastic online auc maximization, in: Advances in Neural Information Processing Systems, 2016.

[17] X. Zhang, A. Saha, S. V. N. Vishwanathan, Smoothing multivariate performance measures, Journal of Machine Learning Research 13 (2012) 3623–3680.

[18] P. Zhao, R. Jin, T. Yang, S. C. Hoi, Online auc maximization, in: Proceedings of the 28th international conference on machine learning (ICML-11), 2011.

[19] C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in: Theory of cryptography conference, Springer, 2006, pp. 265–284.

[20] R. Bassily, V. Feldman, K. Talwar, A. G. Thakurta, Private stochastic convex optimization with optimal rates, in: Advances in Neural Information Processing Systems, 2019, pp. 11279–11288.

[21] R. Bassily, A. Smith, A. Thakurta, Private empirical risk minimization: Efficient algorithms and tight error bounds, in: 2014 IEEE 55th Annual Symposium on Foundations of Computer Science, IEEE, 2014, pp. 464–473.

[22] V. Feldman, T. Koren, K. Talwar, Private stochastic convex optimization: Optimal rates in linear time, arXiv preprint arXiv:2005.04763 (2020).

[23] P. Jain, A. G. Thakurta, (near) dimension independent risk bounds for differentially private learning, in: International Conference on Machine Learning, 2014, pp. 476–484.

[24] B. Jayaraman, L. Wang, D. Evans, Q. Gu, Distributed learning without distress: Privacy-preserving empirical risk minimization, in: Advances in Neural Information Processing Systems, 2018, pp. 6343–6354.

[25] A. G. Thakurta, A. Smith, Differentially private feature selection via stability arguments, and the robustness of the lasso, in: Conference on Learning Theory, 2013, pp. 819–850.

[26] D. Wang, M. Ye, J. Xu, Differentially private empirical risk minimization revisited: Faster and more general, in: Advances in Neural Information Processing Systems, 2017, pp. 2722–2731.

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

[28] A. Guha Thakurta, Differentially private convex optimization for empirical risk minimization and high-dimensional regression, PhD Thesis (2012).

[29] K. Sridharan, S. Shalev-Shwartz, N. Srebro, Fast rates for regularized objectives, in: Advances in neural information processing systems, 2009, pp. 1545–1552.

[30] A. Herschtal, B. Raskutti, Optimising area under the roc curve using gradient descent, in: Proceedings of the twenty-first international conference on Machine learning, ACM, 2004, p. 49.

[31] T. Joachims, A support vector method for multivariate performance measures, in: Proceedings of the 22nd international conference on Machine learning, ACM, 2005, pp. 377–384.

[32] P. Kar, B. Sriperumbudur, P. Jain, H. Karnick, On the generalization ability of online learning algorithms for pairwise loss functions, in: International Conference on Machine Learning, 2013, pp. 441–449.

[33] B. Palaniappan, F. Bach, Stochastic variance reduction methods for saddle-point problems, in: Advances in Neural Information Processing Systems, 2016, pp. 1416–1424.

[34] B. I. Rubinstein, P. L. Bartlett, L. Huang, N. Taft, Learning in a large function space: Privacy-preserving mechanisms for svm learning, arXiv preprint arXiv:0911.5708 (2009).

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

[36] J. Zhang, K. Zheng, W. Mou, L. Wang, Efficient private erm for smooth

objectives, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp. 3922–3928.
[37] N. Agarwal, K. Singh, The price of differential privacy for online learning, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 32–40.
[38] P. Jain, P. Kothari, A. Thakurta, Differentially private online learning, in: Conference on Learning Theory, 2012, pp. 24–1.
[39] Y.-X. Wang, J. Lei, S. E. Fienberg, Learning with differential privacy: Stability, learnability and the sufficiency and necessity of erm principle, Journal of Machine Learning Research 17 (2016) 1–40.
[40] R. Arora, T. V. Marinov, E. Ullah, Private stochastic convex optimization: Efficient algorithms for non-smooth objectives, arXiv preprint arXiv:2002.09609 (2020).
[41] S. Shang, T. Wang, P. Cuff, S. Kulkarni, The application of differential privacy for rank aggregation: Privacy and accuracy, in: 17th International Conference on Information Fusion (FUSION), IEEE, 2014, pp. 1–7.
[42] M. Hay, L. Elagina, G. Miklau, Differentially private rank aggregation, in: Proceedings of the 2017 SIAM International Conference on Data Mining, SIAM, 2017, pp. 669–677.
[43] J. Li, Y. Pan, Y. Sui, I. W. Tsang, Secure metric learning via differential pairwise privacy, IEEE Transactions on Information Forensics and Security (2020).
[44] M. Huai, D. Wang, C. Miao, J. Xu, A. Zhang, Pairwise learning with differential privacy guarantees, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 694–701.
[45] O. Bousquet, A. Elisseeff, Stability and generalization, Journal of machine learning research 2 (2002) 499–526.
[46] M. Hardt, B. Recht, Y. Singer, Train faster, generalize better: Stability of stochastic gradient descent, in: International Conference on Machine Learning, 2016, pp. 1225–1234.
[47] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (2011) 27.
[48] M. Tavallaee, E. Bagheri, W. Lu, A. A. Ghorbani, A detailed analysis of the kdd cup 99 data set, in: 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, IEEE, 2009, pp. 1–6.
[49] S. Song, K. Chaudhuri, A. Sarwate, Learning from data with heterogeneous noise using sgd, in: Artificial Intelligence and Statistics, 2015, pp. 894–902.
[50] X. Wu, F. Li, A. Kumar, K. Chaudhuri, S. Jha, J. Naughton, Bolt-on differential privacy for scalable stochastic gradient descent-based analytics, in: Proceedings of the 2017 ACM International Conference on Management of Data, 2017, pp. 1307–1322.
[51] F. McSherry, K. Talwar, Mechanism design via differential privacy, in: 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07), IEEE, 2007, pp. 94–103.
[52] K. Chaudhuri, S. A. Vinterbo, A stability-based validation procedure for differentially private machine learning, in: Advances in Neural Information Processing Systems, 2013, pp. 2652–2660.
[53] P. L. Bartlett, S. Mendelson, Rademacher and gaussian complexities: Risk bounds and structural results, Journal of Machine Learning Research 3 (2002) 463–482.
[54] C. McDiarmid, On the method of bounded differences, Surveys in combinatorics 141 (1989) 148–188.

Appendix

A. Some Technical Lemmas

Here we list some technical lemmas which are used in our proofs later.

Definition 3. We say the function $f : \prod_{k=1}^n \Omega_k \to \mathbb{R}$ has bounded differences $\{c_k\}_{k=1}^n$ if, for all $1 \le k \le n$,

$$\max_{z_1,...,z_k,z_k',...,z_n} |f(z_1,...,z_{k-1},z_k,z_{k+1},...,z_n) - f(z_1,...,z_{k-1},z_k',z_{k+1},...,z_n)| \le c_k.$$

Lemma 5. *(McDiarmid's inequality [54]) Suppose $f : \prod_{k=1}^n \Omega_k \to \mathbb{R}$ has bounded differences $\{c_k\}_{k=1}^n$, then for all $t > 0$, there holds*

$$P_{\mathbf{z}}(f(\mathbf{z}) - \mathbb{E}_{\mathbf{z}} f(\mathbf{z}) \ge t) \le \exp\Big(-\frac{2t^2}{\sum_{k=1}^n c_k^2}\Big).$$

Lemma 6. *([2, Lemma A.1]) Let $q_\tau : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be real valued function indexed by $\tau \in \mathcal{T}$ where $\mathcal{T}$ is some index set. If $z_1, ..., z_n$ are i.i.d. then we have that*

$$\mathbb{E}\Big[\sup_{\tau\in\mathcal{T}} \frac{1}{n(n-1)} \sum_{i\neq j} q_\tau(z_i, z_j)\Big] \le \mathbb{E}\Big[\sup_{\tau\in\mathcal{T}} \frac{1}{\lfloor \frac{n}{2} \rfloor} \sum_{i\neq j}^{\lfloor \frac{n}{2} \rfloor} q_\tau(z_i, z_{i+\lfloor \frac{n}{2} \rfloor})\Big].$$

Lemma 7. *([53]) Let $\{g_j(\theta)\}$ and $\{h_j(\theta)\}$ be the sets of functions on Θ. If for each j, θ, θ' that $|g_j(\theta) - g_j(\theta')| \le |h_j(\theta) - h_j(\theta')|$, then*

$$\mathbb{E}\Big[\sup_{\theta\in\Theta} \sum_{j=1}^m \nu_j g_j(\theta)\Big] \le \mathbb{E}\Big[\sup_{\theta\in\Theta} \sum_{j=1}^m \nu_j h_j(\theta)\Big],$$

where $\{\nu_j\}$ are independent Rademacher random variables.

