Reading notes for “Principled Reinforcement Learning with Human Feedback from Pairwise or K-wise Comparisons”

Yifan Xu

Sep 2025

1 Introduction

This part answers the questions in `https://zhuanlan.zhihu.com/p/15639428116`.

- **Motivation:** RLHF made large language models useful, but the theory lagged behind. Why does learning from *offline*, biased human comparisons not overfit? What guarantees do we have when a policy moves off the data support?
- **Evolution:** Early work includes pairwise preference learning (Bradley–Terry, Thurstone), contextual bandits with comparisons, and offline RL with *pessimism/conservatism* (UCB/LCB principles, CQL). Recent RLHF practice (SFT → DPO/PPO-KL) exposed the issue of *distribution shift*, pushing theory toward *offline preference-based RL*.
- **This paper's place:** It brings *pessimistic estimation* (lower confidence bounds on value) into preference-based RLHF and proves generalization bounds under coverage conditions, thereby bridging practical RLHF with offline RL/bandit theory.

2 Problem formulation

2.1 Language models

A *language model* is a conditional probability distribution

$$p_\theta(y \mid x),$$

where $x \in \mathcal{X}$ is the input context (e.g., a prompt), $y \in \mathcal{Y}$ is an output sequence (e.g., a text completion), and $\theta \in \mathbb{R}^d$ are model parameters. The model generates tokens $y_i \sim p_\theta(\cdot \mid x)$ sequentially.

For example, let $x =$“hello [eos]”, and the model may output $y =$“hi” in response to my input with probability $p(y \mid x) = p(\text{h} \mid \text{hello [eos]}) \cdot (\text{i} \mid \text{hello [eos] h})$.
As we have talked about MDP before, we can easily see that this generation process can be modeled as action selection with action being tokens. However, today we mainly use a more general formulation of this process, which is contextual bandit. Both formulations are used in RLHF.

2.2 Contextual bandit

A *contextual bandit* is a tuple $(\mathcal{S}, \mathcal{A}, \rho, r)$, where

- $\mathcal{S}$ is a context space with distribution ρ,
- $\mathcal{A}$ is an action space,
- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a reward function.

At each round, the learner observes a context $s \sim \rho$, selects an action $a \in \mathcal{A}$, and receives a reward $r(s,a)$. The goal is to learn a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ that maximizes expected reward

$$J(\pi) := \mathbb{E}_{s\sim\rho,\, a\sim\pi(\cdot|s)}[r(s,a)].$$

This is similar to RL, except that each trajectory contains only one state/context (s) and one action (a).

2.3 Language modeling under contextual bandit

We interpret text generation as a contextual bandit problem:

- Context $x \in \mathcal{X}$ is the prompt.
- Action $y \in \mathcal{Y}$ is the generated sequence/output.
- Reward $r(x,y)$ measures the quality of the output (e.g., human ratings).

Thus a language model $p_\theta(y \mid x)$ is a stochastic policy $\pi_\theta(a \mid s)$ in the contextual bandit setting. So in the following derivation, we use s for input x, and a for output y.

2.4 Comparison model

In practice, human feedback is often obtained by *comparisons*. And we mainly discuss *pairwise comparisons* as a typical example here. Given context x and two candidate outputs $a_1, a_2 \in \mathcal{Y}$, the annotator indicates which is preferred. This is modeled by a Bradley–Terry–Luce (BTL) model:

$$\Pr(a_1 \succ a_2 \mid s, \theta) = \frac{\exp(r_\theta(s,a_1))}{\exp(r_\theta(s,a_1)) + \exp(r_\theta(s,a_2))} = \frac{1}{1+\exp(r_\theta(s,a_1) - r_\theta(s,a_2))} = \sigma(r_\theta(s,a_1) - r_\theta(s,a_2))$$

where $r_\theta(s,a)$ is a reward model with parameter θ[1].

Beyond pairwise comparisons, one can also model *K-wise rankings*, where a human provides a preference ordering over K candidate outputs $a_1, \ldots, a_K \in \mathcal{Y}$ for the same context $s \in \mathcal{S}$. The Plackett–Luce (PL) model specifies the probability of observing a ranking $\sigma \in S_K$ (a permutation of $\{1, \ldots, K\}$) as

$$\Pr(\sigma \mid s, \theta) = \prod_{i=1}^{K} \frac{\exp\left(r_\theta(s, a_{\sigma(i)})\right)}{\sum_{j=i}^{K} \exp\left(r_\theta(s, a_{\sigma(j)})\right)},$$

where $r_\theta(s,a)$ is the parameterized reward function and $\sigma(i)$ denotes the item placed at position i in the ranking.

Intuitively, the PL model generates the ranking sequentially: at each step, the probability of selecting the next most-preferred item is proportional to its exponential score under r_θ.

