Risk sensitive Reinforcement Learning with Low-rank MDPs (In progress)

March 30, 2023

Yulai Zhao

Backgrounds

- What is risk sensitive RL?
- First recall the definition of Conditional Value-at-Risk (CVaR), defined as: $CVaR_\tau(x) = E_{x \sim P}[x | x < P^{-1}(\tau)]$
- Often, τ is set to be 0.05, or 0.01
- It is often used as an empirical **RISK** measure, for example, if a portfolio has a high expected return, is it good?
- Not necessarily! We need to figure out what the return will be with the WORST 5% probability.

Backgrounds

- What is risk sensitive RL?
- Some find we can write $CVaR_\tau(x) = E_{x\sim P}[x|x < P^{-1}(\tau)]$ in another way:

  $CVaR_\tau(R) = \max_b[b - \frac{1}{\tau} E_{R\sim P}\max(b - R, 0)]$, we can understand b as a threshold thing, where R is the normal reward from a process.
- Imagine the randomness comes from policy π in the MDP, and R is the received reward, then in a risk-sensitive RL, we aim at maximizing the CVaR measure

$$
\begin{aligned}
max_\pi \; CVaR_\tau(R) &= max_\pi \max_b [b - \frac{1}{\tau} E_{R\sim P}\max(b - R, 0)] \\
&= \max_b [b - \frac{1}{\tau} \min_\pi E_R \max(b - R, 0)]
\end{aligned}
$$

Note that in a RL language, $R = \sum_{i=1} r(s_i, a_i)$

Setup

- MDP
  - a tuple $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathrm{r}, \gamma)$: A set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a transition probability $\mathcal{P}: \mathcal{S} \times \mathcal{A} \times \mathcal{A} \to \Delta(\mathcal{S})$, a known and deterministic reward function $\mathrm{r}: \mathcal{S} \times \mathcal{A} \times \mathcal{A} \to [0, 1]$, a discounted factor $\gamma \in [0, 1)$.
  - We can easily extend r(s,a) to unknown and stochastic case.
  - Start from the initial state distribution d_0
- For each threshold b, we have

$$V^\pi(s, b) = E\left[\max(0,\ b\ -\ \textstyle\sum_{t=0}^{\infty} \gamma^t\ r(s_t, a_t, b_t)) \mid s_0 = s, \pi, b\ \right]$$
$$J^*(b) = \min_\pi J(b) = \min_\pi E_{s\sim d_0} V^\pi(s, b)$$

The final goal is to maximize CVaR, i.e., $\max_b (b - \frac{1}{\tau} J^*(b))$

<u>So this means we can only have a finite set of b and do this bi-level optimization?</u>

Setup

-
- For each threshold b, we have Q function

$$Q^{\pi}(s,b,a) = E\left[\max(0,\ b\ -\ \sum_{t=0}^{\infty}\gamma^t\, r(s_t,a_t,b_t))\,\middle|\, s_0 = s, a_0 = a, \pi, b\ \right] = E_{s'\sim P(\cdot|s,a)} V^{\pi}(s', b\ - r(s,a))$$

The final goal is to maximize CVaR, i.e., $\max_b (b - \frac{1}{\tau} J^*(b))$

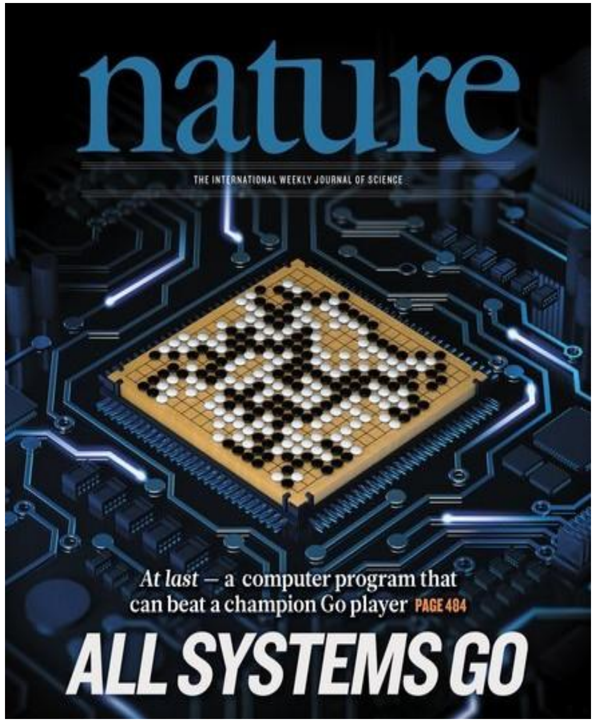
- Assume the MDP admits a low-rank decomposition

