Differentially Private Reward Estimation from Preference Based Feedback

Sayak Ray Chowdhury [*1] **Xingyu Zhou** [*2]

Abstract

Preference-based reinforcement learning (RL) has gained attention as a promising approach to align learning algorithms with human interests in various domains. Instead of relying on numerical rewards, preference-based RL uses feedback from human labelers in the form of pairwise or K-wise comparisons between actions. In this paper, we focus on reward learning in preference-based RL and address the issue of estimating unknown parameters while protecting privacy. We propose two estimators based on the Randomized Response strategy that ensure label differential privacy. The first estimator utilizes maximum likelihood estimation (MLE), while the second estimator employs stochastic gradient descent (SGD). We demonstrate that both estimators achieve an estimation error of $\widetilde{O}(1/\sqrt{n})$ with n number of samples. The additional cost of ensuring privacy for human labelers is proportional to $\frac{e^\varepsilon+1}{e^\varepsilon-1}$ in the best case, where $\varepsilon > 0$ is the privacy budget.

1. Introduction

In an increasing range of applications in modern machine earning, it is of interest to elicit judgments / ratings / feedbacks from humans (Green et al., 1981). For example, in marketing applications, it is common practice to elicit the preferences of consumers about different products. The most used method of preference elicitation is through pairwise comparisons (Shah et al., 2015). For instance, if a consumer chooses one product over another, then it constitutes a pairwise comparison between these two products. Gathering of this comparison data has greatly been facilitated by crowdsourcing platforms like Amazon Mechanical Turk (Khatib et al., 2011). Human workers in crowdsourcing setups like this are often asked to compare pairs of items such as rating the performance of two players in a competitive game (Herbrich et al., 2006), identifying the better of two possible results of an online search engine (Kazai, 2011) etc. The pairwise comparisons can be thought as means of estimating the underlying weights of the items being compared such as skill of players, relevance of search results etc. However, these comparisons are subject to getting corrupted by some noise. For example, noise can arise from the differing levels of expertise of crowd workers. Hence, an important questions is to estimate the latent weights based on noisy data in the form of pairwise comparisons (Shah et al., 2015).

Recently, the AI alignment problem has garnered a lot of interest, where the goal is to steer learning algorithms towards the interest of humans (Glaese et al., 2022). One of the most promising approaches to achieve this is via preference-based reinforcement learning (Christiano et al., 2017), which has gained considerable attention across multiple application domains, including game playing (MacGlashan et al., 2017), large language models (Ouyang et al., 2022) and robot learning (Shin et al., 2023). In the standard RL setting, the agent learns to maximize a numerical reward, which she observes from the environment. However, observing appropriate numerical rewards can often be challenging in the above applications, which could significantly affect the performance of RL algorithms. Preference-based RL with human feedback are able to tackle this effectively (Zhu et al., 2023; Zhan et al., 2023; Pacchiano et al., 2021; Chen et al., 2022).

In preference-based RL, the agent does not receive a numerical reward, instead at every state she receives a feedback from a human labeler in the form of pairwise or K-wise comparisons between actions at a given state. Notably, the language model application InstructGPT (Christiano et al., 2017; Ouyang et al., 2022) is based on this reward model — the comparisons depend purely on the current prompt, which corresponds to the state in a contextual bandit environment (a degenerate RL environment). These comparisons are deployed to learn a reward function based on a pre-trained model, which is then used for downstream policy training (i.e., finetuning the existing pre-trained model). Our focus, in this work, is on reward learning. First, the prompts are first sampled from a pre-collected dataset, and then, for each prompt, a pair of (or K) responses are sampled by executing the pre-trained model. A human labeler then ranks all the responses according to her own preference and based on the current prompt. Finally, the reward model is trained by maximum likelihood estimation, or, equivalently, by cross-entropy minimization.

[*] Equal contributions. [1] Microsoft Research, India. [2] Wayne State University, USA

One important aspect which is ignored in prior literature is protecting private information of crowd workers, which might get revealed from pairwise comparisons provided by them. In fact, after the emergence of ChatGPT several instances of privacy breach including that of human labelers have been reported (Li et al., 2023) and henceforth, efforts have been made to privately fine-tune large language models (Yu et al., 2021; Behnia et al., 2022). In view of this, Differential privacy (DP) (Dwork, 2008) is the most adopted notion to protect the sensitive (private) information of individuals whose data is used during the model training. In the setting of reward training in language models, the output of comparison between a pair of actions or, equivalently, the label is considered sensitive since it can reveal private information (preference) of the human labeler. However, the prompts or the states are not considered sensitive information since they are sampled from a pre-collected dataset. This can be captured by the notion of label differential privacy (Label-DP), which has been studied in the PAC setting (Chaudhuri & Hsu, 2011; Beimel et al., 2013) and in the context of deep learning (Ghazi et al., 2021). Apart from the above example of language models, label-DP captures several other practical scenarios. For example, in recommendations systems the items are known to the service provider but the user ratings or clicks reveal user interest. In computational advertising, the impressions are non-sensitive, but the conversions are considered sensitive information.