2.5 Reinforcement learning from human feedback

Reinforcement learning from human feedback (RLHF) fits the above framework:

1. Train a reward model $r_\theta(s,a)$ from human comparisons using the BTL model.

   $$\ell_{\mathcal{D}}(\theta) = -\mathbb{E}_{s,a_1,a_2\sim\mathcal{D}}[y \log \sigma(r(s,a_1) - r(s,a_2)) + (1-y)\log(1 - \sigma(r(s,a_1) - r(s,a_2)))],$$

   where $y = 1$ means $a_1 \succ a_2 \mid s$ under human annotation, $\mathcal{D}$ is a offline preference dataset, $\mathcal{D} = \{s^i, a_1^i, a_2^i\}_{i=1}^N$.

2. Optimize a policy $\pi(a \mid s)$ to maximize the expected reward

   $$J(\pi) = \mathbb{E}_{x\sim\rho,\, y\sim\pi(\cdot|x)}[r_\theta(s,a)].$$

   This step mainly uses RL algorithms for optimization.

This converts subjective human preferences into a reward function, and then applies contextual bandit optimization to align the language model's outputs with human feedback.

[1] Throughout the note (as well as the original paper), we use θ for reward parameters instead of policy.

3 Main results (Learning from pairwise comparison)

Other interesting results in this paper involve discussion on learning from k-wise comparisons, and results derived using MDP formulation of RLHF, and the authors also discuss the connections with IRL.

3.1 MLE

1. MLE achieves near-optimal convergence to the true reward parameters under the BT model framework.
2. However, MLE provably fails at the performance of the induced policy.

The result is mainly derived under the linear reward assumption (The authors also give results under nonlinear rewards with regularity assumptions at Appendix A).

Assumption 3.1 (Linear reward)**.** The reward lies in the family of linear functions $r_\theta(s,a) = \theta^T\phi(s,a)$ for some known $\phi(s,a) \in \mathbb{R}^d$ with $\max_{s,a} \|\phi(s,a)\|_2 \le L$. Let θ^* be the true parameter. To ensure the identifiability of θ^*, we let $\theta^* \in \Theta_B$, where

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

The MLE estimator:

$$\begin{aligned}
&\hat{\theta}_{\text{MLE}} \in \underset{\theta \in \Theta_B}{\arg\min}\, \ell_{\mathcal{D}}(\theta), \\
&\ell_{\mathcal{D}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log\left(1(y^i=1)\cdot \frac{\exp(r_\theta(s^i,a_1^i))}{\exp(r_\theta(s^i,a_0^i)) + \exp(r_\theta(s^i,a_1^i))} + 1(y^i=0)\cdot \frac{\exp(r_\theta(s^i,a_0^i))}{\exp(r_\theta(s^i,a_0^i)) + \exp(r_\theta(s^i,a_1^i))}\right) \\
&\qquad = -\frac{1}{n}\sum_{i=1}^{n} \log\left(1(y^i=1)\cdot \sigma(\langle \theta, \phi(s^i,a_1^i) - \phi(s^i,a_0^i)\rangle) + 1(y^i=0)\cdot \sigma(\langle \theta, \phi(s^i,a_0^i) - \phi(s^i,a_1^i)\rangle)\right)
\end{aligned}$$

Lemma 3.2 (Estimation error for $\hat{\theta}_{\text{MLE}}$)**.** *For any $\lambda > 0$, with probability at least $1-\delta$,*

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

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

The optimality of this bound can be verified by the minimax bound derived in Shah et al. [2015].

$$\inf_{\hat{\theta}} \sup_{\theta^\star} \mathbb{E}\left[\left\|\hat{\theta} - \theta^\star\right\|_2\right] \ge \Omega\left(\sqrt{\frac{d}{n}}\right).$$

We first summarize the proof sketch adopted in this paper:

Sketch of proof:

- Establish the local convexity of the loss function near the minimizer θ^*, which allows the error term to be expressed via an inner product.
- Apply Hölder's inequality to decompose and bound the resulting inner product.
- Show that the gradient of the loss can be characterized as a sub-Gaussian random variable.
- Bound the gradient via Bernstein's inequality for sub-Gaussian random vectors with positive semi-definite quadratic forms [Hsu et al., 2012], as the semi-norm of covariance matrix is used.

Then we list the detailed proof (with my notes), the original proof can be found in the paper [Zhu et al., 2023][Appendix B.1].

Proof. Based on the MLE loss, we first simplify the notation, let $\boldsymbol{x}_i = \phi(s^i,a_1^i) - \phi(s^i,a_0^i)$. The goal is to bound the estimation error in the squared semi-norm $\|v\|^2_{\Sigma_{\mathcal{D}}+\lambda I} = v^\top(\Sigma_{\mathcal{D}} + \lambda I)v$.

Strong convexity of ℓ. We want to show that ℓ is strongly convex at θ^*, so that the following inequality holds

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

for $\Delta \in \mathbb{R}^d$ such that $\theta^* + \Delta \in \Theta_B$, where $\mu > 0$.