$$\forall s, s', a, \qquad P(s'|s,a) = \mu^*(s')^{\top} \phi^*(s,a)$$

Some Review

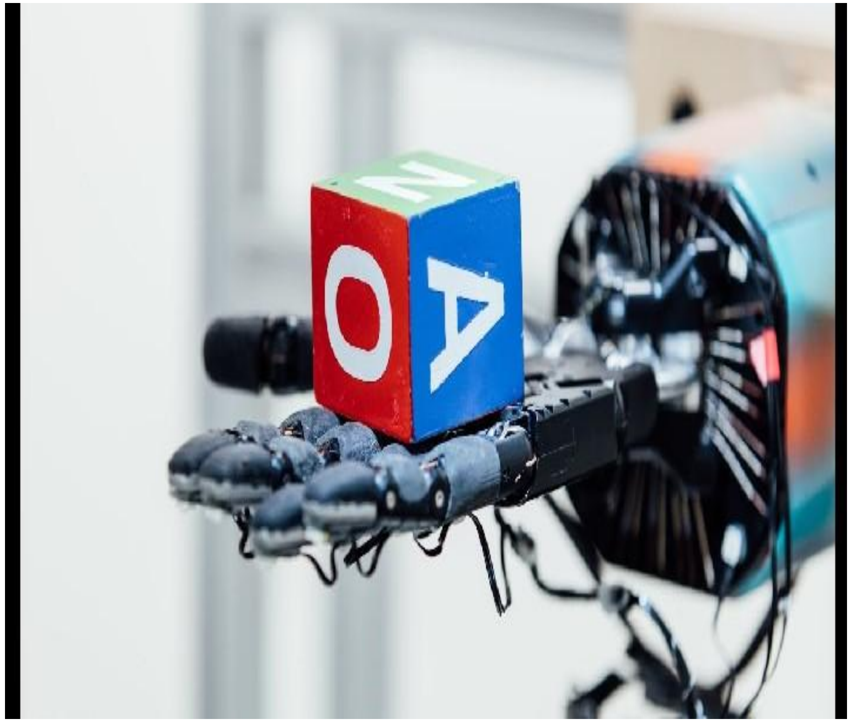
- We wish to learn some useful techniques from [Representation Learning for Online and Offline RL in Low-rank MDPs] by Masatoshi Uehara , Xuezhou Zhang, and Wen Sun
- Now we move on to introduce this paper and briefly sketch its proofs

Empirical RL for large-scale problems



[AlphaGo, Silver et.al, 15]

[OpenAI Five, 18]

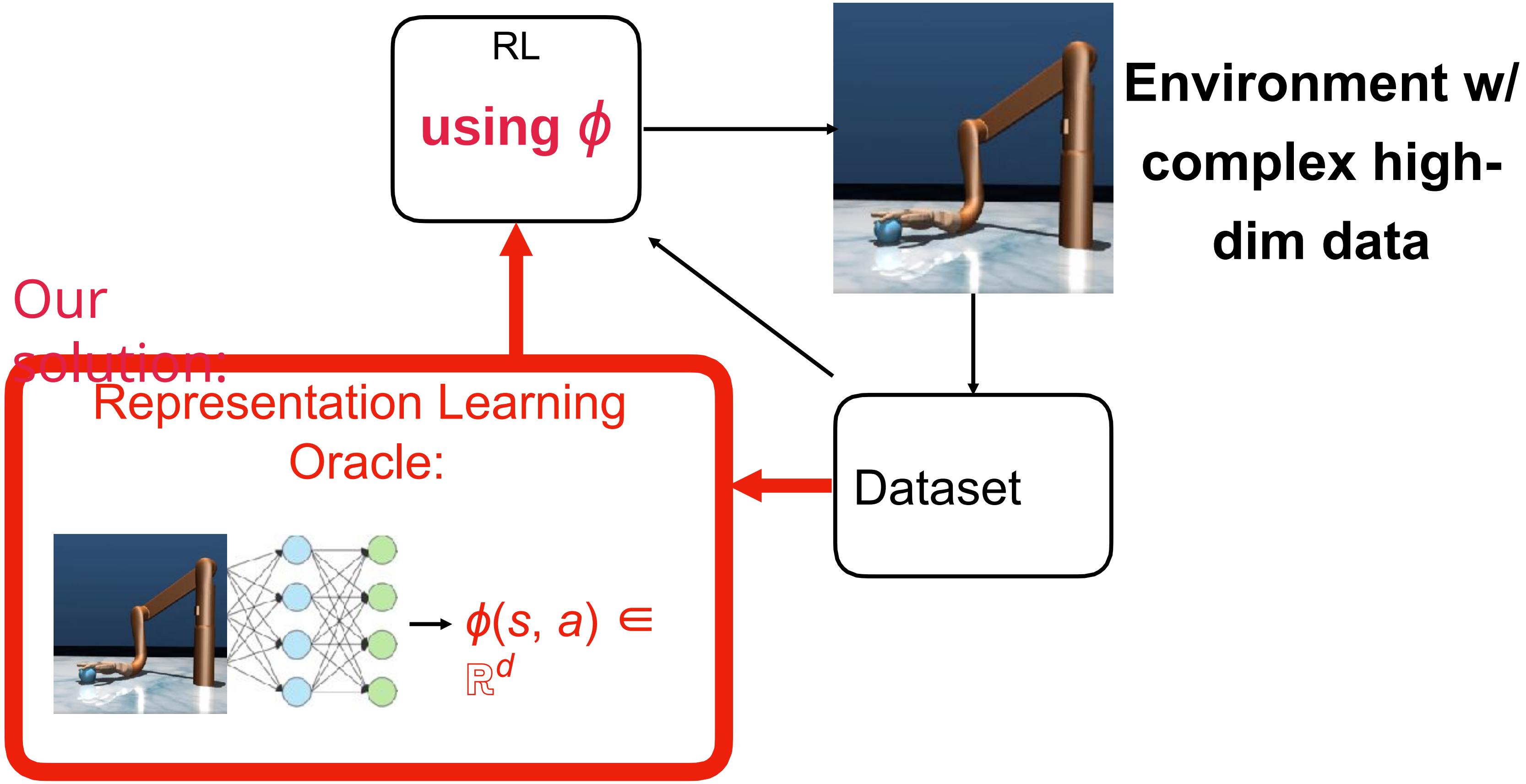


[OpenAI,19]

