



Lecture 5. Online MDPs and One-Pass Bandits

Peng Zhao

School of AI

Nanjing University

July 16, 2026





Outline

- Bandits and MDP
- One-Pass Bandits
- RL Implications



Part 1. Bandits and RL

- MDP
- Tabular MDP
- Function Approximation
- Policy Gradient

MDP

- In RL, the interaction between the agent and environments is usually captured/modelled as an MDP.

Definition 1 (MDPs). An infinite-horizon, discounted Markov decision processes (MDPs) is denoted by $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$, where

- A state space $\mathcal{S}$, which may be finite or infinite.
- An action space $\mathcal{A}$, which also may be discrete or infinite.
- A transition function $P : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ is the space of probability distributions over $\mathcal{S}$. $P(s' | s, a)$ is the probability of transitioning into s' upon taking a in state s .
- A reward function $r : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$.
- A discount factor $\gamma \in [0, 1)$, which defines an effective horizon for the problem.
- An initial state distribution $\mu \in \Delta(\mathcal{S})$, which specifies the initial state s_0 .

RL concepts

- Trajectory: $\tau_t := (s_0, a_0, r_0, s_1, \dots, a_{t-1}, r_{t-1}, s_t)$.
- Policy: π
 - A stationary policy $\pi : S \rightarrow \Delta(A)$ specifies a decision-making strategy in which the agent chooses actions based only on the current state, i.e. $a_t \sim \pi(\cdot | s_t)$.
 - A deterministic, stationary policy is of the form $\pi : S \rightarrow A$.
- Goal: maximize cumulative reward $V_M^\pi(s) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid \pi, s_0 = s \right]$

Tabular MDP

First observe an initial state s_1 from a feasible set $\mathcal{S}$

At each round step $h = 1, 2, \dots H$

1. the agent choose an action a_h from a feasible set $\mathcal{A}$
2. observe the reward $R_h \sim r_h(s_h, a_h)$ and next state $s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, a_h)$

↓ Only contain present s_h and a_h ↓

- **Memoryless** Property : The future depends only on the present, not on the past.
- Full Observability : The agent has access to the complete and true state of the environment at each time step.

Tabular MDP

- Instead optimize one step reward in bandits, we need maximize the **cumulative** reward.

⇒ State-action function and state function under policy π

$$Q_h^\pi(s, a) = \mathbb{E} \left[\sum_{t=h}^H R_t \middle| s_h = s, a_h = a \right], V_h^\pi(s) = \mathbb{E} \left[\sum_{t=h}^H R_t \middle| s_h = s \right]$$

Expectation is over reward, transition, and policy randomness.

We optimize the following regret to achieve the goal.

⇒ Regret is defined as the difference from the optimal policy.

$$REG(\pi) = \max_{\pi'} V_1^{\pi'}(s_1) - V_1^\pi(s_1) \triangleq V_1^*(s_1) - V_1^\pi(s_1)$$

Bellman Equation

- Policy Evaluation via Bellman Recursion

$$Q_h^\pi(s, a) = r_h(s, a) + \sum_{s'} \mathbb{P}_h(s' | s, a) V_{h+1}^\pi(s, a) \triangleq r_h(s, a) + \mathbb{P}_h V_{h+1}^\pi(s, a),$$
$$V_h^\pi(s) = Q_h^\pi(s, \pi_h(s)).$$

Key Idea: Simplify the computation of Q and V into a recursive process.

If $r_h(s, a)$ and $P_h(\cdot | s, a)$ are known, optimize via DP.

DP version

For dynamic programming:

- (1) state: $f[h][s]$, the maximum expected future reward at state s and step h ;
- (2) transition: $f[h][s] = \max_a \{r_h(s, a) + \sum_{s'} \mathbb{P}_h(s' | s, a) f[h+1][s']\}$;
- (3) computation: backward from $h = H$ to $h = 1$.

Exploration vs Exploitation

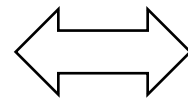
- ❑ **Known model:** dynamic programming yields $\pi_* = \arg \max_{\pi} V_1^{\pi}(s_1)$
- ❑ **Unknown model:** agent must *learn while interacting* with the environment.

Exploration

Try *uncertain actions* to collect information.

- improves future estimates of rewards and transitions

Cost: possibly *lower immediate reward*



trade-off

Exploitation

Choose action with *highest estimated value*.

- maximizes reward under the current model

Risk: inaccurate estimates may cause *persistent mistakes*

The key challenge is to balance *immediate reward* and *information gain*.

Tabular Episodic MDP

- With unknown environments, it's necessary to make episodic trials.
- Episodic MDP: repeat the same task many times, each for H steps.
 - H : horizon length of each episode
 - K : number of episode
 - $T = KH$: total number of interaction steps

For each episode $k = 1, \dots, K$

First observe an initial state s_1 from a feasible set $\mathcal{S}$

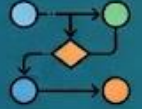
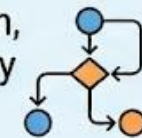
At each round step $h = 1, 2, \dots, H$

1. the agent choose an action $a_{k,h}$ from a feasible set $\mathcal{A}$
2. observe reward $R_{k,h} \sim r_{k,h}(s_{k,h}, a_{k,h})$ and next state $s_{k,h+1} \sim \mathbb{P}_h(\cdot | s_{k,h}, a_{k,h})$

Tabular episodic MDPs vs. MAB

❑ **Regret over K episodes:** $\text{REG}(K) = \sum_{k=1}^K V_1^*(s_{k,1}) - V_1^{\pi_k}(s_{k,1})$.

❑ **Episodic MDP:** $M = \left\{ \mathcal{S}, \mathcal{A}, H, \{r_h\}_{h \in [H]}, \{\mathbb{P}_h\}_{h \in [H]} \right\}$.

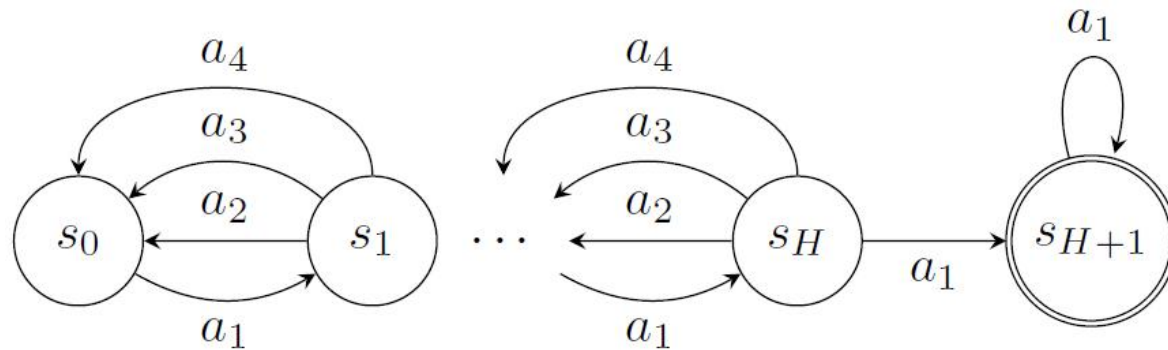
FEATURE	 Episodic MDP 	 Multi-Armed Bandit (MAB) 
	Key Difference Sequential decisions: Actions affect current states, future possibilities, and total rewards within finite episodes.	Key Difference Separate decisions: Actions (arm selections) are independent and only affect immediate rewards.
	Exploration Explore by selecting an action, leading to a random trajectory through multiple states. 	Exploration Explore by selecting any arm (action) in any step; access is direct to any option. 
	Exploitation Learn a policy π to maximize expected total return (V_1^π), prioritizing long-term goals. $V_\pi^\pi =$	Exploitation Select the arm with the highest expected immediate reward (optimal arm). 

Key relation

MAB is a **special case** of episodic MDP with **no state transition**.

UCB for MAB

- We cannot directly visit any state-action pair. Consider this example, we must use **strategic exploration** to visit s_{H+1} . We try to use UCB algorithm in MAB.



s_{H+1} is hard to visit
with random exploration.

UCB Algorithm (known as UCB1)

At each round $t = 1, 2, \dots$

- (1) Choose arm $a_t = \arg \max_{a \in [K]} \mathbf{UCB}_{t-1}(a)$
- (2) Observe reward r_t and update the estimation $\hat{\mu}_t$
- (3) Update upper confidence bounds $\mathbf{UCB}_t(a)$ by new estimation

Employ UCB to MDPs

Assumptions

- Rewards are known and deterministic: $R_{k,h} = r_h(s_{k,h}, a_{k,h})$ and r_h is known.
- The only unknown quantity is transition kernel $\mathbb{P}_h$; Assuming $0 \leq r_h \leq 1$.

➤ UCB principle: **estimate the unknown parameters**, construct confidence sets, and choose the **optimistic model**.

Problem: Given observations $s_{i,h}, a_{i,h}, s_{i,h+1}^k_{i=1}$, how should we estimate the transition probability $\mathbb{P}_h$?

Employ UCB to MDPs

Problem: We have $\forall i \in [k](s_{i,h}, a_{i,h}, s_{i,h+1})$, then how to estimate $\mathbb{P}_h$?

We define

$$N_h^k(s, a) = \sum_{i=1}^k \mathbb{1} \{(s_{i,h}, a_{i,h}) = (s, a)\}$$
$$N_h^k(s, a, s') = \sum_{i=1}^k \mathbb{1} \{(s_{i,h}, a_{i,h}, s_{i,h+1}) = (s, a, s')\}$$

Then estimated transition probability is as follows:

$$\hat{\mathbb{P}}_h^k(s' \mid s, a) := \begin{cases} \frac{N_h^k(s, a, s')}{N_h^k(s, a)} & \text{if } N_h^k(s, a) > 0, \\ \text{any distribution on } \mathcal{S} & \text{if } N_h^k(s, a) = 0. \end{cases}$$

Employ UCB to MDPs

For each $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, define

$$N_h^k(s, a) := \sum_{i=1}^k \mathbb{1}\{(s_{i,h}, a_{i,h}) = (s, a)\}, \quad \Rightarrow \quad \hat{\mathbb{P}}_h^k(s' \mid s, a) := \begin{cases} \frac{N_h^k(s, a, s')}{N_h^k(s, a)}, & N_h^k(s, a) > 0, \\ \text{any distribution over } \mathcal{S}, & N_h^k(s, a) = 0. \end{cases}$$
$$N_h^k(s, a, s') := \sum_{i=1}^k \mathbb{1}\{(s_{i,h}, a_{i,h}, s_{i,h+1}) = (s, a, s')\}.$$

We then analysis the estimation error to build **high confidence interval**.

Lemma 1 (Model error against V^*). Fix $\delta \in (0, 1)$ and let $L = \ln(2SAHK/\delta)$. With probability at least $1 - \delta$, simultaneously for all $h \in [H]$, all (s, a) , and all episodes k with $N_h^k(s, a) \geq 1$,

$$\left| \left(\hat{\mathbb{P}}_h^k(\cdot \mid s, a) - \mathbb{P}_h(\cdot \mid s, a) \right)^\top V_{h+1}^* \right| \leq H \sqrt{\frac{L}{N_h^k(s, a)}}.$$

Proof of Estimation error

Lemma 1 (Model error against V^*). Fix $\delta \in (0, 1)$ and let $L = \ln(2SAHK/\delta)$. With probability at least $1 - \delta$, simultaneously for all $h \in [H]$, all (s, a) , and all episodes k with $N_h^k(s, a) \geq 1$,

$$\left| \left(\hat{\mathbb{P}}_h^k(\cdot | s, a) - \mathbb{P}_h(\cdot | s, a) \right)^\top V_{h+1}^* \right| \leq H \sqrt{\frac{L}{N_h^k(s, a)}}.$$

Proof. Fix (h, s, a) and each time you visit (s, a) in h -step, s' i.i.d samples from $\mathbb{P}_h(\cdot | s, a)$. We set $n = N_h^K(s, a)$ and $s'_1, s'_2, \dots, s'_n$ is the next state according to n samples. Let $Y_i = V_{h+1}^*(s'_i)$, then

$$\hat{\mathbb{P}}_h^k(\cdot | s, a)^\top V_{h+1}^* = \frac{1}{n} \sum_{i=1}^n Y_i, \mathbb{E}[Y_i] = \mathbb{P}_h^*(\cdot | s, a)$$

□

Proof of Estimation error

Proof.

$$\hat{\mathbb{P}}_h^k(\cdot|s, a)^\top V_{h+1}^* = \frac{1}{n} \sum_{i=1}^n Y_i, \mathbb{E}[Y_i] = \mathbb{P}_h^*(\cdot|s, a)$$

Lemma 2. (Hoeffding's inequality). Suppose $X_1, \dots, X_n$ are independent, identically distributed random variables with mean μ . Let $\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[\bar{X}_n]$. If $X_i \in [b_-, b_+]$ almost surely, then for any $\epsilon > 0$,

$$\Pr(|\bar{X}_n - \mu| \geq \epsilon) \leq 2 \exp \left(-\frac{2n\epsilon^2}{(b_+ - b_-)^2} \right).$$

then we have

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{u=1}^n Y_i - \mathbb{E}[Y_i] \right| \geq \epsilon \right) \leq 2 \exp \left(-\frac{2n\epsilon^2}{H^2} \right)$$

Union bound over
(h, s, a) and n gives $\frac{\delta}{2}$

Setting $\epsilon = H \sqrt{\ln(4SAHK/\delta)/2n}$ makes the RHS at most $\frac{\delta}{2SAHK}$. $\square$

Employ UCB to MDPs

To handle case when $N_h^k(s, a) = 0$, we change the lemma as follows.

Lemma 1 (Model error against V^*). Fix $\delta \in (0, 1)$ and let $L = \ln(2SAHK/\delta)$. With probability at least $1 - \delta$, simultaneously for all $h \in [H]$, all (s, a) , and all episodes k with $N_h^k(s, a) \geq 1$,

$$\left| \left(\hat{\mathbb{P}}_h^k(\cdot \mid s, a) - \mathbb{P}_h(\cdot \mid s, a) \right)^\top V_{h+1}^* \right| \leq H \sqrt{\frac{L}{N_h^k(s, a)}}.$$

previous $H \sqrt{\frac{L}{N_h^k(s, a) + 1}}$

Add estimation error to bellman function and get UCB functions as follows.

$$\begin{aligned} \hat{Q}_h^k(s, a) &= r_h(s, a) + 2H \sqrt{\frac{L}{N_h^k(s, a) + 1}} + \hat{\mathbb{P}}_h \hat{V}_{h+1}^k(s, a) \\ \hat{V}_h^k(s) &= \hat{Q}_h^k(s, \pi_h^k(s)) \\ \pi_h^k(s) &= \arg \max_a \hat{Q}_h^k(s, a) \end{aligned}$$

$b_h^k(s, a)$

Optimism

Lemma 3 (Optimistic Value Estimation). *With probability at least $1 - \delta$, uniformly over all $h \in [H]$, (s, a) , and episodes k satisfying $N_h^k(s, a) \geq 1$.*

$$\begin{aligned}\hat{Q}_h^k(s, a) &= r_h(s, a) + 2H \sqrt{\frac{L}{N_h^k(s, a) + 1}} + \hat{\mathbb{P}}_h \hat{V}_{h+1}^k(s, a) \\ \hat{Q}_{H+1}^k(s, a) &= 0, \hat{Q}_h^k(s, a) \geq Q_h^*(s, a)\end{aligned}$$

Proof. We proof $\hat{Q}_h^k(s, a) \geq Q_h^*(s, a)$ by induction, when $h = H + 1$, $\hat{Q}_h^k(s, a) = Q_h^*(s, a) = 0$ $\square$

While $h \leq H$, we have

$$\mathbb{P}V \triangleq \sum_{s'} P(s')V(s')$$

$$\begin{aligned}\hat{Q}_h^k(s, a) - Q_h^*(s, a) &= b_h^k(s, a) + \hat{\mathbb{P}}_h \hat{V}_{h+1}^k(s, a) - \mathbb{P}_h V_{h+1}^* \\ &= b_h^k(s, a) + \left(\hat{\mathbb{P}}_h - \mathbb{P}_h \right) V_{h+1}^* + \hat{\mathbb{P}}_h (\hat{V}_{h+1}^k - V_{h+1}^*) \\ &\geq \hat{\mathbb{P}}_h (\hat{V}_{h+1}^k - V_{h+1}^*) \quad \text{by induction} \\ &\geq 0\end{aligned}$$

Employ UCB to MDPs

Principle: Estimate transitions and plan optimistically with UCB bonuses.

