Differentially Private Reward Estimation with Preference Feedback

Sayak Ray Chowdhury*
Microsoft Research, India

Xingyu Zhou*
Wayne State University, USA

Nagarajan Natarajan
Microsoft Research, India

Abstract

Learning from preference-based feedback has recently gained considerable traction as a promising approach to align generative models with human interests. Instead of relying on numerical rewards, the generative models are trained using reinforcement learning with human feedback (RLHF). These approaches first solicit feedback from human labelers typically in the form of pairwise comparisons between two possible actions, then estimate a reward model using these comparisons, and finally employ a policy based on the estimated reward model. An adversarial attack in any step of the above pipeline might reveal private and sensitive information of human labelers. In this work, we adopt the notion of *label differential privacy* (DP) and focus on the problem of reward estimation from preference-based feedback while protecting privacy of each individual labelers. Specifically, we consider the parametric Bradley-Terry-Luce (BTL) model for such pairwise comparison feedback involving a latent reward parameter $\theta^* \in \mathbb{R}^d$. Within a standard minimax estimation framework, we provide tight upper and lower bounds on the error in estimating θ^* under both *local* and *central* models of DP. We show, for a given privacy budget ε and number of samples n, that the additional cost to ensure label-DP under local model is $\Theta\big(\frac{1}{e^\varepsilon-1}\sqrt{\frac{d}{n}}\big)$, while it is $\Theta\big(\frac{\text{poly}(d)}{\varepsilon n}\big)$ under the weaker central model. We perform simulations on synthetic data that corroborate these theoretical results.

1 INTRODUCTION

In recent years, the problem of aligning generative models to human preferences has garnered a lot of interest (Christiano et al., 2017; Glaese et al., 2022; Ouyang et al., 2022). One of the most promising approaches to achieve this is via preference-based reinforcement learning (Christiano et al., 2017). It has gained considerable attention across multiple application domains such as game playing (MacGlashan et al., 2017), large language models (Ouyang et al., 2022), and robotics (Shin et al., 2023).

Preference-based learning: In standard RL, the agent learns to maximize a numerical reward, which she observes from the environment. In the above applications, however, observing appropriate numerical rewards can be challenging, which could significantly affect the performance of the agent. In such cases, it is of common practice to solicit feedback from a human labeler in the form of pairwise comparisons between two possible actions at every state (Christiano et al., 2017). Notably, the language model application InstructGPT (Ouyang et al., 2022) is based on this feedback model. First, the prompts (states) are sampled from a precollected datasest, and then, for each prompt, a pair of responses (actions) are sampled by deploying the pre-trained model. For each prompt, a human labeler provides pairwise preferences over the responses, which are then used to train a reward model by maximum likelihood estimation, or, equivalently, by cross-entropy minimization (Christiano et al., 2017). Finally, this reward model is used for a downstream policy training (i.e., finetuning the pre-trained model). This complete pipeline forms the basis of preference-based RL, see e.g. Pacchiano et al. (2021); Chen et al. (2022); Zhu et al. (2023); Zhan et al. (2023).

Privacy (or the lack of it): One important aspect which is ignored in the aforementioned learning literature is protecting privacy of human labelers. Potentially sensitive information of an individual can be revealed through the collected (pairwise comparisons) feedback in case of an adversarial attack at any stage of the RL pipeline. In fact, after the emergence of

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s). *Equal contributions

ChatGPT, several instances of privacy breach including that of human labelers have been reported (Li et al., 2023). Since then, efforts have been made to privately fine-tune large language models (Yu et al., 2021; Behnia et al., 2022).

Label-Differential Privacy: In view of this, Differential privacy (DP) (Dwork, 2008) is the most adopted notion to protect the sensitive information of individuals whose preference feedback is used during the model training. The prompts (states) are not considered sensitive since they are typically sampled from a pre-collected dataset which is already public knowledge. In this work, we develop new results for privacy (as well as accuracy) of estimators obtained with such potentially sensitive feedback information via the notion of label differential privacy (Label-DP). This notion of label-DP has been studied previously in deep learning (Ghazi et al., 2021) and in learning theory in general (Chaudhuri and Hsu, 2011; Beimel et al., 2013).

We focus on the problem of reward estimation from pairwise preferences while protecting the privacy of individual labelers. Specifically, we consider the parametric Bradley-Terry-Luce (BTL) model for such feedback involving a latent reward parameter $\theta^* \in \mathbb{R}^d$. We prove upper and lower bounds on the error in estimating θ^* under *local* (where the learner only observes privatized labels) and *central* models (where the learner has access to the raw non-private data) of DP.

Our contributions are summarized below:
(1) We show that the additional cost in estimation for ensuring ε-label-DP under local model is at least $\Omega\big(\frac{1}{e^\varepsilon-1}\sqrt{\frac{d}{n}}\big)$, where ε is a given privacy budget and n is the total number of samples.
(2) For the local model, we design an estimator of θ^* based on the *Randomized Response* (RR) mechanism (Warner, 1965) that satisfies ε label-DP and achieves a matching upper bound on estimation error. To do so, we design a novel loss function tailored to RR, which de-biases the effect of label randomization and can potentially be of independent interest.
(3) For the central model, we show that the additional cost for ensuring (ε, δ)-label-DP under this weaker privacy model is at least $\Omega\big(\frac{1}{\varepsilon+\delta}\frac{\sqrt{d}}{n}\big)$ for $\delta \in (0, 1)$.
(4) Finally, for the central model, we also provide a matching upper bound (in n and ε) by designing an estimator of θ^* based on the classical *objective perturbation* technique with Gaussian privacy noise.

We present numerical simulations on synthetic data to support our theoretical results.

Related work. Our work is inspired by a recent study on reward estimation (and offline bandits/RL) under the linearly parametrized BTL model without privacy protection (Zhu et al., 2023). Our work introduces label-DP into the same setting and provides sharp results on estimation errors as well as some downstream applications. Both Zhu et al. (2023) and our work can be viewed as a generalization of the work on non-private reward estimation under the *tabular* BTL model (Shah et al., 2015), which studies estimation error under both semi-norm and ℓ_2-norm. Label-DP is first introduced by Chaudhuri and Hsu (2011) for private PAC learners. Recently, it has been leveraged to yield better performance for many practical situations where only labels are sensitive data, relative to standard DP which is an overkill (Ghazi et al., 2021; Malek Esmaeili et al., 2021; Esfandiari et al., 2022). As in Esfandiari et al. (2022), we consider label DP under both local and central models. We also remark that our work differs from the vast literature on private logistic regression (or stochastic optimization) (e.g., Chaudhuri and Monteleoni (2008); Song et al. (2021); Bassily et al. (2014)) in the performance metrics, i.e., parameter estimation error vs. generalization/excess population risk. See more details and additional related work in Appendix A.

2 PRELIMINARIES

Let $\mathcal{D} = (s_i, a_i^0, a_i^1, y_i)_{i=1}^n$ be a dataset of n samples, where each sample has a state $s_i \in \mathcal{S}$ (e.g., prompt given to a language model), two actions $a_i^0, a_i^1 \in \mathcal{A}$ (e.g., two responses from the language model), and label $y_i \in \{0, 1\}$ indicating which action is preferred by humans experts. We assume that the state s_i is first sampled from some fixed distribution ρ. The pair of actions (a_i^0, a_i^1) are then sampled from some joint distribution (i.e., a behavior policy) μ conditioned on s_i. Finally, the label y_i is sampled from a Bernoulli distribution conditioned on (s_i, a_i^0, a_i^1), i.e., for $l \in \{0, 1\}$,

$$\mathbb{P}_{\theta^*}[y_i = l | s_i, a_i^0, a_i^1] = \frac{\exp(r_{\theta^*}(s_i, a_i^l))}{\exp(r_{\theta^*}(s_i, a_i^0)) + \exp(r_{\theta^*}(s_i, a_i^1))}.$$

Here $r_{\theta^*}(\cdot, \cdot)$ is the reward model parameterized by an unknown parameter θ^*, which we would want to estimate using $\mathcal{D}$. This model is often called Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952; Luce, 2012).

In this work, we consider a linear reward model $r_{\theta^*}(s, a) = \phi(s, a)^\top \theta^*$, where $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is some known and fixed feature map. For instance, such a ϕ can be constructed by removing the last layer of a pre-trained language model, and in that case, θ^* correspond to the weights of the last layer. With this model, one can equivalently write the probability of sampling $y_i = 1$ given (s_i, a_i^0, a_i^1) as

$$\mathbb{P}_{\theta^*}[y_i = 1 | s_i, a_i^0, a_i^1] = \sigma\Big(\big(\phi(s_i, a_i^1) - \phi(s_i, a_i^0)\big)^\top \theta^*\Big),$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function. We let $x_i = \phi(s_i, a_i^1) - \phi(s_i, a_i^0)$ denote the differential feature of actions a_i^1 and a_i^0 at state s_i. This lets us denote, for any $\theta \in \mathbb{R}^d$, the predicted probabilities of a label y_i given x_i as (we omit dependence on θ for brevity)

$$p_{i,1} := \mathbb{P}_\theta [y_i = 1 | x_i] = \sigma(x_i^\top \theta) \ , \ p_{i,0} := 1 - p_{i,1} \ . \quad (1)$$

We make the following assumption which is standard in the literature (Shah et al., 2015; Zhu et al., 2023).

Assumption 2.1 (Boundedness). (a) θ^* lies in the set $\Theta_B = \{\theta \in \mathbb{R}^d | \langle \mathbf{1}, \theta \rangle = 0, \|\theta\| \le B\}$. The condition $\langle \mathbf{1}, \theta \rangle = 0$ ensures identifiability of θ^*. (b) Features are bounded, i.e., $\|\phi(s, a)\| \le L$, $\forall (s, a)$.

Now, we recall the notion of differential privacy (Dwork, 2008). Roughly, it ensures that the output of an algorithm $\mathcal{M}$ operating on a dataset $\mathcal{D}$ doesn't change much if we change a single example in $\mathcal{D}$. In this paper, we adopt the notion of *label DP* (Ghazi et al., 2021) to protect sensitive information that lies in preference-based feedback y_i. This is motivated by the fact that in most applications, the data (s_i, a_i^0, a_i^1) presented to the human annotator is public (or pre-collected) while the feedback $y_i \in \{0, 1\}$ indicates her personal preference, which needs to be protected.

Definition 2.2 (Label DP in Central Model). Let $\varepsilon \ge 0, \delta \in (0, 1]$. A randomized algorithm $\mathcal{M}$ is said to be (ε, δ)-label differentially private in central model if for any two datasets $\mathcal{D}$ and $\mathcal{D}'$ that differ in the *label* of a single sample and for any subset S in the range of $\mathcal{M}$, it holds that

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

If $\delta = 0$, $\mathcal{M}$ is said to be ε-label DP. We will simply call it central label DP in the following.

For our specific reward estimation problem, Definition 2.2 roughly means that any single change of feedback label will not change the final estimator too much.

The central DP model assumes that the learning agent $\mathcal{A}$ has access to preference feedback given by human labelers in the clear-text. In some applications, however, the individual labelers might not be willing to share their feedback in the clear-text. This motivates us to consider label DP in the local model, where each feedback y_i, before being observed by the agent, is first privatized by some local randomizer $\mathcal{R}$ at each labeler, which is formally defined as follows.

Definition 2.3 (Label DP in Local Model). If each label is first privatized by a local randomizer $\mathcal{R}$, which satisfies for any y, y' and any subset S in the range of $\mathcal{R}$, it holds that

$$\mathbb{P}[\mathcal{R}(y) \in S] \le e^\varepsilon \cdot \mathbb{P}[\mathcal{R}(y') \in S] + \delta,$$

then, we say $\mathcal{R}$ is an (ε, δ)-label differentially private local randomizer, and the entire algorithm (e.g., estimator) that operates with the randomized labels is said to satisfy local label DP.

Remark 2.4. Note that for central label DP, the privacy burden lies in the central agent while for local label DP, the privacy protection relies on local randomizer $\mathcal{R}$. By post-processing of DP (Dwork, 2008), an algorithm that satisfies local label DP also satisfies central local DP. Thus, in the following, we will first focus on the stronger local model of privacy.

The rest of the paper is organized as follows. In the next two sections, we will focus on local label DP and present the lower bound and upper bounds for the estimation error, respectively. In Section 5, we turn to the central label DP and also present corresponding lower and upper bounds.

3 LOCAL MODEL: LOWER BOUND ON ESTIMATION ERROR

In this section, we present the lower bounds on the estimation error under local label DP (cf. Definition 2.3). Let $\Sigma_\mathcal{D} := \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ denote the sample covariance matrix of differential features $x_i = \phi(s_i, a_i^1) - \phi(s_i, a_i^0)$. Then, for any $\lambda > 0$, we have the following lower bound on the estimation error in the semi-norm $\|\cdot\|_{\Sigma_\mathcal{D} + \lambda I}$.

Theorem 3.1 (Semi-norm lower bound). *For a large enough n, any estimator $\widehat{\theta}$ based on n samples from the BTL model that satisfies ε-label DP in the local model has estimation error in semi-norm lower bounded as*

$$\mathbb{E}\left[\left\|\widehat{\theta} - \theta^*\right\|^2_{\Sigma_\mathcal{D} + \lambda I}\right] \ge \Omega\left(\frac{d}{n} + \frac{d}{(e^\varepsilon - 1)^2 n}\right).$$

Shah et al. (2015) shows that (squared) error of estimation under the tabular BTL model of pairwise comparisons is at least $\Omega(d/n)$ without any privacy constraint. In comparison, we pay an additional $\Omega\big(\frac{d}{(e^\varepsilon - 1)^2 n}\big)$ error in estimation in order to ensure ε-label DP. We suffer a similar privacy cost while bounding estimation error in ℓ_2-norm too. The result is formally stated below

Theorem 3.2 (ℓ_2-norm lower bound). *Under the same hypothesis of Theorem 3.1, the estimation error of $\widehat{\theta}$ in ℓ_2-norm is lower bounded as*

$$\mathbb{E}\left[\left\|\widehat{\theta} - \theta^*\right\|^2\right] \ge \Omega\left(\frac{d}{L^2} \cdot \left(\frac{d}{n} + \frac{d}{(e^\varepsilon - 1)^2 n}\right)\right).$$

Note that the lower bound in ℓ_2 norm is an $\Omega(d)$ multiplicative factor higher than the one in semi-norm (when $L = \Theta(1)$). A similar comparative behavior holds for non-private lower bounds too (Shah et al., 2015). Moreover, if the differential features x_i is distributed according to a standard Gaussian, then we have $L = O(\sqrt{d})$. In this case, the first term in ℓ_2-norm lower bound reduces to $\Omega(d/n)$, which recovers the mean-squared-error (MSE) lower bound under Gaussian design without any privacy considerations (Chen et al., 2016; Hsu and Mazumdar, 2023).

Proof summary of lower bounds. For both lower bounds, we leverage the classic reduction from estimation to testing. In particular, for the ℓ_2 norm, we apply a variant of the (private) version of Assouad's lemma (cf. Yu (1997)) by constructing a hypercube over the underlying parameter space. On the other hand, for the semi-norm case, it is somewhat difficult to construct a hypercube. Instead, we turn to (private) version of Fano's lemma, which only requires a packing (in terms of semi-norm). This can be achieved by Varshamov–Gilbert's bound (cf. Guntuboyina (2011)) and vector rotations. For the privacy parts in both bounds, we leverage strong data processing inequality under local DP (cf. Duchi et al. (2018)). The complete proofs for both results are presented in Appendix B.

4 LOCAL MODEL: UPPER BOUND ON ESTIMATION ERROR

In this section, we discuss private estimators of the unknown parameter θ^* and develop a series of results that answer the following questions.

(1) *Is the standard MLE estimator useful under the Randomized Response model, and in what privacy regime?*

(2) *Can we design an estimator for all privacy regimes that achieves the same order of estimation as in the lower bound?*

(3) *How do we compute the estimator efficiently?*

(4) *Can we extend the ideas to other popular preference feedback models such as* Thurstone *and* Plackett-Luce*?*

We first describe the Randomized Response (RR) mechanism (Warner, 1965), which we use to guarantee local label DP.

Randomized Response. Let $\varepsilon \geq 0$ be the privacy budget and $y \in \{0, 1\}$ be the true label. When queried the value of y, the RR mechanism outputs $\widetilde{y}$, which is randomly sampled from the probability distribution

$$\mathbb{P}\left[\widetilde{y} = y\right] = \frac{e^{\varepsilon}}{1 + e^{\varepsilon}} = \sigma(\varepsilon),\ \mathbb{P}\left[\widetilde{y} \neq y\right] = 1 - \sigma(\varepsilon)\ . \quad (2)$$

It is well-known that RR is ε-DP (Dwork, 2008). In the following, we will use RR as $\mathcal{R}$ in Definition 2.3 to achieve label DP in the local model.

We start with a simple maximum likelihood estimator (MLE), which will help us develop intuition for a comparatively complex but a better estimator.

4.1 The Maximum Likelihood Estimator

For any $\theta \in \mathbb{R}^d$, (1) and (2) together define predicted probabilities of a randomized label $\widetilde{y}_i$ given x_i as

$$\widetilde{p}_{i,1} = \sigma(x_i^\top \theta)\sigma(\varepsilon) + (1 - \sigma(x_i^\top \theta))(1 - \sigma(\varepsilon))\ ,$$
$$\widetilde{p}_{i,0} = (1 - \sigma(x_i^\top \theta))\sigma(\varepsilon) + \sigma(x_i^\top \theta)(1 - \sigma(\varepsilon))\ .$$

With n such pairs of features and randomized labels $(x_i, \widetilde{y}_i)_{i=1}^n$, the private MLE $\widetilde{\theta}_{\text{MLE-RR}}$ aims to minimize the negative log-likelihood

$$\widetilde{l}_{\mathcal{D},\varepsilon}(\theta) = -\sum_{i=1}^{n} \Big[\mathbb{1}(\widetilde{y}_i = 1) \log \widetilde{p}_{i,1} + \mathbb{1}(\widetilde{y}_i = 0) \log \widetilde{p}_{i,0}\Big]. \quad (3)$$

As mentioned before, $\widetilde{\theta}_{\text{MLE}}$ is ε-label DP for any $\varepsilon \geq 0$. Recall that $\Sigma_{\mathcal{D}} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$ denotes the sample covariance matrix of differential features and let γ be a constant such that $\sigma'(x^\top \theta) \geq \gamma$ for all $\theta \in \Theta_B$ and for all features x. Under Assumption 2.1, $\gamma = \frac{1}{2 + e^{-2LB} + e^{2LB}}$ satisfies this condition. Then, we have the following estimation error bound for $\widetilde{\theta}_{\text{MLE}}$.

Theorem 4.1 (Estimation error of MLE)**.** *Fix* $\alpha \in (0, 1), \varepsilon > 2LB, \lambda > 0$. *Then, under Assumption 2.1, with probability at least* $1 - \alpha$*, we have*

$$\|\widetilde{\theta}_{MLE} - \theta^*\|_{\Sigma_{\mathcal{D}} + \lambda I} \leq \frac{C}{\gamma} \frac{e^{\varepsilon + 2LB} + 1}{e^{\varepsilon - 2LB} - 1} \sqrt{\frac{d + \log(1/\alpha)}{n}} + \sqrt{\lambda} B,$$

where C *is some absolute constant.*

The above error bound of $\widetilde{\theta}_{\text{MLE}}$ holds only when the privacy budget is higher than a certain threshold (which depends on the norm of θ^* and ϕ), i.e., when $\varepsilon > 2LB$, thus limiting its applicability only to lower privacy regimes (since a high value of ε implies a low level of privacy). This is due to the fact that $\widetilde{\theta}_{\text{MLE}}$ minimizes a noisy objective (3), that is strongly convex in the semi-norm $\|\cdot\|_{\Sigma_{\mathcal{D}}}$ only if $\varepsilon > 2LB$, which is a crucial step in bounding the estimation error (see Appendix C.1 for details). Instead, we want an objective that is strongly convex in the semi-norm for all privacy levels $\varepsilon > 0$. This leads us to an estimator that is specifically tailored to the RR mechanism.

4.2 An Estimator Tailored to RR

For any $\theta \in \mathbb{R}^d$, the logits (log-odds) of the probability that the clear-text label $y_i = 1$ given x_i is

$$\text{logit}(p_{i,1}) = \log \frac{p_{i,1}}{p_{i,0}} = \log \frac{\sigma(x_i^\top \theta)}{1 - \sigma(x_i^\top \theta)},$$

where the same for randomized label $\widetilde{y}_i = 1$ is

$$\operatorname{logit}(\widetilde{p}_{i,1}) = \log \frac{\sigma(x_i^\top\theta)\sigma(\varepsilon)+(1-\sigma(x_i^\top\theta))(1-\sigma(\varepsilon))}{(1-\sigma(x_i^\top\theta))\sigma(\varepsilon)+\sigma(x_i^\top\theta)(1-\sigma(\varepsilon))} \ .$$

It holds that (see Appendix C for details)

$$\operatorname{logit}(\widetilde{p}_{i,1}) \le \sigma(\varepsilon)\cdot \operatorname{logit}(p_{i,1}) \text{ if } p_{i,1} \ge p_{i,0} \ ,$$
$$\operatorname{logit}(\widetilde{p}_{i,0}) \le \sigma(\varepsilon)\cdot \operatorname{logit}(p_{i,0}) \text{ if } p_{i,0} \ge p_{i,1} \ .$$

Since $\sigma(\varepsilon) \in (1/2, 1)$ for any $\varepsilon > 0$, this implies that whenever y_i is more likely to occur than $1 - y_i$ in the clear-text, the log-odds of predicting y_i under ε-randomization given by (2) is at most $\sigma(\varepsilon)$-th fraction of the corresponding log-odds in the clear-text. This makes the objective (3) ill-suited for obtaining a tight estimator for θ^* under randomization of labels.

Essentially, we want to design an objective (or, equivalently a loss function) so that the log-odds of predictions under randomization is same as the log-odds in the clear-text. The following loss achieves this:

$$\widehat{l}_{\mathcal{D},\varepsilon}(\theta) = -\sum_{i=1}^{n}\Big[\mathbb{1}(\widetilde{y}_i = 1)\log \widehat{p}_{i,1} + \mathbb{1}(\widetilde{y}_i = 0)\log \widehat{p}_{i,0}\Big], \quad (4)$$

where we define, for any $\theta \in \mathbb{R}^d$, the predicted *scores* of each randomized label $\widetilde{y}_i$ given x_i as

$$\widehat{p}_{i,1} = \frac{\sigma(x_i^\top\theta)^{\sigma(\varepsilon)}}{(1-\sigma(x_i^\top\theta))^{(1-\sigma(\varepsilon))}}, \ \widehat{p}_{i,0} = \frac{(1-\sigma(x_i^\top\theta))^{\sigma(\varepsilon)}}{\sigma(x_i^\top\theta)^{(1-\sigma(\varepsilon))}}. \quad (5)$$

Although $\widehat{p}_{i,1}$ and $\widehat{p}_{i,0}$ are not probabilities, these satisfy our desired property:

$$\log \frac{\widehat{p}_{i,1}}{\widehat{p}_{i,0}} = \log \frac{\sigma(x_i^\top\theta)}{1-\sigma(x_i^\top\theta)} = \operatorname{logit}(p_{i,1}) \ .$$

Hence the loss function $\widehat{l}_{\mathcal{D},\varepsilon}(\theta)$ essentially de-biases the effect of randomization. This, in turn, yields that $\widehat{l}_{\mathcal{D},\varepsilon}(\theta)$ is $\gamma(2\sigma(\varepsilon)-1)$ strongly convex in the semi-norm $\|\cdot\|_{\Sigma_\mathcal{D}}$ for all $\theta \in \Theta_B$, and importantly, it holds for any $\varepsilon > 0$. This helps us obtain an estimator for θ^* with error bound for all privacy levels $\varepsilon > 0$, defined as

$$\widehat{\theta}_{\text{RR}} \in \operatorname{argmin}_{\theta\in\Theta_B} \widehat{l}_{\mathcal{D},\varepsilon}(\theta) \ . \quad (6)$$

$\widehat{\theta}_{\text{RR}}$ satisfies ε-label DP due to RR and post-processing of DP. Now, for any constant $\lambda > 0$, we have the following estimation error bound for $\widehat{\theta}_{\text{RR}}$. Proof of this result is deferred to Appendix C.2.