B. Analytic solution for the least square loss

In this section, we show that the problem (2) based on the least square loss has a closed-form solution. Furthermore, we do not need to build n_+n_- pairs in solving this problem. For $\ell(t) = (1-t)^2$, the model $\hat{\mathbf{w}}$ in (2) becomes

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n_+n_-} \sum_{i\in I, j\in J} (1 - \mathbf{w}^T(\mathbf{x}_i - \mathbf{x}_j))^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2,$$

where $I = \{i \in \{1,\dots,n\} : y_i = 1\}$ and $J = \{j \in \{1,\dots,n\} : y_j = -1\}$. According to the optimality condition, we know

$$\frac{2}{n_+n_-} \sum_{i\in I, j\in J} (1 - \hat{\mathbf{w}}^T(\mathbf{x}_i - \mathbf{x}_j))(\mathbf{x}_j - \mathbf{x}_i) + \lambda\hat{\mathbf{w}} = 0$$

and therefore

$$\frac{1}{n_+n_-} \sum_{i\in I, j\in J} \Big(\mathbf{x}_i\mathbf{x}_i^T + \mathbf{x}_j\mathbf{x}_j^T - \mathbf{x}_i\mathbf{x}_j^T - \mathbf{x}_j\mathbf{x}_i^T\Big)\hat{\mathbf{w}} + \frac{\lambda}{2}\hat{\mathbf{w}} = \frac{1}{n_+n_-} \sum_{i\in I, j\in J} (\mathbf{x}_i - \mathbf{x}_j) \tag{B.1}$$

It is clear that

$$\sum_{i\in I, j\in J} \mathbf{x}_i\mathbf{x}_i^T = n^- \sum_{i\in I} \mathbf{x}_i\mathbf{x}_i^T, \qquad \sum_{i\in I, j\in J} \mathbf{x}_j\mathbf{x}_j^T = n^+ \sum_{i\in J} \mathbf{x}_j\mathbf{x}_j^T$$

$$\sum_{i\in I, j\in J} (\mathbf{x}_i\mathbf{x}_j^T + \mathbf{x}_j\mathbf{x}_i^T) = (\sum_{i\in I} \mathbf{x}_i)(\sum_{j\in J} \mathbf{x}_j)^T + (\sum_{j\in J} \mathbf{x}_j)(\sum_{i\in I} \mathbf{x}_i)^T$$

$$\sum_{i\in I, j\in J} (\mathbf{x}_i - \mathbf{x}_j) = n_- \sum_{i\in I} \mathbf{x}_i - n_+ \sum_{j\in J} \mathbf{x}_j.$$

We can plug the above identities back into (B.1) and get that

$$\Big(\frac{1}{n_+} \sum_{i\in I} \mathbf{x}_i\mathbf{x}_i^T + \frac{1}{n_-} \sum_{j\in J} \mathbf{x}_j\mathbf{x}_j^T - \bar{\mathbf{x}}_+\bar{\mathbf{x}}_-^T - \bar{\mathbf{x}}_-\bar{\mathbf{x}}_+^T + \frac{\lambda\mathbf{I}}{2}\Big)\hat{\mathbf{w}} = \bar{\mathbf{x}}^+ - \bar{\mathbf{x}}^-,$$

where $\bar{\mathbf{x}}_+ = \frac{1}{n_+}\sum_{i\in I} \mathbf{x}_i$ and $\bar{\mathbf{x}}_- = \frac{1}{n_-}\sum_{j\in J} \mathbf{x}_j$. Therefore,

$$\hat{\mathbf{w}} = \Big(\frac{1}{n_+} \sum_{i\in I} \mathbf{x}_i\mathbf{x}_i^T + \frac{1}{n_-} \sum_{j\in J} \mathbf{x}_j\mathbf{x}_j^T - \bar{\mathbf{x}}_+\bar{\mathbf{x}}_-^T - \bar{\mathbf{x}}_-\bar{\mathbf{x}}_+^T + \frac{\lambda\mathbf{I}}{2}\Big)^{-1}\Big(\bar{\mathbf{x}}^+ - \bar{\mathbf{x}}^-\Big).$$

C. Proofs

C.1. Proof of Lemma 1

Proof. Let $\mathcal{R}_S^\lambda(\mathbf{w}) = \mathcal{R}_S(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2$, and $g(\mathbf{w}) = \mathcal{R}_{S'}^\lambda(\mathbf{w}) - \mathcal{R}_S^\lambda(\mathbf{w})$. Further, $\widehat{\mathbf{w}}(S) = \arg\min_{\mathbf{w}} \mathcal{R}_S^\lambda(\mathbf{w})$, and $\widehat{\mathbf{w}}(S') = \arg\min_{\mathbf{w}} \mathcal{R}_S^\lambda(\mathbf{w}) + g(\mathbf{w})$. From the definition of $\widehat{\mathbf{w}}(S)$ and $\widehat{\mathbf{w}}(S')$, there holds

$$\nabla\mathcal{R}_S^\lambda(\widehat{\mathbf{w}}(S)) = \nabla\mathcal{R}_S^\lambda(\widehat{\mathbf{w}}(S')) + \nabla g(\widehat{\mathbf{w}}(S')) = 0. \tag{C.1}$$

Since $\mathcal{R}_S^\lambda(\mathbf{w})$ is λ-strongly convex, we have

$$\lambda\|\widehat{\mathbf{w}}(S)-\widehat{\mathbf{w}}(S')\|^2 \leq (\nabla\mathcal{R}_S^\lambda(\widehat{\mathbf{w}}(S))-\nabla\mathcal{R}_S^\lambda(\widehat{\mathbf{w}}(S')))^T(\widehat{\mathbf{w}}(S)-\widehat{\mathbf{w}}(S')).$$

Combining this with (C.1) and Cauchy-Schwartz inequality,

$$\begin{aligned}\lambda\|\widehat{\mathbf{w}}(S) - \widehat{\mathbf{w}}(S')\|^2 &\leq \|\nabla\mathcal{R}_S^\lambda(\widehat{\mathbf{w}}(S)) - \nabla\mathcal{R}_S^\lambda(\widehat{\mathbf{w}}(S'))\| \cdot \|\widehat{\mathbf{w}}(S) - \widehat{\mathbf{w}}(S')\| \\ &\leq \|\nabla g(\widehat{\mathbf{w}}(S'))\| \cdot \|\widehat{\mathbf{w}}(S) - \widehat{\mathbf{w}}(S')\|.\end{aligned}$$

Therefore,

$$\|\widehat{\mathbf{w}}(S) - \widehat{\mathbf{w}}(S')\| \leq \frac{1}{\lambda}\|\nabla g(\widehat{\mathbf{w}}(S'))\|. \tag{C.2}$$

Assume that S and S' differ in the first datum, i.e. $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_1', y_1')$, then

$$\begin{aligned}\nabla g(\widehat{\mathbf{w}}(S')) &= \nabla\mathcal{R}_{S'}^\lambda(\widehat{\mathbf{w}}(S')) - \nabla\mathcal{R}_S^\lambda(\widehat{\mathbf{w}}(S')) \\ &= \frac{1}{n_+n_-}\sum_{j=2}^n \nabla\ell(\widehat{\mathbf{w}}(S')^T(\mathbf{x}_1' - \mathbf{x}_j))\mathbb{I}_{[y_1'=1\wedge y_j=-1]} \\ &+ \frac{1}{n_+n_-}\sum_{i=2}^n \nabla\ell(\widehat{\mathbf{w}}(S')^T(\mathbf{x}_i - \mathbf{x}_1'))\mathbb{I}_{[y_i=1\wedge y_1'=-1]} \\ &- \frac{1}{n_+n_-}\sum_{j=2}^n \nabla\ell(\widehat{\mathbf{w}}(S')^T(\mathbf{x}_j - \mathbf{x}_1))\mathbb{I}_{[y_1=1\wedge y_j=-1]} \\ &- \frac{1}{n_+n_-}\sum_{i=2}^n \nabla\ell(\widehat{\mathbf{w}}(S')^T(\mathbf{x}_1 - \mathbf{x}_i))\mathbb{I}_{[y_i=1\wedge y_1=-1]}.\end{aligned} \tag{C.3}$$

For any $\lambda > 0$, by the definition of $\widehat{\mathbf{w}}(S')$, we have $\mathcal{R}_{S'}(\widehat{\mathbf{w}}(S')) + \frac{\lambda}{2}\|\widehat{\mathbf{w}}(S')\|^2 \leq \mathcal{R}_S(\mathbf{0}) + \frac{\lambda}{2}\|\mathbf{0}\|^2 = 1$ as $\ell(0) = 1$. Therefore, $\|\widehat{\mathbf{w}}(S')\| \leq \sqrt{2/\lambda}$. Consequently, for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$,

$$\begin{aligned}\|\nabla\ell(\widehat{\mathbf{w}}(S')^T(\mathbf{x} - \mathbf{x}'))\| &= \|\ell'(\widehat{\mathbf{w}}(S')^T(\mathbf{x} - \mathbf{x}'))(\mathbf{x} - \mathbf{x}')\| \\ &\leq D_{\mathcal{X}}B(\sqrt{2/\lambda}D_{\mathcal{X}}).\end{aligned} \tag{C.4}$$

This combined with (C.3) implies that

$$\|\nabla g(\widehat{\mathbf{w}}(S'))\| \leq 2D_{\mathcal{X}}B(\sqrt{2/\lambda}D_{\mathcal{X}})(\frac{1}{n_+} + \frac{1}{n_-}).$$

Putting this back into (C.2) indicates $\|\widehat{\mathbf{w}}(S) - \widehat{\mathbf{w}}(S')\| \leq \frac{2D_{\mathcal{X}}B(\sqrt{2/\lambda}D_{\mathcal{X}})}{\lambda}(\frac{1}{n_+} + \frac{1}{n_-})$ which completes the proof of the lemma. □

C.2. Proof of Lemma 3

Proof. For any $i \geq 1$, let

$$\mathcal{F}(4^i c) = \{\mathbf{w} \in \mathcal{B} : 4^{i-1}c \leq c + \mathcal{R}^\lambda(\mathbf{w}) - \mathcal{R}^\lambda(\mathbf{w}_\lambda) \leq 4^i c\}.$$

Here, $c > 0$ is a constant to be determined later. Define, for any $i \geq 0$,

$$R_i = \sup_{\mathbf{w}\in\mathcal{F}(4^i c)} [\mathcal{R}(\mathbf{w}) - \mathcal{R}(\mathbf{w}_\lambda) - \mathcal{R}_S(\mathbf{w}) + \mathcal{R}_S(\mathbf{w}_\lambda)].$$

