Paper reading: Near-Optimal Representation Learning for Linear Bandits and Linear RL

Author: Jiachen Hu, Xiaoyu Chen, Chi Jin, Lihong Li, Liwei Wang

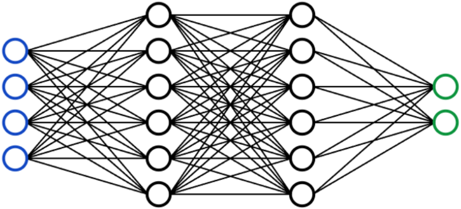
November 11, 2021

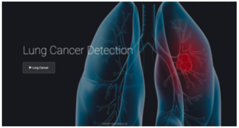
multi-task learning

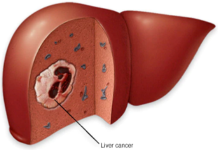
Given M learning tasks $\{T_i\}_{i=1}^{M}$, where all the tasks or a subset of them are related but not identical.

goal: improve the performance of multiple related learning tasks by leveraging useful information among them.

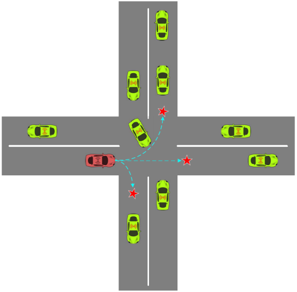
example: computer aided medical diagnosis







example: navigating unsignalized intersection



three navigation tasks, non-identical but related:

- going straight
- turning left
- turning right

How to encode the task relatedness into the learning model?

- Low-rank approach
- Task-clustering approach
- Task-relation learning approach
- Multi-level approach

- Linear stochastic bandit
- Multi-task linear stochastic bandit
- Multi-task reinforcement learning

Linear stochastic bandit

Recall Multi-Armed Bandit model in the class

Algorithm 2 Multi-Armed Bandits

for $k = 0, 1, \ldots, K-1$ **do**
  agent takes action a_k according to π^k
  agent receives a noisy reward $r_k \in [0, 1]$, with $\mathbb{E}[r_k | a_k] = r(a^k)$.
end for

Linear stochastic bandit - Learning model

decision set (given in advance): $D_t \subset \mathbb{R}^d$

choose action: X_t

observe reward: $Y_t = \langle X_t, \theta_* \rangle + \eta_t$
where $\theta_* \in \mathbb{R}^d$ (unknown), η_t noise, centered, tail constrained

goal: $\max \sum_{t=1}^{n} \langle X_t, \theta_* \rangle$

Linear stochastic bandit - OFUL algorithm

for $t := 1, 2, \ldots$ **do**

$(X_t, \widetilde{\theta}_t) = \mathrm{argmax}_{(x,\theta)\in D_t \times C_{t-1}} \langle x, \theta \rangle$

Play X_t and observe reward Y_t

Update C_t

end for

maximise reward $\Leftrightarrow$ minimise regret

$$R_n = \left(\sum_{t=1}^{n} \langle x_t^*, \theta_* \rangle \right) - \left(\sum_{t=1}^{n} \langle X_t, \theta_* \rangle \right) = \sum_{t=1}^{n} \langle x_t^* - X_t, \theta_* \rangle$$

Linear stochastic bandit

Theorem (Self-Normalized Bound for Vector-Valued Martingales)

Let X_t be an $\mathbb{R}^d$-valued stochastic process. Assume that V is a $d \times d$ positive definite matrix. For any t, define

$$\bar{V}_t = V + \sum_{s=1}^{t} X_s X_s^\top, \quad S_t = \sum_{s=1}^{t} \eta_s X_s$$

Then with probability at least $1-\delta$ for all t,

$$\|S_t\|^2_{\bar{V}_t^{-1}} \leq 2R^2 \log\left(\frac{\det\left(\bar{V}_t\right)^{1/2} \det(V)^{-1/2}}{\delta} \right)$$

Linear stochastic bandit - confidence sets

ℓ_2-regularized least-squares estimate of θ_*:

$$\widehat{\theta}_t = \left(\mathbf{X}_{1:t}^\top \mathbf{X}_{1:t} + \lambda I\right)^{-1} \mathbf{X}_{1:t}^\top \mathbf{Y}_{1:t}$$