Our contributions. In this work, we are interested in the sample complexity for learning a reward model from pairwise comparison data under the constraint of label differential privacy. We assume that the reward is linearly parameterized by a weight vector, which is unknown and needs to be learned. We design two estimators based on the *Randomized Response* strategy (Warner, 1965), which ensures label DP for both the estimators. Our first estimator is based on maximum likelihood estimation (MLE) principle, while the second estimator employs stochastic gradient descend (SGD) strategy. We prove that the estimation error for both the estimators goes down as $\widetilde{O}(1/\sqrt{n})$ with the number of samples n, while the cost of ensuring privacy of human labelers is of a multiplicative factor $\frac{e^\varepsilon+1}{e^\varepsilon-1}$ in the best case, where $\varepsilon > 0$ is the privacy budget.

2. Preliminaries

We consider the problem of parameter estimation from preference-based feedback under privacy constraints. Specifically, the preference-based dataset $\mathcal{D} = (s_i, a_i^0, a_i^1, y_i)_{i=1}^n$ consists of n samples, each has one context/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/preference feedback $y_i \in \{0,1\}$ indicating which action is preferred by humans or domain experts. As in Zhu et al. (2023), 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 $l \in \{0,1\}$,

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

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 & Terry, 1952; Luce, 2012).

In this paper, we consider a linear reward model $r_{\theta^*}(s,a) = \phi(s,a)^\top \theta^*$, where $\phi(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ 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

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

where $\sigma(z) = \frac{1}{1+e^{-z}}$ denotes the sigmoid function. For notation simplicity, we let $x_i := \phi(s_i, a_i^1) - \phi(s_i, a_i^0)$ denote the differential feature corresponding to actions a_i^1 and a_i^0. With this notation, we have

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

Throughout the paper, we make the following assumption.

Assumption 2.1 (Boundedness)**.** We assume that θ^* lies in the set $\Theta_B = \{\theta \in \mathbb{R}^d | \langle \mathbf{1}, \theta \rangle = 0, \|\theta\| \leq B\}$. Furthermore, the features are bounded, i.e., $\|\phi(s,a)\| \leq L \; \forall (s,a)$.

These assumptions are standard in the literature (Shah et al., 2015; Zhu et al., 2023). We need the condition $\langle \mathbf{1}, \theta \rangle = 0$ to ensure identifiability of θ^*.

Differential Privacy. First, we recall the definition of differential privacy (DP), which is applicable to any notion of dataset (Dwork, 2008).

Definition 2.2 (DP)**.** Let $\varepsilon \geq 0, \delta \in (0,1]$. A mechanism $\mathcal{M}$ is said to be (ε, δ)-differentially private (DP) if for any two datasets $\mathcal{D}, \mathcal{D}'$ that differ in one single example and for any subset E of the outputs of $\mathcal{M}$, it holds that

$$\mathbb{P}\left[\mathcal{M}(\mathcal{D}) \in E\right] \leq e^\varepsilon \cdot \mathbb{P}\left[\mathcal{M}(\mathcal{D}') \in E\right] + \delta .$$

If $\delta = 0$, $\mathcal{M}$ is said to be ε-DP.

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. In our context, label DP roughly means that any single change of feedback label will not change the final estimator too much, which is what we formalize below.

Definition 2.3 (Label DP). Let $\varepsilon \geq 0, \delta \in (0, 1]$. A randomized algorithm $\mathcal{A}$ is said to be (ε, δ)-label differentially private if for any two datasets $\mathcal{D}$ and $\mathcal{D}'$ that differ in the label of a single sample and for any subset S of the outputs of $\mathcal{A}$, it holds that

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

If $\delta = 0$, $\mathcal{A}$ is said to be ε-label DP.

Performance measure. In this work, we aim to come up with a candidate estimator $\widehat{\theta}$ of the unknown parameter θ^*, which satisfies label DP. The error of this estimator is typically measured by computing its Euclidean distance from θ^*, i.e. the estimation error is given by $\left\|\widehat{\theta} - \theta^*\right\|_2$. In some applications, however, it also makes sense to compute an weighted Euclidean distance or semi-norm $\left\|\widehat{\theta} - \theta^*\right\|_{\Sigma_{\mathcal{D}}}$, where $\Sigma_{\mathcal{D}}$ is some suitable p.s.d. matrix constructed using feature vectors $\phi(s, a)$ from the dataset $\mathcal{D}$.

3. Private Estimation in Semi-Norm

In this section, we introduce a private maximum likelihood estimator (MLE) of the unknown parameter θ^* and bound its estimation error w.r.t. the semi-norm. We first discuss the Randomized Response (RR) mechanism (Warner, 1965), which we use to guarantee label differential privacy. 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

$$\mathbb{P}[\widetilde{y} = y] = \frac{e^{\varepsilon}}{1 + e^{\varepsilon}} \text{ and } \mathbb{P}[\widetilde{y} \neq y] = \frac{1}{1 + e^{\varepsilon}}\,. \tag{2}$$

It is easy to show that RR is ε-DP (Dwork, 2008).

(1) and (2) together imply that each randomized label $\widetilde{y}_i$ is distributed according to the conditional probabilities

$$\mathbb{P}[\widetilde{y}_i = 1 | x_i] = \frac{1 + e^{-\varepsilon} e^{-x_i^\top \theta^*}}{(1 + e^{-x_i^\top \theta^*})(1 + e^{-\varepsilon})}, \tag{3}$$

$$\mathbb{P}[\widetilde{y}_i = 0 | x_i] = \frac{e^{-\varepsilon} + e^{-x_i^\top \theta^*}}{(1 + e^{-x_i^\top \theta^*})(1 + e^{-\varepsilon})}. \tag{4}$$

With n such pairs of features and randomized labels $(x_i, \widetilde{y}_i)_{i=1}^n$, we compute the MLE, defined $\widehat{\theta}_{\text{MLE-RR}}$, which aims to minimize the negative (conditional) log-likelihood, i.e., $\widehat{\theta}_{\text{MLE-RR}} \in \operatorname{argmin}_{\theta \in \Theta_B}\, l_{\mathcal{D},\varepsilon}(\theta)$, where

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

The privacy guarantee of this estimator follows immediately from that of RR due to post-processing property of DP (Dwork, 2008).

