PRINCIPLED REINFORCEMENT LEARNING WITH HUMAN FEEDBACK FROM PAIRWISE OR K-WISE COMPARISONS

Banghua Zhu
Department of EECS
UC Berkeley
banghua@berkeley.edu

Jiantao Jiao
Department of EECS
UC Berkeley
jiantao@berkeley.edu

Michael I. Jordan
Department of EECS
UC Berkeley
jordan@cs.berkeley.edu

ABSTRACT

We provide a theoretical framework for Reinforcement Learning with Human Feedback (RLHF). Our analysis shows that when the true reward function is linear, the widely used maximum likelihood estimator (MLE) converges under both the Bradley-Terry-Luce (BTL) model and the Plackett-Luce (PL) model. However, we show that when training a policy based on the learned reward model, MLE fails while a pessimistic MLE provides policies with improved performance under certain coverage assumptions. Additionally, we demonstrate that under the PL model, the true MLE and an alternative MLE that splits the K-wise comparison into pairwise comparisons both converge. Moreover, the true MLE is asymptotically more efficient. Our results validate the empirical success of existing RLHF algorithms in InstructGPT and provide new insights for algorithm design. Furthermore, our results unify the problem of RLHF and max entropy Inverse Reinforcement Learning (IRL), and provide the first sample complexity bound for max entropy IRL.

1 INTRODUCTION

The alignment problem aims at aligning human values with machine learning systems and steering learning algorithms towards the goals and interests of humans. One of the most promising tools for AI alignment, *Reinforcement Learning with Human Feedback* (RLHF), has delivered significant empirical success in the fields of game playing (Knox and Stone, 2008; MacGlashan et al., 2017; Christiano et al., 2017a; Warnell et al., 2018), robotics (Brown et al., 2019; Shin et al., 2023) and language models (Ziegler et al., 2019; Stiennon et al., 2020; Wu et al., 2021; Nakano et al., 2021; Ouyang et al., 2022; Menick et al., 2022; Glaese et al., 2022; Gao et al., 2022; Bai et al., 2022a; Ganguli et al., 2022; Ramamurthy et al., 2022). Notably, the language model application ChatGPT is based on RLHF and this underlies several of its skills: answering followup questions, admitting its mistakes, challenging incorrect premises, and rejecting inappropriate requests. One of the key capabilities in RLHF is to learn a reward from human feedback, in the form of pairwise or K-wise comparisons between actions (responses). In this paper, we take the first step towards providing a theoretical framework for RLHF, with a specific focus on reward learning. We provide theoretical analysis that justifies the empirical success of RLHF in InstructGPT and ChatGPT, along with new insights for algorithm design.

Taking InstructGPT Ouyang et al. (2022) as an example, a typical deployment of RLHF for language modeling includes the following steps:

(a) Pre-train a Large Language Model (LLM) using supervised training.

(b) Train a reward model based on the pre-trained LLM using human feedback.

(c) Fine-tune the existing LLM based on the learned reward model using Proximal Policy Optimization (PPO).

During the reward training step, the prompts are first sampled from a pre-collected dataset. Then K responses are sampled by executing existing models on the sampled prompts. Based on the prompt provided, a human labeler ranks all the responses according to her own preference. The reward model is trained based on a maximum likelihood estimator (MLE), also known as the learning-to-rank algorithm or cross-entropy minimization (Liu et al., 2009; Xia et al., 2008; Cao et al., 2007; Christiano et al., 2017a; Ouyang et al., 2022).

In the setting of InstructGPT, the ranking of responses is based purely on the current prompt, which can be viewed as the state in a contextual bandit. We accordingly start with the setting of a contextual bandit, and later generalize our results to Markov Decision Process (MDP) where there are transitions between states. Let $\mathcal{S}$ be the set of states (prompts), and $\mathcal{A}$ be the set of actions (responses). For each state-action pair (s, a), we assume that the reward is parametrized by $r_\theta(s, a) = \langle \theta, \phi(s, a) \rangle$ for some known and fixed feature function $\phi(s, a) : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$. In an LLM, such a ϕ is usually derived by removing the last layer of the pre-trained model.[1] We denote the ground-truth reward provided by a human as $r_{\theta^\star}(s, a)$ for some parameter $\theta^\star \in \mathbb{R}^d$.

We are interested in the sample complexity for learning a reward model $r_{\theta^\star}$ from pairwise or K-wise comparison data. For the i-th sample, a state s^i is first sampled from some fixed distribution ρ. Given the state s^i, K actions $(a_0^i, a_1^i, \cdots, a_{K-1}^i)$ are sampled from some joint distribution $\mathbb{P}(a_0, \cdots, a_{K-1} \mid s^i)$. Let $\sigma^i : [K] \mapsto [K]$ denote the output of the human labeller, which is a permutation function representing the ranking of the actions. Here $\sigma^i(0)$ represents the most preferred action. We assume that the distribution of σ^i follows a Plackett-Luce (PL) model (Plackett, 1975; Luce, 2012):

$$\mathbb{P}(\sigma^i \mid s^i, a_0^i, a_1^i, \cdots, a_{K-1}^i) = \prod_{k=0}^{K-1} \frac{\exp(r_{\theta^\star}(s^i, a_{\sigma^i(k)}^i))}{\sum_{j=k}^{K-1} \exp(r_{\theta^\star}(s^i, a_{\sigma^i(j)}^i))}.$$

When $K = 2$, this reduces to the pairwise comparison of the Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952), which is widely applied in existing RLHF algorithms Christiano et al. (2017a); Ouyang et al. (2022).

Since the learned reward model is mainly used for downstream policy training, we measure the correctness of the estimated reward model via the performance of a greedy policy trained from a reward model $r_{\hat{\theta}}$. Concretely, for a greedy policy $\hat{\pi}(s) = \arg\max_a r_{\hat{\theta}}(s, a)$, we compute a performance gap compared to the optimal policy:

$$\mathsf{SubOpt}(\hat{\pi}) \coloneqq \mathbb{E}_{s\sim\rho}[r_{\theta^\star}(s, \pi^\star(s)) - r_{\theta^\star}(s, \hat{\pi}(s)].$$

Here $\pi^\star = \arg\max_a r_{\theta^\star}(s, a)$ is the optimal policy under the true reward $r_{\theta^\star}$.

In this paper, we study the potential sub-optimality of the MLE in the RLHF setting. As a by-product, we also provide guarantee of the estimation error on the semi-norm of the parameter estimation error, $\|\hat{\theta} - \theta^\star\|_\Sigma$, for a query-dependent covariance matrix Σ.

From a broader perspective, the framework of RLHF can be viewed as a special case of reward learning from pre-collected data, which has been a primary focus in Inverse Reinforcement Learning (IRL) and offline reinforcement learning. Our techniques also provide theoretical guarantee for the max-entropy IRL Ziebart et al. (2008) and action-based IRL algorithms Ramachandran and Amir (2007); Neu and Szepesvári (2009); Florence et al. (2022).

[1]In InstructGPT, the function ϕ is still parametrized can be further trained in the reward learning step. However, for simplicity of theoretical analysis we assume in this paper that ϕ is fixed and one only fine-tunes the last layer with parameter θ.

2 MAIN RESULTS

2.1 PAIRWISE COMPARISON

We start with the setting of a contexual bandit with pairwise comparison. We focus on two algorithms, MLE and pessimistic MLE.

We first bound the estimation error for MLE, the most common algorithm in learning to rank and RLHF Liu et al. (2009); Xia et al. (2008); Cao et al. (2007); Christiano et al. (2017a); Ouyang et al. (2022). For any query-observation dataset $\{(s^i, a_1^i, a_2^i, y^i)\}_{i=1}^n$, MLE aims at minimizing the negative log likelihood, defined as:

$$
\begin{aligned}
&\hat{\theta}_{\mathsf{MLE}} \in \arg\min_{\theta \in \Theta_B} \ell_{\mathcal{D}}(\theta), \\
&\ell_{\mathcal{D}}(\theta) = -\sum_{i=1}^n \log\Big(\frac{1(y^i = 1)\cdot \exp(r_\theta(s^i, a_1^i))}{\exp(r_\theta(s^i, a_0^i)) + \exp(r_\theta(s^i, a_1^i))} + \frac{1(y^i = 0)\cdot \exp(r_\theta(s^i, a_0^i))}{\exp(r_\theta(s^i, a_0^i)) + \exp(r_\theta(s^i, a_1^i))}\Big) \\
&\qquad = -\sum_{i=1}^n \log\Big(1(y^i = 1)\cdot \mathsf{sigmoid}(\langle\theta,\, \phi(s^i, a_1^i) - \phi(s^i, a_0^i)\rangle) + 1(y^i = 0)\cdot \mathsf{sigmoid}(\langle\theta,\, \phi(s^i, a_0^i) - \phi(s^i, a_1^i)\rangle)\Big).
\end{aligned}
$$

When the minimizer is not unique, we take any of the $\hat{\theta}$ that achieve the minimum. Let $\mathcal{D} = \{(s^i, a_1^i, a_2^i)\}_{i=1}^n$ denote the queried state-action pairs. In this paper, we study how one can utilize $\mathcal{D}$ to learn a near-optimal reward model and policy. We first present a result on the estimation error conditioned on the data $\mathcal{D}$. The result is a generalization of the upper bound in Shah et al. (2015, Theorem 1) and the analysis follows a similar structure. The main difference is that Shah et al. (2015) focus on the tabular case when $\phi(s, a)$ is always a unit vector, while in our case $\phi(s, a)$ can be an arbitrary d-dimensional vector. This confidence bound guarantee is also similar to the guarantee for dueling bandits and RL in Faury et al. (2020); Pacchiano et al. (2021), except for that we have better rate in logarithmic factors since union bound is not needed in our case.

Lemma 2.1 (Informal). *Under certain regularity conditions, the MLE satisfies the following with probability at least* $1 - \delta$*,*

$$
\|\hat{\theta}_{\mathsf{MLE}} - \theta^\star\|_{\Sigma_{\mathcal{D}}} \leq C \cdot \sqrt{\frac{d + \log(1/\delta)}{n}}.
$$

Here $\Sigma_{\mathcal{D}} = \frac{1}{n}\sum_{i=1}^n (\phi(s^i, a_1^i) - \phi(s^i, a_0^i))(\phi(s^i, a_1^i) - \phi(s^i, a_0^i))^\top$.

However, when we consider the performance of the induced policy, MLE provably fails while pessimistic MLE in Algorithm 1 gives a near-optimal rate. In essence, the pessimism principle discounts actions that are less represented in the observed dataset, and hence is conservative in outputting a policy.

Algorithm 1 Pessimistic MLE

Input: The current estimator $\hat{\theta}$, the data covariance $\Sigma_{\mathcal{D}}$, the regularization parameter λ, the bound on the semi-norm $f(n, d, \delta, \lambda)$, a reference vector $v \in \mathbb{R}^d$.

Construct the confidence set

$$
\Theta(\hat{\theta}, \lambda) = \Big\{\theta \in \Theta_B \mid \|\hat{\theta} - \theta\|_{\Sigma_{\mathcal{D}} + \lambda I} \leq f(n, d, \delta, \lambda)\Big\}.
$$

Compute the pessimistic expected value function

$$
\begin{aligned}
\hat{J}(\pi) &= \min_{\theta \in \Theta(\hat{\theta}_{\mathsf{MLE}}, \lambda)} \mathbb{E}_{s\sim\rho}[\theta^\top(\phi(s, \pi(s)) - v)] \\
&= (\mathbb{E}_{s\sim\rho}[\phi(s, \pi(s))] - v)^\top \hat{\theta} - \|(\Sigma_{\mathcal{D}} + \lambda I)^{-\frac{1}{2}}(\mathbb{E}_{s\sim q}[\phi(s, \pi(s))] - v)\|_2 \cdot f(n, d, \delta, \lambda)
\end{aligned}
$$

Return: $\hat{\pi} = \arg\max_\pi \hat{V}(\pi)$.

Theorem 2.2 (Informal)**.** *Under certain coverage assumption, one can design a pessimistic MLE such that the induced greedy policy $\hat{\pi}_{\mathsf{PE}}$ is good; i.e., with probability at least $1-\delta$,*

$$\mathsf{SubOpt}(\hat{\pi}_{\mathsf{PE}}) = \Theta\left(\sqrt{\frac{d+\log(1/\delta)}{n}}\right).$$

In contrast, under the same assumption, one can find instances such that the greedy policy w.r.t. MLE $\hat{\pi}_{\mathsf{MLE}}$ fails:

$$\forall n > 1, \mathbb{E}[\mathsf{SubOpt}(\hat{\pi}_{\mathsf{MLE}})] \geq 0.1.$$

2.2 K-WISE COMPARISON

For K-wise comparison, we analyze both the MLE and the algorithm in InstructGPT (Ouyang et al., 2022) which splits the ranking data into $K(K-1)/2$ pairwise comparison data and runs an MLE based on the BTL model. We show that both converge in terms of the estimation error under the semi-norm, and give a near-optimal policy when combined with pessimism.

Let the estimated parameter for the splitted estimator be $\hat{\theta}$ and the induced policy be $\hat{\pi}_{\mathsf{PE}}$. We have:

Theorem 2.3 (Informal)**.** *Under certain coverage and regularity conditions, the following holds separately with probability at least $1-\delta$:*

$$\|\hat{\theta}-\theta^{\star}\|_{\Sigma_{\mathcal{D}}} \leq C\cdot\sqrt{\frac{d+\log(1/\delta)}{n}},$$
$$\mathsf{SubOpt}(\hat{\pi}_{\mathsf{PE}}) \leq C'\cdot\sqrt{\frac{d+\log(1/\delta)}{n}},$$

Here $\Sigma_{\mathcal{D}} = \frac{2}{K(K-1)n}(\sum_{i=1}^{n}\sum_{j=0}^{K-1}\sum_{k=j+1}^{K-1}(\phi(s^i,a_j^i)-\phi(s^i,a_k^i))(\phi(s^i,a_j^i)-\phi(s^i,a_k^i))^\top)$.

2.3 EXTENSION TO MDP

We also extend our results to the case of MDP and IRL; see the detailed presentation in Section E and Appendix F. Let the estimated parameter be $\hat{\theta}$ and the induced pessimistic policy be $\hat{\pi}_{\mathsf{PE}}$. For pairwise comparison we have:

Theorem 2.4 (Informal)**.** *In the MDP setting with horizon H, under certain coverage and regularity conditions, the following holds separately with probability at least $1-\delta$:*

$$\|\hat{\theta}-\theta^{\star}\|_{\Sigma_{\mathcal{D}}} \leq C\cdot\sqrt{\frac{d+\log(1/\delta)}{n}},$$
$$\mathsf{SubOpt}(\hat{\pi}_{\mathsf{PE}}) \leq C'\cdot\sqrt{\frac{d+\log(1/\delta)}{n}},$$

Here $\Sigma_{\mathcal{D}} = \frac{1}{n}\sum_{i=1}^{n}(\sum_{h=0}^{H}(\phi(s_h^i,a_h^i)-\phi(s_h^{i\prime},a_h^{i\prime})))\,(\sum_{h=0}^{H}(\phi(s_h^i,a_h^i)-\phi(s_h^{i\prime},a_h^{i\prime})))^\top$.

Our results not only explain the correctness of existing algorithms, but also provide new insights for algorithm design in RLHF. In particular, it suggests the importance of introducing pessimism in the reward learning part, which can be implemented via adding regularization in policy training steps as in Ouyang et al. (2022), or using existing offline RL algorithms, including but not limited to Conservative Q-Learning (Kumar et al., 2020), Implicit Q-Learning (Kostrikov et al., 2021) and Adversarially Trained Actor Critic (Cheng et al., 2022). On the other hand, it also sheds lights on the design of active learning algorithms. Since Lemma 2.1 provides a tight confidence bound on $\hat{\theta}$, one can combine it with G-optimal design (Soare et al., 2014) to achieve a near-optimal rate for pure exploration.

REFERENCES

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, page 1, 2004.

Y. Abdelkareem, S. Shehata, and F. Karray. Advances in preference-based reinforcement learning: A review. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2527–2532. IEEE, 2022.

N. Ailon, Z. S. Karnin, and T. Joachims. Reducing dueling bandits to cardinal bandits. In *ICML*, volume 32, pages 856–864, 2014.

Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

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

J. Bretagnolle and C. Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47(2):119–137, 1979. doi: 10.1007/BF00535278. URL `https://doi.org/10.1007/BF00535278`.

D. Brown, W. Goo, P. Nagarajan, and S. Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International Conference on Machine Learning*, pages 783–792. PMLR, 2019.

R. Busa-Fekete, B. Szörényi, P. Weng, W. Cheng, and E. Hüllermeier. Preference-based reinforcement learning: evolutionary direct policy search using a preference-based racing algorithm. *Machine Learning*, 97(3):327–351, 2014.

Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, pages 129–136, 2007.

X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pages 193–202, 2013.

X. Chen, H. Zhong, Z. Yang, Z. Wang, and L. 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.

Y. Chen and C. Suh. Spectral MLE: top-k rank aggregation from pairwise comparisons. In *International Conference on Machine Learning*, pages 371–380. PMLR, 2015.

C.-A. Cheng, T. Xie, N. Jiang, and A. Agarwal. Adversarially trained actor critic for offline reinforcement learning. *arXiv preprint arXiv:2202.02446*, 2022.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017a.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pages 4299–4307, 2017b.

L. Faury, M. Abeille, C. Calauzènes, and O. Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.

U. Feige, P. Raghavan, D. Peleg, and E. Upfal. Computing with noisy information. *SIAM Journal on Computing*, 23(5):1001–1018, 1994.

P. Florence, C. Lynch, A. Zeng, O. A. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, and J. Tompson. Implicit behavioral cloning. In *Conference on Robot Learning*, pages 158–168. PMLR, 2022.

P. Gajane, T. Urvoy, and F. Clérot. A relative exponential weighing algorithm for adversarial utility-based dueling bandits. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 218–227, 2015.

D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization. *arXiv preprint arXiv:2210.10760*, 2022.

S. Ghoshal and A. Saha. Exploiting correlation to achieve faster learning rates in low-rank preference bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 456–482. PMLR, 2022.

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

V. P. Godambe. An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31(4):1208–1211, 1960.

B. Hajek, S. Oh, and J. Xu. Minimax-optimal inference from partial rankings. *Advances in Neural Information Processing Systems*, 27, 2014.

R. Heckel, M. Simchowitz, K. Ramchandran, and M. Wainwright. Approximate ranking from pairwise comparisons. In *International Conference on Artificial Intelligence and Statistics*, pages 1057–1066. PMLR, 2018.

R. Heckel, N. B. Shah, K. Ramchandran, and M. J. Wainwright. Active ranking from pairwise comparisons and when parametric assumptions do not help. *Annals of Statistics*, 47(6):3099–3126, 2019.

J. Ho and S. Ermon. Generative adversarial imitation learning. *Advances in Neural Information Processing Systems*, 29, 2016.

D. Hsu, S. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:1–6, 2012.

A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

M. Jang, S. Kim, C. Suh, and S. Oh. Optimal sample complexity of m-wise data for top-k ranking. *Advances in Neural Information Processing Systems*, 30, 2017.

Y. Jin, Z. Yang, and Z. Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

J. I. Kiger. The depth/breadth trade-off in the design of menu-driven user interfaces. *International journal of man-machine studies*, 20(2):201–213, 1984.

W. B. Knox and P. Stone. Tamer: Training an agent manually via evaluative reinforcement. In *7th IEEE International Conference on Development and Learning*, pages 292–297. IEEE, 2008.

J. Komiyama, J. Honda, H. Kashima, and H. Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *COLT*, pages 1141–1154, 2015.

I. Kostrikov, A. Nair, and S. Levine. Offline reinforcement learning with implicit Q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

A. Kumar, A. Zhou, G. Tucker, and S. Levine. Conservative Q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

M.-j. Lee. M-estimator and maximum likelihood estimator (mle). In *Micro-Econometrics*, pages 91–132. Springer, 2008.

G. Li, C. Ma, and N. Srebro. Pessimism for offline linear contextual bandits using ℓ_p confidence sets. *arXiv preprint arXiv:2205.10671*, 2022.

T.-Y. Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

R. D. Luce. *Individual Choice Behavior: A Theoretical Analysis*. Courier Corporation, 2012.

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

C. Mao, J. Weed, and P. Rigollet. Minimax rates and efficient algorithms for noisy sorting. In *Algorithmic Learning Theory*, pages 821–847. PMLR, 2018.

J. Menick, M. Trebacz, V. Mikulik, J. Aslanides, F. Song, M. Chadwick, M. Glaese, S. Young, L. Campbell-Gillingham, G. Irving, et al. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022.

G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81, 1956.

R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

S. Negahban, S. Oh, K. K. Thekumparampil, and J. Xu. Learning from comparisons and choices. *The Journal of Machine Learning Research*, 19(1):1478–1572, 2018.