Rich (nonlinear) function approximation + RL can work well w/ enough samples

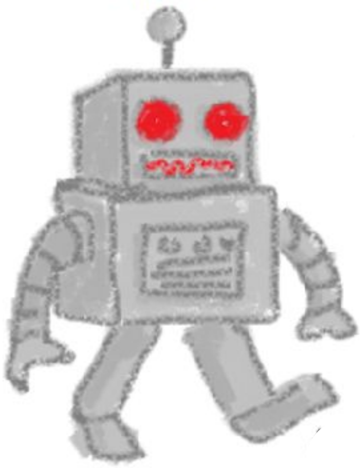
Can we design provably efficient algorithms for

Rich Function Approx + RL ?



Episodic Infinite Horizon Discounted MDPs

agent



Policy: state to action

$\pi(s) \longrightarrow a$

Reward & Next State

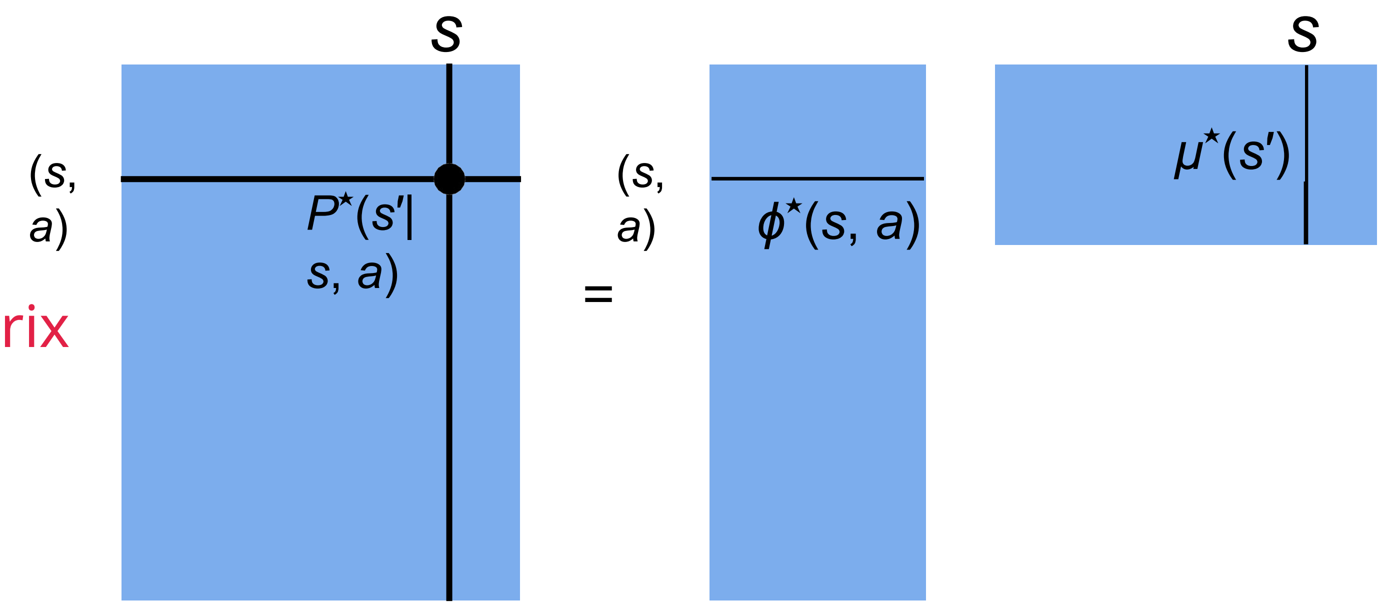
$r(s, a), s' \sim P(\cdot \mid s, a)$

Objective:

$\max_{\pi} J(\pi; P, r)$, where $J(\pi; P, r) := \mathbb{E}\left[r(s_0, a_0) + \gamma r(s_1, a_1) + \gamma^2 r(s_2, a_2) + \ldots \mid a \sim \pi, P\right]$

Assume fixed initial state s

Low-rank MDP



Transition matrix $P \in \mathbb{R}^{SA \times S}$ has rank d

$$\exists \mu^\star, \phi^\star \quad : \quad \forall s, a, s', P^\star(s'|\, s, a) = \mu^\star(s')^\top \phi^\star(s, a)$$

Low-rank MDP **Linear MDPs** (Jin et al, Yang & Wang)