Lemma 3.1 (Privacy of MLE with RR). *For any $\varepsilon \geq 0$, $\widehat{\theta}_{\text{MLE-RR}}$ is ε-label DP.*

We now bound the estimation error of this MLE conditioned on the observed contexts $(s_i)_{i=1}^n$ and queried action pairs $(a_i^0, a_i^1)_{i=1}^n$. We define the sample covariance matrix of differential features as $\Sigma_{\mathcal{D}} = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ and bound the weighted distance $\|\widehat{\theta}_{\text{MLE-RR}} - \theta^*\|_{\Sigma_{\mathcal{D}} + \lambda I}$ for a given $\lambda > 0$.

Theorem 3.2 (Error of estimation in semi-norm). *Fix $\delta \in (0, 1), \varepsilon > 2LB, \lambda > 0$. Then, under Assumption 2.1, with probability at least $1 - \delta$, we have*

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

where C is some absolute constant.

Proof of this result is deferred to Appendix A. Some observations are in order with this result.

Cost of Privacy. First, we compare the error of our private estimator $\widehat{\theta}_{\text{MLE-RR}}$ with the error of the non-private estimator $\widehat{\theta}_{\text{MLE}}$ of Zhu et al. (2023). $\widehat{\theta}_{\text{MLE}}$ minimizes the loss function

$$l_{\mathcal{D}}(\theta) = -\frac{1}{n}\sum_{i=1}^n \Big[\mathbb{1}(y_i = 1) \log \frac{1}{1 + e^{-\theta^\top x_i}} + \mathbb{1}(y_i = 0) \log \frac{e^{-\theta^\top x_i}}{1 + e^{-\theta^\top x_i}}\Big]\,, \tag{5}$$

an achieves an error of estimation of the order $O\big(\sqrt{d/n}\big)$ in the semi-norm. Comparing this with the estimation error of $\widehat{\theta}_{\text{MLE-RR}}$, we observe that the cost of ensuring label DP is a multiplicative factor of the order $O\left(\frac{e^{\varepsilon + LB} + 1}{e^{\varepsilon - 2LB} - 1}\right)$.

Furthermore, the above bound on the estimation error of $\widehat{\theta}_{\text{MLE-RR}}$ holds only when the privacy budget is higher than a certain threshold (which depends on the norm of θ^* and features 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 $\widehat{\theta}_{\text{MLE-RR}}$ minimizes a privacy modulated loss function $l_{\mathcal{D},\varepsilon}(\theta)$, which 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.

Note that Theorem 3.2 immediately implies a bound on the estimation error in ℓ_2-norm.

Corollary 3.3. *Under the same hypothesis of Theorem 3.2, we have, with probability at least $1-\delta$,*

$$\|\widehat{\theta}_{\textit{MLE-RR}} - \theta^*\|_2 \le \frac{C}{\sqrt{\lambda}} \frac{e^{\varepsilon+LB}+1}{e^{\varepsilon-2LB}-1}\sqrt{\frac{d+\log(1/\delta)}{n}} + B\,.$$

As mentioned above, this guarantee only holds for lower privacy regimes, i.e., when $\varepsilon > 2LB$. In the next section, we show that under a coverage assumption on the state-action feature space, one can design a private estimator, whose error guarantee holds for any $\varepsilon > 0$ and which achieves a lesser estimation error than MLE in the ℓ_2-norm. However, in applications such as offline linear contextual bandits (Zhu et al., 2023; Li et al., 2022), where coverage on the entire state-action space is rarely feasible, it makes sense to bound the estimation error in the semi-norm $\|\cdot\|_{\Sigma_{\mathcal{D}}}$. This bound can then be used to learn a downstream pessimistic policy(i.e. an action selection strategy). The pessimistic learning rule selects a policy as

$$\widehat{\pi}_{\Theta} = \operatorname*{argmax}_{\pi\in\Pi} \inf_{\theta\in\Theta} \mathbb{E}_{s\sim\rho}\left[\phi(s,\pi(s))^\top\theta\right]. \qquad (6)$$

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.,

$$\Theta = \left\{\theta \in \Theta_B : \|\widehat{\theta}_{\text{MLE-RR}} - \theta\|_{\Sigma_{\mathcal{D}}+\lambda I} \le f(\varepsilon,\delta,d,n,\lambda)\right\},$$

where $f(\cdot)$ denotes the estimation error of $\widehat{\theta}_{\text{MLE-RR}}$ as given in Theorem 3.2. Similar to Li et al. (2022), one can show that this pessimistic policy achieves a *sub-optimality gap* of the order $O(L\cdot f(\varepsilon,\delta,d,n,\lambda)\left\|(\Sigma_{\mathcal{D}}+\lambda I)^{-1/2}\right\|)$, while guaranteeing label DP.