For $\mathbf{w} \in \mathcal{B}$, consider

$$\begin{aligned}&\frac{\mathcal{R}^\lambda(\mathbf{w}) - \mathcal{R}^\lambda(\mathbf{w}_\lambda) - \mathcal{R}_S^\lambda(\mathbf{w}) + \mathcal{R}_S^\lambda(\mathbf{w}_\lambda)}{c + \mathcal{R}^\lambda(\mathbf{w}) - \mathcal{R}^\lambda(\mathbf{w}_\lambda)} \\ &\leq \sup_{\mathbf{w}\in\mathcal{B}}[\frac{\mathcal{R}^\lambda(\mathbf{w}) - \mathcal{R}^\lambda(\mathbf{w}_\lambda) - \mathcal{R}_S^\lambda(\mathbf{w}) + \mathcal{R}_S^\lambda(\mathbf{w}_\lambda)}{c + \mathcal{R}^\lambda(\mathbf{w}) - \mathcal{R}^\lambda(\mathbf{w}_\lambda)}] \\ &\leq \sup_i \sup_{\mathbf{w}\in\mathcal{F}(4^i c)} [\frac{\mathcal{R}^\lambda(\mathbf{w}) - \mathcal{R}^\lambda(\mathbf{w}_\lambda) - \mathcal{R}_S^\lambda(\mathbf{w}) + \mathcal{R}_S^\lambda(\mathbf{w}_\lambda)}{c + \mathcal{R}^\lambda(\mathbf{w}) - \mathcal{R}^\lambda(\mathbf{w}_\lambda)}] \\ &\leq \sum_{i=1}^\infty \sup_{\mathbf{w}\in\mathcal{F}(4^i c)} [\frac{\mathcal{R}^\lambda(\mathbf{w}) - \mathcal{R}^\lambda(\mathbf{w}_\lambda) - \mathcal{R}_S^\lambda(\mathbf{w}) + \mathcal{R}_S^\lambda(\mathbf{w}_\lambda)}{c + \mathcal{R}^\lambda(\mathbf{w}) - \mathcal{R}^\lambda(\mathbf{w}_\lambda)}] \\ &\leq c^{-1}\sum_{i=1}^\infty 4^{-(i-1)} \sup_{\mathbf{w}\in\mathcal{F}(4^i c)} [\mathcal{R}^\lambda(\mathbf{w}) - \mathcal{R}^\lambda(\mathbf{w}_\lambda) - \mathcal{R}_S^\lambda(\mathbf{w}) + \mathcal{R}_S^\lambda(\mathbf{w}_\lambda)] \\ &= c^{-1}\sum_{i=1}^\infty 4^{-(i-1)}R_i.\end{aligned} \tag{C.5}$$

From the strongly convexity of $\mathcal{R}^\lambda(\cdot)$ and the definition of $\mathbf{w}_\lambda$, we have $\frac{\lambda}{2}\|\mathbf{w} - \mathbf{w}_\lambda\|^2 \leq \mathcal{R}^\lambda(\mathbf{w}) - \mathcal{R}^\lambda(\mathbf{w}_\lambda)$, hence

$$\|\mathbf{w} - \mathbf{w}_\lambda\| \leq \sqrt{\frac{2(\mathcal{R}^\lambda(\mathbf{w}) - \mathcal{R}^\lambda(\mathbf{w}_\lambda))}{\lambda}} \leq 2^i\sqrt{\frac{2c}{\lambda}},\ \forall\ \mathbf{w} \in \mathcal{F}(4^i c). \tag{C.6}$$

Note that R_i is a function of $\{z_1, ..., z_n\}$, where z_k corresponding to data $(\mathbf{x}_k, y_k)$. For any z_j being replaced by z_j', we have

$$\begin{aligned}&|R_i(z_1, ..., z_j, ..., z_n) - R_i(z_1, ..., z_j', ..., z_n)| \\ &\leq \frac{2n}{n_+n_-}BD_{\mathcal{X}}\|\mathbf{w} - \mathbf{w}_\lambda\| \leq \frac{n}{n_+n_-}2^{i+1}BD_{\mathcal{X}}\sqrt{\frac{2c}{\lambda}},\end{aligned}$$

where in the last inequality we use (C.6). Applying McDiarmid's inequality(see Lemma 5 in Appendix A), with probability at least $1 - 2^{-i}\delta$,

$$R_i - \mathbb{E}[R_i] \leq \frac{n\sqrt{n}}{n_+n_-}2^{i+1}BD_{\mathcal{X}}\sqrt{\frac{c\log(\frac{2^i}{\delta})}{\lambda}}.$$

It remains to estimate $\mathbb{E}[R_i] = \mathbb{E}\Big[\sup_{\mathbf{w}\in\mathcal{F}(4^i c)}[\mathcal{E}(\mathbf{w}) - \mathcal{E}_S(\mathbf{w})]\Big]$, where $\mathcal{E}(\mathbf{w}) = \mathcal{R}(\mathbf{w}) - \mathcal{R}(\mathbf{w}_\lambda)$ and $\mathcal{E}_S(\mathbf{w}) = \mathcal{R}_S(\mathbf{w}) - \mathcal{R}_S(\mathbf{w}_\lambda)$. From Lemma A.1 in [2](see Lemma 6 in Appendix A), and let $q_{\mathbf{w}}(z_i, z_j) = \frac{n(n-1)}{n_+n_-}\Big[\mathcal{E}(\mathbf{w}) - \phi(z_i, z_j)\Big]$, where $\phi_{\mathbf{w}}(z_i, z_j) = \ell(\mathbf{w}^T(x_i -$

$x_j))\mathbb{I}_{[y_i=1\wedge y_j=-1]} - \ell(\mathbf{w}_\lambda^T(x_i-x_j))\mathbb{I}_{[y_i=1\wedge y_j=-1]}$. For $n \geq 2$, we have

$$\begin{aligned}\mathbb{E}_S \sup_{\mathbf{w}\in\mathcal{F}(4^i c)}\Big[\mathcal{E}(\mathbf{w}) - \mathcal{E}_{\mathcal{S}}(\mathbf{w})\Big] &\leq \mathbb{E}_S \sup_{\mathbf{w}\in\mathcal{F}(4^i c)}\Big[\frac{1}{n(n-1)}\sum_{i\neq j} q_{\mathbf{w}}(z_i, z_j)\Big] \\ &\leq \mathbb{E}_S \sup_{\mathbf{w}\in\mathcal{F}(4^i c)}\Big[\frac{1}{\lfloor\frac{n}{2}\rfloor}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} q_{\mathbf{w}}(z_i, z_{i+\lfloor\frac{n}{2}\rfloor})\Big] \\ &= \frac{n(n-1)}{n_+n_-}\mathbb{E}_S \sup_{\mathbf{w}\in\mathcal{F}(4^i c)}\Big[\mathcal{E}(\mathbf{w}) - \bar{\mathcal{E}}_{\mathcal{S}}(\mathbf{w})\Big],\end{aligned} \tag{C.7}$$

where $\bar{\mathcal{E}}_{\mathcal{S}}(\mathbf{w}) = \frac{1}{\lfloor\frac{n}{2}\rfloor}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\phi(z_i, z_{i+\lfloor\frac{n}{2}\rfloor})$. For data set $S = \{z_1, ..., z_n\}$, let $S' = \{z_1', ..., z_n'\}$ be the i.i.d. copy of S, then

$$\begin{aligned}\mathbb{E}_S \sup_{\mathbf{w}\in\mathcal{F}(4^i c)}\Big[\mathcal{E}(\mathbf{w}) - \bar{\mathcal{E}}_{\mathcal{S}}(\mathbf{w})\Big] &= \mathbb{E}_S \sup_{\mathbf{w}\in\mathcal{F}(4^i c)}\Big[\mathbb{E}_{S'}[\bar{\mathcal{E}}_{\mathcal{S}'}(\mathbf{w})] - \bar{\mathcal{E}}_{\mathcal{S}}(\mathbf{w})\Big] \\ &\leq \mathbb{E}_{S,S'} \sup_{\mathbf{w}\in\mathcal{F}(4^i c)}\Big[\bar{\mathcal{E}}_{\mathcal{S}'}(\mathbf{w}) - \bar{\mathcal{E}}_{\mathcal{S}}(\mathbf{w})\Big],\end{aligned} \tag{C.8}$$

By standard symmetrization techniques (see e.g.[53]), for i.i.d. Rademacher variables $\{\nu_i \in \{\pm 1\} : i \in \mathbb{N}_{\lfloor\frac{n}{2}\rfloor}\}$, we have

$$\begin{aligned}&\mathbb{E}_{S,S'} \sup_{\mathbf{w}\in\mathcal{F}(4^i c)}\Big[\bar{\mathcal{E}}_{\mathcal{S}'}(\mathbf{w}) - \bar{\mathcal{E}}_{\mathcal{S}}(\mathbf{w})\Big] \\ &= \mathbb{E}_{S,S'}\frac{1}{\lfloor\frac{n}{2}\rfloor} \sup_{\mathbf{w}\in\mathcal{F}(4^i c)}\Big[\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}[\phi(z_i', z'_{i+\lfloor\frac{n}{2}\rfloor}) - \phi(z_i, z_{i+\lfloor\frac{n}{2}\rfloor})\Big] \\ &= \mathbb{E}_{S,S'}\frac{1}{\lfloor\frac{n}{2}\rfloor} \sup_{\mathbf{w}\in\mathcal{F}(4^i c)}\Big[\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\nu_i[\phi(z_i', z'_{i+\lfloor\frac{n}{2}\rfloor}) - \phi(z_i, z_{i+\lfloor\frac{n}{2}\rfloor})]\Big] \\ &\leq 2\mathbb{E}_{S,\nu}\frac{1}{\lfloor\frac{n}{2}\rfloor} \sup_{\mathbf{w}\in\mathcal{F}(4^i c)}\Big[\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\nu_i\phi(z_i, z_{i+\lfloor\frac{n}{2}\rfloor})\Big],\end{aligned} \tag{C.9}$$

Applying the contraction property of Rademacher averages (see Lemma 7 in Appendix A) with $\theta = \mathbf{w}$, $g_i(\theta) = \phi(z_i, z_{i+\lfloor\frac{n}{2}\rfloor})$, and $h_i(\theta) = BD_{\mathcal{X}}\|\mathbf{w} - \mathbf{w}_\lambda\|$,

$$\begin{aligned}&\mathbb{E}_{S,\nu}\frac{1}{\lfloor\frac{n}{2}\rfloor} \sup_{\mathbf{w}\in\mathcal{F}(4^i c)}\Big[\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\nu_i\phi(z_i, z_{i+\lfloor\frac{n}{2}\rfloor})\Big] \\ &\leq \mathbb{E}_{S,\nu}\frac{1}{\lfloor\frac{n}{2}\rfloor} \sup_{\mathbf{w}\in\mathcal{F}(4^i c)}\Big[\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\nu_i BD_{\mathcal{X}}\|\mathbf{w} - \mathbf{w}_\lambda\|\Big] \\ &= \frac{1}{\lfloor\frac{n}{2}\rfloor}BD_{\mathcal{X}}2^i\sqrt{\frac{2c}{\lambda}}\mathbb{E}_\nu\Big[\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\nu_i\Big] \leq BD_{\mathcal{X}}2^i\sqrt{\frac{2c}{\lambda\lfloor\frac{n}{2}\rfloor}},\end{aligned} \tag{C.10}$$

where in the last inequality we have used

$$\mathbb{E}_\nu\Big[\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\nu_i\Big] \leq \Big(\mathbb{E}_\nu\Big[(\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor}\nu_i)^2\Big]\Big)^{1/2} \leq \sqrt{\lfloor\frac{n}{2}\rfloor}.$$

Combine (C.7), (C.8), (C.9) and (C.10), we have

$$\begin{aligned}R_i &\leq \frac{n\sqrt{n}}{n_+n_-}2^{i+1}BD_{\mathcal{X}}\sqrt{\frac{c\log(\frac{2^i}{\delta})}{\lambda}} + \frac{n(n-1)}{n_+n_-}BD_{\mathcal{X}}2^{i+1}\sqrt{\frac{2c}{\lambda\lfloor\frac{n}{2}\rfloor}} \\ &\leq \frac{n\sqrt{n}}{n_+n_-}2^{i+1}BD_{\mathcal{X}}\sqrt{\frac{c}{\lambda}}\Big(2\sqrt{2} + \sqrt{\log(\frac{2^i}{\delta})}\Big).\end{aligned} \tag{C.11}$$

with probability at least $1 - 2^{-i}\delta$.