Theorem 4.2 (Estimation error of $\widehat{\theta}_{\text{RR}}$)**.** *Fix $\alpha \in (0,1), \varepsilon > 0, \lambda > 0$. Then, under Assumption 2.1, with probability at least $1-\alpha$, we have*

$$\|\widehat{\theta}_{RR} - \theta^*\|_{\Sigma_\mathcal{D}+\lambda I} \le \frac{C}{\gamma}\frac{e^\varepsilon+1}{e^\varepsilon-1}\sqrt{\frac{d+\log(1/\alpha)}{n}} + C'\sqrt{\lambda}B, \quad (7)$$

where $\gamma = \frac{1}{2+e^{-2LB}+e^{2LB}}$, C, C' are absolute constants.

Cost of Privacy. We compare the error of our private estimator $\widehat{\theta}_{\text{RR}}$ with that of the clear-text (i.e., non-private) estimator θ_{MLE}, which minimizes the following non-private negative log-likelihood

$$l_\mathcal{D}(\theta) = -\sum_{i=1}^{n}\Big[\mathbb{1}(y_i = 1)\log p_{i,1} + \mathbb{1}(y_i = 0)\log p_{i,0}\Big]. \quad (8)$$

As shown in Zhu et al. (2023), θ_{MLE} achieves an estimation error of $O\big(\sqrt{d/n}\big)$ in semi-norm. Comparing this with Theorem 4.2, we observe that the cost of ensuring label DP for $\widehat{\theta}_{\text{RR}}$ is of the order $O\left(\frac{1}{e^\varepsilon-1}\sqrt{\frac{d}{n}}\right)$, which almost matches lower bound discussed below.

Comparison with Lower Bound. We now compare the upper bound in Theorem 4.2 with the semi-norm lower bound in Theorem 3.1. Setting $\lambda = \left(\frac{e^\varepsilon+1}{e^\varepsilon-1}\right)^2 \frac{d+\log(1/\alpha)}{B^2\gamma^2 n}$, we see that the upper bound matches the lower bound up to a factor of $O(1/\gamma) \approx e^{LB}$. Hence, if both $L = O(1)$ and $B = O(1)$, the bounds are tight up to a constant factor.

Applications in Contextual Bandits. In applications such as offline linear contextual bandits (Li et al., 2022), this bound can then be used to learn a downstream pessimistic policy

$$\widehat{\pi}_\Theta = \operatorname*{argmax}_{\pi\in\Pi} \inf_{\theta\in\Theta} \mathbb{E}_{s\sim\rho}\left[\phi(s,\pi(s))^\top\theta\right]. \quad (9)$$

Here Π is the set of all action selection policies $\pi : \mathcal{S} \to \mathcal{A}$ and Θ is a high-probability confidence set for θ^*, i.e., it is a set of all $\theta \in \Theta_B$ that satisfies (7). Similar to Li et al. (2022), one can show that this pessimistic policy achieves a *sub-optimality gap* of $O\Big(\frac{L}{\gamma}\frac{e^\varepsilon+1}{e^\varepsilon-1}\sqrt{\frac{d}{n}}\,\|\Sigma_\mathcal{D}+\lambda I\|^{-1/2} + \sqrt{\lambda}LB\Big)$ in high probability for any $\lambda > 0$ while guaranteeing label DP.

4.3 Efficient Computation via SGD

It is evident that computing the exact minimizer $\widehat{\theta}_{\text{RR}}$ in (4) is impractical in practice – an issue shared by the non-private estimator θ_{MLE} of Zhu et al. (2023) as well. Note that even if an approximate solution is allowed, it still requires solving the optimization problem up to a certain accuracy level so as to preserve the same estimation error bound. This motivates us to consider the (one-pass) SGD algorithm, which iterates over each sample once. In particular, we replace (6) by a sequential update rule:

$$\widehat{\theta}_1 = 0 \ , \ \widehat{\theta}_{t+1} = \Pi_{\Theta_B}\Big(\widehat{\theta}_t - \eta_t \widehat{g}_t\Big), \ 1 \le t \le n \ . \quad (10)$$

Here Π_{Θ_B} is a projection operator onto the set Θ_B, η_t is a suitable learning rate and $\widehat{g}_t = -\nabla_{\widehat{\theta}_t} \log \widehat{p}_{t,\widetilde{y}_t}$ is the (negative) gradient of the log-predicted score of randomized label $\widetilde{y}_t$ computed at current estimate $\widehat{\theta}_t$, where $\widehat{p}_{t,y}, y \in 0,1$ is given by (5). We denote the estimate after n iterations as $\widehat{\theta}_{\text{SGD-RR}}$.

Although the error bound under semi-norm as proved in Theorem 4.2 does not hold for this SGD variant, we can bound the estimation error of $\widehat{\theta}_{\text{SGD-RR}}$ in ℓ_2-norm. To begin with, we note that Theorem 4.2 implies a bound on the estimation error of $\widehat{\theta}_{\text{RR}}$ in ℓ_2-norm.

Corollary 4.3. *Under the same hypothesis of Theorem 4.2, we have, with probability at least $1-\alpha$,*

$$\|\widehat{\theta}_{RR}-\theta^*\|_2 \le \frac{C}{\gamma\sqrt{\lambda_{\min}(\Sigma_{\mathcal{D}})}}\frac{e^\varepsilon+1}{e^\varepsilon-1}\sqrt{\frac{d+\log(1/\alpha)}{n}} \ .$$

where $\lambda_{\min}(\Sigma_{\mathcal{D}})$ is the minimum eigenvalue of $\Sigma_{\mathcal{D}}$.

This bound is non-trivial only if assume that sample covariance matrix $\Sigma_{\mathcal{D}}$ is positive definite. One can also relax this assumption to the population covariance matrix of differential state-action features $x = \phi(s,a^1)-\phi(s,a^0)$, defined as

$$\Sigma = \mathbb{E}_{s\sim\rho(\cdot),(a^0,a^1)\sim\mu(\cdot|s)}\left[xx^\top\right] \ .$$

Assumption 4.4 (Coverage of feature space)**.** The data distributions ρ,μ are such that $\lambda_{\min}(\Sigma) \ge \kappa$ for some constant $\kappa > 0$.

This is essentially a coverage assumption on the state-action feature space, which is standard in offline bandits and RL (Yin et al., 2022). The next result bounds the estimation error of $\widehat{\theta}_{\text{SGD-RR}}$ in ℓ_2-norm.

Theorem 4.5 (Estimation error of $\widehat{\theta}_{\text{SGD-RR}}$)**.** *Fix $\alpha \in (0,1/e)$ and $\varepsilon \ge 0$. Then, under Assumptions 2.1 and 4.4 and setting $\eta_t = \frac{1}{\gamma\kappa}$, we have, with probability at least $1-\alpha$,*

$$\left\|\widehat{\theta}_{SGD\text{-}RR}-\theta^*\right\|_2 \le C\cdot\frac{L}{\gamma\kappa}\cdot\frac{e^\varepsilon+1}{e^\varepsilon-1}\sqrt{\frac{\log\left(\log(n)/\alpha\right)}{n}},$$

where $\gamma = \frac{1}{2+e^{-2LB}+e^{2LB}}$, C is an absolute constant.

The complete algorithm and proof of the theorem is deferred to Appendix C.3. In fact, we prove a stronger and general result than Theorem 4.5 by bounding the estimator error uniformly for all intermediate parameter estimates $\widehat{\theta}_{t+1}$, $1 \le t \le n$, with $\sqrt{1/n}$ replaced by $\sqrt{1/t}$ in the bound. The $\log\log n$ term is the (minimal) cost to ensure uniform concentration over all $t \le n$. The bound for $\widehat{\theta}_{\text{SGD-RR}}$ follows by setting $t = n$ in the general result. The key idea behind this result is to show the gradient $\widehat{g}_t$ in the SGD update (10) is an unbiased estimate of the gradient (except some scaling) in the clear-text $g_t = -\nabla_{\widehat{\theta}_t}\log p_{t,y_t}$, where $p_{t,y}, y \in \{0,1\}$ denotes the probability of observing y at round t, see (1). Specifically, we have

$$\widehat{g}_t = \frac{\sum_{y\in\{0,1\}}\nabla_{\widehat{\theta}_t}\log p_{t,y}}{e^\varepsilon+1} - \nabla_{\widehat{\theta}_t}\log p_{t,\widetilde{y}_t} \ ,$$

which, in turn, gives $\mathbb{E}\left[\widehat{g}_t|x_t,y_t,\widehat{\theta}_t\right] = (2\sigma(\varepsilon)-1)\,g_t$, where the expectation is over the ε-randomization of clear-text label y_t given by (2). This, along with the coverage assumption and the fact that $\sigma'(x_t^\top\theta) \ge \gamma$ for all $\theta \in \Theta_B$ help us achieve the desired error bound.

Comparison with Semi-norm Bound. The main difference compared to the semi-norm bound in Theorem 4.2 is the inverse dependence on coverage parameter κ – estimation error increases as κ decreases. Another apparent difference is the dependence (or the lack of it) on the feature dimension d in the error bound. However, κ is a problem dependent quantity. It depends implicitly on the dimension d of feature space (Wang et al., 2020), thereby capturing the dependence of error bound on d. For example, since $\|x\| \le L$, we have $\kappa = O(L^2/d)$ under Assumption 2.1. In the best case when $\kappa = \Theta(L^2/d)$, the error bound in ℓ_2-norm scales as $\widetilde{O}\left(\frac{1}{\gamma}\frac{e^\varepsilon+1}{e^\varepsilon-1}\frac{d}{\sqrt{n}}\right)$, which is $\sqrt{d}$ factor higher than that in semi-norm $\|\cdot\|_{\Sigma_{\mathcal{D}}}$. Finally, due to the coverage assumption, instead of employing a pessimistic policy as in (9) for a downstream offline contextual bandit task, we can design a greedy (plug-in) policy $\widehat{\pi}_{\text{Greedy}}(s) = \text{argmax}_{a\in\mathcal{A}}\,\phi(s,a)^\top\widehat{\theta}_{\text{SGD-RR}}$, which achieves a *sub-optimality gap* of $\widetilde{O}\left(\frac{L^2}{\gamma\kappa}\frac{e^\varepsilon+1}{e^\varepsilon-1}\frac{1}{\sqrt{n}}\right)$ in high-probability while ensuring label-DP.

Comparison with Lower Bound. We now compare the upper bound in Theorem 4.5 with the ℓ_2-norm lower bound in Theorem 3.2. First, we note that $\kappa = O(L^2/d)$ under Assumption 2.1. That is, in the best case when $\kappa = \Theta(L^2/d)$, the upper bound in Theorem 4.5 becomes $\widetilde{O}\left(\frac{d}{L\gamma\sqrt{n}}\frac{e^\varepsilon+1}{e^\varepsilon-1}\right)$. This matches the lower bound up to a factor of $O(e^{LB})$. Hence, similar to the semi-norm bounds, if L, B are $\Theta(1)$, the ℓ_2-norm bounds are also tight up to a constant factor.

Remark 4.6. A similar SGD update as (10) is used in Ghazi et al. (2021) in the context of private stochastic convex optimization (SCO). They bound the excess population risk of the SGD estimate in expectation under the clear-text distribution $(x_t, y_t) \sim P$. Natarajan et al. (2013) consider a similar objective function as (4) in the context of binary classification with noisy labels. They obtain a classifier by minimizing the empirical risk on the noisy samples $(x_t,\widetilde{y}_t)_{t=1}^n$ (which is equivalent to maximizing (4)) and bound its excess population risk under the clear-text distribution. In contrast to both works, we aim to bound the estimation error of the parameter estimate in high probability, which brings additional challenges in the analysis.

4.4 Extensions to Other Preference Models

Thurstone Model. Our result can be extended to any pairwise comparison model of the form

$$\mathbb{P}_{\theta^*}\left[y_i = 1 | s_i, a_i^0, a_i^1\right] = F(x_i^\top \theta^*)$$

if F satisfies following two properties: (i) F is an even function, i.e., $F(z) = 1 - F(-z)$ for all z and (ii) F is strongly log-concave in an interval around $z = 0$, i.e., there is a curvature parameter $\gamma > 0$ and a range parameter $c > 0$ such that

$$\frac{d^2}{dz^2}\left(-\log(F(z))\right) \geq \gamma \quad \forall\, z \in [-c, c]\ .$$

For the BTL model, where F is specified by the sigmoid function, these properties hold for $c = 2LB$ and $\gamma = \frac{1}{2+\exp(-2LB)+\exp(2LB)}$ under Assumption 2.1.

In the Thurstone model (Thurstone, 1927), each label $y_i \in \{0, 1\}$ is sampled from the conditional distribution

$$\mathbb{P}_{\theta^*}\left[y_i = 1 | s_i, a_i^0, a_i^1\right] = \Phi(x_i^\top \theta^*)\ ,$$

where Φ is the CDF of standard Gaussian distribution. It holds that Φ is strongly log-concave for all $\theta \in \Theta_B$ under Assumption 2.1 (Tsukida et al., 2011). Hence, a similar error bound as Theorem 4.5 for the SGD-based estimator holds under Thurstone model with a proper choice of the curvature parameter γ.

Placket-Luce Model. One practical extension of our results is to privately learn the reward parameter θ^* from K-wise comparisons between actions, which is captured by the Placket-Luce (PL) model (Plackett, 1975; Luce, 2012). Let s be a state and $a_1, \ldots, a_K$ be K actions to be compared at that state. Let the label/preference feedback $y \in \{1, 2, \ldots, K\}$ indicates which action is most preferred by human labeler.[1] Under the Placket-Luce model, the label y is sampled according to the probability distribution

$$\mathbb{P}_{\theta^*}[y = k | s, a_1, \ldots, a_K] = \frac{\exp(\phi(s, a_k)^\top \theta^*)}{\sum_{j=1}^K \exp(\phi(s, a_j)^\top \theta^*)}. \quad (11)$$

In this case, label-DP is ensured by employing the K-Randomized Response (K-RR) mechanism, which, when queried the value of y, outputs $\widetilde{y}$ that is randomly sampled from the probability distribution:

$$\mathbb{P}\left[\widetilde{y} = y\right] = \frac{e^\varepsilon}{e^\varepsilon + K - 1},\ \mathbb{P}\left[\widetilde{y} = y'\right] = \frac{1}{e^\varepsilon + K - 1} \forall y' \neq y\ . \quad (12)$$

Given n samples $(s_t, a_{t,1}, \ldots a_{t,k}, y_t)_{t=1}^n$, we estimate θ^* using the SGD update of (10), with the gradient

$$\widehat{g}_t = \frac{\sum_{y=1}^K \nabla_{\widehat{\theta}_t} \log p_{t,y}}{e^\varepsilon + K - 1} - \nabla_{\widehat{\theta}_t} \log p_{t,\widetilde{y}_t}\ ,$$

where $p_{t,y}, y \in [K]$ is the probability of observing y at round t, see (11).

Let $x_{i,j} = \phi(s, a_i) - \phi(s, a_j)$ be the feature difference between actions a_i and a_j at state and $\Sigma_{i,j} = \mathbb{E}[x_{i,j} x_{i,j}^\top]$ be the corresponding population covariance matrix. Assume there exists a coverage parameter $\kappa > 0$ such that $\Sigma_{ij} \geq \kappa$ for all pair of actions (a_i, a_j). Then, similar to Theorem 4.5, we can prove an error bound for this SGD-based estimator with K-RR, denoted by $\widehat{\theta}_{\text{SGD-KRR}}$. Specifically, we have

$$\left\|\widehat{\theta}_{\text{SGD-KRR}} - \theta^*\right\|_2 = \widetilde{O}\left(\frac{L}{\gamma\kappa} \cdot \frac{e^\varepsilon + K - 1}{e^\varepsilon - 1} \frac{1}{\sqrt{n}}\right)$$

with high probability, where $\gamma = \frac{1}{e^{4LB}}$. See Appendix C.4 for a precise statement and complete proof.

5 CENTRAL MODEL: ESTIMATION ERROR BOUNDS

In this section, we turn to study label DP in the central model where the learning agent has access to the clear-text dataset $\mathcal{D}$ and it only needs to guarantee the estimator is "insensitive" with respect to any single change of the label. Under this weaker privacy model, we show that the estimation error can be greatly improved compared to those in the local model. In the main paper, we will mainly focus on ℓ_2-norm bounds and leave semi-norm bounds to Appendix D.4.

5.1 Lower Bound

We first have the following lower bound on estimation error, the proof of which is given in Appendix D.1.

Theorem 5.1. *For a large enough n, any estimator $\widehat{\theta}$ based on samples form the BTL model that satisfies (ε, δ)-label DP in the central model has the estimation error in ℓ_2-norm lower bounded as*

$$\mathbb{E}\left[\left\|\widehat{\theta} - \theta^*\right\|^2\right] \geq \Omega\left(\frac{d^2}{nL^2} + \frac{d}{n^2(\varepsilon + \delta)^2}\right).$$

Let us compare our lower bound with a similar one (although via a different approach) established in Cai et al. (2023), which enforces privacy protection for both label and features (i.e., standard DP notion rather than label DP). If $L = O(\sqrt{d})$ (which holds for Gaussian design), then the first term is the same as in Cai et al. (2023) and is equal to the standard non-private mean-square-error (MSE) lower bound (Hsu and Mazumdar, 2023). The main difference is the dependence on dimension in the second term, i.e., d in our bound vs. d^2 in Cai et al. (2023). This improvement is due to the fact that our privacy protection is only for the scalar label, whereas Cai et al. (2023)

[1] We differ here from the standard PL model, where human labeler outputs the entire ranking between K actions.

also protects d-dimensional feature vectors (albeit in a different application than ours).

5.2 Algorithm and Upper Bound

In this section, we present our algorithm and its privacy and estimation error guarantees.

Our algorithm builds upon the classic technique – objective perturbation (Kifer et al., 2012), i.e., it adds an additional noise term in the objective function. In particular, our estimator is given by

$$\widehat{\theta}_{\rm obj} = \mathop{\rm argmin}_{\theta\in\Theta_B} l_{\mathcal{D}}(\theta) + \frac{\beta}{2}\left\|\theta\right\|_2^2 + w^\top\theta \ ,$$

where $l_{\mathcal{D}}(\theta)$ is the negative log-likelihood defined in (8), $\beta > 0$ is some regularizer and $w \sim \mathcal{N}(0, \sigma^2 I)$ is an independent Gaussian noise. We then have the following privacy guarantee. See Appendix D.2 for the proof and Algorithm 2 for pseudo-code.

Theorem 5.2 (Privacy). *Let $\varepsilon > 0$, $\delta \in (0,1)$. Then, setting $\sigma = \frac{L\sqrt{8\log(2/\delta)+4\varepsilon}}{\varepsilon}$ under Assumption 2.1, Algorithm 2 satisfies (ε,δ)-label DP in the central model.*

The next theorem provides the estimation error of $\widehat{\theta}_{\rm obj}$ in ℓ_2-norm. See Appendix D.3 for the full proof.

Theorem 5.3 (Estimation error). *Let $\alpha \in (0,1)$. Then, under Assumptions 2.1 and 4.4, with probability at least $1-\alpha$, $\widehat{\theta}_{obj}$ satisfies*

$$\left\|\widehat{\theta}_{obj}-\theta^*\right\|_2 \le O\left(\frac{L}{\kappa\gamma}\sqrt{\frac{\log(1/\alpha)}{n}} + \frac{\sigma\big(\sqrt{d}+\sqrt{\log(1/\alpha)}\big)}{n\kappa\gamma}\right)$$

where $\gamma := \frac{1}{2+\exp(-2LB)+\exp(2LB)}$ and κ is the coverage coefficient in Assumption 4.4.

Corollary 5.4. *For $\varepsilon \in (0,1]$ and $\delta \in (0,1)$, let $\sigma = \frac{L\sqrt{8\log(2/\delta)+4\varepsilon}}{\varepsilon}$ as in Theorem 5.2, then the estimation error is of the order*

$$\left\|\widehat{\theta}_{obj}-\theta^*\right\|_2 \le \widetilde{O}\left(\frac{L}{\kappa\gamma\sqrt{n}} + \frac{L\sqrt{d\log(1/\delta)}}{n\varepsilon\kappa\gamma}\right).$$

Remark 5.5 (Central vs. Local Models). A key observation here is that the cost to ensure label DP is an additive lower-order term of the order $\frac{1}{\varepsilon n}$ under the central model. In contrast, under the local model, the privacy cost is of the order $\frac{1}{(e^\varepsilon-1)\sqrt{n}}$ (cf. Theorem 4.5), which is approximately $\frac{1}{\varepsilon\sqrt{n}}$ for high privacy regime (i.e., $\varepsilon < 1$). This sharp decrease in privacy cost under the central model is due to the fact that it is a weaker privacy model; hence, instead of randomizing each label, one needs to add noise only once in the loss function.

Comparison with Lower Bound. We now compare the upper bound in Corollary 5.4 with lower bound in Theorem 5.1. Under Assumption 2.1 and in the best case when $\kappa = \Theta(L^2/d)$, we observe that the upper bound matches the lower bound up to a factor of de^{LB}. If $L = B = \Theta(1)$, the gap between the bounds is on the order of d. Closing this gap and obtaining optimal error bounds is an open question.

Extension to approximate minimizer. Currently, the privacy guarantee in Theorem 5.2 only holds for the exact minimizer $\widehat{\theta}_{\rm obj}$. One can also add an additional output perturbation as in Bassily et al. (2019); Iyengar et al. (2019) to guarantee that an approximate minimizer (e.g., obtained by SGD) is private with the same order of estimation error.

6 SIMULATIONS

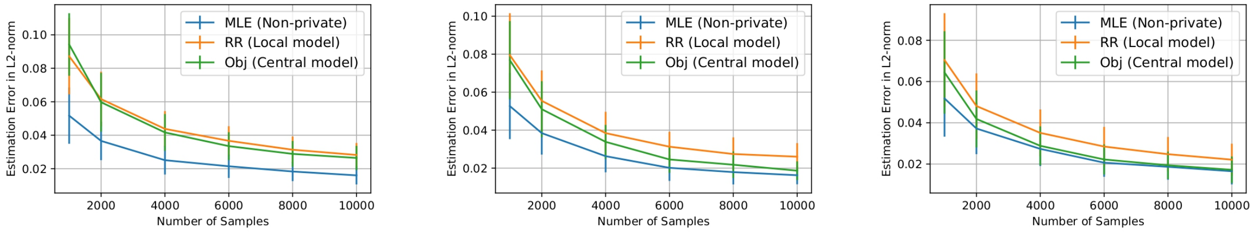
We numerically evaluate the errors of our estimators under local and central models of label DP; and present comparisons with that of the non-private estimator of Zhu et al. (2023). Our simulations are proof-of-concept only; we do not tune any hyper-parameters.