4. Private Estimation in ℓ_2 Norm

Our main algorithm for reward estimation under ℓ_2-norm is given by Algorithm 1, which can be viewed as one particular instantiation (i.e., with log loss) of Algorithm 5 proposed in Ghazi et al. (2021). The key difference is that Ghazi et al. (2021) focus on establishing the population risk bound for general stochastic convex optimization under the label DP, while we aim to establish a high probability concentration bound for the parameter estimate.

Algorithm 1 runs one-pass SGD over the entire data set $\mathcal{D}$ with private labels only. In particular, at each iteration t, the algorithm first uses Randomized Response (RR) to privatize the label (line 5). Then, it computes the noisy gradient based on noisy label $\tilde{y}_t$ and performs shifting and scaling to obtain $\widehat{g}_t$, which can be shown to be an unbiased estimate of the true gradient. Finally, it performs a standard SGD update. We denote the estimator returned by Algorithm, 1 as $\widehat{\theta}_{\text{SGD-RR}}$.

Algorithm 1 SGD with Randomized Response

```
1: Parameters: privacy budget $\varepsilon$; i.i.d dataset $\mathcal{D} = (x_i, y_i)_{i=1}^n$; parameter space $\Theta_B$; log loss $\ell$
2: Initialize: $\theta_1 = 0$
3: for $t = 1, \dots, n$ do
4:     Take data point $(x_t, y_t)$ from the dataset $\mathcal{D}$
5:     Let $\tilde{y}_t$ be the output of RR mechanism on $y_t$, i.e.,
           $\mathbb{P}[\tilde{y}_t = y_t] = \frac{e^\varepsilon}{1+e^\varepsilon}$ and $\mathbb{P}[\tilde{y}_t \neq y_t] = \frac{1}{1+e^\varepsilon}$
6:     Compute the gradient $\tilde{g}_t = \nabla_\theta \ell(\theta_t, (x_t, \tilde{y}_t))$ and let
           $\widehat{g}_t = \frac{e^\varepsilon+1}{e^\varepsilon-1} \cdot \left(\tilde{g}_t - \frac{\sum_{l=0}^{1} \nabla_\theta \ell(\theta_t, (x_t, l))}{e^\varepsilon+1}\right)$
7:     Update the estimate $\theta_{t+1} = \Pi_{\Theta_B}(\theta_t - \eta_t \widehat{g}_t)$
8: end for
9: Output $\widehat{\theta}_{\text{SGD-RR}} = \theta_{n+1}$
```

The privacy guarantee of Algorithm 1 follows directly from Randomized Response (Warner, 1965).

Lemma 4.1 (Privacy of SGD with RR)**.** *For any $\varepsilon \ge 0$, $\widehat{\theta}_{\textit{SGD-RR}}$ is ε-label DP.*

Remark 4.2 (Central vs. Local Label DP)*.* Our current definition of label DP follows from the standard one in Ghazi et al. (2021), which implicitly considers a central trust model. That is, the learning agent has access to non-private raw data of human labelers. It is worth noting that our Algorithm 1 also works under the stronger local model where the learning agent only has access to private labels. To achieve this, one can simply replace lines 4-5 in Algorithm 1 by requiring each labeler t to privatize her label y_t using RR before sending it to the learning agent.

In the following, we will establish that the final output of Algorithm 1 (i.e., $\widehat{\theta}_{\text{SGD-RR}}$) is close to the true parameter θ^* in ℓ_2 norm with high probability. In fact, our concentration result holds for all t, i.e., for all intermediate parameter estimates. To establish our result, we need the following coverage assumption on the state-action feature space, which is standard for offline bandits and RL, see Yin et al. (2022). To begin with, we define the population covariance matrix of differential state-action features

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

Assumption 4.3 (Coverage of state-action space)**.** There exists a $\kappa > 0$, such that the data distributions ρ, μ satisfy the minimum eigenvalue condition $\lambda_{\min}(\Sigma) \ge \kappa$.

Note that the coverage parameter κ implicitly depends on the parameter dimension d, and hence, it is a problem-dependent quantity (Wang et al., 2020).

The following theorem gives our main result – estimation error in ℓ_2 norm under label DP.

Theorem 4.4 (Private Estimation in ℓ_2-norm)**.** *Fix $\delta \in (0, 1/e)$ and $\varepsilon \geq 0$. The, under Assumptions 2.1 and 4.3, running Algorithm 1 with $\eta_t = \frac{1}{\gamma\kappa}$, we have, with probability at least $1-\delta$, for all $t \leq n$, the following:*

$$\|\theta_t - \theta^*\| \leq C \cdot \frac{L}{\gamma\kappa} \cdot \frac{e^\varepsilon + 1}{e^\varepsilon - 1}\sqrt{\frac{\log\log(n) + \log(1/\delta)}{t}},$$

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

Proof of this theorem is deferred to Appendix B. The above result implies an estimation error of $g(\varepsilon, \delta, L, n, \kappa) = \widetilde{O}\left(\frac{L}{\gamma\kappa}\frac{e^\varepsilon+1}{e^\varepsilon-1}\sqrt{\frac{\log(1/\delta)}{n}}\right)$ for $\widehat{\theta}_{\text{SGD-RR}}$. Several remarks are in order with this observation.