G. Neu and C. Szepesvári. Training parsers by inverse reinforcement learning. *Machine Learning*, 77 (2):303–337, 2009.

A. Y. Ng, S. Russell, et al. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, volume 1, page 2, 2000.

E. R. Novoseller, Y. Sui, Y. Yue, and J. W. Burdick. Dueling posterior sampling for preference-based reinforcement learning. *arXiv preprint arXiv:1908.01289*, 2019.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

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

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

A. Rajkumar and S. Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *International conference on machine learning*, pages 118–126. PMLR, 2014.

D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. In *IJCAI*, volume 7, pages 2586–2591, 2007.

R. Ramamurthy, P. Ammanabrolu, K. Brantley, J. Hessel, R. Sifa, C. Bauckhage, H. Hajishirzi, and Y. Choi. Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*, 2022.

P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34: 11702–11716, 2021.

T. L. Saaty and M. S. Ozdemir. Why the magic number seven plus or minus two. *Mathematical and computer modelling*, 38(3-4):233–244, 2003.

A. Saha and A. Gopalan. Battle of bandits. In *Uncertainty in Artificial Intelligence*, 2018a.

A. Saha and A. Gopalan. Active ranking with subset-wise preferences. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018b.

A. Saha and A. Gopalan. PAC Battling Bandits in the Plackett-Luce Model. In *Algorithmic Learning Theory*, pages 700–737, 2019.

A. Saha and A. Krishnamurthy. Efficient and optimal algorithms for contextual dueling bandits under realizability. In *International Conference on Algorithmic Learning Theory*, pages 968–994. PMLR, 2022.

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

N. B. Shah and M. J. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *The Journal of Machine Learning Research*, 18(1):7246–7283, 2017.

R. M. Shiffrin and R. M. Nosofsky. Seven plus or minus two: a commentary on capacity limitations. 1994.

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

M. Soare, A. Lazaric, and R. Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

A. W. Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge university press, 2000.

G. Warnell, N. Waytowich, V. Lawhern, and P. Stone. Deep tamer: Interactive agent shaping in high-dimensional state spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

C. Wirth, J. Furnkranz, G. Neumann, et al. Model-free preference-based reinforcement learning. In *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, pages 2222–2228, 2016.

C. Wirth, R. Akrour, G. Neumann, and J. Fürnkranz. A survey of preference-based reinforcement learning methods. *The Journal of Machine Learning Research*, 18(1):4945–4990, 2017.

J. Wu, L. Ouyang, D. M. Ziegler, N. Stiennon, R. Lowe, J. Leike, and P. Christiano. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*, 2021.

F. Xia, T.-Y. Liu, J. Wang, W. Zhang, and H. Li. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1192–1199, 2008.

T. Xie, C.-A. Cheng, N. Jiang, P. Mineiro, and A. Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:6683–6694, 2021a.

T. Xie, N. Jiang, H. Wang, C. Xiong, and Y. Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in Neural Information Processing Systems*, 34:27395–27407, 2021b.

T. Xu and Y. Liang. Provably efficient offline reinforcement learning with trajectory-wise reward. *arXiv preprint arXiv:2206.06426*, 2022.

Y. Xu, R. Wang, L. Yang, A. Singh, and A. Dubrawski. Preference-based reinforcement learning with finite-time guarantees. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18784–18794. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d9d3837ee7981e8c064774da6cdd98bf-Paper.pdf.

B. Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

Y. Yue and T. Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1201–1208. ACM, 2009.

Y. Yue and T. Joachims. Beat the mean bandit. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 241–248, 2011.

Y. Yue, J. Broder, R. Kleinberg, and T. Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

A. Zanette. When is realizability sufficient for off-policy reinforcement learning? *arXiv preprint arXiv:2211.05311*, 2022.

A. Zanette, M. J. Wainwright, and E. Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:13626–13640, 2021.

B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey, et al. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

M. Zoghi, S. Whiteson, R. Munos, M. d. Rijke, et al. Relative upper confidence bound for the k-armed dueling bandit problem. In *JMLR Workshop and Conference Proceedings*, number 32, pages 10–18. JMLR, 2014a.

M. Zoghi, S. A. Whiteson, M. De Rijke, and R. Munos. Relative confidence sampling for efficient on-line ranker evaluation. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 73–82. ACM, 2014b.

A RELATED WORK

Learning and Estimation from Pairwise Comparison and Ranking. The problem of estimation and ranking from pairwise or K-wise comparisons has been studied extensively in the literature. In the literature of *dueling bandit*, one compares two actions and aims to minimize regret based on pairwise comparisons (Yue et al., 2012; Zoghi et al., 2014b; Yue and Joachims, 2009; 2011; Saha and Krishnamurthy, 2022; Ghoshal and Saha, 2022; Saha and Gopalan, 2018a; Ailon et al., 2014; Zoghi et al., 2014a; Komiyama et al., 2015; Gajane et al., 2015; Saha and Gopalan, 2018b; 2019; Faury et al., 2020). Novoseller et al. (2019); Xu et al. (2020) analyze the sample complexity of dueling RL under the tabular case, which is extended to linear case and function approximation by the recent work Pacchiano et al. (2021); Chen et al. (2022). However, most of the work focus on pairwise comparisons between trajectories with regret minimization as the target. We take a first step towards the theoretical anlaysis for function approximation for K-wise comparisons with policy learning as the target.

On the other hand, in the literature of ranking, most of the theoretical work focuses on the tabular case where the rewards for different actions are uncorrelated (Feige et al., 1994; Shah et al., 2015; Shah and Wainwright, 2017; Heckel et al., 2018; Mao et al., 2018; Jang et al., 2017; Chen et al., 2013; Chen and Suh, 2015; Rajkumar and Agarwal, 2014; Negahban et al., 2018; Hajek et al., 2014; Heckel et al., 2019). And a majority of the empirical literature focuses on the framework of learning to rank (MLE) under general function approximation, especially when the reward is parameterized by a neural network (Liu et al., 2009; Xia et al., 2008; Cao et al., 2007; Christiano et al., 2017a; Ouyang et al., 2022; Brown et al., 2019; Shin et al., 2023; Busa-Fekete et al., 2014; Wirth et al., 2016; 2017; Christiano et al., 2017b; Abdelkareem et al., 2022). Similar idea of RL with AI feedback also learns a reward model from preference Bai et al. (2022b), except for that the preference is labeled by another AI model instead of human.

Inverse Reinforcement Learning and Offline Reinforcement Learning. RLHF, IRL and offline learning are all approaches that can be used to incorporate human preferences or expertise into the decision-making process of an agent. However, they differ in the way that they use human input to guide the agent's behavior. In IRL and imitation learning, we only observe an expert's behavior and would like to infer the expert's preferences or goals (Ng et al., 2000; Abbeel and Ng, 2004; Ziebart et al., 2008; Ramachandran and Amir, 2007; Neu and Szepesvári, 2009; Ho and Ermon, 2016; Florence et al., 2022; Hussein et al., 2017). In offline learning, we directly observe the cardinal rewards for the state. But the actions are likely to be sub-optimal. In RLHF, we observe ordinal comparisons between pairs or a set of actions. In one of the popular IRL frameworks, max-entropy IRL (Ziebart et al., 2008), it is also assumed that human choice follows a PL model. We unify the problem of RLHF and max-entropy IRL, and provide the first sample complexity analysis for both problems.

Pessimism in Offline RL. The idea of introducing pessimism for offline RL has been studied in recent year (Jin et al., 2021; Rashidinejad et al., 2021; Li et al., 2022; Xie et al., 2021b; Zanette, 2022; Zanette et al., 2021; Xie et al., 2021a; Xu and Liang, 2022). In this paper, we connect RLHF with offline RL and show that pessimism also helps in RLHF.

B PRELIMINARIES

We begin with the notation that we use in the paper. We discuss our formulations of contextual bandits and Markov decision processes in Section B.1. We introduce the data collection model and the BTL and PL models in Section B.2.

Notations. We use calligraphic letters for sets, e.g., $\mathcal{S}$ and $\mathcal{A}$. Given a set $\mathcal{S}$, we write $|\mathcal{S}|$ to represent the cardinality of $\mathcal{S}$. For vectors x and y, we use $\langle x, y\rangle = x^\top y$ to denote their inner product. We use $[K]$ to denote the set of integers from 0 to $K-1$. We write $\|x\|_\Sigma = \sqrt{x^\top \Sigma x}$ as a semi-norm of x when Σ is some positive-semidefinite matrix. We write $\Sigma \succeq \Sigma'$ if $\Sigma - \Sigma'$ is positive semidefinite.

B.1 MARKOV DECISION PROCESSES

We consider a finite-horizon MDP described by a tuple $M = (\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, \{R_h\}_{h=1}^H, \rho)$, where $\mathcal{S}$ is a (possibly infinite) state space, $\mathcal{A}$ is a (possibly infinite) action space, H is the horizon length, $P_h : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$ is a probability transition matrix at step h, $R_h : \mathcal{S} \times \mathcal{A} \mapsto \Delta([0, 1])$ encodes a family of reward distributions with $r_h : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ as the expected reward function, $\rho : \mathcal{S} \mapsto \Delta(\mathcal{S})$ is the initial state distribution. At step h, upon executing action a from state s, the agent receives a deterministic reward $r_h(s, a)$ and transits to the next state s' with probability $P_h(s'|s, a)$. The MDP transits to an absorbing termination state with zero reward at step H. When $H = 1$ and there is no transition, the model reduces to the contextual bandit problem.

A deterministic policy $\pi_h : \mathcal{S} \mapsto \mathcal{A}$ is a function that maps a state to an action at step $h \in [H]$. We use π to denote the family of policies $\{\pi_h\}_{h=1}^H$. Correspondingly, the value function $V^\pi : \mathcal{S} \mapsto \mathbb{R}$ of the policy family $\{\pi_h\}_{h\in[H]}$ is defined as the expected sum of rewards starting at state s and following policy π_h at step h. More precisely, we have for any $s \in \mathcal{S}$, $V^\pi(s) := \mathbb{E}\left[\sum_{h=0}^H r_h(s_h, a_h) \mid s_0 = s, a_h = \pi_h(s_h), \forall h \geq 0\right]$, where the expectation is taken over the trajectory generated according to the transition kernel $s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$ and reward distribution $r_h \sim R_h(\cdot \mid s_h, a_h)$. The Q-function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ of policy π is defined analogously: $Q^\pi(s, a) := \mathbb{E}\left[\sum_{h=0}^H r_h(s_h, a_h) \mid s_0 = s, a_0 = a, a_h = \pi_h(s_h), \forall h \geq 0\right]$. Note that although we work with undiscounted episodic case, it is straightforward to extend the framework and analysis to discounted MDP. We define the expected value of a policy π:

$$J(\pi) := \mathbb{E}_{s\sim\rho}[V^\pi(s)] = \sum_{s\in\mathcal{S}} \rho(s) V^\pi(s).$$

We use shorthands $V^\star := V^{\pi^\star}$ and $Q^\star := Q^{\pi^\star}$ to denote the optimal value function and the optimal Q-function. We define the sub-optimality of any policy π as

$$\mathsf{SubOpt}(\pi) := J(\pi^\star) - J(\hat{\pi}).$$

We use shorthands $V^\star := V^{\pi^\star}$ and $Q^\star := Q^{\pi^\star}$ to denote the optimal value function and the optimal Q-function. We define the sub-optimality of any policy π as

$$\mathsf{SubOpt}(\pi) := J(\pi^\star) - J(\hat{\pi}).$$

We also define the state occupancy measures $d^\pi : \mathcal{S} \mapsto [0, H]$ and state-action occupancy measures $d^\pi : \mathcal{S} \times \mathcal{A} \mapsto [0, H]$ as $d^\pi(s) := \sum_{h=0}^H \mathbb{P}_h(s_h = s \mid \pi), d^\pi(s, a) := \sum_{h=0}^H \mathbb{P}_h(s_h = s; a_h = a \mid \pi)$, where we use $\mathbb{P}_h(s_h = s \mid \pi)$ to denote the probability of visiting state $s_h = s$ (and similarly $s_h = s, a_h = a$) at step h after executing policy π and starting from $s_0 \sim \rho(\cdot)$. Throughout the paper, we make the following assumption on the parameterization of the reward:

Assumption B.1. The reward lies in the family of linear functions $r_\theta(s, a) = \theta^\top \phi(s, a)$ for some known $\phi(s, a)$ with $\max_{s,a} \|\phi(s, a)\|_2 \leq L$. Let $\theta^\star$ be the true parameter. To ensure the identifiability of $\theta^\star$, we let $\theta^\star \in \Theta_B$, where

$$\Theta_B = \{\theta \in \mathbb{R}^d \mid \langle 1, \theta\rangle = 0, \|\theta\|_2 \leq B\}.$$

B.2 SAMPLING PROCEDURE AND COMPARISON MODEL

As in Ouyang et al. (2022), we assume that both the states and actions in the training set come from a pre-collected dataset. In a contextual bandit, for the i-th sample, a state (prompt) s^i is first sampled from some fixed distribution ρ. Given the state s^i, K actions $(a_0^i, a_1^i, \cdots, a_{K-1}^i)$ are sampled from some joint distribution $\mathbb{P}(a_0, \cdots, a_{K-1} \mid s^i)$[2]. Let $\sigma^i : [K] \mapsto [K]$ be the output of the human labeller, which is a permutation function that denotes the ranking of the actions. Here $\sigma^i(0)$ represents the most preferred action. We use $a_0 > a_1$ to denote the event that the action a_0 is more preferred compared to a_1. A common model on the distribution of σ under K-ary comparisons is a Plackett-Luce model (Plackett, 1975; Luce, 2012). The Plackett-Luce model defines the probability

[2]Indeed, it is not necessary to only compare actions under the same state. Our results can be easily generalized to the case when the states for K queries are completely different.

of a state-action pair (s, a_i) being the largest among a given set $\{(s, a_i)\}_{i=0}^{K-1}$ as

$$\mathbb{P}(a_i > a_j, \forall j \neq i \mid s) = \frac{\exp(r_\theta(s, a_i))}{\sum_{j=0}^{K-1} \exp(r_\theta(s, a_j))}.$$

Moreover, one can calculate the probability of observing the permutation σ as[3]

$$\mathbb{P}(\sigma \mid s, \{a_i\}_{i=0}^{K-1}) = \prod_{i=0}^{K-1} \frac{\exp(r_{\theta^\star}(s, a_{\sigma(i)}))}{\sum_{j=i}^{K-1} \exp(r_{\theta^\star}(s, a_{\sigma(j)}))}.$$

When $K = 2$, this reduces to the pairwise comparison considered in the BTL model, which is used in existing RLHF algorithms. In this case, the permutation σ can be reduced to a Bernoulli random variable, representing whether a_0 is preferred compared to a_1. Concretely, for each queried state-actions pair (s, a_0, a_1), we observe a sample y from a Bernoulli distribution with parameter $\frac{\exp(r_{\theta^\star}(s,a_1))}{\exp(r_{\theta^\star}(s,a_0))+\exp(r_{\theta^\star}(s,a_1))}$; i.e., for any $l \in \{0, 1\}$,

$$\mathbb{P}(y = l \mid s, a_0, a_1) = \frac{\exp(r_{\theta^\star}(s, a_l))}{\exp(r_{\theta^\star}(s, a_0)) + \exp(r_{\theta^\star}(s, a_1))}.$$

B.3 Organization

Section C presents the problem of learning with pairwise comparisons under the contextual bandit framework, we provide upper and lower bounds for MLE and pessimistic MLE. We extend the result into K-wise comparisons in Section D and MDP in Section E. We discuss the guarantee for IRL in Section F. We present our experimental results on simulated dataset in Section G. We also discuss the analysis for nonlinear rewards in Section I .

C Learning from Pairwise Comparison

We begin with the problem of learning from pairwise comparisons under the BTL model.

C.1 Algorithms: MLE and Pessimistic MLE

We first bound the estimation error for MLE, the most common algorithm in learning to rank and RLHF Liu et al. (2009); Xia et al. (2008); Cao et al. (2007); Christiano et al. (2017a); Ouyang et al. (2022). For any query-observation dataset $\{(s^i, a_1^i, a_2^i, y^i)\}_{i=1}^n$, MLE aims at minimizing the negative log likelihood, defined as:

$$\begin{aligned}
&\hat{\theta}_{\mathsf{MLE}} \in \arg\min_{\theta \in \Theta_B} \ell_{\mathcal{D}}(\theta),\\
&\ell_{\mathcal{D}}(\theta) = -\sum_{i=1}^n \log\Big(\frac{1(y^i = 1) \cdot \exp(r_\theta(s^i, a_1^i))}{\exp(r_\theta(s^i, a_0^i)) + \exp(r_\theta(s^i, a_1^i))} + \frac{1(y^i = 0) \cdot \exp(r_\theta(s^i, a_0^i))}{\exp(r_\theta(s^i, a_0^i)) + \exp(r_\theta(s^i, a_1^i))}\Big)\\
&\quad = -\sum_{i=1}^n \log\Big(1(y^i = 1) \cdot \mathsf{sigmoid}(\langle\theta,\, \phi(s^i, a_1^i) - \phi(s^i, a_0^i)\rangle) + 1(y^i = 0) \cdot \mathsf{sigmoid}(\langle\theta,\, \phi(s^i, a_0^i) - \phi(s^i, a_1^i)\rangle)\Big).
\end{aligned}$$

When the minimizer is not unique, we take any of the $\hat{\theta}$ that achieve the minimum. Let $\mathcal{D} = \{(s^i, a_1^i, a_2^i)\}_{i=1}^n$ denote the queried state-action pairs. In this paper, we study how one can utilize $\mathcal{D}$ to learn a near-optimal reward model and policy. We first present a lemma on the estimation error conditioned on the data $\mathcal{D}$. The lemma is a generalization of the upper bound in Shah et al. (2015, Theorem 1) and the analysis follows a similar structure. The main difference is that Shah et al. (2015) focus on the tabular case when $\phi(s, a)$ is always a unit vector, while in our case $\phi(s, a)$ can be an arbitrary d-dimensional vector.

[3] In practice, one may introduce an extra temperature parameter σ and replace all $r_{\theta^\star}$ with $r_{\theta^\star}/\sigma$. Here we take $\sigma = 1$.

Algorithm 2 Pessimistic MLE

Input: The current estimator $\hat{\theta}$, the data covariance $\Sigma_{\mathcal{D}}$, the regularization parameter λ, the bound on the semi-norm $f(n, d, \delta, \lambda)$, a reference vector $v \in \mathbb{R}^d$, state distribution q

Construct the confidence set

$$\Theta(\hat{\theta}, \lambda) = \Big\{\theta \in \Theta_B \mid \|\hat{\theta} - \theta\|_{\Sigma_{\mathcal{D}}+\lambda I} \leq f(n, d, \delta, \lambda)\Big\}.$$

Compute the pessimistic expected value function

$$\begin{aligned}\hat{J}(\pi) &= \min_{\theta \in \Theta(\hat{\theta}, \lambda)} \mathbb{E}_{s \sim q}[\theta^\top(\phi(s, \pi(s)) - v)] \\ &= (\mathbb{E}_{s\sim q}[\phi(s, \pi(s))] - v)^\top \hat{\theta} - \|(\Sigma_{\mathcal{D}} + \lambda I)^{-\frac{1}{2}}(\mathbb{E}_{s\sim q}[\phi(s, \pi(s))] - v)\|_2 \cdot f(n, d, \delta, \lambda)\end{aligned}$$

Return: $\hat{\pi} = \arg\max_\pi \hat{J}(\pi)$.

Lemma C.1. *For any $\lambda > 0$, with probability at least $1 - \delta$,*

$$\|\hat{\theta}_{\mathsf{MLE}} - \theta^\star\|_{\Sigma_{\mathcal{D}}+\lambda I} \leq C \cdot \sqrt{\frac{d + \log(1/\delta)}{\gamma^2 n} + \lambda B^2}.$$

Here $\Sigma_{\mathcal{D}} = \frac{1}{n}\sum_{i=1}^n (\phi(s^i, a_1^i) - \phi(s^i, a_0^i))(\phi(s^i, a_1^i) - \phi(s^i, a_0^i))^\top$, $\gamma = 1/(2 + \exp(-LB) + \exp(LB))$.

The proof is deferred to Appendix J.1. The optimality of the bound can be seen via a lower-bound argument akin to that in Shah et al. (2015, Theorem 1).