Theorem (Confidence Ellipsoid)

With probability at least $1-\delta$ for all t, θ_* lies in the set

$$C_t = \left\{ \theta \in \mathbb{R}^d : \left\| \widehat{\theta}_t - \theta \right\|_{\bar{V}_t} \leq R\sqrt{2\log\left(\frac{\det\left(\bar{V}_t\right)^{1/2} \det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{1/2} S \right\}$$

Linear stochastic bandit - confidence sets

Theorem (Confidence Ellipsoid Cont'd)

Furthermore, if for all t $||X_t||_2 \leq L$, the with probability at least $1-\delta$ for all t, θ_* lies in the set

$$C_t = \left\{ \theta \in \mathbb{R}^d : \left\| \widehat{\theta}_t - \theta \right\|_{\bar{V}_t} \leq R \sqrt{d \log\left(\frac{1+tL^2/\lambda}{\delta}\right)} + \lambda^{1/2} S \right\}$$

As comparison:

- Dani et al. (2008): $\left\| \widehat{\theta}_t - \theta_* \right\|_{\bar{V}_t} \leq R \max\left\{ \sqrt{128 d \log(t) \log\left(\frac{t^2}{\delta}\right)}, \frac{8}{3} \log\left(\frac{t^2}{\delta}\right) \right\}$
- Rusmevichientong and Tsitsiklis (2010): $\left\| \widehat{\theta}_t - \theta_* \right\|_{\bar{V}_t} \leq 2\kappa^2 R \sqrt{\log t} \sqrt{d \log(t) + \log\left(t^2/\delta\right)} + \lambda^{1/2} S$

Linear stochastic bandit - regret

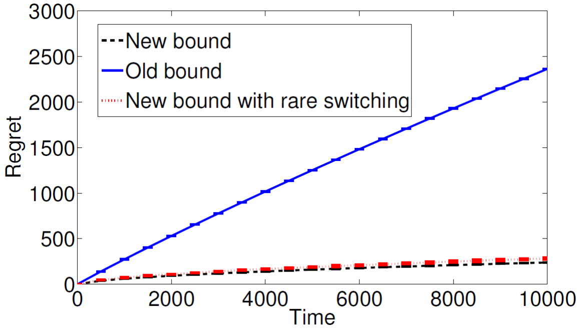
Theorem

Assume that for all t and all $x \in D_t$ $\langle x, \theta_* \rangle \in [-1, 1]$, the with probability at least $1 - \delta$ the regret of the OFUL algorithm satisfies,

$$R_t \leq 4\sqrt{td \log\left(\lambda + \frac{tL}{d}\right)} \left\{ \sqrt{\lambda} S + R\sqrt{d \log\left(1 + \frac{tL}{\lambda d}\right) + 2 \log \frac{1}{\delta}} \right\}$$

Almost matches lower bound by Rusmevichientong and Tsitsiklis, which is $\Omega(d\sqrt{t})$

Multi-task Linear stochastic bandit - regret



Multi-task Linear stochastic bandit - Learning model

play M tasks concurrently for T steps each

decision set (given in advance): $A_{t,i} \subset \mathbb{R}^d$

choose action: $\boldsymbol{x}_{t,i}$ for $i \in [M]$

observe reward: $Y_{t,i} = \langle \boldsymbol{x}_{t,i}, \boldsymbol{\theta}_i \rangle + \eta_{t,i}$ for $i \in [M]$

goal: $\min \mathrm{Reg}(T) \stackrel{\text{def}}{=} \sum_{t=1}^{T} \sum_{i=1}^{M} \left(\left\langle \boldsymbol{x}_{t,i}^{\star}, \boldsymbol{\theta}_i \right\rangle - \langle \boldsymbol{x}_{t,i}, \boldsymbol{\theta}_i \rangle \right)$
where, $\boldsymbol{x}_{t,i}^{\star} = \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{A}_{t,i}} \langle \boldsymbol{x}, \boldsymbol{\theta}_i \rangle$.