Algorithm 1 UCBVI (episodic, tabular)

Input: failure probability δ

- 1: Initialize $N_h^0(s, a) \leftarrow 0$ and $N_h^0(s, a, s') \leftarrow 0$ for all h, s, a, s' .
 - 2: **for** $k = 0, 1, \dots, K - 1$ **do**
 - 3: Form $\hat{\mathbb{P}}_h^k(\cdot \mid s, a)$ from counts (arbitrary if $N_h^k(s, a) = 0$).
 - 4: Set bonuses $b_h^k(s, a) \leftarrow 2H \sqrt{\frac{L}{N_h^k(s, a) + 1}}$ where $L = \ln(SAHK/\delta)$.
 - 5: Compute $(\hat{V}_h^k, \hat{Q}_h^k, \pi_h^k)_{h=0}^{H-1}$ by backward induction:
 $\hat{Q}_h^k(s, a) \leftarrow r_h(s, a) + b_h^k(s, a) + \hat{\mathbb{P}}_h^k \hat{V}_{h+1}^k(s, a)$
 $\pi_h^k(s) \leftarrow \arg \max_a \hat{Q}_h^k(s, a), \quad \hat{V}_h^k(s) \leftarrow \hat{Q}_h^k(s, \pi_h^k(s))$
 - 6: Execute π^k : $s_0^k = s_0$; **for** $h = 0, \dots, H - 1$ take $a_h^k = \pi_h^k(s_h^k)$ and sample $s_{h+1}^k \sim \mathbb{P}_h^*(\cdot \mid s_h^k, a_h^k)$.
 - 7: Update counts: $N_h^{k+1}(\cdot) \leftarrow N_h^k(\cdot)$ then increment the visited $(s_h^k, a_h^k, s_{h+1}^k)$.
 - 8: **end for**
-

Employ UCB to MDPs

Theorem 1. (*Upper Bound for UCBVI*) With $\delta \in (0, 1)$ and $L = \ln\left(\frac{2SAHK}{\delta}\right)$, then UCBVI guarantees with the probability at least $1 - \delta$

$$\text{REG} = \tilde{O}\left(H^2 S \sqrt{AKL}\right).$$

Theorem 2. (*Lower bound of MDP*) For any algorithm $\mathbb{A}$, there exist an MDP M with S states and A actions such that for any initial state $s \in \mathcal{S}$, the expected regret of $\mathbb{A}$ after $T = KH$ steps is

$$\mathbb{E}[\text{REG}] = \Omega\left(H \sqrt{SAT}\right).$$

Jaksch et al., Near-optimal Regret Bounds for Reinforcement Learning, JMLR, 2010.

Proof of regret bound

Theorem 1. (Upper Bound for UCBVI) With $\delta \in (0, 1)$ and $L = \ln\left(\frac{2SAHK}{\delta}\right)$, then UCBVI guarantees with the probability at least $1 - \delta$

$$\text{REG} = \tilde{\mathcal{O}}\left(H^2 S \sqrt{AKL}\right).$$

Proof.

$$\text{REG} = \sum_{k=1} V_1^*(s_{k,1}) - V_1^{\pi_k}(s_{k,h}) \leq \hat{V}_1^k(s_{k,1}) - V_1^{\pi_k}(s_{k,1})$$

Fix k, h , we have

$$\begin{aligned} \hat{V}_h^k(s_{k,h}) - V_h^{\pi_k}(s_{k,h}) &= b_h^k(s_{k,h}, a_{k,h}) + \hat{\mathbb{P}} \hat{V}_{h+1}^k(s_{k,h}, a_{k,h}) - \mathbb{P}_h^* V_{h+1}^{\pi_k}(s_{k,h}, a_{k,h}) \\ &= b_h^k(s, a) + \hat{\mathbb{P}}_h^*(\hat{V}_{h+1}^k - V_{h+1}^{\pi_k})(s_{k,h}, a_{k,h}) + (\hat{\mathbb{P}}_h^k - \mathbb{P}_h^*) \hat{V}_{h+1}^k(s_{k,h}, a_{k,h}) \end{aligned}$$

Proof of regret bound

Proof.

$$\begin{aligned}\hat{V}_h^k(s_{k,h}) - V_h^{\pi_k}(s_{k,h}) &= b_h^k(s_{k,h}, a_{k,h}) + \hat{\mathbb{P}}_h^*(\hat{V}_{h+1}^k - V_{h+1}^{\pi_k})(s_{k,h}, a_{k,h}) + (\hat{\mathbb{P}}_h^k - \mathbb{P}_h^*)\hat{V}_{h+1}^k(s_{k,h}, a_{k,h}) \\ &= b_h^k(s_{k,h}, a_{k,h}) + \mathbb{E} \left[(\hat{V}_{h+1}^k - V_{h+1}^{\pi_k})(s_{k,h}, a_{k,h}) \right] + (\hat{\mathbb{P}}_h^k - \mathbb{P}_h^*)\hat{V}_{h+1}^k(s_{k,h}, a_{k,h})\end{aligned}$$

So we have by induction

$$\hat{V}_1^k(s_{k,1}) - V_h^{\pi_k}(s_{k,1}) = \mathbb{E} \left[\sum_{h=1}^H \overset{\text{bonus term}}{b_h^k(s_{k,h}, a_{k,h})} + \overset{\text{transition estimation error}}{(\hat{\mathbb{P}}_h^k - \mathbb{P}_h^*)\hat{V}_{h+1}^k(s_{k,h}, a_{k,h})} \right].$$

where the expectation take on **random trajectory**.

Proof of regret bound

Theorem 3 (Concentration for discrete distributions). *Let Z be a discrete random variable supported on $\{1, \dots, d\}$ with distribution $q \in \Delta([d])$, and let $\hat{q}$ be the empirical distribution from N i.i.d. samples. Then with probability at least $1 - \delta$,*

$$\|\hat{q} - q\|_2 \leq \frac{1}{\sqrt{N}} + \sqrt{\frac{\log(1/\delta)}{2N}}.$$

Consequently,

$$\|\hat{q} - q\|_1 \leq \sqrt{d} \left(\frac{1}{\sqrt{N}} + \sqrt{\frac{\log(1/\delta)}{2N}} \right).$$

Proof of regret bound

Proof.

$$\hat{V}_1^k(s_{k,1}) - V_h^{\pi_k}(s_{k,1}) = \mathbb{E} \left[\sum_{h=1}^H b_h^k(s_{k,h}, a_{k,h}) + (\hat{\mathbb{P}}_h^k - \mathbb{P}_h^*) \hat{V}_{h+1}^k(s_{k,h}, a_{k,h}) \right].$$

Directly use with $\delta' = \frac{\delta}{2\mathcal{S}\mathcal{A}HK}$, we have with the probability $1 - \frac{\delta}{2}$ for all $(s, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$

union bound

$$\left\| \hat{P}_h^k(\cdot|s, a) - P_h^*(\cdot|s, a) \right\|_1 \leq \sqrt{\mathcal{S}} \left(\frac{1}{\sqrt{n}} + \sqrt{\frac{\ln(1/\delta')}{2n}} \right) \leq 2\sqrt{\frac{\mathcal{S} \ln(1/\delta')}{n}}$$

then we can bound transition estimation error on $\hat{V}$ by

$$(\hat{\mathbb{P}}_h^k - \mathbb{P}_h^*) \hat{V}_{h+1}^k(s_{k,h}, a_{k,h}) \leq \left\| \hat{\mathbb{P}}_h^k - \mathbb{P}_h^* \right\|_1 \left\| \hat{V}_{h+1}^k \right\|_\infty \leq 2H \sqrt{\frac{\mathcal{S}L}{N_h^k(s, a)}} \leq 4H \sqrt{\frac{\mathcal{S}L}{N_h^k(s, a) + 1}}.$$

Proof of regret bound

Proof.

$$\begin{aligned}\hat{V}_1^k(s_{k,1}) - V_h^{\pi_k}(s_{k,1}) &= \mathbb{E} \left[\sum_{h=1}^H b_h^k(s_{k,h}, a_{k,h}) + (\hat{\mathbb{P}}_h^k - \mathbb{P}_h^*) \hat{V}_{h+1}^k(s_{k,h}, a_{k,h}) \right] \\ &\leq \sum_{h=1}^H 2H \sqrt{\frac{L}{N_h^k(s_{k,h}, a_{k,h}) + 1}} + 4H \sqrt{\frac{SL}{N_h^k(s_{k,h}, a_{k,h}) + 1}} \\ &\leq 6H \sqrt{\frac{SL}{N_h^k(s_{k,h}, a_{k,h}) + 1}} \\ \text{REG} = \sum_{k=1}^K \hat{V}_1^k(s_{k,1}) - V_h^{\pi_k}(s_{k,1}) &\leq 6H \sqrt{SL} \sum_{k=1}^K \sum_{h=1}^H \sqrt{\frac{1}{N_h^k(s_{k,h}, a_{k,h}) + 1}}\end{aligned}$$

Proof of regret bound

Proof.

$$\begin{aligned}\text{REG} &\leq 6H\sqrt{SL} \sum_{k=1}^K \sum_{h=1}^H \sqrt{\frac{1}{N_h^k(s_{k,h}, a_{k,h}) + 1}} \\&\leq 6H\sqrt{SL} \sum_{h=1}^H \sum_{s,a} \sum_{i=1}^{N_h^K(s,a)+1} \sqrt{\frac{1}{i}} \leq 6H\sqrt{SL} \sum_{h=1}^H \sum_{s,a} \sqrt{N_h^K(s,a) + 1} \\&\leq 12H\sqrt{SL} \sum_{h=1}^H \sum_{s,a} \sqrt{N_h^K(s,a)} \leq 12H\sqrt{SL} \sum_{h=1}^H \sqrt{\mathcal{SA} \sum_{s,a} N_h^K(s,a)} \\&\leq 12H\sqrt{SL} \sum_{h=1}^H \sqrt{\mathcal{SA}K} = \tilde{\mathcal{O}}\left(H^2 S \sqrt{\mathcal{A}K}\right)\end{aligned}$$

relabeling summation index

Optimality Concern

Instead of bounding the error term $(\hat{\mathbb{P}}_h^k - \mathbb{P}_h)\hat{V}_{h+1}^k$ by $\left\|\hat{\mathbb{P}}_h^k - \mathbb{P}_h\right\|_1 \left\|\hat{V}_{h+1}^k\right\|_\infty$.

we decompose it into two terms: $(\hat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^*$ and $(\hat{\mathbb{P}}_h^k - \mathbb{P}_h)(\hat{V}_{h+1}^k - V_{h+1}^*)$.

Theorem 4. (Azar et al. (2017)) Consider a parameter $\delta \geq 0$. Then the regret of UCBVI is bounded w.p. at least $1 - \delta$, by

$$REG = \tilde{O}\left(H^2\sqrt{SAK}\right)$$

- Improved the regret bound by $\sqrt{S}$.
- For optimal regret, **estimate the transition variance** and build a weighted estimator; this gives a further $\sqrt{H}$ improvement.

Azar et al., Minimax Regret Bounds for Reinforcement Learning, ICML, 2017.

A Reflection

Theorem 2. (Lower bound of MDP) For any algorithm $\mathbb{A}$, there exist an MDP M with S states and A actions such that for any initial state $s \in \mathcal{S}$, the expected regret of $\mathbb{A}$ after $T = KH$ steps is

$$\mathbb{E}[\text{REG}] = \Omega \left(H\sqrt{SAT} \right).$$

- A trivial upper bound is KH since $r_h(s, a) \leq 1$.
- When $K \geq HSA$, the lower bound may no longer be informative.

Takeaway: The regret scales with both S and A . In real problems, large state and action spaces imply that **many learning episodes** may be required.

Jaksch et al., Near-optimal Regret Bounds for Reinforcement Learning, JMLR, 2010.



Part 1. Bandits and RL

- MDP
- Tabular MDP
- Function Approximation
- Policy Gradient

Function Approximation

Tabular MDPs: usually maintain a table to store values for all states (or state-action pairs), which **scales with state number S and action number A** .



Figure 1

We discover through experience that this state is bad



Figure 2

In tabular methods, we know nothing about this state.



Figure 3

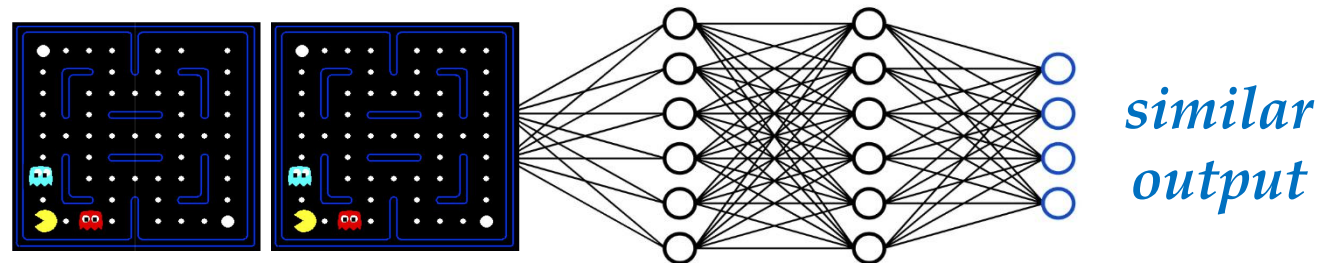
*We know **nothing** about this state either!*

But this has a poor scalability in practical scenarios; and many structures yet to exploit...

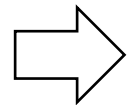
Function Approximation

RL Function approximation: approximate using a parameterized function.

- To avoid bad dependence on #states S , #action A in tabular MDPs
- Describe states (or state-actions) using feature representations in $\mathbb{R}^d$.
- A modern choice: DNN as a feature representer



parameterize MDP model with a low-dimensional representation



regret bound should not dependent on S or A , but rather the intrinsic dimension d

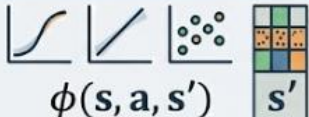
Linear MDP and Linear Mixture MDP

Random transition with

$$s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, a_h)$$

$\mathbb{P}_h(\cdot | s, a) = \langle \phi(s, a), \mu(\cdot) \rangle$ Linear MDP

$\mathbb{P}_h(\cdot | s, a) = \langle \psi(\cdot | s, a), \mathbf{w}_h^* \rangle$ Linear Mixture MDP

	LINEAR MDP	LINEAR MIXTURE MDP
IS KNOWLEDGE OF s' INFORMATION REQUIRED?	  NO. Only send features $\phi(s, a)$ to an “unknown” box. No need to analyze actual s'.	  YES. Features depend on s' explicitly, e.g., using $\phi(s, a, s')$.
IS IT EASIER TO DESCRIBE NONLINEAR RELATIONS?	  NO. Principally models linear relationships.	  YES. Easily describes nonlinear relations by mixing diverse feature components.

- ϕ and ψ are **known** function.
- μ and w are **unknown**.
- The difference is whether model next step to known function.
- We will use **Linear MDP**.

Linear Mixture MDP

- Linear Mixture MDPs

$$r_h(x, a) = \langle \phi(x, a), \theta_h^* \rangle$$

$$\mathbb{P}_h(s' | s, a) = \langle \psi(s' | s, a), \mathbf{w}_h^* \rangle$$

- $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ is known feature map
- $\psi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}^d$ is known feature map
- $\{\theta_h^*\}_{h=1}^H$ is the **unknown** reward parameter
- $\{\mathbf{w}_h^*\}_{h=1}^H$ is the **unknown** transition parameter

- Linear Bandits

(1) the player first chooses an arm X_t from arm set $\mathcal{X}$;

(2) and then environment reveals a reward $r_t \in \mathbb{R}$.

- Linear modeling assumption: $r_t(x) = x^\top \theta_* + \eta_t$

Linear bandits serve as a foundational tool for understanding linear mixture MDPs

Linear Mixture MDP

Using the parametrization of Linear Mixture MDPs where $\mathbb{P}(s' | s, a) = \langle \psi(s' | s, a), \theta_h^* \rangle$, we have the following

$$Q_h^\pi(s, a) = r_h(s, a) + \sum_{s'} \mathbb{P}(s' | s, a) V_{h+1}^\pi(s') = r_h(s, a) + \langle \sum_{s'} \psi(s' | s, a) V_{h+1}^\pi(s'), \theta_h^* \rangle$$

$$\hat{\mathbf{w}}_h^k = \arg \min_{\mathbf{w} \in \mathcal{W}} \lambda_{\mathbf{w}} \|\mathbf{w}\|_2^2 + \sum_{i=1}^{k-1} (V_{h+1}^i(s_{i,h+1}) - \langle \psi_{i,h}, \mathbf{w} \rangle)^2 \text{ where } \psi_{i,h} = \sum_{s'} \psi(s' | s_{i,h}, a_{i,h}) V_{h+1}^i(s')$$

V_{h+1}^i instead of V_{h+1}^k .
Thus can do one-pass and history-free update.

Computation cost will contains S without more assumption.

Zhou et al., Nearly Minimax Optimal RL for Linear Mixture MDPs, COLT, 2021.

Linear Mixture MDPs

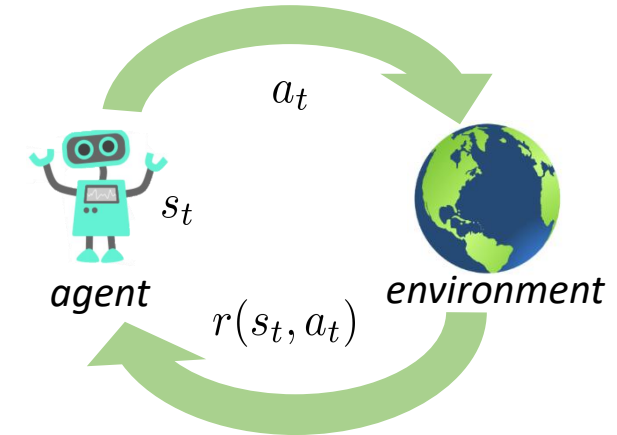
- Least square for parameter estimation

Reward estimation

$$\hat{\theta}_h = \arg \min_{\theta \in \mathbb{R}^d} \left\{ \frac{\lambda_\theta}{2} \|\theta\|_2^2 + \sum_{j=1}^{k-1} (r_h(s_h, a_h) - \phi(s_h, a_h)^\top \theta)^2 \right\}$$

Transition estimation

$$\hat{\mathbf{w}}_h = \arg \min_{\mathbf{w} \in \mathbb{R}^d} \left\{ \frac{\lambda_{\mathbf{w}}}{2} \|\mathbf{w}\|_2^2 + \sum_{j=1}^{k-1} (\langle \psi_{h+1}(s_h, a_h), \mathbf{w} \rangle - V_{h+1}(s_{h+1}))^2 \right\}$$



$$V_h^\pi(s) = \mathbb{E}_\pi \left[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}) \mid s_h = s \right]$$

Estimation error

$$\|\hat{\mathbf{w}}_h - \mathbf{w}_h\|_{\Sigma_h} \leq \mathcal{O} \left(\sqrt{dH} (\log(t/\delta))^2 \right)$$

Regret bound

$$\text{REG}_T \leq \tilde{\mathcal{O}} \left(d\sqrt{H^3 K} \right)$$

K : the number of episodes
 H : the length of each episode

Linear MDP and Tabular MDP

Linear MDP assume there exist known feature map

$$\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d, \|\phi(s, a)\|_2 \leq 1,$$

and unknown transition matrix

$$\mu_h^* \in \mathbb{R}^{|\mathcal{S}| \times d}.$$

rank of transition
matrix less than d .