Now consider the set of parameters

$$\Theta(\hat{\theta}_{\mathsf{MLE}}, \lambda) = \Big\{\theta \in \Theta_B \mid \|\hat{\theta}_{\mathsf{MLE}} - \theta\|_{\Sigma_{\mathcal{D}}+\lambda I} \leq C \cdot \sqrt{\frac{d + \log(\frac{1}{\delta})}{\gamma^2 n} + \lambda B^2}\Big\}.$$

Lemma C.1 shows that with probability at least $1 - \delta$, one has $\theta^\star \in \Theta(\hat{\theta}_{\mathsf{MLE}})$. We thus consider the pessimistic MLE in Algorithm 2, which takes the lower confidence bound (LCB) as the reward estimate. In the context of LLM, the features of meaningful prompts and responses usually lie on a low-dimensional manifold. The idea of pessimism is to assign larger reward for the responses that lie on the manifold, and penalize the rarely seen responses that do not lie on manifold. We have the following guarantee for pessimistic MLE:

Theorem C.2. *Let $\hat{\pi}_{\mathsf{PE}}$ be the output of Algorithm 2 when taking $\hat{\theta} = \hat{\theta}_{\mathsf{MLE}}$, $f(n, d, \delta, \lambda) = C \cdot \sqrt{\frac{d+\log(1/\delta)}{\gamma^2 n} + \lambda B^2}$, $q = \rho$. For any $\lambda > 0$ and $v \in \mathbb{R}^d$, with probability at least $1 - \delta$,*

$$\mathsf{SubOpt}(\hat{\pi}_{\mathsf{PE}}) \leq C \cdot \sqrt{\frac{d + \log(1/\delta)}{\gamma^2 n} + \lambda B^2} \cdot \|(\Sigma_{\mathcal{D}} + \lambda I)^{-1/2}\mathbb{E}_{s\sim\rho}[(\phi(s, \pi^\star(s)) - v)]\|_2.$$

The proof is deferred to Appendix J.2. We make several remarks.

Remark C.3 (**The single concentratability coefficient assumption**). When $v = 0$, the term $\|(\Sigma_{\mathcal{D}} + \lambda I)^{-1/2}\mathbb{E}_{s\sim\rho}[\phi(s, \pi^\star(s))]\|_2$ is referred to as a "single concentratability coefficient", which is assumed to be bounded in most of the literature on offline learning (Rashidinejad et al., 2021; Li et al., 2022; Xie et al., 2021b; Zanette, 2022; Zanette et al., 2021). A bounded concentratability coefficient can be understood as certifying good coverage of the target vector $\mathbb{E}_{s\sim\rho}[\phi(s, \pi^\star(s))]$ from the dataset $\mathcal{D}$ in the feature space. The performance guarantee also holds when we replace $\pi^\star$ with any reference policy π on both sides.

Remark C.4 (**The choice of** λ). When $\Sigma_{\mathcal{D}}$ is invertible, or when any $\theta \in \Theta_B$ is orthogonal to the nullspace of $\Sigma_{\mathcal{D}}$, the above inequality holds for the case of $\lambda = 0$. In other cases, one may minimize λ on the right-hand side, or simply take $\lambda = (d + \log(1/\delta)/(B^2\gamma^2 n))$ to achieve a near-optimal rate up to a constant factor.

Remark C.5 (**The choice of** v). Compared to the traditional pessimism principle (Rashidinejad et al., 2021; Li et al., 2022; Xie et al., 2021b; Zanette, 2022; Zanette et al., 2021), we subtract an extra reference vector v in all the feature vectors ϕ. Subtracting a constant vector in feature space will not change the induced policy, but may affect the concentratability coefficient $\|(\Sigma_{\mathcal{D}} + \lambda I)^{-1/2}(\mathbb{E}_{s\sim\rho}[\phi(s, \pi(s))] - v)\|_2$.

We briefly describe the reason for introducing v here. Consider the case where the differences between features lie in the same subspace, while the feature ϕ itself does not. As a concrete example, consider a single state s and two actions a_0, a_1, we let $\phi(s, a_0) = (1, 1)$ and $\phi(s, a_1) = (1, 0)$. The data covariance is $(\phi(s^i, a_1^i) - \phi(s^i, a_0^i))(\phi(s^i, a_1^i) - \phi(s^i, a_0^i))^\top = [0, 0; 0, 1]$. Thus $\|(\Sigma_{\mathcal{D}} + \lambda I)^{-1/2}\phi(s, a_0)\|_2$ can be arbitrarily large as $\lambda \to 0$ when $v = 0$. On the other hand, when we take $v = \phi(s, a_1)$, one can verify that $\|(\Sigma_{\mathcal{D}} + \lambda I)^{-1/2}(\phi(s, a_0) - v)\|_2 \le 1$.

The above example illustrates the importance of choosing an appropriate v. A good rule of thumb for choosing v is the most common feature vector ϕ that appears in the data, so that more features can be covered. This also affords additional design latitude for other pessimism algorithms.

Remark C.6 (**Implementation for neural network**). When r_θ is a neural network, Algorithm 2 may not be directly implementable. As an alternative, there has been a number of heuristic approximations considered, including Conservative Q-Learning (Kumar et al., 2020), Implicit Q-Learning (Kostrikov et al., 2021) and Adversarially Trained Actor Critic (Cheng et al., 2022). Furthermore, one may also introduce pessimism in the policy training procedure. For example, Ouyang et al. (2022) add regularization terms in policy training, which enforces that the policy stays close to the original policy, and within the coverage of the pre-trained dataset. Our analysis supplies a theoretical rationale for such regularization terms.

Remark C.7 (**Implications for online learning**). Although we mainly focus on offline learning, Lemma C.1 also gives a straightforward online learning algorithm when combined with an optimism-based algorithm. In particular, a pure exploration-based active learning scheme would seek to compare pairs of actions whose feature difference is poorly covered by the past observations; i.e., find (s, a_1, a_2) such that $\|\phi(s, a_1) - \phi(s, a_2)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}$ is maximized. As a corollary of Lemma C.1 and exploration results for linear bandits (Abbasi-Yadkori et al., 2011; Soare et al., 2014), one can derive tight regret bound for online learning.

Remark C.8 (**Special Case: Multi-Armed Bandit**). For multi-armed bandits we have only a single state, such that the feature $\phi(s, a)$ reduces to $\vec{1}_a$, which is a unit vector with 1 on its a-th element. In this case, the data covariance reduces to a Laplacian matrix, defined as $\Sigma_{\mathcal{D}} = \frac{1}{n}\sum_{i=1}^n(\vec{1}_{a_1} - \vec{1}_{a_0})(\vec{1}_{a_1} - \vec{1}_{a_0})^\top$. This is precisely the problem considered in Shah et al. (2015). The Laplacian matrix is positive semidefinite and always has a zero eigenvalue, corresponding to an all ones eigenvector. When the graph induced by the Laplacian matrix is connected, any θ with $\langle 1, \theta\rangle = 0$ is orthogonal to the nullspace of $\Sigma_{\mathcal{D}}$, thus the theorem holds for the case of $\lambda = 0$.

C.2 Failure of MLE and Lower Bounds

We also show that there exists a simple linear bandit where MLE fails and pessimistic MLE succeeds. Let $\hat{\pi}_{\mathsf{MLE}} = \arg\max_\pi \mathbb{E}[r_{\hat{\theta}_{\mathsf{MLE}}}(s, \pi(s))]$ be the greedy policy with respect to the MLE.

Theorem C.9. *There exists a linear bandit with four actions and a sampling distribution such that for any* $n > 1$,

$$\mathbb{E}[\mathsf{SubOpt}(\hat{\pi}_{\mathsf{MLE}})] \ge 0.1.$$

On the other hand, with probability at least $1 - \delta$,

$$\mathsf{SubOpt}(\hat{\pi}_{\mathsf{PE}}) \le \frac{C \cdot \log(1/\delta)}{\sqrt{n}}.$$

Here C *is some universal constant.*

The proof is deferred to Appendix J.3. The results show a separation between MLE and pessimistic MLE when the concentratability coefficient is bounded.

We also show that for the problems with bounded concentratability coefficient, pessimistic MLE is minimax-rate optimal up to a constant factor. Consider the family of contextual bandit instances as follows:

$$\mathsf{CB}(\Lambda) = \{\rho, \{(s_i, a_{i1}, a_{i2})\}_{i=1}^n, \theta^\star \mid \|\Sigma_{\mathcal{D}}^{-1/2}\mathbb{E}_{s\sim\rho}[\phi(s, \pi^\star(s))]\|_2 \leq \Lambda\}.$$

Here we assume that $\Sigma_{\mathcal{D}}$ is invertible to simplify the presentation of the lower bound. For any $\mathcal{Q} \in \mathsf{CB}(\Lambda)$, we let $\mathsf{SubOpt}_{\mathcal{Q}}(\pi)$ be the sub-optimality under instance $\mathcal{Q}$. We have the following lower bound result, the proof of which is deferred to Appendix J.4.

Theorem C.10. *For any $d > 6, n \geq Cd\Lambda^2, \Lambda \geq 2$, there exists a feature mapping ϕ such that the following lower bound holds.*

$$\inf_{\hat{\pi}} \sup_{\mathcal{Q}\in\mathsf{CB}(\Lambda)} \mathsf{SubOpt}_{\mathcal{Q}}(\hat{\pi}) \geq C\Lambda \cdot \sqrt{\frac{d}{n}}.$$

Comparing with the upper bound in Theorem C.2, we see that the pessimistic MLE is minimax-optimal up to constant factors for the sub-optimality of induced policy.

D LEARNING FROM K-WISE COMPARISONS

We now consider learning from K-wise comparisons under the PL model. In this case, we design two different estimators based on MLE. One involves directly maximizing the likelihood under the PL model, denoted as MLE_K. The other involves splitting the K-wise comparison data with pairwise comparisons and running MLE for pairwise comparisons. We denote this estimator as MLE_2.

D.1 ALGORITHMS

Guarantee for MLE_K. Let $\mathcal{D} = \{(s^i, a_0^i, \cdots, a_K^i)\}_{i=1}^n$ be the set of queried states and actions, and the permutation function σ^i be the output of the i-th query. We can compute its maximum likelihood estimator as

$$\hat{\theta}_{\mathsf{MLE}_K} \in \arg\min_{\theta\in\Theta_B} \ell_{\mathcal{D}}(\theta),$$
$$\text{where } \ell_{\mathcal{D}}(\theta) = -\frac{1}{n}\sum_{i=1}^n \sum_{j=0}^{K-1} \log\left(\frac{\exp(\langle\theta,\, \phi(s^i, a^i_{\sigma_i(j)})\rangle)}{\sum_{k=j}^{K-1}\exp(\langle\theta,\, \phi(s^i, a^i_{\sigma_i(k)})\rangle)}\right).$$

Similar to Shah et al. (2015), we restrict our attention to $K = \mathcal{O}(1)$ since it is known that it is difficult for human to compare more than a small number of items due to a limited information storage and processing capacity (Miller, 1956; Kiger, 1984; Shiffrin and Nosofsky, 1994; Saaty and Ozdemir, 2003). For instance, Saaty and Ozdemir (2003) recommend eliciting preferences over no more than seven options. We have the following result for K-wise comparisons.

Theorem D.1. *Under the K-wise PL model, for any $\lambda > 0$, with probability at least $1 - \delta$,*

$$\|\hat{\theta}_{\mathsf{MLE}_K} - \theta^\star\|_{\Sigma_{\mathcal{D}}+\lambda I} \leq C \cdot \sqrt{\frac{K^4(d + \log(1/\delta))}{\gamma^2 n} + \lambda B^2}.$$

Here $\Sigma_{\mathcal{D}} = \frac{2}{K(K-1)n}(\sum_{i=1}^n \sum_{j=0}^{K-1}\sum_{k=j+1}^{K-1}(\phi(s^i, a_j^i) - \phi(s^i, a_k^i))(\phi(s^i, a_j^i) - \phi(s^i, a_k^i))^\top)$, and $\gamma = \exp(-4LB)$. As a consequence, let $\hat{\pi}_{\mathsf{PE}_K}$ be the output of Algorithm 2 when taking $\hat{\theta} = \hat{\theta}_{\mathsf{MLE}_K}$, $f(n, d, \delta, \lambda) = C \cdot \sqrt{\frac{K^4(d+\log(1/\delta))}{\gamma^2 n} + \lambda B^2}$. For any $\lambda > 0$ and $v \in \mathbb{R}^d$, with probability at least $1 - \delta$,

$$\mathsf{SubOpt}(\hat{\pi}_{\mathsf{PE}_K}) \leq C \cdot \sqrt{\frac{K^4(d + \log(1/\delta))}{\gamma^2 n} + \lambda B^2} \cdot \|(\Sigma_{\mathcal{D}} + \lambda I)^{-1/2}\mathbb{E}_{s\sim\rho}[(\phi(s, \pi^\star(s)) - v)]\|_2.$$

The proof of Theorem D.1 is provided in Appendix J.5. Shah et al. (2015) also study the extension from pairwise to K-wise comparisons. However, they focus on the setting where only the maximum is selected, where we assume a complete ranking among K items is given. Also, they only provide an expectation bound while we provide a high-probability bound.

Compared to the pairwise comparison result in Theorem C.2, the covariance matrix $\Sigma_{\mathcal{D}}$ now takes the sum over the feature differences between all pairs of actions among K-wise comparisons. As a cost, the right-hand side bound also introduces extra dependence on K. Our bound is likely to be loose in terms of the dependence on K. However, since we mainly focus on the case of $K = \mathcal{O}(1)$, such a bound is still near-optimal due to the minimax lower bound for pairwise comparisons. Furthermore, the gap between MLE and pessimistic MLE for sub-optimality still exists since Theorem C.9 holds as a special case of K-wise comparison.

Guarantee for MLE_2 Besides the standard MLE approach, another option is to replace the joint distribution of K-ranking data with $K(K-1)/2$ pairs of pairwise comparisons. This can be understood as replacing the true probability in MLE_K with the product of marginals:

$$\hat{\theta}_{\mathsf{MLE}_2} \in \arg\min_{\theta\in\Theta_B} \ell_{\mathcal{D}}(\theta),$$
$$\text{where } \ell_{\mathcal{D}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{K-1}\sum_{k=j+1}^{K-1}\log\left(\frac{\exp(\langle\theta,\ \phi(s^i, a^i_{\sigma_i(j)})\rangle)}{\exp(\langle\theta,\ \phi(s^i, a^i_{\sigma_i(j)})\rangle) + \exp(\langle\theta,\ \phi(s^i, a^i_{\sigma_i(k)})\rangle)}\right).$$

This estimator is also applied in the current RLHF for LLM (see, e.g., Ouyang et al., 2022). We show that it also leads to a good induced policy, as is shown in the theorem below.

Theorem D.2. *Under the K-wise PL model, for any $\lambda > 0$, with probability at least $1-\delta$,*

$$\|\hat{\theta}_{\mathsf{MLE}_2} - \theta^\star\|_{\Sigma_{\mathcal{D}}+\lambda I} \le C\cdot\sqrt{\frac{d+\log(1/\delta)}{\gamma^2 n} + \lambda B^2}.$$

Here $\Sigma_{\mathcal{D}} = \frac{2}{K(K-1)n}(\sum_{i=1}^{n}\sum_{j=0}^{K-1}\sum_{k=j+1}^{K-1}(\phi(s^i,a^i_j) - \phi(s^i,a^i_k))(\phi(s^i,a^i_j) - \phi(s^i,a^i_k))^\top)$, and $\gamma = 1/(2+\exp(-2LB)+\exp(2LB))$. As a consequence, let $\hat{\pi}_{\mathsf{PE}_2}$ be the output of Algorithm 2 when taking $\hat{\theta} = \hat{\theta}_{\mathsf{MLE}_2}$, $f(n,d,\delta,\lambda) = C\cdot\sqrt{\frac{d+\log(1/\delta)}{\gamma^2 n} + \lambda B^2}$, $q=\rho$. For any $\lambda > 0$ and $v\in\mathbb{R}^d$, with probability at least $1-\delta$,

$$\mathsf{SubOpt}(\hat{\pi}_{\mathsf{PE}_2}) \le C\cdot\sqrt{\frac{d+\log(1/\delta)}{\gamma^2 n} + \lambda B^2}\cdot\|(\Sigma_{\mathcal{D}}+\lambda I)^{-1/2}\mathbb{E}_{s\sim\rho}[(\phi(s,\pi^\star(s)) - v)]\|_2.$$

The proof of Theorem D.2 is provided in Appendix J.6. Our theoretical analysis validates the empirical performance of MLE_2 in Ouyang et al. (2022). Compared to the guarantee for MLE_K, MLE_2 seems to has better nonasymptotic upper bound in terms of the dependence on K. However, it is likely that this comes from a loose analysis of MLE_K. The MLE_2 belongs to the family of the M-estimators, whose asymptotic variance is known to be larger than that of MLE Godambe (1960); Lee (2008). Thus, asymptotically, MLE_K is more efficient than MLE_2. We can calculate the asymptotic variance of both estimators as follows:

Theorem D.3. *We have*

$$\sqrt{n}(\hat{\theta}_{\mathsf{MLE}_K} - \theta^\star) \to \mathcal{N}(0, \mathcal{I}(\theta^\star)^{-1});$$
$$\sqrt{n}(\hat{\theta}_{\mathsf{MLE}_2} - \theta^\star) \to \mathcal{N}(0, V).$$

where

$$\mathcal{I}(\theta^\star) = \mathbb{E}_{\theta^\star}\Bigg[\sum_{j=0}^{K-1}\sum_{k=j}^{K-1}\sum_{k'=j}^{K-1}\frac{\exp(\langle\theta^\star,\, \phi(s^i, a^i_{\sigma_i(k)}) + \phi(s^i, a^i_{\sigma_i(k')})\rangle)}{(\sum_{k'=j}^{K-1}\exp(\langle\theta^\star,\, \phi(s^i, a^i_{\sigma_i(k')})\rangle))^2}$$
$$\cdot\,(\phi(s^i, a^i_{\sigma_i(k)}) - \phi(s^i, a^i_{\sigma_i(k')}))(\phi(s^i, a^i_{\sigma_i(k)}) - \phi(s^i, a^i_{\sigma_i(k')}))^\top\Bigg],$$

$$V = \Sigma^{-1}\mathbb{E}_{\theta^\star}\left[GG^\top\right]\Sigma^{-1},$$

$$\Sigma = \mathbb{E}_{\theta^\star}\left[\sum_{j=0}^{K-1}\sum_{k=j}^{K-1}\frac{\exp(-\langle\theta^\star,\, x^i_{\sigma_i(j)\sigma_i(k)}\rangle)}{(1+\exp(-\langle\theta^\star,\, x^i_{\sigma_i(j)\sigma_i(k)}\rangle))^2}\cdot\Big(\phi(s^i, a^i_{\sigma_i(j)}) - \phi(s^i, a^i_{\sigma_i(k)})\Big)\Big(\phi(s^i, a^i_{\sigma_i(j)}) - \phi(s^i, a^i_{\sigma_i(k)})\Big)^\top\right],$$

$$G = \sum_{j=0}^{K-1}\sum_{k=j+1}^{K-1}\frac{\exp(-\langle\theta^\star,\, x^i_{\sigma_i(j)\sigma_i(k)}\rangle)}{1+\exp(-\langle\theta^\star,\, x^i_{\sigma_i(j)\sigma_i(k)}\rangle)}\cdot\Big(\phi(s^i, a^i_{\sigma_i(j)}) - \phi(s^i, a^i_{\sigma_i(k)})\Big)$$

The proof follows directly the gradient and Hessian computed in Appendix J.5 and J.6, combined with Van der Vaart (2000, Section 5.3). We also empirically verify the performances of both estimators in Section G.

E EXTENSION TO MDPS

Thus far we have considered only contextual bandits. We now extend our results to the MDP setting. Depending on whether the comparison is based on a single action or a whole trajectory, we have two regimes, namely action-based comparison and trajectory-based comparison.

E.1 TRAJECTORY-BASED COMPARISON

In trajectory-based comparison, we assume that two trajectories that start from the same initial state are given, and the comparison is based on the cumulative reward of the two trajectories. Concretely, we first sample the initial state s_0 from some fixed distribution ρ, and then sample two trajectories $\tau_0 = (a_0, s_1, a_1, \cdots, s_H, a_H)$ and $\tau_1 = (a_0', s_1', a_1', \cdots, s_H', a_H')$ from joint distributions $P_l(a_0, s_1, a_1, \cdots, s_H, a_H|s_0) = \prod_i \pi_l(a_i|s_i)P(s_{i+1}|s_i, a_i)$, where $l \in \{0, 1\}$. For each queried state-trajectory pair, we observe a sample y from a Bernoulli distribution as follows:

$$\mathbb{P}(y = 1 \mid s, \tau_0, \tau_1) = \frac{\exp(\sum_{h=0}^H r_{\theta^\star}(s_h, a_h))}{\exp(\sum_{h=0}^H r_{\theta^\star}(s_h, a_h)) + \exp(\sum_{h=0}^H r_{\theta^\star}(s_h', a_h'))}.$$

Given the dataset $\{(s^i, \tau_0^i, \tau_1^i, y^i\}_{i=1}^n$, the MLE is

$$\hat{\theta}_{\mathsf{MLE}} \in \mathop{\arg\min}_{\theta\in\Theta_B} \ell_{\mathcal{D}}(\theta),$$

where $\ell_{\mathcal{D}}(\theta) = -\sum_{i=1}^n \log\Big(\frac{1(y^i=1)\cdot\exp(\sum_{h=0}^H r_\theta(s_h^i, a_h^i))}{\exp(\sum_{h=0}^H r_\theta(s_h^i, a_h^i)) + \exp(\sum_{h=0}^H r_\theta(s_h^{i\prime}, a_h^{i\prime}))} + \frac{1(y^i=0)\cdot\exp(\sum_{h=0}^H r_\theta(s_h^{i\prime}, a_h^{i\prime}))}{\exp(\sum_{h=0}^H r_\theta(s_h^i, a_h^i)) + \exp(\sum_{h=0}^H r_\theta(s_h^{i\prime}, a}$

Compared to the pairwise comparison in the contextual bandit, the exponent changes from a single reward to the cumulative reward. Similarly, we provide the following guarantee for the estimation error of MLE:

Lemma E.1. *Assume that $\|\phi(\cdot,\cdot)\|_\infty \le L$ for any s, a. Then for any $\lambda > 0$, with probability at least $1-\delta$,*

$$\|\hat{\theta}_{\mathsf{MLE}} - \theta^\star\|_{\Sigma_{\mathcal{D}}+\lambda I} \le C\cdot\sqrt{\frac{d\log(1/\delta)}{\gamma^2 n} + \lambda B^2}.$$

Here $\Sigma_{\mathcal{D}} = \frac{1}{n}\sum_{i=1}^n(\sum_{h=0}^H(\phi(s_h^i, a_h^i) - \phi(s_h^{i\prime}, a_h^{i\prime})))\,(\sum_{h=0}^H(\phi(s_h^i, a_h^i) - \phi(s_h^{i\prime}, a_h^{i\prime})))^\top$, and $\gamma = 1/(2+\exp(-2HLB)+\exp(2HLB))$.