Linear MDP = low-rank + known $\phi^\star$

The formulation

1. Realizable hypothesis classes Ψ, Φ, and

$$\mu^* \in \Psi, \phi^* \in \Phi$$

2. Computation oracle for learning representations: Maximum Likelihood Estimation (MLE): the sum is taken over a collection of tuples: $D = (s_i, a_i, s_i'), i = 1, \cdots, n$

$$(\hat{\mu}, \hat{\phi}) = \arg\max_{\mu,\phi} \sum_{i}^{n} \ln(\mu(s_i')^\top \phi(s_i, a_i))$$

Our algorithm: Rep-UCB

(UCB-driven Representation Learning for online RL)

At iteration n:

Data generation from π^n:

$s \sim d^{\pi^n}, a \sim \text{Uniform}(\mathscr{A}), s' \sim P^\star(.\,|\,s,a)$

$a' \sim \text{Uniform}(\mathscr{A}), s'' \sim P^\star(.\,|\,s',a')$

Data Aggregation:

$\mathscr{D}_n = \mathscr{D}_{n-1} + \{s, a, s'\}$

$\mathscr{D}'_n = \mathscr{D}'_{n-1} + \{s', a', s''\}$

Representation / model Learning (MLE)

$\hat{P} := (\hat{\mu}, \hat{\phi}) = \arg\max_{\mu,\phi} \mathbb{E}_{\mathscr{D}_n + \mathscr{D}'_n} \ln(\mu(s')^\top \phi(s,a))$

(Linear bandit style) bonus under $\hat{\phi}$:

$b(s,a) = c\sqrt{\hat{\phi}(s,a)\Sigma^{-1}\hat{\phi}(s,a)}$

$\Sigma = \sum_{s,a \in \mathscr{D}_n} \hat{\phi}(s,a)\hat{\phi}(s,a)^\top$

Plan w/ reward + bonus

$\pi^{n+1} = \max_\pi J(\pi; \hat{P}, r + b)$

Algorithm 1 UCB-driven representation learning, exploration, and exploitation (REP-UCB)

Input: Regularizer λ_n, parameter α_n, Models $\mathcal{M} = \{(\mu, \phi) : \mu \in \Psi, \phi \in \Phi\}$, Iteration N
Initialize $\pi_0(\cdot \mid s)$ to be uniform; set $\mathcal{D}_0 = \emptyset$, $\mathcal{D}'_0 = \emptyset$
for episode $n = 1, \cdots, N$ **do**
  Collect a tuple $(s, a, s', a', \tilde{s})$ with

$$s \sim d_{P^\star}^{\pi_{n-1}}, a \sim U(\mathcal{A}), s' \sim P^\star(\cdot|s,a), a' \sim U(\mathcal{A}), \tilde{s} \sim P^\star(\cdot|s',a')$$

  Update datasets by adding triples (s, a, s') and $(s', a', \tilde{s})$:

$$\mathcal{D}_n = \mathcal{D}_{n-1} + \{(s,a,s')\}, \quad \mathcal{D}'_n = \mathcal{D}'_{n-1} + \{(s',a',\tilde{s})\}$$

  Learn representation via ERM (i.e., MLE):

$$\hat{P}_n := (\hat{\mu}_n, \hat{\phi}_n) = \arg\max_{(\mu,\phi)\in\mathcal{M}} \mathbb{E}_{\mathcal{D}_n + \mathcal{D}'_n}\left[\ln \mu^\top(s')\phi(s,a)\right]$$

  Update empirical covariance matrix $\hat{\Sigma}_n = \sum_{s,a \in \mathcal{D}_n} \hat{\phi}_n(s,a)\hat{\phi}_n(s,a)^\top + \lambda_n I$
  Set the exploration bonus:

$$\hat{b}_n(s,a) := \min\left(\alpha_n \sqrt{\hat{\phi}_n(s,a)^\top \hat{\Sigma}_n^{-1} \hat{\phi}_n(s,a)}, 2\right) \tag{1}$$

  Update policy $\pi_n = \arg\max_\pi V^\pi_{\hat{P}_n, r + \hat{b}_n}$
end for
Return $\pi_1, \cdots, \pi_N$

PAC-Bound of Rep-UCB in low-rank MDP

Assume trajectory-reward is normalized in [0,1]. W/ high probability, it finds an ϵ near optimal policy, with # of samples:

$$\widetilde{O}\left(\frac{d^4 A^2}{\epsilon^2(1-\gamma)^5} \cdot \ln\left(|\Gamma||\Phi|\right)\right)$$

For reference, prior SOTA
FLAMBE has the following bound:

$$\widetilde{O}\left(\frac{d^7 A^9}{\epsilon^{10}(1-\gamma)^{22}} \cdot \ln\left(|\Gamma||\Phi|\right)\right)$$

Extension to offline RL

- We note that this algorithm can be extended to offline setting where the set D is pre-collected.
- Similar to other offline theories, a dataset coverage coefficient is needed

Coverage condition of the offline data

A comparator policy π is covered by offline data if the relative condition number is bounded:

$$C_{\pi^*} := \max_x \frac{x^\top \left(\mathbb{E}_{s,a\sim d^\pi} \phi^\star(s,a)\phi^\star(s,a)^\top\right) x}{x^\top \left(\mathbb{E}_{s,a\sim d^{\pi_b}} \phi^\star(s,a)\phi^\star(s,a)^\top\right) x} < \infty$$

Note coverage is wrt true representation only!