Cost of privacy. The privacy cost of our estimator $\widehat{\theta}_{\text{SGD-RR}}$ is a multiplicative factor of $\frac{e^\varepsilon+1}{e^\varepsilon-1}$. This cost of privacy is standard for RR mechanism (Duchi et al., 2018; Chan et al., 2012) and improves over the privacy cost suffered by our MLE based estimator $\widehat{\theta}_{\text{MLE-RR}}$. Not only that, the error bound of $\widehat{\theta}_{\text{SGD-RR}}$ holds for all privacy budgets $\varepsilon > 0$ rather than that of $\widehat{\theta}_{\text{MLE-RR}}$, which holds only when ε is higher than a certain threshold. Hence, Theorem 4.4 significantly boosts the applicability of our method in all practical privacy regimes, which comes with an expense of a coverage assumption on the state-action feature space.

Comparison with Zhu et al. (2023). The non-private estimator $\widehat{\theta}_{\text{MLE}}$ of Zhu et al. (2023) minimizes the loss function (5) and achives an estimation error $O\left(\frac{1}{\gamma}\sqrt{\frac{d}{n}}\right)$ in the semi-norm. We have the same $1/\gamma$ dependency in the estimation error as in Zhu et al. (2023). The main difference compared to Zhu et al. (2023) is that we bound estimation error under l_2 norm and hence get hit by the coverage parameter κ – our error increases as κ decreases. Another apparent difference is dependence (or the lack of it) on the feature dimension d in the estimation. However, it is often the case that feature norm bound L is of the order $O(\sqrt{d})$ yielding a similar dependence on d as Zhu et al. (2023). Finally, armed with the coverage assumption, instead of employing a pessimistic policy as in (6) for a downstream offline contextual bandit task, we can design a greedy (plug-in) policy

$$\widehat{\pi}_{\text{Greedy}}(s) = \underset{a\in\mathcal{A}}{\operatorname{argmax}}\, \phi(s,a)^\top \widehat{\theta}_{\text{SGD-RR}},$$

which achieves a *sub-optimality gap* of the order $O(L \cdot g(\varepsilon, \delta, L, n, \kappa)$, while ensuring label-DP.

Comparison with Cai et al. (2023). One closest work to ours is Cai et al. (2023), which studied the BTL model under the constraint of label DP. They leverage the objective perturbation technique of Kifer et al. (2012) to design a private estimator, which only suffers an additive privacy cost rather than a multiplicative one which we get. This is mainly because their algorithm (via objective perturbation) *only* works under the central model of DP, i.e., when the agent is trusted and she has access to all the raw data. In contrast, as discussed in Remark 4.2, our Algorithm 1 also works under the stronger local model (i.e., when each labeler does not trust the central agent and only sends randomized label $\tilde{y}$ to her), and achieves the same error guarantee as in Theorem 4.4. The multiplicative cost of privacy is what we pay for designing an algorithm which works simultaneously under both central and local model of label DP. Another important difference is that their result holds only in the *tabular* setting, i.e., when $\mathcal{S}, \mathcal{A}$ are finite and each $\phi(s,a)$ corresponds to a standard basis vector. One important future work is to employ objective perturbation to achive central DP with additive privacy cost under the linear BTL model as considered in this work.

Extension to K-wise comparison data. One possible extension of our results is to privately learn the reward function r_{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 preferred by human labeler. Under the Placket-Luce model, the label y is sampled according to the probability distribution, for each $l \in \{1, \ldots, K\}$,

$$\mathbb{P}\left[y = l | s, a_1, \ldots, a_K\right] = \frac{\exp(r_{\theta^*}(s, a_l))}{\sum_{j=1}^K \exp(r_{\theta^*}(s, a_j))}.$$

When $K = 2$, this reduces to the pairwise comparison considered in this work, i.e., the BTL model. One approach to extend our results to the PL model is by splitting the K-wise comparison data to pairwise comparisons and running MLE (or SGD) for total $K(K-1)/2$ number of pairwise comparisons. In this case, privacy can be ensured by employing the K-Randomized Response (K-RR) mechanism. When queried the value of y, the K-RR mechanism outputs $\widetilde{y}$, which is randomly sampled from the probability distribution:

$$\mathbb{P}\left[\widetilde{y} = y\right] = \frac{e^\varepsilon}{e^\varepsilon + K - 1} \text{ and } \mathbb{P}\left[\widetilde{y} \neq y\right] = \frac{1}{e^\varepsilon + K - 1}.$$

Again K-RR reduces to RR when $K = 2$. This approach would achieve an estimation error roughly of the same order as in Theorem 3.2 for the private MLE estimator under the semi-norm defined by the sample covariance matrix $\Sigma_{\mathcal{D}} = \frac{2}{K(K-1)n}\sum_{i=1}^n \sum_{j=1}^K \sum_{k=j+1}^K x_{i,(jk)} x_{i,(jk)}^\top$, where $x_{i,(jk)} = \phi(s_i, a_{i,j}) - \phi(s_i, a_{i,k})$ denotes the feature difference between j-th and k-th action for the i-th data point. Another approach is to directly run MLE or SGD on the K-wise comparison data. For example, for the SGD-based approach, one can replace the binary logistic loss with the cross-entropy loss for multi-class classification. Then, armed with K-RR and a new gradient estimate $\widehat{g}_t$, one can establish a similar bound as in Theorem 4.4 with a new multiplicative factor of the order $\frac{e^\varepsilon+K-1}{e^\varepsilon-1}$.