Then

$$\mathbb{P}_h^*(\cdot \mid s, a) = \mu_h^* \phi(s, a)$$

If we set $\mu_h^*(s') = (\mu_h^*)_{s'}$ and $\mathbb{P}_h^*(s' \mid s, a) = (\mathbb{P}_h^*(\cdot \mid s, a))_{s'}$, then

$$\mathbb{P}_h^*(s' \mid s, a) = \mu_h^*(s') \phi(s, a)$$

To achieve tabular mdp, we set $d = |\mathcal{S}| |\mathcal{A}|$ and $\phi(s, a) = e_{(s, a)}$ (one-hot encoding), then $(\mu_h^*)_{(s', (s, a))} = \mathbb{P}_h^*(s' \mid s, a)$.

Linear Bellman Equation

Key property: For any bounded $f : \mathcal{S} \rightarrow \mathbb{R}$, $(\mathbb{P}_h^* f)(s, a)$ is **linear in $\phi(s, a)$**

Lemma 4. Fix h and bounded f . Let $w_h^* = (\mu_h^*)^\top f \in \mathbb{R}^d$. Then

$$\mathbb{E}_{s' \sim \mathbb{P}_h^*(\cdot | s, a)}[f(s')] = (\mathbb{P}_h^* f)(s, a) = \phi(s, a)^\top \mathbf{w}_h^*$$

If $\|f\|_\infty \leq H$, then $\|w_h^*\|_2 \leq H\sqrt{d}$.

Proof:

$$\begin{aligned} [\mathbb{P}_h^* f](s, a) &= \sum_{s'} f(s') \mathbb{P}_h^*(s' | s, a) = \sum_{s'} f(s') \mu_h^*(s')^\top \phi(s, a) \\ &= \left[\sum_{s'} f(s') \mu_h^*(s') \right]^\top \phi(s, a) = [(\mu_h^*)^\top f]^\top \phi(s, a) = \phi(s, a)^\top w_h^*. \end{aligned}$$

Planning with a Known Model

If μ_h^* is known, then all transition parameter can be compute. We can find the optimal policy by a **backward DP**.

DP version

For dynamic programming:

- (1) state: $f[h][s]$, the maximum expected future reward at state s and step h ;
- (2) transition: $f[h][s] = \max_a \{r_h(s, a) + \phi(s, a)^\top \mathbf{w}_h^*\}$;
- (3) computation: backward from $h = H$ to $h = 1$.

$$w_h^* = \sum_{s'} \mu_h^*(s') f[h+1][s']$$

We don't need to estimate **all** μ_h^* , we just need to estimate $\mathbf{w}_h^*$

Transition Estimator

Follow the backward dp, we estimate value function $\hat{V}$
From $h = H$ to 1 when k episode end.

$$w_h^* = \sum_{s'} \mu_h^*(s') \hat{V}_{h+1}$$

unknown part

known since backward

We use **history transition** $\{s_{i,h}, a_{i,h}, s_{i,h+1}\}_{i \in [k]}$ to estimate transition.

$$\hat{\mathbf{w}}_h = \arg \min_{\mathbf{w}} \lambda \|\mathbf{w}\|_2^2 + \sum_{i=1}^k \left(\hat{V}_{h+1}(s_{i,h+1}) - \sum_{s'} \langle \mathbf{w}, \phi(s_{i,h}, a_{i,h}) \rangle \right)^2.$$

$$\hat{\mathbf{w}}_h = \Lambda_{k,h}^{-1} \sum_{i=1}^k \phi(s_{i,h}, a_{i,h}) \hat{V}_{h+1}(s_{i,h+1})$$

$$\Lambda_{k,h} = \lambda I + \sum_{i=1}^k \phi(s_{i,h}, a_{i,h})^\top \phi(s_{i,h}, a_{i,h})$$

Estimation Error

Lemma 5 (Concentration Bound for Linear Bellman Function). *Let $\Lambda_{k,h} = \lambda \mathbf{I} + \sum_{i=1}^k \phi(s_{i,h}, a_{i,h}) \phi(s_{i,h}, a_{i,h})^\top$ be the covariance matrix at stage h . For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, the following inequality holds uniformly for all state-action pairs $(s, a) \in \mathcal{S} \times \mathcal{A}$ and all stages h :*

$$\left| \langle \hat{\mathbf{w}}_h^k, \phi(s, a) \rangle - \mathbb{P}_h^* \hat{V}_{h+1}^k(s, a) \right| \leq \beta \sqrt{\phi(s, a)^\top \Lambda_{k,h}^{-1} \phi(s, a)} \triangleq \beta \|\phi(s, a)\|_{\Lambda_{k,h}^{-1}},$$

where the scaling factor β satisfies:

$$\beta = \tilde{\mathcal{O}}(Hd).$$

In sub-gaussian noise, we assume $r(s, a) = \langle \phi(s, a), \theta \rangle + \eta$, and there we can consider $\mathbb{P}_h^k(s, a) = \langle w_h^k, \phi(s, a) \rangle + \eta$. So the analysis is similar.

Estimation Error analysis

Proof. Let $\mathbf{w}_h^k = \sum_{s'} \mu_h(s') \hat{V}_{h+1}^k(s')$, $\eta_h^k = \hat{V}_{h+1}^k(s_{i,h+1}) - \mathbb{P}_h^* \hat{V}_{h+1}^k(s_{i,h}, a_{i,h})$ and $\phi_{k,h} = \phi(s_{k,h}, a_{k,h})$, then

$$\langle \hat{\mathbf{w}}_h^k, \phi(s, a) \rangle - \mathbb{P}_h^* \hat{V}_{h+1}^k(s, a) = \phi(s, a)^\top (\hat{\mathbf{w}}_h^k - \mathbf{w}_h^k) \leq \|\phi(s, a)\|_{\Lambda_{k,h}^{-1}} \|\hat{\mathbf{w}}_h^k - \mathbf{w}_h^k\|_{\Lambda_{k,h}}$$

with $\hat{\mathbf{w}}_h^k = \Lambda_{k,h}^{-1} \sum_{i=1}^k \phi_{i,h} \hat{V}_{h+1}^k(s_{i,h+1})$, we have

$$\|\hat{\mathbf{w}}_h^k - \mathbf{w}_h^k\|_{\Lambda_{k,h}} = \left\| \Lambda_{k,h}^{-1} \sum_{i=1}^k \phi_{i,h} \hat{V}_{h+1}^k(s_{i,h+1}) - \mathbf{w}_h^k \right\|$$

Estimation Error analysis

Proof.

$$\|\hat{\mathbf{w}}_h^k - \mathbf{w}_h^k\|_{\Lambda_{k,h}} = \left\| \Lambda_{k,h}^{-1} \sum_{i=1}^k \phi_{i,h} \hat{V}_{h+1}^k(s_{i,h+1}) - \mathbf{w}_h^k \right\|$$

we have

$$\hat{V}_{h+1}^k(s_{i,h+1}) = \langle \phi_{i,h}, \mathbf{w}_h^k \rangle + \eta_h^k$$

Then

$$\|\hat{\mathbf{w}}_h^k - \mathbf{w}_h^k\|_{\Lambda_{k,h}} = \left\| \Lambda_{k,h}^{-1} \left[\sum_{i=1}^k \phi_{i,h} (\phi_{i,h}^\top \mathbf{w}_h^k + \eta_h^k) - \Lambda_{k,h} \mathbf{w}_h^k \right] \right\|_{\Lambda_{k,h}} \leq \left\| \sum_{i=1}^k \phi_{i,h} \eta_{i,h} \right\|_{\Lambda_{k,h}^{-1}} + \|\lambda \mathbf{w}_h^k\|_{\Lambda_{k,h}^{-1}} = \tilde{\mathcal{O}}(Hd)$$

Theorem 4 (Self-normalized concentration). *For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, for all $t \geq 0$,*

$$\left\| \sum_{s=1}^t \eta_s X_s \right\|_{V_t^{-1}}^2 \leq 2R^2 \log \left(\frac{\det(V_t)^{\frac{1}{2}} \det(V_0)^{-\frac{1}{2}}}{\delta} \right).$$



Construct UCB algorithm

- We add **estimation error as bonus** part and get

$$\begin{aligned}\hat{Q}_h^k(s, a) &= r_h(s, a) + \beta \sqrt{\phi(s, a)^\top \Lambda_{h,k} \phi(s, a)} + \mathbb{P}_h V_{h+1}^\pi(s, a) \\ \hat{V}_h^k(s) &= \hat{Q}_h^k(s, \pi_h^k(s)) \\ \pi_h^k(s) &= \arg \max_a \hat{Q}_h^k(s, a)\end{aligned}$$

Lin-UCBVI

For $k = 1$ to K

1. Compute all Q, V function and estimated $\hat{w}_h^k$ from $h = H$ to 1
2. Choose action a_h according to current state s_h and Q function.

Lin-UCBVI

Algorithm 2 Lin-UCBVI (episodic, linear MDP)

Input: regularization parameter λ , exploration bonus coefficient β

- 1: Initialize $\Lambda_{h,0} \leftarrow \lambda \mathbf{I}$ and $\mathbf{b}_{h,0} \leftarrow \mathbf{0}$ for all $h = 0, 1, \dots, H - 1$.
 - 2: **for** $k = 0, 1, \dots, K - 1$ **do**
 - 3: Compute $\hat{\mathbf{w}}_h^k \leftarrow \Lambda_{h,k}^{-1} \mathbf{b}_{h,k}$ for all $h = 0, 1, \dots, H - 1$ (set $\hat{\mathbf{w}}_h^k \leftarrow \mathbf{0}$ if $k = 0$).
 - 4: Compute $(\hat{V}_h^k, \hat{Q}_h^k, \pi_h^k)_{h=0}^{H-1}$ by backward induction: $\hat{V}_H^k(s) \leftarrow 0$, and for $h = H - 1, \dots, 0$:

$$\hat{Q}_h^k(s, a) \leftarrow \min \left\{ H, r_h(s, a) + \langle \hat{\mathbf{w}}_h^k, \phi(s, a) \rangle + \beta \sqrt{\phi(s, a)^\top \Lambda_{h,k}^{-1} \phi(s, a)} \right\}$$

$$\pi_h^k(s) \leftarrow \arg \max_a \hat{Q}_h^k(s, a), \quad \hat{V}_h^k(s) \leftarrow \hat{Q}_h^k(s, \pi_h^k(s))$$
 - 5: Execute π^k : $s_0^k = s_0$; for $h = 0, \dots, H - 1$ take $a_h^k = \pi_h^k(s_h^k)$ and sample $s_{h+1}^k \sim \mathbb{P}_h^\star(\cdot \mid s_h^k, a_h^k)$.
 - 6: Update design matrices: $\Lambda_{h,k+1} \leftarrow \Lambda_{h,k} + \phi(s_h^k, a_h^k) \phi(s_h^k, a_h^k)^\top$ for all h .
 - 7: Update target vectors: $\mathbf{b}_{h,k+1} \leftarrow \mathbf{b}_{h,k} + \phi(s_h^k, a_h^k) \hat{V}_{h+1}^k(s_{h+1}^k)$ for all h .
 - 8: **end for**
-

Regret analysis

Theorem 5. Fix $\delta \in (0, 1)$ and assume $r_h(s, a) \in [0, 1], \|\phi(s, a)\|_2 \leq 1$, then Lin-UCBVI with regularizer λ has the regret with the probability at least $1 - \delta$

$$\mathcal{O}(\beta H \sqrt{K d \log \left(1 + \frac{K}{d\lambda}\right)})$$

With our estimator, $\beta = \tilde{\mathcal{O}}(dH)$, then Jin et al., “Provably Efficient RL with Linear Function Approximation,” COLT, 2020.

$$REG = \tilde{\mathcal{O}}\left(H^2 \sqrt{d^3 K}\right)$$

If we use **weighted estimator** (Weighted Ridge Regression, Jin et al. (2020)) and get $\beta = \tilde{\mathcal{O}}(\sqrt{d})$, then we get the optimal regret.

$$REG = \tilde{\mathcal{O}}\left(dH \sqrt{K}\right)$$

Regret analysis

Proof. Optimism:

$$\begin{aligned}\hat{Q}_h^k(s, a) - Q_h^*(s, a) &= \beta \|\phi(s, a)\|_{\Lambda_{k,h}} + \hat{\mathbb{P}}_h \hat{V}_{h+1}^k(s, a) - \mathbb{P}_h V_{h+1}^* \\ &= \beta \|\phi(s, a)\|_{\Lambda_{k,h}} + \left(\hat{\mathbb{P}}_h - \mathbb{P}_h \right) \hat{V}_{h+1}^k + \mathbb{P}_h (\hat{V}_{h+1}^k - V_{h+1}^*) \\ &\geq \mathbb{P}_h (\hat{V}_{h+1}^k - V_{h+1}^*) \\ &\geq 0\end{aligned}$$

Looking back to Tabular MDP, Linear MDP **add more assumption** on Tabular MDP, so we can directly use

$$\hat{V}_1^k(s_{k,1}) - V_h^{\pi_k}(s_{k,1}) = \mathbb{E} \left[\sum_{h=1}^H \beta \|\phi_{k,h}\|_{\Lambda_{k,h}} + (\hat{\mathbb{P}}_h^k - \mathbb{P}_h^*) \hat{V}_{h+1}^k(s_{k,h}, a_{k,h}) \right].$$

where $\phi_{k,h} = \phi(s_{k,h}, a_{k,h})$.

Regret analysis

Proof.

$$\begin{aligned}\hat{V}_1^k(s_{k,1}) - V_h^{\pi_k}(s_{k,1}) &= \mathbb{E} \left[\sum_{h=1}^H \beta \|\phi_{k,h}\|_{\Lambda_{k,h}} + (\hat{\mathbb{P}}_h^k - \mathbb{P}_h^*) \hat{V}_{h+1}^k(s_{k,h}, a_{k,h}) \right] \\ &\leq \mathbb{E} \left[\sum_{h=1}^H 2\beta \|\phi_{k,h}\|_{\Lambda_{k,h}} \right]\end{aligned}$$

$$\text{REG} \leq \sum_{k=1}^K \hat{V}_1^k(s_{k,1}) - V_h^{\pi_k}(s_{k,1}) \leq 2\beta \sum_{k=1}^K \sum_{h=1}^H \|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}} \leq 2\beta \sum_{h=1}^H \sqrt{K \sum_{k=1}^K \|\phi_{k,h}\|_{\Lambda_{k,h}^{-1}}^2} = \mathcal{O} \left(\beta H \sqrt{K d \log \left(1 + \frac{K}{\lambda d} \right)} \right)$$

Lemma 4 (Elliptical Potential Lemma). *For any sequence $\{X_1, \dots, X_T\} \in \mathbb{R}^{d \times T}$, suppose $V_0 = \lambda I$, $V_t = V_{t-1} + X_t X_t^\top$, and $\|X_t\|_2 \leq L$, then*

$$\sum_{t=1}^T \|X_t\|_{V_t^{-1}}^2 \leq d \log \left(1 + \frac{L^2 T}{\lambda d} \right) \quad \text{proved in Lecture 6}$$

Efficient Concern

$$\begin{aligned}\hat{\mathbf{w}}_h^\pi &= \Lambda_{k,h}^{-1} \sum_{i=1}^k \phi(s_{i,h}, a_{i,h}) \hat{V}_{h+1}^\pi(s_{i,h+1}) \\ \Lambda_{k,h} &= \lambda I + \sum_{i=1}^k \phi(s_{i,h}, a_{i,h})^\top \phi(s_{i,h}, a_{i,h})\end{aligned}$$

Once policy π change, we need to again calculate all $\hat{V}_{h+1}^\pi(s_{i,h+1})$. So we must save all history data and cannot do one-pass update.