The proof is deferred to Appendix J.7. Compared to the guarantee for contextual bandits in Lemma C.1, the features in the covariance is now the difference between the cumulative feature in trajectory τ and the cumulative feature in trajectory τ'. The result reduces to Lemma C.1 when $H = 1$.

In order to bound the sub-optimality of the induced policy, one needs to plug-in a pessimistic version of the reward estimate. Note that from the definition of d^π, one has

$$\mathbb{E}_{s\sim\rho}[V^\pi(s)] = \mathbb{E}_{s,a\sim d^\pi}[r(s,a)].$$

In the case when the transition distribution P is known, one may directly compute d^π for any policy π and replace the initial distribution ρ in the algorithm for contextual bandit. This gives the following result:

Theorem E.2. *Let $\hat{\pi}_{\mathsf{PE}}$ be the output of Algorithm 2 when taking $\hat{\theta} = \hat{\theta}_{\mathsf{MLE}}$, $f(n, d, \delta, \lambda) = C \cdot \sqrt{\frac{d+\log(1/\delta)}{\gamma^2 n} + \lambda B^2}$, $q = d^\pi$. For any $\lambda > 0$ and $v \in \mathbb{R}^d$, with probability at least $1-\delta$,*

$$\mathsf{SubOpt}(\hat{\pi}_{\mathsf{PE}}) \leq C \cdot \sqrt{\frac{d+\log(1/\delta)}{\gamma^2 n} + \lambda B^2} \cdot \|(\Sigma_{\mathcal{D}} + \lambda I)^{-1/2} \mathbb{E}_{s\sim d^{\pi^\star}}[(\phi(s, \pi^\star(s)) - v)]\|_2.$$

The proof is deferred to Appendix J.8. The result can be generalized to the case of K-wise comparisons following the same argument in Section D.

E.2 Action-based Comparison

In action-based comparison, we assume that two actions are sampled for each state, and the comparison is based on the expected cumulative return starting from such state-action pair.

Concretely, assume that the optimal Q-function is parameterized as $Q^\star_\theta(s,a) = \theta^\top \phi(s,a)$ for some given $\phi(s,a)$. Let $\theta^\star$ be the true parameter. During the training, we first sample the state s from some fixed distribution ρ, and then sample a pair of actions a_0, a_1 from a joint distribution $P(a_0, a_1|s)$. For each queried state-actions pair (s, a_0, a_1), we observe a sample y from a Bernoulli distribution with parameter $\frac{\exp(Q_{\theta^\star}(s,a_1))}{\exp(Q_{\theta^\star}(s,a_0))+\exp(Q_{\theta^\star}(s,a_1))}$, i.e.

$$\mathbb{P}(y = 1 \mid s, a_0, a_1) = \frac{\exp(Q_{\theta^\star}(s,a_1))}{\exp(Q_{\theta^\star}(s,a_0)) + \exp(Q_{\theta^\star}(s,a_1))} \quad \text{and} \quad \mathbb{P}(y = 0 \mid s, a_0, a_1) = \frac{\exp(Q_{\theta^\star}(s,a_0))}{\exp(Q_{\theta^\star}(s,a_0)) + \exp(Q_{\theta^\star}(s,a_1))}$$

In this case, one may use the same MLE to estimate $\theta^\star$, which results in an estimator $\hat{Q}$ for the $Q^\star$-function. The following lemma follows exactly the same analysis as Lemma C.1:

Lemma E.3. *Under the BTL model for action-based RLHF, for any $\lambda > 0$, with probability at least $1-\delta$,*

$$\|\hat{\theta}_{\mathsf{MLE}} - \theta^\star\|_{\Sigma_{\mathcal{D}}+\lambda I} \leq C \cdot \sqrt{\frac{d+\log(1/\delta)}{\gamma^2 n} + \lambda B^2}.$$

Here $\gamma = 1/(2 + \exp(-LB) + \exp(LB))$. $\Sigma_{\mathcal{D}} = \frac{1}{n}\sum_{i=1}^n (\phi(s^i, a_1^i) - \phi(s^i, a_0^i))(\phi(s^i, a_1^i) - \phi(s^i, a_0^i))^\top$.

When $\Sigma_{\mathcal{D}}$ is invertible and covers all the directions well, this will lead to a valid confidence bound for $Q^\star$, which implies a good performance of the induced greedy policy without pessimism. However, when $\Sigma_{\mathcal{D}}$ does not provide good coverage, introducing pessimism in this case can be hard. The reason is that one needs to construct lower confidence bound for Q^π for any π. However, given such confidence bound of $\hat{\theta}_{\mathsf{MLE}}$, one can only construct confidence bound for $Q^\star$.

F Connection with Inverse Reinforcement Learning

In Inverse Reinforcement Learning (IRL), BTL and PL model are also popular model of human behavior. However, in IRL it is assumed that we only observe the human behavior, which is sampled from the distribution under PL model. Thus no comparison is queried. Depending on the comparison is action-based on trajectory-based, one has max-entropy IRL or action-based IRL, discussed in details below.

F.1 TRAJECTORY-BASED IRL

In max-entropy IRL Ziebart et al. (2008), it is also assumed that the human selection of trajectory follows a PL model. A common assumption in IRL or IL is that the observed trajectory collected by human behavior is likely to be the optimal policy. Assumee that the transitions are deterministic. For any trajectory $\tau = (s_0, a_0, \cdots, s_H, a_H)$, it is assumed that the expert chooses trajectory τ under the following model:

$$\mathbb{P}(\tau) = \frac{\exp(\sum_{h=0}^{H} \langle \theta^\star,\ \phi(s_h, a_h) \rangle)}{\sum_{\tau' \in \mathcal{T}(s_0)} \exp(\sum_{h=0}^{H} \langle \theta^\star,\ \phi(s'_h, a'_h) \rangle)}.$$

Here the set $\mathcal{T}(s_0)$ denotes the set for all possible trajectories that start from s_0. Each trajectory is represented by $\tau' = \{(s'_h, a'_h)\}_{h=1}^{H}$. Assume that we are given a set of trajectories $\{s_h^i, a_h^i\}_{i \in [n], h \in [H]}$ that are sampled from the distribution $\mathbb{P}(\tau)$. When the denominator can be computed exactly, the algorithm of max entropy IRL also reduces to the MLE, which can be written as

$$\hat{\theta}_{\mathsf{MLE}} \in \arg\min_{\theta \in \Theta_B} \ell_{\mathcal{D}}(\theta),$$
$$\text{where } \ell_{\mathcal{D}}(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{\exp(\sum_{h=0}^{H} \langle \theta,\ \phi(s_h^i, a_h^i) \rangle)}{\sum_{\tau' \in \mathcal{T}(s_0^i)} \exp(\sum_{h=0}^{H} \langle \theta,\ \phi(s'_h, a'_h) \rangle)} \right).$$

Although the enumeration of all trajectories $\mathcal{T}(s_0^i)$ is not possible due to exponential growth of the possible trajectories with respect to horizon H, Ziebart et al. (2008) provides an alternative way of computing the gradient via calculating the expected state frequency. This enables the efficient implementation of MLE. One can show the performance guarantee for max entropy IRL as follows:

Lemma F.1. *Under the PL model, for any $\lambda > 0$, with probability at least $1 - \delta$,*

$$\|\hat{\theta}_{\mathsf{MLE}} - \theta^\star\|_{\Sigma_{\mathcal{D}} + \lambda I} \le C \cdot \sqrt{\frac{\sup_s |\mathcal{T}(s)|^2 \cdot (d + \log(1/\delta))}{\gamma^2 n} + \lambda B^2}.$$

Here $\Sigma_{\mathcal{D}} = \frac{1}{n \sup_s |\mathcal{T}(s)|^2} \sum_{i=1}^{n} \sum_{\{(s_h, a_h)\} \in \mathcal{T}(s_0^i)} \sum_{\{(s'_h, a'_h)\} \in \mathcal{T}(s_0^i)} (\sum_{h=0}^{H} (\phi(s_h, a_h) - \phi(s'_h, a'_h)))(\sum_{h=0}^{H} (\phi(s_h, a_h) - \phi(s'_h, a'_h)))^\top$, *and* $\gamma = \exp(-4LB)/2$.

Given such guarantee for MLE, we also show that IRL, when combined with pessimism principle, will lead to a good policy.

Theorem F.2. *Let $\hat{\pi}_{\mathsf{PE}}$ be the output of Algorithm 2 when taking $\hat{\theta} = \hat{\theta}_{\mathsf{MLE}}$, $f(n, d, \delta, \lambda) = C \cdot \sqrt{\frac{\sup_s |\mathcal{T}(s)|(d + \log(1/\delta))}{\gamma^2 n} + \lambda B^2}$, $q = d^\pi$. For any $\lambda > 0$ and $v \in \mathbb{R}^d$, with probability at least $1 - \delta$,*

$$\mathsf{SubOpt}(\hat{\pi}_{\mathsf{PE}}) \le C \cdot \sqrt{\frac{\sup_s |\mathcal{T}(s)|^2 (d + \log(1/\delta))}{\gamma^2 n} + \lambda B^2} \cdot \|(\Sigma_{\mathcal{D}} + \lambda I)^{-1/2} \mathbb{E}_{s \sim \rho}[(\phi(s, \pi^\star(s)) - v)]\|_2.$$

The proof of Lemma F.1 and Theorem F.2 is provided in Appendix J.9. For IRL we have the dependence of $\sup_s |\mathcal{T}(s)|$ in our bound, which can be much larger than d. Similar to the case of K-wise comparison, one may also split the one observation into $\sup_s |\mathcal{T}(s)|$ pairwise comparisons, which can help improve the dependence on $\sup_s |\mathcal{T}(s)|$ in the current analysis.

F.2 ACTION-BASED IRL

Similar to action-based RLHF, action-based IRL also models human choice based on $Q^\star$ instead of cumulative reward Ramachandran and Amir (2007); Neu and Szepesvári (2009); Florence et al. (2022). Concretely, the human behavior is assumed to be based on the Q function $Q^\star(s, a) = \langle \theta^\star,\ \phi(s, a) \rangle$, i.e.

$$\pi^\star(a|s) = \frac{\exp(\langle \theta^\star,\ \phi(s, a) \rangle)}{\sum_{a' \in \mathcal{A}} \exp(\langle \theta^\star,\ \phi(s, a') \rangle)}.$$

Here the denominator takes all possible actions. Unlike RLHF where a pair of actions are observed, in IRL or IL, only a single human behavior is observed in each round and there is no comparison, i.e. the observed actions a are sampled from $\pi^\star(a \mid s)$. Given such observation, one can still run MLE and gives similar performance guarantee. In particular, the MLE is given by

$$\hat{\theta}_{\mathsf{MLE}} \in \arg\min_{\theta \in \Theta_B} \ell_{\mathcal{D}}(\theta),$$
$$\text{where } \ell_{\mathcal{D}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\exp(\langle\theta,\, \phi(s^i, a^i)\rangle)}{\sum_{a' \in \mathcal{A}} \exp(\langle\theta,\, \phi(s^i, a')\rangle)}\right).$$

The following lemma follows a similar analysis as Lemma C.1 and Lemma F.1:

Lemma F.3. *Under the PL model for action-based IRL, for any $\lambda > 0$, with probability at least $1 - \delta$,*

$$\|\hat{\theta}_{\mathsf{MLE}} - \theta^\star\|_{\Sigma_{\mathcal{D}} + \lambda I} \leq C \cdot \sqrt{\frac{|\mathcal{A}|^2(d + \log(1/\delta))}{\gamma^2 n} + \lambda B^2}.$$

Here $\Sigma_{\mathcal{D}} = \frac{1}{n|\mathcal{A}|^2}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}}\sum_{a'\in\mathcal{A}}(\phi(s^i, a) - \phi(s^i, a'))(\phi(s^i, a) - \phi(s^i, a'))^\top$, *and* $\gamma = \exp(-4LB)/2$.

Similar to the case of action-based RLHF, it remains an interesting open problem how one can introduce provable lower confidence bound algorithm for policy learning.

G Experiments

There has been a large amount of empirical work that demonstrates the success of MLE and pessimistic MLE in RLHF for game playing (Knox and Stone, 2008; MacGlashan et al., 2017; Christiano et al., 2017a; Warnell et al., 2018), robotics (Brown et al., 2019; Shin et al., 2023) and language models (Ziegler et al., 2019; Stiennon et al., 2020; Wu et al., 2021; Nakano et al., 2021; Ouyang et al., 2022; Menick et al., 2022; Glaese et al., 2022; Gao et al., 2022; Bai et al., 2022a; Ganguli et al., 2022; Ramamurthy et al., 2022). Notably, the concurrent work Shin et al. (2023) proposes Offline Preference-Based Reward Learning (OPRL), which trains pessimistic policy from the learned reward and shows empirically the superior performance of pessimistic based method (which can be viewed as an approximation of pessimistic MLE).

In this section, we provide experiments for the contextual bandit case. In particular, we conduct both MLE and pessimistic MLE on the example constructed in Appendix J.3. The results are included in Fig. 1. We range the number of samples n from 10 to 500. Each sample size is repeated 100 times. The result verifies our theoretical analysis: MLE converges under the semi-norm but fails to give good policy. On the other hand, pessimistic MLE gives vanishing rate when considering the sub-optimality of the induced policy. Note that in the left figure we do not include pessimistic MLE, since both MLE and pessimistic MLE rely on the same parameter $\hat{\theta}_{\mathsf{MLE}}$, and they only defer in how the induced policy is trained.

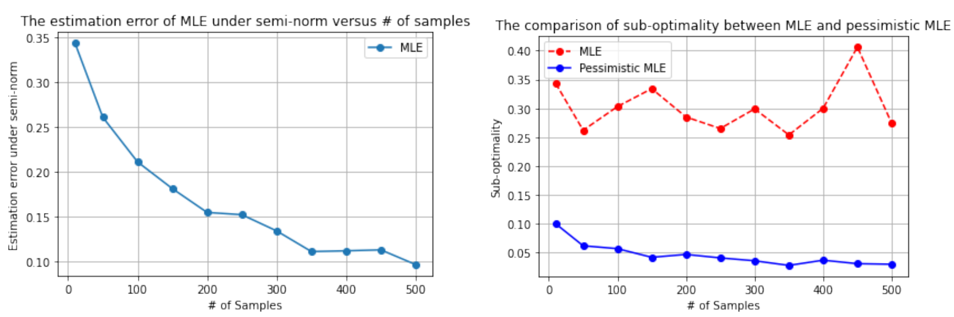

Figure 1: Left: the convergence of MLE under the semi-norm $\|\cdot\|_\Sigma$; Right: the comparison between MLE and pessimistic MLE under sub-optimality metric.

On the other hand, we compare the performance of MLE_2 and MLE_K when learning from K-wise comparisons. We take $K = 4$ and $K = 9$, and range samples from 10 to 500. We randomly generate ϕ and $\theta^\star$ as independent samples from 3-dimensional Gaussian distribution. The result is shown in Figure 2. One can see that as n grows larger, both estimators converge, while MLE_K has smaller estimation error than MLE_2. The gap grows larger when K becomes larger. This is consistent with our theoretical prediction in Section D: since MLE_K is the true MLE and MLE_2 belongs to the family of M-estimators, asymptotically MLE_K shall be more efficient than MLE_2.

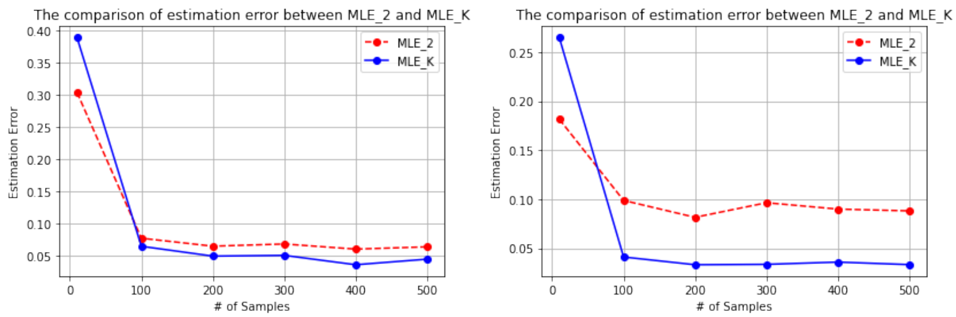

Figure 2: The comparison of estimation error between MLE_2 and MLE_K, with $K = 4$ in the left and $K = 9$ in the right.

H Conclusion

We have provided a theoretical analysis of the sample complexity of RLHF. Our main results involve two insights: (i) pessimism is important to guarantee a good policy; (ii) in K-wise comparison, both MLE_K and MLE_2 converge. Moreover, MLE_K is asymptotically more efficient.

While we have made progress in understanding the reward learning aspect of RLHF, there are many additional questions that remain to be answered.

1. We assumed that the policy trained is greedy with respect to the learned reward. However, in practice the reward is mostly used to fine-tune the pre-trained policy. This requires a more extensive theory that considers the whole procedure of pre-training the policy, learning a reward model and then fine-tuning the policy with policy gradient or PPO.
2. Although we focused on the BTL and PL models, there have been a number of other models considered for the modeling of human behavior, including the Thurstone model and cardinal models. It would be interesting to extend our analysis to cover these additional models and begin to provide a general characterization of behavioral models for RLHF.
3. Our constructed confidence bound is based on a fixed feature ϕ. In the practical fine-tuning scenario, ϕ is not fixed but may change slowly. It is interesting to see how the constructed confidence bound helps in the practical fine-tuning scenario for online (active) learning or offline learning, and how one can design valid confidence bound for slowly changing ϕ.

I Analysis for nonlinear r_θ

Consider the case of pairwise comparison when r_θ is not linear, the MLE can be written as

$$\hat{\theta}_{\mathsf{MLE}} \in \arg\min_{\theta \in \Theta_B} \ell_{\mathcal{D}}(\theta),$$

where $\ell_{\mathcal{D}}(\theta) = -\sum_{i=1}^{n} \log\left(1(y^i = 1) \cdot \frac{\exp(r_\theta(s^i, a_1^i))}{\exp(r_\theta(s^i, a_0^i)) + \exp(r_\theta(s^i, a_1^i))} + 1(y^i = 0) \cdot \frac{\exp(r_\theta(s^i, a_0^i))}{\exp(r_\theta(s^i, a_0^i)) + \exp(r_\theta(s^i, a_1^i))}\right).$

Here we provide a guarantee for the case when r_θ is nonlinear and non-convex. We first make the following boundedness and smoothness assumption on r_θ:

Assumption I.1. Assume that for any $\theta \in \Theta_B, s \in \mathcal{S}, a_0 \in \mathcal{A}, a_1 \in \mathcal{A}$ with $a_0 \neq a_1$, we have,

$$\begin{aligned} |r_\theta(s,a)| &\leq \alpha_0, \quad \text{(Bounded value)} \\ \|\nabla r_\theta(s,a)\|_2 &\leq \alpha_1, \quad \text{(Bounded gradient)} \\ \|\nabla^2 r_\theta(s,a)\|_2 &\leq \alpha_2. \quad \text{(Bounded Hessian / Lipschitz gradient)} \end{aligned}$$

One can verify that our linear reward satisfies the above assumption with $\alpha_0 = LB, \alpha_1 = L, \alpha_2 = 0$. Under this assumption, we have

Theorem I.2. *For any $\lambda > 0$, with probability at least $1 - \delta$,*

$$\|\hat{\theta}_{\mathsf{MLE}} - \theta^\star\|_{\Sigma_\mathcal{D}+\lambda I} \leq C \cdot \sqrt{\frac{d + \log(1/\delta)}{\gamma^2 n} + (\lambda + \alpha_2/\gamma + \alpha_1\alpha_2 B)B^2}.$$

Here $\gamma = \frac{1}{2+\exp(-2\alpha_0)+\exp(2\alpha_0)}$, $\Sigma_\mathcal{D} = \frac{1}{n}\sum_{i=1}^n \nabla(r_{\theta^\star}(s^i,a_1^i) - r_{\theta^\star}(s^i,a_0^i))\nabla(r_{\theta^\star}(s^i,a_1^i) - r_{\theta^\star}(s^i,a_0^i))^\top$.