Goal is to learn **robustly**, i.e., as long as there is a high quality policy that is covered by d^{π_b}, we want to compete against it!

- However, the proof techniques are basically the same with the online algorithm.

Summary

1. Improved online Representation Learning algorithm for low-rank MDP:

Oracle-efficient + tight sample complexity

2. New offline RL algorithm for low-rank MDP:

Partial coverage + Oracle-efficient

Rep-UCB / LCB: https://arxiv.org/pdf/2110.04652.pdf

Proof Sketch

- Now we go over the proof sketch for this algorithm

Algorithm 1 UCB-driven representation learning, exploration, and exploitation (REP-UCB)

Input: Regularizer λ_n, parameter α_n, Models $\mathcal{M} = \{(\mu, \phi) : \mu \in \Psi, \phi \in \Phi\}$, Iteration N

Initialize $\pi_0(\cdot \mid s)$ to be uniform; set $\mathcal{D}_0 = \emptyset$, $\mathcal{D}'_0 = \emptyset$

for episode $n = 1, \cdots, N$ **do**

  Collect a tuple $(s, a, s', a', \tilde{s})$ with

$$s \sim d_{P^\star}^{\pi_{n-1}}, a \sim U(\mathcal{A}), s' \sim P^\star(\cdot|s,a), a' \sim U(\mathcal{A}), \tilde{s} \sim P^\star(\cdot|s', a')$$

  Update datasets by adding triples (s, a, s') and $(s', a', \tilde{s})$:

$$\mathcal{D}_n = \mathcal{D}_{n-1} + \{(s, a, s')\}, \quad \mathcal{D}'_n = \mathcal{D}'_{n-1} + \{(s', a', \tilde{s})\}$$

  Learn representation via ERM (i.e., MLE):

$$\hat{P}_n := (\hat{\mu}_n, \hat{\phi}_n) = \arg\max_{(\mu,\phi)\in\mathcal{M}} \mathbb{E}_{\mathcal{D}_n + \mathcal{D}'_n}\left[\ln \mu^\top(s')\phi(s,a)\right]$$

  Update empirical covariance matrix $\hat{\Sigma}_n = \sum_{s,a\in\mathcal{D}_n} \hat{\phi}_n(s,a)\hat{\phi}_n(s,a)^\top + \lambda_n I$

  Set the exploration bonus:

$$\hat{b}_n(s,a) := \min\left(\alpha_n \sqrt{\hat{\phi}_n(s,a)^\top \hat{\Sigma}_n^{-1} \hat{\phi}_n(s,a)}, 2\right) \tag{1}$$

  Update policy $\pi_n = \arg\max_\pi V^\pi_{\hat{P}_n, r+\hat{b}_n}$

end for

Return $\pi_1, \cdots, \pi_N$

Define ρ_n and ρ'_n as the marginal distributions for s in D_n, D_n', respectively.

Therefore, we have: for (s, a, s') in D_n, $(s, a, s') \sim \rho_n(s)U(a)P^*(s'|s,a)$

Useful auxiliary lemmas

- Capture model error

Lemma 18 (MLE guarantee). *For a fixed episode n, with probability $1-\delta$,*

$$\mathbb{E}_{s\sim\{0.5\rho_n+0.5\rho'_n\},a\sim U(\mathcal{A})}[\|\hat{P}_n(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1^2] \lesssim \zeta, \quad \zeta := \frac{\ln(|\mathcal{M}|/\delta)}{n}.$$

As a straightforward corollary, with probability $1-\delta$,

$$\forall n \in \mathbb{N}^+, \mathbb{E}_{s\sim\{0.5\rho_n+0.5\rho'_n\},a\sim U(\mathcal{A})}[\|\hat{P}_n(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1^2] \lesssim 0.5\zeta_n, \quad \zeta_n := \frac{\ln(|\mathcal{M}|n/\delta)}{n}.$$

Algorithm 1 UCB-driven representation learning, exploration, and exploitation (REP-UCB)

Input: Regularizer λ_n, parameter α_n, Models $\mathcal{M} = \{(\mu,\phi) : \mu \in \Psi, \phi \in \Phi\}$, Iteration N
Initialize $\pi_0(\cdot \mid s)$ to be uniform; set $\mathcal{D}_0 = \emptyset$, $\mathcal{D}'_0 = \emptyset$
for episode $n = 1, \cdots, N$ **do**
  Collect a tuple $(s, a, s', a', \tilde{s})$ with

$$s \sim d_{P^\star}^{\pi_{n-1}}, a \sim U(\mathcal{A}), s' \sim P^\star(\cdot|s,a), a' \sim U(\mathcal{A}), \tilde{s} \sim P^\star(\cdot|s',a')$$

  Update datasets by adding triples (s,a,s') and $(s',a',\tilde{s})$:

$$\mathcal{D}_n = \mathcal{D}_{n-1} + \{(s,a,s')\}, \quad \mathcal{D}'_n = \mathcal{D}'_{n-1} + \{(s',a',\tilde{s})\}$$

  Learn representation via ERM (i.e., MLE):

$$\hat{P}_n := (\hat{\mu}_n, \hat{\phi}_n) = \underset{(\mu,\phi)\in\mathcal{M}}{\arg\max}\, \mathbb{E}_{\mathcal{D}_n+\mathcal{D}'_n}\left[\ln \mu^\top(s')\phi(s,a)\right]$$