Lin-UCBVI algorithm has the total computation cost

$$\mathcal{O}(Hd^3 + HKd^2 + KHA d)$$

Computation contains action size A , which is always very large in real world. Thus strategic exploration is not a typical choice in current practical applications.(instead policy optimization algorithms)



Part 1. Bandits and RL

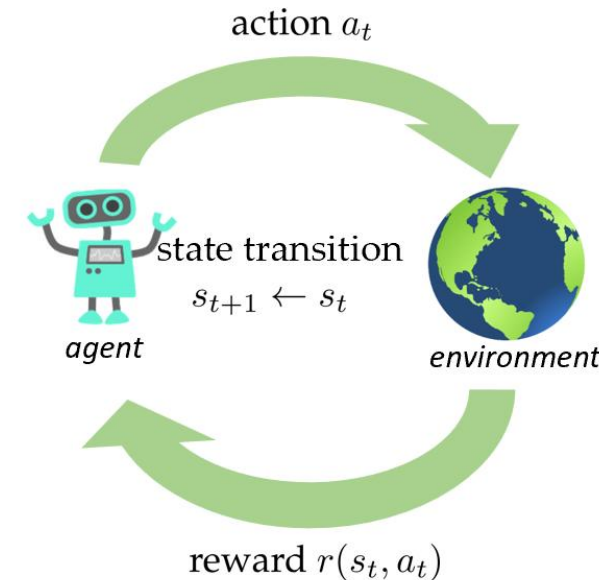
- MDP
- Tabular MDP
- Function Approximation
- Policy Gradient

Policy Optimization

- Consider infinite-horizon discounted MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho)$, where ρ is the initial state distribution and $\gamma \in (0, 1)$ is the discount factor.
- A policy is parameterized by θ : $\pi_\theta(a \mid s)$.

Initial state $s_0 \sim \rho$. At each round $t = 0, 1, 2, \dots$

- (1) agent take action $a_t \sim \pi_\theta(\cdot \mid s_t)$;
- (2) environment reveals a (deterministic) reward $r_t(s_t, a_t)$;
- (3) state transits to $s_{t+1} \sim P(\cdot \mid s_t, a_t)$.



- The goal is to maximize expected return $J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} [\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$, where a trajectory is $\tau = (s_0, a_0, s_1, a_1, \dots)$.

Vanilla Policy Gradient

- The goal is to maximize expected return $J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} [\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$, where a trajectory is $\tau = (s_0, a_0, s_1, a_1, \dots)$.
- In general, **policy optimization** means *directly updating* θ to increase $J(\theta)$.

A vanilla idea: do *gradient ascent* to $J(\theta)$ although the optimization problem is **non-convex** in many cases).

Vanilla Policy Gradient (REINFORCE)

At round $k = 1, 2, \dots$, update:

$$\theta_{k+1} = \theta_k + \eta \sum_t \nabla_\theta \log \pi_\theta(a_t | s_t) G_t, \quad \text{where } G_t = \sum_{k=t}^{\infty} \gamma^{k-t} r(s_k, a_k).$$

Vanilla Policy Gradient

$$\theta_{k+1} = \theta_k + \eta \sum_t \gamma^t \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) G_t, \quad \text{where } G_t = \sum_{k=t}^{\infty} \gamma^{k-t} r(s_k, a_k).$$

- Calculating the policy gradient estimator:

$$\begin{aligned} \nabla_{\theta} J(\theta) &= \nabla_{\theta} \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \\ &= \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\nabla_{\theta} \mathbb{P}^{\pi_{\theta}}[\tau] \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] && \text{(log-derivative trick: } \nabla \mathbb{P}[\tau] = \mathbb{P}[\tau] \nabla \log \mathbb{P}[\tau]) \\ &= \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_t \nabla_{\theta} \pi_{\theta}(a_t | s_t) \cdot \left(\sum_{k=0}^{\infty} \gamma^k r(s_k, a_k) \right) \right] && (\mathbb{P}^{\pi_{\theta}}[\tau] = \rho(s_0) \prod_t \pi_{\theta}(a_t | s_t) P(s_{t+1} | s_t, a_t)) \\ &= \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_t \gamma^t \nabla_{\theta} \pi_{\theta}(a_t | s_t) \cdot G_t \right]. && \text{(action } a_t \text{ only affects rewards-to-go } r(s_k, a_k) \text{ for } k \geq t) \end{aligned}$$

Baseline Policy Gradient

- Value function $Q^\pi(s, a) = \mathbb{E}_\pi [\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a]$.
- Discounted state occupancy $d^{\pi_\theta}(s) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}^{\pi_\theta}[s_t = s]$.

Theorem 1 (Policy Gradient Theorem).

$$\nabla_\theta J(\theta) = \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d^{\pi_\theta}, a \sim \pi_\theta(\cdot | s)} [\nabla_\theta \log \pi_\theta(a | s) Q^{\pi_\theta}(s, a)] .$$

- Usually we replace Q^π with the **advantage function** $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$, where $V^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot | s)} [Q^\pi(s, a)]$. Then,

$$\nabla_\theta J(\theta) = \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d^{\pi_\theta}, a \sim \pi_\theta(\cdot | s)} [\nabla_\theta \log \pi_\theta(a | s) A^{\pi_\theta}(s, a)] .$$

Baseline Policy Gradient

- **Vanilla policy gradient:** $\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau} \left[\sum_t \gamma^t \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \cdot G_t \right].$
- We can subtract any state-dependent baseline $b(s)$:

$$\begin{aligned} \mathbb{E}_{a \sim \pi_{\theta}(\cdot | s)} [\nabla_{\theta} \log \pi_{\theta}(a | s) b(s)] &= \sum_a \pi_{\theta}(a | s) \nabla_{\theta} \log \pi_{\theta}(a | s) b(s) \\ &= b(s) \sum_a \nabla_{\theta} \pi_{\theta}(a | s) = b(s) \nabla_{\theta} 1 = 0. \end{aligned}$$

So we can use $G_t - b(s_t)$ instead of G_t for any such baselines. The most common choice is $b(s) = V^{\pi}(s)$. Then $G_t - V^{\pi}(s_t) := \hat{A}_t$ estimates the **advantage**.

- **Advantage policy gradient:** $\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau} \left[\sum_t \gamma^t \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \cdot \hat{A}_t \right].$

Actor-Critic

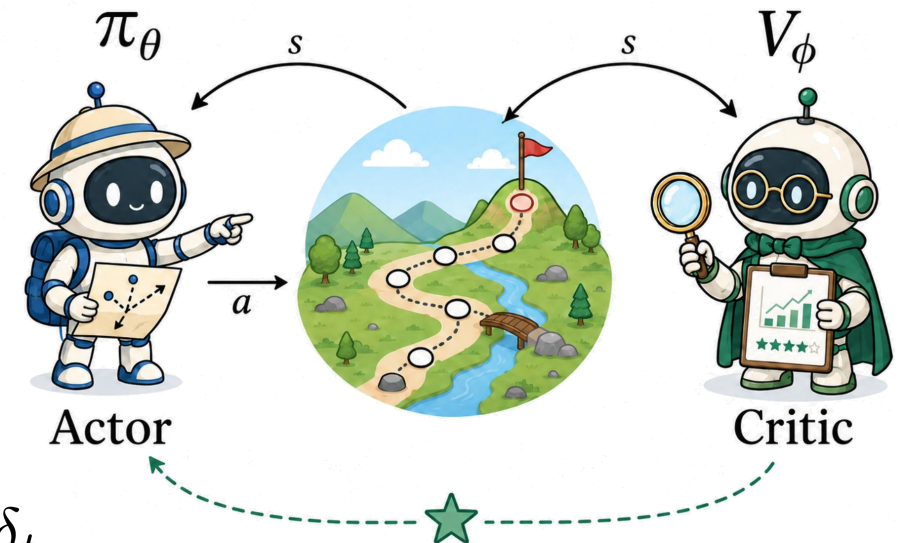
$$\theta_{k+1} = \theta_k + \eta \sum_t \gamma^t \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \hat{A}_t, \quad \text{where } \hat{A}_t = \sum_{k=t}^{\infty} \gamma^{k-t} r(s_k, a_k) - V^{\pi}(s_t).$$

- **Actor** (policy model): $\pi_{\theta}(a | s)$
- **Critic** (value model): $V_{\phi}(s)$

Example: estimator via *TD error*

$$\hat{A}_t \approx \delta_t := r_t + \gamma V_{\phi}(s_{t+1}) - V_{\phi}(s_t).$$

- **Actor update:** $\theta_{k+1} = \theta_k + \eta_{\theta} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \delta_t$
- **Critic update** minimizes $L_V(\phi) = (V_{\phi}(s_t) - (r_t + \gamma V_{\phi}(s_{t+1})))^2$.



Natural Policy Gradient

- **Vanilla policy gradient:** $\theta_{k+1} = \theta_k + \eta \nabla_{\theta} J(\theta_k)$.

Note that policies are probability distributions. A small change in θ can cause a large change in π_{θ} , or vice versa.

Optimizing over <i>parameter space</i>	Optimizing over <i>policy space</i>
$\theta_{k+1} = \theta_k + \eta \nabla_{\theta} J(\theta_k)$	?
$\begin{aligned} \max_{\Delta\theta} \quad & \nabla_{\theta} J(\theta)^{\top} \Delta\theta \\ \text{s.t.} \quad & \ \Delta\theta\ _2^2 \leq \delta. \end{aligned}$	$\begin{aligned} \max_{\Delta\theta} \quad & \nabla_{\theta} J(\theta)^{\top} \Delta\theta \\ \text{s.t.} \quad & D_{\text{KL}}(\pi_{\theta} \parallel \pi_{\theta+\Delta\theta}) \leq \delta. \end{aligned}$

where $D_{\text{KL}}(\pi_{\theta} \parallel \pi_{\theta+\Delta\theta}) = \mathbb{E}_{s \sim d^{\pi_{\theta}}} [D_{\text{KL}}(\pi_{\theta}(\cdot|s) \parallel \pi_{\theta+\Delta\theta}(\cdot|s))]$.

Natural Policy Gradient

**Constrained
form:**

$$\begin{aligned} \max_{\Delta\theta} \quad & \nabla_{\theta} J(\theta)^{\top} \Delta\theta \\ \text{s.t.} \quad & D_{\text{KL}}(\pi_{\theta} \parallel \pi_{\theta+\Delta\theta}) \leq \delta. \end{aligned}$$

**Mirror descent
form:**

$$\theta_{k+1} = \arg \max_{\theta'} \left\{ \langle \nabla_{\theta} J(\theta_k), \theta' - \theta_k \rangle - \frac{1}{\eta} D_{\text{KL}}(\pi_{\theta_k} \parallel \pi_{\theta'}) \right\}$$

- Second-order approximation of D_{KL} : $D_{\text{KL}}(\pi_{\theta} \parallel \pi_{\theta+\Delta\theta}) \approx \frac{1}{2} \Delta\theta^{\top} F(\theta) \Delta\theta$, where $F(\theta) = \mathbb{E}_{s \sim d^{\pi_{\theta}}, a \sim \pi_{\theta}(\cdot|s)} [\nabla_{\theta} \log \pi_{\theta}(a|s) \nabla_{\theta} \log \pi_{\theta}(a|s)^{\top}]$ is known as the **Fisher information matrix**.
- **Natural policy gradient**: $\theta_{k+1} = \theta_k + \eta F(\theta_k)^{\dagger} \nabla_{\theta} J(\theta_k)$.

Trust Region Policy Optimization (TRPO)



Constrained form of NPG:

$$\begin{aligned} \max_{\Delta\theta} \quad & \nabla_{\theta} J(\theta)^{\top} \Delta\theta \\ \text{s.t.} \quad & D_{\text{KL}}(\pi_{\theta} \parallel \pi_{\theta+\Delta\theta}) \leq \delta. \end{aligned}$$

- It's a KL-constrained linearized problem, since $J(\theta+\Delta\theta) \approx J(\theta) + \nabla_{\theta} J(\theta)^{\top} \Delta\theta$.
- TRPO maximizes the surrogate objective $L_{\theta_{\text{old}}}(\theta) = \mathbb{E}_{s,a \sim \pi_{\text{old}}} \left[r_{\theta}(s,a) \hat{A}^{\pi_{\theta_{\text{old}}}}(s,a) \right]$,
where $r_{\theta}(s,a) = \frac{\pi_{\theta}(s,a)}{\pi_{\theta_{\text{old}}}(s,a)}$ is the probability ratio.



TRPO:

$$\begin{aligned} \max_{\theta} \quad & L_{\theta_{\text{old}}}(\theta) \\ \text{s.t.} \quad & D_{\text{KL}}(\pi_{\theta_{\text{old}}} \parallel \pi_{\theta}) \leq \delta. \end{aligned}$$

(performance difference lemma:

$$J(\theta) = J(\theta_{\text{old}}) + \frac{1}{1-\gamma} \mathbb{E}_{s,a \sim \pi_{\theta}} [A^{\pi_{\theta_{\text{old}}}}(s,a)]$$

Proximal Policy Optimization (PPO)

TRPO:

$$\begin{aligned} \max_{\theta} \quad & \mathbb{E}_{s,a \sim \pi_{\text{old}}} \left[r_{\theta}(s, a) \hat{A}^{\pi_{\theta_{\text{old}}}}(s, a) \right] \\ \text{s.t.} \quad & D_{\text{KL}}(\pi_{\theta_{\text{old}}} \| \pi_{\theta}) \leq \delta. \end{aligned}$$

PPO-KL Penalty:

$$\max_{\theta} \quad \mathbb{E}_{s,a \sim \pi_{\text{old}}} \left[r_{\theta}(s, a) \hat{A}^{\pi_{\theta_{\text{old}}}}(s, a) - \beta D_{\text{KL}}(\pi_{\theta_{\text{old}}}(\cdot|s) \| \pi_{\theta}(\cdot|s)) \right].$$

PPO-Clip:

$$\max_{\theta} \quad \mathbb{E}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right].$$

- If $\hat{A}_t > 0$, then action a_t is better than expected (so we want to increase its probability). But PPO prevents $r_t(\theta) = \frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ from becoming too large.

Proximal Policy Optimization (PPO)

PPO-Clip:

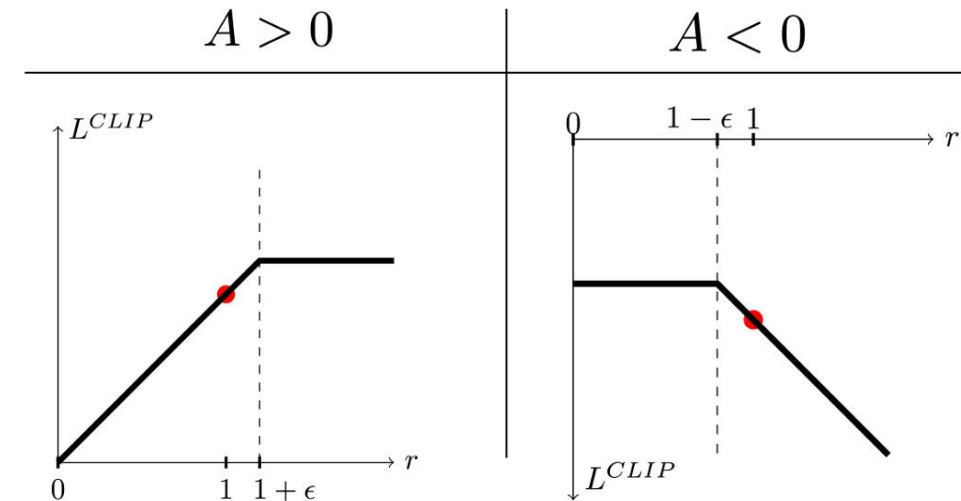
$$\max_{\theta} L^{\text{clip}}(\theta) := \mathbb{E}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right].$$

PPO in practice:

$$L^{\text{PPO}}(\theta, \phi) = L^{\text{clip}}(\theta) - c_1 L_V(\phi) + c_2 H(\pi_{\theta})$$

where

- value loss: $L_V(\phi) = \mathbb{E}_t \left[\left(V_{\phi}(s_t) - \hat{V}_t \right)^2 \right];$
- entropy bonus: $H(\pi_{\theta}) = \mathbb{E}_s \left[- \sum_a \pi_{\theta}(a|s) \log \pi_{\theta}(a|s) \right]$
- In practice, $\hat{A}_t$ is estimated via **generalized advantage estimation** (GAE).