We consider the BTL preference model of pairwise comparisons, with the number of samples $1000 \leqslant n \leqslant 10000$. Each sample size is repeated 100 times. We randomly generate θ^* from d-dimensional standard Gaussian, where we vary $d \in \{3,5,10\}$. The state-action features ϕ are sampled iid, also from a d-dimensional standard Gaussian. The results for $d=5$ is shown in Figure 1.

For the local model, we use our SGD-based estimator $\widehat{\theta}_{\text{SGD-RR}}$. For the central model, we implement our objective perturbation based estimator $\widehat{\theta}_{\rm obj}$ using SGD updates. To ensure consistency, we also implement the non-private estimator $\theta_{\rm MLE}$ of Zhu et al. (2023) using SGD updates. We use learning rate $\eta = 0.1$ for all three estimators. We compare estimation error in ℓ_2-norm for all the estimators for varying privacy levels $\varepsilon \in \{0.1, 0.5, 1\}$. We fix $\delta = 0.001$ for $\widehat{\theta}_{\rm obj}$.

We observe that estimation error decreases for all the estimators as the number of samples grows larger. Moreover, we observe that the non-private MLE-based estimator has the smallest estimation error, while the error in RR based estimator under local model is higher than the objective perturbation based estimator under the central model. This is consistent with our theoretical results in Sections 4 and 5. [2]

[2] Codes can be found at https://github.com/sayakrc/Differentialy_Private_Estimation.



(a) $\varepsilon = 0.1$ (b) $\varepsilon = 0.5$ (c) $\varepsilon = 1$

Figure 1: Comparison of estimation error in ℓ_2-norm between non-private MLE-based estimator, RR based estimator in local model and objective perturbation based estimator in central model for different privacy levels ε.

7 CONCLUSION

We provided a systematic study of reward estimation via human feedback under the label DP. We also discuss the generalization to standard DP in Appendix E. For future directions, it is instructive to establish upper bounds with a dependency milder than e^{LB}. Along the lines of Bach (2010), where risk bounds are considered, it might be possible to leverage self-concordance properties of log-loss to obtain tighter estimation error bounds. Another important direction is to empirically evaluate the performance of a downstream policy trained using the estimated reward model along the lines of Zhu et al. (2023) under different privacy notions and budgets.

Acknowledgments

XZ is supported in part by NSF CNS-2153220 and CNS-2312835. The authors would like to thank Adam Smith, Daniel Kifer, and Abhradeep Guha Thakurta for discussions on the objective perturbation paper (Kifer et al., 2012). XZ would also like to thank Abhradeep Guha Thakurta, Satyen Kale, and Karan Singh for discussions on Agarwal et al. (2023), which highlights some subtlety in proving privacy guarantee in objective perturbation. XZ would also like to thank Yichen Wang for the discussions on Cai et al. (2023).

References

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. Interactive learning from policy-dependent human feedback. In *International Conference on Machine Learning*, pages 2285–2294. PMLR, 2017.

Daniel Shin, Anca D Dragan, and Daniel S Brown. Benchmarks and algorithms for offline preference-based reward learning. *arXiv preprint arXiv:2301.01392*, 2023.

Aldo Pacchiano, Aadirupa Saha, and Jonathan Lee. Dueling rl: reinforcement learning with trajectory preferences. *arXiv preprint arXiv:2111.04850*, 2021.

Xiaoyu Chen, Han Zhong, Zhuoran Yang, Zhaoran Wang, and Liwei Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, pages 3773–3793. PMLR, 2022.

Banghua Zhu, Jiantao Jiao, and Michael I Jordan. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. *arXiv preprint arXiv:2301.11270*, 2023.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable offline reinforcement learning with human feedback. *arXiv preprint arXiv:2305.14816*, 2023.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*, 2023.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*, 2021.

Rouzbeh Behnia, Mohammadreza Reza Ebrahimi, Jason Pacheco, and Balaji Padmanabhan. Ew-tune: A framework for privately fine-tuning large language models with differential privacy. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 560–566. IEEE, 2022.

Cynthia Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation: 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings 5*, pages 1–19. Springer, 2008.

Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning with label differential privacy. *Advances in neural information processing systems*, 34:27131–27145, 2021.

Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 155–186. JMLR Workshop and Conference Proceedings, 2011.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques: 16th International Workshop, APPROX 2013, and 17th International Workshop, RANDOM 2013, Berkeley, CA, USA, August 21-23, 2013. Proceedings*, pages 363–378. Springer, 2013.

Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309): 63–69, 1965.

Nihar Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In *Artificial intelligence and statistics*, pages 856–865. PMLR, 2015.

Mani Malek Esmaeili, Ilya Mironov, Karthik Prasad, Igor Shilov, and Florian Tramer. Antipodes of label differential privacy: Pate and alibi. *Advances in Neural Information Processing Systems*, 34:6934–6945, 2021.

Hossein Esfandiari, Vahab Mirrokni, Umar Syed, and Sergei Vassilvitskii. Label differential privacy via clustering. In *International Conference on Artificial Intelligence and Statistics*, pages 7055–7075. PMLR, 2022.

Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. *Advances in neural information processing systems*, 21, 2008.

Shuang Song, Thomas Steinke, Om Thakkar, and Abhradeep Thakurta. Evading the curse of dimensionality in unconstrained private glms. In *International Conference on Artificial Intelligence and Statistics*, pages 2638–2646. PMLR, 2021.

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

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2012.

Xi Chen, Adityanand Guntuboyina, and Yuchen Zhang. On bayes risk lower bounds. *The Journal of Machine Learning Research*, 17(1):7687–7744, 2016.

Daniel Hsu and Arya Mazumdar. On the sample complexity of estimation in logistic regression. *arXiv preprint arXiv:2307.04191*, 2023.

Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pages 423–435. Springer, 1997.

Adityanand Guntuboyina. Lower bounds for the minimax risk using f-divergences, and applications. *IEEE Transactions on Information Theory*, 57(4): 2386–2399, 2011.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

Gene Li, Cong Ma, and Nati Srebro. Pessimism for offline linear contextual bandits using lp confidence sets. *Advances in Neural Information Processing Systems*, 35:20974–20987, 2022.

Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

Ruosong Wang, Dean P Foster, and Sham M Kakade. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

Louis L Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.

Kristi Tsukida, Maya R Gupta, et al. How to analyze paired comparison data. *Department of Electrical Engineering University of Washington, Tech. Rep. UWEETR-2011-0004*, 1, 2011.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

T Tony Cai, Yichen Wang, and Linjun Zhang. Score attack: A lower bound technique for optimal differentially private learning. *arXiv preprint arXiv:2303.07152*, 2023.

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

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019.

Roger Iyengar, Joseph P Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 299–316. IEEE, 2019.

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4(none): 384 – 414, 2010. doi: 10.1214/09-EJS521. URL https://doi.org/10.1214/09-EJS521.

Naman Agarwal, Satyen Kale, Karan Singh, and Abhradeep Thakurta. Differentially private and lazy online convex optimization. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4599–4632. PMLR, 2023.

Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *The Journal of Machine Learning Research*, 22(1):278–385, 2021.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(none):1 – 6, 2012. doi: 10.1214/ECP.v17-2079. URL https://doi.org/10.1214/ECP.v17-2079.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private assouad, fano, and le cam. In *Algorithmic Learning Theory*, pages 48–78. PMLR, 2021.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Roshan Shariff and Or Sheffet. Differentially private contextual linear bandits. *Advances in Neural Information Processing Systems*, 31, 2018.

Sayak Ray Chowdhury and Xingyu Zhou. Shuffle private linear contextual bandits. In *Proceedings of the 39th International Conference on Machine Learning*, pages 3984–4009. PMLR, 2022.

Adam Smith, Abhradeep Thakurta, and Jalaj Upadhyay. Is interaction necessary for distributed private learning? In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 58–77. IEEE, 2017.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A ADDITIONAL RELATED WORK

Preference-based Learning. Shah et al. (2015) study the problem of reward estimation under the pairwise (BTL and Thurstone) model and K-wise (PL) comparisons model. They work in the tabular setting and provide minimax error bounds for estimated rewards under both semi-norm and ℓ_2-norm. Zhu et al. (2023) consider linearly parameterized rewards under the BTL and PL comparisons model, and prove the error bound of the maximum likelihood estimator. They further use this estimator to learn a pessimistic policy for an offline contextual bandit task and bound its sub-optimality gap. Multi-armed bandits (both tabular and parametric) under dueling/preference-based feedback are considered in a range of work, and different notions of regret guarantees are established; see Bengs et al. (2021) for a comprehensive survey. Pacchiano et al. (2021) consider the episodic online RL problem under the BTL comparison model with tabular latent rewards and prove sublinear *regret* guarantees in the number of episodes. Chen et al. (2022) generalize this to latent rewards with function approximation and establish sublinear regret bounds. Zhan et al. (2023) consider the offline RL problem in the function approximation framework and prove corresponding performance guarantees. In contrast to all these prior work, we consider learning under the privacy of human labelers, where the pairwise comparisons provided by them is considered to be sensitive information. It is worth noting that Cai et al. (2023) also consider privacy protection under the BTL model. Some key differences are as follows: (i) they study the tabular case rather than our linear reward model; (ii) their privacy notion is also different from ours in that it protects all the outcomes of a single item, rather than our label DP notion.

Label Differential Privacy. Privacy of pairwise comparisons can be protected using the notion of Label Differential Privacy. Chaudhuri et al. (2011) introduce this notion for the first time in learning theory and design private PAC learners under label DP. Since then, several follow-up work consider this notion of DP (see, e.g. Beimel et al. (2013); Ghazi et al. (2021); Esfandiari et al. (2022)) where standard DP seems to be an overkill. The most related work to ours is Ghazi et al. (2021), which considers training deepNNs under label DP. Our work differs from theirs as well as from the literature on private stochastic optimization Chaudhuri and Monteleoni (2008); Bassily et al. (2014); Kifer et al. (2012); Bassily et al. (2019); Song et al. (2021) in the utility guarantees (i.e. performance metrics); we consider bounding the error of parameter estimates under some metric, whereas these papers bound generalization errors or population risks.

B ADDITIONAL DETAILS ON SECTION 3

We first introduce some necessary backgrounds and notations for our proofs on lower bounds, i.e., the lower model in this section and the central model in Appendix D.

Let $\mathcal{P}$ be a family of distributions over $\mathcal{X}^n$, where $\mathcal{X}$ is the data universe and n is the sample size. Let $\theta : \mathcal{P} \to \Theta$ be the parameter of the distribution that we aim to estimate and let $\rho : \Theta \times \Theta \to \mathbb{R}_+$ be a pseudo-metric that is the loss function for estimating θ. The minimax risk of estimation under loss ρ for the class $\mathcal{P}$ is

$$R(\mathcal{P},\rho) := \min_{\widehat{\theta}} \max_{P\in\mathcal{P}} \mathbb{E}_{X\sim P}\left[\rho(\widehat{\theta}(X),\theta(P))\right]. \tag{13}$$

For (ε,δ)-label DP in the central model, the minimax risk is

$$R_c(\mathcal{P},\rho,\varepsilon,\delta) := \min_{\widehat{\theta}\text{ is}(\varepsilon,\delta)\text{-label DP}} \max_{P\in\mathcal{P}} \mathbb{E}_{X\sim P}\left[\rho(\widehat{\theta}(X),\theta(P))\right].$$

For (ε,δ)-label DP in the local model, the minimax risk is

$$R_l(\mathcal{P},\rho,\varepsilon,\delta) := \min_{Q\text{ is }(\varepsilon,\delta)\text{-label DP mechanism}} \min_{\widehat{\theta}} \max_{P\in\mathcal{P}} \mathbb{E}_{X\sim P,Q}\left[\rho(\widehat{\theta}(X),\theta(P))\right].$$

for pure-DP where $\delta = 0$, we simply write $R_c(\mathcal{P},\rho,\varepsilon)$ and $R_l(\mathcal{P},\rho,\varepsilon)$.

In this work, under BTL model, our goal is essentially to estimate the unknown parameter θ in the logistic model/distribution. In particular, we consider a fixed design (i.e., $x_i \in \mathbb{R}^d$ is known) and the goal is to infer unknown θ after observing (private) sequence y_i, where the non-private y_i is drawn from

$$\mathbb{P}\left[y_i = 1 | x_i\right] = \sigma(\theta^\top x_i) = \frac{1}{1+\exp(-\theta^\top x_i)} \ , \quad \mathbb{P}\left[y_i = 0 | x_i\right] = 1 - \sigma(\theta^\top x_i).$$

We denote this family of distribution by $\mathcal{P}_{\log}$. More specifically, for the local model, the learner/agent aims to estimate θ from a sequence of private $\widetilde{y}_i$ generated by an LDP mechanism Q, while under the central model, the goal is to output an estimate $\widehat{\theta}$ that is close to θ while guaranteeing label DP. Here, ρ will be either squared ℓ_2 norm or squared semi-norm.

The following result will be useful in our proofs.

Claim B.1. *Let $p_a := 1/(1+e^a)$ and $p_b := 1/(1+e^b)$, we have*

$$\mathrm{kl}(p_a \| p_b) + \mathrm{kl}(p_b \| p_a) \le (a-b)^2,$$

where $\mathrm{kl}(p\|q) := D_{\mathrm{KL}}\left(\mathrm{Bernoulli}(p) \| \mathrm{Bernoulli}(q)\right)$ denotes KL-divergence between Bernoulli distributions with parameters p and q.

Proof. By a direct calculation, we have

$$\mathrm{kl}(p_a \| p_b) + \mathrm{kl}(p_b \| p_a) = (p_a - p_b) \log\left(\frac{p_a}{1-p_a}\frac{1-p_b}{p_b}\right).$$

Further, by the definition of p_a, p_b, we have

$$(p_a - p_b) \log\left(\frac{p_a}{1-p_a}\frac{1-p_b}{p_b}\right) = \left(\frac{1}{1+e^a} - \frac{1}{1+e^b}\right) \cdot (b-a).$$

Without loss of generality, we assume $b \ge a$, then

$$\left(\frac{1}{1+e^a} - \frac{1}{1+e^b}\right) \le \frac{e^b - e^a}{e^b} = 1 - e^{a-b} \le 1 - (1 + a - b) = b - a.$$

Combining the above, yields the result. □

B.1 Proof of Theorem 3.1

Before we state the main proof, let us first present some useful lemmas. In particular, as mentioned in the main paper, for the semi-norm part, we will rely on Fano's lemma to derive the minimax lower bound. The key idea is to construct a proper packing rather than restricting to hypercubes as in Assouad's lemma. Let us first recall the non-private Fano's lemma (Yu, 1997) as follows.

Lemma B.2 (Fano's Lemma)**.** *Let $\mathcal{V} = \{P_1, P_2, \ldots, P_M\} \subseteq \mathcal{P}$ such that for all $i \neq j$,*

$$D_{\mathrm{KL}}\left(P_i \| P_j\right) \le \beta \ , \quad \rho'(\theta(P_i), \theta(P_j)) \ge \tau$$

for a semi-metric ρ' and some $\tau, \beta > 0$. Then, we have

$$R(\mathcal{P}, (\rho')^2) \ge \frac{\tau^2}{4}\left(1 - \frac{\beta + \log 2}{\log M}\right).$$

To construct a proper packing, Varshamov–Gilbert's bound (cf. Guntuboyina (2011)) will be useful.

Lemma B.3 (Varshamov–Gilbert's bound)**.** *For any $\xi \in (0, 1/2)$ and for every dimension $d \ge 1$, there exist $M \ge e^{\frac{\xi^2 d}{2}}$ and $w_1, \ldots, w_M \in \{0,1\}^d$ such that*

$$d_{\mathrm{ham}}(w_i, w_j) \ge (1/2 - \xi)d, \quad \forall i \neq j.$$

Now, we are ready to prove Theorem 3.1.

Proof of Theorem 3.1. We divide it into non-private and private parts.

Non-private part. Let $\xi = 1/4$ in Lemma B.3, then there exist $M \geq e^{\frac{\xi^2 d}{2}}$ and $w_1, \ldots, w_M \in \{0,1\}^d$ such that

$$\forall i \neq j, \quad \frac{d}{4} \leq \|w_i - w_j\|_2^2 \leq d.$$

Now, let the eigenvalue decomposition of $\Sigma_{\mathcal{D}} + \lambda I$ be $U^\top \Lambda U$ and $\theta_i := \frac{\Delta}{\sqrt{d}} U^\top \sqrt{\Lambda^{-1}} w_i$, then we have

$$\begin{aligned}\|\theta_i - \theta_j\|_{\Sigma_{\mathcal{D}}+\lambda I}^2 &= (\theta_i - \theta_j)^\top (\Sigma_{\mathcal{D}} + \lambda I)(\theta_i - \theta_j)\\ &= \frac{\Delta^2}{d}(w_i - w_j)^\top \sqrt{\Lambda^{-1}} U (U^\top \Lambda U) U^\top \sqrt{\Lambda^{-1}} (w_i - w_j)\\ &= \frac{\Delta^2}{d}\|w_i - w_j\|_2^2 .\end{aligned}$$

Thus, we have constructed a packing such that for any $i, j \in [M]$ and $i \neq j$,

$$\Delta \geq \|\theta_i - \theta_j\|_{\Sigma_{\mathcal{D}}+\lambda I} \geq \frac{\Delta}{2}.$$

Now, let us turn to the KL divergence part. Let P_i^n be the product distribution when $\theta^* = \theta_i$. Then, by chain rule of KL divergence and Claim B.1, we have

$$D_{\text{KL}}\left(P_i^n \| P_j^n\right) \leq \sum_{k=1}^n (x_k^\top(\theta_i - \theta_j))^2 = n(\theta_i - \theta_j)^\top \Sigma_{\mathcal{D}} (\theta_i - \theta_j) \leq n\|\theta_i - \theta_j\|_{\Sigma_{\mathcal{D}}+\lambda I}^2 \leq n\Delta^2. \tag{14}$$

Thus, by Fano's lemma, we have

$$R(\mathcal{P}_{\log}, \|\cdot\|_{\Sigma_{\mathcal{D}}+\lambda I}^2) \geq \frac{\Delta^2}{8}\left(1 - 32 \cdot \frac{n\Delta^2 + \log 2}{d}\right).$$

Thus, choosing $\Delta^2 = c\frac{d}{n}$ for some constant c, we have for large d,

$$R(\mathcal{P}_{\log}, \|\cdot\|_{\Sigma_{\mathcal{D}}+\lambda I}^2) \geq \Omega\left(\frac{d}{n}\right).$$

Finally, we also need to check that $\|\theta_i\| \leq B$ when n is large. To this end, by the fact that $w_i \in \{0,1\}^d$ for any i, we have that

$$\|\theta_i\| \leq \frac{\Delta}{\sqrt{d}}\sqrt{\operatorname{tr}(\Lambda^{-1})} = \frac{\Delta}{\sqrt{d}}\sqrt{\operatorname{tr}((\Sigma_{\mathcal{D}} + \lambda I)^{-1})} \leq B,$$

where the last step holds when $n \geq \frac{c\,\operatorname{tr}((\Sigma_{\mathcal{D}}+\lambda I)^{-1})}{B^2}$, since $\Delta^2 = c\frac{d}{n}$. We also note that the centered condition $\langle 1, \theta\rangle = 0$ can be simply achieved by reducing d to $d/2$.

Private part. Let M_i^n be the product distribution of private view when $\theta^* = \theta_i$. The only change here is the KL divergence part. By Corollary 3 in Duchi et al. (2018), Pinsker's inequality and Claim B.1, we can obtain that

$$D_{\text{KL}}\left(M_i^n \| M_j^n\right) \leq n(e^\varepsilon - 1)^2 (\theta_i - \theta_j)^\top \Sigma_{\mathcal{D}} (\theta_i - \theta_j) \leq n(e^\varepsilon - 1)^2 \Delta^2.$$

Thus, a similar analysis as the non-private case gives

$$R(\mathcal{P}_{\log}, \|\cdot\|_{\Sigma_{\mathcal{D}}+\lambda I}^2, \varepsilon) \geq \Omega\left(\frac{d}{n(e^\varepsilon - 1)^2}\right).$$

□

B.2 Proof of Theorem 3.2

Before we present the proof, we first introduce some useful lemmas. A convenient way to establish a lower bound in ℓ_2 norm is via Assouad's lemma. We restate it below with proof for completeness and some additional implications, which are useful for our proof.