We then compute the Hessian for $\ell_{\mathcal{D}}$, as the NLL loss equals the standard CE loss under binary classification, we can re-write the loss function as

$$\ell_{\mathcal{D}} = -\frac{1}{n} \sum_{i=1}^{n} \left(y^i \log \sigma(\boldsymbol{z}_i) + (1 - y^i) \log(1 - \sigma(\boldsymbol{z}_i)) \right),$$

where $\boldsymbol{z}_i = \langle \theta^\top, \boldsymbol{x}_i \rangle$.

The derivative

$$\nabla_\theta \ell_{\mathcal{D}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} (\sigma(\boldsymbol{z}_i) - y^i) \boldsymbol{x}_i,$$

as

$$\begin{aligned} \sigma(-x) &= 1 - \sigma(x) \\ \sigma(x)' &= \sigma(x)(1 - \sigma(x)) \\ \log \sigma(x)' &= 1 - \sigma(x) \end{aligned}$$

The Hessian

$$\nabla^2_\theta \ell_{\mathcal{D}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sigma(\boldsymbol{z}_i)(1 - \sigma(\boldsymbol{z}_i)) \cdot \boldsymbol{x}_i \boldsymbol{x}_i^\top.$$

Based on the linear reward assumption, we can obtain that $\langle \theta, \boldsymbol{x}_i \rangle \in [-2LB, 2LB]$, so that

$$\sigma(\boldsymbol{z}_i)(1 - \sigma(\boldsymbol{z}_i)) \geq \frac{1}{2 + \exp(-2LB) + \exp(2LB)}.$$

Therefore, we have

$$\begin{aligned} \boldsymbol{v}^\top \nabla^2 \ell_{\mathcal{D}}(\theta) \boldsymbol{v} &\geq \frac{1}{n} \cdot \frac{1}{2 + \exp(-2LB) + \exp(2LB)} \cdot \sum_{i=1}^{n} \boldsymbol{v}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{v} \\ &= \frac{\gamma}{n} \|X\boldsymbol{v}\|_2^2 \quad \text{for all } \boldsymbol{v}, \end{aligned}$$

where $\gamma = \frac{1}{2+\exp(-2LB)+\exp(2LB)}$, $X \in \mathbb{R}^{n \times d}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the i^{th} row of X.

So ℓ is convex. Let $\Delta = \hat{\theta}_{\text{MLE}} - \theta^*$, and based on convexity, we have

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

Step 1: Function convexity

Bounding the estimation error. Since $\hat{\theta}_{\text{MLE}}$ is the minimizer for $\ell_{\mathcal{D}}$, we have $\ell_{\mathcal{D}}(\hat{\theta}_{\text{MLE}}) \leq \ell_{\mathcal{D}}(\theta^*)$.

$$\begin{aligned} &\ell_{\mathcal{D}}(\theta^* + \Delta) - \ell_{\mathcal{D}}(\theta^*) \leq 0 \\ \Rightarrow\ &\ell_{\mathcal{D}}(\theta^* + \Delta) - \ell_{\mathcal{D}}(\theta^*) - \langle \nabla \ell_{\mathcal{D}}(\theta^*), \Delta \rangle \leq -\langle \nabla \ell_{\mathcal{D}}(\theta^*), \Delta \rangle. \end{aligned}$$

As the LHS is lower-bounded by γ-convexity,

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

The second inequality holds due to Hölder's inequality, as $\|\cdot\|_M$ and $\|\cdot\|_{M^{-1}}$ are dual norms, which we'll prove later.

Step 2: Dual norm

Now to further bound the term $\|\nabla \ell_{\mathcal{D}}(\theta^*)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}$, we recall the gradient

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

We can observe that the term inside the paratheses can be seen as another random variable, as y^i is sampled from the Bernoulli distribution, $y^i \sim \text{Bernoulli}(\frac{1}{1+\exp(-\langle \theta^*, x_i \rangle)})$. So we construct a random vector $V \in \mathbb{R}^n$ such that

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

Based on this random vector, we can re-write the gradient term as

$$\nabla \ell_{\mathcal{D}}(\theta^\star) = -\frac{1}{n} X^\top V.$$

Also the random vector satisfies the sub-Gaussianity.

$$\begin{aligned} &\mathbb{E}[V_i] = \mathbb{E}[\mathbb{E}[V_i \mid x_i]] = 0, \Rightarrow \mathbb{E}[V] = \mathbf{0}, \\ &|V_i| \leq 1. \end{aligned}$$

Step 3: Sub-Gaussian variable

Define an n-dimensional matrix $M := \frac{1}{n^2} X(\Sigma_{\mathcal{D}} + \lambda I)^{-1} X^\top$, we can further write

$$\|\nabla \ell_{\mathcal{D}}(\theta^*)\|_{(\sum_{\mathcal{D}}+\lambda I)^{-1}} = \frac{1}{n^2} V^T X \left(\sum_{\mathcal{D}} + \lambda I \right)^{-1} X^T V = V^T M V.$$

As V is a sub-Gaussian random vector, we want to use the Bernstein's inequality for sub-Gaussian random variables in quadratic form [Hsu et al., 2012]. Therefore, we bound $\text{Tr}(M)$, $\text{Tr}(M^2)$ and the operator norm of M, let $X^\top X = U \Lambda U^\top$ be the eigenvalue decomposition for $X^\top X$, and the SVD for X can be written as $QSU^\top$, we can obtain the following results by applying cyclic trace property.