Group Relative Policy Optimization

GRPO can be viewed as a *critic-free* variance of PPO:

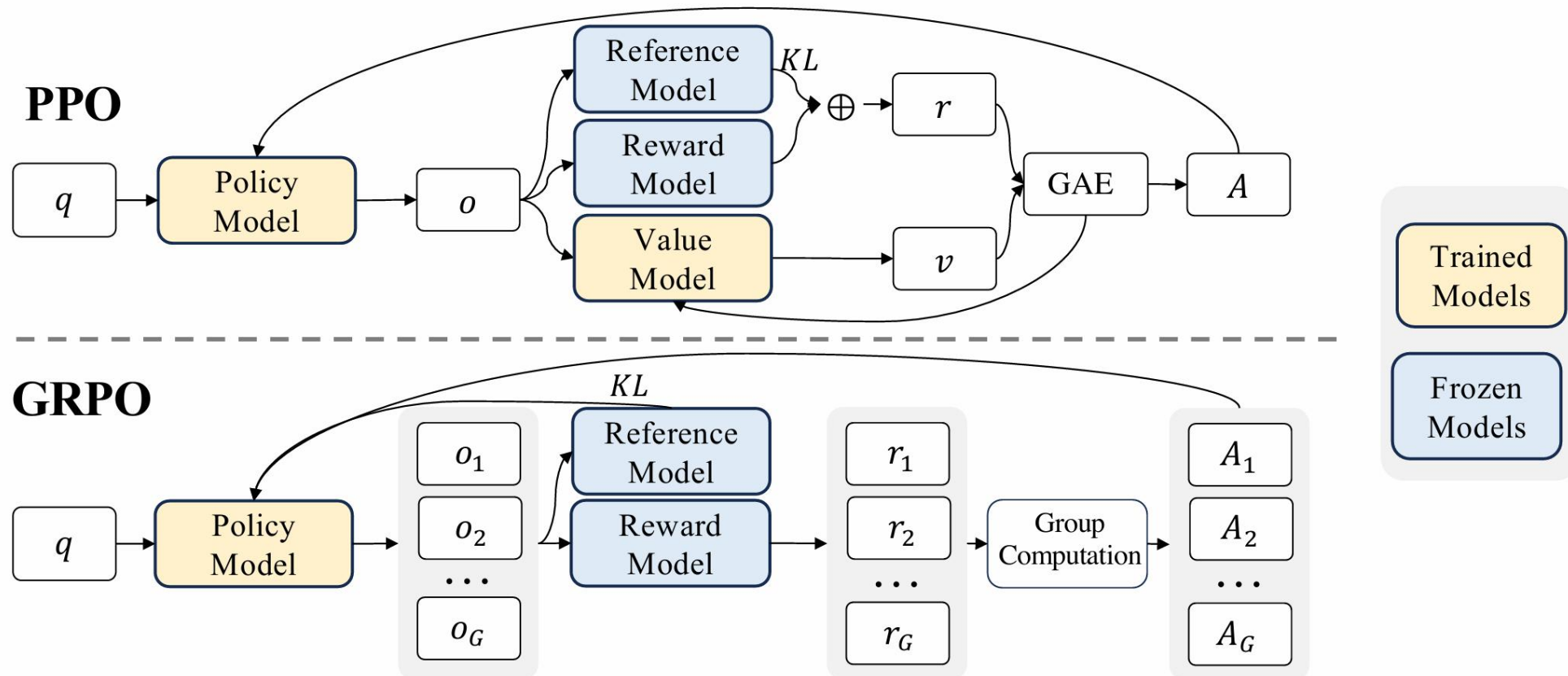
- For each prompt q , sample a group of G responses from the old policy: $\{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)$, where $o_i = (o_{i,1}, o_{i,2}, \dots, o_{i,|o_i|})$. Each response receives a sequence-level reward $R_i = R(q, o_i)$.
- Then GRPO defines the group-relative advantage as

$$\hat{A}_{i,t} = \hat{A}_i = \frac{R_i - \frac{1}{G} \sum_{j=1}^G R_j}{\sqrt{\frac{1}{G} \sum_{j=1}^G \left(R_j - \frac{1}{G} \sum_{\ell=1}^G R_\ell \right)^2}}.$$

- Accordingly, the token-level probability ratio is $r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t})}$.

Group Relative Policy Optimization

GRPO can be viewed as a *critic-free* variance of PPO:



A Good Reference for RL Theory

- Reinforcement Learning: Theory and Algorithms
- Agarwal, Brantley, Jiang, Kakade, and Sun
- In Progress

Reinforcement Learning: Theory and Algorithms

Alekh Agarwal Kianté Brantley Nan Jiang

Sham M. Kakade Wen Sun

June 27, 2026

<https://rltheorybook.github.io>

- ✓ II Strategic Exploration
 - Multi-Armed & Linear Bandits
 - Strategic Exploration in Tabular MDPs
 - Linearly Parameterized MDPs
 - Generalization with Bounded Bellman Rank
- ✓ III Policy Optimization
 - Policy Gradient Methods and Non-Convex Optimization
 - Optimality
 - Function Approximation and the Natural Policy Gradient
 - Conservative Policy Optimization

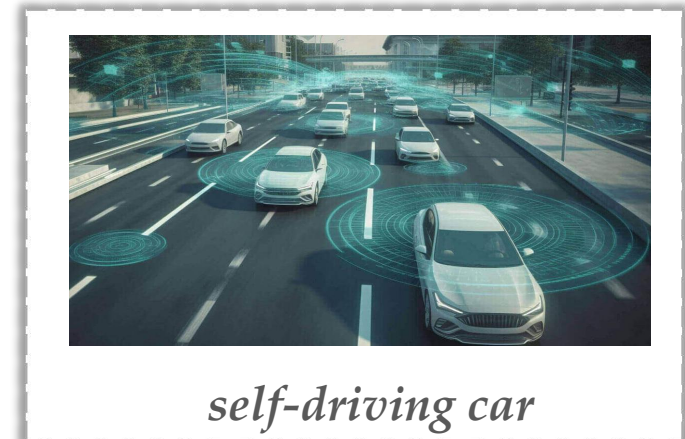
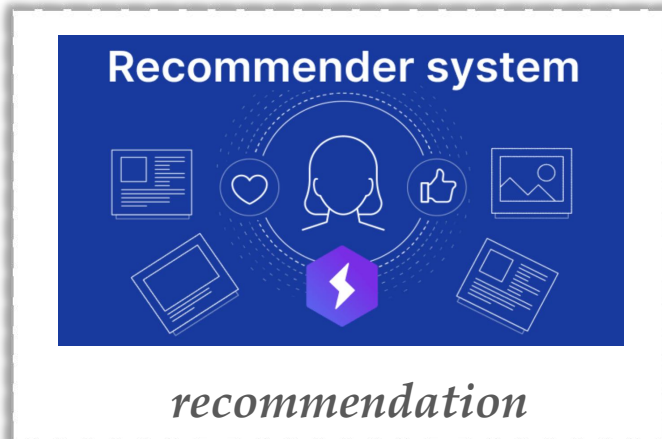


Part 2. One-Pass Bandits

- One-Pass Learning
- Generalized Linear Bandits
- Heavy-Tailed Linear Bandits
- Implication to RL

Interactive Machine Learning

- Many ML systems are now *interactive* with environment/user...



- Interactive Learning: *effective* and *efficient* over **long horizon**
 - online horizon T can be very large (millions/billions of rounds), or even unbounded
 - impossible to store all historical interactions (memory cost, privacy, etc).

One-Pass Learning

- “Re-fitting” fails at scale: a standard estimator at round t

regularized ERM/MLE: $\hat{\theta}_t \in \arg \min_{\theta \in \Theta} \sum_{s=1}^{t-1} \ell_s(\theta) + \lambda \Omega(\theta).$

- One-Pass Learning

Definition (One-Pass Estimator) An estimator $\hat{\theta}_t$ is called *one-pass* if there exists a state S_t such that:

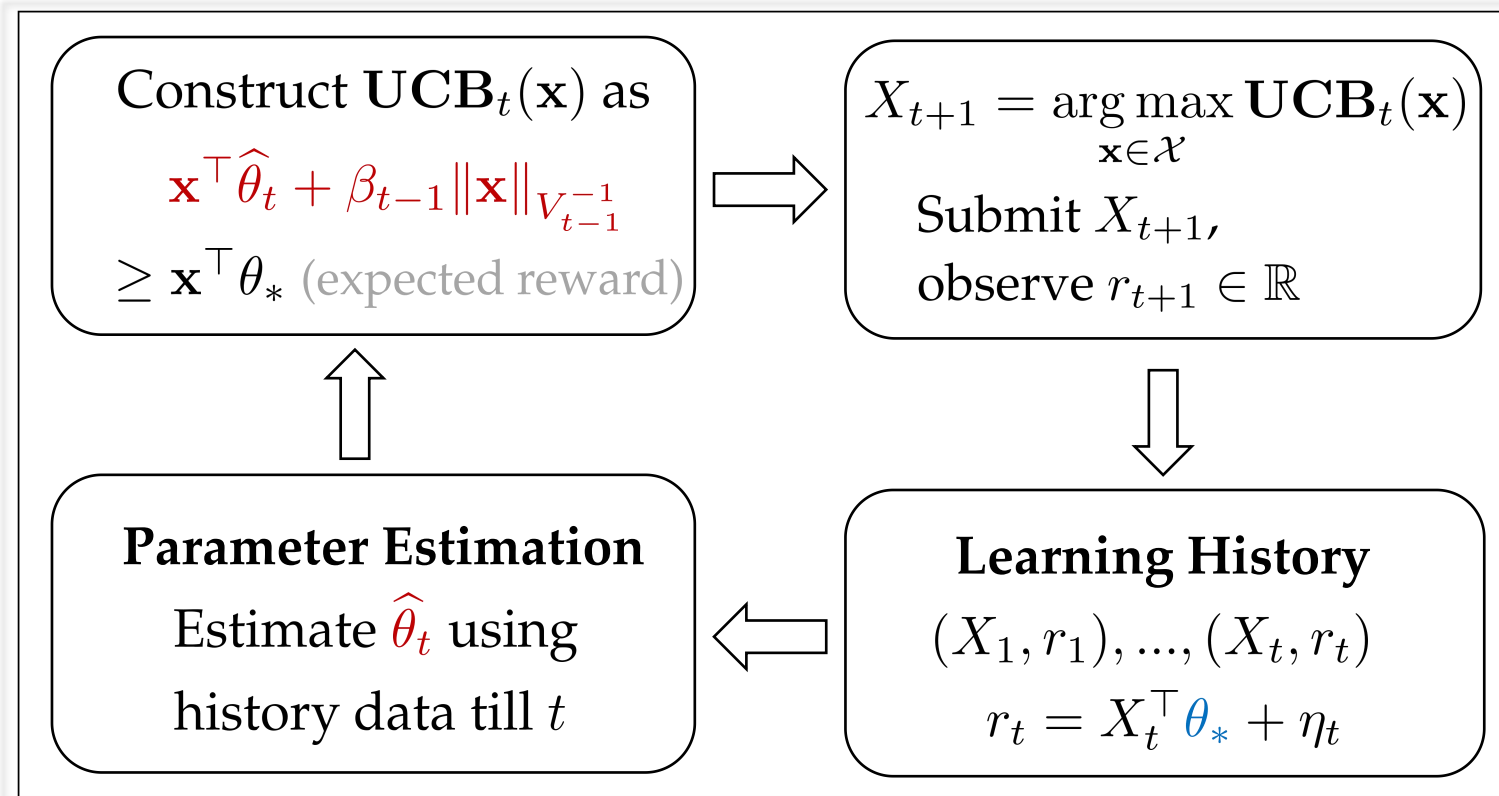
- S_t has a fixed size of $\text{poly}(d)$;
- $S_{t+1} = \text{OnlineUpdate}(S_t, \mathbf{z}_t)$, and then *discard* $\mathbf{z}_t$;
- $\hat{\theta}_t = \mathcal{G}(S_t)$, and both update and computation are *independent of t* .

- ✓ per-round time/storage memory: $\text{poly}(d)$, independent of t
- ✓ no historical data kept (esp. privacy issue..)

Bandits: Estimate-Action

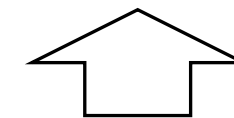
- **LinUCB** [Abbasi-Yadkori et al., NIPS 2011]

Parameter estimation + Upper Confidence Bound Selection; $\tilde{O}(d\sqrt{T})$ near-optimal regret



➤ **Action:** upper confidence bound

$$X_{t+1} = \arg \max_{\mathbf{x} \in \mathcal{X}} \left\{ \underbrace{\mathbf{x}^\top \hat{\theta}_t}_{\text{exploit}} + \underbrace{\beta_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}}}_{\text{explore}} \right\}$$



*estimation error
relates to exploration*

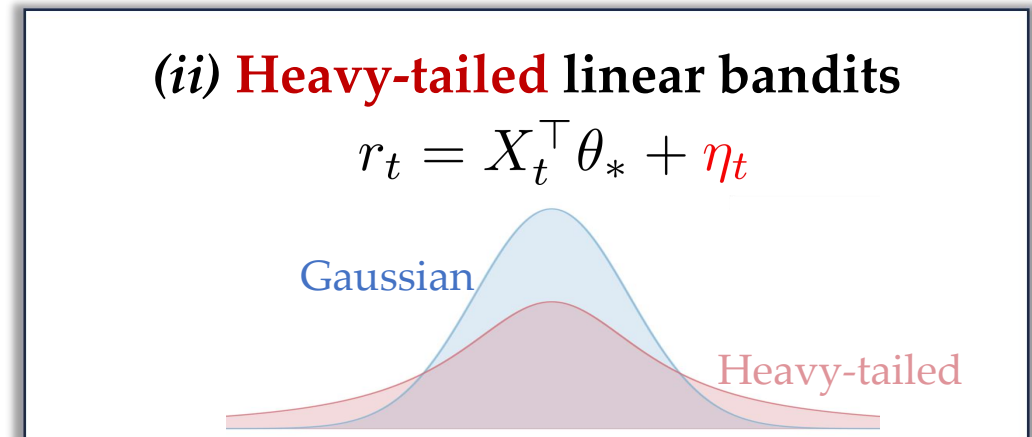
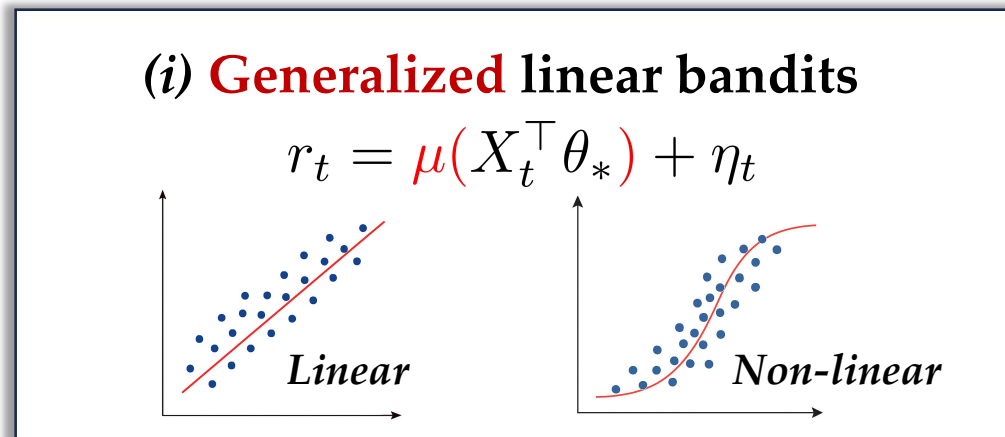
➤ **Estimator:** regularized least-square

$$\hat{\theta}_t = \arg \min_{\theta \in \mathbb{R}^d} \lambda \|\theta\|_2^2 + \sum_{s=1}^{t-1} (X_s^\top \theta - r_s)^2$$

LinUCB is Naturally One-Pass

- Linear Bandits: $\hat{\theta}_t = \arg \min_{\theta \in \mathbb{R}^d} \lambda \|\theta\|_2^2 + \sum_{s=1}^{t-1} (X_s^\top \theta - r_s)^2$ a special gift of quadratic loss  **GIFT**
- | | | |
|---|--|---|
| <p><i>sufficient statistics</i></p> $V_t = \lambda I + \sum_{s \leq t} X_s X_s^\top$ $b_t = \sum_{s \leq t} r_s X_s^\top$ | <p>✓ Online Update</p> $\hat{\theta}_{t+1} = V_t^{-1} b_t$ $V_t = V_{t-1} + X_t X_t^\top$ $b_t = b_{t-1} + r_t X_t^\top$ | <p>✓ Rank-1 Update: only $O(d^2)$ cost</p> $\hat{\theta}_{t+1} = \hat{\theta}_t + K_{t+1} (r_{t+1} - X_{t+1}^\top \hat{\theta}_t)$ $P_t = P_{t-1} - K_t X_t^\top P_{t-1},$ $K_t = P_{t-1} X_t \cdot (1 + X_t^\top P_{t-1} X_t)^{-1}$ |
|---|--|---|

- Beyond linear bandits: **More Expressivity**



① Generalized Linear Bandits

- Generalized linear reward modeling:

$$\mathbb{E}[r \mid x; \theta] = \mu(x^\top \theta)$$

- Linear: $\mu(z) = z$
- Logistic: $\mu(z) = 1/(1 + \exp(-z))$
- Poisson: $\mu(z) = \exp(z)$

Logistic model: binary feedback (click/not-click, prefer/not-prefer)

$$r_t \in \{0, 1\}, \quad \mathbb{P}[r_t = 1 \mid x_t] = \mu(x_t^\top \theta_*).$$

logistic (Bradley–Terry) modeling $\mu(z) = 1/(1 + \exp(-z))$

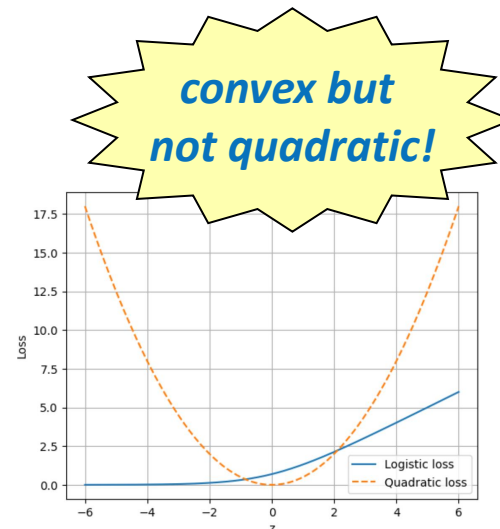
- GLM-UCB [Filippi et al., NIPS 2010]

- **Estimator:** regularized maximum likelihood estimator

$$\hat{\theta}_t = \arg \min_{\theta \in \Theta} \frac{\lambda}{2} \|\theta\|_2^2 + \sum_{s=1}^{t-1} \ell_s^{\text{GLB}}(\theta), \text{ with } \ell_s^{\text{GLB}}(\theta) = -\log \mathbb{P}_\theta[r_{s+1} \mid X_s]$$

$$\ell_s^{\text{logistic}}(\theta) = \log(1 + e^{x_s^\top \theta}) - r_{s+1} x_s^\top \theta$$

- **Action:** UCB $X_{t+1} = \arg \max_{\mathbf{x} \in \mathcal{X}} \left\{ \mu(\mathbf{x}^\top \hat{\theta}_t) + \beta_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}} \right\}$



② Heavy-Tailed Bandits

- Heavy-tailed reward modeling:

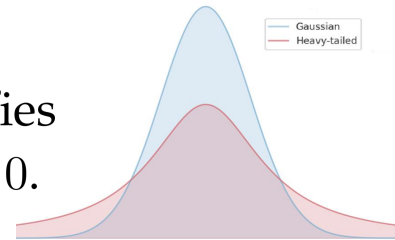
$$r_t = X_t^\top \theta_* + \eta_t$$

with heavy-tailed noise $\{\eta_t\}$

Esp. data contamination/large noise (e.g., in finance, sensor/IoT data).