  Update empirical covariance matrix $\hat{\Sigma}_n = \sum_{s,a\in\mathcal{D}_n} \hat{\phi}_n(s,a)\hat{\phi}_n(s,a)^\top + \lambda_n I$
  Set the exploration bonus:

$$\hat{b}_n(s,a) := \min\left(\alpha_n\sqrt{\hat{\phi}_n(s,a)^\top\hat{\Sigma}_n^{-1}\hat{\phi}_n(s,a)}, 2\right) \tag{1}$$

  Update policy $\pi_n = \arg\max_\pi V^\pi_{\hat{P}_n, r+\hat{b}_n}$
end for
Return $\pi_1, \cdots, \pi_N$

Useful auxiliary lemmas

- Concentration for covariance

Lemma 11 (Concentration of the bonus term). *Set $\lambda_n = \Theta(d \ln(n|\Phi|/\delta))$ for any n. Define*

$$\Sigma_{\rho_n,\phi} = n\mathbb{E}_{s\sim\rho_n, a\sim U(\mathcal{A})}[\phi(s,a)\phi^\top(s,a)] + \lambda_n I, \quad \hat{\Sigma}_{n,\phi} = \sum_{i=0}^{n-1} \phi(s^{(i)}, a^{(i)})\phi^\top(s^{(i)}, a^{(i)}) + \lambda_n I.$$

With probability $1-\delta$, we have

$$\forall n \in \mathbb{N}^+, \forall \phi \in \Phi, c_1\|\phi(s,a)\|_{\Sigma^{-1}_{\rho_n \times U(\mathcal{A}),\phi}} \leq \|\phi(s,a)\|_{\hat{\Sigma}^{-1}_{n,\phi}} \leq c_2\|\phi(s,a)\|_{\Sigma^{-1}_{\rho_n \times U(\mathcal{A}),\phi}}.$$

Algorithm 1 UCB-driven representation learning, exploration, and exploitation (REP-UCB)

1: **Input:** Regularizer λ_n, parameter α_n, Models $\mathcal{M} = \{(\mu,\phi) : \mu \in \Psi, \phi \in \Phi\}$, Iteration N
2: Initialize $\pi_0(\cdot \mid s)$ to be uniform; set $\mathcal{D}_0 = \emptyset$, $\mathcal{D}'_0 = \emptyset$
3: **for** episode $n = 1, \cdots, N$ **do**
4: Collect a tuple $(s, a, s', a', \tilde{s})$ with

$$s \sim d^{\pi_{n-1}}_{P^\star}, a \sim U(\mathcal{A}), s' \sim P^\star(\cdot|s,a), a' \sim U(\mathcal{A}), \tilde{s} \sim P^\star(\cdot|s',a')$$

5: Update datasets by adding triples (s,a,s') and $(s',a',\tilde{s})$:

$$\mathcal{D}_n = \mathcal{D}_{n-1} + \{(s,a,s')\}, \quad \mathcal{D}'_n = \mathcal{D}'_{n-1} + \{(s',a',\tilde{s})\}$$

Learn representation via ERM (i.e., MLE):

$$\hat{P}_n := (\hat{\mu}_n, \hat{\phi}_n) = \arg\max_{(\mu,\phi)\in\mathcal{M}} \mathbb{E}_{\mathcal{D}_n + \mathcal{D}'_n}\left[\ln \mu^\top(s')\phi(s,a)\right]$$

Update empirical covariance matrix $\hat{\Sigma}_n = \sum_{s,a\in\mathcal{D}_n} \hat{\phi}_n(s,a)\hat{\phi}_n(s,a)^\top + \lambda_n I$
Set the exploration bonus:

$$\hat{b}_n(s,a) := \min\left(\alpha_n\sqrt{\hat{\phi}_n(s,a)^\top\hat{\Sigma}_n^{-1}\hat{\phi}_n(s,a)}, 2\right) \tag{1}$$

9: Update policy $\pi_n = \arg\max_\pi V^\pi_{\hat{P}_n, r+\hat{b}_n}$
10: **end for**
11: **Return** $\pi_1, \cdots, \pi_N$

Performance difference lemma is always useful

- Substitute the two lemmas into the performance difference lemma then we got

$$V^{\pi}_{\hat{P}_n, r+\hat{b}_n} - V^{\pi}_{P^\star, r} = (1-\gamma)^{-1}\mathbb{E}_{s,a\sim d^{\pi}_{\hat{P}_n}}\left[\hat{b}_n(s,a) + \gamma\mathbb{E}_{s'\sim\hat{P}_n(\cdot|s,a)}V^{\pi}_{P^\star}(s') - \gamma\mathbb{E}_{s'\sim P^\star(\cdot|s,a)}V^{\pi}_{P^\star}(s')\right]$$
$$\geq (1-\gamma)^{-1}\mathbb{E}_{s,a\sim d^{\pi}_{\hat{P}_n}}\left[\hat{b}_n(s,a) - \|\hat{P}_n(\cdot|s,a) - P^\star(\cdot|s,a)\|_1\right],$$

$$\hat{b}_n(s,a) := \min\left(\alpha_n\sqrt{\hat{\phi}_n(s,a)^\top\hat{\Sigma}_n^{-1}\hat{\phi}_n(s,a)}, 2\right)$$