Multi-task Linear stochastic bandit

Key assumption:

There exists a linear feature extractor $\boldsymbol{B} \in \mathbb{R}^{d\times k}$ and and a set of k-dimensional coefficients $\{\boldsymbol{w}_i\}_{i=1}^M$ such that $\{\boldsymbol{\theta}_i\}_{i=1}^M$ satisfies $\boldsymbol{\theta}_i = \boldsymbol{B}\boldsymbol{w}_i$.

Other standard regularity assumptions

$$\|\boldsymbol{\theta}_i\|_2 \leq 1, \forall i \in [M]$$

$$\|\boldsymbol{x}\|_2 \leq 1, \forall \boldsymbol{x} \in \mathcal{A}_{t,i}, t \in [T], i \in [M]$$

Multi-task Linear stochastic bandit

How about we run OFUL algorithm for the M tasks independently?

Recall the confidence set from OFUL algorithm:

$$C_t = \left\{ \theta \in \mathbb{R}^d : \left\| \widehat{\theta}_t - \theta \right\|_{\bar{V}_t} \leq R \sqrt{d \log \left(\frac{1 + tL^2/\lambda}{\delta} \right)} + \lambda^{1/2} S \right\}$$

Multi-task Linear stochastic bandit - Multi-Task Low-Rank OFUL

Algorithm 1 Multi-Task Low-Rank OFUL

for step $t = 1, 2, \cdots, T$ **do**
  Calculate the confidence interval $\mathcal{C}_t$ by Eqn 8
  $\tilde{\boldsymbol{\Theta}}_t, \boldsymbol{x}_{t,i} = \operatorname{argmax}_{\boldsymbol{\Theta} \in \mathcal{C}_t, \boldsymbol{x}_i \in \mathcal{A}_{t,i}} \sum_{i=1}^{M} \langle \boldsymbol{x}_i, \boldsymbol{\theta}_i \rangle$
  for task $i = 1, 2, \cdots, M$ **do**
    Play $\boldsymbol{x}_{t,i}$ for task i, and obtain the reward $y_{t,i}$
  end for
end for

Multi-task Linear stochastic bandit - confidence sets

optimism in the face of uncertainty principle

choose an optimistic estimation

$$\tilde{\boldsymbol{\theta}}_t = \mathrm{argmax}_{\boldsymbol{\theta} \in \mathcal{C}_t} \left(\max_{\boldsymbol{x} \in \mathcal{A}_t} \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle \right)$$

multi-task setting:

$$\tilde{\boldsymbol{\Theta}}_t = \mathrm{argmax}_{\boldsymbol{\Theta} \in \mathcal{C}_t} \left(\max_{\{x_i \in \mathcal{A}_{t,i}\}_{i=1}^{M}} \sum_{i=1}^{M} \langle \boldsymbol{x}_i, \boldsymbol{\theta}_i \rangle \right)$$

where $\boldsymbol{\Theta} \stackrel{\text{def}}{=} [\boldsymbol{\theta}_1, \boldsymbol{\theta_2}, \cdots, \boldsymbol{\theta_M}]$

Multi-task Linear stochastic bandit - confidence sets

Suppose we have samples collected till $t-1$, calculate by least-square problem:

$$\begin{aligned} \arg\min_{\boldsymbol{B}\in\mathbb{R}^{d\times k},\boldsymbol{w}_{1..M}\in\mathbb{R}^{k\times M}} & \sum_{i=1}^{M} \left\| \boldsymbol{y}_{t-1,i} - \boldsymbol{X}_{t-1,i}^{\top} \boldsymbol{B}\boldsymbol{w}_i \right\|_2^2 \\ \text{s.t.} \quad & \left\| \boldsymbol{B}\boldsymbol{w}_i \right\|_2 \leq 1, \forall i \in [M] \end{aligned}$$