Combine (C.5) and (C.11), there holds

$$\begin{aligned}&\sup_{\mathbf{w}\in\mathcal{B}}\Big[\frac{\mathcal{R}^\lambda(\mathbf{w}) - \mathcal{R}^\lambda(\mathbf{w}_\lambda) - \mathcal{R}_S^\lambda(\mathbf{w}) + \mathcal{R}_S^\lambda(\mathbf{w}_\lambda)}{c + \mathcal{R}^\lambda(\mathbf{w}) - \mathcal{R}^\lambda(\mathbf{w}_\lambda)}\Big] \\ &\leq 8\frac{n\sqrt{n}}{n_+n_-}BD_{\mathcal{X}}\sqrt{\frac{1}{c\lambda}}\sum_{i=1}^{\infty}2^{-i}[2\sqrt{2} + \sqrt{\log(\frac{2^i}{\delta})}] \\ &\leq 8\frac{n\sqrt{n}}{n_+n_-}BD_{\mathcal{X}}\sqrt{\frac{1}{c\lambda}}[2\sqrt{2} + \sqrt{\log(\frac{1}{\delta})} + \sum_{i=1}^{\infty}2^{-i}\sqrt{i}] \\ &\leq 8\frac{n\sqrt{n}}{n_+n_-}BD_{\mathcal{X}}\sqrt{\frac{1}{c\lambda}}[2\sqrt{2} + 2 + \sqrt{\log(\frac{1}{\delta})}],\end{aligned} \tag{C.12}$$

where in the last inequality used $\sum_{i=1}^{\infty}2^{-i}\sqrt{i} \leq 2$. Let $M = 8\frac{n\sqrt{n}}{n_+n_-}BD_{\mathcal{X}}\sqrt{\frac{1}{\lambda}}[2\sqrt{2} + 2 + \sqrt{\log(\frac{1}{\delta})}]$, and for any $0 < \tau < 1$, select $c = \frac{M^2}{(1-\tau)^2}$, we have with probability at least $1 - \delta$,

$$\sup_{\mathbf{w}\in\mathcal{B}}\Big[\frac{\mathcal{R}^\lambda(\mathbf{w}) - \mathcal{R}^\lambda(\mathbf{w}_\lambda) + \frac{M^2}{(1-\tau)^2}}{\mathcal{R}_S^\lambda(\mathbf{w}) - \mathcal{R}_S^\lambda(\mathbf{w}_\lambda) + \frac{M^2}{(1-\tau)^2}}\Big] \leq \frac{1}{\tau}. \tag{C.13}$$

that is, for all $\mathbf{w} \in \mathcal{B}$,

$$\mathcal{R}^\lambda(\mathbf{w}) - \mathcal{R}^\lambda(\mathbf{w}_\lambda) \leq \frac{1}{\tau}(\mathcal{R}_S^\lambda(\mathbf{w}) - \mathcal{R}_S^\lambda(\mathbf{w}_\lambda)) + O\Big(\frac{B^2D_{\mathcal{X}}^2\log(\frac{1}{\delta})n^3}{\lambda\tau(1-\tau)(n_+n_-)^2}\Big).$$

This completes the proof of the lemma. □

C.3. Proof of Theorem 3

Proof. Obviously, if ℓ is L-Lipschitz continuous, $B \leq L$ for all $\mathbf{w}$ in Lemma 3. Hence, the error decomposition (18) and Lemma 3 with $\tau = \frac{1}{2}$ imply, with probability at least $1 - \xi$, that

$$\begin{aligned}&\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}}\mathcal{R}(\mathbf{w}) \\ &\leq 2\mathbb{E}_{\mathbf{b}}[\mathcal{R}_S^\lambda(\mathbf{w}_{\text{priv}}) - \mathcal{R}_S^\lambda(\widehat{\mathbf{w}})] + \frac{\lambda}{2}\|\mathbf{w}^*\|^2 + O\Big(\frac{L^2D_{\mathcal{X}}^2\log(\frac{1}{\xi})n^3}{\lambda(n_+n_-)^2}\Big).\end{aligned} \tag{C.14}$$

From the strongly smoothness of ℓ and the definition of $\widehat{\mathbf{w}}$, there holds

$$\begin{aligned}\mathcal{R}_S^\lambda(\mathbf{w}_{\text{priv}}) - \mathcal{R}_S^\lambda(\widehat{\mathbf{w}}) &\leq \langle\nabla\mathcal{R}_S^\lambda(\widehat{\mathbf{w}}), \mathbf{w}_{\text{priv}} - \widehat{\mathbf{w}}\rangle + \frac{\beta+\lambda}{2}\|\mathbf{w}_{\text{priv}} - \widehat{\mathbf{w}}\|^2 \\ &= \frac{\beta+\lambda}{2}\|\mathbf{w}_{\text{priv}} - \widehat{\mathbf{w}}\|^2.\end{aligned}$$

Therefore,

$$\mathbb{E}_{\mathbf{b}}[\mathcal{R}_S^\lambda(\mathbf{w}_{\text{priv}}) - \mathcal{R}_S^\lambda(\widehat{\mathbf{w}})] \leq \frac{\beta+\lambda}{2}\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|^2]. \tag{C.15}$$

For ϵ-DP, Since $\nu_1(\mathbf{b}) = \frac{1}{\alpha}\exp(-\frac{\|\mathbf{b}\|}{\gamma})$, then $\|\mathbf{b}\|$ is a random vector drawn from Gamma distribution $\Gamma(d, \gamma)$. Thus $\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|] = d\gamma$, $\text{Var}(\|\mathbf{b}\|) = d\gamma^2$, and $\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|^2] = (d + d^2)\gamma^2$. Notice that ℓ is L-Lipschitz, then $\gamma = \frac{2D_{\mathcal{X}}L}{\epsilon\lambda}(\frac{1}{n_+} + \frac{1}{n_-})$. Hence,

$$\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|^2] \leq \frac{4D_{\mathcal{X}}^2L^2(d+d^2)}{\epsilon^2\lambda^2}\Big(\frac{1}{n_+} + \frac{1}{n_-}\Big)^2, \tag{C.16}$$

Therefore,

$$\mathbb{E}_{\mathbf{b}}[\mathcal{R}_S^\lambda(\mathbf{w}_{\text{priv}}) - \mathcal{R}_S^\lambda(\widehat{\mathbf{w}})] \leq \frac{2(\beta+\lambda)(d+d^2)D_{\mathcal{X}}^2L^2n^2}{\epsilon^2\lambda^2(n_+n_-)^2}. \quad \text{(C.17)}$$

Combine (C.14) and (C.17), we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}} \mathcal{R}(\mathbf{w}) \\
&\leq 2\mathbb{E}_{\mathbf{b}}[\mathcal{R}_S^\lambda(\mathbf{w}_{\text{priv}}) - \mathcal{R}_S^\lambda(\mathbf{w}_\lambda)] + \frac{\lambda}{2}\|\mathbf{w}^*\|^2 + \mathcal{O}\Big(\frac{L^2D_{\mathcal{X}}^2\log(\frac{1}{\xi})n^3}{\lambda(n_+n_-)^2}\Big) \\
&\leq \frac{\lambda}{2}\|\mathbf{w}^*\|^2 + \frac{4(\beta+\lambda)(d+d^2)D_{\mathcal{X}}^2L^2n^2}{\epsilon^2\lambda^2(n_+n_-)^2} + \mathcal{O}\Big(\frac{L^2D_{\mathcal{X}}^2\log(\frac{1}{\xi})n^3}{\lambda(n_+n_-)^2}\Big) \\
&= \frac{\lambda}{2}\|\mathbf{w}^*\|^2 + \mathcal{O}\Big(\max\Big\{\frac{\beta L^2D_{\mathcal{X}}^2n^2d^2}{\epsilon^2\lambda^2(n_+n_-)^2}, \frac{L^2D_{\mathcal{X}}^2\log(\frac{1}{\xi})n^3}{\lambda(n_+n_-)^2}\Big\}\Big). \quad \text{(C.18)}
\end{aligned}$$

If $\lambda \leq \frac{\beta d^2}{\epsilon^2 n\log(\frac{1}{\xi})}$, we have

$$\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}} \mathcal{R}(\mathbf{w}) \leq \frac{\lambda}{2}\|\mathbf{w}^*\|^2 + \mathcal{O}\Big(\frac{\beta L^2D_{\mathcal{X}}^2n^2d^2}{\epsilon^2\lambda^2(n_+n_-)^2}\Big),$$

else,

$$\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}} \mathcal{R}(\mathbf{w}) \leq \frac{\lambda}{2}\|\mathbf{w}^*\|^2 + \mathcal{O}\Big(\frac{L^2D_{\mathcal{X}}^2\log(\frac{1}{\xi})n^3}{\lambda(n_+n_-)^2}\Big).$$