5. Conclusion

We presented the first results on private reward estimation from preference-based feedback. In particular, we showed that for both semi-norm and ℓ_2 norm, there exist estimators of sample efficiency $\widetilde{O}(1/\sqrt{n})$ while guaranteeing label DP. Our algorithms for both cases even offer privacy protection in the local trust model where each human labeler does not trust the agent. We believe that our private estimators will be useful in many emerging preference-based learning scenarios, such as preference-based RL and learning from human feedback in general. Several interesting future research directions are in order. First, it is instructive to establish a tight lower bound for label DP under both central and local models. For the central model, it would be interesting to study how to adapt the lower bound technique in Cai et al. (2023) from the tabular case to our linear case. On the other hand, for the local model, one promising approach is to leverage the techniques in Shah et al. (2015) and Duchi et al. (2018). It would also be interesting to study how to apply our private estimators to trajectory-based comparison in offline RL and establish the private counterpart of the results in Zhu et al. (2023).

References

Behnia, R., Ebrahimi, M. R., Pacheco, J., and Padmanabhan, B. Ew-tune: A framework for privately fine-tuning large language models with differential privacy. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 560–566. IEEE, 2022.

Beimel, A., Nissim, K., and Stemmer, U. 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*, pp. 363–378. Springer, 2013.

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

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

Chan, T. H., Shi, E., and Song, D. Optimal lower bound for differentially private multi-party aggregation. In *Algorithms–ESA 2012: 20th Annual European Symposium, Ljubljana, Slovenia, September 10-12, 2012. Proceedings 20*, pp. 277–288. Springer, 2012.

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

Chen, X., Zhong, H., Yang, Z., Wang, Z., and Wang, L. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, pp. 3773–3793. PMLR, 2022.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

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

Dwork, C. 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*, pp. 1–19. Springer, 2008.

Ghazi, B., Golowich, N., Kumar, R., Manurangsi, P., and Zhang, C. Deep learning with label differential privacy. *Advances in neural information processing systems*, 34: 27131–27145, 2021.

Glaese, A., McAleese, N., Trebacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L., Chadwick, M., Thacker, P., et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

Green, P. E., Carroll, J. D., and DeSarbo, W. S. Estimating choice probabilities in multiattribute decision making. *Journal of Consumer Research*, 8(1):76–84, 1981.

Herbrich, R., Minka, T., and Graepel, T. Trueskill™: a bayesian skill rating system. *Advances in neural information processing systems*, 19, 2006.

Hsu, D., Kakade, S., and Zhang, T. A tail inequality for quadratic forms of subgaussian random vectors. 2012.

Kazai, G. In search of quality in crowdsourcing for search engine evaluation. In *Advances in Information Retrieval: 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings 33*, pp. 165–176. Springer, 2011.

Khatib, F., DiMaio, F., Group, F. C., Group, F. V. C., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., et al. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology*, 18(10):1175–1177, 2011.

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

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

Li, H., Guo, D., Fan, W., Xu, M., and Song, Y. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*, 2023.

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

MacGlashan, J., Ho, M. K., Loftin, R., Peng, B., Wang, G., Roberts, D. L., Taylor, M. E., and Littman, M. L. Interactive learning from policy-dependent human feedback. In *International Conference on Machine Learning*, pp. 2285–2294. PMLR, 2017.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Pacchiano, A., Saha, A., and Lee, J. Dueling rl: reinforcement learning with trajectory preferences. *arXiv preprint arXiv:2111.04850*, 2021.

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

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

Shah, N., Balakrishnan, S., Bradley, J., Parekh, A., Ramchandran, K., and Wainwright, M. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In *Artificial intelligence and statistics*, pp. 856–865. PMLR, 2015.

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

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

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

Yin, M., Duan, Y., Wang, M., and Wang, Y.-X. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

Yu, D., Naik, S., Backurs, A., Gopi, S., Inan, H. A., Kamath, G., Kulkarni, J., Lee, Y. T., Manoel, A., Wutschitz, L., et al. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*, 2021.

Zhan, W., Uehara, M., Kallus, N., Lee, J. D., and Sun, W. Provable offline reinforcement learning with human feedback. *arXiv preprint arXiv:2305.14816*, 2023.

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

A. Derivation of estimation error bound in semi-norm

We are given a query-observation dataset $\mathcal{D} = (s_i, a_i^0, a_i^1, y_i)_{i=1}^n$. We define $x^i = \phi(s_i, a_i^1) - \phi(s_i, a_i^0)$. We privatize human feedback using randomized response mechanism. We define $\widetilde{y}_i$ to be the output of RR given input y_i. In this case, we have

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

Based on this, we define negative log-likelihood

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

Now we compute its gradient $\nabla l_{\mathcal{D},\varepsilon}(\theta) = -\frac{1}{n}\sum_{i=1}^n V_i x_i = -\frac{1}{n}X^\top V$, where

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

Then, we have

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

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

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

Assume that $-c \leqslant \theta^\top x_i \leqslant c$. (Note that $c = LB$ in our setting.) Then both $\alpha_{1,i}, \alpha_{0,i} \geqslant \gamma$, where

$$\gamma = \frac{(e^\varepsilon - 1)(e^\varepsilon e^{-2c} - 1)}{e^c(1+e^c)^2(e^\varepsilon e^c + 1)^2} > 0$$

if $\varepsilon > 2c$. This implies that $l_{\mathcal{D},\varepsilon}$ is strongly convex around θ with parameter γ and with respect to the semi-norm $\left\|\cdot\right\|_{\Sigma_{\mathcal{D}}}$. Then, if we introduce the error vector $\Delta = \widehat{\theta}_n - \theta$, we conclude that