Heavy-tailed noise, with $(1 + \varepsilon)$ -moment bounded

The noise $\{\eta_t\}$ is martingale difference, and satisfies that $\mathbb{E}[|\eta_t|^{1+\varepsilon} | \mathcal{F}_{t-1}] \leq \nu_t^{1+\varepsilon}$ for some $\varepsilon \in (0, 1]$, $\nu_t > 0$.



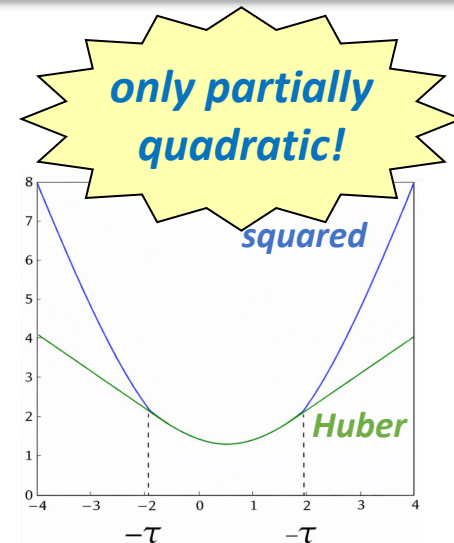
- Heavy-tail: only a low-order moment is bounded (even variance unstable)
- Sub-Gaussian: exponential tail decay $\rightarrow$ all moments are controlled

- HEAVY-OFUL [Huang et al., NeurIPS 2023]

- **Estimator:** adaptive Huber regression (reduce penalty for large deviation)

$$\hat{\theta}_t = \arg \min_{\theta \in \Theta} \frac{\lambda}{2} \|\theta\|_2^2 + \sum_{s=1}^{t-1} \ell_s^{\text{Hvt}}(\theta), \text{ with } \ell_s^{\text{Hvt}}(\theta) = \begin{cases} \frac{z_s(\theta)^2}{2} & \text{if } |z_s(\theta)| \leq \tau_s, \\ \tau_s |z_s(\theta)| - \frac{\tau_s^2}{2} & \text{if } |z_s(\theta)| > \tau_s, \end{cases}$$

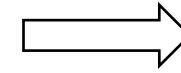
Huber loss is defined using a threshold $\tau_s > 0$, with $z_s(\theta) = \frac{r_s - X_s^\top \theta}{\sigma_s}$.



Efficiency Concerns

- **Stochastic LB:** least squares (closed-form solution)

$$\hat{\theta}_t = \arg \min_{\theta \in \mathbb{R}^d} \frac{\lambda}{2} \|\theta\|_2^2 + \sum_{s=1}^{t-1} (X_s^\top \theta - r_s)^2$$



one-pass update

$$\begin{aligned} \hat{\theta}_t &= V_{t-1}^{-1} \left(\sum_{s=1}^{t-1} r_s X_s \right) \\ V_{t-1} &= \lambda I + \sum_{s=1}^{t-1} X_s X_s^\top \end{aligned}$$

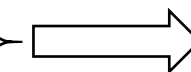
- **Generalized LB:** maximum likelihood estimator

$$\hat{\theta}_t = \arg \min_{\theta \in \Theta} \frac{\lambda}{2} \|\theta\|_2^2 + \sum_{s=1}^{t-1} \ell_s^{\text{GLB}}(\theta)$$

convex but non-quadratic loss

- **Heavy-tailed LB:** adaptive Huber regression

$$\hat{\theta}_t = \arg \min_{\theta \in \Theta} \frac{\lambda}{2} \|\theta\|_2^2 + \sum_{s=1}^t \ell_s^{\text{Hvt}}(\theta)$$



The cost at round t

Computational cost: $\mathcal{O}(t \log T)$

Storage cost: $\mathcal{O}(t)$

infeasible!

Question: Can Generalized/Heavy-tailed LB enjoy one-pass algorithms?

SGD for One-Pass Update?

Can we use SGD to incrementally solve MLE?

e.g., $\hat{\theta}_{t+1} = \hat{\theta}_t - \eta_t \nabla \ell_t(\hat{\theta}_t)$

Why plain SGD is not enough?

- 1) optimization error typically in the form of $\|\hat{\theta}_t - \theta_*\|_2^2$
- 2) i.i.d. over samples $\{(x_s, r_s)\}_{s \leq t}$

bandits need estimation for:

- 1) action selection needs a “*self-normalized*” V_t -exploration:

$$\left| \mathbf{x}^\top (\hat{\theta}_t - \theta_*) \right| \leq \beta_t \|\mathbf{x}\|_{V_t^{-1}} \quad \text{where } V_t = \lambda I + \sum_{s \leq t} X_s X_s^\top$$

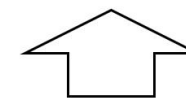
- 2) *on-policy* issue in bandits/RL

$$X_t = \pi_t(\hat{\theta}_1, \dots, \hat{\theta}_{t-1}), \quad \hat{\theta}_t = \text{Update}(\hat{\theta}_{t-1}, X_t)$$

no longer i.i.d., so classical guarantees are not applied

➤ **Action:** upper confidence bound

$$X_{t+1} = \arg \max_{\mathbf{x} \in \mathcal{X}} \left\{ \underbrace{\mathbf{x}^\top \hat{\theta}_t}_{\text{exploit}} + \underbrace{\beta_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}}}_{\text{explore}} \right\}$$



*estimation error
relates to exploration*

➤ **Estimator:** regularized least-square

$$\hat{\theta}_t = \arg \min_{\theta \in \mathbb{R}^d} \lambda \|\theta\|_2^2 + \sum_{s=1}^{t-1} (X_s^\top \theta - r_s)^2$$

Our Method: Online Mirror Descent

- OMD is a powerful framework for online regret optimization.

consider both *gradient update* and *geometry*

$$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \right\}$$

where $\mathcal{D}_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$ is the Bregman divergence.

	$\psi(\mathbf{x})$	$\mathcal{D}_\psi(\mathbf{x}, \mathbf{y})$
Squared L_2 -distance	$\ \mathbf{x}\ _2^2$	$\ \mathbf{x} - \mathbf{y}\ _2^2$
Mahalanobis distance	$\ \mathbf{x}\ _A^2$	$\ \mathbf{x} - \mathbf{y}\ _A^2$
Negative entropy	$\sum_i x_i \log x_i$	$\text{KL}(\mathbf{x} \parallel \mathbf{y})$

Online Mirror Descent

- Our previous mentioned algorithms can **all be covered** by OMD.

Algo.	OMD/proximal form	$\psi(\cdot)$	η_t	REG_T
OGD for convex	$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2} \ \mathbf{x} - \mathbf{x}_t\ _2^2$	$\frac{1}{2} \ \mathbf{x}\ _2^2$	$\frac{1}{\sqrt{t}}$	$\mathcal{O}(\sqrt{T})$
OGD for strongly c.	$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2} \ \mathbf{x} - \mathbf{x}_t\ _2^2$	$\frac{1}{2} \ \mathbf{x}\ _2^2$	$\frac{1}{\sigma t}$	$\mathcal{O}(\frac{1}{\sigma} \log T)$
ONS for exp-concave	$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \frac{1}{2} \ \mathbf{x} - \mathbf{x}_t\ _{A_t}^2$	$\frac{1}{2} \ \mathbf{x}\ _{A_t}^2$	$\frac{1}{\gamma}$	$\mathcal{O}(\frac{d}{\gamma} \log T)$
Hedge for PEA	$\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in \Delta_N} \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \text{KL}(\mathbf{x} \parallel \mathbf{x}_t)$	$\sum_{i=1}^N x_i \log x_i$	$\sqrt{\frac{\ln N}{T}}$	$\mathcal{O}(\sqrt{T \log N})$

Curvature-aware OMD Estimator

OMD with local norm (curvature-aware geometry matrix)

- local norm: typically choose $\psi_t(\mathbf{x}) = \frac{1}{2} \|\mathbf{x}\|_{H_t}^2$

$$\theta_{t+1} = \Pi_{\Theta}^{H_t} [\theta_t - \eta H_t^{-1} g_t], \quad g_t = \nabla \ell_t(\theta_t)$$

- Exploitation: using g_t
- Exploration: encoding H_t

- curvature geometry: using a **weighted** rank-1 increment to keep as a state

$$H_{t+1} = H_t + \alpha_t X_t X_t^\top \quad \text{require a careful design}$$

- complexity: independent of t

rank-1 updates $\Rightarrow$ time per round $O(d^2)$, memory $O(d^2)$

- ✓ **GLB**: use OMD and exploit self-concordance property to design the geometry matrix.
- ✓ **Hvt-LB**: use OMD and adaptively adjust Huber loss regions to set the weight.

“Self-Normalized” Error Analysis

- OMD: from regret optimization to statistical estimation.

Lemma 1 (Mirror Descent Lemma). *For $\forall \mathbf{u} \in \mathcal{X}$, the following inequality holds*

$$f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \leq \frac{1}{\eta_t} (\mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_t) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{t+1})) + \frac{\eta_t}{\lambda} \|\nabla f_t(\mathbf{x}_t)\|_\star^2 - \frac{1}{\eta_t} \mathcal{D}_\psi(\mathbf{x}_{t+1}, \mathbf{x}_t)$$

$$\theta_{t+1} = \arg \min_{\theta \in \Theta} \left\{ g_t(\theta) + \frac{1}{2\eta} \|\theta - \theta_t\|_{H_t}^2 \right\}$$

Lemma 2. *For OMD estimator, we have*

$$\frac{1}{2\eta} \|\theta_{t+1} - \theta_*\|_{H_t}^2 \leq \langle \nabla g_t(\theta_t), \theta_* - \theta_{t+1} \rangle + \frac{1}{2\eta} \|\theta_t - \theta_*\|_{H_t}^2 - \frac{1}{2\eta} \|\theta_{t+1} - \theta_t\|_{H_t}^2$$

“Self-Normalized” Error Analysis

- Standard regret analysis of OMD with twist yields

$$\theta_{t+1} = \arg \min_{\theta \in \Theta} \left\{ g_t(\theta) + \frac{1}{2\eta} \|\theta - \theta_t\|_{H_t}^2 \right\}$$

where $g_t(\theta)$ is the surrogate loss and H_t is the local norm.

Lemma 2. *For OMD estimator, we have*

$$\frac{1}{2\eta} \|\theta_{t+1} - \theta_*\|_{H_t}^2 \leq \langle \nabla g_t(\theta_t), \theta_* - \theta_{t+1} \rangle + \frac{1}{2\eta} \|\theta_t - \theta_*\|_{H_t}^2 - \frac{1}{2\eta} \|\theta_{t+1} - \theta_t\|_{H_t}^2$$

A proper choice of the local norm H_t and the surrogate loss $g_t(\theta)$ become highly crucial.

Importantly: compatible with self-normalized concentration
key to address *on-policy dependence* (details omitted...)

① Generalized Linear Bandits

- OMD-based estimator: *curvature-aware* local norm design

$$\theta_{t+1} = \arg \min_{\theta \in \Theta} \tilde{\ell}_t(\theta) + \frac{1}{2\eta} \|\theta - \theta_t\|_{H_t}^2,$$

$$\tilde{\ell}_t(\theta) \triangleq \langle \nabla \ell_t(\theta_t), \theta - \theta_t \rangle + \frac{1}{2} \|\theta - \theta_t\|_{\nabla^2 \ell_t(\theta_t)}^2$$

$$H_t \triangleq \lambda I_d + \sum_{s=1}^{t-1} \nabla^2 \ell_s(\theta_{s+1})$$

Computational Efficiency

$$\zeta_{t+1} = \theta_t - \eta \tilde{H}_t^{-1} \nabla \ell_t(\theta_t),$$

$$\theta_{t+1} = \arg \min_{\theta \in \Theta} \|\theta - \zeta_{t+1}\|_{\tilde{H}_t}^2.$$

$$\tilde{H}_t = H_t + \eta \nabla^2 \ell_t(\theta_t)$$

Technique: self-concordance property, second-order approximation, lookahead regularizer, etc.

Theorem 3 (Estimation Error). Let the regularization parameter $\lambda = 2 \max\{7d\eta R^2, \max\{3\eta RS, 1\}C_\mu/g(\tau)\}$ and the stepsize $\eta = 1 + RS$. Then, with probability at least $1 - \delta$, $\forall t > 1$, we have with

$$\|\theta_* - \theta_t\|_{H_t} \leq \beta_t(\delta) \triangleq \sqrt{4\lambda S^2 + 2\eta \ln\left(\frac{1}{\delta}\right) + 6d\eta^2 \ln\left(2 + \frac{2C_\mu t}{\lambda g(\tau)}\right)} = \mathcal{O}\left(SR \sqrt{d\left(S^2 R + \log \frac{t}{\delta}\right)}\right).$$

① Generalized Linear Bandits

Lemma 3. Let the stepsize $\eta = 1 + RS$ and the regularization parameter $\lambda \geq 6\eta C_\mu RS / g(\tau)$, then the online estimator satisfies

$$\|\theta_{t+1} - \theta_*\|_{H_{t+1}}^2 \leq (2 + 2RS) \left(\sum_{s=1}^t \ell_s(\theta_*) - \sum_{s=1}^t \ell_s(\theta_{s+1}) \right) - \frac{2}{3} \sum_{s=1}^t \|\theta_s - \theta_{s+1}\|_{H_s}^2 + 4\lambda S^2.$$

Proof.

$$\ell_s(\theta_{s+1}) - \ell_s(\theta_*) = \langle \nabla \ell_s(\theta_{s+1}), \theta_{s+1} - \theta_* \rangle - \|\theta_{s+1} - \theta_*\|_{\tilde{h}_s}^2 \quad \tilde{h}_s = \frac{\tilde{\alpha}(\theta_{s+1}, \theta_*, X_s)}{g(\tau)} X_s X_s^\top$$

$$\leq \langle \nabla \ell_s(\theta_{s+1}), \theta_{s+1} - \theta_* \rangle - \frac{1}{2 + 2RS} \|\theta_{s+1} - \theta_*\|_{\nabla^2 \ell_s(\theta_{s+1})}^2 \quad \tilde{h}_s \succcurlyeq \frac{1}{2 + 2RS} \nabla^2 \ell_s(\theta_{s+1})$$

$$\begin{aligned} \langle \nabla \ell_s(\theta_{s+1}), \theta_{s+1} - \theta_* \rangle &= \langle \nabla \ell_s(\theta_{s+1}) - \nabla \tilde{\ell}_s(\theta_{s+1}), \theta_{s+1} - \theta_* \rangle + \langle \nabla \tilde{\ell}_s(\theta_{s+1}), \theta_{s+1} - \theta_* \rangle \\ &\leq \frac{RSC_\mu}{g(\tau)} \|\theta_s - \theta_{s+1}\|_2^2 + \frac{1}{2\eta} \left(\|\theta_s - \theta_*\|_{H_s}^2 - \|\theta_{s+1} - \theta_*\|_{H_s}^2 - \|\theta_s - \theta_{s+1}\|_{H_s}^2 \right) \end{aligned}$$

Lemma 1

$$\sum_{s=1}^t \ell_s(\theta_{s+1}) - \ell_s(\theta_*) \leq \frac{1}{2 + 2RS} \left(\|\theta_1 - \theta_*\|_{H_1}^2 - \|\theta_{t+1} - \theta_*\|_{H_{t+1}}^2 - \sum_{s=1}^t \|\theta_s - \theta_{s+1}\|_{H_s}^2 \right) + \sum_{s=1}^t \frac{RSC_\mu}{g(\tau)} \|\theta_s - \theta_{s+1}\|_2^2$$

① Generalized Linear Bandits

Lemma 3. Let the stepsize $\eta = 1 + RS$ and the regularization parameter $\lambda \geq 6\eta C_\mu RS / g(\tau)$, then the online estimator satisfies

inverse regret

$$\|\theta_{t+1} - \theta_*\|_{H_{t+1}}^2 \leq (2 + 2RS) \left(\sum_{s=1}^t \ell_t(\theta_*) - \sum_{s=1}^t \ell_s(\theta_{s+1}) \right) - \frac{2}{3} \sum_{s=1}^t \|\theta_s - \theta_{s+1}\|_{H_s}^2 + 4\lambda S^2.$$

$$\sum_{s=1}^t \ell_t(\theta_*) - \sum_{s=1}^t \ell_s(\theta_{s+1}) = \underbrace{\sum_{s=1}^t \ell_s(\theta_*)}_{\text{term (a)}} - \underbrace{\sum_{s=1}^t m_s(P_s) + \sum_{s=1}^t m_s(P_s)}_{\text{mix loss}} - \underbrace{\sum_{s=1}^t \ell_s(\theta_{s+1})}_{\text{term (b)}}$$

$P_s = \mathcal{N}(\theta_s, \alpha H_s^{-1})$ as a d -dimensional multivariate Gaussian distribution

$m_s(P_s) = -\ln(\mathbb{E}_{\theta \sim P_s} [\exp(-\ell_s(\theta))])$ mapping distribution P_s to a real number value