Now, setting $\lambda = \min\Big\{\frac{\beta^{\frac{1}{3}}(LD_{\mathcal{X}}nd)^{\frac{2}{3}}}{(\|\mathbf{w}^*\|\epsilon n_+n_-)^{\frac{2}{3}}}, \frac{LD_{\mathcal{X}}\sqrt{\log(\frac{1}{\xi})}n^{\frac{3}{2}}}{\|\mathbf{w}^*\|n_+n_-}\Big\}$, and recalling that imbalanced ratio $\rho = \frac{n_+}{n}$,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}} \mathcal{R}(\mathbf{w}) \\
&= \mathcal{O}\Big(\max\Big\{\frac{\beta^{\frac{1}{3}}(LD_{\mathcal{X}}d)^{\frac{2}{3}}\|\mathbf{w}^*\|^{\frac{4}{3}}}{(\rho(1-\rho)\epsilon n)^{\frac{2}{3}}}, \frac{LD_{\mathcal{X}}\sqrt{\log(\frac{1}{\xi})}\|\mathbf{w}^*\|}{\rho(1-\rho)\sqrt{n}}\Big\}\Big).
\end{aligned}$$

For (ϵ,δ)-DP, notice that $\mathbf{b} \sim \mathcal{N}(0,\sigma^2\mathbf{I})$, we have $\mathbb{E}[\|\mathbf{b}\|^2] = d\sigma^2$. From (C.15), we have

$$\mathbb{E}_{\mathbf{b}}[\mathcal{R}_S^\lambda(\mathbf{w}_{\text{priv}}) - \mathcal{R}_S^\lambda(\widehat{\mathbf{w}})] \leq \frac{4(\beta+\lambda)L^2D_{\mathcal{X}}^2n^2d\log(\frac{1.25}{\delta})}{\epsilon^2\lambda^2(n_+n_-)^2}. \quad \text{(C.19)}$$

Plugging (C.19) in (C.14),

$$\begin{aligned}
&\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}} \mathcal{R}(\mathbf{w}) \\
&\leq 2\mathbb{E}_{\mathbf{b}}[\mathcal{R}_S^\lambda(\mathbf{w}_{\text{priv}}) - \mathcal{R}_S^\lambda(\mathbf{w}_\lambda)] + \frac{\lambda}{2}\|\mathbf{w}^*\|^2 + \mathcal{O}\Big(\frac{L^2D_{\mathcal{X}}^2\log(\frac{1}{\xi})n^3}{\lambda(n_+n_-)^2}\Big) \\
&\leq \frac{\lambda}{2}\|\mathbf{w}^*\|^2 + \frac{4(\beta+\lambda)L^2D_{\mathcal{X}}^2n^2d\log(\frac{1.25}{\delta})}{\epsilon^2\lambda^2(n_+n_-)^2} + \mathcal{O}\Big(\frac{L^2D_{\mathcal{X}}^2\log(\frac{1}{\xi})n^3}{\lambda(n_+n_-)^2}\Big) \\
&= \frac{\lambda}{2}\|\mathbf{w}^*\|^2 + \mathcal{O}\Big(\max\Big\{\frac{\beta(LD_{\mathcal{X}})^2n^2d\log(\frac{1}{\delta})}{\epsilon^2\lambda^2(n_+n_-)^2}, \frac{(LD_{\mathcal{X}})^2\log(\frac{1}{\xi})n^3}{\lambda(n_+n_-)^2}\Big\}\Big).
\end{aligned}$$

If $\lambda \leq \frac{\beta d\log(\frac{1}{\delta})}{\epsilon^2 n\log(\frac{1}{\xi})}$, we have

$$\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}} \mathcal{R}(\mathbf{w}) \leq \frac{\lambda}{2}\|\mathbf{w}^*\|^2 + \mathcal{O}\Big(\frac{\beta(LD_{\mathcal{X}})^2n^2d\log(\frac{1}{\delta})}{\epsilon^2\lambda^2(n_+n_-)^2}\Big),$$

else

$$\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}} \mathcal{R}(\mathbf{w}) \leq \frac{\lambda}{2}\|\mathbf{w}^*\|^2 + \mathcal{O}\Big(\frac{(LD_{\mathcal{X}})^2\log(\frac{1}{\xi})n^3}{\lambda(n_+n_-)^2}\Big).$$

Setting $\lambda = \min\Big\{\frac{(\log(\frac{1}{\delta})\beta)^{\frac{1}{3}}(LD_{\mathcal{X}}n)^{\frac{2}{3}}d^{\frac{1}{3}}}{(\|\mathbf{w}^*\|\epsilon(n_+n_-))^{\frac{2}{3}}}, \frac{LD_{\mathcal{X}}\sqrt{\log(\frac{1}{\delta})}n^{\frac{3}{2}}}{\|\mathbf{w}^*\|(n_+n_-)}\Big\}$, there holds

$$\begin{aligned}
&\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}} \mathcal{R}(\mathbf{w}) \\
&= \mathcal{O}\Big(\max\Big\{\frac{(\log(\frac{1}{\delta})\beta)^{\frac{1}{3}}(LD_{\mathcal{X}})^{\frac{2}{3}}d^{\frac{1}{3}}\|\mathbf{w}^*\|^{\frac{4}{3}}}{(\rho(1-\rho)\epsilon n)^{\frac{2}{3}}}, \frac{LD_{\mathcal{X}}\sqrt{\log(\frac{1}{\xi})}\|\mathbf{w}^*\|}{\rho(1-\rho)\sqrt{n}}\Big\}\Big).
\end{aligned}$$

This yields the desired results. □

C.4. Proof of Lemma 4

Proof. *(a)* By the strong smoothness of ℓ,

$$\mathcal{R}(\mathbf{w}_{\text{priv}}) - \mathcal{R}(\widehat{\mathbf{w}}) \leq \langle\nabla\mathcal{R}(\widehat{\mathbf{w}}), \widehat{\mathbf{w}}_{\text{priv}} - \widehat{\mathbf{w}}\rangle + \frac{\beta}{2}\|\mathbf{w}_{\text{priv}} - \widehat{\mathbf{w}}\|^2.$$

Therefore,

$$\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}}) - \mathcal{R}(\widehat{\mathbf{w}})] \leq \mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|]\|\nabla\mathcal{R}(\widehat{\mathbf{w}})\| + \frac{\beta}{2}\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|^2]. \quad \text{(C.20)}$$

Below we will estimate the terms on the right hand side of (C.20) one by one.

Firstly, we estimate terms $\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|]$ and $\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|^2]$. Specifically, since $\nu_1(\mathbf{b}) \propto \frac{1}{\alpha}\exp(-\frac{\|\mathbf{b}\|}{\gamma})$, then $\|\mathbf{b}\|$ is a random vector drawn from the distribution $\Gamma(d,\gamma)$. Thus $\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|] = d\gamma$, $\text{Var}(\|\mathbf{b}\|) = d\gamma^2$, and $\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|^2] = (d+d^2)\gamma^2$. Using the properties of strongly smoothness, we have $|\ell'(s) - \ell'(0)| \leq \beta|s|$ and $|\ell'(0)| \leq 2\sqrt{\beta\ell(0)}$, thus $|\ell'(s)| \leq 2\sqrt{\beta} + \beta|s|$ as $\ell(0) = 1$. Thus $B(\sqrt{2/\lambda}D_{\mathcal{X}}) \leq (\beta D_{\mathcal{X}}\sqrt{2/\lambda} + 2\sqrt{\beta})$. Recall that $\gamma = \frac{2D_{\mathcal{X}}B(\sqrt{2/\lambda}D_{\mathcal{X}})}{\epsilon\lambda}\Big(\frac{1}{n_+} + \frac{1}{n_-}\Big)$, there holds

$$\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|] \leq \frac{2D_{\mathcal{X}}d(\beta D_{\mathcal{X}}\sqrt{2/\lambda} + 2\sqrt{\beta})}{\epsilon\lambda}\Big(\frac{1}{n_+} + \frac{1}{n_-}\Big), \quad \text{(C.21)}$$

and

$$\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|^2] \leq \frac{4D_{\mathcal{X}}^2(\beta D_{\mathcal{X}}\sqrt{2/\lambda} + 2\sqrt{\beta})^2}{\epsilon^2\lambda^2}\Big(\frac{1}{n_+} + \frac{1}{n_-}\Big)^2(d+d^2). \quad \text{(C.22)}$$

Secondly, we estimate the term $\|\nabla\mathcal{R}(\widehat{\mathbf{w}})\|$. To this end, by the strong smoothness of ℓ, there holds $|\ell'(s)| \leq 2\sqrt{\beta\ell(s)}$,

$$\begin{aligned}
&\|\nabla\ell(\mathbf{w}^T(\mathbf{x}-\mathbf{x}')|y=1,y'=-1)\| \\
&\leq D_{\mathcal{X}}|\ell'(\mathbf{w}^T(\mathbf{x}-\mathbf{x}')|y=1,y'=-1)| \\
&\leq 2\sqrt{\beta}D_{\mathcal{X}}(\ell(\mathbf{w}^T(\mathbf{x}-\mathbf{x}')|y=1,y'=-1))^{\frac{1}{2}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\nabla\mathcal{R}(\widehat{\mathbf{w}})\| &\leq \mathbb{E}\|\nabla\ell(\widehat{\mathbf{w}}^T(\mathbf{x}-\mathbf{x}')|y=1,y'=-1)\| \\
&\leq 2\sqrt{\beta}D_{\mathcal{X}}(\mathcal{R}(\widehat{\mathbf{w}}))^{\frac{1}{2}}. \quad \text{(C.23)}
\end{aligned}$$

Hence it suffices to estimate $\mathcal{R}(\widehat{\mathbf{w}})$. Indeed, using the fact that $\ell((\mathbf{w}^*)^T(\mathbf{x}-\mathbf{x}')) \le \ell(0)+\ell'(0)D_{\mathcal{X}}\|\mathbf{w}^*\|+\frac{\beta D_{\mathcal{X}}^2}{2}\|\mathbf{w}^*\|^2$, and $\ell(0)=1$, $|\ell'(0)| \le 2\sqrt{\beta}$, and the definition of $\mathcal{R}(\mathbf{w}^*)$, we have that

$$\mathcal{R}(\mathbf{w}^*) \le 1 + 2\sqrt{\beta}D_{\mathcal{X}}\|\mathbf{w}^*\| + \frac{\beta D_{\mathcal{X}}^2}{2}\|\mathbf{w}^*\|^2.$$