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

Finally we can apply the results in [Hsu et al., 2012] to derive the bound

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

Note that the term $\frac{1}{n}\sqrt{d \log(1/\delta)}$ in the original bound is simplified into $d + \log(1/\delta)$ with the constant C_1.

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

Solving the above inequality gives the final result with some constant C_2.

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

Step 4: Bernstein's inequality

□

This lemma gives a near-optimal convergence result of the MLE estimator under the linear reward setting in RLHF, which provides theoretical guarantee for RLHF algorithms under the preference model assumption (BT or PL).

The mathematic techniques used in this proof include:

Function convexity First recall the definition for function convexity.

For a differentiable function f, if f is μ-strongly convex ($\mu > 0$), it means that

$$f(x+v) \geq f(x) + \langle \nabla f(x), v \rangle + \frac{\mu}{2}\|v\|^2 \qquad \text{for all } x, y \in \mathbb{R}^d.$$

For a twice differentiable function f, if f is μ-strongly convex, it means that

$$\nabla^2 f(x) \succeq \mu I \qquad \text{for all } x \in \mathbb{R}^d.$$

Dual norm The dual norm of a norm $\|\cdot\|$ on $\mathbb{R}^n$ is defined for $\boldsymbol{z} \in \mathbb{R}^n$ by

$$\|\boldsymbol{z}\|_* = \sup\{\boldsymbol{z}^\top \boldsymbol{x} : \|\boldsymbol{x}\| \leq 1\}.$$

This supremum is achieved over the closed unit ball of the primal norm.

A useful result for dual norm is *Hölder's inequality*, which states that for any norm $\|\cdot\|$ with dual norm $\|\cdot\|_*$, and for vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|_*.$$

This inequality determines ℓ_p-norm duals. The dual of $\|\cdot\|_1$ is $\|\cdot\|_\infty$:

$$\|\boldsymbol{z}\|_* = \sup_{\|\boldsymbol{x}\|_1 \leq 1} \boldsymbol{z}^\top \boldsymbol{x} = \max_i |z_i| = \|\boldsymbol{z}\|_\infty.$$

The dual of $\|\cdot\|_\infty$ is $\|\cdot\|_1$:

$$\|\boldsymbol{z}\|_* = \sup_{\|\boldsymbol{x}\|_\infty \leq 1} \boldsymbol{z}^\top \boldsymbol{x} = \sum_{i=1}^n |z_i| = \|\boldsymbol{z}\|_1.$$

For $p \in (1, \infty)$, the dual of $\|\cdot\|_p$ is $\|\cdot\|_q$ with $q = p/(p-1)$:

$$\|\boldsymbol{z}\|_* = \sup_{\|\boldsymbol{x}\|_p \leq 1} \boldsymbol{z}^\top \boldsymbol{x} = \left(\sum_{i=1}^n |z_i|^q\right)^{1/q} = \|\boldsymbol{z}\|_q,$$

where the supremum is attained at $x_i = \operatorname{sgn}(z_i)|z_i|^{q/p}/\|\boldsymbol{z}\|_q^{q/p}$.

In this paper, the proof uses the dual for norm defined on matrix M, which is $\|\cdot\|_M, \|\cdot\|_{M^{-1}}$. We prove that the Hölder's inequality holds under this circumstance.

Lemma 3.3. *Let $M \in \mathbb{R}^{n \times n}$ be positive definite. Define $\|x\|_M := \sqrt{x^\top M x}$. Then the dual norm of $\|\cdot\|_M$ is*

$$\|y\|_* = \sup_{x \neq 0} \frac{x^\top y}{\|x\|_M} = \|y\|_{M^{-1}} = \sqrt{y^\top M^{-1} y}.$$

Consequently (Hölder/Cauchy–Schwarz for this dual pair),

$$|\langle x, y \rangle| \leq \|x\|_M \, \|y\|_{M^{-1}} \qquad \forall\, x, y \in \mathbb{R}^n.$$

Proof. Since $M \succ 0$, there exists $M^{1/2} \succ 0$ with $M^{1/2}M^{1/2} = M$ and $M^{-1/2} = (M^{1/2})^{-1}$. For $y \in \mathbb{R}^n$,

$$\|y\|_* = \sup_{x \neq 0} \frac{x^\top y}{\sqrt{x^\top M x}} = \sup_{\|x\|_M = 1} x^\top y.$$

Let $z := M^{1/2}x$ so that $x = M^{-1/2}z$ and $\|x\|_M = \|z\|_2$. Then

$$x^\top y = (M^{-1/2}z)^\top y = z^\top M^{-1/2} y.$$

Therefore

$$\|y\|_* = \sup_{\|z\|_2 = 1} z^\top M^{-1/2} y = \|M^{-1/2}y\|_2 = \sqrt{y^\top M^{-1} y}.$$

The supremum is achieved at $z^* = \frac{M^{-1/2}y}{\|M^{-1/2}y\|_2}$, i.e., $x^* = \frac{M^{-1}y}{\|y\|_{M^{-1}}}$.