Lemma B.4 (Assouad's lemma)**.** *Let $\mathcal{V} \subseteq \mathcal{P}$ be a set of distributions indexed by the hypercube $\mathcal{E}_d = \{\pm 1\}^d$. Suppose there exists a $\tau \in \mathbb{R}$ and $\alpha > 0$, such that ρ satisfies: (i) for all $u, v, w \in \mathcal{E}_d$, $\rho(\theta(P_u), \theta(P_v)) \geq 2\tau \cdot \sum_{i=1}^d \mathbb{1}(u_i \neq v_i)$ and (ii) $\rho(\theta(P_u), \theta(P_v)) \leq \alpha(\rho(\theta(P_u), \theta(P_w)) + \rho(\theta(P_v), \theta(P_w)))$, i.e., α-triangle inequality. For each $i \in [d]$, define the mixture distributions:*

$$P_{+i} := \frac{2}{|\mathcal{E}_d|} \sum_{e \in \mathcal{E}_d : e_i = 1} P_e \text{ and } P_{+i} := \frac{2}{|\mathcal{E}_d|} \sum_{e \in \mathcal{E}_d : e_i = -1} P_e.$$

Then, we have

$$R(\mathcal{P}, \rho) \geq \frac{\tau}{2\alpha} \sum_{i=1}^d \left(1 - \|P_{+i} - P_{-i}\|_{\mathrm{TV}}\right).$$

Proof. Let $P \in \mathcal{V} \subseteq \mathcal{P}$ and $X \sim P$. For any estimator $\widehat{\theta}(X)$, define $\psi^* = \operatorname{argmin}_{e \in \mathcal{E}_d} \rho(\widehat{\theta}, \theta(P_e))$. Thus, we have

$$\rho(\theta(P), \theta(P_{\psi^*})) \overset{(a)}{\leq} \alpha\left(\rho(\theta(P), \widehat{\theta}) + \rho(\theta(P_{\psi^*}), \widehat{\theta})\right) \overset{(b)}{\leq} 2\alpha \cdot \rho(\theta(P), \widehat{\theta}),$$

where (a) holds by the α-triangle inequality of ρ; (b) holds by the definition of ψ^*. As a result, we have

$$R(\mathcal{P}, \rho) \geq R(\mathcal{V}, \rho) = \min_{\widehat{\theta}} \max_{P \in \mathcal{V}} \mathbb{E}_{X \sim P}\left[\rho(\widehat{\theta}(X), \theta(P))\right] \geq \frac{1}{2\alpha} \min_{\widehat{\theta}} \max_{P \in \mathcal{V}} \mathbb{E}_{X \sim P}\left[\rho(\theta(P), \theta(P_{\psi^*}))\right].$$

Now, by condition (i) of the loss ρ, we have

$$\begin{aligned}
\max_{P \in \mathcal{V}} \mathbb{E}_{X \sim P}\left[\rho(\theta(P), \theta(P_{\psi^*}))\right] &\overset{(a)}{\geq} \frac{1}{|\mathcal{E}_d|} \sum_{e \in \mathcal{E}_d} \mathbb{E}_{X \sim P_e}\left[\rho(\theta(P_e), \theta(P_{\psi^*}))\right] \\
&\overset{(b)}{\geq} \frac{2\tau}{|\mathcal{E}_d|} \sum_{e \in \mathcal{E}_d} \sum_{i=1}^d \mathbb{E}_{X \sim P_e}\left[\mathbb{1}(\psi_i^* \neq e_i)\right] \\
&\overset{(c)}{=} \frac{2\tau}{|\mathcal{E}_d|} \sum_{e \in \mathcal{E}_d} \sum_{i=1}^d \mathbb{P}_e\left[\psi_i^* \neq e_i\right] \\
&\overset{(d)}{=} \frac{2\tau}{|\mathcal{E}_d|} \sum_{i=1}^d \sum_{e \in \mathcal{E}_d} \mathbb{P}_e\left[\psi_i^* \neq e_i\right],
\end{aligned}$$

where (a) holds by maximum is larger than average; (b) holds by condition (i) of ρ; in (c), $\mathbb{P}_e$ is the probability measure when samples are generated from P_e; (d) follows by swapping the two sums. For each $i \in d$, we divide the set $\mathcal{E}_d$ into two parts based on the value at i, i.e.,

$$\begin{aligned}
\frac{2\tau}{|\mathcal{E}_d|} \sum_{e \in \mathcal{E}_d} \mathbb{P}_e\left[\psi_i^* \neq e_i\right] &= \frac{2\tau}{|\mathcal{E}_d|} \sum_{e \in \mathcal{E}_d : e_i = +1} \mathbb{P}_e\left[\psi_i^* \neq e_i\right] + \frac{2\tau}{|\mathcal{E}_d|} \sum_{e \in \mathcal{E}_d : e_i = -1} \mathbb{P}_e\left[\psi_i^* \neq e_i\right] \\
&= \tau \cdot \left(\mathbb{P}_{X \sim P_{+i}}\left[\psi_i^*(X) \neq 1\right] + \mathbb{P}_{X \sim P_{-i}}\left[\psi_i^*(X) \neq -1\right]\right).
\end{aligned}$$

Combining the above together, we have

$$\begin{aligned}
R(P, \rho) &\geq \frac{\tau}{2\alpha} \sum_{i=1}^d \left(\mathbb{P}_{X \sim P_{+i}}\left[\psi_i^*(X) \neq 1\right] + \mathbb{P}_{X \sim P_{-i}}\left[\psi_i^*(X) \neq -1\right]\right) \\
&\geq \frac{\tau}{2\alpha} \sum_{i=1}^d \inf_{\Psi} \left(\mathbb{P}_{X \sim P_{+i}}\left[\Psi(X) \neq 1\right] + \mathbb{P}_{X \sim P_{-i}}\left[\Psi(X) \neq -1\right]\right) \\
&\overset{(a)}{=} \frac{\tau}{2\alpha} \sum_{i=1}^d \left(1 - \|P_{+i} - P_{-i}\|_{\mathrm{TV}}\right),
\end{aligned}$$

where (a) holds by Le Cam's first lemma. □

Corollary B.5. *Under the same conditions of Lemma B.4, we have*

$$R(\mathcal{P},\rho) \geq \frac{d\tau}{2\alpha}\left[1-\left(\frac{1}{d}\sum_{i=1}^{d}\frac{1}{2^d}\sum_{e\in\mathcal{E}_d}\|P_e - P_{\bar{e}^i}\|_{\mathrm{TV}}^2\right)^{1/2}\right],$$

where $\bar{e}^i$ is a vector in $\mathcal{E}_d$ that flips the i-th coordinate of e.

Proof. To start with, we introduce the following additional notations. For any $e \in \mathcal{E}_d$, let $P_{e,+i}$ be the distribution indexed by first choosing e and then letting $e_i = +1$. Similarly, we have $P_{e,-i}$. By this definition, we can rewrite P_{+i} and P_{-i} above as follows:

$$P_{+i} = \frac{1}{|\mathcal{E}_d|}\sum_{e\in\mathcal{E}_d} P_{e,+i} \text{ and } P_{-i} = \frac{1}{|\mathcal{E}_d|}\sum_{e\in\mathcal{E}_d} P_{e,-i}\,. \tag{15}$$

By Lemma B.4, we have

$$R(\mathcal{P},\rho) \geq \frac{\tau}{2\alpha}\sum_{i=1}^{d}\left(1 - \|P_{+i} - P_{-i}\|_{\mathrm{TV}}\right). \tag{16}$$

Now note that

$$\begin{aligned}
\sum_{i=1}^{d}\|P_{+i} - P_{-i}\|_{\mathrm{TV}} &\overset{(a)}{\leq} \sqrt{d}\left(\sum_{i=1}^{d}\|P_{+i} - P_{-i}\|_{\mathrm{TV}}^2\right)^{1/2}\\
&\overset{(b)}{\leq} \sqrt{d}\left(\sum_{i=1}^{d}\frac{1}{|\mathcal{E}_d|}\sum_{e\in\mathcal{E}_d}\|P_{e,+i} - P_{e,-i}\|_{\mathrm{TV}}^2\right)^{1/2}\\
&= \sqrt{d}\left(\sum_{i=1}^{d}\frac{1}{2^d}\sum_{e\in\mathcal{E}_d}\|P_{e,+i} - P_{e,-i}\|_{\mathrm{TV}}^2\right)^{1/2}.
\end{aligned}$$

where (a) holds by Cauchy-Schwarz inequality; (b) holds by (15) and joint convexity of $\|\cdot\|_{\mathrm{TV}}^2$. Plugging it back to (16) and rearranging, we have

$$R(\mathcal{P},\rho) \geq \frac{\tau d}{2\alpha}\left[1-\left(\frac{1}{d}\sum_{i=1}^{d}\frac{1}{2^d}\sum_{e\in\mathcal{E}_d}\|P_{e,+i} - P_{e,-i}\|_{\mathrm{TV}}^2\right)^{1/2}\right],$$

which finishes the first part. The final result in the corollary simply follows from that TV distance is symmetric. □

Now, we are ready to prove Theorem 3.2.

Proof of Theorem 3.2. We also divide it into two parts: the non-private part and the private part.

Non-private part. Choose some $\Delta > 0$ and for each $e \in \mathcal{E}_d = \{\pm 1\}^d$, let $\theta_e = \Delta e$. Now we need to check the two conditions in Lemma B.4. First note that $\rho = \|\cdot\|_2^2$ satisfies 2-triangle inequality, i.e., $\alpha = 2$. Also, note that for any $u, v \in \mathcal{E}_d$, $\|\theta_u - \theta_v\|_2^2 = 4\Delta^2\sum_{i=1}^{d}\mathbb{1}(u_i \neq v_i)$, i.e., $\tau = 2\Delta^2$. Thus, let P_e^n be the distribution for the n independent samples of (non-private) y_i when $\theta = \theta_e$ and then by Corollary B.5, we have

$$R_l(\mathcal{P}_{\log}, \|\cdot\|_2^2, \varepsilon, \delta) \geq R(\mathcal{P}_{\log}, \|\cdot\|_2^2) \geq \frac{d\Delta^2}{2}\left[1-\left(\frac{1}{d}\sum_{i=1}^{d}\frac{1}{2^d}\sum_{e\in\mathcal{E}_d}\|P_e^n - P_{\bar{e}^i}^n\|_{\mathrm{TV}}^2\right)^{1/2}\right].$$

Thus, it remains to bound the part of TV distance. By Pinsker's inequality and chain rule of KL-divergence, we have for any $u, v \in \mathcal{E}_d$

$$\begin{aligned}\left\|P_u^n - P_v^n\right\|_{\mathrm{TV}}^2 &\le \frac{1}{4}\left(D_{\mathrm{KL}}\left(P_u^n \| P_v^n\right) + D_{\mathrm{KL}}\left(P_v^n \| P_u^n\right)\right)\\ &= \frac{1}{4}\sum_{k=1}^{n}\left(\mathrm{kl}(p_u(x_k)\|p_v(x_k)) + \mathrm{kl}(p_v(x_k)\|p_u(x_k))\right),\end{aligned}$$

Then, by Claim B.1, we can bound the TV-distance term as follows.

$$\left\|P_u^n - P_v^n\right\|_{\mathrm{TV}}^2 \le \frac{\Delta^2}{4}\sum_{k=1}^{n}\left(x_k^\top(u-v)\right)^2.$$

This directly implies that

$$\frac{1}{d2^d}\sum_{i=1}^{d}\sum_{e\in\mathcal{E}_d}\left\|P_e^n - P_{\bar{e}^i}^n\right\|_{\mathrm{TV}}^2 \le \frac{\Delta^2}{4d}\frac{1}{2^d}\sum_{e\in\mathcal{E}_d}\sum_{i=1}^{d}\sum_{k=1}^{n}(2x_{ki})^2 = \frac{\Delta^2}{d}\frac{1}{2^d}\sum_{e\in\mathcal{E}_d}\sum_{i=1}^{d}\sum_{k=1}^{n}x_{ki}^2 = \frac{\Delta^2}{d}\frac{1}{2^d}\sum_{e\in\mathcal{E}_d}\left\|X\right\|_{\mathrm{F}}^2,$$

where $X \in \mathbb{R}^{n\times d}$ and $x_k^\top \in \mathbb{R}^d$ is the k-th row and $\left\|\cdot\right\|_{\mathrm{F}}$ is the Frobenius norm. Hence, we obtain that

$$R_l(\mathcal{P}_{\log}, \left\|\cdot\right\|_2^2, \varepsilon, \delta) \ge \frac{d\Delta^2}{2}\left[1 - \left(\frac{\Delta^2}{d}\left\|X\right\|_{\mathrm{F}}^2\right)^{1/2}\right].$$

Finally, choosing $\Delta^2 = \frac{d}{4\|X\|_{\mathrm{F}}^2}$, we have

$$R_l(\mathcal{P}_{\log}, \left\|\cdot\right\|_2^2, \varepsilon, \delta) \ge \frac{d^2}{16\left\|X\right\|_{\mathrm{F}}^2} = \frac{d}{n}\cdot\frac{1}{16\frac{1}{dn}\sum_{k=1}^{n}\left\|x_k\right\|^2}.$$

Since $\left\|x_k\right\|^2 \le L^2$, one can further simplify it as

$$R_l(\mathcal{P}_{\log}, \left\|\cdot\right\|_2^2, \varepsilon, \delta) \ge \Omega\left(\frac{d}{L^2}\cdot\frac{d}{n}\right).$$

Finally, note that one can indeed easily check that for large enough n, $\|\theta\| \le B$ and also $\langle 1, \theta\rangle = 0$ by halving the dimension d.

Private part. Now, let us turn to the private part. In particular, let M_e^n be the distribution for the n independent samples of private view $\widetilde{y}_i$ when $\theta = \theta_e$ and then by Corollary B.5, we have

$$R_l(\mathcal{P}_{\log}, \left\|\cdot\right\|_2^2, \varepsilon, \delta) \ge \frac{d\Delta^2}{2}\left[1 - \left(\frac{1}{d}\sum_{i=1}^{d}\frac{1}{2^d}\sum_{e\in\mathcal{E}_d}\left\|M_e^n - M_{\bar{e}^i}^n\right\|_{\mathrm{TV}}^2\right)^{1/2}\right].$$

Again, the key is to bound the TV-distance term. To this end, by Corollary 3 in Duchi et al. (2018) and Pinsker's inequality, we have

$$\left\|M_u^n - M_v^n\right\|_{\mathrm{TV}}^2 \le \frac{(e^\varepsilon - 1)^2}{2}\sum_{k=1}^{n}\left(\mathrm{kl}(p_u(x_k)\|p_v(x_k)) + \mathrm{kl}(p_v(x_k)\|p_u(x_k))\right).$$

Then, following the same analysis as the non-private case, we can obtain that

$$R_l(\mathcal{P}_{\log}, \left\|\cdot\right\|_2^2, \varepsilon, \delta) \ge \frac{d\Delta^2}{2}\left[1 - \left(\frac{2(e^\varepsilon - 1)^2\Delta^2}{d}\left\|X\right\|_{\mathrm{F}}^2\right)^{1/2}\right].$$

Finally, choosing $\Delta^2 = \frac{d}{8(e^\varepsilon - 1)^2\|X\|_{\mathrm{F}}^2}$, we have

$$R_l(\mathcal{P}_{\log}, \left\|\cdot\right\|_2^2, \varepsilon, \delta) \ge \frac{d^2}{32(e^\varepsilon - 1)^2\left\|X\right\|_{\mathrm{F}}^2} = \frac{d}{n(e^\varepsilon - 1)^2}\cdot\frac{1}{32\frac{1}{dn}\sum_{k=1}^{n}\left\|x_k\right\|^2}.$$

Again, noting that $\|x_k\|^2 \leq L^2$, one can simplify it as

$$R_l(\mathcal{P}_{\log}, \|\cdot\|_2^2, \varepsilon, \delta) \geq \Omega\left(\frac{d}{L^2} \cdot \frac{d}{n(e^\varepsilon - 1)^2}\right).$$

□

C ADDITIONAL DETAILS ON SECTION 4

We are given a query-observation dataset $\mathcal{D} = (s_i, a_i^0, a_i^1, y_i)_{i=1}^n$. Define $x_i = \phi(s_i, a_i^1) - \phi(s_i, a_i^0)$. Under the BTL model, the labels $y_i \in \{0, 1\}$ be drawn from the distribution

$$p_{i,1} = \mathbb{P}\left[y_i = 1|x_i\right] = \sigma(\theta^\top x_i) = \frac{1}{1 + \exp(-\theta^\top x_i)}\ , \quad p_{i,0} = \mathbb{P}\left[y_i = 0|x_i\right] = 1 - \sigma(\theta^\top x_i)\ .$$

When queried the value of y_i, the RR mechanism outputs $\widetilde{y}_i$ with probability

$$\mathbb{P}\left[\widetilde{y}_i = y_i\right] = \sigma(\varepsilon) = \frac{1}{1 + \exp(-\varepsilon)}\ , \quad \mathbb{P}\left[\widetilde{y}_i \neq y_i\right] = 1 - \sigma(\varepsilon)\ .$$

Then, for any $\theta \in \mathbb{R}^d$, the predicted probabilities of a randomized label $\widetilde{y}_i$ given x_i are

$$\begin{aligned}
\widetilde{p}_{i,1} &= \mathbb{P}\left[\widetilde{y}_i = 1|x_i\right] = \sigma(\theta^\top x_i)\sigma(\varepsilon) + (1 - \sigma(\theta^\top x_i))(1 - \sigma(\varepsilon))\ , \\
\widetilde{p}_{i,0} &= \mathbb{P}\left[\widetilde{y}_i = 0|x_i\right] = (1 - \sigma(\theta^\top x_i))\sigma(\varepsilon) + \sigma(\theta^\top x_i)(1 - \sigma(\varepsilon))\ .
\end{aligned}$$

C.1 Proof of Theorem 4.1

First, the predicted probabilities for any $\theta \in \mathbb{R}^d$ can be computed as

$$\begin{aligned}
\mathbb{P}\left[\widetilde{y}_i = 1|x_i\right] &= \frac{1}{1 + \exp(-\langle\theta, x_i\rangle)} \cdot \frac{\exp(\varepsilon)}{1 + \exp(\varepsilon)} + \frac{\exp(-\langle\theta, x_i\rangle)}{1 + \exp(-\langle\theta, x_i\rangle)} \cdot \frac{1}{1 + \exp(\varepsilon)} = \frac{1 + e^{-\varepsilon}e^{-\theta^\top x_i}}{(1 + e^{-\theta^\top x_i})(1 + e^{-\varepsilon})} \\
\mathbb{P}\left[\widetilde{y}_i = 0|x_i\right] &= \frac{\exp(-\langle\theta, x_i\rangle)}{1 + \exp(-\langle\theta, x_i\rangle)} \cdot \frac{\exp(\varepsilon)}{1 + \exp(\varepsilon)} + \frac{1}{1 + \exp(-\langle\theta, x_i\rangle)} \cdot \frac{1}{1 + \exp(\varepsilon)} = \frac{e^{-\varepsilon} + e^{-\theta^\top x_i}}{(1 + e^{-\theta^\top x_i})(1 + e^{-\varepsilon})}.
\end{aligned}$$

Based on this, the negative log-likelihood in (3) takes the form

$$\widetilde{l}_{\mathcal{D},\varepsilon}(\theta) = -\frac{1}{n}\sum_{i=1}^n \left[\mathbb{1}(\widetilde{y}_i = 1)\log\frac{1 + e^{-\varepsilon}e^{-\theta^\top x_i}}{(1 + e^{-\theta^\top x_i})(1 + e^{-\varepsilon})} + \mathbb{1}(\widetilde{y}_i = 0)\log\frac{e^{-\varepsilon} + e^{-\theta^\top x_i}}{(1 + e^{-\theta^\top x_i})(1 + e^{-\varepsilon})}\right]\ .$$

Now the gradient of negative log-likelihood is given by $\nabla l_{\mathcal{D},\varepsilon}(\theta) = -\frac{1}{n}\sum_{i=1}^n V_{\theta,i}x_i = -\frac{1}{n}X^\top V_\theta$, where

$$V_{\theta,i} = \mathbb{1}(\widetilde{y}_i = 1)\left(\frac{e^{-\theta^\top x_i}}{1 + e^{-\theta^\top x_i}} - \frac{e^{-\varepsilon}e^{-\theta^\top x_i}}{1 + e^{-\varepsilon}e^{-\theta^\top x_i}}\right) + \mathbb{1}(\widetilde{y}_i = 0)\left(\frac{e^{-\theta^\top x_i}}{1 + e^{-\theta^\top x_i}} - \frac{e^{-\theta^\top x_i}}{e^{-\varepsilon} + e^{-\theta^\top x_i}}\right)\ .$$

It holds that

$$\begin{aligned}
\mathbb{E}_\theta[V_{\theta,i}|x_i] &= \frac{e^{-\theta^\top x_i}}{1 + e^{-\theta^\top x_i}} - \left(\frac{e^{-\varepsilon}e^{-\theta^\top x_i}}{1 + e^{-\varepsilon}e^{-\theta^\top x_i}} \cdot \frac{1 + e^{-\varepsilon}e^{-\theta^\top x_i}}{(1 + e^{-\theta^\top x_i})(1 + e^{-\varepsilon})} + \frac{e^{-\theta^\top x_i}}{e^{-\varepsilon} + e^{-\theta^\top x_i}} \cdot \frac{e^{-\varepsilon} + e^{-\theta^\top x_i}}{(1 + e^{-\theta^\top x_i})(1 + e^{-\varepsilon})}\right) \\
&= \frac{e^{-\theta^\top x_i}}{1 + e^{-\theta^\top x_i}} - \frac{e^{-\theta^\top x_i}}{1 + e^{-\theta^\top x_i}} = 0
\end{aligned}$$

Now, under Assumption 2.1, we have $-c \leqslant \theta^\top x_i \leqslant c$, where $c = O(LB)$. Hence, we have

$$\begin{aligned}
V_{\theta,i}|(\widetilde{y}_i = 1) &= \frac{e^{-\theta^\top x_i}(e^\varepsilon - 1)}{(1 + e^{-\theta^\top x_i})(e^\varepsilon + e^{-\theta^\top x_i})} \leq \frac{e^\varepsilon - 1}{(e^\varepsilon + e^{-c})} = \frac{e^c(e^\varepsilon - 1)}{(e^\varepsilon e^c + 1)}\ , \\
V_{\theta,i}|(\widetilde{y}_i = 0) &= \frac{e^{-\theta^\top x_i}(e^\varepsilon - 1)}{(1 + e^{-\theta^\top x_i})(1 + e^\varepsilon e^{-\theta^\top x_i})} \leq \frac{(e^\varepsilon - 1)}{(e^\varepsilon + e^{-c})} = \frac{e^c(e^\varepsilon - 1)}{(e^\varepsilon e^c + 1)}\ .
\end{aligned}$$

Therefore, it holds that $V_{\theta,i}$ is zero-mean and $v = \frac{e^c(e^\varepsilon-1)}{(e^\varepsilon e^c+1)}$-sub-Gaussian under the conditional distribution $\mathbb{P}_\theta[\cdot|x_i]$ and under Assumption 2.1.