The proof is deferred to Appendix J.10. Our result recovers Lemma C.1 when $\alpha_2 = 0$ and reveals how the gradient of r plays a role in the bound for estimation error. However, the dependence on α_2 will not vanish as $n \to \infty$. It remains open how to get vanishing rate for nonlinear reward functions. Similar argument can also be applied to the case of K-wise comparison and MDP.

On the other hand, we can show that the true parameter $\theta^\star$ is a global minimum of the population negative log likelihood even when r_θ is nonlinear and we use MLE_2 for K-wise comparison. Recall that the MLE_2 splits K-wise comparisons into pairwise comparisons, and is given by

$$\begin{aligned} &\hat{\theta}_{\mathsf{MLE}_2} \in \arg\min_{\theta\in\Theta_B} \ell_\mathcal{D}(\theta), \\ \text{where } \ell_\mathcal{D}(\theta) &= -\frac{1}{n}\sum_{i=1}^n\sum_{j=0}^{K-1}\sum_{k=j+1}^{K-1}\log\left(\frac{\exp(r_\theta(s^i,a^i_{\sigma_i(j)}))}{\exp(r_\theta(s^i,a^i_{\sigma_i(j)}))+\exp(r_\theta(s^i,a^i_{\sigma_i(k)}))}\right). \end{aligned}$$

When there is infinite number of data, the loss become

$$\begin{aligned} \mathbb{E}[\ell(\theta)] = -\sum_s \rho(s) \sum_{a_0,a_1\in\mathcal{A}} \rho(a_0,a_1 \mid s) \cdot \Big(&\frac{\exp(r_{\theta^\star}(s,a_0))}{\exp(r_{\theta^\star}(s,a_0))+\exp(r_{\theta^\star}(s,a_1))}\log\Big(\frac{\exp(r_\theta(s,a_0))}{\exp(r_\theta(s,a_0))+\exp(r_\theta(s,a_1))}\Big) \\ &+ \frac{\exp(r_{\theta^\star}(s,a_1))}{\exp(r_{\theta^\star}(s,a_0))+\exp(r_{\theta^\star}(s,a_1))}\log\Big(\frac{\exp(r_\theta(s,a_1))}{\exp(r_\theta(s,a_0))+\exp(r_\theta(s,a_1))}\Big)\Big). \end{aligned}$$

Here $\rho(a_0, a_1 \mid s)$ is the probability that actions a_0, a_1 are included in the K-comparison when the state is s. Now we show

$$\theta^\star \in \arg\min_\theta \mathbb{E}[\ell(\theta)]. \tag{1}$$

To see this, note that we have

$$\begin{aligned} \mathbb{E}[\ell(\theta)] &= -\sum_s \rho(s) \sum_{a_0,a_1\in\mathcal{A}} \rho(a_0,a_1 \mid s) \cdot \Big(\frac{\exp(r_{\theta^\star}(s,a_0))}{\exp(r_{\theta^\star}(s,a_0))+\exp(r_{\theta^\star}(s,a_1))}\log\Big(\frac{\exp(r_\theta(s,a_0))}{\exp(r_\theta(s,a_0))+\exp(r_\theta(s,a_1))}\Big) \\ &\quad + \frac{\exp(r_{\theta^\star}(s,a_1))}{\exp(r_{\theta^\star}(s,a_0))+\exp(r_{\theta^\star}(s,a_1))}\log\Big(\frac{\exp(r_\theta(s,a_1))}{\exp(r_\theta(s,a_0))+\exp(r_\theta(s,a_1))}\Big)\Big) \\ &= \sum_s \rho(s) \sum_{a_0,a_1\in\mathcal{A}} p(a_0,a_1 \mid s) \cdot (H(p_{\theta^\star}(s,a_0,a_1)) + \mathsf{KL}(p_{\theta^\star}(s,a_0,a_1)\|p_\theta(s,a_0,a_1))). \end{aligned}$$

Here $H(p) = p\log(1/p) + (1-p)\log(1/(1-p))$ is the entropy of a Bernoulli distribution with parameter p. And $p_\theta(s,a_0,a_1) = \frac{\exp(r_\theta(s,a_1))}{\exp(r_\theta(s,a_0))+\exp(r_\theta(s,a_1))}$. Now note that KL is lower bounded by 0, with equality when $\theta = \theta^\star$. This proves Equation (1).

J REMAINING PROOFS

J.1 PROOF OF LEMMA C.1

Recall that the MLE is given by

$$\hat{\theta}_{\mathsf{MLE}} \in \arg\min_{\theta \in \Theta_B} \ell_{\mathcal{D}}(\theta),$$

where
$$\begin{aligned}\ell_{\mathcal{D}}(\theta) &= -\sum_{i=1}^{n} \log\left(1(y^i=1)\cdot\frac{\exp(r_\theta(s^i,a_1^i))}{\exp(r_\theta(s^i,a_0^i))+\exp(r_\theta(s^i,a_1^i))} + 1(y^i=0)\cdot\frac{\exp(r_\theta(s^i,a_0^i))}{\exp(r_\theta(s^i,a_0^i))+\exp(r_\theta(s^i,a_1^i))}\right) \\ &= -\sum_{i=1}^{n} \log\left(1(y^i=1)\cdot\frac{1}{1+\exp(r_\theta(s^i,a_0^i)-r_\theta(s^i,a_1^i))} + 1(y^i=0)\cdot\left(1-\frac{1}{1+\exp(r_\theta(s^i,a_0^i)-r_\theta(s^i,a_1^i))}\right)\right) \\ &= -\sum_{i=1}^{n} \log\left(1(y^i=1)\cdot\frac{1}{1+\exp(\theta^\top(\phi(s^i,a_0^i)-\phi(s^i,a_1^i)))} + 1(y^i=0)\cdot\left(1-\frac{1}{1+\exp(\theta^\top(\phi(s^i,a_0^i)-\phi(s^i,a_1^i)))}\right)\right)\end{aligned}$$

To simplify the notation, we let $x_i = \phi(s^i, a_1^i) - \phi(s^i, a_0^i)$. Our goal is to bound the estimation error of the MLE in the squared semi-norm $\|v\|^2_{\Sigma_{\mathcal{D}}+\lambda I} = v^T(\Sigma_{\mathcal{D}} + \lambda I)v$.

Strong convexity of ℓ. We first show that $\ell_{\mathcal{D}}$ is strongly convex at $\theta^\star$ with respect to the semi-norm $\|\cdot\|_{\Sigma_{\mathcal{D}}}$, meaning that there is some constant $\gamma > 0$ such that

$$\ell_{\mathcal{D}}(\theta^\star + \Delta) - \ell_{\mathcal{D}}(\theta^\star) - \langle \nabla \ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle \geq \gamma \|\Delta\|^2_{\Sigma_{\mathcal{D}}} \tag{2}$$

for all perturbations $\Delta \in \mathbb{R}^d$ such that $\theta^\star + \Delta \in \Theta_B$.

One can directly calculate the Hessian of ℓ as

$$\begin{aligned}\nabla^2 \ell_{\mathcal{D}}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\left(1(y^i=1)\cdot\frac{\exp(-\langle\theta,\, x_i\rangle)}{(\exp(-\langle\theta,\, x_i\rangle)+1)^2} + 1(y^i=0)\cdot\frac{\exp(\langle\theta,\, x_i\rangle)}{(\exp(\langle\theta,\, x_i\rangle)+1)^2}\right)\cdot x_i x_i^T \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\exp(-\langle\theta,\, x_i\rangle)}{(\exp(-\langle\theta,\, x_i\rangle)+1)^2}\cdot x_i x_i^T\end{aligned}$$

Observe that $\langle\theta,\, x_i\rangle \in [-2LB, 2LB]$, which gives that

$$\frac{\exp(-\langle\theta,\, x_i\rangle)}{(\exp(-\langle\theta,\, x_i\rangle)+1)^2} \geq \frac{1}{2+\exp(-2LB)+\exp(2LB)}.$$

Putting together the pieces, we conclude that

$$v^T \nabla^2 \ell_{\mathcal{D}}(\theta) v \geq \frac{\gamma}{n}\|Xv\|_2^2 \qquad \text{for all } v,$$

where $\gamma = 1/(2 + \exp(-2LB) + \exp(2LB))$, $X \in \mathbb{R}^{n\times d}$ has the differencing vector $x_i \in \mathbb{R}^d$ as its i^{th} row. Thus, if we introduce the error vector $\Delta := \hat{\theta}_{\mathsf{MLE}} - \theta^\star$, then we may conclude that

$$\ell_{\mathcal{D}}(\theta^\star + \Delta) - \ell_{\mathcal{D}}(\theta^\star) - \langle \nabla \ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle \geq \frac{\gamma}{n}\|X\Delta\|_2^2 \;=\; \gamma\|\Delta\|^2_{\Sigma_{\mathcal{D}}},$$

showing that $\ell_{\mathcal{D}}$ is strongly convex around $\theta^\star$ with parameter γ.

Bounding the estimation error. Now we aim at bounding the estimation error $\|\hat{\theta}_{\mathsf{MLE}} - \theta^\star\|_{\Sigma_{\mathcal{D}}}$. Since $\hat{\theta}_{\mathsf{MLE}}$ is optimal for $\ell_{\mathcal{D}}$, we have $\ell_{\mathcal{D}}(\hat{\theta}_{\mathsf{MLE}}) \leq \ell_{\mathcal{D}}(\theta^\star)$. (When $\hat{\theta}_{\mathsf{MLE}}$ is approximately optimal, i.e. $\ell_{\mathcal{D}}(\hat{\theta}_{\mathsf{MLE}}) \leq \min_\theta \ell_{\mathcal{D}}(\theta) + \epsilon$, the same argument also holds up to an extra additive term ϵ.) Defining the error vector $\Delta = \hat{\theta}_{\mathsf{MLE}} - \theta^\star$, adding and subtracting the quantity $\langle \nabla \ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle$ yields the bound

$$\ell_{\mathcal{D}}(\theta^\star + \Delta) - \ell_{\mathcal{D}}(\theta^\star) - \langle \nabla \ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle \leq -\langle \nabla \ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle.$$

By the γ-convexity condition, the left-hand side is lower bounded by $\gamma\|\Delta\|^2_{\Sigma_\mathcal{D}}$. As for the right-hand side, note that $|\langle\nabla\ell_\mathcal{D}(\theta^\star),\, \Delta\rangle| \le \|\nabla\ell_\mathcal{D}(\theta^\star)\|_{(\Sigma_\mathcal{D}+\lambda I)^{-1}}\ \|\Delta\|_{\Sigma_\mathcal{D}+\lambda I}$ for any $\lambda > 0$. Altogether we have

$$\gamma\|\Delta\|^2_{\Sigma_\mathcal{D}} \le \|\nabla\ell_\mathcal{D}(\theta^\star)\|_{(\Sigma_\mathcal{D}+\lambda I)^{-1}}\ \|\Delta\|_{\Sigma_\mathcal{D}+\lambda I}.$$

Now we further bound the term $\|\nabla\ell_\mathcal{D}(\theta^\star)\|_{(\Sigma_\mathcal{D}+\lambda I)^{-1}}$. Observe that the gradient takes the form

$$\nabla\ell_\mathcal{D}(\theta^\star) = \frac{-1}{n}\sum_{i=1}^n \left[\mathbf{1}[y^i = 1]\frac{\exp(-\langle\theta^\star,\, x_i\rangle)}{1+\exp(-\langle\theta^\star,\, x_i\rangle))} - \mathbf{1}[y^i = 0]\frac{1}{1+\exp(-\langle\theta^\star,\, x_i\rangle))}\right] x_i.$$

Define a random vector $V \in \mathbb{R}^n$ with independent components as

$$V_i = \begin{cases} \frac{\exp(-\langle\theta^\star,\, x_i\rangle)}{1+\exp(-\langle\theta^\star,\, x_i\rangle))} & \text{w.p.} \quad \frac{1}{1+\exp(-\langle\theta^\star,\, x_i\rangle))} \\ \frac{-1}{1+\exp(-\langle\theta^\star,\, x_i\rangle))} & \text{w.p.} \quad \frac{\exp(-\langle\theta^\star,\, x_i\rangle)}{1+\exp(-\langle\theta^\star,\, x_i\rangle))}. \end{cases}$$

With this notation, we have $\nabla\ell_\mathcal{D}(\theta^\star) = -\frac{1}{n}X^T V$. One can verify that $\mathbb{E}[V] = 0$ and $|V_i| \le 1$.

Defining the n-dimensional square matrix $M := \frac{1}{n^2}X(\Sigma_\mathcal{D} + \lambda I)^{-1}X^T$, we have $\|\nabla\ell_\mathcal{D}(\theta^\star)\|^2_{(\Sigma_\mathcal{D}+\lambda I)^{-1}} = V^T M V$. Let the eigenvalue decomposition of $X^\top X$ be $X^\top X = U\Lambda U^\top$. We can bound the trace and operator norm of M as

$$\begin{aligned}
\mathsf{Tr}(M) &= \frac{1}{n^2}\mathsf{Tr}(U(\Lambda/n + \lambda I)^{-1}U^\top U\Lambda U^\top) \le \frac{d}{n} \\
\mathsf{Tr}(M^2) &= \frac{1}{n^4}\mathsf{Tr}(U(\Lambda/n + \lambda I)^{-1}U^\top U\Lambda U^\top U(\Lambda/n + \lambda I)^{-1}U^\top U\Lambda U^\top) \le \frac{d}{n^2} \\
\|M\|_{\mathsf{op}} &= \lambda_{\mathsf{max}}(M) \le \frac{1}{n},
\end{aligned}$$

Moreover, since the components of V are independent and of zero mean, and $|V_i| \le 1$, the variables are 1-sub-Gaussian, and hence 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 $1-\delta$,

$$\|\nabla\ell_\mathcal{D}(\theta^\star)\|^2_{(\Sigma_\mathcal{D}+\lambda I)^{-1}} = V^\top M V \le C_1 \cdot \frac{d + \log(1/\delta)}{n}.$$

Here C_1 is some universal constant. This gives us

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

Solving the above inequality gives us that for some constant C_2,

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

J.2 Proof of Theorem C.2

Proof. Let $J'(\pi) = J(\pi) - \langle\theta^\star,\, v\rangle$. We have

$$\begin{aligned}
\mathsf{SubOpt}(\hat\pi_{\mathsf{PE}}) &= J(\pi^\star) - J(\hat\pi_{\mathsf{PE}}) \\
&= J'(\pi^\star) - J'(\hat\pi_{\mathsf{PE}}) \\
&= (J'(\pi^\star) - \hat J(\pi^\star)) + (\hat J(\pi^\star) - \hat J(\hat\pi_{\mathsf{PE}})) + (\hat J(\hat\pi_{\mathsf{PE}}) - J'(\hat\pi_{\mathsf{PE}})).
\end{aligned}$$

Since $\hat\pi_{\mathsf{PE}}$ is the optimal policy under expected value $J'(\pi)$, we know that the second difference satisfies $\hat J(\pi^\star) - \hat J(\hat\pi_{\mathsf{PE}}) \le 0$. For the third difference, we have

$$\hat J(\hat\pi_{\mathsf{PE}}) - J'(\hat\pi_{\mathsf{PE}}) = \min_{\theta\in\Theta(\hat\theta_{\mathsf{MLE}},\lambda)} \mathbb{E}_{s\sim\rho}[\theta^\top(\phi(s,\pi(s)) - v)] - \mathbb{E}_{s\sim\rho}[\theta^{\star\top}(\phi(s,\pi(s)) - v)].$$

From Lemma C.1 we know that $\theta^\star \in \Theta(\hat{\theta}_{\mathsf{MLE}}, \lambda)$ with probability at least $1-\delta$. Thus we know that with probability at least $1-\delta$, $\hat{J}(\hat{\pi}_{\mathsf{PE}}) - J'(\hat{\pi}_{\mathsf{PE}}) \leq 0$. Now combining everything together and condition on the above event, we have

$$\begin{aligned}
\mathsf{SubOpt}(\hat{\pi}_{\mathsf{PE}}) &\leq J'(\pi^\star) - \hat{J}(\pi^\star) \\
&= \sup_{\theta \in \Theta(\hat{\theta}_{\mathsf{MLE}}, \lambda)} \mathbb{E}_{s\sim\rho}[(\theta^\star - \theta)^\top(\phi(s, \pi^\star(s)) - v)] \\
&= \sup_{\theta \in \Theta(\hat{\theta}_{\mathsf{MLE}}, \lambda)} \mathbb{E}_{s\sim\rho}[(\theta^\star - \hat{\theta}_{\mathsf{MLE}} + \hat{\theta}_{\mathsf{MLE}} - \theta)^\top(\phi(s, \pi^\star(s)) - v)] \\
&= \mathbb{E}_{s\sim\rho}[(\theta^\star - \hat{\theta}_{\mathsf{MLE}})^\top(\phi(s, \pi^\star(s)) - v)] + \sup_{\theta \in \Theta(\hat{\theta}_{\mathsf{MLE}}, \lambda)} \mathbb{E}_{s\sim\rho}[(\hat{\theta}_{\mathsf{MLE}} - \theta)^\top(\phi(s, \pi^\star(s)) - v)].
\end{aligned}$$

By the definition of $\Theta(\hat{\theta}_{\mathsf{MLE}}, \lambda)$, we know that for any $\theta \in \Theta(\hat{\theta}_{\mathsf{MLE}}, \lambda)$, one has $\mathbb{E}_{s\sim\rho}[(\hat{\theta}_{\mathsf{MLE}} - \theta)^\top(\phi(s, \pi^\star(s)) - v)] \leq C \cdot \sqrt{\frac{d+\log(1/\delta)}{\gamma^2 n} + \lambda B^2} \cdot \|(\Sigma_{\mathcal{D}} + \lambda I)^{-1/2}\mathbb{E}_{s\sim\rho}[\phi(s, \pi^\star(s)) - v]\|_2$. Furthermore, we know that $\theta^\star \in \Theta(\hat{\theta}_{\mathsf{MLE}}, \lambda)$ from Lemma C.1. Altogether we have with probability $1-\delta$

$$\mathsf{SubOpt}(\hat{\pi}_{\mathsf{PE}}) \leq 2C \cdot \sqrt{\frac{d+\log(1/\delta)}{\gamma^2 n} + \lambda B^2} \cdot \|(\Sigma_{\mathcal{D}} + \lambda I)^{-1/2}\mathbb{E}_{s\sim\rho}[\phi(s, \pi^\star(s)) - v]\|_2.$$

□

J.3 Proof of Theorem C.9

Proof. Consider 4 actions with parameter $\phi(a_1) = [1,1,0]$, $\phi(a_2) = [1,0,0]$, $\phi(a_3) = [0,0,0]$, $\phi(a_4) = [0,1,0]$. Let the true reward be $\theta^\star = [-1, 0.1, 0.9] \in \Theta_B$ with $B = 2$. We query $n-1$ times a_1, a_2 and 1 time a_2, a_3. For the single pairwise comparison result $Y_{2>3}$ between a_2 and a_3, we know that

$$P(Y_{2>3} = 1) = \frac{\exp((\phi(a_2) - \phi(a_3))^\top \theta^\star)}{1 + \exp((\phi(a_2) - \phi(a_3))^\top \theta^\star)} > 0.26.$$

Now conditioned on the event that $Y_{2>3} = 1$, we know that the MLE aims to find

$$\hat{\theta}_{\mathsf{MLE}} = \arg\min_{\theta \in \Theta_B} \ell_{\mathcal{D}}(\theta),$$

$$\begin{aligned}
\text{where } \ell_{\mathcal{D}}(\theta) &= -n_{1>2} \cdot \log\left(\frac{\exp((\phi(a_1) - \phi(a_2))^\top \theta)}{1 + \exp((\phi(a_1) - \phi(a_2))^\top \theta)}\right) - n_{1<2} \cdot \log\left(\frac{\exp((\phi(a_2) - \phi(a_1))^\top \theta)}{1 + \exp((\phi(a_2) - \phi(a_1))^\top \theta)}\right) \\
&\quad - \log\left(\frac{\exp((\phi(a_2) - \phi(a_3))^\top \theta)}{1 + \exp((\phi(a_2) - \phi(a_3))^\top \theta)}\right) \\
&= -n_{1>2} \cdot \log\left(\frac{\exp(\theta_2)}{1 + \exp(\theta_2)}\right) - n_{1<2} \cdot \log\left(\frac{\exp(-\theta_2)}{1 + \exp(-\theta_2)}\right) - \log\left(\frac{\exp(\theta_1)}{1 + \exp(\theta_1)}\right).
\end{aligned}$$

By concentration of $n_{1>2}$, we know that when $n > 500$, with probability at least 0.5, we have

$$n_{1<2} > 0.45n.$$

Under this case, the MLE will satisfy at $\hat{\theta}_1 > 0, \hat{\theta}_2 < 0.5$. Thus the policy based on MLE estimator will choose action a_1 or a_2 instead of the optimal action a_4 under the events above. The expected suboptimality is

$$\mathbb{E}[V^\star(s) - V^{\hat{\pi}_{\mathsf{MLE}}}(s)] \geq 0.26 * 0.5 * 1 > 0.1.$$

On the other hand, one can calculate the coverage as

$$\|\Sigma_{\mathcal{D}}^{-1/2}\mathbb{E}_{s\sim\rho}[\phi(s, \pi^\star(s))]\|_2 = \frac{n}{n-1}.$$

Thus by Theorem C.2 we know that pessimistic MLE achieves vanishing error. □

J.4 Proof of Theorem C.10

Proof. Assume without loss of generality that $d/3$ is some integer. We set $\mathcal{S} = [d/3]$, $\mathcal{A} = \{a_1, a_2, a_3, a_4\}$. For each of the s, a_i, we set $\phi(s, a_1) = e_{3s+1} + e_{3s+2}, \phi(s, a_2) = e_{3s+1}, \phi(s, a_3) = 0, \phi(s, a_4) = e_{3s+2}$. We set the initial distribution of states as $\rho = \mathsf{Unif}([1, 2, \cdots, S])$, the query times $n(s, a_1, a_2) = n/S \cdot (1 - 2/\Lambda^2), n(s, a_2, a_3) = n/S \cdot (2/\Lambda^2)$.