For the inequality, using the Euclidean Cauchy–Schwarz inequality,

$$|x^\top y| = |(M^{1/2}x)^\top (M^{-1/2}y)| \leq \|M^{1/2}x\|_2 \, \|M^{-1/2}y\|_2 = \|x\|_M \, \|y\|_{M^{-1}}.$$

This is Hölder's inequality for the dual pair $(\| \cdot \|_M, \| \cdot \|_{M^{-1}})$. □

Sub-Gaussian variable Recall the gradient for a binary cross entropy loss,

$$\nabla_\theta \ell_{\mathcal{D}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} (\sigma(\boldsymbol{z}_i) - y^i)\boldsymbol{x}_i,$$

where $y^i \sim \text{Bernoulli}(p_i)$.

We can directly define a new random variable V_i,

$$V_i = \sigma(\boldsymbol{z}_i) - y^i,$$

and its distribution

$$V_i = \begin{cases} \sigma(\boldsymbol{z}_i), & \text{w.p. } 1 - p_i, \\ \sigma(\boldsymbol{z}_i) - 1, & \text{w.p. } p_i. \end{cases}$$

When the model is well-specified, we can conclude that V_i is sub-Gaussian.

Bernstein's inequality The paper uses a result on bounding the quadratic form of sub-Gaussian random vectors in Hsu et al. [2012], which I find a useful bound in many scenarios.

3.2 Pessimistic MLE

To introduce the pessimistic MLE, the paper first shows the **failure of MLE w.r.t. policy performance**, measured under sub-optimality.

$$\text{SubOpt}(\hat{\pi}) := \mathbb{E}_{s \sim \rho} \left[r_{\theta^\star}(s, \pi^\star(s)) - r_{\theta^\star}(s, \hat{\pi}(s)) \right].$$

Theorem 3.4 (Failure of MLE in policy optimization)**.** *There exists a linear bandit with four actions and a sampling distribution such that for any $n > 1$,*

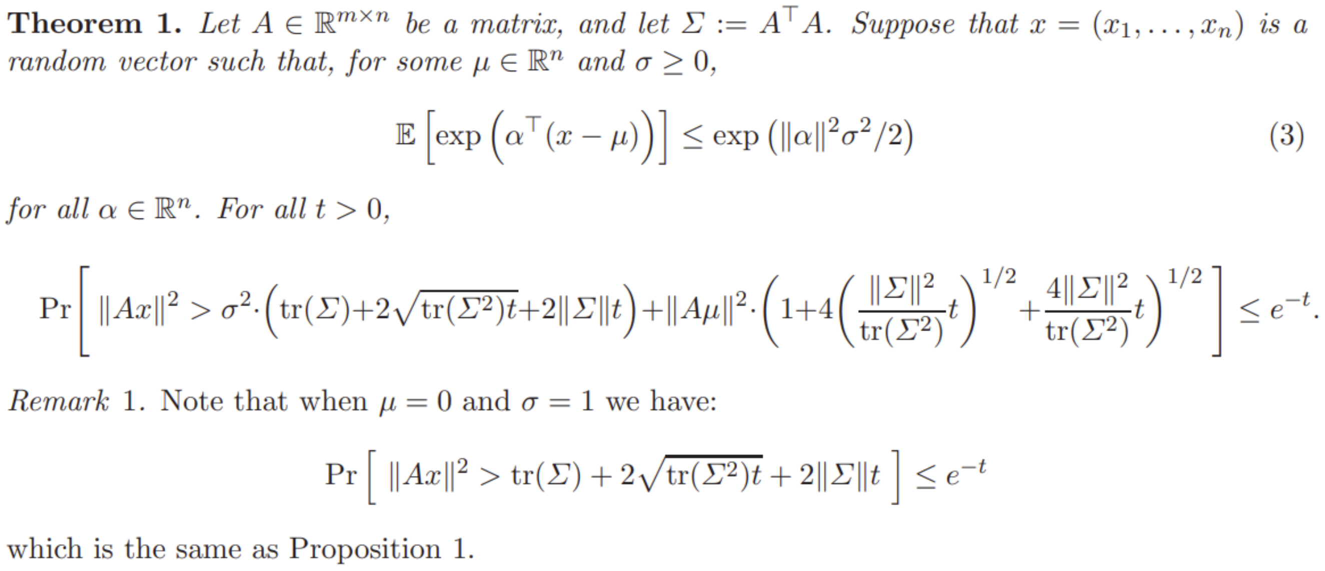
$$\mathbb{E}[\text{SubOpt}(\hat{\pi}_{\text{MLE}})] \geq 0.1.$$

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

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

Here C is some universal constant.