Now the Hessian of negative log-likelihood is given by $\nabla^2 l_{\mathcal{D},\varepsilon}(\theta) = \frac{1}{n}\sum_{i=1}^n \left[\mathbb{1}(\widetilde{y}_i = 1)\alpha_{1,i} + \mathbb{1}(\widetilde{y}_i = 0)\alpha_{0,i}\right] x_i x_i^\top$, where

$$\alpha_{1,i} = \frac{e^{-\theta^\top x_i}}{(1+e^{-\theta^\top x_i})^2} - \frac{e^{-\varepsilon}e^{-\theta^\top x_i}}{(1+e^{-\varepsilon}e^{-\theta^\top x_i})^2} = \frac{e^{-\theta^\top x_i}}{(1+e^{-\theta^\top x_i})^2}\cdot\frac{(e^\varepsilon-1)(e^\varepsilon e^{2\theta^\top x_i}-1)}{(1+e^\varepsilon e^{\theta^\top x_i})^2}\ ,$$
$$\alpha_{0,i} = \frac{e^{-\theta^\top x_i}}{(1+e^{-\theta^\top x_i})^2} - \frac{e^{-\theta^\top x_i}}{(e^{-\varepsilon}+e^{-\theta^\top x_i})^2} = \frac{e^{-\theta^\top x_i}}{(1+e^{-\theta^\top x_i})^2}\cdot\frac{(e^\varepsilon-1)(e^\varepsilon e^{-2\theta^\top x_i}-1)}{(1+e^\varepsilon e^{-\theta^\top x_i})^2}\ .$$

Under Assumption 2.1, both $\alpha_{1,i}, \alpha_{0,i} \geq \widetilde{\gamma}_\varepsilon$ for all $\theta \in \Theta_B$, where

$$\widetilde{\gamma}_\varepsilon = \gamma\frac{(e^\varepsilon-1)(e^\varepsilon e^{-2c}-1)}{(e^\varepsilon e^c+1)^2}\ .$$

Now $\widetilde{\gamma}_\varepsilon > 0$ only when $\varepsilon > 2c$. This implies that $\widetilde{l}_{\mathcal{D},\varepsilon}$ is $\widetilde{\gamma}_\varepsilon$ strongly convex in θ_B when $\varepsilon > 2c$ with respect to the semi-norm $\|\cdot\|_{\Sigma_\mathcal{D}}$. Since $\theta^* \in \Theta_B$, introducing the error vector $\Delta = \widetilde{\theta}_{\text{MLE}} - \theta^*$, we conclude that

$$\widetilde{\gamma}_\varepsilon \|\Delta\|^2_{\Sigma_\mathcal{D}} \leqslant \left\|\nabla\widetilde{l}_{\mathcal{D},\varepsilon}(\theta^*)\right\|_{(\Sigma_\mathcal{D}+\lambda I)^{-1}}\|\Delta\|_{(\Sigma_\mathcal{D}+\lambda I)}$$

for some $\lambda > 0$. Introducing $M = \frac{1}{n^2}X(\Sigma_\mathcal{D}+\lambda I)^{-1}X^\top$, we have $\left\|\nabla\widetilde{l}_{\mathcal{D},\varepsilon}(\theta^*)\right\|^2_{(\Sigma_\mathcal{D}+\lambda I)^{-1}} = V_{\theta^*}^\top M V_{\theta^*}$. Then, the Bernstein's inequality for sub-Gaussian random variables in quadratic form (see e.g. Hsu et al. (2012, Theorem 2.1)) implies that with probability at least $1-\alpha$,

$$\begin{aligned}\left\|\nabla\widetilde{l}_{\mathcal{D},\varepsilon}(\theta^*)\right\|^2_{(\Sigma_\mathcal{D}+\lambda I)^{-1}} = V_{\theta^*}^\top M V_{\theta^*} &\leqslant v^2\left(\operatorname{tr}(M) + 2\sqrt{\operatorname{tr}(M^\top M)\log(1/\alpha)} + 2\|M\|\log(1/\delta)\right)\\ &\leqslant C_1\cdot v^2\cdot\frac{d+\log(1/\alpha)}{n}\end{aligned}$$

for some constant $C_1 > 0$. This gives us

$$\begin{aligned}\widetilde{\gamma}_\varepsilon\|\Delta\|^2_{\Sigma_\mathcal{D}+\lambda I} &\leqslant \left\|\nabla\widetilde{l}_{\mathcal{D},\varepsilon}(\theta^*)\right\|_{(\Sigma_\mathcal{D}+\lambda I)^{-1}}\|\Delta\|_{(\Sigma_\mathcal{D}+\lambda I)} + 4\lambda\widetilde{\gamma}_\varepsilon B^2\\ &\leqslant \sqrt{C_1\cdot v^2\cdot\frac{d+\log(1/\alpha)}{n}}\|\Delta\|_{(\Sigma_\mathcal{D}+\lambda I)} + 4\lambda\widetilde{\gamma}_\varepsilon B^2\ .\end{aligned}$$

Solving for the above inequality, we get

$$\|\Delta\|_{(\Sigma_\mathcal{D}+\lambda I)} \leqslant C_2\cdot\sqrt{\frac{v^2}{\widetilde{\gamma}_\varepsilon^2}\cdot\frac{d+\log(1/\alpha)}{n} + \lambda B^2}$$

for some constant $C_2 > 0$. Now note that $\frac{v}{\widetilde{\gamma}_\varepsilon} = \frac{e^c(e^\varepsilon e^c+1)}{\gamma(e^\varepsilon e^{-2c}-1)} \leq C_3\cdot\frac{(e^\varepsilon e^{2c}+1)}{\gamma(e^\varepsilon e^{-2c}-1)}$ for some constant $C_3 > 0$. Substituting this, we get

$$\left\|\widetilde{\theta}_{\text{MLE}} - \theta^*\right\|_{(\Sigma_\mathcal{D}+\lambda I)} \leqslant C\cdot\frac{(e^\varepsilon e^{2c}+1)}{\gamma(e^\varepsilon e^{-2c}-1)}\sqrt{\frac{d+\log(1/\alpha)}{n}} + C'\cdot\sqrt{\lambda}B$$

for some constants $C, C' > 0$, which holds for any $\varepsilon > 2c$. Setting $c = O(LB)$ completes the proof.

C.2 Proof of Theorem 4.2

First recall from (4) our de-biased loss function

$$\begin{aligned}\widehat{l}_{\mathcal{D},\varepsilon}(\theta) = -\frac{1}{n}\sum_{i=1}^n\Bigg[&\mathbb{1}(\widetilde{y}_i = 1)\left(\sigma(\varepsilon)\log\sigma(\theta^\top x_i) - (1-\sigma(\varepsilon))\log(1-\sigma(\theta^\top x_i))\right)\\ &+ \mathbb{1}(\widetilde{y}_i = 0)\left(\sigma(\varepsilon)\log(1-\sigma(\theta^\top x_i)) - (1-\sigma(\varepsilon))\log\sigma(\theta^\top x_i)\right)\Bigg]\ .\end{aligned}$$

The gradient of the loss function is given by $\nabla \widehat{l}_{\mathcal{D},\varepsilon}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} V_{\theta,i} x_i = -\frac{1}{n} X^\top V_\theta$, where

$$V_{\theta,i} = \mathbb{1}(\widetilde{y}_i = 1)\left(\frac{\sigma'(\theta^\top x_i)}{\sigma(\theta^\top x_i)}\sigma(\varepsilon) + \frac{\sigma'(\theta^\top x_i)}{1-\sigma(\theta^\top x_i)}(1-\sigma(\varepsilon))\right) - \mathbb{1}(\widetilde{y}_i = 0)\left(\frac{\sigma'(\theta^\top x_i)}{1-\sigma(\theta^\top x_i)}\sigma(\varepsilon) + \frac{\sigma'(\theta^\top x_i)}{\sigma(\theta^\top x_i)}(1-\sigma(\varepsilon))\right) .$$

It holds that

$$\begin{aligned}
\mathbb{E}_\theta[V_{\theta,i}|x_i] &= \big(\sigma(\theta^\top x_i)\sigma(\varepsilon) + (1-\sigma(\theta^\top x_i))(1-\sigma(\varepsilon))\big)\left(\frac{\sigma'(\theta^\top x_i)}{\sigma(\theta^\top x_i)}\sigma(\varepsilon) + \frac{\sigma'(\theta^\top x_i)}{1-\sigma(\theta^\top x_i)}(1-\sigma(\varepsilon))\right) \\
&\quad - \big((1-\sigma(\theta^\top x_i))\sigma(\varepsilon) + \sigma(\theta^\top x_i)(1-\sigma(\varepsilon))\big)\left(\frac{\sigma'(\theta^\top x_i)}{1-\sigma(\theta^\top x_i)}\sigma(\varepsilon) + \frac{\sigma'(\theta^\top x_i)}{\sigma(\theta^\top x_i)}(1-\sigma(\varepsilon))\right) \\
&= 0 .
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
|V_{\theta,i}|_{\widetilde{y}_i=1} &= \frac{\sigma'(\theta^\top x_i)}{\sigma(\theta^\top x_i)}\sigma(\varepsilon) + \frac{\sigma'(\theta^\top x_i)}{1-\sigma(\theta^\top x_i)}(1-\sigma(\varepsilon)) , \\
|V_{\theta,i}|_{\widetilde{y}_i=0} &= \frac{\sigma'(\theta^\top x_i)}{1-\sigma(\theta^\top x_i)}\sigma(\varepsilon) + \frac{\sigma'(\theta^\top x_i)}{\sigma(\theta^\top x_i)}(1-\sigma(\varepsilon)) .
\end{aligned}$$

The first derivative of the logistic function $\sigma(\cdot)$ is given by $\sigma'(z) = \sigma(z)(1-\sigma(z))$, which gives us

$$\begin{aligned}
|V_{\theta,i}|_{\widetilde{y}_i=1} &= (1-\sigma(\theta^\top x_i))\sigma(\varepsilon) + \sigma(\theta^\top x_i)(1-\sigma(\varepsilon)) = \mathbb{P}_\theta[\widetilde{y}_i = 0|x_i] \\
|V_{\theta,i}|_{\widetilde{y}_i=0} &= \sigma(\theta^\top x_i)\sigma(\varepsilon) + (1-\sigma(\theta^\top x_i))(1-\sigma(\varepsilon)) = \mathbb{P}_\theta[\widetilde{y}_i = 1|x_i] .
\end{aligned}$$

Therefore, it holds that $V_{\theta,i}$ is zero-mean and $v = 1$ sub-Gaussian under the conditional distribution $\mathbb{P}_\theta[\cdot|x_i]$.

Now the Hessian of the loss function is given by

$$\begin{aligned}
\nabla^2 \widehat{l}_{\mathcal{D},\varepsilon}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\Bigg[&\mathbb{1}(\widetilde{y}_i = 1)\big((1-\sigma(\varepsilon))\nabla^2 \log(1-\sigma(\theta^\top x_i)) - \sigma(\varepsilon)\nabla^2 \log \sigma(\theta^\top x_i)\big) \\
&+ \mathbb{1}(\widetilde{y}_i = 0)\big((1-\sigma(\varepsilon))\nabla^2 \log \sigma(\theta^\top x_i) - \sigma(\varepsilon)\nabla^2 \log(1-\sigma(\theta^\top x_i))\big)\Bigg] ,
\end{aligned}$$

where

$$\begin{aligned}
\nabla^2 \log \sigma(\theta^\top x_i) &= \frac{\sigma''(\theta^\top x_i)\sigma(\theta^\top x_i) - \sigma'(\theta^\top x_i)^2}{\sigma(\theta^\top x_i)^2} x_i x_i^\top , \\
\nabla^2 \log(1-\sigma(\theta^\top x_i)) &= -\frac{\sigma''(\theta^\top x_i)(1-\sigma(\theta^\top x_i)) + \sigma'(\theta^\top x_i)^2}{(1-\sigma(\theta^\top x_i))^2} x_i x_i^\top .
\end{aligned}$$

Now the second derivative of the logistic function $\sigma(\cdot)$ is given by $\sigma''(z) = \sigma'(z)(1-2\sigma(z))$, which gives us

$$\nabla^2 \log \sigma(\theta^\top x_i) = \nabla^2 \log(1-\sigma(\theta^\top x_i)) = -\sigma'(\theta^\top x_i) x_i x_i^\top .$$

Hence, the Hessian of the loss function takes the form

$$\nabla^2 \widehat{l}_{\mathcal{D},\varepsilon}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{1}(\widetilde{y}_i = 1)(2\sigma(\varepsilon) - 1)\sigma'(\theta^\top x_i) + \mathbb{1}(\widetilde{y}_i = 0)(2\sigma(\varepsilon) - 1)\sigma'(\theta^\top x_i)\right] x_i x_i^\top .$$

Now, under Assumption 2.1, observe that $\sigma'(\theta^\top x_i) \geq \gamma$ for all $\theta \in \Theta_B$, where $\gamma = \frac{1}{2+\exp(-2LB)+\exp(2LB)}$. This implies that $\widehat{l}_{\mathcal{D},\varepsilon}$ is $\gamma_\varepsilon := \gamma(2\sigma(\varepsilon) - 1)$ strongly convex in Θ_B for all $\varepsilon > 0$ with respect to the semi-norm $\|\cdot\|_{\Sigma_\mathcal{D}}$. Since $\theta^* \in \Theta_B$, introducing the error vector $\Delta = \widehat{\theta}_{\mathrm{RR}} - \theta^*$, we conclude that

$$\gamma_\varepsilon \|\Delta\|_{\Sigma_\mathcal{D}}^2 \leqslant \left\|\nabla \widehat{l}_{\mathcal{D},\varepsilon}(\theta^*)\right\|_{(\Sigma_\mathcal{D}+\lambda I)^{-1}} \|\Delta\|_{(\Sigma_\mathcal{D}+\lambda I)}$$

for some $\lambda > 0$. Introducing $M = \frac{1}{n^2} X(\Sigma_{\mathcal{D}} + \lambda I)^{-1} X^\top$, we now have $\|\nabla l_{\mathcal{D},\varepsilon}(\theta^*)\|^2_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} = V_{\theta^*}^\top M V_{\theta^*}$. Then, the Bernstein's inequality for sub-Gaussian random variables in quadratic form (see e.g. Hsu et al. (2012, Theorem 2.1)) implies that with probability at least $1-\alpha$,

$$\left\|\nabla \widehat{l}_{\mathcal{D},\varepsilon}(\theta^*)\right\|^2_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} = V_{\theta^*}^\top M V_{\theta^*} \leqslant v^2\left(\operatorname{tr}(M) + 2\sqrt{\operatorname{tr}(M^\top M)\log(1/\alpha)} + 2\|M\|\log(1/\alpha)\right)$$
$$\leqslant C_1 \cdot v^2 \cdot \frac{d+\log(1/\alpha)}{n}$$

for some $C_1 > 0$. This gives us

$$\gamma_\varepsilon \|\Delta\|^2_{\Sigma_{\mathcal{D}}+\lambda I} \leqslant \left\|\nabla \widehat{l}_{\mathcal{D},\varepsilon}(\theta^*)\right\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} \|\Delta\|_{(\Sigma_{\mathcal{D}}+\lambda I)} + 4\lambda\gamma_\varepsilon B^2$$
$$\leqslant \sqrt{C_1 \cdot v^2 \cdot \frac{d+\log(1/\alpha)}{n}}\,\|\Delta\|_{(\Sigma_{\mathcal{D}}+\lambda I)} + 4\lambda\gamma_\varepsilon B^2\ .$$

Solving for the above inequality, we get

$$\|\Delta\|_{(\Sigma_{\mathcal{D}}+\lambda I)} \leqslant C_2 \cdot \sqrt{\frac{v^2}{\gamma_\varepsilon^2}\cdot\frac{d+\log(1/\alpha)}{n} + \lambda B^2}$$

for some constant $C_2 > 0$. Now note that $\frac{v}{\gamma_\varepsilon} = \frac{1}{\gamma}\cdot\frac{e^\varepsilon+1}{e^\varepsilon-1}$. Hence, we get

$$\left\|\widehat{\theta}_{\mathrm{RR}} - \theta^*\right\|_{(\Sigma_{\mathcal{D}}+\lambda I)} \leqslant \frac{C}{\gamma}\cdot\frac{e^\varepsilon+1}{e^\varepsilon-1}\sqrt{\frac{d+\log(1/\alpha)}{n}} + C'\cdot\sqrt{\lambda}B,$$

for some $C, C' > 0$, which holds for any $\varepsilon \in (0,\infty)$. This completes our proof.

C.2.1 Logits

For any $\theta \in \mathbb{R}^d$, the logits (log-odds) of the probability that the clear-text label $y_i = 1$ given x_i is

$$\operatorname{logit}(p_{i,1}) = \log\frac{p_{i,1}}{p_{i,0}} = \log\frac{\sigma(x_i^\top\theta)}{1-\sigma(x_i^\top\theta)},$$

where the same for randomized label $\widetilde{y}_i = 1$ is

$$\operatorname{logit}(\widetilde{p}_{i,1}) = \log\frac{\widetilde{p}_{i,1}}{\widetilde{p}_{i,0}} = \log\frac{\sigma(x_i^\top\theta)\sigma(\varepsilon)+(1-\sigma(x_i^\top\theta))(1-\sigma(\varepsilon))}{(1-\sigma(x_i^\top\theta))\sigma(\varepsilon)+\sigma(x_i^\top\theta)(1-\sigma(\varepsilon))}\ .$$

By Jensen's inequality and basics of linear programming, we get

$$\operatorname{logit}(\widetilde{p}_{i,1}) \leqslant \log\left(\frac{\sigma(x_i^\top\theta)\sigma(\varepsilon)+(1-\sigma(x_i^\top\theta))(1-\sigma(\varepsilon))}{(1-\sigma(x_i^\top\theta))^{\sigma(\varepsilon)}\sigma(x_i^\top\theta)^{(1-\sigma(\varepsilon))}}\right) \leqslant \log\left(\frac{\max\{\sigma(\theta^\top x_i), 1-\sigma(\theta^\top x_i)\}}{(1-\sigma(x_i^\top\theta))^{\sigma(\varepsilon)}\sigma(x_i^\top\theta)^{(1-\sigma(\varepsilon))}}\right)\ .$$

Now, if $p_{i,1} \geq p_{i,0}$, then we have

$$\operatorname{logit}(\widetilde{p}_{i,1}) \leqslant \log\left(\frac{\sigma(\theta^\top x_i)}{(1-\sigma(x_i^\top\theta))^{\sigma(\varepsilon)}\sigma(x_i^\top\theta)^{(1-\sigma(\varepsilon))}}\right) \leqslant \sigma(\varepsilon)\log\left(\frac{\sigma(x_i^\top\theta)}{1-\sigma(x_i^\top\theta)}\right) = \sigma(\varepsilon)\cdot\operatorname{logit}(p_{i,1})\ .$$

Similarly, observe that

$$\operatorname{logit}(\widetilde{p}_{i,0}) \leqslant \log\left(\frac{(1-\sigma(x_i^\top\theta))\sigma(\varepsilon)+\sigma(x_i^\top\theta)(1-\sigma(\varepsilon))}{\sigma(x_i^\top\theta)^{\sigma(\varepsilon)}(1-\sigma(x_i^\top\theta))^{(1-\sigma(\varepsilon))}}\right) \leqslant \log\left(\frac{\max\{\sigma(\theta^\top x_i), 1-\sigma(\theta^\top x_i)\}}{\sigma(x_i^\top\theta)^{\sigma(\varepsilon)}(1-\sigma(x_i^\top\theta))^{(1-\sigma(\varepsilon))}}\right)\ .$$

Now, if $p_{i,0} \geq p_{i,1}$, then we have

$$\operatorname{logit}(\widetilde{p}_{i,0}) \leqslant \log\left(\frac{1-\sigma(\theta^\top x_i)}{\sigma(x_i^\top\theta)^{\sigma(\varepsilon)}(1-\sigma(x_i^\top\theta))^{(1-\sigma(\varepsilon))}}\right) \leqslant \sigma(\varepsilon)\log\left(\frac{1-\sigma(x_i^\top\theta)}{\sigma(x_i^\top\theta)}\right) = \sigma(\varepsilon)\cdot\operatorname{logit}(p_{i,0})\ .$$

Algorithm 1 SGD with Randomized Response

```
Parameters: privacy budget $\varepsilon$; i.i.d dataset $\mathcal{D} = (x_i, y_i)_{i=1}^n$; parameter space $\Theta_B$; learning rate $(\eta_t)_{t\geq 1}$.
Initialize: $\widehat{\theta}_1 = 0$.
for $t = 1, \ldots, n$ do
    Take data point $(x_t, y_t)$ from the dataset $\mathcal{D}$.
    Let $\widetilde{y}_t$ be the output of RR mechanism on $y_t$, i.e.,
        $\mathbb{P}\left[\tilde{y}_t = y_t\right] = \frac{e^\varepsilon}{1+e^\varepsilon}$ and $\mathbb{P}\left[\tilde{y}_t \neq y_t\right] = \frac{1}{1+e^\varepsilon}$ .
    Compute the gradient
        $\widehat{g}_t = \frac{\sum_{y\in\{0,1\}} \nabla_{\widehat{\theta}_t} \log p_{t,y}}{e^\varepsilon + 1} - \nabla_{\widehat{\theta}_t} \log p_{t,\widetilde{y}_t}$ ,
    where $p_{t,y}$ denotes the probability of observing $y \in \{0,1\}$ at round $t$, see (1).
    Update the estimate $\widehat{\theta}_{t+1} = \Pi_{\Theta_B}(\widehat{\theta}_t - \eta_t \widehat{g}_t)$
end for
Output $\widehat{\theta}_{\text{SGD-RR}} = \widehat{\theta}_{n+1}$.
```

Since $\sigma(\varepsilon) \in (1/2, 1)$ for any $\varepsilon > 0$, this impies that whenever y_i is more likely to occur than $1 - y_i$ in the clear-text, the log-odds of predicting y_i under ε-randomization given by (2) is at most $\sigma(\varepsilon)$-th fraction of the corresponding log-odds in the clear-text.

Therefore, we work with the predicted *scores* of randomized labels:

$$\widehat{p}_{i,1} = \frac{\sigma(x_i^\top \theta)^{\sigma(\varepsilon)}}{(1-\sigma(x_i^\top \theta))^{(1-\sigma(\varepsilon))}}, \quad \widehat{p}_{i,0} = \frac{(1-\sigma(x_i^\top \theta))^{\sigma(\varepsilon)}}{\sigma(x_i^\top \theta)^{(1-\sigma(\varepsilon))}} ,$$

which have the property

$$\log \frac{\widehat{p}_{i,1}}{\widehat{p}_{i,0}} = \log \frac{\sigma(x_i^\top \theta)}{1-\sigma(x_i^\top \theta)} = \text{logit}(p_{i,1}) ,$$

i.e., the log-odds of predicting y_i under ε-randomization is same as the corresponding log-odds in the clear-text.

C.3 Proof of Theorem 4.5

We divide the proof of Theorem 4.5 into the following steps. For ease of presentation, the complete algorithm for computing $\widehat{\theta}_{\text{SGD-RR}}$ is given in Algorithm 1.

Step 1: For each $t \geq 1$, we aim to show that there exists some constants λ, G and random variable $\widehat{z}_t$ such that

$$\left\|\widehat{\theta}_{t+1} - \theta^*\right\|^2 \leq (1 - 2/t) \left\|\widehat{\theta}_t - \theta^*\right\|^2 + \frac{2}{\lambda t}\langle \widehat{z}_t, \theta_t - \theta*\rangle + \left(\frac{G}{\lambda t}\right)^2 . \tag{17}$$

To this end, first recall that that the gradient at round t is given by

$$\widehat{g}_t = \frac{\sum_{y\in\{0,1\}} \nabla_{\widehat{\theta}_t} \log p_{t,y}}{e^\varepsilon + 1} - \nabla_{\widehat{\theta}_t} \log p_{t,\widetilde{y}_t} ,$$

where $p_{t,y}, y \in \{0,1\}$ denotes the probability of observing y at round t, see (1).

Now, we define $\widehat{z}_t := \mathbb{E}[\widehat{g}_t | \mathcal{F}_{t-1}] - \widehat{g}_t$, where $\mathcal{F}_{t-1} = \sigma\left(\{x_s, y_s, \widetilde{y}_s\}_{s=1}^{t-1}\right)$ is the σ-algebra generated by all the random variables up to and including round $t-1$. This conditioning is necessary since $\widehat{\theta}_t$ depends on randomness till round $t-1$. Then, we have

$$\begin{aligned}\left\|\widehat{\theta}_{t+1}-\theta^*\right\|^2 &= \left\|\Pi_{\Theta_B}(\widehat{\theta}_t-\eta_t\widehat{g}_t)-\theta^*\right\|^2\\ &\le \left\|\widehat{\theta}_t-\eta_t\widehat{g}_t-\theta^*\right\|^2\\ &= \left\|\widehat{\theta}_t-\theta^*\right\|^2-2\eta_t\langle\widehat{g}_t,\widehat{\theta}_t-\theta^*\rangle+\eta_t^2\left\|\widehat{g}_t\right\|^2\\ &= \left\|\widehat{\theta}_t-\theta^*\right\|^2-2\eta_t\langle\mathbb{E}[\widehat{g}_t|\mathcal{F}_{t-1}],\widehat{\theta}_t-\theta^*\rangle+2\eta_t\langle\widehat{z}_t,\widehat{\theta}_t-\theta^*\rangle+\eta_t^2\left\|\widehat{g}_t\right\|^2 \qquad (18)\end{aligned}$$

where the last equality holds by definition of $\widehat{z}_t$, i.e., $\widehat{g}_t=\mathbb{E}[\widehat{g}_t|\mathcal{F}_{t-1}]-\widehat{z}_t$.