Let $v_{-1} = [1/d, 1/d + \Delta, -2/d - \Delta]$, $v_{+1} = [1/d + 2\Delta, 1/d + \Delta, -2/d - 3\Delta]$. We construct 2^S instances, indexed by $\tau \in \{\pm 1\}^S$, where each $\theta_\tau = [v_{\tau_1}, v_{\tau_2}, \cdots, v_{\tau_S}]$. One can see that $\mathbb{E}[V_{\mathcal{Q}}(\pi^\star) - V^\star_{\mathcal{Q}}(\hat{\pi})] = 1/S \cdot \sum_{s \in \mathcal{S}} (r_{\mathcal{Q}}(s, \pi^\star(s)) - r_{\mathcal{Q}}(s, \hat{\pi}(s)))$. Under each θ_τ, the optimal policy $\pi(s)$ is either a_2 or a_4. One can verify that $\|\Sigma_{\mathcal{D}}^{-1/2} \mathbb{E}_{s \sim \rho}[\phi(s, \pi^\star(s))]\|_2 \leq \Lambda$ and that $\theta_\tau \in \Theta_B$ with $B = 1$ when $d > 6$ and $\Delta < 1/6\sqrt{d}$.

Furthermore, for any $\theta_\tau, \theta_{\tau'}$ that differs only in the j-th coordinate of τ, we have

$$1/S \cdot (r_{\mathcal{Q}_\tau}(j, \pi^\star(j)) - r_{\mathcal{Q}_\tau}(j, \hat{\pi}(j)) + r_{\mathcal{Q}_{\tau'}}(j, \pi^\star(j)) - r_{\mathcal{Q}_{\tau'}}(j, \hat{\pi}(j))) \geq \Delta/S.$$

Thus by Assouad's lemma (see e.g. Yu (1997)), we have

$$\begin{aligned}\inf_{\hat{\pi}} \sup_{\mathcal{Q} \in \mathsf{CB}(\lambda)} \mathbb{E}[V_{\mathcal{Q}}(\pi^\star) - V^\star_{\mathcal{Q}}(\hat{\pi})] &\geq S \cdot \frac{\Delta}{2S} \min_{\tau \sim \tau'} (1 - \mathsf{TV}(\mathbb{P}_{\theta_\tau}, \mathbb{P}_{\theta_{\tau'}})) \\ &\geq \frac{\Delta}{4} \min_{\tau \sim \tau'} \exp(-D_{\mathsf{KL}}(\mathbb{P}_{\theta_\tau}, \mathbb{P}_{\theta_{\tau'}})).\end{aligned}$$

Here $\tau \sim \tau'$ refers to any τ, τ' that only differs in one element. And the last inequality is due to the Bretagnolle–Huber inequality Bretagnolle and Huber (1979). To bound the KL divergence, we have the following lemma from Shah et al. (2015):

Lemma J.1 (Shah et al. (2015)). *For any pair of quality score vectors θ_τ and $\theta_{\tau'}$, we have*

$$D_{\mathrm{KL}}(\mathbb{P}_{\theta_\tau} \| \mathbb{P}_{\theta_\tau}) \leq Cn(\theta_\tau - \theta_{\tau'})^\top \Sigma_{\mathcal{D}} (\theta_\tau - \theta_{\tau'}). \tag{3}$$

From the lemma, we have

$$\begin{aligned}\inf_{\hat{\pi}} \sup_{\mathcal{Q} \in \mathsf{CB}(\lambda)} \mathbb{E}[V_{\mathcal{Q}}(\pi^\star) - V^\star_{\mathcal{Q}}(\hat{\pi})] &\geq \frac{\Delta}{2} \min_{\tau \sim \tau'} \exp(-D_{\mathsf{KL}}(\mathbb{P}_{\theta_\tau}, \mathbb{P}_{\theta_{\tau'}})) \\ &\geq \frac{\Delta}{2} \exp(-Cn\Delta^2/(S\Lambda^2))\end{aligned}$$

Taking $\Delta = \Lambda\sqrt{S/n}$ and noting that $S = d/3$ finishes the proof. ☐

J.5 Proof of Theorem D.1

This section presents the proof of Theorem D.1 for the setting of K-wise comparisons. We first prove the following lemma on the estimation error:

Lemma J.2. *Under the K-wise PL model, for any $\lambda > 0$, with probability at least $1 - \delta$,*

$$\|\hat{\theta}_{\mathsf{MLE}} - \theta^\star\|_{\Sigma_{\mathcal{D}} + \lambda I} \leq C \cdot \sqrt{\frac{K^4(d + \log(1/\delta))}{\gamma^2 n} + \lambda B^2}.$$

Recall that the MLE is given by

$$\hat{\theta}_{\mathsf{MLE}} \in \arg\min_{\theta \in \Theta_B} \ell_{\mathcal{D}}(\theta),$$
$$\text{where } \ell_{\mathcal{D}}(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=0}^{K-1} \log \left(\frac{\exp(\langle \theta,\, \phi(s^i, a^i_{\sigma_i(j)}) \rangle)}{\sum_{k=j}^{K-1} \exp(\langle \theta,\, \phi(s^i, a^i_{\sigma_i(k)}) \rangle)} \right).$$

Our goal is to bound the estimation error of the MLE in the squared semi-norm $\|v\|^2_{\Sigma_{\mathcal{D}} + \lambda I} = v^T (\Sigma_{\mathcal{D}} + \lambda I) v$.

Strong convexity of ℓ. We first show that $\ell_{\mathcal{D}}$ is strongly convex at $\theta^\star$ with respect to the semi-norm $\|\cdot\|_{\Sigma_{\mathcal{D}}}$, meaning that there is some constant $\gamma > 0$ such that

$$\ell_{\mathcal{D}}(\theta^\star + \Delta) - \ell_{\mathcal{D}}(\theta^\star) - \langle \nabla \ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle \geq \gamma \|\Delta\|_{\Sigma_{\mathcal{D}}}^2 \tag{4}$$

for all perturbations $\Delta \in \mathbb{R}^d$ such that $\theta^\star + \Delta \in \Theta_B$.

The gradient of the negative log likelihood is

$$\nabla \ell_{\mathcal{D}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{K-1}\sum_{k=j}^{K-1}\frac{\exp(\langle \theta,\, \phi(s^i, a^i_{\sigma_i(k)})\rangle)}{\sum_{k'=j}^{K-1}\exp(\langle \theta,\, \phi(s^i, a^i_{\sigma_i(k')})\rangle)} \cdot (\phi(s^i, a^i_{\sigma_i(j)}) - \phi(s^i, a^i_{\sigma_i(k)})).$$

The Hessian of the negative log likelihood can be written as

$$\begin{aligned}
&\nabla^2 \ell_{\mathcal{D}}(\theta) \\
=&\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{K-1}\sum_{k=j}^{K-1}\sum_{k'=j}^{K-1}\frac{\exp(\langle \theta,\, \phi(s^i, a^i_{\sigma_i(k)}) + \phi(s^i, a^i_{\sigma_i(k')})\rangle)}{2(\sum_{k'=j}^{K-1}\exp(\langle \theta,\, \phi(s^i, a^i_{\sigma_i(k')})\rangle))^2} \cdot (\phi(s^i, a^i_{\sigma_i(k)}) - \phi(s^i, a^i_{\sigma_i(k')}))(\phi(s^i, a^i_{\sigma_i(k)}) - \phi(s^i, a^i_{\sigma_i(k')}))
\end{aligned}$$

Since $\exp(\langle \theta,\, \phi\rangle) \in [\exp(-LB), \exp(LB)]$, we know that the coefficients satisfy

$$\frac{\exp(\langle \theta,\, \phi(s^i, a^i_{\sigma_i(k)}) + \phi(s^i, a^i_{\sigma_i(k')})\rangle)}{(\sum_{k'=j}^{K-1}\exp(\langle \theta,\, \phi(s^i, a^i_{\sigma_i(k')})\rangle))^2} \geq \frac{\exp(-4LB)}{2(K-j)^2}.$$

Set $\gamma = \exp(-4LB)/2$. We can verify that for any vector $v \in \mathbb{R}^K$, one has

$$\begin{aligned}
v^\top \nabla^2 \ell_{\mathcal{D}}(\theta) v &\geq \frac{\gamma}{n} v^\top \left(\sum_{i=1}^{n}\sum_{j=0}^{K-1}\frac{1}{(K-j)^2}\sum_{k=j}^{K-1}\sum_{k'=k}^{K-1}(\phi(s^i, a^i_{\sigma_i(k)}) - \phi(s^i, a^i_{\sigma_i(k')}))(\phi(s^i, a^i_{\sigma_i(k)}) - \phi(s^i, a^i_{\sigma_i(k')}))^\top\right) v \\
&\geq \frac{\gamma}{n} v^\top \left(\sum_{i=1}^{n}\min_{\sigma_i \in \Pi[K]}\sum_{j=0}^{K-1}\frac{1}{(K-j)^2}\sum_{k=j}^{K-1}\sum_{k'=k}^{K-1}(\phi(s^i, a^i_{\sigma_i(k)}) - \phi(s^i, a^i_{\sigma_i(k')}))(\phi(s^i, a^i_{\sigma_i(k)}) - \phi(s^i, a^i_{\sigma_i(k')}))^\top\right) \\
&\geq \gamma v^\top \Sigma_{\mathcal{D}} v \\
&= \gamma \|v\|_{\Sigma_{\mathcal{D}}}^2.
\end{aligned}$$

Thus we know that ℓ is γ-strongly convex with respect to the semi-norm $\|\cdot\|_{\Sigma_{\mathcal{D}}}$.

Bounding the estimation error. Now we aim at bounding the estimation error $\|\hat{\theta}_{\mathsf{MLE}} - \theta^\star\|_{\Sigma_{\mathcal{D}+\lambda I}}$.

Since $\hat{\theta}_{\mathsf{MLE}}$ is optimal for $\ell_{\mathcal{D}}$, we have $\ell_{\mathcal{D}}(\hat{\theta}_{\mathsf{MLE}}) \leq \ell_{\mathcal{D}}(\theta^\star)$. Defining the error vector $\Delta = \hat{\theta}_{\mathsf{MLE}} - \theta^\star$, adding and subtracting the quantity $\langle \nabla \ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle$ yields the bound

$$\ell_{\mathcal{D}}(\theta^\star + \Delta) - \ell_{\mathcal{D}}(\theta^\star) - \langle \nabla \ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle \leq -\langle \nabla \ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle.$$

By the γ-convexity condition, the left-hand side is lower bounded by $\gamma\|\Delta\|_{\Sigma_{\mathcal{D}}}^2$. As for the right-hand side, note that $|\langle \nabla \ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle| \leq \|\nabla \ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}\, \|\Delta\|_{\Sigma_{\mathcal{D}}+\lambda I}$ for any $\lambda > 0$. Altogether we have

$$\gamma\|\Delta\|_{\Sigma_{\mathcal{D}}}^2 \leq \|\nabla \ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}\, \|\Delta\|_{\Sigma_{\mathcal{D}}+\lambda I}.$$

Now we further bound the term $\|\nabla \ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}$. Observe that the gradient takes the form

$$\nabla \ell_{\mathcal{D}}(\theta^\star) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{K-1}\sum_{k=j}^{K-1}\frac{\exp(\langle \theta^\star,\, \phi(s^i, a^i_{\sigma_i(k)})\rangle)}{\sum_{k'=j}^{K-1}\exp(\langle \theta^\star,\, \phi(s^i, a^i_{\sigma_i(k')})\rangle)} \cdot (\phi(s^i, a^i_{\sigma_i(j)}) - \phi(s^i, a^i_{\sigma_i(k)})). \tag{5}$$

We set $x^i_{jk} = \phi(s^i, a^i_j) - \phi(s^i, a^i_k)$. $X \in \mathbb{R}^{(nK(K-1)/2)\times d}$ has the differencing vector x^i_{jk} as its $(iK(K-1)/2 + k + \sum_{l=K-j+1}^{K} l)^{th}$ row. We also define V^i_{jk} be the random variable of the coefficient of x_{jk}^i in Equation (5) under the PL model, i.e. conditioned on an arbitrary permutation σ_i,

$$V_{jk}^i = \begin{cases} \frac{\exp(\langle\theta^\star,\, \phi(s^i, a_k^i)\rangle)}{\sum_{k'=\sigma_i^{-1}(j)}^{K-1} \exp(\langle\theta^\star,\, \phi(s^i, a_{\sigma_i(k')}^i)\rangle)}, & \text{if } \sigma_i^{-1}(j) < \sigma_i^{-1}(k) \\ -\frac{\exp(\langle\theta^\star,\, \phi(s^i, a_j^i)\rangle)}{\sum_{k'=\sigma_i^{-1}(k)}^{K-1} \exp(\langle\theta^\star,\, \phi(s^i, a_{\sigma_i(k')}^i)\rangle)}, & \text{otherwise.} \end{cases}$$

Here $\sigma_i^{-1}(j) < \sigma_i^{-1}(k)$ means that the j-th item ranks higher than the k-th item. Let $\tilde{V}_i \in \mathbb{R}^{K(K-1)/2}$ be the concatenated random vector of $\{V_{jk}^i\}_{0 \le j < k \le K-1}$, $V \in \mathbb{R}^{nK(K-1)/2}$ be the concatenated random vector of $\{\tilde{V}_i\}_{i=1}^n$. We know that $\tilde{V}_i$ and $\tilde{V}_j$ are independent for each $i \neq j$ due to the independent sampling procedure. We can also verify that the mean of $\tilde{V}_i$ is 0, the proof of which is deferred to the end of this section. Furthermore, since under any permutation, the sum of absolute value of each element in $\tilde{V}_i$ is at most K, we know that $\tilde{V}_i$ is sub-Gaussian with parameter K. Thus we know that V is also sub-Gaussian with mean 0 and parameter K. Now we know that the term $\|\nabla \ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}^2$ can be written as

$$\|\nabla \ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}^2 = \frac{1}{n^2} V^\top X (\Sigma_{\mathcal{D}} + \lambda I)^{-1} X^\top V.$$

Let $M = \frac{K^2}{n} I$. One can verify that $M \succeq \frac{1}{n^2} X(\Sigma_{\mathcal{D}} + \lambda I)^{-1} X^\top$ almost surely since $\lambda_{\mathsf{max}}(X(\Sigma_{\mathcal{D}} + \lambda I)^{-1} X^\top / n^2) \le K^2/n$. Thus we can upper bound the original term as

$$\|\nabla \ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}^2 \le \frac{K^2}{n} \|V\|_2^2.$$

By Bernstein's inequality for sub-Gaussian random variables in quadratic form (see e.g. Hsu et al. (2012, Theorem 2.1)), we know that with probability at least $1 - \delta$,

$$\|V\|_2^2 \le CK^2 \cdot (d + \log(1/\delta)).$$

Thus altogether, we have

$$\gamma \|\Delta\|_{\Sigma_{\mathcal{D}}}^2 \le \sqrt{\frac{CK^4 \cdot (d + \log(1/\delta))}{n}} \|\Delta\|_{\Sigma_{\mathcal{D}}+\lambda I}.$$

Similar to the pairwise comparison analysis in Appendix J.1, we can derive that with probability at least $1 - \delta$,

$$\|\hat{\theta}_{\mathsf{MLE}} - \theta^\star\|_{\Sigma_{\mathcal{D}}+\lambda I} \le C \cdot \sqrt{\frac{K^4(d + \log(1/\delta))}{n} + \lambda B^2}.$$

The rest of the proof on the sub-optimality upper bound follows the same argument as Theorem C.2.

Lastly, we verify that the mean of $\tilde{V}_i$ is 0. For any fixed $j, k \in [K]$, let $\mathcal{P}$ be the ordered set of all elements which are ranked higher than both j and k. Now conditioned on $\mathcal{P}$, we have

$$\begin{aligned} \mathbb{E}[V_{jk}^i \mid \mathcal{P}] &= \mathbb{P}(j \text{ follows } \mathcal{P} \mid \mathcal{P}) \cdot \frac{\exp(\langle\theta^\star,\, \phi(s^i, a_k^i)\rangle)}{\sum_{k' \in \bar{\mathcal{P}}} \exp(\langle\theta^\star,\, \phi(s^i, a_{k'}^i)\rangle)} - \mathbb{P}(k \text{ follows } \mathcal{P} \mid \mathcal{P}) \cdot \frac{\exp(\langle\theta^\star,\, \phi(s^i, a_j^i)\rangle)}{\sum_{k' \in \bar{\mathcal{P}}} \exp(\langle\theta^\star,\, \phi(s^i, a_{k'}^i)\rangle)} \\ &= \frac{1}{\sum_{k' \in \bar{\mathcal{P}}} \exp(\langle\theta^\star,\, \phi(s^i, a_{k'}^i)\rangle)} \cdot \Bigg(\frac{\exp(\langle\theta^\star,\, \phi(s^i, a_j^i)\rangle) \exp(\langle\theta^\star,\, \phi(s^i, a_k^i)\rangle) - \exp(\langle\theta^\star,\, \phi(s^i, a_j^i)\rangle) \exp(\langle\theta^\star,\, \phi(s}{\exp(\langle\theta^\star,\, \phi(s^i, a_j^i)\rangle) + \exp(\langle\theta^\star,\, \phi(s^i, a_k^i)\rangle)} \\ &= 0. \end{aligned}$$

Here the second equality uses the fact that j follows $\mathcal{P}$ is equivalent to the event that j is larger than k and either j, k is the largest among $\bar{\mathcal{P}}$. Taking expectation over $\mathcal{P}$ gives us that $\mathbb{E}[V_{jk}^i] = 0$.

J.6 Proof of Theorem D.2

This section presents the proof of Theorem D.2 for the setting of K-wise comparisons. We first prove the following lemma on the estimation error.

Lemma J.3. *Under the K-wise PL model, for any $\lambda > 0$, with probability at least $1-\delta$,*

$$\|\hat{\theta}_{\mathsf{MLE}_2} - \theta^\star\|_{\Sigma_{\mathcal{D}}+\lambda I} \le C \cdot \sqrt{\frac{d+\log(1/\delta)}{\gamma^2 n} + \lambda B^2}.$$

Recall that the pairwise compairson based estimator is given by

$$\hat{\theta}_{\mathsf{MLE}_2} \in \arg\min_{\theta\in\Theta_B} \ell_{\mathcal{D}}(\theta),$$

$$\text{where } \ell_{\mathcal{D}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{K-1}\sum_{k=j+1}^{K-1}\log\left(\frac{\exp(\langle\theta,\,\phi(s^i, a^i_{\sigma_i(j)})\rangle)}{\exp(\langle\theta,\,\phi(s^i, a^i_{\sigma_i(j)})\rangle) + \exp(\langle\theta,\,\phi(s^i, a^i_{\sigma_i(k)})\rangle)}\right).$$

Our goal is to bound the estimation error of the MLE in the squared semi-norm $\|v\|^2_{\Sigma_{\mathcal{D}}+\lambda I} = v^T(\Sigma_{\mathcal{D}}+\lambda I)v$.