Multi-task Linear stochastic bandit - confidence sets

Theorem

With probability at least $1-\delta$ for all t, the true parameter $\boldsymbol{\theta} = \boldsymbol{B}\boldsymbol{w}$ is always contained in the confidence set

$$\mathcal{C}_t \overset{\text{def}}{=} \left\{ \boldsymbol{\Theta} = \boldsymbol{B}\boldsymbol{W} : \sum_{i=1}^{M} \left\| \hat{\boldsymbol{B}}_t \hat{\boldsymbol{w}}_{t,i} - \boldsymbol{B}\boldsymbol{w}_i \right\|^2_{\tilde{\boldsymbol{V}}_{t-1,i}(\lambda)} \leq L \text{ , } \boldsymbol{B} \in \mathbb{R}^{d\times k}, \boldsymbol{w}_i \in \mathbb{R}^k, \|\boldsymbol{B}\boldsymbol{w}_i\|_2 \leq 1, \forall i \in [M] \right\}$$

where. $L = \tilde{O}(Mk + kd)$.

Multi-task Linear stochastic bandit - regret

Theorem

With probability at least $1-\delta$ for all t, the regret of Multi-Task Low-Rank OFUL algorithm is bounded by:

$$\mathrm{Reg}(T) = \tilde{O}(M\sqrt{dkT} + d\sqrt{kMT} + MT\sqrt{d}\zeta)$$

Theorem

For any k, M, d, T, with $k < d < T$ and $k < M$, and any learning algorithm. There exist a multi-task linear bandit instance such that the regret of algorithm is lower bounded by

$$\mathrm{Reg}(T) \geq \Omega(Mk\sqrt{T} + d\sqrt{kMT} + MT\sqrt{d}\zeta)$$

Multi-task RL

undiscounted episodic MDP:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, H)$$

multi-task episodic MDP:

$$\mathcal{M}^1, \mathcal{M}^1, \cdots, \mathcal{M}^m$$

share the same state space and action space, but have different rewards and transitions
The total regret of M tasks in T episodes:

$$\mathrm{Reg}(T) \stackrel{\mathsf{def}}{=} \sum_{t=1}^{T} \sum_{i=1}^{M} \left(V_1^{i*} - V_1^{\pi_t^i} \right) \left(s_{1t}^i \right)$$

Multi-task RL - approximate linear value functions

at $h \in [H]$, define the following function space

$$\mathcal{Q}'_h = \left\{ Q_h\left(\boldsymbol{\theta}_h\right) \mid \boldsymbol{\theta}_h \in \Theta'_h \right\}$$

where $Q_h\left(\boldsymbol{\theta}_h\right)(s,a) \stackrel{\text{def}}{=} \boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_h$.

$$\mathcal{V}'_h = \left\{ V_h\left(\boldsymbol{\theta}_h\right) \mid \boldsymbol{\theta}_h \in \Theta'_h \right\}$$

where $V_h\left(\theta_h\right)(s) \stackrel{\text{def}}{=} \max_a \boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_h$

Multi-task RL - approximate linear value functions

inherent Bellman error (Zanette et al., 2020a):

$$\mathcal{I}_h \stackrel{\text{def}}{=} \sup_{Q_{h+1} \in \mathcal{Q}_{h+1}} \inf_{Q_h \in \mathcal{Q}_h} \sup_{s \in \mathcal{S}, a \in \mathcal{A}} \left| \left(Q_h - \mathcal{T}_h \left(Q_{h+1} \right) \right) (s, a) \right|$$

where, Bellman optimality operator:

$$\mathcal{T}_h \left(Q_{h+1} \right) (s, a) \stackrel{\text{def}}{=} r_h(s, a) + \mathbb{E}_{s' \sim p_h(\cdot | s, a)} \max_{a'} Q_{h+1} \left(s', a' \right)$$