Remark:

- The bonus term is defined using elliptical potential function under $\hat{\phi}$
- We need to bound $|| \hat{P} - P^* ||$ with the elliptical potential function, this is the following lemma

Relate model error with potential function

Lemma 12 (One-step back inequality for the learned model). *Take any $g \in \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that $\|g\|_\infty \le B$. We condition on the event where the MLE guarantee (17):*

$$\mathbb{E}_{s\sim\rho_n, a\sim U(\mathcal{A})}[f_n(s,a)] \lesssim \zeta_n,$$

holds. Then, for any policy π, we have

$$\begin{aligned}
&|\mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{P}_n}}\{g(s,a)\}| \\
&\le \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^{\pi}_{\hat{P}_n}} \|\hat{\phi}_n(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_n\times U(\mathcal{A}),\hat{\phi}_n}} \sqrt{\{n|\mathcal{A}|\mathbb{E}_{s\sim\rho'_n, a\sim U(\mathcal{A})}[g^2(s,a)]\} + B^2\lambda_n d + nB^2\zeta_n} \\
&+ \sqrt{(1-\gamma)|\mathcal{A}|\mathbb{E}_{s\sim\rho_n, a\sim U(\mathcal{A})}[g^2(s,a)]}.
\end{aligned}$$

Recall $\Sigma_{\rho_n\times U(\mathcal{A}),\hat{\phi}_n} = n\mathbb{E}_{s\sim\rho_n, a\sim U(\mathcal{A})}[\hat{\phi}_n(s,a)\hat{\phi}_n^\top(s,a)] + \lambda_n I.$

Remark:

- Instantiate this lemma with $g = \|\hat{P}(\cdot|s,a) - P^*(\cdot|s,a)\|_1$ and $B = 2$
- This lemma is quite hard to prove, it requires both formulations of D_n, D_n'
- I think it's an issue to extend this result to risk sensitive setting

Back to regret: Optimism

$$V^{\pi}_{\hat{P}_n, r+\hat{b}_n} - V^{\pi}_{P^\star, r} = (1-\gamma)^{-1}\mathbb{E}_{s,a\sim d^{\pi}_{\hat{P}_n}}\left[\hat{b}_n(s,a) + \gamma\mathbb{E}_{s'\sim\hat{P}_n(\cdot|s,a)}V^{\pi}_{P^\star}(s') - \gamma\mathbb{E}_{s'\sim P^\star(\cdot|s,a)}V^{\pi}_{P^\star}(s')\right]$$

$$\geq (1-\gamma)^{-1}\mathbb{E}_{s,a\sim d^{\pi}_{\hat{P}_n}}\left[\hat{b}_n(s,a) - \|\hat{P}_n(\cdot|s,a) - P^\star(\cdot|s,a)\|_1\right],$$

Use the previous lemma to Relate model error with potential function

$$\hat{b}_n(s,a) := \min\left(\alpha_n\sqrt{\hat{\phi}_n(s,a)^\top\hat{\Sigma}_n^{-1}\hat{\phi}_n(s,a)}, 2\right)$$

$$\min\left(\alpha_n\|\hat{\phi}_n(s,a)\|_{\Sigma^{-1}_{\rho_n\times U(\mathcal{A}),\hat{\phi}_n}} + \sqrt{(1-\gamma)|\mathcal{A}|\zeta_n}, 2\right)$$

Finally we have:

Lemma 5 (Almost Optimism at the Initial State Distribution). *Set the parameters as in Theorem 4. With probability $1-\delta$,*

$$\forall n \in [1,\cdots,N], \forall\pi\in\Pi, V^{\pi}_{\hat{P}_n, r+\hat{b}_n} - V^{\pi}_{P^\star, r} \geq -c_1\sqrt{\frac{|\mathcal{A}|\ln(|\mathcal{M}|n/\delta)(1-\gamma)^{-1}}{n}}.$$

Remark: From this result, we can see that the bonus term b is designed to provide near-optimism.

Regret

$$
\begin{aligned}
& V^{\pi^\star}_{P^\star,r} - V^{\pi_n}_{P^\star,r} \\
& \le V^{\pi^\star}_{\hat{P}_n, r+\hat{b}_n} - V^{\pi_n}_{P^\star,r} + \sqrt{|\mathcal{A}|\zeta_n(1-\gamma)^{-1}} && \text{(Lemma 8)} \\
& \le V^{\pi_n}_{\hat{P}_n, r+\hat{b}_n} - V^{\pi_n}_{P^\star,r} + \sqrt{|\mathcal{A}|\zeta_n(1-\gamma)^{-1}} && (\pi_n = \arg\max_\pi V^{\pi}_{\hat{P}_n, r+\hat{b}_n}) \\
& = (1-\gamma)^{-1}\mathbb{E}_{(s,a)\sim d^{\pi_n}_{P^\star}}[\hat{b}_n(s,a) + \gamma \mathbb{E}_{\hat{P}_n(s'|s,a)}[V^{\pi_n}_{\hat{P}_n, r+\hat{b}_n}(s')] - \gamma \mathbb{E}_{P^\star(s'|s,a)}[V^{\pi_n}_{\hat{P}_n, r+\hat{b}_n}(s')]] + \sqrt{|\mathcal{A}|\zeta_n(1-\gamma)^{-1}}.
\end{aligned}
$$