Consequently,

$$\begin{aligned}\mathcal{R}(\widehat{\mathbf{w}}) &= [\mathcal{R}^\lambda(\widehat{\mathbf{w}}) - \mathcal{R}^\lambda(\mathbf{w}_\lambda)] + [\mathcal{R}^\lambda(\mathbf{w}_\lambda) - \mathcal{R}^\lambda(\mathbf{w}^*)] + \mathcal{R}^\lambda(\mathbf{w}^*) - \frac{\lambda}{2}\|\widehat{\mathbf{w}}\|^2 \\ &\le [\mathcal{R}^\lambda(\widehat{\mathbf{w}}) - \mathcal{R}^\lambda(\mathbf{w}_\lambda)] + 1 + 2\sqrt{\beta}D_{\mathcal{X}}\|\mathbf{w}^*\| + \frac{(\beta D_{\mathcal{X}}^2+\lambda)}{2}\|\mathbf{w}^*\|^2. \end{aligned} \tag{C.24}$$

Applying Lemma 3 with $B(\sqrt{2/\lambda}D_{\mathcal{X}}) = (\beta D_{\mathcal{X}}\sqrt{2/\lambda} + 2\sqrt{\beta})$ implies, with probability at least $1-\xi$, that

$$\mathcal{R}^\lambda(\widehat{\mathbf{w}}) - \mathcal{R}^\lambda(\mathbf{w}_\lambda) \le O\Big(\frac{\beta^2 D_{\mathcal{X}}^4 \log(\frac{1}{\xi})n^3}{\lambda^2(n_+n_-)^2}\Big).$$

Putting this back into (C.24), and then combining it with (C.23), we have that

$$\begin{aligned}\|\nabla\mathcal{R}(\widehat{\mathbf{w}})\| \le 2\sqrt{\beta}D_{\mathcal{X}}\Big(\frac{(\beta D_{\mathcal{X}}^2+\lambda)}{2}\|\mathbf{w}^*\|^2 + 2\sqrt{\beta}D_{\mathcal{X}}\|\mathbf{w}^*\| + 1\Big)^{\frac{1}{2}} \\ + O\Big(\frac{n^{\frac{3}{2}}\beta^{\frac{3}{2}}D_{\mathcal{X}}^3\sqrt{\log(\frac{1}{\xi})}}{n_+n_-\lambda}\Big).\end{aligned} \tag{C.25}$$

Now putting (C.21), (C.22) and (C.25) back into (C.20) yields that

$$\begin{aligned}\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}}) - \mathcal{R}(\widehat{\mathbf{w}})] \le O\Big(\frac{nD_{\mathcal{X}}^4\beta^2 d\|\mathbf{w}^*\|}{(n_+n_-)\epsilon\lambda^{\frac{3}{2}}} + \frac{n^{\frac{5}{2}}D_{\mathcal{X}}^5\beta^{\frac{5}{2}}d\sqrt{\log(\frac{1}{\xi})}}{(n_+n_-)^2\epsilon\lambda^{\frac{5}{2}}} \\ + \frac{n^2D_{\mathcal{X}}^4\beta^3(d+d^2)}{(n_+n_-)^2\epsilon^2\lambda^3}\Big).\end{aligned}$$

This completes the proof of part (a).

(b). The proof for part (b) is very similar to that for part (a) given above. The only difference is the estimation of the noise for $\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|$ and $\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|^2]$ which given by $\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|] \le \sigma\sqrt{d}$ and $\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|^2] = \sigma^2 d$. Recall $\sigma = \frac{2\sqrt{2\log(1.25/\delta)}D_{\mathcal{X}}B(\sqrt{2/\lambda}D_{\mathcal{X}})}{\epsilon\lambda}(\frac{1}{n_+}+\frac{1}{n_-})$ and $B(\sqrt{2/\lambda}D_{\mathcal{X}}) \le (\beta D_{\mathcal{X}}\sqrt{2/\lambda} + 2\sqrt{\beta})$. Consequently,

$$\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|] \le \frac{2\sqrt{2\log(1.25/\delta)}D_{\mathcal{X}}\sqrt{d}(\beta D_{\mathcal{X}}\sqrt{2/\lambda}+2\sqrt{\beta})}{\epsilon\lambda}\Big(\frac{1}{n_+}+\frac{1}{n_-}\Big), \tag{C.26}$$

and

$$\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|^2] \le \frac{8\log(1.25/\delta)D_{\mathcal{X}}^2 d(\beta D_{\mathcal{X}}\sqrt{2/\lambda}+2\sqrt{\beta})^2}{\epsilon^2\lambda^2}\Big(\frac{1}{n_+}+\frac{1}{n_-}\Big)^2. \tag{C.27}$$

Putting (C.25), (C.26) and (C.27) back into (C.20), we have that

$$\begin{aligned}&\mathbb{E}[\mathcal{R}(\mathbf{w}_{\text{priv}}) - \mathcal{R}(\widehat{\mathbf{w}})] \\ &= O\Big(\frac{nD_{\mathcal{X}}^4\beta^2\|\mathbf{w}^*\|\sqrt{d\log(\frac{1}{\delta})}}{(n_+n_-)\epsilon\lambda^{\frac{3}{2}}} + \frac{n^{\frac{5}{2}}D_{\mathcal{X}}^5\beta^{\frac{5}{2}}\sqrt{d\log(\frac{1}{\delta})\log(\frac{1}{\xi})}}{(n_+n_-)^2\epsilon\lambda^{\frac{5}{2}}} \\ &\quad + \frac{n^2D_{\mathcal{X}}^4\beta^2 d\log(\frac{1}{\delta})}{(n_+n_-)^2\epsilon^2\lambda^3}\Big).\end{aligned}$$

This completes the proof of the lemma. □

C.5. Proof of Theorem 5

Proof. In this theorem, we consider the case $\Delta = 0$. Since ℓ is L-Lipschitz, the error decomposition (18) and Lemma 3 with $\tau = \frac{1}{2}$ imply, with probability at least $1-\xi$, that

$$\begin{aligned}\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}}\mathcal{R}(\mathbf{w}) \le 2\mathbb{E}_{\mathbf{b}}[\mathcal{R}_S^\lambda(\mathbf{w}_{\text{priv}}) - \mathcal{R}_S^\lambda(\widehat{\mathbf{w}})] + \frac{\lambda}{2}\|\mathbf{w}^*\|^2 \\ + O\Big(\frac{L^2D_{\mathcal{X}}^2\log(\frac{1}{\xi})n^3}{\lambda(n_+n_-)^2}\Big).\end{aligned} \tag{C.28}$$

From the definition of $\mathbf{w}_{\text{priv}} = \arg\inf_{\mathbf{w}}\{\mathcal{R}_S^\lambda(\mathbf{w}) + \mathbf{b}^T\mathbf{w}\}$ and Cauchy–Schwarz inequality, there holds

$$\mathcal{R}_S^\lambda(\mathbf{w}_{\text{priv}}) - \mathcal{R}_S^\lambda(\widehat{\mathbf{w}}) \le \mathbf{b}^T(\widehat{\mathbf{w}} - \mathbf{w}_{\text{priv}}) \le \|\mathbf{b}\|\|\mathbf{w}_{\text{priv}} - \widehat{\mathbf{w}}\|. \tag{C.29}$$

To estimate $\|\mathbf{w}_{\text{priv}} - \widehat{\mathbf{w}}\|$, noticing that $\mathcal{R}_S^\lambda(\mathbf{w})$ is λ-strongly convex, we have that

$$(\nabla\mathcal{R}_S^\lambda(\mathbf{w}_{\text{priv}}) - \nabla\mathcal{R}_S^\lambda(\widehat{\mathbf{w}}))^T(\mathbf{w}_{\text{priv}} - \widehat{\mathbf{w}}) \ge \lambda\|\mathbf{w}_{\text{priv}} - \widehat{\mathbf{w}}\|^2,$$

This, by the Cauchy–Schwarz inequality, implies that

$$\|\mathbf{w}_{\text{priv}} - \widehat{\mathbf{w}}\| \le \frac{1}{\lambda}\|\nabla\mathcal{R}_S^\lambda(\mathbf{w}_{\text{priv}}) - \nabla\mathcal{R}_S^\lambda(\widehat{\mathbf{w}})\|.$$

From the definition of $\mathbf{w}_{\text{priv}}$ and $\widehat{\mathbf{w}}$, we know $\nabla\mathcal{R}_S^\lambda(\widehat{\mathbf{w}}) = \nabla\mathcal{R}_S^\lambda(\mathbf{w}_{\text{priv}}) + \mathbf{b} = 0$ which indicates that

$$\nabla\mathcal{R}_S^\lambda(\widehat{\mathbf{w}}) - \nabla\mathcal{R}_S^\lambda(\mathbf{w}_{\text{priv}}) = \mathbf{b}.$$

Consequently,

$$\|\mathbf{w}_{\text{priv}} - \widehat{\mathbf{w}}\| \le \|\mathbf{b}\|/\lambda. \tag{C.30}$$

Combining (C.29) and (C.30) implies that

$$\mathcal{R}_S^\lambda(\mathbf{w}_{\text{priv}}) - \mathcal{R}_S^\lambda(\widehat{\mathbf{w}}) \le \frac{\|\mathbf{b}\|^2}{\lambda}.$$

Now, plugging the estimation of $\mathcal{R}_S^\lambda(\mathbf{w}_{\text{priv}}) - \mathcal{R}_S^\lambda(\widehat{\mathbf{w}})$ into (C.28) yields that

$$\begin{aligned}\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}}\mathcal{R}(\mathbf{w}) \le \frac{2\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|^2]}{\lambda} + O\Big(\frac{L^2D_{\mathcal{X}}^2\log(\frac{1}{\xi})n^3}{\lambda(n_+n_-)^2}\Big) \\ + \frac{\lambda}{2}\|\mathbf{w}^*\|^2.\end{aligned} \tag{C.31}$$

Here we only consider the case $\Delta = 0$ which, as shown in Algorithm 2, will require the condition $\epsilon - n\log(1+\frac{\beta D_{\mathcal{X}}^2}{n_+n_-\lambda}) > 0$. In particular, we consider the following stronger condition

$$\epsilon' := \epsilon - n\log(1+\frac{\beta D_{\mathcal{X}}^2}{n_+n_-\lambda}) \ge \frac{\epsilon}{2} > 0. \tag{C.32}$$

The above condition is identical to $n\log(1+\frac{\beta D_{\mathcal{X}}^2}{n_+n_-\lambda}) \le \frac{\epsilon}{2}$. Using the elementary inequality that $\log(1+x) \le x$ for any $x > 0$, the condition (C.32) holds true if $\frac{n\beta D_{\mathcal{X}}^2}{n_+n_-\lambda} \le \frac{\epsilon}{2}$ or equivalently

$$\beta \le \frac{n_+n_-\lambda\epsilon}{2nD_{\mathcal{X}}^2}. \tag{C.33}$$

In the sequel, we will prove the generalization bounds for ϵ-DP and (ϵ, δ)-DP respectively, using the (C.31) under the condition (C.33) which ensures that $\Delta = 0$.