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

Now note that

$$\begin{aligned}
V_i|(\widetilde{y}_i = 1) &= \frac{e^{-\theta^\top x_i}(e^\varepsilon - 1)}{(1+e^{-\theta^\top x_i})(e^\varepsilon + e^{-\theta^\top x_i})} \leqslant \frac{e^c(e^\varepsilon - 1)}{(1+e^{-c})(e^\varepsilon + e^{-c})} = \frac{e^{3c}(e^\varepsilon - 1)}{(1+e^c)(e^\varepsilon e^c + 1)} \\
V_i|(\widetilde{y}_i = 0) &= \frac{e^{-\theta^\top x_i}(e^\varepsilon - 1)}{(1+e^{-\theta^\top x_i})(1+e^\varepsilon e^{-\theta^\top x_i})} \\
&= \frac{e^{\theta^\top x_i}(e^\varepsilon - 1)}{(1+e^{\theta^\top x_i})(e^\varepsilon + e^{\theta^\top x_i})} \leqslant \frac{e^c(e^\varepsilon - 1)}{(1+e^{-c})(e^\varepsilon + e^{-c})} = \frac{e^{3c}(e^\varepsilon - 1)}{(1+e^c)(e^\varepsilon e^c + 1)}
\end{aligned}$$

Therefore each V_i is $\sigma = \frac{e^{3c}(e^{\varepsilon}-1)}{(1+e^c)(e^{\varepsilon}e^c+1)}$-sub-Gaussian.

Introducing $M = \frac{1}{n^2}X(\Sigma_{\mathcal{D}}+\lambda I)^{-1}X^\top$, we have $\|\nabla l_{\mathcal{D},\varepsilon}(\theta)\|^2_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} = V^\top MV$. 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 $1-\delta$,

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

This gives us

$$\begin{aligned}\gamma\|\Delta\|^2_{\Sigma_{\mathcal{D}}+\lambda I} &\leqslant \|\nabla l_{\mathcal{D},\varepsilon}(\theta)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}\|\Delta\|_{(\Sigma_{\mathcal{D}}+\lambda I)} + 4\lambda\gamma B^2\\ &\leqslant \sqrt{C_1\cdot\sigma^2\cdot\frac{d+\log(1/\delta)}{n}}\|\Delta\|_{(\Sigma_{\mathcal{D}}+\lambda I)} + 4\lambda\gamma B^2\end{aligned}$$

Solving for the above inequality, we get

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

Now note that $\frac{\sigma}{\gamma} = \frac{e^{4c}(1+e^c)(e^{\varepsilon}e^c+1)}{(e^{\varepsilon}e^{-2C}-1)}$. Hence we get

$$\|\Delta\|_{(\Sigma_{\mathcal{D}}+\lambda I)} \leqslant C\cdot\frac{(e^{\varepsilon}e^c+1)}{(e^{\varepsilon}e^{-2c}-1)}\sqrt{\frac{d+\log(1/\delta)}{n}} + C'\cdot\sqrt{\lambda}B,$$

which holds for any $\varepsilon > 2c$, where $\left|\theta^\top x_i\right| \leqslant c$ for all $i\in[n]$. This proves Theorem 3.2.

B. Derivation of estimation error bound in ℓ_2-norm

Proof. We divide the proof of Theorem 4.4 into the following steps.

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

$$\|\theta_{t+1}-\theta^*\|^2 \le (1-2/t)\|\theta_t-\theta^*\|^2 + \frac{2}{\lambda t}\langle\widehat{z}_t, \theta_t-\theta*\rangle + \left(\frac{G}{\lambda t}\right)^2. \tag{7}$$

To this end, we first define $\widehat{z}_t := \mathbb{E}[\widehat{g}_t|\mathcal{F}_{t-1}] - \widehat{g}_t$, where $\mathcal{F}_{t-1}$ is the filtration up to the end of $t-1$. Note that this condition is necessary since θ_t also depends on previous randomness in gradient computation. Then, we have

$$\begin{aligned}\|\theta_{t+1}-\theta^*\|^2 &= \|\Pi_\Theta(\theta_t-\eta_t\widehat{g}_t)-\theta^*\|^2\\ &\le \|\theta_t-\eta_t\widehat{g}_t-\theta^*\|^2\\ &= \|\theta_t-\theta^*\|^2 - 2\eta_t\langle\widehat{g}_t,\theta_t-\theta^*\rangle + \eta_t^2\|\widehat{g}_t\|^2\\ &\overset{(a)}{=} \|\theta_t-\theta^*\|^2 - 2\eta_t\langle\mathbb{E}[\widehat{g}_t|\mathcal{F}_{t-1}],\theta_t-\theta^*\rangle + 2\eta_t\langle\widehat{z}_t,\theta_t-\theta^*\rangle + \eta_t^2\|\widehat{g}_t\|^2\end{aligned} \tag{8}$$

where (a) 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 $\langle \mathbb{E}[\widehat{g}_t|\mathcal{F}_{t-1}], \theta_t - \theta^*\rangle$, which can be bounded as follows.