Optimism

Relate model error with potential function under ϕ^*

Algorithm 1 UCB-driven representation learning, exploration, and exploitation (REP-UCB)

Input: Regularizer λ_n, parameter α_n, Models $\mathcal{M} = \{(\mu, \phi) : \mu \in \Psi, \phi \in \Phi\}$, Iteration N
Initialize $\pi_0(\cdot \mid s)$ to be uniform; set $\mathcal{D}_0 = \emptyset$, $\mathcal{D}'_0 = \emptyset$
for episode $n = 1, \cdots, N$ **do**
  Collect a tuple $(s, a, s', a', \tilde{s})$ with

$$s \sim d^{\pi_{n-1}}_{P^\star}, a \sim U(\mathcal{A}), s' \sim P^\star(\cdot|s,a), a' \sim U(\mathcal{A}), \tilde{s} \sim P^\star(\cdot|s', a')$$

  Update datasets by adding triples (s, a, s') and $(s', a', \tilde{s})$:

$$\mathcal{D}_n = \mathcal{D}_{n-1} + \{(s,a,s')\}, \quad \mathcal{D}'_n = \mathcal{D}'_{n-1} + \{(s', a', \tilde{s})\}$$

  Learn representation via ERM (i.e., MLE):

$$\hat{P}_n := (\hat{\mu}_n, \hat{\phi}_n) = \arg\max_{(\mu,\phi)\in\mathcal{M}} \mathbb{E}_{\mathcal{D}_n + \mathcal{D}'_n}\left[\ln \mu^\top(s')\phi(s,a)\right]$$

  Update empirical covariance matrix $\hat{\Sigma}_n = \sum_{s,a\in\mathcal{D}_n} \hat{\phi}_n(s,a)\hat{\phi}_n(s,a)^\top + \lambda_n I$
  Set the exploration bonus:

$$\hat{b}_n(s,a) := \min\left(\alpha_n \sqrt{\hat{\phi}_n(s,a)^\top \hat{\Sigma}_n^{-1} \hat{\phi}_n(s,a)}, 2\right) \tag{1}$$

Policy update

  Update policy $\pi_n = \arg\max_\pi V^{\pi}_{\hat{P}_n, r+\hat{b}_n}$
end for
Return $\pi_1, \cdots, \pi_N$

Finally we get some terms like: $||\phi^*||_{\Sigma^{-1}}$, which are controlled by some data-dependent quantities. Q.E.D.

Back to risk sensitive RL

- For each threshold b, we have

$$V^{\pi}(s,b) = E\left[\max(0,\ b\ -\ \textstyle\sum_{t=0}^{\infty}\gamma^{t}\ r(s_t,a_t,b_t))\,|\ s_0 = s,\pi,b\ \right]$$

$$J^*(b) = \min_{\pi} J(b) = \min_{\pi} E_{s\sim d_0} V^{\pi}(s,b)$$

The final goal is to maximize CVaR, i.e., $\max_{b}(b - \frac{1}{\tau}J^*(b))$

Another loop for threshold b ?

Algorithm 1 UCB-drivern CVaR learning

Input: Regularizer λ_n, parameter α_n, models $\mathcal{M} = \{(\mu, \phi) : \mu \in \Psi, \phi \in \Phi\}$, number of iterations N.

for $b \in \mathcal{B}$ **do**

  for episode $n = 1, \ldots, N$ **do**

    Collect a tuple $(s, a, s', a', \tilde{s})$

    Update datasets: $\mathcal{D}_n = \mathcal{D}_{n-1} + \{(s, a, s')\}$, $\mathcal{D}'_n = \mathcal{D}'_{n-1} + \{(s', a', \tilde{s})\}$

    Learn representations via MLE

$$\hat{P}_n = \arg \max_{\mu, \phi \in \mathcal{M}} \mathbb{E} \ln \mu^\top(s')\phi(s, a)$$

    Update empirical covariance $\hat{\sum}_n = \sum_{s,a \in \mathcal{D}_n} \hat{\phi}_n(s, a)\hat{\phi}_n(s, a)^\top + \lambda_n I_n$

    Set the exploration bonus:

$$\hat{b}_n = ???$$

    Update policy $\pi_n = \arg\max_\pi V_{\hat{P}_n, r + \hat{b}_n}$

  end for

  Store $\pi_1(|b), \cdots, \pi_N(|b)$

end for

Output: $b = \arg\max_b (b - \frac{1}{\tau}\mathbb{E}_{s \sim d_0} V^\pi(s, b))$ and $\pi_1(|b), \cdots, \pi_N(|b)$

Final remarks

- Concentrations should still hold
- Performance difference lemma changes because of the threshold parameter, recall that $V^{\pi}(s,b) = E\ [\max(0,\ b\ -\ \sum_{t=0}^{\infty}\gamma^t\ r(s_t,a_t,b_t))|\ s_0 = s,\pi,b\ \ )]$
- Optimism results does not apply, need new techniques
- The relation between model error and potential function is built upon the MDP transition which still holds, might not change much but need double check.