Firstly, for ϵ-DP, note that the condition (C.33) ensures that the choice of γ stated in Algorithm 2, i.e. $\gamma = \frac{2nLD_{\mathcal{X}}}{n_+n_-(\epsilon - n\log\left(1+\frac{\beta D_{\mathcal{X}}^2}{n_+n_-\lambda}\right)}$, satisfies that

$$\gamma = \frac{2nLD_{\mathcal{X}}}{n_+n_-(\epsilon - n\log(1+\frac{\beta D_{\mathcal{X}}^2}{n_+n_-\lambda})} \leq \frac{4nLD_{\mathcal{X}}}{n_+n_-\epsilon}.$$

In addition, notice that $\|\mathbf{b}\|$ is drawn from $\Gamma(d, \gamma)$, and then we have $\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|^2] = (d + d^2)\gamma^2$. Putting these estimations into (C.31) implies that

$$\begin{aligned}&\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}} \mathcal{R}(\mathbf{w})\\ &\leq \frac{32n^2L^2D_{\mathcal{X}}^2(d+d^2)}{(\epsilon n_+n_-)^2\lambda} + O(\frac{L^2D_{\mathcal{X}}^2\log(\frac{1}{\xi})n^3}{\lambda(n_+n_-)^2}) + \frac{\lambda}{2}\|\mathbf{w}^*\|^2.\end{aligned}$$

Now, setting $\lambda = \frac{LD_{\mathcal{X}}n\sqrt{d^2+n\log(\frac{1}{\xi})\epsilon^2}}{n_+n_-\epsilon\|\mathbf{w}^*\|}$ and recalling that imbalanced ratio $\rho = \frac{n_+}{n}$, we have

$$\begin{aligned}&\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}} \mathcal{R}(\mathbf{w})\\ &= O\Big(\max\Big\{\frac{LD_{\mathcal{X}}\sqrt{d^2+n\log(\frac{1}{\xi})\epsilon^2}\|\mathbf{w}^*\|}{\rho(1-\rho)\epsilon n}, \frac{LD_{\mathcal{X}}\sqrt{\log(\frac{1}{\xi})}\|\mathbf{w}^*\|}{\rho(1-\rho)\sqrt{n}}\Big\}\Big).\end{aligned}$$

Notice that

$$\begin{aligned}&\frac{LD_{\mathcal{X}}\sqrt{d^2+n\log(\frac{1}{\xi})\epsilon^2}\|\mathbf{w}^*\|}{\rho(1-\rho)\epsilon n}\\ &\leq 2\max\Big\{\frac{LD_{\mathcal{X}}d\|\mathbf{w}^*\|}{\rho(1-\rho)\epsilon n}, \frac{LD_{\mathcal{X}}\sqrt{\log(\frac{1}{\xi})}\|\mathbf{w}^*\|}{\rho(1-\rho)\sqrt{n}}\Big\}.\end{aligned}$$

Consequently,

$$\begin{aligned}&\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}} \mathcal{R}(\mathbf{w})\\ &= O\Big(\max\Big\{\frac{LD_{\mathcal{X}}d\|\mathbf{w}^*\|}{\rho(1-\rho)\epsilon n}, \frac{LD_{\mathcal{X}}\sqrt{\log(\frac{1}{\xi})}\|\mathbf{w}^*\|}{\rho(1-\rho)\sqrt{n}}\Big\}\Big).\end{aligned}$$

Notice the above estimation is true under the condition (C.33). This, by recalling the choice of $\lambda = \frac{LD_{\mathcal{X}}n\sqrt{d^2+n\log(\frac{1}{\xi})\epsilon^2}}{n_+n_-\epsilon\|\mathbf{w}^*\|}$, means that

$$\beta \leq \frac{n_+n_-\lambda\epsilon}{2nD_{\mathcal{X}}^2} = \frac{L\sqrt{d^2+n\log(\frac{1}{\xi})\epsilon^2}}{2D_{\mathcal{X}}\|\mathbf{w}^*\|}.$$

This completes the proof of part (a).

We now move on to the proof of the case of (ϵ, δ)-DP. In this case, $\mathbf{b} \sim \mathcal{N}(0, \sigma^2\mathbb{I})$, where $\sigma = (2\sqrt{2\log(\frac{1}{\delta})} + \sqrt{2\epsilon'})nLD_{\mathcal{X}}/(n_+n_-\epsilon')$. Recalling that (C.33) ensures that (C.32) which means that $\epsilon' \geq \frac{\epsilon}{2}$. Therefore,

$$\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|^2] = \sigma^2 d \leq \frac{4(2\sqrt{2\log(\frac{1}{\delta})} + \sqrt{\epsilon})^2n^2L^2D_{\mathcal{X}}^2d}{(n_+n_-\epsilon)^2}. \quad \text{(C.34)}$$

Putting this back into (C.31) implies that

$$\begin{aligned}\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}} \mathcal{R}(\mathbf{w}) &\leq \frac{8(2\sqrt{2\log(\frac{1}{\delta})} + \sqrt{\epsilon})^2n^2L^2D_{\mathcal{X}}^2d}{(n_+n_-\epsilon)^2\lambda}\\ &+ O(\frac{L^2D_{\mathcal{X}}^2\log(\frac{1}{\xi})n^3}{\lambda(n_+n_-)^2}) + \frac{\lambda}{2}\|\mathbf{w}^*\|^2.\end{aligned}$$

Setting $\lambda = \frac{LD_{\mathcal{X}}n\sqrt{(\sqrt{\log(\frac{1}{\delta})}+\sqrt{\epsilon})^2d+n\log(\frac{1}{\xi})\epsilon^2}}{\epsilon n_+n_-\|\mathbf{w}^*\|}$ and note that $\rho = \frac{n_+}{n}$, we have

$$\begin{aligned}&\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}} \mathcal{R}(\mathbf{w})\\ &= O\Big(\max\Big\{\frac{LD_{\mathcal{X}}\sqrt{(\sqrt{\log(\frac{1}{\delta})}+\sqrt{\epsilon})^2d+n\log(\frac{1}{\xi})\epsilon^2}\|\mathbf{w}^*\|}{\rho(1-\rho)\epsilon n},\\ &\quad\frac{LD_{\mathcal{X}}\sqrt{\log(\frac{1}{\xi})}\|\mathbf{w}^*\|}{\rho(1-\rho)\sqrt{n}}\Big\}\Big).\end{aligned}$$

Note that

$$\begin{aligned}&\frac{LD_{\mathcal{X}}\sqrt{(\sqrt{\log(\frac{1}{\delta})}+\sqrt{\epsilon})^2d+n\log(\frac{1}{\xi})\epsilon^2}\|\mathbf{w}^*\|}{\rho(1-\rho)\epsilon n}\\ &\leq 2\max\Big\{\frac{LD_{\mathcal{X}}(\sqrt{\log(\frac{1}{\delta})}+\sqrt{\epsilon})\sqrt{d}\|\mathbf{w}^*\|}{\rho(1-\rho)\epsilon n}, \frac{LD_{\mathcal{X}}\sqrt{\log(\frac{1}{\xi})}\|\mathbf{w}^*\|}{\rho(1-\rho)\sqrt{n}}\Big\},\end{aligned}$$

Consequently,

$$\begin{aligned}&\mathbb{E}_{\mathbf{b}}[\mathcal{R}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}} \mathcal{R}(\mathbf{w})\\ &= O\Big(\max\Big\{\frac{LD_{\mathcal{X}}(\sqrt{\log(\frac{1}{\delta})}+\sqrt{\epsilon})\sqrt{d}\|\mathbf{w}^*\|}{\rho(1-\rho)\epsilon n}, \frac{LD_{\mathcal{X}}\sqrt{\log(\frac{1}{\xi})}\|\mathbf{w}^*\|}{\rho(1-\rho)\sqrt{n}}\Big\}\Big).\end{aligned}$$

Notice the above estimation is true under the condition (C.33). This, by recalling the choice of $\lambda = \frac{LD_{\mathcal{X}}n\sqrt{(\sqrt{\log(\frac{1}{\delta})}+\sqrt{\epsilon})^2d+n\log(\frac{1}{\xi})\epsilon^2}}{\epsilon n_+n_-\|\mathbf{w}^*\|}$, means that

$$\beta \leq \frac{n_+n_-\lambda\epsilon}{2nD_{\mathcal{X}}^2} = \frac{L\sqrt{(\sqrt{\log(\frac{1}{\delta})}+\sqrt{\epsilon})^2d+n\log(\frac{1}{\xi})\epsilon^2}}{2D_{\mathcal{X}}\|\mathbf{w}^*\|}.$$

This completes the proof of part (b). □

C.6. Proof of Theorem 6

Proof. Assume S and S' differs in the first datum, i.e. $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_1', y_1')$. To show the differential privacy, it suffices to estimate the density ratio of $\frac{\text{pdf}(\mathbf{w}_{\text{priv}}|S)}{\text{pdf}(\mathbf{w}_{\text{priv}}|S')}$. This ratio can be written as

$$\frac{\text{pdf}(\mathbf{w}_{\text{priv}}|S)}{\text{pdf}(\mathbf{w}_{\text{priv}}|S')} = \frac{\text{pdf}(\mathbf{b}|S)}{\text{pdf}(\mathbf{b}'|S')} \cdot \frac{|\det(\mathbf{J}(\mathbf{w}_{\text{priv}} \to \mathbf{b}'|S'))|}{|\det(\mathbf{J}(\mathbf{w}_{\text{priv}} \to \mathbf{b}|S))|}. \quad \text{(C.35)}$$