Multi-task RL

in the case of multi-task RL, redefine the parameter space:

$$\Theta_h \overset{\text{def}}{=} \left\{ \left(\boldsymbol{B}_h \boldsymbol{w}_h^1, \boldsymbol{B}_h \boldsymbol{w}_h^2, \cdots, \boldsymbol{B}_h \boldsymbol{w}_h^M \right) : \boldsymbol{B}_h \in \mathcal{O}^{d \times k}, \boldsymbol{w}_h^i \in \mathcal{B}^k, \boldsymbol{B}_h \boldsymbol{w}_h^i \in \Theta_h^{i\prime} \right\}$$

a generalization of inherent Bellman error:

$$\mathcal{I}_h^{\text{mul}} \overset{\text{def}}{=} \sup_{\left\{Q_{h+1}^i\right\}_{i=1}^M \in \mathcal{Q}_{h+1}} \inf_{\left\{Q_h^i\right\}_{i=1}^M \in \mathcal{Q}_h} \sup_{s \in \mathcal{S}, a \in \mathcal{A}, i \in [M]} \left| \left(Q_h^i - \mathcal{T}_h^i \left(Q_{h+1}^i \right) \right) (s, a) \right|$$

Multi-task RL - MTLR-LSVI

Algorithm 2 Multi-Task Low-Rank LSVI

Input: low-rank parameter k, failure probability δ, regularization $\lambda = 1$, inherent Bellman error $\mathcal{I}$
Initialize $\tilde{\boldsymbol{V}}_{h1} = \lambda \boldsymbol{I}$ for $h \in [H]$
for episode $t = 1, 2, \cdots$ **do**
    Compute α_{ht} for $h \in [H]$. (see Lemma 9)
    Solve the global optimization problem 1
    Compute $\pi_{ht}^i(s) = \operatorname{argmax}_a \boldsymbol{\phi}(s,a)^\top \bar{\boldsymbol{\theta}}_{ht}^i$
    Execute π_{ht}^i for task i at step h
    Collect $\left\{s_{ht}^i, a_{ht}^i, r\left(s_{ht}^i, a_{ht}^i\right)\right\}$ for episode t.
end for

---

optimization procedure in every episode:

$$\max_{\overline{\boldsymbol{\xi}}_h^i, \hat{\boldsymbol{\theta}}_h^i, \overline{\boldsymbol{\theta}}_h^i} \sum_{i=1}^M \max_{a^i} \left(\phi\left(s_1^i, a^i\right)\right)^\top \overline{\boldsymbol{\theta}}_1^i$$

constraints:

- $\hat{\boldsymbol{B}}_h \left[\begin{array}{cccc} \hat{\boldsymbol{w}}_h^1 & \hat{\boldsymbol{w}}_h^2 & \cdots & \hat{\boldsymbol{w}}_h^M \end{array}\right] \leq \underset{\|\boldsymbol{B}_h \boldsymbol{w}_h^i\|_2 \leq D}{\operatorname{argmin}} \sum_{i=1}^M \sum_{j=1}^{t-1} L\left(\boldsymbol{B}_h, \boldsymbol{w}_h^i\right)$
- $\overline{\boldsymbol{\theta}}_h^i = \hat{\boldsymbol{\theta}}_h^i + \overline{\boldsymbol{\xi}}_h^i; \quad \sum_{i=1}^M \left\|\overline{\boldsymbol{\xi}}_h^i\right\|_{\tilde{\boldsymbol{V}}_{ht}^i(\lambda)}^2 \leq \alpha_{ht}; \ \left(\overline{\boldsymbol{\theta}}_h^1, \overline{\boldsymbol{\theta}}_h^2, \cdots, \overline{\boldsymbol{\theta}}_h^M\right) \in \Theta_h$

Multi-task RL - regret

Theorem

With probability at least $1-\delta$ the regret after T episodes is bounded by:

$$\mathrm{Reg}(T) = \tilde{O}(HM\sqrt{dkT} + Hd\sqrt{kMT} + HMT\sqrt{d}\mathcal{I})$$

Key step to the result:

$$\sum_{i=1}^{M} \left\| \hat{\boldsymbol{\theta}}_h^i - \dot{\boldsymbol{\theta}}_h^i \right\|^2_{\tilde{\boldsymbol{V}}_{ht}^i(\lambda)} = \tilde{O}\left(Mk + kd + MT\mathcal{I}^2\right)$$

Multi-task RL - lower bound

Theorem

The expected regret of any algorithm where $d, k, H > 10, |\mathcal{A}| \geq 3, M \geq k, T = \Omega(d^2 H), \mathcal{I} \leq 1/4H$ *is*

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

Multi-task RL - MTLR-LSVI

Thanks!