This theorem corresponds to Zhu et al. [2023][Theorem 3.9]. I recommend the readers to read the proof from the original paper, so I directly put the screen shot of the original proof below for convenience.



Theorem 1. *Let $A \in \mathbb{R}^{m\times n}$ be a matrix, and let $\Sigma := A^\top A$. Suppose that $x = (x_1, \ldots, x_n)$ is a random vector such that, for some $\mu \in \mathbb{R}^n$ and $\sigma \geq 0$,*

$$\mathbb{E}\left[\exp\left(\alpha^\top(x-\mu)\right)\right] \leq \exp\left(\|\alpha\|^2\sigma^2/2\right) \tag{3}$$

for all $\alpha \in \mathbb{R}^n$. For all $t > 0$,

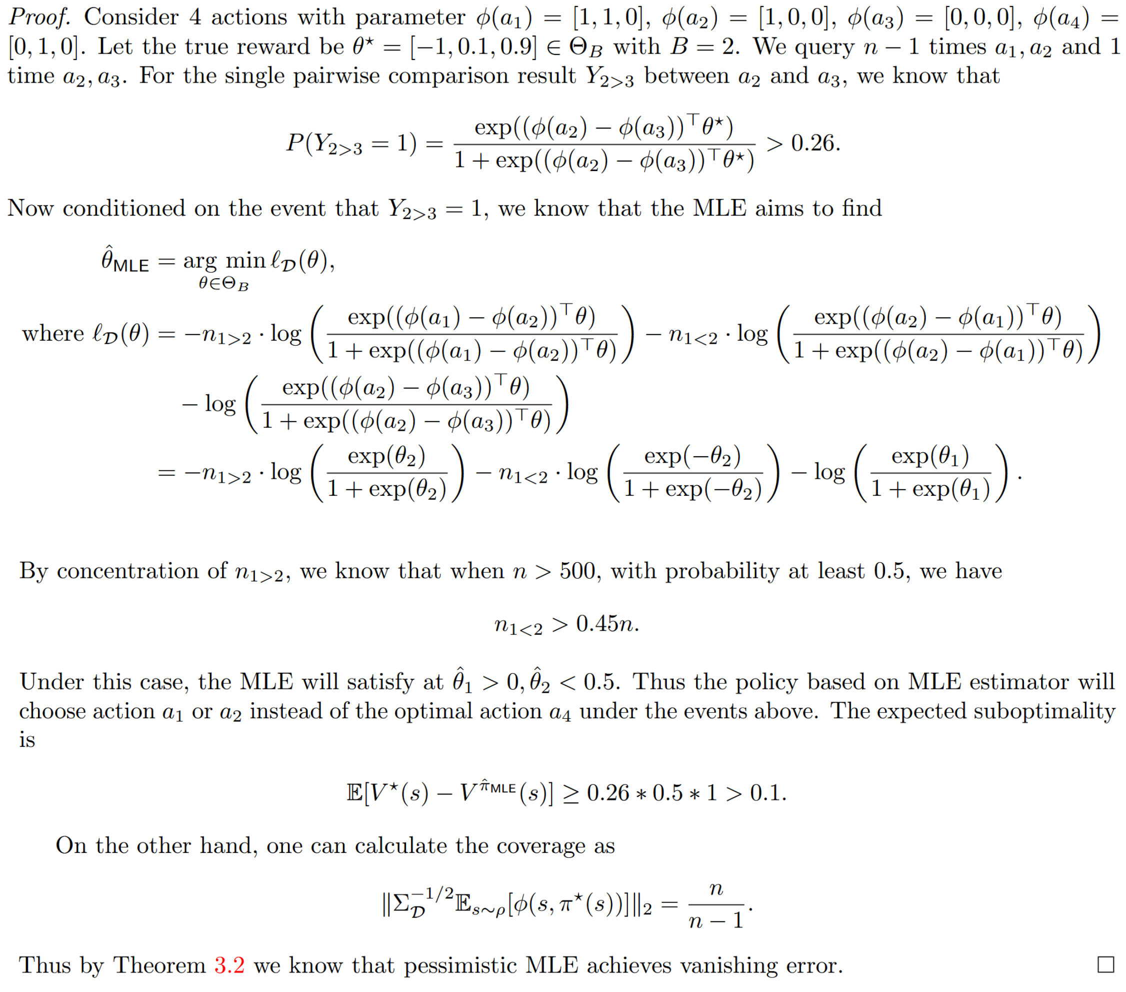
$$\Pr\left[\ \|Ax\|^2 > \sigma^2\cdot\left(\operatorname{tr}(\Sigma)+2\sqrt{\operatorname{tr}(\Sigma^2)t}+2\|\Sigma\|t\right)+\|A\mu\|^2\cdot\left(1+4\left(\frac{\|\Sigma\|^2}{\operatorname{tr}(\Sigma^2)}t\right)^{1/2}+\frac{4\|\Sigma\|^2}{\operatorname{tr}(\Sigma^2)}t\right)^{1/2}\ \right] \leq e^{-t}.$$

Remark 1. Note that when $\mu = 0$ and $\sigma = 1$ we have:

$$\Pr\left[\ \|Ax\|^2 > \operatorname{tr}(\Sigma) + 2\sqrt{\operatorname{tr}(\Sigma^2)t} + 2\|\Sigma\|t\ \right] \leq e^{-t}$$

which is the same as Proposition 1.