We will estimate (C.35) in two steps. First, we bound the ratio of Jacobian determinants. Notice that

$$\mathbf{J}(\mathbf{w}_{\text{priv}} \to \mathbf{b}|S) = -\Big(\sum_{i=1}^{n} \ell''(\mathbf{w}_{\text{priv}}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^T + n(\lambda+\Delta)\mathbb{I}\Big)$$

Let $A = \sum_{i=2}^{n} \ell''(\mathbf{w}_{\text{priv}}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^T + n(\lambda+\Delta)\mathbb{I}$, $E = \ell''(\mathbf{w}_{\text{priv}}^T\mathbf{x}_1)\mathbf{x}_1\mathbf{x}_1^T$, $E' = \ell''(\mathbf{w}_{\text{priv}}^T\mathbf{x}_1')\mathbf{x}_1'\mathbf{x}_1'^T$. Then $\mathbf{J}(\mathbf{w}_{\text{priv}} \to \mathbf{b}|S) = -(A+E)$, and $\mathbf{J}(\mathbf{w}_{\text{priv}} \to \mathbf{b}'|S') = -(A+E')$. Then,

$$\begin{aligned}
&\frac{|\det(\mathbf{J}(\mathbf{w}_{\text{priv}} \to \mathbf{b}'|S'))|}{|\det(\mathbf{J}(\mathbf{w}_{\text{priv}} \to \mathbf{b}|S))|} \\
&= \frac{\det(A+E')}{\det(A+E)} = \frac{\det(A+E')}{\det(A)} \cdot \frac{\det(A)}{\det(A+E)} \\
&\le \frac{\det(A+E')}{\det(A)} \le 1 + \frac{\beta R_2^2}{n(\lambda+\Delta)},
\end{aligned} \tag{C.36}$$

where the first inequality follows from the PSD of E, and the last inequality used (8).

Second, we bound the ratio of the densities of noise. Let $\Gamma = \mathbf{b} - \mathbf{b}'$, from the definition of $\mathbf{w}_{\text{priv}}$, there holds $\mathbf{b} = -\Big(\sum_{i=1}^{n} y_i\ell'(\mathbf{w}^T\mathbf{x}_i)\mathbf{x}_i + n(\lambda+\Delta)\mathbf{w}_{\text{priv}}\Big)$, which implies that $\|\Gamma\| = \|\mathbf{b} - \mathbf{b}'\| \le 2LR_2$. Therefore,

$$\begin{aligned}
\frac{\text{pdf}(\mathbf{b}|S)}{\text{pdf}(\mathbf{b}'|S')} &= \frac{e^{\frac{-\|\mathbf{b}\|^2}{2\sigma^2}}}{e^{\frac{-\|\mathbf{b}'\|^2}{2\sigma^2}}} = \exp\Big(\frac{1}{2\sigma^2}(\|\Gamma\|^2 - 2\langle\mathbf{b},\Gamma\rangle)\Big) \\
&\le \exp\Big(\frac{1}{2\sigma^2}((2LR_2)^2 + 2|\langle\mathbf{b},\Gamma\rangle|)\Big).
\end{aligned} \tag{C.37}$$

Let the event $\mathcal{E} = \{\mathbf{b} \in \mathbb{R}^d : |\langle\mathbf{b},\Gamma\rangle| \ge 2LR_2\sigma t\}$. Since $\langle\mathbf{b},\Gamma\rangle \sim \mathcal{N}(0, \|\Gamma\|^2\sigma^2)$, there holds $\Pr(\mathcal{E}) = \Pr(|\langle\mathbf{b},\Gamma\rangle| \ge 2LR_2\sigma t) \le e^{-\frac{t^2}{2}}$. Plugging this inequality into (C.37) and choosing $t = \sqrt{2\log\frac{1}{\delta}}$, then there holds $\frac{\text{pdf}(\mathbf{b}|S)}{\text{pdf}(\mathbf{b}'|S')} \le e^{\frac{2(LR_2)^2}{\sigma^2} + \frac{2LR_2}{\sigma}\sqrt{2\log(\frac{1}{\delta})}}$. For any $\epsilon' > 0$, if $\sigma \ge \frac{(2\sqrt{2\log(\frac{1}{\delta})} + \sqrt{2\epsilon'})LR_2}{\epsilon'}$, we have $\frac{\text{pdf}(\mathbf{b}|S)}{\text{pdf}(\mathbf{b}'|S')} \le e^{\epsilon'}$ on the event $\mathcal{E}^c$.

Combining (C.35) (C.36) and the above estimation, and choosing $\sigma \ge \frac{(2\sqrt{2\log(\frac{1}{\delta})} + \sqrt{2\epsilon'})LR_2}{\epsilon'}$, there holds

$$\frac{\text{pdf}(\mathbf{w}_{\text{priv}}|S)}{\text{pdf}(\mathbf{w}_{\text{priv}}|S')} \le \exp\Big(\log\Big(1 + \frac{\beta R_2^2}{n(\lambda+\Delta)}\Big) + \epsilon'\Big). \tag{C.38}$$

on the event $\mathcal{E}^c$. We now consider two cases. If $\log\Big(1 + \frac{\beta R_2^2}{n\lambda}\Big) < \epsilon$, letting $\Delta = 0$. And we choose $\epsilon' = \epsilon - \log\Big(1 + \frac{\beta R_2^2}{n\lambda}\Big)$. If $\log\Big(1 + \frac{\beta R_2^2}{n\lambda}\Big) \ge \epsilon$, letting $\epsilon' = \frac{\epsilon}{2}$, and $\Delta = \frac{\beta R_2^2}{n(e^{\epsilon/2}-1)} - \lambda$. Therefore, for any set $E \subseteq \mathbb{R}^d$

$$\begin{aligned}
\Pr(\mathbf{w}_{\text{priv}}(S) \in E) &= \Pr(\mathbf{w}_{\text{priv}}(S) \in E \cap \mathcal{E}) + \Pr(\mathbf{w}_{\text{priv}}(S) \in E \cap \mathcal{E}^c) \\
&\le \Pr(\mathcal{E}) + \Pr(\mathbf{w}_{\text{priv}}(S) \in E \cap \mathcal{E}^c) \\
&\le \delta + \int_{E\cap\mathcal{E}^c} \text{pdf}(\mathbf{w}_{\text{priv}} = \alpha|S))d\alpha \\
&\le \delta + e^{\epsilon}\int_{E\cap\mathcal{E}^c} \text{pdf}(\mathbf{w}_{\text{priv}} = \alpha|S'))d\alpha \\
&\le \delta + e^{\epsilon}\Pr(\mathbf{w}_{\text{priv}}(S') \in E).
\end{aligned}$$

This completes the proof of the theorem. □

C.7. Proof of Theorem 7

Proof. From the proof of Lemma 19 in [4], we can see that

$$J(\mathbf{w}_{\text{priv}}, S) - J(\widehat{\mathbf{w}}, S) \le \frac{1}{n}\mathbf{b}^T(\widehat{\mathbf{w}} - \mathbf{w}_{\text{priv}}) \le \frac{\|\mathbf{b}\|^2}{n^2\lambda}. \tag{C.39}$$

Now, putting (C.39) into (21), and taking expectation over $\mathbf{b}$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{b}}[\mathcal{L}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}} \mathcal{L}(\mathbf{w}) \le &\frac{2\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|^2]}{n^2\lambda} + O\Big(\frac{L^2R_2^2\log(1/\xi)}{\lambda n}\Big) \\
&+ \frac{\lambda}{2}\|\mathbf{w}^*\|^2.
\end{aligned} \tag{C.40}$$

In this theorem, we only consider the case $\Delta = 0$ which, as shown in Algorithm 3, will require the condition $\epsilon - \log(1 + \frac{\beta R_2^2}{n\lambda}) > 0$. Here, we consider the stronger condition, $\epsilon' := \epsilon - \log(1 + \frac{\beta R_2^2}{n\lambda}) \ge \frac{\epsilon}{2} > 0$. Notice that $\log(1+x) \le x$ for any $x > 0$, the above condition holds true if $\beta \le \frac{\lambda\epsilon n}{2R_2^2}$.

Now, we present the generalization bounds for (ϵ,δ)-DP. Note that the condition of β ensures that $\Delta = 0$, and thus noise $\mathbf{b}$ is drawn from $\nu(\mathbf{b};\epsilon,\delta,\sigma)$ with $\sigma = (2\sqrt{2\log(\frac{1}{\delta})} + \sqrt{2\epsilon'})LR_2/\epsilon'$. Therefore, $\mathbb{E}_{\mathbf{b}}[\|\mathbf{b}\|^2] = \sigma^2 d \le 4(2\sqrt{2\log(\frac{1}{\delta})} + \sqrt{\epsilon})^2L^2R_2^2d/\epsilon^2$. Putting this back into (C.40) and setting $\lambda = \frac{8LR_2\sqrt{(\sqrt{\log(\frac{1}{\delta})}+\sqrt{\epsilon})^2d + n\log(\frac{1}{\xi})\epsilon^2}}{\epsilon n\|\mathbf{w}^*\|}$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{b}}[\mathcal{L}(\mathbf{w}_{\text{priv}})] - \inf_{\mathbf{w}} \mathcal{L}(\mathbf{w}) \\
&= O\Big(\max\Big\{\frac{LR_2(\sqrt{\log(\frac{1}{\delta})} + \sqrt{\epsilon})\sqrt{d}\|\mathbf{w}^*\|}{\epsilon n}, \frac{LR_2\sqrt{\log(\frac{1}{\xi})}\|\mathbf{w}^*\|}{\sqrt{n}}\Big\}\Big).
\end{aligned}$$

Notice the above estimation is true under the condition $\beta \le \frac{n\lambda\epsilon}{2R_2^2}$. This, by recalling the choice of $\lambda = \frac{8LR_2\sqrt{(\sqrt{\log(\frac{1}{\delta})}+\sqrt{\epsilon})^2d + n\log(\frac{1}{\xi})\epsilon^2}}{\epsilon n\|\mathbf{w}^*\|}$, means that

$$\beta \le \frac{\lambda\epsilon n}{2R_2^2} = \frac{4L\sqrt{(\sqrt{\log(\frac{1}{\delta})} + \sqrt{\epsilon})^2d + n\log(\frac{1}{\xi})\epsilon^2}}{R_2\|\mathbf{w}^*\|}.$$

□