To bound the above, we need to study the term $\langle\mathbb{E}[\widehat{g}_t|\mathcal{F}_{t-1}],\widehat{\theta}_t-\theta^*\rangle$. First note that $\widehat{g}_t$ is an unbiased and scaled estimate of the clear-text gradient $g_t=-\nabla_{\widehat{\theta}_t}\log p_{t,y_t}$ as

$$\begin{aligned}\mathbb{E}[\widehat{g}_t|\mathcal{F}_{t-1},x_t,y_t] &= \frac{\sum_{y\in\{0,1\}}\nabla_{\widehat{\theta}_t}\log p_{t,y}}{e^\varepsilon+1}-\left(\frac{e^\varepsilon}{e^\varepsilon+1}\nabla_{\widehat{\theta}_t}\log p_{t,y_t}+\frac{1}{e^\varepsilon+1}\nabla_{\widehat{\theta}_t}\log p_{t,1-y_t}\right)\\ &= -\frac{e^\varepsilon-1}{e^\varepsilon+1}\nabla_{\widehat{\theta}_t}\log p_{t,y_t}=(2\sigma(\varepsilon)-1)g_t\ .\end{aligned}$$

Then, by tower property of conditional expectation, we have

$$\mathbb{E}[\widehat{g}_t|\mathcal{F}_{t-1}]=\mathbb{E}[\mathbb{E}[\widehat{g}_t|\mathcal{F}_{t-1},x_t,y_t]|\mathcal{F}_{t-1}]=(2\sigma(\varepsilon)-1)\mathbb{E}[g_t|\mathcal{F}_{t-1}]=(2\sigma(\varepsilon)-1)\mathbb{E}[(\sigma(x_t^\top\widehat{\theta}_t)-y_t)x_t|\mathcal{F}_{t-1}]\ ,$$

where the final equality holds by definition of g_t. One more application of tower property gives us

$$\mathbb{E}[(\sigma(x_t^\top\widehat{\theta}_t)-y_t)x_t|\mathcal{F}_{t-1}]=\mathbb{E}[\mathbb{E}[(\sigma(x_t^\top\widehat{\theta}_t)-y_t)x_t|\mathcal{F}_{t-1},x_t]|\mathcal{F}_{t-1}]=\mathbb{E}[(\sigma(x_t^\top\widehat{\theta}_t)-\sigma(x_t^\top\widehat{\theta}^*))x_t|\mathcal{F}_{t-1}]\ .$$

Since $\widehat{\theta}_t$ is deterministic given $\mathcal{F}_{t-1}$, we can bound $\langle\mathbb{E}[\widehat{g}_t|\mathcal{F}_{t-1}],\widehat{\theta}_t-\theta^*\rangle$ using the above two equation as

$$\begin{aligned}\langle\mathbb{E}[\widehat{g}_t|\mathcal{F}_{t-1}],\widehat{\theta}_t-\theta^*\rangle &= (2\sigma(\varepsilon)-1)\mathbb{E}[\langle(\sigma(x_t^\top\widehat{\theta}_t)-\sigma((x_t^\top\theta^*))x_t,\widehat{\theta}_t-\theta^*\rangle|\mathcal{F}_{t-1}]\\ &\overset{(a)}{\ge}\gamma(2\sigma(\varepsilon)-1)\mathbb{E}[(x_t^\top(\widehat{\theta}_t-\theta^*))^2|\mathcal{F}_{t-1}]\\ &\overset{(b)}{\ge}\gamma_\varepsilon(\widehat{\theta}_t-\theta^*)^\top\mathbb{E}[x_tx_t^\top|\mathcal{F}_{t-1}](\widehat{\theta}_t-\theta^*)\\ &\overset{(c)}{\ge}\gamma_\varepsilon\kappa\left\|\widehat{\theta}_t-\theta^*\right\|^2\ . \qquad (19)\end{aligned}$$

Here (a) holds by mean-value theorem and by noting that $\sigma'(\theta^\top x_i)\ge\gamma$ for all $\theta\in\Theta_B$ under Assumption 2.1, where $\gamma=\frac{1}{2+\exp(-2LB)+\exp(2LB)}$; (b) holds by defining $\gamma_\varepsilon=(2\sigma(\varepsilon)-1)\gamma$; (c) holds by Assumption 4.4 and by noting that x_t is independent of $\mathcal{F}_{t-1}$.

Now, plugging (19) into (18), yields

$$\begin{aligned}\left\|\widehat{\theta}_{t+1}-\theta^*\right\|^2 &\le \left\|\widehat{\theta}_t-\theta^*\right\|^2(1-2\eta_t\gamma\kappa)+2\eta_t\langle z_t,\widehat{\theta}_t-\theta^*\rangle+\eta_t^2\left\|\widehat{g}_t\right\|^2\\ &\overset{(a)}{\le}\left\|\widehat{\theta}_t-\theta^*\right\|^2(1-2\eta_t\gamma_\varepsilon\kappa)+2\eta_t\langle z_t,\widehat{\theta}_t-\theta^*\rangle+\eta_t^2G^2\\ &\overset{(b)}{=}(1-2/t)\left\|\widehat{\theta}_t-\theta^*\right\|^2+\frac{2}{\lambda t}\langle\widehat{z}_t,\widehat{\theta}_t-\theta*\rangle+\left(\frac{G}{\lambda t}\right)^2\end{aligned}$$

where (a) holds by bounding $\|\widehat{g}_t\|\le G:=4L$ under Assumption 2.1; (b) holds by letting $\lambda:=\gamma_\varepsilon\kappa$ and $\eta_t:=\frac{1}{\lambda t}$; Hence, we have established (17).

Step 2: We aim to show that for all $t\ge 2$

$$\left\|\widehat{\theta}_{t+1}-\theta^*\right\|^2\le\frac{2}{\lambda(t-1)t}\sum_{i=2}^{t}(i-1)\langle\widehat{z}_i,\widehat{\theta}_i-\theta^*\rangle+\frac{G^2}{\lambda^2t^2}. \qquad (20)$$

To this end, we basically expand the recursion in (17) till $t = 2$ and simple algebra leads to the result. This step directly follows from Rakhlin et al. (2011).

Step 3: We will apply one particular version of Freedman's inequality to control the concentration of $\sum_{i=2}^{t}(i-1)\langle \widehat{z}_i, \widehat{\theta}_i - \theta^* \rangle$ in (20). In particular, we will apply (Rakhlin et al., 2011, Lemma 3) to bound this sum of martingale differences for all $t \leq n$. This needs to hold for all t since we will rely on induction later.

To start with, we let $Z_i = \langle \widehat{z}_i, \widehat{\theta}_i - \theta^* \rangle$. Then, we have the conditional expectation of Z_i given $\mathcal{F}_{i-1}$ is $\mathbb{E}[Z_i|\mathcal{F}_{i-1}] = 0$ and the conditional variance is $\mathrm{Var}[Z_i|\mathcal{F}_{i-1}] \leq 4G^2 \left\|\widehat{\theta}_i - \theta^*\right\|^2$, which holds by $\|\widehat{z}_i\| \leq 2G$. Now consider the sum $\sum_{i=2}^{t}(i-1)\langle \widehat{z}_i, \theta_i - \theta^* \rangle$ in (20). We need to check two conditions: (i) The sum of conditional variance satisfies

$$\sum_{i=2}^{t} \mathrm{Var}[(i-1)Z_i|\mathcal{F}_{i-1}] \leq 4G^2 \sum_{i=2}^{t}(i-1)^2 \left\|\widehat{\theta}_i - \theta^*\right\|^2 \;;$$

(ii) Uniform upper bound on each term satisfies

$$|(i-1)Z_i| \leq 2G(t-1)\left\|\widehat{\theta}_i - \theta^*\right\| \overset{(a)}{\leq} \frac{2G^2(t-1)}{\lambda},$$

where (a) comes from (19) and substituting $\lambda = \gamma_\varepsilon \kappa$. To see this, by Cauchy-Schwartz inequality, we have $\gamma_\varepsilon \kappa \left\|\widehat{\theta}_t - \theta^*\right\|^2 \leq G\left\|\widehat{\theta}_t - \theta^*\right\|$, and hence $\left\|\widehat{\theta}_t - \theta^*\right\| \leq G/\lambda$ for all t. We can then apply (Rakhlin et al., 2011, Lemma 3) to obtain that for $n \geq 4$ and $\alpha \in (0, 1/e)$, with probability at least $1-\alpha$, it holds for all $t \leq n$ that

$$\sum_{i=2}^{t}(i-1)Z_i \leq 8G \max\left\{ \sqrt{\sum_{i=2}^{t}(i-1)^2\left\|\widehat{\theta}_i - \theta^*\right\|^2}, \frac{G(t-1)}{\lambda}\sqrt{\log(\log n/\alpha)} \right\} \sqrt{\log(\log n/\alpha)}. \tag{21}$$

Step 4: Once we obtain (21), the remaining step is all about induction and algebra, which follows the same procedures as in Rakhlin et al. (2011). After all, we will obtain with probability at least $1-\alpha$ that for all $t \leq n$,

$$\begin{aligned}\left\|\widehat{\theta}_{t+1} - \theta^*\right\|^2 &\leq \frac{(624\log(\log n/\alpha)+1)G^2}{\lambda^2 t} \\ &= CL^2\left(\frac{e^\varepsilon+1}{e^\varepsilon-1}\right)^2 \cdot \frac{\log(\log n/\alpha)+1}{\gamma^2\kappa^2 t},\end{aligned}$$

for some absolute constant C. Setting $t = n$ completes the proof.

C.4 Estimation Error under Placket-Luce Model

Let s be a state and $a_1, \ldots, a_K$ be K actions to be compared at that state. Let the label/preference feedback $y \in \{1, 2, \ldots, K\}$ indicates which action is most preferred by human labeler. Let $x_{i,j} = \phi(s, a_i) - \phi(s, a_j), 1 \leq i \neq j \leq K$ be the feature difference between actions a_i and a_j at state s. Define the population covariance matrix

$$\Sigma_{i,j} = \mathbb{E}_{s\sim\rho(\cdot),(a_1,\ldots,a_K)\sim\mu(\cdot|s)}\left[x_{i,j}x_{i,j}^\top\right] \;.$$

Assumption C.1 (Coverage of feature space)**.** The data distributions ρ, μ are such that $\lambda_{\min}(\Sigma_{i,j}) \geq \kappa$ for some constant $\kappa > 0$ for all $1 \leq i \neq j \leq K$.

This is a coverage assumption on the state-action feature space. The next result bounds the estimation error of $\widehat{\theta}_{\text{SGD-KRR}}$ in ℓ_2-norm.

Theorem C.2 (Estimation error of $\widehat{\theta}_{\text{SGD-KRR}}$ under Placket-Luce model)**.** *Fix $\alpha \in (0, 1/e)$ and $\varepsilon > 0$. Then, under the Placket Luce model* (11) *and under Assumptions 2.1 and C.1 and setting $\eta_t = \frac{1}{\gamma\kappa}$, we have, with probability at least $1-\alpha$,*

$$\left\|\widehat{\theta}_{SGD\text{-}KRR} - \theta^*\right\|_2 \leq C \cdot \frac{L}{\gamma\kappa} \cdot \frac{e^\varepsilon + K - 1}{e^\varepsilon - 1}\sqrt{\frac{\log(\log(n)/\alpha)}{n}},$$

where $\gamma = \frac{e^{-4LB}}{2}$, C is an absolute constant.

Proof. We will first show that for all $t \geq 1$, the parameter updates satisfy

$$\langle \mathbb{E}[\widehat{g}_t | \mathcal{F}_{t-1}], \widehat{\theta}_t - \theta^* \rangle \geq \gamma_{K,\varepsilon} \kappa \left\| \widehat{\theta}_t - \theta^* \right\|^2 \ , \tag{22}$$

where $\widehat{g}_t = \frac{\sum_{y=1}^K \nabla_{\widehat{\theta}_t} \log p_{t,y}}{e^\varepsilon + K - 1} - \nabla_{\widehat{\theta}_t} \log p_{t,\widetilde{y}_t}$ is the gradient, $\mathcal{F}_{t-1} = \sigma\big(\{x_s, y_s, \widetilde{y}_s\}_{s=1}^{t-1}\big)$ is the σ-algebra generated by all the random variables up to and including round $t-1$, $\gamma_{K,\varepsilon} := \gamma \frac{e^\varepsilon - 1}{e^\varepsilon + K - 1}$ and $\gamma = e^{-4LB}/2$. Then, one can follow the steps used in the proof of Theorem 4.5 to derive this result.

Let's now establish (22). To this end, let Π be the set of all permutations $\pi : [K] \to [K]$ that denotes a ranking over all K actions given by a human labeler, where $a_{\pi(1)}$ denotes the highest-ranked action. Under the Placket-Luce model, one can compute the probability of observing the permutation $\pi \in \Pi$ as

$$\mathbb{P}_{\theta^*}[\pi | s, a_1, \ldots, a_K] = \prod_{j=1}^K \frac{\exp(\phi(s, a_{\pi(j)})^\top \theta^*)}{\sum_{k'=j}^K \exp(\phi(s, a_{\pi(k')})^\top \theta^*)} \ .$$

Define, with an abuse of notation, $x = (s, a_1, \ldots, a_K)$ and $x_{\pi(j)} = \phi(s, a_{\pi(j)})$ for all $j \in [K]$. This lets us denote for any $\theta \in \mathbb{R}^d$:

$$\mathbb{P}_\theta[\pi | x] = \prod_{j=1}^K \mathbb{P}_\theta[\pi(j) | x] \ , \text{ where } \mathbb{P}_\theta[\pi(j)|x] = \frac{\exp(x_{\pi(j)}^\top \theta)}{\sum_{k'=j}^K \exp(x_{\pi(k')}^\top \theta)} \ .$$

The negative log-likelihood (log-loss) of predicting the the highest-ranked action $a_{\pi(1)}$ given x is

$$l_\theta(a_{\pi(1)}, x) := -\log \mathbb{P}_\theta[\pi(1)|x] = -\log \frac{\exp(x_{\pi(1)}^\top \theta)}{\sum_{k'=1}^K \exp(x_{\pi(k')}^\top \theta)} \ .$$

The expected log-loss takes the form

$$G_\theta(x) := \mathbb{E}_{\pi \sim \mathbb{P}_{\theta^*}[\cdot|x]} \left[l_\theta(a_{\pi(1)}, x) \right] = \sum_{\pi \in \Pi} \mathbb{P}_{\theta^*}[\pi | x] \, l_\theta\big(a_{\pi(1)}, x\big) \ .$$

This yields the following:

$$\nabla^2 G_\theta(x) = \sum_{\pi \in \Pi} \mathbb{P}_{\theta^*}[\pi|x] \, \nabla^2 l_\theta(a_{\pi(1)}, x) \ .$$

Note that the following holds (Zhu et al., 2023):

$$\nabla l_\theta(a_{\pi(1)}, x) = \sum_{k=1}^K \frac{\exp(x_{\pi(k)}^\top \theta)}{\sum_{k'=1}^K \exp(x_{\pi(k')}^\top \theta)} \big(x_{\pi(1)} - x_{\pi(k)}\big) \ ,$$

$$\nabla^2 l_\theta(a_{\pi(1)}, x) = \sum_{k=1}^K \sum_{k'=1}^K \frac{\exp(x_{\pi(k)}^\top \theta) \cdot \exp(x_{\pi(k')}^\top \theta)}{2 \left(\sum_{k'=1}^K \exp(x_{\pi(k')}^\top \theta) \right)^2} \big(x_{\pi(k)} - x_{\pi(k')}\big) \big(x_{\pi(k)} - x_{\pi(k')}\big)^\top \ .$$

Under Assumption 2.1, we have $-LB \leq \phi(s,a)^\top \theta \leq LB$ for all $\theta \in \Theta_B$. Define $x_{\pi,k,k'} = x_{\pi(k)} - x_{\pi(k')}$ for all $k, k' \in [K]$. Then, for any $v \in \mathbb{R}^d$ and $\theta \in \Theta_B$, we have

$$v^\top \nabla^2 l_\theta(a_{\pi(1)}, x) v \geq \frac{e^{-4LB}}{2} \cdot v^\top \left(\frac{1}{K^2} \sum_{k=1}^K \sum_{k'=1}^K x_{\pi,k,k'} x_{\pi,k,k'}^\top \right) v \ .$$

Define the matrix

$$\Sigma(\pi, x) := \frac{1}{K^2} \sum_{k=1}^K \sum_{k'=1}^K x_{\pi,k,k'} x_{\pi,k,k'}^\top \ .$$

Then, for $\theta \in \Theta_B$ and $\pi \in \Pi$, the loss function $l_\theta(a_{\pi(1)}, x)$ is $\gamma = \frac{e^{-4LB}}{2}$ strongly convex w.r.t. the semi-norm $\|\cdot\|_{\Sigma(\pi,x)}$. This further implies that $G_\theta(x)$ is γ strongly convex w.r.t. the semi-norm $\|\cdot\|_{\Sigma(x)}$, where $\sigma(x) := \sum_{\pi\in\Pi} \mathbb{P}_{\theta^*}[\pi|x]\, \Sigma(\pi, x)$.

Since $\theta^* \in \Theta_B$, we have from definition of strong convexity,

$$G_{\theta^*}(x) \geq G_\theta(x) + \langle \nabla G_\theta(x), \theta^* - \theta\rangle + \frac{\gamma}{2}\|\theta - \theta^*\|^2_{\Sigma(x)} \implies \langle \nabla G_\theta(x), \theta - \theta^*\rangle \geq G_\theta(x) - G_{\theta^*}(x) + \frac{\gamma}{2}\|\theta - \theta^*\|^2_{\Sigma(x)} \;.$$

Since $\theta^* \in \operatorname{argmin}_{\theta\in\Theta_B} G_\theta(x)$, we have from definition of first-order optimality of convex functions,

$$G_\theta(x) - G_{\theta^*}(x) \geq \langle \nabla G_{\theta^*}(x), \theta - \theta^*\rangle + \frac{\gamma}{2}\|\theta - \theta^*\|^2_{\Sigma(x)} \geq \frac{\gamma}{2}\|\theta - \theta^*\|^2_{\Sigma(x)} \;.$$

Combining the above, we have for any $\theta \in \Theta_B$:

$$\langle \nabla G_\theta(x), \theta - \theta^*\rangle \geq \gamma \|\theta - \theta^*\|^2_{\Sigma(x)} \implies \langle \mathbb{E}_{\pi\sim\mathbb{P}_{\theta^*}}\left[\nabla l_\theta(a_{\pi(1)}, x)|x\right], \theta - \theta^*\rangle \geq \gamma \|\theta - \theta^*\|^2_{\Sigma(x)} \;.$$

Now, taking expectation over $x \sim (\rho \times \mu)$, we have

$$\begin{aligned}\langle \mathbb{E}_{x\sim(\rho\times\mu),\pi\sim\mathbb{P}_{\theta^*}[\cdot|x]}\left[\nabla l_\theta(a_{\pi(1)}, x)\right], \theta - \theta^*\rangle &\geq \gamma(\theta - \theta^*)^\top \mathbb{E}_x\left[\Sigma(x)\right](\theta - \theta^*)\\ &= \gamma(\theta - \theta^*)^\top \mathbb{E}_x\left[\sum_{\pi\in\Pi}\mathbb{P}_{\theta^*}[\pi|x]\Sigma(\pi, x)\right](\theta - \theta^*) \;.\end{aligned}$$

Note that, by the coverage Assumption C.1, we have $\mathbb{E}_x\left[\Sigma(\pi, x)\right] \geq \kappa$ for all $\pi \in \Pi$. This yields for any $v \in \mathbb{R}^d$,

$$\forall\pi \in \Pi,\; v^\top \mathbb{E}_x\left[\Sigma(\pi, x)\right] v \geq \kappa\|v\|^2 \implies \mathbb{E}_x\left[\min_{\pi\in\Pi} v^\top\Sigma(\pi, x)v\right] \geq \kappa\|v\|^2 \;.$$

This further yields

$$\begin{aligned}v^\top \mathbb{E}_x\left[\sum_{\pi\in\Pi}\mathbb{P}_{\theta^*}[\pi|x]\Sigma(\pi, x)\right] v &= \mathbb{E}_x\left[\sum_{\pi\in\Pi}\mathbb{P}_{\theta^*}[\pi|x]v^\top\Sigma(\pi, x)v\right]\\ &\geq \mathbb{E}_x\left[\min_{\pi\in\Pi} v^\top\Sigma(\pi, x)v\right] \geq \kappa\|v\|^2 \;.\end{aligned}$$

This implies for any $\theta \in \Theta_B$, the following:

$$\langle \mathbb{E}_{x\sim(\rho\times\mu),\pi\sim\mathbb{P}_{\theta^*}[\cdot|x]}\left[\nabla l_\theta(a_{\pi(1)}, x)\right], \theta - \theta^*\rangle \geq \gamma\kappa\|\theta - \theta^*\|^2 \;. \tag{23}$$

Now, let π_t be the permutation (ranking) given by human labeler at round t, i.e. $\pi_t(1) = y_t$, and $\widetilde{\pi}_t$ be the (noisy) ranking after randomization by KRR mechanism (12), i.e. $\widetilde{\pi}_t(1) = \widetilde{y}_t$. Note that, we have

$$\mathbb{P}\left[\widetilde{\pi}_t(1) = \pi_t(1)\right] = \frac{e^\varepsilon}{e^\varepsilon + K - 1} \;,\; \mathbb{P}\left[\widetilde{\pi}_t(1) = y\right] = \frac{1}{e^\varepsilon + K - 1} \;, \forall y \neq \pi_t(1) \;.$$

Using this, we can re-write the gradient as

$$\widehat{g}_t = \frac{\sum_{y=1}^K \nabla_{\widehat{\theta}_t}\log p_{t,y}}{e^\varepsilon + K - 1} - \nabla_{\widehat{\theta}_t}\log p_{t,\widetilde{\pi}_t(1)} \;,$$

where $p_{t,y}, y \in [K]$ is the probability of observing y at round t, see (11). Then, we have

$$\begin{aligned}\mathbb{E}[\widehat{g}_t|\mathcal{F}_{t-1}, x_t, y_t] &= \frac{\sum_{y\in\{0,1\}}\nabla_{\widehat{\theta}_t}\log p_{t,y}}{e^\varepsilon + K - 1} - \left(\frac{e^\varepsilon}{e^\varepsilon + 1}\nabla_{\widehat{\theta}_t}\log p_{t,y_t} + \frac{1}{e^\varepsilon + 1}\nabla_{\widehat{\theta}_t}\log p_{t,1-y_t}\right)\\ &= -\frac{e^\varepsilon - 1}{e^\varepsilon + K - 1}\nabla_{\widehat{\theta}_t}\log p_{t,y_t} = \frac{e^\varepsilon - 1}{e^\varepsilon + K - 1}\nabla l_{\widehat{\theta}_t}(a_{\pi_t(1)}, x_t) \;.\end{aligned}$$

Since $\widehat{\theta}_t$ is deterministic given $\mathcal{F}_{t-1}$, by tower property of conditional expectation, we have

$$\begin{aligned}\langle \mathbb{E}[\widehat{g}_t|\mathcal{F}_{t-1}], \widehat{\theta}_t - \theta^*\rangle &= \frac{e^\varepsilon - 1}{e^\varepsilon + K - 1}\mathbb{E}[\langle \nabla l_{\widehat{\theta}_t}(a_{\pi_t(1)}, x_t), \widehat{\theta}_t - \theta^*\rangle|\mathcal{F}_{t-1}]\\ &\geq \gamma\kappa\frac{e^\varepsilon - 1}{e^\varepsilon + K - 1}\left\|\widehat{\theta}_t - \theta^*\right\|^2 \;,\end{aligned}$$

where the last step follows from (23) and by noting that x_t, π_t are independent of $\mathcal{F}_{t-1}$. Defining $\gamma_{K,\varepsilon} := \gamma\frac{e^\varepsilon - 1}{e^\varepsilon + K - 1}$, we get (22). This completes our proof. □

D ADDITIONAL DETAILS ON SECTION 5

D.1 Proof of Theorem 5.1

Before presenting the proof, let us introduce the following useful lemma. For the label-DP in the central model, we will leverage the DP version of Assouad's lemma in Acharya et al. (2021), which is re-stated as follows[3].