$$\begin{aligned} \langle \mathbb{E}[\widehat{g}_t|\mathcal{F}_{t-1}], \theta_t - \theta^*\rangle &\overset{(a)}{=} \mathbb{E}[\langle g_t, \theta_t-\theta^*\rangle|\mathcal{F}_{t-1}] \\ &\overset{(b)}{=} \mathbb{E}[\langle(\sigma(x_t^\top\theta_t)-y_t)x_t, \theta_t-\theta^*\rangle|\mathcal{F}_{t-1}]\\ &\overset{(c)}{=} \mathbb{E}[\langle(\sigma(x_t^\top\theta_t)-\sigma((x_t^\top\theta^*))x_t,\theta_t-\theta^*\rangle|\mathcal{F}_{t-1}] \\ &\overset{(d)}{\geq}\gamma\mathbb{E}[(x_t^\top(\theta_t-\theta^*))^2|\mathcal{F}_{t-1}]\\ &= \gamma(\theta_t-\theta^*)^\top \mathbb{E}[x_tx_t^\top|\mathcal{F}_{t-1}](\theta_t-\theta^*)\\ &\overset{(e)}{\geq}\gamma\kappa\left\|\theta_t-\theta^*\right\|^2 \end{aligned} \tag{9}$$

where (a) holds by the fact that $\widehat{g}_t$ is an unbiased estimate of true gradient $g_t := \nabla\ell(\theta_t, (x_t, y_t))$, where ℓ is the log-loss. To see this, by the definitions of $\tilde{g}_t$ and $\widehat{g}_t$, and for any given (x_t, y_t), we have

$$\begin{aligned} &\mathbb{E}[\widehat{g}_t|\mathcal{F}_{t-1}]\\ =&\mathbb{E}_{\tilde{y}_t}[\widehat{g}_t|\mathcal{F}_{t-1}]\\ =&\frac{e^\varepsilon+1}{e^\varepsilon-1}\cdot\left(\frac{e^\varepsilon}{1+e^\varepsilon}\nabla\ell(\theta_t,(x_t,y_t))+\frac{1}{1+e^\varepsilon}\nabla\ell(\theta_t,(x_t,1-y_t))-\sum_{y=0}^{1}\frac{1}{1+e^\varepsilon}\cdot\nabla\ell(\theta_t,(x_t,y))\right)\\ =&\nabla\ell(\theta_t,(x_t,y_t)) = g_t \end{aligned}$$

(b) holds by definition of g_t and $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function; (c) holds by definition of y_t; (d) holds by mean-value theorem and note that $\sigma'(z) = \sigma(z)(1-\sigma(z))$ and hence $\inf_{z\in[-2LB,2LB]}\sigma'(z) \geq \gamma := \frac{1}{2+\exp(-2LB)+\exp(2LB)}$, where we utilize Assumption 2.1; (e) holds by Assumption 4.3 and x_t is independent of $\mathcal{F}_{t-1}$.

Thus, plugging (9) into (8), yields

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

where (a) holds by $\left\|\widehat{g}_t\right\|^2 \leq G^2 := 36L^2\left(\frac{e^\varepsilon+1}{e^\varepsilon-1}\right)^2$, which again utilizes Assumption 2.1; (b) holds by letting $\eta_t := \frac{1}{\lambda t}$ and $\lambda := \gamma\kappa$. Hence, we have established (7).

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

$$\left\|\theta_{t+1}-\theta^*\right\|^2 \leq \frac{2}{\lambda(t-1)t}\sum_{i=2}^{t}(i-1)\langle\widehat{z}_i,\theta_i-\theta^*\rangle+\frac{G^2}{\lambda^2t^2}. \tag{10}$$

To this end, we basically expand the recursion in (7) till $t = 2$ and simple algebra leads to the result. This step also directly follows from (Rakhlin et al., 2011).

Step 3: We will apply one particular version of Freedman's inequality to control the concentration of $\sum_{i=2}^t(i-1)\langle\widehat{z}_i,\theta_i-\theta^*\rangle$ in (10). In particular, we will apply Lemma 3 in (Rakhlin et al., 2011) 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 $Z_i = \langle\widehat{z}_i,\theta_i-\theta^*\rangle$. 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 conditional variance $\text{Var}[Z_i|\mathcal{F}_{i-1}] \leq 4G^2\left\|\theta_i-\theta^*\right\|^2$, which holds by $\left\|\widehat{z}_i\right\| \leq 2G$. Now consider the sum $\sum_{i=2}^t(i-1)\langle\widehat{z}_i,\theta_i-\theta^*\rangle$ in (10). We need to check two conditions: (i) The sum of conditional variance satisfies

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

(ii) Uniform upper bound on each term

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

where (a) comes from (9) and recall that $\lambda = \gamma\kappa$. To see it, by Cauchy-Schwartz inequality, we have $\gamma\kappa\left\|\theta_t - \theta^*\right\|^2 \leq G\left\|\theta_t - \theta^*\right\|$, and hence $\left\|\theta_t - \theta^*\right\| \leq G/\lambda$ for all t. We can then apply Lemma 3 in (Rakhlin et al., 2011) to obtain that for $n \geq 4$ and $\delta \in (0, 1/e)$, then with probability at least $1-\delta$, for all $t \leq n$

$$\sum_{i=2}^{t}(i-1)Z_i \leq 8G\max\left\{\sqrt{\sum_{i=2}^{t}(i-1)^2\left\|\theta_i - \theta^*\right\|^2}, \frac{G(t-1)}{\lambda}\sqrt{\log(\log n/\delta)}\right\}\sqrt{\log(\log n/\delta)}. \tag{11}$$

Step 4: Once we obtain (11), 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 that for all $t \leq n$,

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

for some absolute constant C. Hence, we have completed the proof. □