# Near-Optimal Representation Learning for Linear Bandits and Linear RL

Jiachen Hu [1] Xiaoyu Chen [1] Chi Jin [2] Lihong Li [3] Liwei Wang [1]

[1]Peking University [2]Princeton University [3]Amazon

June 20, 2021

# Multi-Task Low Rank Linear Bandits

- The agent chooses arms for $M$ linear bandits instance concurrently
  - At each step, he pulls $M$ arms and then receives $M$ rewards
  - Linear rewards $r_{t,i} = \langle a_{t,i}, \theta_i^\star \rangle + \text{noise}$
- The low rank property: There exists $k \ll d$ such that

$$\operatorname{rank}\left(\begin{bmatrix} \theta_1^\star & \theta_2^\star & \cdots & \theta_M^\star \end{bmatrix}\right) = k$$

$$\theta_i^\star = B^\star w_i^\star, \forall i \in [M]$$

- $B^\star \in \mathbb{R}^{d \times k}$: the shared $k$-dim representation. $w_i^\star \in \mathbb{R}^k$: the task-dependent parameter

# Multi-Task Low Rank Linear Bandits

- The agent chooses arms for $M$ linear bandits instance concurrently
  - At each step, he pulls $M$ arms and then receives $M$ rewards
  - Linear rewards $r_{t,i} = \langle a_{t,i}, \theta_i^\star \rangle + \text{noise}$
- The low rank property: There exists $k \ll d$ such that

$$\operatorname{rank}\left(\begin{bmatrix} \theta_1^\star & \theta_2^\star & \cdots & \theta_M^\star \end{bmatrix}\right) = k$$

$$\theta_i^\star = B^\star w_i^\star, \forall i \in [M]$$

- $B^\star \in \mathbb{R}^{d \times k}$: the shared $k$-dim representation. $w_i^\star \in \mathbb{R}^k$: the task-dependent parameter
- We can estimate $B^\star$ using the data from all tasks!

# Algorithm

## Multi-task linear bandits

For step $t = 1, 2, \ldots, T$:

1. Estimate $\{\theta_i^\star\}_{i=1}^M$ using constrained least-square linear regression

$$\{\hat{\theta}_{t,i}\}_{i=1}^M = \{\hat{B}_t \hat{w}_{t,i}\}_{i=1}^M \stackrel{\text{def}}{=} \arg\min_{\|Bw_i\|_2 \le 1} \sum_{\tau=1}^{t-1} \sum_{i=1}^{M} (\langle a_{\tau,i}, Bw_i \rangle - r_{\tau,i})^2$$

2. Construct tighter confidence set $\mathcal{C}_t \subset (\mathbb{R}^d)^M$ centered at $\{\hat{\theta}_{t,i}\}_{i=1}^M$ such that $\{\theta_i^\star\}_{i=1}^M \in \mathcal{C}_t$ with high probability
3. Choose actions optimistically and observe reward $\{r_{t,i}\}_{i=1}^M$

$$\{a_{t,i}\}_{i=1}^M \leftarrow \arg\max_{a_i \in \mathcal{A}_{t,i}} \max_{\{\theta_i\}_{i=1}^M \in \mathcal{C}_t} \sum_{i=1}^{M} \langle a_i, \theta_i \rangle$$

# Results

- Independent single-task algorithm for each instance

$$\tilde{O}\left(Md\sqrt{T}\right)$$

- Multi-task algorithm

$$\tilde{O}\left(M\sqrt{dkT} + d\sqrt{MkT}\right)$$

- Known lower bound

$$\Omega\left(Mk\sqrt{T} + d\sqrt{MkT}\right)$$

# Multi-task Low Rank Episodic RL

- The multi-task LSVI condition: We have access to a joint linear function class $\mathcal{Q}$ with low inherent Bellman error [1]
- Linear parameters in $\mathcal{Q}$ share a $k$-dim subspace
  - $k$ is much smaller than the feature dimension $d$

[1] See Zanette et al. Learning Near Optimal Policies with Low Inherent Bellman Error

# Multi-task Low Rank Episodic RL

- The multi-task LSVI condition: We have access to a joint linear function class $\mathcal{Q}$ with low inherent Bellman error [1]
- Linear parameters in $\mathcal{Q}$ share a $k$-dim subspace
  - $k$ is much smaller than the feature dimension $d$
- Independent single-task regret

$$\tilde{O}\left(HMd\sqrt{T} + HMT\sqrt{d}\mathcal{I}\right)$$

- Multi-task regret

$$\tilde{O}\left(HM\sqrt{dkT} + Hd\sqrt{kMT} + HMT\sqrt{d}\mathcal{I}\right)$$

- Lower bound

$$\Omega\left(Mk\sqrt{HT} + d\sqrt{HkMT} + HMT\sqrt{d}\mathcal{I}\right)$$

[1] See Zanette et al. Learning Near Optimal Policies with Low Inherent Bellman Error