Strong convexity of ℓ. Let $x^i_{jk} = \phi(s^i, a^i_j) - \phi(s^i, a^i_k)$. The gradient of the negative log likelihood is

$$\nabla\ell_{\mathcal{D}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{K-1}\sum_{k=j+1}^{K-1}\frac{\exp(-\langle\theta,\,x^i_{\sigma_i(j)\sigma_i(k)}\rangle)}{1+\exp(-\langle\theta,\,x^i_{\sigma_i(j)\sigma_i(k)}\rangle)}\cdot x^i_{\sigma_i(j)\sigma_i(k)}.$$

The Hessian of the negative log likelihood can be written as

$$\nabla^2\ell_{\mathcal{D}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{K-1}\sum_{k=j}^{K-1}\frac{\exp(-\langle\theta,\,x^i_{\sigma_i(j)\sigma_i(k)}\rangle)}{(1+\exp(-\langle\theta,\,x^i_{\sigma_i(j)\sigma_i(k)}\rangle))^2}\cdot x^i_{\sigma_i(j)\sigma_i(k)}x^{i\top}_{\sigma_i(j)\sigma_i(k)}.$$

Since $\exp(\langle\theta,\,x^i_{\sigma_i(j)\sigma_i(k)}\rangle) \in [\exp(-2LB), \exp(2LB)]$, we know that the coefficients satisfy

$$\frac{\exp(-\langle\theta,\,x^i_{\sigma_i(j)\sigma_i(k)}\rangle)}{(1+\exp(-\langle\theta,\,x^i_{\sigma_i(j)\sigma_i(k)}\rangle))^2} \ge \frac{1}{2+\exp(2LB)+\exp(-2LB)}.$$

Set $\gamma = \frac{1}{2+\exp(2LB)+\exp(-2LB)}$. We can verify that for any vector $v \in \mathbb{R}^K$, one has

$$\begin{aligned}
v^\top\nabla^2\ell_{\mathcal{D}}(\theta)v &\ge \frac{\gamma}{n}v^\top\left(\sum_{i=1}^{n}\sum_{j=0}^{K-1}\sum_{k=j+1}^{K-1}x^i_{\sigma_i(j)\sigma_i(k)}x^{i\top}_{\sigma_i(j)\sigma_i(k)}\right)v \\
&= \frac{\gamma}{n}v^\top\left(\sum_{i=1}^{n}\sum_{j=0}^{K-1}\sum_{k=j+1}^{K-1}x^i_{jk}x^{i\top}_{jk}\right)v \\
&= \gamma K(K-1)v^\top\Sigma_{\mathcal{D}}v/2 \\
&= \gamma K(K-1)\|v\|^2_{\Sigma_{\mathcal{D}}}/2.
\end{aligned}$$

Thus we know that ℓ is γ-strongly convex with respect to the semi-norm $\|\cdot\|_{\Sigma_{\mathcal{D}}}$.

Bounding the estimation error. Now we aim at bounding the estimation error $\|\hat{\theta}_{\mathsf{MLE}_2} - \theta^\star\|_{\Sigma_{\mathcal{D}+\lambda I}}$.

Since $\hat{\theta}_{\mathsf{MLE}_2}$ is optimal for $\ell_{\mathcal{D}}$, we have $\ell_{\mathcal{D}}(\hat{\theta}_{\mathsf{MLE}_2}) \le \ell_{\mathcal{D}}(\theta^\star)$. Defining the error vector $\Delta = \hat{\theta}_{\mathsf{MLE}_2} - \theta^\star$, adding and subtracting the quantity $\langle\nabla\ell_{\mathcal{D}}(\theta^\star),\,\Delta\rangle$ yields the bound

$$\ell_{\mathcal{D}}(\theta^\star+\Delta) - \ell_{\mathcal{D}}(\theta^\star) - \langle\nabla\ell_{\mathcal{D}}(\theta^\star),\,\Delta\rangle \le -\langle\nabla\ell_{\mathcal{D}}(\theta^\star),\,\Delta\rangle.$$

By the γ-convexity condition, the left-hand side is lower bounded by $\gamma K(K-1)\|\Delta\|^2_{\Sigma_{\mathcal{D}}}/2$. As for the right-hand side, note that $|\langle\nabla\ell_{\mathcal{D}}(\theta^\star),\,\Delta\rangle| \le \|\nabla\ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}\;\|\Delta\|_{\Sigma_{\mathcal{D}}+\lambda I}$ for any $\lambda > 0$. Altogether we have

$$\gamma\|\Delta\|^2_{\Sigma_{\mathcal{D}}} \le 2\|\nabla\ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}\;\|\Delta\|_{\Sigma_{\mathcal{D}}+\lambda I}/K(K-1).$$

Now we further bound the term $\|\nabla \ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}$. Observe that the gradient takes the form

$$\nabla \ell_{\mathcal{D}}(\theta^\star) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{K-1}\sum_{k=j+1}^{K-1} \frac{\exp(-\langle \theta,\, x^i_{\sigma_i(j)\sigma_i(k)}\rangle)}{1+\exp(-\langle \theta,\, x^i_{\sigma_i(j)\sigma_i(k)}\rangle)} \cdot x^i_{\sigma_i(j)\sigma_i(k)}. \tag{6}$$

We set $X \in \mathbb{R}^{(nK(K-1)/2)\times d}$ with the differencing vector x^i_{jk} as its $(iK(K-1)/2 + k + \sum_{l=K-j+1}^{K} l)^{th}$ row. We also define V^i_{jk} be the random variable of the coefficient of x^i_{jk} in Equation (6) under the PL model, i.e. conditioned on an arbitrary permutation σ_i,

$$V^i_{jk} = \begin{cases} \frac{\exp(-\langle \theta,\, x^i_{jk}\rangle)}{1+\exp(-\langle \theta,\, x^i_{jk}\rangle)}, & \text{if } \sigma_i^{-1}(j) < \sigma_i^{-1}(k) \\ -\frac{1}{1+\exp(-\langle \theta,\, x^i_{jk}\rangle)}, & \text{otherwise.} \end{cases}$$

Let $\tilde{V}_i \in \mathbb{R}^{K(K-1)/2}$ be the concatenated random vector of $\{V^i_{jk}\}_{0\le j<k\le K-1}$, $V \in \mathbb{R}^{nK(K-1)/2}$ be the concatenated random vector of $\{\tilde{V}_i\}_{i=1}^n$. We know that $\tilde{V}_i$ is independent for each i, and that V is sub-Gaussian with mean $\mathbf{0}$ and parameter $\sqrt{K(K-1)/2}$ since the PL model reduces to BTL model when considering pairwise comparisons. Now we know that the term $\|\nabla \ell_{\mathcal{D}}(\theta^\star)\|^2_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}$ can be written as

$$\|\nabla \ell_{\mathcal{D}}(\theta^\star)\|^2_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} = \frac{1}{n^2} V^\top X(\Sigma_{\mathcal{D}}+\lambda I)^{-1} X^\top V.$$

Let $M = \frac{K^2}{n} I$. One can verify that $M \succeq \frac{1}{n^2} X(\Sigma_{\mathcal{D}}+\lambda I)^{-1}X^\top$ almost surely since $\lambda_{\max}(X(\Sigma_{\mathcal{D}}+\lambda I)^{-1}X^\top/n^2) \le K^2/n$. Thus we can upper bound the original term as

$$\|\nabla \ell_{\mathcal{D}}(\theta^\star)\|^2_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} \le \frac{K^2}{n}\|V\|_2^2.$$

By Bernstein's inequality for sub-Gaussian random variables in quadratic form (see e.g. Hsu et al. (2012, Theorem 2.1)), we know that with probability at least $1-\delta$,

$$\|V\|_2^2 \le CK(K-1)\cdot(d+\log(1/\delta)).$$

Thus altogether, we have

$$\gamma\|\Delta\|^2_{\Sigma_{\mathcal{D}}} \le \sqrt{\frac{C\cdot(d+\log(1/\delta))}{n}}\|\Delta\|_{\Sigma_{\mathcal{D}}+\lambda I}.$$

Similar to the pairwise comparison, we can derive that with probability at least $1-\delta$,

$$\|\hat{\theta}_{\mathsf{MLE}_2} - \theta^\star\|_{\Sigma_{\mathcal{D}}+\lambda I} \le C\cdot\sqrt{\frac{d+\log(1/\delta)}{n} + \lambda B^2}.$$

The rest of the proof on the sub-optimality upper bound follows the same argument as Theorem C.2.

J.7 Proof of Lemma E.1

Recall that the MLE is given by

$$\hat{\theta}_{\mathsf{MLE}} \in \arg\min_{\theta \in \Theta_B} \ell_{\mathcal{D}}(\theta),$$

$$\begin{aligned}
\text{where } \ell_{\mathcal{D}}(\theta) &= -\sum_{i=1}^{n} \log\Big(1(y^i = 1) \cdot \frac{\exp(\sum_{h=1}^{H} r_\theta(s_h^i, a_h^i))}{\exp(\sum_{h=1}^{H} r_\theta(s_h^i, a_h^i)) + \exp(\sum_{h=1}^{H} r_\theta(s_h^{i\prime}, a_h^{i\prime}))} \\
&\quad + 1(y^i = 0) \cdot \frac{\exp(\sum_{h=1}^{H} r_\theta(s_h^{i\prime}, a_h^{i\prime}))}{\exp(\sum_{h=1}^{H} r_\theta(s_h^i, a_h^i)) + \exp(\sum_{h=1}^{H} r_\theta(s_h^{i\prime}, a_h^{i\prime}))}\Big) \\
&= -\sum_{i=1}^{n} \log\Big(1(y^i = 1) \cdot \frac{1}{\exp(-\sum_{h=1}^{H}(r_\theta(s_h^i, a_h^i) - r_\theta(s_h^{i\prime}, a_h^{i\prime}))) + 1} \\
&\quad + 1(y^i = 0) \cdot \frac{1}{\exp(\sum_{h=1}^{H}(r_\theta(s_h^i, a_h^i) - r_\theta(s_h^{i\prime}, a_h^{i\prime}))) + 1}\Big) \\
&= -\sum_{i=1}^{n} \log\Big(1(y^i = 1) \cdot \frac{1}{\exp(-\langle\theta, \sum_{h=1}^{H}(\phi(s_h^i, a_h^i) - \phi(s_h^{i\prime}, a_h^{i\prime}))\rangle) + 1} \\
&\quad + 1(y^i = 0) \cdot \frac{1}{\exp(\langle\theta, \sum_{h=1}^{H}(\phi(s_h^i, a_h^i) - \phi(s_h^{i\prime}, a_h^{i\prime}))\rangle) + 1}\Big)
\end{aligned}$$

To simplify the notation, we let $x_i = \sum_{h=1}^{H}(\phi(s_h^i, a_h^i) - \phi(s_h^{i\prime}, a_h^{i\prime}))$. Our goal is to bound the estimation error of the MLE in the squared semi-norm $\|v\|_{\Sigma_{\mathcal{D}}+\lambda I}^2 = v^T(\Sigma_{\mathcal{D}} + \lambda I)v$.

Strong convexity of ℓ. We first show that $\ell_{\mathcal{D}}$ is strongly convex at $\theta^\star$ with respect to the semi-norm $\|\cdot\|_{\Sigma_{\mathcal{D}}}$, meaning that there is some constant $\gamma > 0$ such that

$$\ell_{\mathcal{D}}(\theta^\star + \Delta) - \ell_{\mathcal{D}}(\theta^\star) - \langle\nabla\ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle \geq \gamma\|\Delta\|_{\Sigma_{\mathcal{D}}}^2 \tag{7}$$

for all perturbations $\Delta \in \mathbb{R}^d$ such that $\theta^\star + \Delta \in \Theta_B$.

One can directly calculate the Hessian of ℓ as

$$\nabla^2\ell_{\mathcal{D}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(1(y^i = 1) \cdot \frac{\exp(-\langle\theta,\, x_i\rangle)}{(\exp(-\langle\theta,\, x_i\rangle) + 1)^2} + 1(y^i = 0) \cdot \frac{\exp(\langle\theta,\, x_i\rangle)}{(\exp(\langle\theta,\, x_i\rangle) + 1)^2}\right) \cdot x_i x_i^T,$$

Observe that $\langle\theta,\, x_i\rangle \in [-2HLB, 2HLB]$, we have

$$v^T\nabla^2\ell_{\mathcal{D}}(\theta)v \geq \frac{\gamma}{n}\|Xv\|_2^2 \qquad \text{for all } v,$$

where $\gamma = 1/(2 + \exp(-2HLB) + \exp(2HLB))$, $X \in \mathbb{R}^{n\times d}$ has the differencing vector $x_i \in \mathbb{R}^d$ as its i^{th} row.

Thus, if we introduce the error vector $\Delta := \hat{\theta}_{\mathsf{MLE}} - \theta^\star$, then we may conclude that

$$\ell_{\mathcal{D}}(\theta^\star + \Delta) - \ell_{\mathcal{D}}(\theta^\star) - \langle\nabla\ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle \geq \frac{\gamma}{n}\|X\Delta\|_2^2 \;=\; \gamma\|\Delta\|_{\Sigma_{\mathcal{D}}}^2,$$

showing that $\ell_{\mathcal{D}}$ is strongly convex around $\theta^\star$ with parameter γ.

Bounding the estimation error. Now we aim at bounding the estimation error $\|\hat{\theta}_{\mathsf{MLE}} - \theta^\star\|_{\Sigma_{\mathcal{D}}}$.

Since $\hat{\theta}_{\mathsf{MLE}}$ is optimal for $\ell_{\mathcal{D}}$, we have $\ell_{\mathcal{D}}(\hat{\theta}_{\mathsf{MLE}}) \leq \ell_{\mathcal{D}}(\theta^\star)$. Defining the error vector $\Delta = \hat{\theta}_{\mathsf{MLE}} - \theta^\star$, adding and subtracting the quantity $\langle\nabla\ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle$ yields the bound

$$\ell_{\mathcal{D}}(\theta^\star + \Delta) - \ell_{\mathcal{D}}(\theta^\star) - \langle\nabla\ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle \leq -\langle\nabla\ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle.$$

By the γ-convexity condition, the left-hand side is lower bounded by $\gamma\|\Delta\|_{\Sigma_{\mathcal{D}}}^2$. As for the right-hand side, note that $|\langle\nabla\ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle| \leq \|\nabla\ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}\, \|\Delta\|_{\Sigma_{\mathcal{D}}+\lambda I}$ for any $\lambda > 0$. Altogether we have

$$\gamma\|\Delta\|_{\Sigma_{\mathcal{D}}}^2 \leq \|\nabla\ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}\, \|\Delta\|_{\Sigma_{\mathcal{D}}+\lambda I}.$$

Now we further bound the term $\|\nabla \ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}$. Observe that the gradient takes the form

$$\nabla \ell_{\mathcal{D}}(\theta^\star) = \frac{-1}{n}\sum_{i=1}^{n}\left[\mathbf{1}[y^i=1]\frac{\exp(-\langle\theta^\star,\, x_i\rangle)}{1+\exp(-\langle\theta^\star,\, x_i\rangle))} - \mathbf{1}[y^i=0]\frac{1}{1+\exp(-\langle\theta^\star,\, x_i\rangle))}\right] x_i.$$

Define a random vector $V \in \mathbb{R}^n$ with independent components as

$$V_i = \begin{cases} \frac{\exp(-\langle\theta^\star,\, x_i\rangle)}{1+\exp(-\langle\theta^\star,\, x_i\rangle))} & \text{w.p.} \quad \frac{1}{1+\exp(-\langle\theta^\star,\, x_i\rangle))} \\ \frac{-1}{1+\exp(-\langle\theta^\star,\, x_i\rangle))} & \text{w.p.} \quad \frac{\exp(-\langle\theta^\star,\, x_i\rangle)}{1+\exp(-\langle\theta^\star,\, x_i\rangle))}. \end{cases}$$

With this notation, we have $\nabla \ell_{\mathcal{D}}(\theta^\star) = -\frac{1}{n}X^T V$. One can verify that $\mathbb{E}[V]=0$ and $|V_i|\le 1$.

Defining the n-dimensional square matrix $M := \frac{1}{n^2}X(\Sigma_{\mathcal{D}}+\lambda I)^{-1}X^T$, we have $\|\nabla \ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} = V^T M V$. Let the eigenvalue decomposition of $XX^\top$ be $XX^\top = U\Lambda U^\top$. We can bound the trace and operator norm of M as

$$\begin{aligned}\mathsf{Tr}(M) &= \frac{1}{n^2}\mathsf{Tr}(U(\Lambda/n+\lambda I)^{-1}U^\top U\Lambda U^\top) \le \frac{d}{n} \\ \|M\|_{\mathsf{op}} &= \lambda_{\mathsf{max}}(M) \le \frac{1}{n},\end{aligned}$$

Moreover, since the components of V are independent and of zero mean, and $|V_i|\le 1$, the variables are 1-sub-Gaussian, and hence 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 $1-\delta$,

$$\|\nabla \ell_{\mathcal{D}}(\theta^\star)\|^2_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} = V^\top M V \le C_1 \cdot \frac{d+\log(1/\delta)}{n}.$$

Here C_1 is some universal constant. This gives us

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

Solving the above inequality gives us that for some constant C_2,

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

J.8 PROOF OF THEOREM E.2

Proof. From Lemma E.1, we know that with probability at least $1-\delta$,

$$\|\hat{\theta}_{\mathsf{MLE}} - \theta^\star\|_{\Sigma_{\mathcal{D}}+\lambda I} \le C \cdot \sqrt{\frac{d+\log(1/\delta)}{\gamma^2 n} + \lambda B^2}.$$

Let $J'(\pi) = J(\pi) - H\langle\theta^\star,\, v\rangle$. We have

$$\begin{aligned}\mathsf{SubOpt}(\hat{\pi}_{\mathsf{PE}}) &= J(\pi^\star) - J(\hat{\pi}_{\mathsf{PE}}) \\ &= J'(\pi^\star) - J'(\hat{\pi}_{\mathsf{PE}}) \\ &= (J'(\pi^\star) - \hat{J}(\pi^\star)) + (\hat{J}(\pi^\star) - \hat{J}(\hat{\pi}_{\mathsf{PE}})) + (\hat{J}(\hat{\pi}_{\mathsf{PE}}) - J'(\hat{\pi}_{\mathsf{PE}})).\end{aligned}$$

Since $\hat{\pi}_{\mathsf{PE}}$ is the optimal policy under expected value $\hat{J}(\pi)$, we know that the second difference satisfies $\hat{J}(\pi^\star) - \hat{J}(\hat{\pi}_{\mathsf{PE}}) \le 0$. For the third difference, we have

$$\hat{J}(\hat{\pi}_{\mathsf{PE}}) - J'(\hat{\pi}_{\mathsf{PE}}) = \mathbb{E}_{s\sim d^{\hat{\pi}_{\mathsf{PE}}}}[\hat{r}(s, \hat{\pi}_{\mathsf{PE}}(s)) - r(s, \hat{\pi}_{\mathsf{PE}}(s))].$$

From Lemma E.1 we know that $\theta^\star \in \Theta(\hat{\theta}_{\mathsf{MLE}}, \lambda)$ with probability at least $1-\delta$. Thus we know that with probability at least $1-\delta$, $\hat{J}(\hat{\pi}_{\mathsf{PE}}) - J'(\hat{\pi}_{\mathsf{PE}}) \leq 0$. Now combining everything together, we have

$$
\begin{aligned}
\mathsf{SubOpt}(\hat{\pi}_{\mathsf{PE}}) &\leq J'(\pi^\star) - \hat{J}(\pi^\star) \\
&= \sup_{\theta \in \Theta(\hat{\theta}_{\mathsf{MLE}}, \lambda)} \mathbb{E}_{s\sim d^{\pi^\star}}[(\theta^\star - \theta)^\top(\phi(s, \pi^\star(s)) - v)] \\
&= \sup_{\theta \in \Theta(\hat{\theta}_{\mathsf{MLE}}, \lambda)} \mathbb{E}_{s\sim d^{\pi^\star}}[(\theta^\star - \hat{\theta}_{\mathsf{MLE}} + \hat{\theta}_{\mathsf{MLE}} - \theta)^\top(\phi(s, \pi^\star(s)) - v)] \\
&= \mathbb{E}_{s\sim d^{\pi^\star}}[(\theta^\star - \hat{\theta}_{\mathsf{MLE}})^\top(\phi(s, \pi^\star(s)) - v)] + \sup_{\theta \in \Theta(\hat{\theta}_{\mathsf{MLE}}, \lambda)} \mathbb{E}_{s\sim d^{\pi^\star}}[(\hat{\theta}_{\mathsf{MLE}} - \theta)^\top(\phi(s, \pi^\star(s)) - v)].
\end{aligned}
$$

By the definition of $\Theta(\hat{\theta}_{\mathsf{MLE}}, \lambda)$, we know that for any $\theta \in \Theta(\hat{\theta}_{\mathsf{MLE}}, \lambda)$, one has $\mathbb{E}_{s\sim d^{\pi^\star}}[(\hat{\theta}_{\mathsf{MLE}} - \theta)^\top(\phi(s, \pi^\star(s)) - v)] \leq C \cdot \sqrt{\frac{d+\log(1/\delta)}{\gamma^2 n} + \lambda B^2} \cdot \|(\Sigma_{\mathcal{D}} + \lambda I)^{-1/2}\mathbb{E}_{s\sim d^{\pi^\star}}[\phi(s, \pi^\star(s)) - v]\|_2$. Furthermore, we know that $\hat{\theta}^\star \in \Theta(\hat{\theta}_{\mathsf{MLE}}, \lambda)$ from Lemma E.1. Altogether we have with probability $1-2\delta$

$$
\mathsf{SubOpt}(\hat{\pi}_{\mathsf{PE}}) \leq 2C \cdot \sqrt{\frac{d+\log(1/\delta)}{\gamma^2 n} + \lambda B^2} \cdot \|(\Sigma_{\mathcal{D}} + \lambda I)^{-1/2}\mathbb{E}_{s\sim d^{\pi^\star}}[\phi(s, \pi^\star(s)) - v]\|_2.
$$

□

J.9 Proof of Theorem F.2

Proof. Here We mainly prove Lemma F.1, since Theorem F.2 is a direct corollary when combined with the proof in Theorem E.2.