Lemma D.1 (Assouad's lemma for central DP)**.** *Let the same conditions of Lemma B.4 hold. If for all $i \in [d]$, there exists a coupling (X, Y) between P_{+i} and P_{-i} with $\mathbb{E}\left[d_{\mathrm{Ham}}(X,Y)\right] \le D$ for some $D \ge 0$, then*

$$R_c(\mathcal{P}, \rho, \varepsilon, \delta) \ge \frac{d\tau}{2\alpha} \cdot \left(0.9 e^{-10\varepsilon D} - 10 D\delta\right).$$

Now, we are well-prepared to prove Theorem 5.1.

Proof of Theorem 5.1. First note that the non-private part is the same as before. Thus, we only need to focus on the second private part.

Choose some $\Delta > 0$ and for each $e \in \mathcal{E}_d = \{\pm 1\}^d$, let $\theta_e = \Delta e$. Now we need to check the two conditions in Lemma B.4. First note that $\rho = \left\|\cdot\right\|_2^2$ satisfies 2-triangle inequality, i.e., $\alpha = 2$. Also, note that for any $u, v \in \mathcal{E}_d$, $\left\|\theta_u - \theta_v\right\|_2^2 = 4\Delta^2 \sum_{i=1}^d \mathbb{1}(u_i \neq v_i)$, i.e., $\tau = 2\Delta^2$. Thus, let P_{+i}^n be the product distribution of P_{+i} and similarly for P_{-i}^n, then by Lemma D.1, we have

$$\begin{aligned} R_c(\mathcal{P}_{\log}, \left\|\cdot\right\|_2^2, \varepsilon, \delta) &\ge \frac{d\Delta^2}{2}\left(0.9 e^{-10\varepsilon D} - 10 D\delta\right) \\ &\overset{(a)}{\ge} \frac{d\Delta^2}{2}\left(0.9 - 10 D(\varepsilon + \delta)\right) \end{aligned}$$

where D is the bound on the expected hamming distance between (X, Y), which is a coupling between P_{+i}^n and P_{-i}^n; (a) holds by the fact that $e^x \ge 1 + x$.

Thus, it remains to determine D in our case. That is, we need to bound the expected hamming distance between two product distributions P_{+i}^n and P_{-i}^n. Note that for the lower bound, it suffices to consider $x_k = x \in \mathbb{R}^d$ with $\left\|x\right\|_\infty \le 1$ for all $k \in [n]$. In this case, by the standard result on maximal coupling, we have that for the maximal coupling (X, Y) between P_{+i}^n and P_{-i}^n,

$$\mathbb{E}\left[d_{\mathrm{Ham}}(X,Y)\right] = n\left\|P_{+i} - P_{-i}\right\|_{\mathrm{TV}}.$$

Now, it remains to bound the TV-distance. To this end, by (15) and joint convexity of TV distance, we have

$$\begin{aligned} \left\|P_{+i} - P_{-i}\right\|_{\mathrm{TV}} &= \left\|\frac{1}{|\mathcal{E}_d|}\sum_{e \in \mathcal{E}_d} P_{e,+i} - P_{e,-i}\right\|_{\mathrm{TV}} \\ &\le \frac{1}{|\mathcal{E}_d|}\sum_{e \in \mathcal{E}_d}\left\|P_{e,+i} - P_{e,-i}\right\|_{\mathrm{TV}} \\ &\le \max_{e \in \mathcal{E}_d, i \in [d]}\left\|P_{e,+i} - P_{e,-i}\right\|_{\mathrm{TV}} \\ &= \max_{e \in \mathcal{E}_d, i \in [d]}\left\|P_e - P_{\bar{e}^i}\right\|_{\mathrm{TV}}, \end{aligned}$$

where recall that $\bar{e}^i$ is a vector in $\mathcal{E}_d$ that flips the i-th coordinate of e. To proceed, for any $i \in [d]$, by Pinsker's inequality, we have

$$\left\|P_e - P_{\bar{e}^i}\right\|_{\mathrm{TV}}^2 \le \frac{1}{4}\left(D_{\mathrm{KL}}\left(P_e \| P_{\bar{e}^i}\right) + D_{\mathrm{KL}}\left(P_{\bar{e}^i} \| P_e\right)\right) \overset{(a)}{\le} \Delta^2,$$

where (a) follows from Claim B.1 and the choice of x such that $\left\|x\right\|_\infty \le 1$. Thus, putting everything together, yields that

$$\mathbb{E}\left[d_{\mathrm{Ham}}(X,Y)\right] = n\left\|P_{+i} - P_{-i}\right\|_{\mathrm{TV}} \le n\Delta := D.$$

[3] We correct some constant factor error in the original statement.

Algorithm 2 Objective Perturbation with Gaussian Noise

```
1: Parameters: privacy budget $\varepsilon > 0, \delta \in (0,1)$; regularization parameter $\beta$; i.i.d dataset $\mathcal{D} = (x_i, y_i)_{i=1}^n$; parameter space $\Theta_B$; log loss $\ell$
2: Sample $w \in \mathcal{N}(0, \sigma^2 I)$
3: Return $\widehat{\theta}_{\text{obj}} = \text{argmin}_{\theta \in \Theta_B} l_{\mathcal{D}}(\theta) + \frac{\beta}{2n} \|\theta\|_2^2 + \frac{w^\top \theta}{n}$, where $l_{\mathcal{D}}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\theta, (x_i, y_i))$
```

With this value of D, we finally obtain that

$$R_c(\mathcal{P}_{\log}, \|\cdot\|_2^2, \varepsilon, \delta) \geq \frac{d\Delta^2}{2}\left(0.9 - 10n\Delta(\varepsilon+\delta)\right).$$

Thus, choosing $\Delta = \frac{0.04}{n(\varepsilon+\delta)}$, we obtain that

$$R_c(\mathcal{P}_{\log}, \|\cdot\|_2^2, \varepsilon, \delta) \geq c \cdot \frac{d}{n^2(\varepsilon+\delta)^2},$$

for some universal constant c. Finally, combined with the non-private part, we have finished the proof. □

D.2 Proof of Theorem 5.2

Before we present the proof, we first highlight some differences between our proof and the one in Kifer et al. (2012). In particular, we note that one cannot simply follow the one in Kifer et al. (2012) as there exists a gap in Lemma 16 of Kifer et al. (2012) due to non-independence. Thus, we carefully handle this subtlety under our model. We also explicitly write down the two-step procedures of Successive Approximation to handle the minimization over a constrained set.

Now, we are ready to present the proof.

Proof of Theorem 5.2. Our goal is to show that $\widehat{\theta}_{\text{obj}}$ is (ε, δ)-label DP in the central model, where

$$\widehat{\theta}_{\text{obj}} = \operatorname*{argmin}_{\theta \in \Theta_B} l_{\mathcal{D}}(\theta) + \frac{\beta}{2n}\|\theta\|_2^2 + \frac{w^\top \theta}{n}. \tag{24}$$

To this end, we will first use Successive Approximation (Theorem 1 in Kifer et al. (2012)), which allows us to only focus on the following sequence of unconstrained problems (indexed by $i \in \mathbb{N}$).

$$\widehat{\theta}_{\text{obj}}^{(i)} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} l_{\mathcal{D}}(\theta) + \frac{\beta}{2n}\|\theta\|_2^2 + \frac{w^\top \theta}{n} + \frac{if(\theta)}{n}, \tag{25}$$

where $f(\theta) = \min_{z \in \Theta_B} \|\theta - z\|_2$, which is a convex function (but not necessarily differentiable everywhere). The technique of Successive Approximation (SA) says that it suffices to show that for each i, the computation in (25) is (ε, δ)-label DP. To show this, we will have to use SA again as $f(\theta)$ in (25) is not differentiable everywhere. To handle this, for each i, we will consider another sequence of problems (indexed by $j \in \mathbb{N}$) as follows

$$\widehat{\theta}_{\text{obj}}^{(i,j)} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} l_{\mathcal{D}}(\theta) + \frac{\beta}{2n}\|\theta\|_2^2 + \frac{w^\top \theta}{n} + \frac{1}{n} r^{(i,j)}(\theta), \tag{26}$$

where $r^{(i,j)}(\theta)$ be the convolution between $if(\theta)$ and K_j (defined in Eq.(5) of Kifer et al. (2012)). Now, we have $r^{(i,j)}(\theta)$ is differentiable everywhere and convex. Thus, it only remains to show that the computation in (26) is (ε, δ)-label DP, for all i, j.

Fix a pair (i, j), we simplify notation in (26) by focusing on the following problem.

$$\widetilde{\theta}_{\mathcal{D}} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} l_{\mathcal{D}}(\theta) + \frac{\beta}{2n}\|\theta\|_2^2 + \frac{w^\top \theta}{n} + \frac{1}{n} r(\theta). \tag{27}$$

Step 1: Establish the PDF for $\widetilde{\theta}_{\mathcal{D}}$.

By differentiability, we have $n\nabla_\theta l_{\mathcal{D}}(\widetilde{\theta}_{\mathcal{D}}) + \beta\widetilde{\theta}_{\mathcal{D}} + w + \nabla r(\widetilde{\theta}_{\mathcal{D}}) = 0$. Define $\psi_{\mathcal{D}}(\theta) := n\nabla_\theta l_{\mathcal{D}}(\theta) + \beta\theta + \nabla r(\theta)$. By change of random variables and $w \sim \mathcal{N}(0, \sigma^2 I_d)$, we have that the probability density of $\widetilde{\theta}_{\mathcal{D}}$ is given by for $t \in \mathbb{R}^d$

$$f_{\widetilde{\theta}_{\mathcal{D}}}(t) = \underbrace{C\exp\left(-\frac{\|\psi_{\mathcal{D}}(t)\|_2^2}{2\sigma^2}\right)}_{\mathcal{T}_{1,\mathcal{D}}} \cdot \underbrace{\left|\det\left[\frac{d\psi_{\mathcal{D}}(\theta)}{d\theta}|_{\theta=t}\right]\right|}_{\mathcal{T}_{2,\mathcal{D}}} \tag{28}$$

where we use the fact that if X has density function f_X, $Y = H(X)$ for some bijective, differentiable function H, the Y has density

$$f_Y(y) = f_X(H^{-1}(y))\left|\det\left[\frac{dH^{-1}(z)}{dz}|_{z=y}\right]\right|.$$

Note that here the bijective relation holds by the strong convexity thanks to the regularization term $\beta > 0$.

Step 2: Bound the PDF ratio under two neighboring datasets.

By definition of DP, it suffices to show that for all $t \in \mathbb{R}^d$, with probability at least $1-\delta$

$$e^{-\varepsilon} f_{\widetilde{\theta}_{\mathcal{D}'}}(t) \le f_{\widetilde{\theta}_{\mathcal{D}}}(t) \le e^{\varepsilon} f_{\widetilde{\theta}_{\mathcal{D}'}}(t),$$

for all neighboring datasets D, D'. To this end, we first look at the ratio of $\mathcal{T}_{2,\mathcal{D}}/\mathcal{T}_{2,\mathcal{D}'}$. Note that the matrix inside the determinant in (28) is the Hessian of $l_{\mathcal{D}}$ plus some common terms, given by

$$\nabla^2 r(\theta)|_{\theta=t} + \beta I + \nabla^2 l_{\mathcal{D}}(\theta)|_{\theta=t} = \nabla^2 r(\theta)|_{\theta=t} + \beta I + \sum_{i=1}^n \sigma(x_i^\top t)(1-\sigma(x_i^\top t))x_i x_i^\top,$$

which does not depend on labels $\{y_i\}_{i=1}^n$. Thus, $\mathcal{T}_{2,\mathcal{D}}/\mathcal{T}_{2,\mathcal{D}'} = 1$.

Now, we turn to the ratio of $\mathcal{T}_{1,\mathcal{D}}/\mathcal{T}_{1,\mathcal{D}'}$. In particular, we have

$$\begin{aligned}\frac{\mathcal{T}_{1,\mathcal{D}}}{\mathcal{T}_{1,\mathcal{D}'}} &= \exp\left(\frac{\|\psi_{\mathcal{D}'}(t)\|_2^2 - \|\psi_{\mathcal{D}}(t)\|_2^2}{2\sigma^2}\right)\\ &= \exp\left(\frac{2\langle\psi_{\mathcal{D}}(t), \psi_{\mathcal{D}'(t)} - \phi_{\mathcal{D}}(t)\rangle + \|\phi_{\mathcal{D}'}(t) - \psi_{\mathcal{D}}(t)\|_2^2}{2\sigma^2}\right)\\ &= \exp\left(\frac{2\langle -w, \psi_{\mathcal{D}'(t)} - \psi_{\mathcal{D}}(t)\rangle + \|\psi_{\mathcal{D}'}(t) - \psi_{\mathcal{D}}(t)\|_2^2}{2\sigma^2}\right).\end{aligned} \tag{29}$$

where we know that $\psi_{\mathcal{D}}(t) = -w$, which is distributed according to a normal. However, one needs to be careful here to show that $\phi_{\mathcal{D}'}(t) - \phi_{\mathcal{D}}(t)$ is *independent* of w so that one can claim that the inner product is also distributed according to a normal. In fact, this is not true in general[4]! Fortunately, in our case, for two neighboring datasets $\mathcal{D}, \mathcal{D}'$ that differs only in y_j, y_j' we have

$$\begin{aligned}\psi_{\mathcal{D}'}(t) - \psi_{\mathcal{D}}(t) &= \left(\frac{1}{1+\exp(-\langle x_j, t\rangle)} - y_j'\right)x_j - \left(\frac{1}{1+\exp(-\langle x_j, t\rangle)} - y_j\right)x_j\\ &= x_j y_j - x_j y_j',\end{aligned}$$

which is independent of the sampled noise w. Thus, we now can safely follow a similar approach in Kifer et al. (2012). That is, by the concentration of normal distribution, we have with probability at least $1-\delta$,

$$|\langle -w, \psi_{\mathcal{D}'}(t) - \psi_{\mathcal{D}}(t)\rangle| \le \|\psi_{\mathcal{D}'}(t) - \psi_{\mathcal{D}}(t)\|\,\sigma\sqrt{2\log(2/\delta)}.$$

[4] This is why Lemma 16 in Kifer et al. (2012) does not hold in general.

Meanwhile, we have $\|\psi_{\mathcal{D}'}(t) - \psi_{\mathcal{D}}(t)\| \le 2L$ by Assumption 2.1. Putting everything back to (29), yields that with prbability at least $1-\delta$

$$\frac{\mathcal{T}_{1,\mathcal{D}}}{\mathcal{T}_{1,\mathcal{D}'}} \le \exp\left(\frac{2L\sigma\sqrt{8\log(2/\delta)} + (2L)^2}{2\sigma^2}\right) \overset{(a)}{\le} \exp(\varepsilon),$$

where (a) holds if $\sigma \ge \frac{L\sqrt{8\log(2/\delta)+4\varepsilon}}{\varepsilon}$. Combining this with $\mathcal{T}_{2,\mathcal{D}}/\mathcal{T}_{2,\mathcal{D}'} = 1$, yields the required result, hence finishing the proof. □

D.3 Proof of Theorem 5.3

Before presenting the proof, we first introduce the following useful lemma.

Lemma D.2 (Theorm 5.1.1 in Tropp et al. (2015))**.** *Consider a finite sequence $\{X_i\}$ of independent random, symmetric matrices in $\mathbb{R}^d$. Assume that $\lambda_{\min}(X_i) \ge 0$ and $\lambda_{\max}(X_i) \le H$ for each i. Let $Y = \sum_i X_i$ and $\mu_{\min}$ denote the minimum eigenvalue of the expectation $\mathbb{E}[Y]$, i.e., $\mu_{\min} = \lambda_{\min}(\sum_i \mathbb{E}[X_i])$. Then, for any $\varepsilon \in (0,1)$, it holds*

$$\mathbb{P}\left[\lambda_{\min}(Y) \le \varepsilon\mu_{\min}\right] \le d \cdot \exp\left(-(1-\varepsilon)^2 \frac{\mu_{\min}}{2H}\right).$$

Now, we are ready to prove Theorem 5.3.

Proof of Theorem 5.3. Let $\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta) := l_{\mathcal{D}}(\theta) + \frac{\beta}{2n}\|\theta\|_2^2 + \frac{w^\top\theta}{n}$. We divide the proof into the following steps.

Step 1: Let $\Delta := \widehat{\theta}_{\text{obj}} - \theta^*$. Show that $c\|\Delta\| \le \left\|\nabla\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*)\right\|$ for some positive constant c.

To this end, note that we always have

$$\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^* + \Delta) - \widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*) - \langle\nabla\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*), \Delta\rangle \le -\langle\nabla\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*), \Delta\rangle,$$

since $\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^* + \Delta) = \widetilde{\mathcal{L}}_{\mathcal{D}}(\widehat{\theta}_{\text{obj}}) \le \widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*)$ by the optimality of $\widehat{\theta}_{\text{obj}}$. The RHS of above inequality can be upper bounded by $\left\|\nabla\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*)\right\|\|\Delta\|$. Thus, it remains to lower bound the LHS, which motivates us to show that $\widetilde{\mathcal{L}}_{\mathcal{D}}$ is strongly convex with respect to the ℓ_2-norm $\|\cdot\|_2$. That is, we need to show that for all v, θ,

$$v^\top \nabla^2 \widetilde{\mathcal{L}}_{\mathcal{D}}(\theta) v \ge c\|v\|_2^2$$

for some positive constant $c > 0$. Now, let us look at the Hessian matrix of $\widetilde{\mathcal{L}}_{\mathcal{D}}$ at any θ,

$$\nabla^2 \widetilde{\mathcal{L}}_{\mathcal{D}}(\theta) = \frac{\beta}{n}I + \frac{1}{n}\sum_{i=1}^n \sigma(x_i^\top\theta)(1-\sigma(x_i^\top\theta))x_i x_i^\top. \tag{30}$$

To proceed, we will leverage Lemma D.2. In particular, to apply it to our case, we have $X_i = x_i x_i^\top$ with $H = L^2$, $\mu_{\min} = n\kappa$ by Assumptions 2.1 and 4.4. Hence, as a result of Lemma D.2, with probability at least $1-\alpha$,

$$\lambda_{\min}(\sum_i x_i x_i^\top) \ge \frac{n\kappa}{2}, \tag{31}$$

when $n \ge \frac{8L^2\log(d/\alpha)}{\kappa}$. Thus, condition on the good event, plugging (31) into (30) and noting that $\inf_{z\in[-2LB,2LB]} \sigma(z)(1-\sigma(z)) \ge \gamma := \frac{1}{2+\exp(-2LB)+\exp(2LB)}$, yields that

$$v^\top \nabla^2 \widetilde{\mathcal{L}}_{\mathcal{D}}(\theta) v \ge \left(\frac{\beta}{n} + \frac{\kappa\gamma}{2}\right)\|v\|_2^2.$$

Thus, we have so far established that

$$\left(\frac{\kappa\gamma}{2}\right)\|\Delta\| \le \left(\frac{\beta}{n} + \frac{\kappa\gamma}{2}\right)\|\Delta\| \le \left\|\nabla\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*)\right\|, \tag{32}$$

where $\beta > 0$.

Step 2: Bound $\left\|\nabla\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*)\right\|$ with high probability.

To this end, we note that

$$\nabla\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(x_i(\sigma(x_i^\top\theta^*) - y_i)\right)}_{\mathcal{T}_1} + \underbrace{\frac{\beta}{n}\theta^*}_{\mathcal{T}_2} + \underbrace{\frac{w}{n}}_{\mathcal{T}_3}.$$

Thus, we need to bound each of the terms on the RHS. We start with $\mathcal{T}_1$. Let $V_i := \sigma(x_i^\top\theta^*) - y_i$ and hence we have $\mathbb{E}[V_i] = 0$. Thus, we can write $\mathcal{T}_1 = -\frac{1}{n}X^TV$, where $X \in \mathbb{R}^{n\times d}$ is the data matrix and $x_i^\top \in \mathbb{R}^d$ is the i-th row of it. Hence, we have $\|\mathcal{T}_1\|^2 = \frac{1}{n^2}V^\top XX^\top V$. To analyze the concentration for this quadratic form, we will resort to the classic Hanson-Wright inequality. In particular, we will apply the explicit bound in Theorem 2.1 of Hsu et al. (2012). To this end, we need to check the following quantities of $M := \frac{1}{n^2}XX^\top$:

$$\begin{aligned}
&\mathrm{tr}(M) \le 4L^2/n\\
&\mathrm{tr}(M^2) \le 16L^4/n^2\\
&\|M\|_{\mathrm{op}} = \lambda_{\max}(M) \le 4L^2/n,
\end{aligned}$$

where the above inequalities hold by simple linear algebra and the boundedness assumption. Thus, by Theorem 2.1 of Hsu et al. (2012), we have with probability at least $1-\alpha$

$$\|\mathcal{T}_1\|^2 = V^\top MV \le C_1L^2\frac{1+\log(1/\alpha)}{n},$$

where C_1 is some universal constant.

For $\mathcal{T}_2$, we have $\|\mathcal{T}_2\| \le \frac{\beta B}{n}$ by boundedness assumption. For $\mathcal{T}_3$, by the standard concentration of the norm of Gaussian vector (cf. Theorem 3.1.1 in Vershynin (2018)), we have with probability at least $1-\alpha$,

$$\|\mathcal{T}_3\| \le C_3\frac{1}{n}\sigma\left(\sqrt{d} + \sqrt{\log(1/\alpha)}\right),$$

where C_3 is again form universal constant.

Putting all these bounds together and choosing $\beta = \sqrt{n}/B$, yields that with probability at least $1-\alpha$,

$$\left\|\nabla\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*)\right\| \le C\cdot\left(L\sqrt{\frac{1+\log(1/\alpha)}{n}} + \sigma\frac{1}{n}\sqrt{d} + \sigma\frac{1}{n}\sqrt{\log(1/\alpha)}\right), \tag{33}$$

where C is some universal constant.

Step 3: Derive the final bound.