Figure 1: Theorem adopted from Hsu et al. [2012][Theorem 1.]



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

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

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 3.2 we know that pessimistic MLE achieves vanishing error. □

Figure 2: Proof adopted from Zhu et al. [2023][Appendix B.3]

Sketch of proof: The proof idea is to **construct a bandit where data heavily oversamples comparisons between suboptimal actions and provides almost no information about the optimal one**. A single misleading outcome then shifts the MLE into a parameter region that consistently favors the wrong actions, so the greedy policy stays suboptimal with constant probability regardless of sample size. This construction shows that the failure stems not from limited data but from coverage gaps and greedy amplification of estimation errors, while pessimistic LCB estimates prevent such over-optimism in unexplored directions.

As the reward would most likely to be inaccurate with OOD data, so the failure is mainly due to the greedy policy selection w.r.t. MLE. **This is the risk if you learn reward from offline data!**

To address this problem, *pessimism principle* is proposed, which is a concept adopted in offline RL (Offline trajectories normally lack full coverage of the entire state-action space, causing problems of overfitting).

We first show how pessimistic MLE is constructed, then we explain why.

First, based on lemma 3.2, we know that with probability at least $1-\delta$, θ^* lies in the following region:

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

So for the estimator $\hat{\theta}$, we can construct a confidence set, where the true parameter falls in with high probability, we simply denote it as

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

where $f(n,d,\delta,\lambda)$ corresponds to the bound in lemma 3.2.

And the pessimistic objective becomes the *minimal return value* within the confidence set

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

Derivation of the second equality. We start from the definition

$$\hat{J}(\pi) = \min_{\theta \in \Theta(\hat{\theta},\lambda)} \mathbb{E}_{s\sim q}\left[\theta^\top(\phi(s,\pi(s)) - v)\right],$$

where the confidence set is

$$\Theta(\hat{\theta},\lambda) = \left\{ \theta : \|\theta - \hat{\theta}\|_{\Sigma_D + \lambda I} \le f(n,d,\delta,\lambda) \right\}.$$

Let

$$z := \mathbb{E}_{s\sim q}[\phi(s,\pi(s))] - v,$$

then the objective becomes

$$\hat{J}(\pi) = \min_{\theta \in \Theta(\hat{\theta},\lambda)} z^\top \theta.$$

Now write $\theta = \hat{\theta} + \Delta$, where Δ is the deviation. The constraint is

$$\|\Delta\|_{\Sigma_D+\lambda I} \le f(n,d,\delta,\lambda),$$

and the objective becomes

$$z^\top\theta = z^\top\hat{\theta} + z^\top\Delta.$$

Thus,

$$\hat{J}(\pi) = z^\top\hat{\theta} + \min_{\|\Delta\|_{\Sigma_D+\lambda I} \le f} z^\top\Delta.$$

By Cauchy–Schwarz inequality under the $(\Sigma_D + \lambda I)$-norm,

$$z^\top\Delta \ge -\|z\|_{(\Sigma_D+\lambda I)^{-1}} \cdot \|\Delta\|_{\Sigma_D+\lambda I}.$$

Hence,

$$\min_{\|\Delta\|_{\Sigma_D+\lambda I}\le f} z^\top \Delta = -f(n,d,\delta,\lambda)\cdot \|z\|_{(\Sigma_D+\lambda I)^{-1}}.$$

Combining terms, we obtain

$$\hat{J}(\pi) = z^\top \hat{\theta} - f(n,d,\delta,\lambda)\cdot \|z\|_{(\Sigma_D+\lambda I)^{-1}},$$

which can be equivalently written as

$$\hat{J}(\pi) = \big(\mathbb{E}_{s\sim q}[\phi(s,\pi(s))] - v\big)^\top \hat{\theta} - \Big\|(\Sigma_D+\lambda I)^{-1/2}\big(\mathbb{E}_{s\sim q}[\phi(s,\pi(s))] - v\big)\Big\|_2 \cdot f(n,d,\delta,\lambda).$$

□

Why the LCB avoids overfitting. The lower confidence bound $\hat{J}(\pi)$ subtracts a penalty term

$$f(n,d,\delta,\lambda)\cdot \|z\|_{(\Sigma_D+\lambda I)^{-1}},$$

which explicitly accounts for statistical uncertainty. If the data distribution $\mathcal{D}$ does not cover the feature direction z well, the Mahalanobis norm $\|z\|_{(\Sigma_D+\lambda I)^{-1}}$ becomes large, and the algorithm would lower the reward estimate by adding this penalty. Therefore, policies that exploit poorly supported directions in the data receive low $\hat{J}(\pi)$ values, even if their empirical estimate reward is high. This conservatism prevents the algorithm from over-optimistic reward estimation against OOD data.