① Generalized Linear Bandits

$$\sum_{s=t}^t \ell_s(\theta_*) - \sum_{s=1}^t \ell_s(\theta_{s+1}) = \underbrace{\sum_{s=1}^t \ell_s(\theta_*) - \sum_{s=1}^t m_s(P_s)}_{\text{term (a)}} + \underbrace{\sum_{s=1}^t m_s(P_s) - \sum_{s=1}^t \ell_s(\theta_{s+1})}_{\text{term (b)}}$$

Lemma 4. Let $\{\mathcal{F}_t\}_{t=1}^\infty$ be a filtration defined by $\mathcal{F}_t = \sigma\left(\{(X_s, r_s)\}_{s=1}^{t-1}\right)$. Let $\{P_t\}_{t=1}^\infty$ be a stochastic process such that the random variable P_t is a distribution over $\mathbb{R}^d$ and is $\mathcal{F}_t$ -measurable. Moreover, assume that the loss function ℓ_t is $\mathcal{F}_{t+1}$ -measurable. For any $t \geq 1$, define

$$L_t(\theta_*) = \sum_{s=1}^t \ell_s(\theta_*) \text{ and } F_t = -\sum_{s=1}^t \ln\left(\mathbb{E}_{\theta \sim P_s}\left[e^{-\ell_s(\theta)}\right]\right)$$

Then, for any $\delta \in (0, 1]$, we have

$$\Pr\left[\forall t \geq 1, L_t(\theta_*) \leq F_t + \log\left(\frac{1}{\delta}\right)\right] \geq 1 - \delta$$

Proof. $M_t = \exp(L_t(\theta_*) - F_t) \longrightarrow$ sequence $\{M_t\}_{t=1}^\infty$ is a non-negative (super)martingale

① Generalized Linear Bandits

Proof. $M_t = \exp(L_t(\theta_*) - F_t)$ $\longrightarrow$ sequence $\{M_t\}_{t=1}^\infty$ is a non-negative (super)martingale

$$M_t = \frac{\prod_{s=1}^t \mathbb{E}_{\theta \sim P_s} [\exp(-\ell_s(\theta))]}{\prod_{s=1}^t \exp(-\ell_s(\theta_*))} = M_{t-1} \cdot \frac{\mathbb{E}_{\theta \sim P_t} [\exp(-\ell_t(\theta))]}{\exp(-\ell_t(\theta_*))} = M_{t-1} \cdot \frac{\mathbb{E}_{\theta \sim P_t} [p(r_t | X_t^\top \theta)]}{p(r_t | X_t^\top \theta_*)}$$

$$\mathbb{E}[M_t | \mathcal{F}_t] = M_{t-1} \cdot \mathbb{E} \left[\frac{\mathbb{E}_{\theta \sim P_t} [p(r_t | X_t^\top \theta)]}{p(r_t | X_t^\top \theta_*)} \middle| \mathcal{F}_t \right] = M_{t-1}$$

Lemma 5. (Ville [1939]). Let $\{M_t\}_{t=0}^\infty$ be a supermartingale with $M_t \geq 0$ almost surely for all $t \geq 0$. Then, for any $\varepsilon > 0$,

$$\Pr \left[\sup_{t \in \mathbb{N}} M_t \geq \varepsilon \right] \leq \frac{\mathbb{E}[X_0]}{\varepsilon}$$

$$\Pr \left[\sup_{t \in \mathbb{N}} L_t(\theta_*) - F_t \geq \log \frac{1}{\delta} \right] \leq \delta M_0 = \delta$$

① Generalized Linear Bandits

$$\sum_{s=1}^t \ell_s(\theta_*) - \sum_{s=1}^t \ell_s(\theta_{s+1}) = \underbrace{\sum_{s=1}^t \ell_s(\theta_*) - \sum_{s=1}^t m_s(P_s)}_{\text{term (a)}} + \underbrace{\sum_{s=1}^t m_s(P_s) - \sum_{s=1}^t \ell_s(\theta_{s+1})}_{\text{term (b)}}$$

Lemma 6. Setting $\lambda \geq 64d\alpha R^2/7$ we have

$$\sum_{s=1}^t m_s(P_s) \leq \sum_{s=1}^t \ell_s(\theta_{s+1}) + \frac{1}{2\alpha} \sum_{s=1}^t \|\theta_{s+1} - \theta_s\|_{H_s}^2 + d \left(2\alpha + \frac{1}{2}\right) \ln \left(1 + \frac{2C_\mu t}{\lambda g(\tau)}\right)$$

Proof. $m_s(P_s) = -\log(\mathbb{E}_{\theta \sim P_s}[e^{-\ell_s(\theta)}]) = \underbrace{\mathbb{E}_{\theta \sim Q_s}[\ell_s(\theta)]}_{\text{term (b1)}} + \underbrace{\text{KL}(Q_s \| P_s)}_{\text{term (b2)}} - \text{KL}(Q_s \| P_s^*) \quad P_s^*(\theta) \propto P_s(\theta) \cdot e^{-\ell_s(\theta)}$

① Generalized Linear Bandits

Proof. $m_s(P_s) = -\log(\mathbb{E}_{\theta \sim P_s}[e^{-\ell_s(\theta)}]) = \underbrace{\mathbb{E}_{\theta \sim Q_s}[\ell_s(\theta)]}_{\text{term (b1)}} + \underbrace{\text{KL}(Q_s \| P_s)}_{\text{term (b2)}} - \text{KL}(Q_s \| P_s^*) \quad P_s^*(\theta) \propto P_s(\theta) \cdot e^{-\ell_s(\theta)}$

$$\begin{aligned} \text{term(b1)} &= \mathbb{E}_{\theta \sim Q_s}[\ell_s(\theta_{s+1}) + \langle \nabla \ell_s(\theta_{s+1}), \theta - \theta_{s+1} \rangle] + \mathbb{E}_{\theta \sim Q_s}[\mathcal{D}_{\ell_s}(\theta, \theta_{s+1})] \\ &= \ell_s(\theta_{s+1}) + \mathbb{E}_{\theta \sim Q_s}[\mathcal{D}_{\ell_s}(\theta, \theta_{s+1})] \\ &\leq \ell_s(\theta_{s+1}) + 2\alpha(\log \det(H_{s+1}) - \log \det(H_s)) \end{aligned} \quad \text{choose } Q_s = \mathcal{N}(\theta_{s+1}, \alpha H_{s+1}^{-1})$$

$$\begin{aligned} \text{term(b2)} &= \frac{1}{2} \left(\log \det(H_{s+1}) - \log \det(H_s) + \text{Tr}(H_s H_{s+1}^{-1}) + \frac{\|\theta_s - \theta_{s+1}\|_{H_s}^2}{\alpha} - d \right) \\ &\leq \frac{1}{2} (\log \det(H_{s+1}) - \log \det(H_s)) + \frac{1}{2\alpha} \|\theta_s - \theta_{s+1}\|_{H_s}^2 \end{aligned}$$

$$\sum_{s=1}^t m_s(P_s) \leq \sum_{s=1}^t \ell_s(\theta_{s+1}) + \frac{1}{2\alpha} \sum_{s=1}^t \|\theta_{s+1} - \theta_s\|_{H_s}^2 + \left(2\alpha + \frac{1}{2}\right) \ln \left(\frac{\det(H_{t+1})}{\det(\lambda I_d)} \right)$$

① Generalized Linear Bandits

GLM-UCB

$$\text{MLE } \hat{\theta}_{t+1} = \arg \min_{\theta \in \Theta} \frac{\lambda}{2} \|\theta\|_2^2 + \sum_{s=1}^t \ell_s(\theta)$$

Comp. cost per round $\mathcal{O}(t)$

Estimation error $\mathcal{O}(\kappa \sqrt{d \log t})$

GLB-OMD

$$\text{OMD } \hat{\theta}_{t+1} = \arg \min_{\theta \in \Theta} \tilde{\ell}_t(\theta) + \frac{1}{2\eta} \|\theta - \hat{\theta}_t\|_{H_t}^2$$

Comp. cost per round $\mathcal{O}(1)$

Estimation error $\mathcal{O}(\sqrt{d \log t})$

one-pass!

Theorem 4. With probability at least $1 - \delta$, the regret of GLB-OMD with parameter $\eta = 1 + RS$ and $\lambda = 2 \max\{7d\eta R^2, \max\{3\eta RS, 1\}C_\mu/g(\tau)\}$ ensures

$$\text{REG}_T \lesssim dSR\sqrt{S^2R + \log T} \sqrt{\frac{T \log T}{\kappa_*}} + \kappa d^2 S^2 R^3 \log T (S^2 R + \log T),$$

The first one-pass GLB algorithm with (almost) optimal regret guarantee!

 [Zhang-Xu-Z-Sugiyama, NeurIPS'25] Generalized Linear Bandits: Almost Optimal Regret with One-Pass Update.

① Generalized Linear Bandits

Proof.

$$\begin{aligned} \text{REG}_T &= \sum_{t=1}^T \mu(\mathbf{x}_{t,*}^\top \theta_*) - \sum_{t=1}^T \mu(X_t^\top \theta_*) \leq \sum_{t=1}^T \mu(X_t^\top \tilde{\theta}_t) - \sum_{t=1}^T \mu(X_t^\top \theta_*) \\ &= \underbrace{\sum_{t=1}^T \mu'(X_t^\top \theta_*) X_t^\top (\tilde{\theta}_t - \theta_*)}_{\text{term (a)}} + \underbrace{\frac{1}{2} \sum_{t=1}^T \tilde{\alpha}(\theta_*, \tilde{\theta}_t, X_t) \left(X_t^\top (\tilde{\theta}_t - \theta_*) \right)^2}_{\text{term (b)}} \end{aligned}$$

$$\text{term(a)} \leq 2\gamma_T^{(1)}(\delta) \sqrt{T/\kappa_* + R \cdot \text{REG}_T} + \gamma_T^{(2)}(\delta)$$

$$\text{term(b)} \leq \frac{2RC_\mu \beta_T^2(\delta)}{g(\tau)} \sum_{t=1}^T \|X_t\|_{V_t^{-1}}^2 \leq \gamma_T^{(2)}(\delta)$$

$$(X_t, \tilde{\theta}_t) = \arg \max_{\mathbf{x} \in \mathcal{X}_t, \theta \in \mathcal{C}_t(\delta)} \mu(\mathbf{x}^\top \theta)$$

$$\tilde{\alpha}(\theta_1, \theta_2, X_s) = \int_0^1 (1-v) \mu''(X_s^\top \theta_1 + v X_s^\top (\theta_2 - \theta_1)) dv$$

$$\begin{cases} \gamma_T^{(1)}(\delta) = \frac{2\beta_T(\delta) \sqrt{2d(1+C_\mu)}}{\sqrt{g(\tau)}} \sqrt{\log \left(1 + \frac{TC_\mu}{g(\tau)d\lambda} \right)} = \mathcal{O}(\beta_T(\delta) \sqrt{d \log T}) \\ \gamma_T^{(2)}(\delta) = \frac{8\kappa d R C_\mu \beta_{T+1}^2(\delta)}{g(\tau)} \log \left(1 + \frac{T}{g(\tau)\kappa d\lambda} \right) = \mathcal{O}(\beta_{T+1}(\delta)^2 \kappa d R \log T) \end{cases}$$

$$\text{REG}_T \leq 2\gamma_T^{(1)}(\delta) \sqrt{T/\kappa_* + R \cdot \text{REG}_T} + 2\gamma_T^{(2)}(\delta) \leq \mathcal{O} \left(dSR \sqrt{S^2 R + \log T} \sqrt{\frac{T \log T}{\kappa_*}} + \kappa d^2 S^2 R^3 \log T (S^2 R + \log T) \right)$$

② Heavy-Tailed Bandits

- OMD-based estimator: *curvature-aware* local norm design

$$\hat{\theta}_{t+1} = \arg \min_{\theta \in \Theta} \left\{ \left\langle \theta, \nabla \ell_t(\hat{\theta}_t) \right\rangle + \mathcal{D}_{\psi_t}(\theta, \hat{\theta}_t) \right\}$$

$$\psi_t(\theta) = \frac{1}{2} \|\theta\|_{V_t}^2 \text{ with } V_t \triangleq \lambda I + \frac{1}{\alpha} \sum_{s=1}^t \frac{X_s X_s^\top}{\sigma_s^2}$$

Computational Efficiency

$$\tilde{\theta}_{t+1} = \hat{\theta}_t - V_t^{-1} \nabla \ell_t(\hat{\theta}_t)$$

$$\hat{\theta}_{t+1} = \arg \min_{\theta \in \Theta} \left\| \theta - \tilde{\theta}_{t+1} \right\|_{V_t}$$

Technique: adaptively adjust the threshold/renormalized factor in Huber loss, exploit curvature of in/out-liers

Theorem 5 (Estimation error). If σ_t, τ_t, τ_0 are set as where $w_t \triangleq \frac{1}{\sqrt{\alpha}} \left\| \frac{X_t}{\sigma_t} \right\|_{V_{t-1}^{-1}}$ and let the step size $\alpha = 4$, then with probability at least $1 - 4\delta, \forall t \geq 1$, we have

$$\|\hat{\theta}_{t+1} - \theta_*\|_{V_t} \leq \beta_t \triangleq 107 \log \frac{2T^2}{\delta} \tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} + \sqrt{\lambda(2 + 4S^2)}$$

② Heavy-Tailed Bandits

HEAVY-OFUL

$$\text{MLE } \hat{\theta}_{t+1} = \arg \min_{\theta \in \Theta} \frac{\lambda}{2} \|\theta\|_2^2 + \sum_{s=1}^t \ell_s(\theta)$$

Comp. cost per round $\mathcal{O}(t)$

Estimation error $\tilde{\mathcal{O}} \left(t^{\frac{1-\epsilon}{2(1+\epsilon)}} \right)$

Hvt-UCB

$$\text{OMD } \hat{\theta}_{t+1} = \arg \min_{\theta \in \Theta} \left\{ \langle \theta, \nabla \ell_t(\hat{\theta}_t) \rangle + \frac{1}{2} \|\theta - \hat{\theta}_t\|_{V_t}^2 \right\}$$

Comp. cost per round $\mathcal{O}(1)$

Estimation error $\tilde{\mathcal{O}} \left(t^{\frac{1-\epsilon}{2(1+\epsilon)}} \right)$

one-pass!

Theorem 6. By setting $\sigma_t, \tau_t, \tau_0, \alpha$ as in Lemma 1, and let $\lambda = d, \sigma_{\min} = \frac{1}{\sqrt{T}}, \delta = \frac{1}{8T}$, with probability at least $1 - 1/T$, the regret of Hvt-UCB is bounded by

$$\text{REG}_T \leq \tilde{\mathcal{O}} \left(dT^{\frac{1-\epsilon}{2(1+\epsilon)}} \sqrt{\sum_{t=1}^T \nu_t^2} + dT^{\frac{1-\epsilon}{2(1+\epsilon)}} \right).$$

When $\nu_t = \nu$, this can recover to optimal regret bound $\text{REG}_T \leq \tilde{\mathcal{O}} \left(dT^{\frac{1}{1+\epsilon}} \right)$

The first one-pass algorithm for heavy-tailed linear bandits with (almost) optimal regret!



Part 3. RL Implications

- Function Approximation
- RLHF

① Function Approximation

□ Linear Function Approximation

- Linear mixture MDPs [Ayoub et al., 2020]: $\mathbb{P}_h(s'|s, a) = \phi(s'|s, a)^\top \theta_h^*$
- Linear / low-rank MDPs [Jin et al., 2020]: $\mathbb{P}_h(s'|s, a) = \phi(s, a)^\top \mu^*(s'), r_h(s, a) = \phi(s, a)^\top \theta_h^*$
- ...

linearity is hard to satisfy in practice!

Technically, this "**linear**" MDP parametrization arises because it can be reduced to and solved by **stochastic linear bandits**, which is well-understood.

□ General Function Approximation

- Eluder dimension [Russo and Roy, 2013, Jin et al., 2021]
- Decision-Estimation Coefficient (DEC) [Foster et al., 2021]
- Admissible Bellman Characterization (ABC) [Chen et al., 2023]
- ...

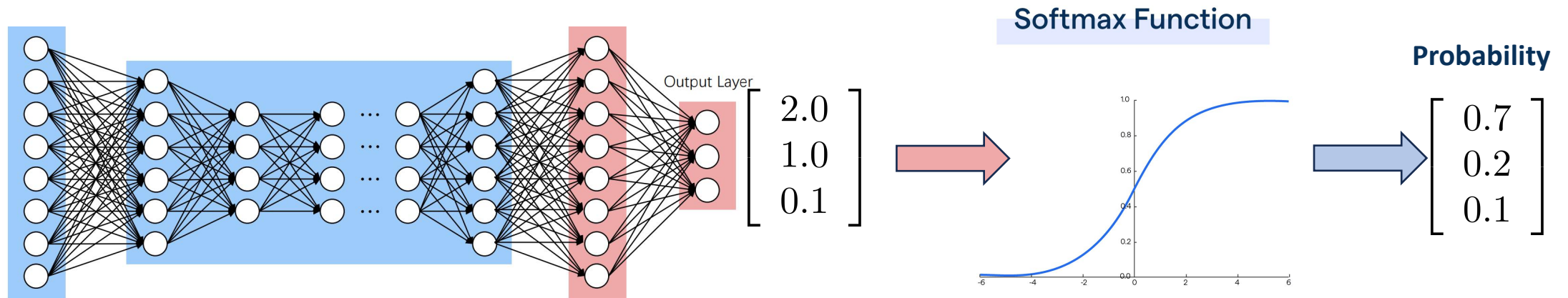
usually no computationally efficient algorithms provided

computationally efficient beyond linearity?