Plugging the bound in (33) into (32), yields

$$\|\Delta\| \le C'\left(\frac{L}{\kappa\gamma}\sqrt{\frac{1+\log(1/\alpha)}{n}} + \frac{\sigma\left(\sqrt{d}+\sqrt{\log(1/\alpha)}\right)}{n\kappa\gamma}\right).$$

Recall that $\sigma = \frac{L\sqrt{8\log(2/\delta)+4\varepsilon}}{\varepsilon}$, and hence we finally have the bound

$$\|\Delta\| \le C'\left(\frac{L}{\kappa\gamma}\sqrt{\frac{1+\log(1/\alpha)}{n}} + \frac{\left(\sqrt{d}+\sqrt{\log(1/\alpha)}\right)}{n\kappa\gamma}\cdot\frac{L\sqrt{8\log(2/\delta)+4\varepsilon}}{\varepsilon}\right).$$

□

D.4 Semi-norm Error Bounds under Central Label DP

In this section, we prove bounds on the estimation error in semi-norm under central label DP.

D.4.1 Lower Bound

We have the following result for the lower bound on the estimation error.

Theorem D.3. *For a large enough n, any estimator $\widehat{\theta}$ based on samples form the BTL model that satisfies $(\varepsilon, 0)$-label DP in the central model has the estimation error in semi-norm lower bounded as*

$$\mathbb{E}\left[\left\|\widehat{\theta} - \theta^*\right\|^2_{\Sigma_{\mathcal{D}}+\lambda I}\right] \geq \Omega\left(\frac{d}{n} + \frac{d^2}{n^2\varepsilon^2}\right).$$

To prove the theorem, we will leverage the following useful result, i.e., DP version of Fano's lemma[5].

Lemma D.4 (Fano's lemma for central DP (Acharya et al., 2021))**.** *Let $\mathcal{V} = \{P_1, P_2, \ldots, P_M\} \subseteq \mathcal{P}$ such that for all $i \neq j$,*

$$D_{\mathrm{KL}}\left(P_i \| P_j\right) \leq \beta\ , \quad \rho'(\theta(P_i), \theta(P_j)) \geq \tau\ .$$

for a semi-metric ρ' and for some $\tau, \beta > 0$. Moreover, let there exists a coupling between P_i and P_j such that $\mathbb{E}\left[d_{\mathrm{ham}}(X, Y)\right] \leq D$ for some $D > 0$. Then, we have

$$R(\mathcal{P}, (\rho')^2, \varepsilon) \geq \max\left\{\frac{\tau^2}{4}\left(1 - \frac{\beta + \log 2}{\log M}\right), 0.2\tau^2 \min\left\{1, \frac{M}{e^{10\varepsilon D}}\right\}\right\}.$$

Now, we are ready to prove Theorem D.3.

Proof of Theorem D.3. The non-private part is the same as before, i.e., the proof for Theorem 3.1. For the private part, we follow the same packing construction as in the proof of Theorem 3.1. Let (X, Y) be the coupling between P_i^n and P_j^n, since n samples are observed. Again, we utilize the maximal coupling property to obtain

$$\begin{aligned}
\mathbb{E}\left[d_{\mathrm{Ham}}(X, Y)\right] &= \sum_{k=1}^n \|P_{i,k} - P_{j,k}\|_{\mathrm{TV}} \\
&\leq \sqrt{n}\sqrt{\sum_{k=1}^n \|P_{i,k} - P_{j,k}\|^2_{\mathrm{TV}}} \\
&\leq \sqrt{n/2}\sqrt{\sum_k D_{\mathrm{KL}}\left(P_{i,k} \| P_{j,k}\right)} \\
&\overset{(a)}{\leq} \sqrt{n/2}\sqrt{n\Delta^2} \leq 1/\sqrt{2}n\Delta := D,
\end{aligned}$$

where (a) follows from (14). Now, noting that $\tau^2 = \Theta(\Delta^2)$, $M = \Theta(e^d)$, letting $\Delta = c \cdot \frac{d}{n\varepsilon}$, we obtain

$$R_c(\mathcal{P}_{\log}, \|\cdot\|^2_{\Sigma_{\mathcal{D}}+\lambda I}, \varepsilon) \geq \Omega\left(\frac{d^2}{n^2\varepsilon^2}\right).$$

□

D.4.2 Upper Bound

Theorem D.5. *Let $\alpha \in (0, 1)$. Then, under Assumptions 2.1 and 4.4, $\widehat{\theta}_{obj}$ satisfies*

$$\left\|\widehat{\theta}_{obj} - \theta^*\right\|_{\Sigma_{\mathcal{D}}+\lambda' I} \leq O\left(\frac{1}{\gamma}\sqrt{\frac{d + \log(1/\alpha)}{n}} + \frac{\sqrt{\sigma}(d\log(1/\alpha))^{1/4}}{\sqrt{n\gamma B}}\right)$$

[5] As before, we correct some constant factor errors in the original statement in Acharya et al. (2021).

with probability at least $1-\alpha$*, where* $\gamma := \frac{1}{2+\exp(-2LB)+\exp(2LB)}$ *and* $\lambda' := \frac{\sigma\sqrt{d\log(1/\alpha)}}{\gamma n B}$*. Thus, setting noise parameter* $\sigma = \frac{L\sqrt{8\log(2/\delta)+4\varepsilon}}{\varepsilon}$*, it satisfies* (ε,δ)*-label DP in the central model and has estimation error*

$$\left\|\widehat{\theta}_{obj} - \theta^*\right\|_{\Sigma_{\mathcal{D}}+\lambda' I} \le O\left(\frac{1}{\gamma}\sqrt{\frac{d+\log(1/\alpha)}{n}} + \frac{\sqrt{L}\left((\log(2/\delta)+4\varepsilon)d\log(1/\alpha)\right)^{1/4}}{\sqrt{n\varepsilon\gamma B}}\right).$$

Proof. Let $\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta) := l_{\mathcal{D}}(\theta) + \frac{\beta}{2n}\|\theta\|_2^2 + \frac{w^\top\theta}{n}$. We divide the proof into the following steps.

Step 1: Let $\Delta := \widehat{\theta}_{\text{obj}} - \theta^*$. Show that $c\|\Delta\|_{\Sigma_{\mathcal{D}}}^2 \le \left\|\nabla\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*)\right\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} \|\Delta\|_{\Sigma_{\mathcal{D}}+\lambda I}$, for some positive constant c.

To this end, note that we always have

$$\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*+\Delta) - \widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*) - \langle\nabla\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*),\Delta\rangle \le -\langle\nabla\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*),\Delta\rangle,$$

since $\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*+\Delta) = \widetilde{\mathcal{L}}_{\mathcal{D}}(\widehat{\theta}_{\text{obj}}) \le \widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*)$ by the optimality of $\widehat{\theta}_{\text{obj}}$. The RHS of above inequality can be upper bounded by $\left\|\nabla\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*)\right\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} \|\Delta\|_{\Sigma_{\mathcal{D}}+\lambda I}$ for any $\lambda > 0$. Thus, it remains to lower bound the Hessian. That is, we aim to show that for all $v, \theta \in \Theta_B$,

$$v^\top\nabla^2\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta)v \ge c\|v\|_{\Sigma_{\mathcal{D}}}^2$$

for some positive constant $c > 0$. By definition, the Hessian of $\widetilde{\mathcal{L}}_{\mathcal{D}}$ at any $\theta \in \Theta_B$ is

$$\begin{aligned}\nabla^2\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta) &= \frac{\beta}{n}I + \frac{1}{n}\sum_{i=1}^n \sigma(x_i^\top\theta)(1-\sigma(x_i^\top\theta))x_i x_i^\top \\ &\ge \gamma\|v\|_{\Sigma_{\mathcal{D}}+\lambda' I}^2,\end{aligned}$$

where the inequality follows from $\beta > 0$ and $\inf_{z\in[-2LB,2LB]}\sigma(z)(1-\sigma(z)) \ge \gamma := \frac{1}{2+\exp(-2LB)+\exp(2LB)}$ and $\lambda' := \beta/(\gamma n)$ Thus, for all Δ such that $\theta^*+\Delta \in \Theta_B$, by Taylor expansion, we have

$$\frac{\gamma}{2}\|\Delta\|_{\Sigma_{\mathcal{D}}+\lambda' I}^2 \le \left\|\nabla\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*)\right\|_{(\Sigma_{\mathcal{D}}+\lambda' I)^{-1}} \|\Delta\|_{\Sigma_{\mathcal{D}}+\lambda' I}.$$

Step 2: Bound $\left\|\nabla\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*)\right\|_{(\Sigma_{\mathcal{D}}+\lambda' I)^{-1}}$ with high probability.

To this end, we note that

$$\nabla\widetilde{\mathcal{L}}_{\mathcal{D}}(\theta^*) = \underbrace{\frac{1}{n}\sum_{i=1}^n\left(x_i(\sigma(x_i^\top\theta^*)-y_i)\right)}_{\mathcal{T}_1} + \underbrace{\frac{\beta}{n}\theta^*}_{\mathcal{T}_2} + \underbrace{\frac{w}{n}}_{\mathcal{T}_3}.$$

By the same analysis as in Zhu et al. (2023), we have with probability at least $1-\alpha$

$$\|\mathcal{T}_1\|_{(\Sigma_{\mathcal{D}}+\lambda' I)^{-1}} \le C\cdot\sqrt{\frac{d+\log(1/\alpha)}{n}},$$

for some absolute constant C. For $\mathcal{T}_2$, we have

$$\|\mathcal{T}_2\|_{(\Sigma_{\mathcal{D}}+\lambda' I)^{-1}} \le \frac{\beta}{n\sqrt{\lambda'}}\|\theta^*\| \le \frac{\beta B}{n\sqrt{\lambda'}}.$$

For $\mathcal{T}_3$, by the concentration of the norm of the Gaussian vector, we have with probability at least $1-\alpha$,

$$\|\mathcal{T}_3\|_{(\Sigma_{\mathcal{D}}+\lambda' I)^{-1}} \le \frac{1}{n\sqrt{\lambda'}}\|w\| \le O\left(\frac{\sigma}{n\sqrt{\lambda'}}\left(\sqrt{d}+\sqrt{\log(1/\alpha)}\right)\right).$$

Putting all of them together, we have

$$\frac{\gamma}{2}\left\|\Delta\right\|_{\Sigma_{\mathcal{D}}+\lambda' I}^2 \le C'\left(\sqrt{\frac{d+\log(1/\alpha)}{n}} + \frac{\beta B}{n\sqrt{\lambda'}} + \frac{\sigma\sqrt{d\log(1/\alpha)}}{n\sqrt{\lambda'}}\right)\left\|\Delta\right\|_{\Sigma_{\mathcal{D}}+\lambda' I},$$

which directly implies that

$$\left\|\Delta\right\|_{\Sigma_{\mathcal{D}}+\lambda' I} \le C_1\cdot\frac{1}{\gamma}\sqrt{\frac{d+\log(1/\alpha)}{n}} + C_1\frac{1}{\gamma}\left(\frac{\beta B}{n\sqrt{\lambda'}} + \frac{\sigma\sqrt{d\log(1/\alpha)}}{n\sqrt{\lambda'}}\right).$$

Thus, choosing $\lambda' = \frac{\sigma\sqrt{d\log(1/\alpha)}}{\gamma n B}$ (i.e., $\beta = \frac{\sigma\sqrt{d\log(1/\alpha)}}{B}$), yields that

$$\left\|\Delta\right\|_{\Sigma_{\mathcal{D}}+\lambda' I} \le O\left(\frac{1}{\gamma}\sqrt{\frac{d+\log(1/\alpha)}{n}} + \frac{\sqrt{\sigma}(d\log(1/\alpha))^{1/4}}{\sqrt{n\gamma B}}.\right)$$

Finally, plugging in noise value $\sigma = \frac{L\sqrt{8\log(2/\delta)+4\varepsilon}}{\varepsilon}$, yileds our final result

$$\left\|\widehat{\theta}_{\mathrm{obj}} - \theta^*\right\|_{\Sigma_{\mathcal{D}}+\lambda' I} \le O\left(\frac{1}{\gamma}\sqrt{\frac{d+\log(1/\alpha)}{n}} + \frac{\sqrt{L}\left((\log(2/\delta)+4\varepsilon)d\log(1/\alpha)\right)^{1/4}}{\sqrt{n\varepsilon\gamma B}}\right).$$

Note that the privacy guarantee follows the same as before, hence completing the proof. □

E Generalization to Standard DP

In the main paper, we mainly focus on protecting the labels via label DP, which is well-motivated by many practical situations. It turns out that our technique can also be generalized to handle privacy protection of both features and labels, i.e., the standard DP notion.

We start with the central model. Since objective perturbation (Kifer et al., 2012) was originally proposed to achieve standard DP in the central model, it would be natural to adopt it in our case. However, as before, we cannot directly employ the results in Kifer et al. (2012) to prove privacy guarantee due to the gap in their Lemma 16. Instead, we found that for log loss, one can get rid of the independence issue in their Lemma 16, and hence establish the privacy guarantee with the same order of Gaussian noise. This is not true in general for arbitrary convex losses (where an additional $\sqrt{d}$ factor is required), as also observed in Agarwal et al. (2023).

Privacy. In the following, we will show that with minor constant changes in the noise parameter of Theorem 5.2, Algorithm 2 also achieves standard DP in the central model, i.e., the neighboring relation is now about a change of (x_i, y_i) rather than only y_i under label DP as considered in Theorem 5.2.

Theorem E.1 (Privacy under standard DP). *Let $\varepsilon > 0$, $\delta \in (0,1)$ and Assumption 2.1 hold. Then, setting $\sigma \ge \frac{4L\sqrt{8\log(4/\delta)+2\varepsilon}}{\varepsilon}$ and $\beta \ge \frac{4L^2}{\varepsilon}$, Algorithm 2 satisfies (ε,δ)-DP in the central model.*

Proof. As in the proof of Theorem 5.2, we will use two Successive Approximations, which allows us to only focus on the following problem

$$\widetilde{\theta}_{\mathcal{D}} = \operatorname*{argmin}_{\theta\in\mathbb{R}^d} l_{\mathcal{D}}(\theta) + \frac{\beta}{2n}\left\|\theta\right\|_2^2 + \frac{w^\top\theta}{n} + \frac{1}{n}r(\theta).$$

Also, as before, we have the following PDF.

$$f_{\widetilde{\theta}_{\mathcal{D}}}(t) = C\underbrace{\exp\left(-\frac{\left\|\psi_{\mathcal{D}}(t)\right\|_2^2}{2\sigma^2}\right)}_{\mathcal{T}_{1,\mathcal{D}}}\cdot\underbrace{\left|\det\left[\frac{d\psi_{\mathcal{D}}(\theta)}{d\theta}|_{\theta=t}\right]\right|}_{\mathcal{T}_{2,\mathcal{D}}} \tag{34}$$

We are again left to bound the two ratios. To this end, we first look at the ratio of $\mathcal{T}_{2,\mathcal{D}}/\mathcal{T}_{2,\mathcal{D}'}$. Note that the matrix inside the determinant in (34) is the Hessian of $l_{\mathcal{D}}$ plus some common terms, given by

$$A_{\mathcal{D}} := \nabla^2 r(\theta)|_{\theta=t} + \beta I + \nabla^2 l_{\mathcal{D}}(\theta)|_{\theta=t} = \nabla^2 r(\theta)|_{\theta=t} + \beta I + \sum_{i=1}^{n} \sigma(x_i^\top t)(1-\sigma(x_i^\top t))x_i x_i^\top,$$

which now depends on x_i. Hence, we need some additional steps to bound this ratio under standard DP. In particular, we define $E := A_{\mathcal{D}} - A_{\mathcal{D}'} = \sigma(x_s^\top t)(1-\sigma(x_s^\top t))x_s x_s^\top - \sigma(x_s'^\top t)(1-\sigma(x_s'^\top t))x_s' x_s'^\top$, where $\mathcal{D}, \mathcal{D}'$ differs in one single sample at index s. Thus, the rank of E is most two and moreover the sum of largest and second largest eigenvalue of E satisfies

$$|\lambda_1(E)| + |\lambda_2(E)| \le \frac{1}{4} \cdot 4L^2 + \frac{1}{4} \cdot 4L^2 = 2L^2,$$

where we have used the boundedness assumption. This also implies that

$$|\lambda_1(E)| \cdot |\lambda_2(E)| \le L^4.$$

To proceed, we will leverage the following result.

Claim E.2 (Lemma 10 in Chaudhuri et al. (2011))**.** *If A is full rank and if E has rank at most 2, then*

$$\frac{\det(A+E) - \det(A)}{\det(A)} = \lambda_1(A^{-1}E) + \lambda_2(A^{-1}E) + \lambda_1(A^{-1}E) \cdot \lambda_2(A^{-1}E),$$

where $\lambda_j(Z)$ is the j-th largest eigenvalue of matrix Z.

Note that for $j = 1, 2$, $|\lambda_j(A_{\mathcal{D}'}^{-1}E)| \le \frac{|\lambda_j(E)|}{\beta}$ due to the fact that the minimal eigenvalue of $A_{\mathcal{D}'}$ is at least β. Thus, by Claim E.2, we have

$$\begin{aligned}
\frac{\mathcal{T}_{2,\mathcal{D}}}{\mathcal{T}_{2,D'}} = \frac{|\det(A_{\mathcal{D}'} + E)|}{|\det(A_{\mathcal{D}'})|} &= \left|1 + \lambda_1(A_{\mathcal{D}'}^{-1}E) + \lambda_2(A_{\mathcal{D}'}^{-1}E) + \lambda_1(A_{\mathcal{D}'}^{-1}E) \cdot \lambda_2(A_{\mathcal{D}'}^{-1}E)\right| \\
&\le 1 + \frac{2L^2}{\beta} + \frac{L^4}{\beta^2} \\
&= \left(1 + \frac{L^2}{\beta}\right)^2 \\
&\le e^{2L^2/\beta}.
\end{aligned}$$

Thus, when $\beta \ge \frac{4L^2}{\varepsilon}$, we have $\frac{\mathcal{T}_{2,\mathcal{D}}}{\mathcal{T}_{2,D'}} \le e^{\varepsilon/2}$.

Now, let us turn to bound $\frac{\mathcal{T}_{1,\mathcal{D}}}{\mathcal{T}_{1,D'}}$. In particular, we have

$$\begin{aligned}
\frac{\mathcal{T}_{1,\mathcal{D}}}{\mathcal{T}_{1,\mathcal{D}'}} &= \exp\left(\frac{\|\psi_{\mathcal{D}'}(t)\|_2^2 - \|\psi_{\mathcal{D}}(t)\|_2^2}{2\sigma^2}\right) \\
&= \exp\left(\frac{2\langle \psi_{\mathcal{D}}(t), \psi_{\mathcal{D}'(t)} - \phi_{\mathcal{D}}(t)\rangle + \|\psi_{\mathcal{D}'}(t) - \psi_{\mathcal{D}}(t)\|_2^2}{2\sigma^2}\right) \\
&= \exp\left(\frac{2\langle -w, \psi_{\mathcal{D}'(t)} - \psi_{\mathcal{D}}(t)\rangle + \|\psi_{\mathcal{D}'}(t) - \psi_{\mathcal{D}}(t)\|_2^2}{2\sigma^2}\right).
\end{aligned} \tag{35}$$

where we know that $\psi_{\mathcal{D}}(t) = -w$, which is distributed according to a normal. However, one needs to be careful here to show that $\psi_{\mathcal{D}'}(t) - \psi_{\mathcal{D}}(t)$ is *independent* of w so that one can claim that the inner product is also distributed according to a normal. In our case, for two neighboring datasets $\mathcal{D}, \mathcal{D}'$ that differs only in (x_s, y_s) and (x_s', y_s') we have

$$\psi_{\mathcal{D}'}(t) - \psi_{\mathcal{D}}(t) = \left(\frac{1}{1 + \exp(-\langle x_s', t\rangle)} - y_s'\right) x_s' - \left(\frac{1}{1 + \exp(-\langle x_s, t\rangle)} - y_s\right) x_s.$$

Then, we have

$$|\langle -w, \psi_{\mathcal{D}'}(t) - \psi_{\mathcal{D}}(t)\rangle| \leq |\langle w, x_s'\rangle| + |\langle w, x_s\rangle|,$$

which combined with the concentration of normal distribution and boundedness of x_s, x_s', leads to that with probability at least $1-\delta$,

$$|\langle -w, \phi_{\mathcal{D}'}(t) - \phi_{\mathcal{D}}(t)\rangle| \leq 4L\sigma\sqrt{2\log(4/\delta)}.$$

Meanwhile, we have $\|\psi_{\mathcal{D}'}(t) - \psi_{\mathcal{D}}(t)\| \leq 4L$ by Assumption 2.1

Putting everything back to (35), yields that with probability at least $1-\delta$

$$\frac{\mathcal{T}_{1,\mathcal{D}}}{\mathcal{T}_{1,\mathcal{D}'}} \leq \exp\left(\frac{2L\sigma\sqrt{8\log(4/\delta)} + (4L)^2}{2\sigma^2}\right) \overset{(a)}{\leq} \exp(\varepsilon/2),$$

where (a) holds if $\sigma \geq \frac{4L\sqrt{8\log(4/\delta)+2\varepsilon}}{\varepsilon}$. Combining this with $\mathcal{T}_{2,\mathcal{D}}/\mathcal{T}_{2,\mathcal{D}'} = e^{\varepsilon/2}$, yields the required result, hence finishing the proof. □

Utility. For the estimation error under ℓ_2 norm, one can follow the same proof of Theorem 5.3. One difference is to remember to check the condition of β in Theorem E.1, which can be satisfied by conditions on n and ε. For the estimation error in semi-norm, one needs additional steps compared to the proof of Theorem D.5, since now it needs to establish the concentration of $\|\cdot\|_{\widetilde{\Sigma}_{\mathcal{D}}+\lambda I}$, where $\widetilde{\Sigma}_{\mathcal{D}}$ is the private covariance matrix. First, one can privatize the covariance matrix $\Sigma_{\mathcal{D}}$ via the standard Gaussian mechanism. Then, to guarantee a semi-positive nature of $\widetilde{\Sigma}_{\mathcal{D}} + \lambda I$, one needs to choose λ properly, which can be done by following the routine in previous DP linear bandits (see Shariff and Sheffet (2018); Chowdhury and Zhou (2022)). Finally, one can translate the concentration $\|\cdot\|_{\Sigma_{\mathcal{D}}+\lambda I}$ in Theorem D.5 to $\|\cdot\|_{\widetilde{\Sigma}_{\mathcal{D}}+\lambda I}$ in Theorem D.5 via standard Gaussian concentration and the property of linear summation of Gaussian. Ignoring all other factors (γ, B, L), the final cost of privacy should be on the order of $\frac{(d\log(1/\delta))^{1/4}}{\sqrt{n\varepsilon}}$ for $\varepsilon \in (0,1)$. One subtlety again is that one needs to check β satisfies the condition in Theorem E.1.

Remark E.3 (Remark on local DP). In contrast to local label DP in the main paper, establishing local standard DP is challenging in our offline reward estimation setting, which is *non-interactive*. This is different from interactive online logistic regression in Duchi et al. (2018). In fact, it is in general not straightforward to derive an efficient algorithm even for ERM under the non-interactive setting Smith et al. (2017), let alone the parameter estimation problem in our setting. We leave it to one of our future research directions.