Remark 3.5 (Concentration coefficient). The term $\left\|(\Sigma_{\mathcal{D}}+\lambda I)^{-\frac{1}{2}}\left(\mathbb{E}_{s\sim q}[\phi(s,\pi(s))]\right)\right\|_2$ is referred to as "concentration coefficient" in offline RL, bounding this term means that the offline dataset $\mathcal{D}$ satisfies good coverage of the target vector $\mathbb{E}_{s\sim q}[\phi(s,\pi(s))]$ in the feature space. Otherwise, the model may not behave well on data not seen during training.

Remark 3.6 (Baseline $\boldsymbol{v}$). Subtracting a constant vector in feature space will not change the induced policy, but may affect the concentratability coefficient. Consider the case where the differences between features lie in the same subspace, while the feature ϕ itself does not.

Remark 3.7 (Complex r_θ). When r_θ is a neural network, the confidence set may not be tractable, so other forms of pessimism can be introduced, a widely used technique in RLHF is to introduce a KL constraint at the objective function, to keep the updated policy model close to the initial policy.

Theorem 3.8 (Sub-optimality of pessimistic MLE). *Let $\hat{\pi}_{PE}$ be the output of Algorithm 1 when taking $\hat{\theta} = \hat{\theta}_{MLE}$, $f(n,d,\delta,\lambda) = C\cdot\sqrt{\frac{d+\log(1/\delta)}{\gamma^2 n} + \lambda B^2}$, $q=\rho$. For any $\lambda>0$ and $v\in\mathbb{R}^d$, with probability at least $1-\delta$,*

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

Proof. First, we clarify some notations.

$$\begin{aligned}
J(\pi) &= \mathbb{E}_{s\sim\rho}[\theta^{*\top}(\phi(s,\pi(s)))]\\
\hat{J}(\pi) &= \min_{\theta\in\Theta(\hat{\theta},\lambda)} \mathbb{E}_{s\sim\rho}[\theta^\top(\phi(s,\pi(s)) - v)]\\
J'(\pi) &= J(\pi) - \langle\theta^*, v\rangle = \mathbb{E}_{s\sim\rho}[\theta^{*\top}(\phi(s,\pi(s)) - v)].
\end{aligned}$$

π^* is the global optimal policy, $\hat{\pi}_{\mathrm{PE}}$ is the optimal policy under $\hat{J}(\pi)$.

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

As $\hat{\pi}_{\mathrm{PE}}$ is the minimizer for $\hat{J}(\pi)$, so for the second difference term above, we have

$$\hat{J}(\pi^\star) - \hat{J}(\hat{\pi}_{\mathrm{PE}}) \le 0.$$

The for the third difference term, we can write:

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

As $\theta^{*}\in\Theta(\hat{\theta}_{\mathrm{MLE}},\lambda)$ w.p. at least $1-\delta$ from lemma 3.2. So w.p. at least $1-\delta$ such that

$$\hat{J}(\pi)-J'(\pi)\leq 0.$$

Then we have

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

We bound the second term by observing that it can be decomposed using Cauchy-Schwarz inequality.

$$\begin{aligned}
\mathbb{E}_{s\sim\rho}\left[\left(\hat{\theta}_{\mathrm{MLE}}-\theta\right)^{\top}\left(\phi\left(s,\pi^{\star}(s)\right)-v\right)\right] &=\left(\hat{\theta}_{\mathrm{MLE}}-\theta\right)^{\top}\mathbb{E}_{s\sim\rho}\left(\phi\left(s,\pi^{\star}(s)\right)-v\right)\\
&\leq\|\hat{\theta}_{\mathrm{MLE}}-\theta\|_{\Sigma_{\mathcal{D}}+\lambda I}\|\mathbb{E}_{s\sim\rho}\left(\phi\left(s,\pi^{\star}(s)\right)-v\right)\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}\\
&=f(n,d,\delta,\lambda)\|(\Sigma_{\mathcal{D}}+\lambda I)^{-1/2}\mathbb{E}_{s\sim\rho}\left(\phi\left(s,\pi^{\star}(s)\right)-v\right)\|_{2}
\end{aligned}$$

Besides, as $\theta^{*}\in\Theta(\hat{\theta}_{\mathrm{MLE}})$ w.p. $1-\delta$, we can conclude that

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

□

References

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. 2012.

Nihar Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin Wainwright. Estimation from Pairwise Comparisons: Sharp Minimax Bounds with Topology Dependence. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 856–865, San Diego, California, USA, 09–12 May 2015. PMLR. URL `https://proceedings.mlr.press/v38/shah15.html`.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 43037–43067. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/zhu23f.html`.