Our goal is to bound the estimation error of the MLE in the squared semi-norm $\|v\|^2_{\Sigma_{\mathcal{D}}+\lambda I} = v^T(\Sigma_{\mathcal{D}} + \lambda I)v$.

Strong convexity of ℓ. We first show that $\ell_{\mathcal{D}}$ is strongly convex at $\theta^\star$ with respect to the semi-norm $\|\cdot\|_{\Sigma_{\mathcal{D}}}$, meaning that there is some constant $\gamma > 0$ such that

$$
\ell_{\mathcal{D}}(\theta^\star + \Delta) - \ell_{\mathcal{D}}(\theta^\star) - \langle \nabla \ell_{\mathcal{D}}(\theta^\star),\, \Delta \rangle \geq \gamma \|\Delta\|^2_{\Sigma_{\mathcal{D}}} \tag{8}
$$

for all perturbations $\Delta \in \mathbb{R}^d$ such that $\theta^\star + \Delta \in \Theta_B$.

The gradient of the negative log likelihood is

$$
\nabla \ell_{\mathcal{D}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{\tau' \in \mathcal{T}(s_0^i)} \frac{\exp(\sum_{h=0}^{H}\langle \theta,\, \phi(s_h', a_h')\rangle)}{\sum_{\tau'' \in \mathcal{T}(s_0^i)} \exp(\sum_{h=0}^{H}\langle \theta,\, \phi(s_h'', a_h'')\rangle)} \cdot \left(\sum_{h=0}^{H}(\phi(s_h^i, a_h^i) - \phi(s_h', a_h'))\right).
$$

Let $x^i_{\tau,\tau'} = \sum_{h=0}^{H}(\phi(s_h, a_h) - \phi(s_h', a_h'))$, where $\tau = \{(s_h, a_h)\}_{h\in[H]}$, $\tau' = \{(s_h', a_h')\}_{h\in[H]}$. The Hessian of the negative log likelihood can be written as

$$
\begin{aligned}
&\nabla^2 \ell_{\mathcal{D}}(\theta) \\
=&\frac{1}{n}\sum_{i=1}^{n}\sum_{\tau \in \mathcal{T}(s_0^i)}\sum_{\tau' \in \mathcal{T}(s_0^i)} \frac{\exp(\sum_{h=0}^{H}\langle \theta,\, \phi(s_h, a_h) + \phi(s_h', a_h')\rangle)}{2(\sum_{\tau'' \in \mathcal{T}(s_0^i)} \exp(\sum_{h=0}^{H}\langle \theta,\, \phi(s_h'', a_h'')\rangle))^2} \cdot x^i_{\tau,\tau'} x^{i\top}_{\tau,\tau'}.
\end{aligned}
$$

Since $\exp(\langle \theta,\, \phi\rangle) \in [\exp(-LB), \exp(LB)]$, we know that the coefficients satisfy

$$
\frac{\exp(\sum_{h=0}^{H}\langle \theta,\, \phi(s_h, a_h) + \phi(s_h', a_h')\rangle)}{2(\sum_{\tau'' \in \mathcal{T}(s_0^i)} \exp(\sum_{h=0}^{H}\langle \theta,\, \phi(s_h'', a_h'')\rangle))^2} \geq \frac{\exp(-4LB)}{2\sup_s |\mathcal{T}(s)|^2}.
$$

Set $\gamma = \exp(-4LB)/2$. We can verify that for any vector $v \in \mathbb{R}^K$, one has

$$
v^\top \nabla^2 \ell_{\mathcal{D}}(\theta) v \geq \gamma v^\top \Sigma_{\mathcal{D}} v = \gamma \|v\|^2_{\Sigma_{\mathcal{D}}}.
$$

Thus we know that ℓ is γ-strongly convex with respect to the semi-norm $\|\cdot\|_{\Sigma_{\mathcal{D}}}$.

Bounding the estimation error. Now we aim at bounding the estimation error $\|\hat{\theta}_{\mathsf{MLE}} - \theta^\star\|_{\Sigma_{\mathcal{D}}+\lambda I}$.

Since $\hat{\theta}_{\mathsf{MLE}}$ is optimal for $\ell_{\mathcal{D}}$, we have $\ell_{\mathcal{D}}(\hat{\theta}_{\mathsf{MLE}}) \le \ell_{\mathcal{D}}(\theta^\star)$. Defining the error vector $\Delta = \hat{\theta}_{\mathsf{MLE}} - \theta^\star$, adding and subtracting the quantity $\langle \nabla \ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle$ yields the bound

$$\ell_{\mathcal{D}}(\theta^\star + \Delta) - \ell_{\mathcal{D}}(\theta^\star) - \langle \nabla \ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle \le -\langle \nabla \ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle.$$

By the γ-convexity condition, the left-hand side is lower bounded by $\gamma\|\Delta\|^2_{\Sigma_{\mathcal{D}}}$. As for the right-hand side, note that $|\langle \nabla \ell_{\mathcal{D}}(\theta^\star),\, \Delta\rangle| \le \|\nabla \ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}\, \|\Delta\|_{\Sigma_{\mathcal{D}}+\lambda I}$ for any $\lambda > 0$. Altogether we have

$$\gamma\|\Delta\|^2_{\Sigma_{\mathcal{D}}} \le \|\nabla \ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}\, \|\Delta\|_{\Sigma_{\mathcal{D}}+\lambda I}.$$

Now we further bound the term $\|\nabla \ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}$. Observe that the gradient takes the form

$$\nabla \ell_{\mathcal{D}}(\theta^\star) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{\tau'\in\mathcal{T}(s_0^i)} \frac{\exp(\sum_{h=0}^{H}\langle \theta^\star,\, \phi(s'_h, a'_h)\rangle)}{\sum_{\tau''\in\mathcal{T}(s_0^i)}\exp(\sum_{h=0}^{H}\langle\theta^\star,\, \phi(s''_h, a''_h)\rangle)} \cdot \left(\sum_{h=0}^{H}(\phi(s_h^i, a_h^i) - \phi(s'_h, a'_h))\right). \tag{9}$$

We set X as the concatenated differencing vector $x^i_{\tau,\tau'}$ where τ, τ' are distinct and ordered. We also define $V^i_{\tau,\tau'}$ be the random variable of the coefficient of $x^i_{\tau,\tau'}$ in Equation (9), i.e.

$$V^i_{\tau,\tau'} = \begin{cases} \frac{\exp(\sum_{h=0}^{H}\langle\theta^\star,\, \phi(s'_h, a'_h)\rangle)}{\sum_{\tau''\in\mathcal{T}(s_0^i)}\exp(\sum_{h=0}^{H}\langle\theta^\star,\, \phi(s''_h, a''_h)\rangle)}, & \text{if } \tau = \{(s_h^i, a_h^i)\}_{h\in[H]}, \\ -\frac{\exp(\sum_{h=0}^{H}\langle\theta^\star,\, \phi(s_h, a_h)\rangle)}{\sum_{\tau''\in\mathcal{T}(s_0^i)}\exp(\sum_{h=0}^{H}\langle\theta^\star,\, \phi(s''_h, a''_h)\rangle)}, & \text{if } \tau' = \{(s_h^i, a_h^i)\}_{h\in[H]}, \\ 0 & \text{otherwise.}\end{cases}$$

Let $\tilde{V}_i$ be the concatenated random vector of $\{V^i_{\tau,\tau'}\}$, V be the concatenated random vector of $\{\tilde{V}_i\}_{i=1}^n$. We know that $\tilde{V}_i$ and $\tilde{V}_j$ are independent for each $i \ne j$ due to the independent sampling procedure. We can also verify that the mean of $\tilde{V}_i$ is 0. We know that $\tilde{V}_i$ has almost $\sup_s|\mathcal{T}(s)|$ non-zero elements. And the sum of their absolute value is bounded by 1. we know $\tilde{V}_i$ is 1-sub-Gaussian. Now we know that the term $\|\nabla \ell_{\mathcal{D}}(\theta^\star)\|^2_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}$ can be written as

$$\|\nabla \ell_{\mathcal{D}}(\theta^\star)\|^2_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} = \frac{1}{n^2}V^\top X(\Sigma_{\mathcal{D}}+\lambda I)^{-1}X^\top V.$$

Let $M = \frac{\sup_s|\mathcal{T}(s)|^2}{n}I$. One can verify that $M \succeq \frac{1}{n^2}X(\Sigma_{\mathcal{D}}+\lambda I)^{-1}X^\top$ almost surely since $\lambda_{\mathsf{max}}(X(\Sigma_{\mathcal{D}}+\lambda I)^{-1}X^\top/n^2) \le \sup_s|\mathcal{T}(s)|^2/n$. Thus we can upper bound the original term as

$$\|\nabla \ell_{\mathcal{D}}(\theta^\star)\|^2_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} \le \frac{\sup_s|\mathcal{T}(s)|^2}{n}\|V\|_2^2.$$

By Bernstein's inequality for sub-Gaussian random variables in quadratic form (see e.g. Hsu et al. (2012, Theorem 2.1)), we know that with probability at least $1-\delta$,

$$\|V\|_2^2 \le C\cdot(d + \log(1/\delta)).$$

Thus altogether, we have

$$\gamma\|\Delta\|^2_{\Sigma_{\mathcal{D}}} \le \sqrt{\frac{C\sup_s|\mathcal{T}(s)|^2\cdot(d+\log(1/\delta))}{n}}\|\Delta\|_{\Sigma_{\mathcal{D}}+\lambda I}.$$

Similar to the pairwise comparison analysis in Appendix J.1, we can derive that with probability at least $1-\delta$,

$$\|\hat{\theta}_{\mathsf{MLE}} - \theta^\star\|_{\Sigma_{\mathcal{D}}+\lambda I} \le C\cdot\sqrt{\frac{\sup_s|\mathcal{T}(s)|^2(d+\log(1/\delta))}{n} + \lambda B^2}.$$

The rest of the proof on the sub-optimality upper bound follows the same argument as Theorem E.2.

□

J.10 PROOF OF THEOREM I.2

To simplify the notation, we let $f_\theta^i = r_\theta(s^i, a_1^i) - r_\theta(s^i, a_0^i)$. We can see that the gradient of ℓ takes the form

$$\nabla \ell_{\mathcal{D}}(\theta) = \frac{-1}{n} \sum_{i=1}^{n} \left[\mathbf{1}[y^i = 1] \frac{\exp(-f_\theta^i)}{1 + \exp(-f_\theta^i)} - \mathbf{1}[y^i = 0] \frac{1}{1 + \exp(-f_\theta^i)} \right] \nabla f_\theta^i.$$

And the Hessian of ℓ is

$$\nabla^2 \ell_{\mathcal{D}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\exp(f_\theta^i)}{(\exp(f_\theta^i)+1)^2} \cdot \nabla f_\theta^i \nabla f_\theta^{i\top} - \frac{1(y^i = 1) \cdot \exp(-f_\theta^i)}{1 + \exp(-f_\theta^i)} \cdot \nabla^2 f_\theta^i + \frac{1(y^i = 0) \cdot \exp(f_\theta^i)}{1 + \exp(f_\theta^i)} \cdot \nabla^2 f_\theta^i \right).$$

Now from Assumption I.1, we have

$$\nabla^2 \ell_{\mathcal{D}}(\theta) \succeq \frac{1}{n} \sum_{i=1}^{n} \gamma \nabla f_\theta^i \nabla f_\theta^{i\top} - 2\alpha_2 I.$$

where $\gamma = \frac{1}{2+\exp(-2LB)+\exp(2LB)}$. Now from the Lipschitz gradient assumption we also know that $\|\nabla f_\theta^i - \nabla f_{\theta^\star}^i\| \leq 2\alpha_2 \|\theta^\star - \theta\|$. Let $u = \nabla f_\theta^i - \nabla f_{\theta^\star}^i$, we have

$$\begin{aligned}
\nabla^2 \ell_{\mathcal{D}}(\theta) &\succeq \frac{1}{n} \sum_{i=1}^{n} \gamma (\nabla f_{\theta^\star}^i + u)(\nabla f_{\theta^\star}^i + u)^\top - 2\alpha_2 I \\
&\succeq \frac{1}{n} \sum_{i=1}^{n} \gamma \nabla f_{\theta^\star}^i \nabla f_{\theta^\star}^{i\top} + \gamma(\nabla f_{\theta^\star}^i u^\top + u \nabla f_{\theta^\star}^{i\top}) - 2\alpha_2 I.
\end{aligned}$$

Since $u^\top v \leq \|u\|_2 \|v\|_2 \leq 2\alpha_2 B \|v\|_2$, $v^\top \nabla f_{\theta^\star}^i \leq \alpha_1 \|v\|_2$, this gives that

$$v^T \nabla^2 \ell_{\mathcal{D}}(\theta) v \geq \frac{\gamma}{n} \|Xv\|_2^2 - 2\alpha_2(1 + 2\gamma\alpha_1 B)\|v\|_2^2 \qquad \text{for all } v,$$

where $X \in \mathbb{R}^{n \times d}$ has the vector $\nabla f_{\theta^\star}^i \in \mathbb{R}^d$ as its i^{th} row. Thus, if we introduce the error vector $\Delta := \hat{\theta}_{\mathsf{MLE}} - \theta^\star$, then we may conclude that

$$\ell_{\mathcal{D}}(\theta^\star + \Delta) - \ell_{\mathcal{D}}(\theta^\star) - \langle \nabla \ell_{\mathcal{D}}(\theta^\star), \Delta \rangle \geq \frac{\gamma}{n} \|X\Delta\|_2^2 - 2\alpha_2(1 + 2\gamma\alpha_1 B)\|\Delta\|_2^2 = \gamma \|\Delta\|_{\Sigma_{\mathcal{D}}}^2 - 2\alpha_2(1 + 2\gamma\alpha_1 B)\|\Delta\|_2^2.$$

Bounding the estimation error. Now we aim at bounding the estimation error $\|\hat{\theta}_{\mathsf{MLE}} - \theta^\star\|_{\Sigma_{\mathcal{D}}}$. Since $\hat{\theta}_{\mathsf{MLE}}$ is optimal for $\ell_{\mathcal{D}}$, we have $\ell_{\mathcal{D}}(\hat{\theta}_{\mathsf{MLE}}) \leq \ell_{\mathcal{D}}(\theta^\star)$. (When $\hat{\theta}_{\mathsf{MLE}}$ is approximately optimal, i.e. $\ell_{\mathcal{D}}(\hat{\theta}_{\mathsf{MLE}}) \leq \min_\theta \ell_{\mathcal{D}}(\theta) + \epsilon$, the same argument also holds up to an extra additive term ϵ.) Defining the error vector $\Delta = \hat{\theta}_{\mathsf{MLE}} - \theta^\star$, adding and subtracting the quantity $\langle \nabla \ell_{\mathcal{D}}(\theta^\star), \Delta \rangle$ yields the bound

$$\ell_{\mathcal{D}}(\theta^\star + \Delta) - \ell_{\mathcal{D}}(\theta^\star) - \langle \nabla \ell_{\mathcal{D}}(\theta^\star), \Delta \rangle \leq -\langle \nabla \ell_{\mathcal{D}}(\theta^\star), \Delta \rangle.$$

We know the left-hand side is lower bounded by $\gamma \|\Delta\|_{\Sigma_{\mathcal{D}}}^2 - 2\alpha_2(1 + 2\gamma\alpha_1 B)\|\Delta\|_2^2$. As for the right-hand side, note that $|\langle \nabla \ell_{\mathcal{D}}(\theta^\star), \Delta \rangle| \leq \|\nabla \ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}} + \lambda I)^{-1}} \ \|\Delta\|_{\Sigma_{\mathcal{D}} + \lambda I}$ for any $\lambda > 0$. Altogether we have

$$\gamma \|\Delta\|_{\Sigma_{\mathcal{D}}}^2 \leq \|\nabla \ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}} + \lambda I)^{-1}} \ \|\Delta\|_{\Sigma_{\mathcal{D}} + \lambda I} + \beta \|\Delta\|_2^2,$$

where $\beta = 2\alpha_2(1 + 2\gamma\alpha_1 B)$. Now we further bound the term $\|\nabla \ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}} + \lambda I)^{-1}}$. Observe that the gradient takes the form

$$\nabla \ell_{\mathcal{D}}(\theta^\star) = \frac{-1}{n} \sum_{i=1}^{n} \left[\mathbf{1}[y^i = 1] \frac{\exp(-f_{\theta^\star}^i)}{1 + \exp(-f_{\theta^\star}^i)} - \mathbf{1}[y^i = 0] \frac{1}{1 + \exp(-f_{\theta^\star}^i)} \right] \nabla f_{\theta^\star}^i.$$

Define a random vector $V \in \mathbb{R}^n$ with independent components as

$$V_i = \begin{cases} \frac{\exp(-f_{\theta^\star}^i)}{1+\exp(-f_{\theta^\star}^i)} & \text{w.p.} \quad \frac{1}{1+\exp(-f_{\theta^\star}^i)} \\ \frac{-1}{1+\exp(-f_{\theta^\star}^i)} & \text{w.p.} \quad \frac{\exp(-f_{\theta^\star}^i)}{1+\exp(-f_{\theta^\star}^i)}. \end{cases}$$

With this notation, we have $\nabla \ell_{\mathcal{D}}(\theta^\star) = -\frac{1}{n} X^T V$. One can verify that $\mathbb{E}[V] = 0$ and $|V_i| \leq 1$.

Defining the n-dimensional square matrix $M := \frac{1}{n^2} X(\Sigma_{\mathcal{D}} + \lambda I)^{-1} X^T$, we have $\|\nabla \ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} = V^T M V$. Let the eigenvalue decomposition of $X^\top X$ be $X^\top X = U \Lambda U^\top$. We can bound the trace and operator norm of M as

$$
\begin{aligned}
\mathsf{Tr}(M) &= \frac{1}{n^2} \mathsf{Tr}(U(\Lambda/n + \lambda I)^{-1} U^\top U \Lambda U^\top) \leq \frac{d}{n} \\
\mathsf{Tr}(M^2) &= \frac{1}{n^4} \mathsf{Tr}(U(\Lambda/n + \lambda I)^{-1} U^\top U \Lambda U^\top U(\Lambda/n + \lambda I)^{-1} U^\top U \Lambda U^\top) \leq \frac{d}{n^2} \\
\|M\|_{\mathsf{op}} &= \lambda_{\mathsf{max}}(M) \leq \frac{1}{n},
\end{aligned}
$$

Moreover, since the components of V are independent and of zero mean, and $|V_i| \leq 1$, the variables are 1-sub-Gaussian, and hence 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 $1 - \delta$,

$$
\|\nabla \ell_{\mathcal{D}}(\theta^\star)\|^2_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} = V^\top M V \leq C_1 \cdot \frac{d + \log(1/\delta)}{n}.
$$

Here C_1 is some universal constant. This gives us

$$
\begin{aligned}
\gamma \|\Delta\|^2_{\Sigma_{\mathcal{D}}+\lambda I} &\leq \|\nabla \ell_{\mathcal{D}}(\theta^\star)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} \|\Delta\|_{\Sigma_{\mathcal{D}}+\lambda I} + 4(\lambda\gamma + 2\alpha_2(1 + 2\gamma\alpha_1 B))B^2 \\
&\leq \sqrt{C_1 \cdot \frac{d + \log(1/\delta)}{n}} \|\Delta\|_{\Sigma_{\mathcal{D}}+\lambda I} + 4(\lambda\gamma + 2\alpha_2(1 + 2\gamma\alpha_1 B))B^2.
\end{aligned}
$$

Solving the above inequality gives us that for some constant C_2,

$$
\|\Delta\|_{\Sigma_{\mathcal{D}}+\lambda I} \leq C_2 \cdot \sqrt{\frac{d + \log(1/\delta)}{\gamma^2 n} + (\lambda + \alpha_2/\gamma + \alpha_1\alpha_2 B)B^2}.
$$