MNL Function Approximation

□ A new class: **Multinomial Logit (MNL)** function approximation

[Hwang and Oh, 2023]



MNL mixture MDPs:

$$\mathbb{P}_h(s' | s, a) = \frac{\exp(\phi(s' | s, a)^\top \theta_h^*)}{\sum_{\tilde{s} \in \mathcal{S}_{h,s,a}} \exp(\phi(\tilde{s} | s, a)^\top \theta_h^*)}$$

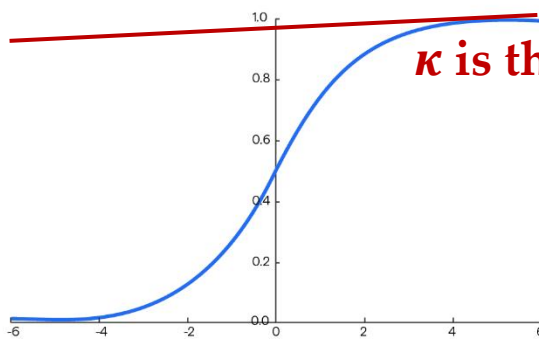
- $\phi(s' | s, a)$ is the known feature mapping
- $\{\theta_h^*\}_{h=1}^H$ is the **unknown** transition parameter
- $\mathcal{S}_{h,s,a} \triangleq \{s' \in \mathcal{S} \mid \mathbb{P}_h(s' | s, a) \neq 0\}$ is reachable states

Key Challenge: non-linearity

Linear mixture MDPs: $\mathbb{P}_h(s'|s, a) = \phi(s'|s, a)^\top \theta_h^*$

MNL mixture MDPs: $\mathbb{P}_h(s' | s, a) = \frac{\exp(\phi(s' | s, a)^\top \theta_h^*)}{\sum_{\tilde{s} \in \mathcal{S}_{h,s,a}} \exp(\phi(\tilde{s} | s, a)^\top \theta_h^*)}$

Softmax Function



κ is the minimum slope

*even two vastly different inputs
will have much similar outputs*

Regularity assumption:

$$\inf_{\theta \in \Theta} p_{s,a}^{s'}(\theta) p_{s,a}^{s''}(\theta) \geq \kappa$$

$$\text{where } p_{s,a}^{s'}(\theta) = \frac{\exp(\phi(s'|s,a)^\top \theta)}{\sum_{\tilde{s} \in \mathcal{S}_{s,a}} \exp(\phi(\tilde{s}|s,a)^\top \theta)}$$

Define $U = \max_{(h,s,a)} S_{h,s,a} \Rightarrow \kappa \leq 1/U^2$.

in the worst case, $\kappa^{-1} = \Omega(S^2)$

MNL Mixture MDPs

- OMD for one-pass estimation

$$\tilde{\theta}_{k+1,h} = \arg \min_{\theta \in \Theta} \left\{ \langle \nabla \ell_{k,h}(\tilde{\theta}_{k,h}), \theta \rangle + \frac{1}{2\eta} \|\theta - \tilde{\theta}_{k,h}\|_{\tilde{\mathcal{H}}_{k,h}}^2 \right\},$$

one-pass!

where $\tilde{\mathcal{H}}_{k,h} = \eta H_{k,h}(\tilde{\theta}_{k,h}) + \sum_{i=1}^{k-1} H_{i,h}(\tilde{\theta}_{i+1,h})$ incorporates additional second-order quantity.

Reference	Model	Upper Bound	Lower Bound
Zhou et al. [2021]	Linear mixture MDP	$\tilde{\mathcal{O}}(dH^{3/2}\sqrt{K})$	$\Omega(dH^{3/2}\sqrt{K})$
Hwang and Oh [2023]	MNL mixture MDP	$\tilde{\mathcal{O}}(\kappa^{-1}dH^2\sqrt{K})$	—
Our work	MNL mixture MDP	$\tilde{\mathcal{O}}(dH^2\sqrt{K} + \kappa^{-1}d^2H^2)$	$\Omega(dH\sqrt{K})$

*in the worst case,
 $\kappa^{-1} = \Omega(S^2)$*

Match the results for linear mixture MDPs except for the dependence on H .

② RLHF

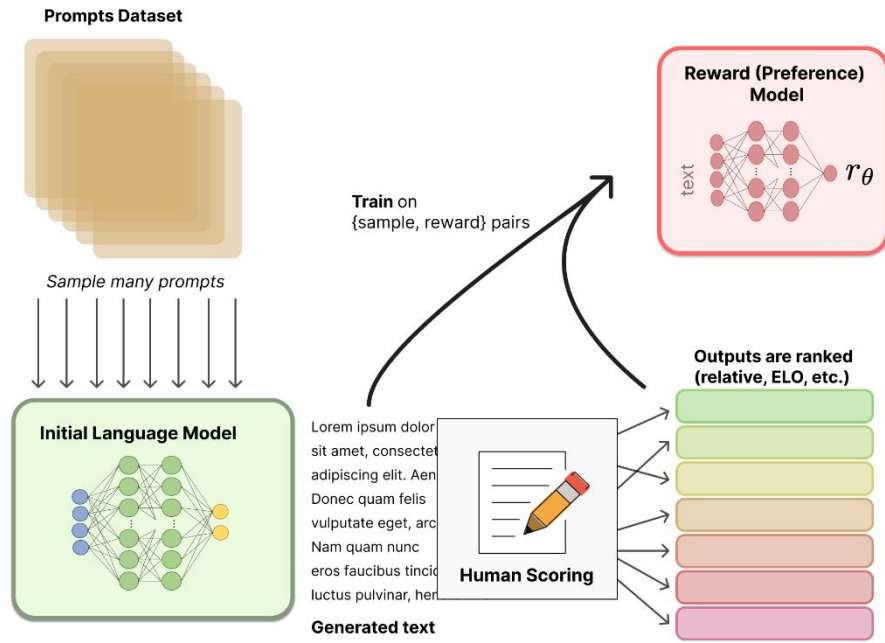
RLHF (or *preference OPT*): align model towards human preferences or values.

- **Input:** a 4-argument preference tuple (x, a, a', y)
 - x : the prompt: `"Please write a joke for me."`
 - a : the first response: `"Sorry, I can't."`
 - a' : the second response: `"Here is a joke for you: ..."`
 - $y \in \{0, 1\}$: the label (human's preference): `$a'$`
- RLHF wants to use input to improve LLM
i.e., *align LLM with human's preference or value (encoded in the preference data)*
- **Output:** a fine-tuned LLM with better aligned preference

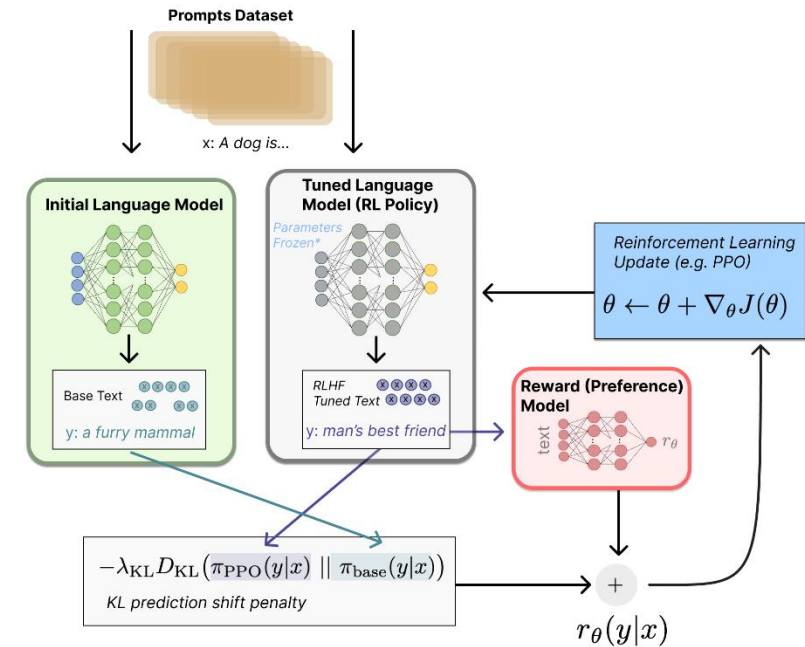
RLHF for Alignment

- A standard pipeline of RLHF: reward modelling + PPO

(i) reward model learning



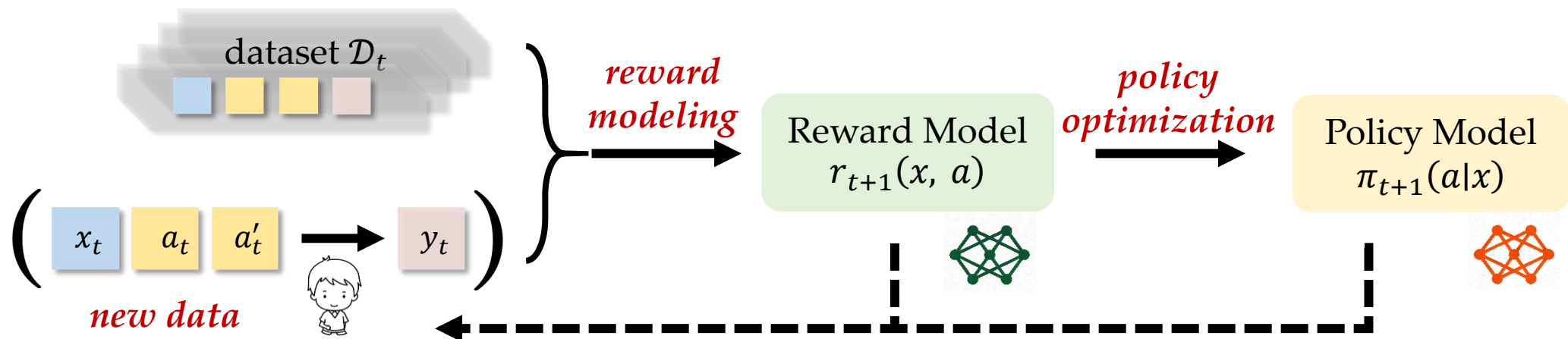
(ii) policy optimization (guided by reward model)



Online RLHF

General Framework of Online RLHF

- 1: **New data collection:** sample a tuple (x_t, a_t, a'_t) , obtain the preference label y_t , expand the dataset: $\mathcal{D}_{t+1} = \mathcal{D}_t \cup (x_t, a_t, a'_t, y_t)$
- 2: **Reward Modeling:** Train reward model r_{t+1} based on dataset $\mathcal{D}_{t+1}$
- 3: **Policy Optimization:** Update the policy π_{t+1} using the learned reward model r_{t+1}



Reward Model Learning

- How to model the underlying reward based on observed data?

Definition 1 (Bradley-Terry Model). Given a context $x \in \mathcal{X}$ and two actions $a, a' \in \mathcal{A}$, the probability of the human preferring action a over action a' is given by

$$\mathbb{P}(a \succ a' \mid x) = \frac{\exp(r(x, a))}{\exp(r(x, a)) + \exp(r(x, a'))}$$

where r is the latent function.

- **Reward Modeling:** Maximum Likelihood Estimation (MLE)

Define feature difference: $z_t = \phi(x_t, a_t) - \phi(x_t, a'_t)$

$$\hat{\theta}_{t+1} = \arg \min_{\theta \in \mathbb{R}^d} \sum_{s=1}^t \ell_s(\theta),$$

where $\ell_t(\theta) = -y_t \log(\sigma(z_t^\top \theta)) - (1 - y_t) \log(1 - \sigma(z_t^\top \theta))$

At iteration t :

*per-round time: $O(t \log t)$,
storage memory: $O(t)$*

Deploying bandits techniques

- Linear reward model assumption

$$r(x, a) = \phi(x, a)^\top \theta^*$$

BT model



$$\mathbb{P}(a \succ a' \mid x) = \frac{\exp(\phi(x, a)^\top \theta^*)}{\exp(\phi(x, a)^\top \theta^*) + \exp(\phi(x, a')^\top \theta^*)}$$

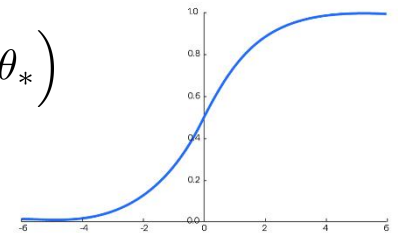
- $\phi(x, a)$ is the known feature mapping
- θ^* is the **unknown** parameter

- Contextual dueling bandits

At each round $t = 1, 2, \dots$

- (1) the learner first chooses two arms $\mathbf{x}_t, \mathbf{y}_t \in \mathcal{X} \subseteq \mathbb{R}^d$;
- (2) and then environment reveals a preference feedback o_t .

$$\mathbb{P}(o_t = 1) = \mu((\mathbf{x}_t - \mathbf{y}_t)^\top \theta_*)$$
$$\mu(z) = \frac{1}{1 + \exp(-z)}$$



One-Pass Reward Modeling

- OMD for one-pass estimation

Define gradient and Hessian: $g_t(\theta) = (\sigma(z_t^\top \theta) - y_t) z_t$, $H_t(\theta) = \dot{\sigma}(z_t^\top \theta) z_t z_t^\top$

$$\tilde{\theta}_{t+1} = \arg \min_{\theta \in \Theta} \left\{ \langle g_t(\tilde{\theta}_t), \theta \rangle + \frac{1}{2\eta} \|\theta - \tilde{\theta}_t\|_{\tilde{\mathcal{H}}_t}^2 \right\}, \text{ where } \tilde{\mathcal{H}}_t = \sum_{i=1}^{t-1} H_i(\tilde{\theta}_{i+1}) + \eta H_t(\tilde{\theta}_t) + \lambda I.$$

*constant time and storage cost,
independent of t*

one-pass!

*look-ahead
local norm*

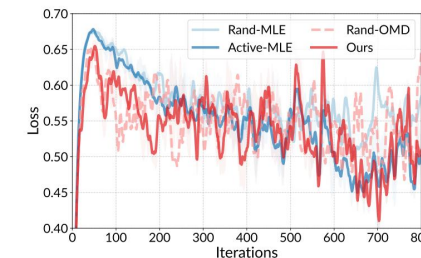
*second-order
approximation*

Estimation error

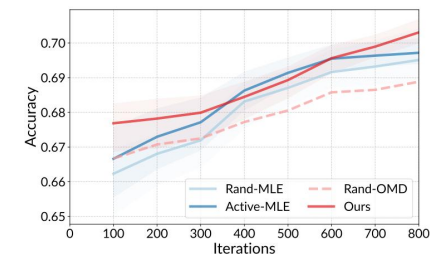
$$\|\theta - \tilde{\theta}_t\|_{\mathcal{H}_t} \leq \mathcal{O}(\sqrt{d}(\log(t/\delta))^2)$$

Regret bound

$$\text{Reg}_T \leq \tilde{\mathcal{O}} \left(d \sqrt{\frac{T}{\kappa}} \right)$$



(a) training loss



(b) evaluation accuracy



[Li*-Qian*-Z-Zhou, NeurIPS'25] Provably Efficient Online RLHF with One-Pass Reward Modeling.

Summary

❑ How to do interactive learning without revisiting history with guarantee?

❑ One-Pass Bandits

- Beyond linear bandits: For non-quadratic loss, MLE doesn't enjoy the one-pass property
- *Generalized linear bandits*: exploit the self-concordance property of the link function
- *Heavy-tailed linear bandits*: adaptively set Huber threshold to adjust curvatures such that outliers fall in the linear region, while normal data remain in the quadratic region

❑ OMD Estimator

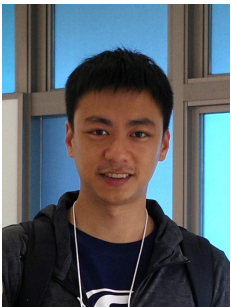
- Online Mirror Descent as a statistical estimator, where the *curvature-aware adaptivity* is crucial for the local norm design; similar to “from SGD to AdaGrad/Adam”

❑ RL Implications

- *RL with function approximation*: MNL mixture MDPs (related to GLB)
- *RLHF*: Bradley-Terry model naturally relates to logistic bandits, etc.

Reference

-  [NeurIPS 2025] Yu-Jie Zhang, Sheng-An Xu, Peng Zhao, Masashi Sugiyama. Generalized Linear Bandits: Almost Optimal Regret with **One-Pass** Update.
-  [ICML 2025] Jing Wang, Yu-Jie Zhang, Peng Zhao, and Zhi-Hua Zhou. Heavy-Tailed Linear Bandits: Huber Regression with **One-Pass** Update.
-  [NeurIPS 2025] Long-Fei Li*, Yu-Yang Qian*, Peng Zhao, Zhi-Hua Zhou. Provably Efficient Online RLHF with **One-Pass** Reward Modeling.
-  [NeurIPS 2024] Long-Fei Li, Yu-Jie Zhang, Peng Zhao, Zhi-Hua Zhou. Provably Efficient Reinforcement Learning with Multinomial Logit Function Approximation. **Thanks!**



Yu-Jie Zhang
(NJU → U Tokyo → UW)



Jing Wang
(NJU)



Long-Fei Li
(NJU → Noah's Ark Lab)



Yu-Yang Qian
(NJU)



Sheng-An Xu
(NJU → UCB)



Masashi Sugiyama
(RIKEN & U Tokyo)



Zhi-Hua Zhou
(NJU)