FROM API TO POLITEX

CMPUT653: THEORETICAL FOUNDATIONS OF RL

TABLE OF CONTENTS

PSEUDO-ALGORITHM OF LSPI-G

1. Given the feature map ϕ, find $\mathcal{C}$ and ρ.
2. Let $\theta_{-1} = 0$
3. For $k = 0, 1, 2, \ldots, K-1$ do
4. Let π_k be a greedy policy wrt $\Phi\theta_{k-1}$
5. For each $z \in \mathcal{C}$ do
6. Get rollouts with π_k for H steps from z
7. Compute return estimate $\hat{R}_m(z)$
8. $\theta_k = G_\varrho^{-1} \sum_{z\sim\rho} \varrho(z)\hat{R}_m(z)\phi(z)$
9. Return a greedy policy wrt $\Phi\theta_{K-1}$

Recall that $G_\varrho = \sum_{z\in\mathcal{C}} \phi(z)\phi(z)^\top$.

PERFORMANCE GUARANTEE

For any MDP feature-map pair (M, ϕ) and any $\varepsilon' > 0$,

LSPI-G can produce a policy π such that its suboptimality gap δ satisfies

$$\delta \leq \frac{2(1+\sqrt{d})}{(1-\gamma)^{2}} \tilde{\varepsilon}(M, \phi) + \varepsilon' ,$$

where $\tilde{\varepsilon}(M, \phi) = \sup_{\pi \in \Pi_\phi} \inf_\theta ||\Phi\theta - q^\pi||_\infty$

with a total runtime of

$$\text{poly}\left(d, \frac{1}{1-\gamma}, A, \frac{1}{\varepsilon'}\right).$$

We saw $\sqrt{d}$ is inevitable

but

is $\frac{1}{(1-\gamma)^2}$ also inevitable?

YES 😭

Theorem (LSPI error amplification lower bound)

For every $\gamma \in [0,1)$,

there is a featurized MDP (M,ϕ), its policy π_0, and a distribution μ over $\mathcal{S}$ s.t.

LSPI *with access to true Q-functions* produces a sequence of policies $\pi_0, \pi_1, ...$ satisfying

$$\mu v^* - \frac{c\tilde{\varepsilon}(M,\phi)}{(1-\gamma)^2} \geq \sup_{k \geq 0} \mu v^{\pi_k},$$

where c is a universal constant independent of other variables.

STATE-AGGREGATION

Suppose that $\mathcal{S} = \{1, \ldots, S\}$, $\mathcal{A} = \{1, \ldots, A\}$, and $\mathcal{S}$ has a partition $\{\mathcal{S}_i\}_{i=1}^d$.

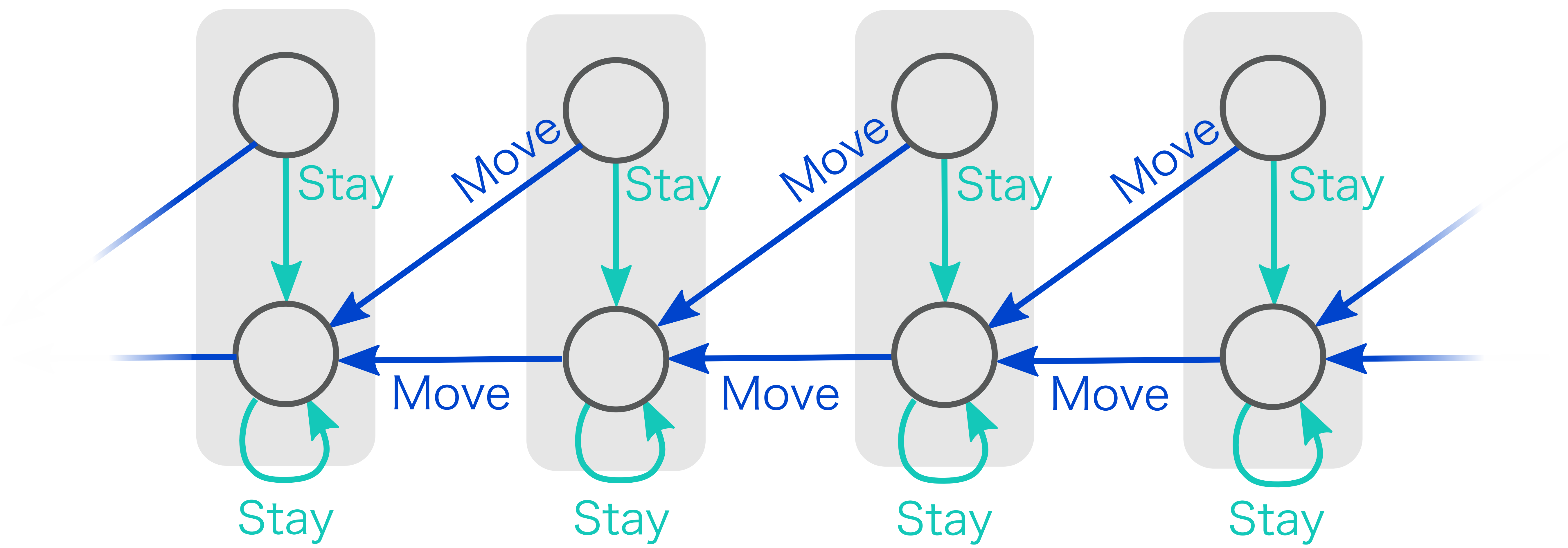
State-aggregation is the following feature map:

$$\phi_j(s, a) = \mathbb{I}\{s \in \mathcal{S}_{\mathrm{ceil}(j/A)}\}\mathbb{I}\{\mathrm{rem}(j-1, A) + 1 = a\},$$

where $\phi(s, a) \in \mathbb{R}^{Ad}$ and $\mathrm{rem}(x, y)$ is the remainder of x/y.

```python
import numpy as np

A = 3; d = 2

phi = np.arange(A * d) + 1 # [ 1, ..., 6])

ceil = np.ceil(phi / A).astype(np.int) # [1, 1, 1, 2, 2, 2]
rem_1 = np.remainder(phi - 1, A) + 1 # [1, 2, 3, 1, 2, 3]

Is = ceil == 2 # [0, 0, 0, 1, 1, 1]
Ia = rem_1 == 1 # [1, 0, 0, 1, 0, 0]
Is * Ia # [0, 0, 0, 1, 0, 0]
```

PROOF IDEA


Stay
Move
Stay
Move
Stay
Move
Stay
Move
Move
Move
Stay
Stay
Stay
Stay

OK, so $\frac{1}{(1-\gamma)^2}$ is real.

Can we avoid it with a different algorithm?

Yes! Politex can!

PSEUDO-ALGORITHM OF POLITEX

1. Given the feature map ϕ, find $\mathcal{C}$ and ρ.
2. Let $\theta_{-1} = 0$ and $\bar{q}_{-1} = \hat{q}_{-1} = \Pi\Phi\theta_{-1}$
3. For $k = 0, 1, 2, \ldots, K-1$ do
4. Let π_k be a Boltzmann policy induced by $\bar{q}_{k-1}$
5. For each $z \in \mathcal{C}$ do
6. Get rollouts with π_k for H steps from z
7. Compute return estimate $\hat{R}_m(z)$
8. $\theta_k = G_\varrho^{-1} \sum_{z\sim\rho} \varrho(z)\hat{R}_m(z)\phi(z)$
9. Let $\hat{q}_k = \Pi\Phi\theta_k$ and $\bar{q}_k = \bar{q}_{k-1} + \hat{q}_k$
10. Return all policies $(\pi_k)_{k=0}^{K-1}$

SUBTLE BUT IMPORTANT CHANGE

$$\pi_k(a|s) = \frac{E_k(s,a)}{\sum_{b\in\mathcal{A}} E_k(s,b)}$$

where

$$E_k = \exp\left(\eta \bar{q}_{k-1}\right) = \exp\left(\eta \sum_{j=0}^{k-1} \hat{q}_j\right)$$

This seemingly simple modification has an interesting connection to online algorithms.

WHY POLITEX PERFORMS BETTER?

$$v^* - v^{\bar{\pi}_k} = \underbrace{\frac{1}{k}(I - \gamma P_{\pi^*}) \sum_{j=0}^{k-1} (M_{\pi^*} - M_{\pi_j}) \hat{q}_j}_{T_1}$$

$$+ \underbrace{\frac{1}{k}(I - \gamma P_{\pi^*}) \sum_{j=0}^{k-1} (M_{\pi^*} - M_{\pi_j})(q^{\pi_j} - \hat{q}_j)}_{T_2 \le \frac{2}{1-\gamma} \max_j \|q^{\pi_j} - \hat{q}_j\|_\infty}$$

NOTES

STATIONARY POINTS OF A POLICY SEARCH OBJECTIVE

Let $J(\pi) = \mu v^\pi$. A stationary point of J with respect to some set of memoryless policies Π is any $\pi \in \Pi$ such that

$$\langle \nabla J(\pi), \pi' - \pi \rangle \leq 0, .$$

If ϕ are state-aggregation features, any stationary point π satisfies

$$\mu v^\pi \geq \mu v^* - \frac{4\varepsilon_{\mathrm{apx}}}{1-\gamma},$$

where $\varepsilon_{\mathrm{apx}} = \sup_{\pi \in \Pi_\phi} \inf_\theta \|\Phi\theta - q^\pi\|_\infty$.

LAST-ITERATE CONVERGENCE OF POLITEX?

$$v^* - v^{\pi_k} = \underbrace{v^* - \frac{1}{k} M_{\pi_k} \bar{q}_{k-1}}_{T_1} + \underbrace{\frac{1}{k} M_{\pi_k} \bar{q}_{k-1} - v^{\pi_k}}_{T_2}$$

Key: A Boltzmann policy is entropy regularized greedy policy!

$$
\begin{aligned}
T_1 &= v^* - \frac{1}{k} M_{\pi_k} \bar{q}_{k-1} \\
&\le M_{\pi^*} \left(q^* - \frac{1}{k} \bar{q}_{k-1} \right) + \frac{\log A}{\eta}
\end{aligned}
$$

Then, we can use the almost same argument we had before.

$$
\begin{aligned}
T_2 &= \frac{1}{k} M_{\pi_k} \bar{q}_{k-1} - v^{\pi_k} \\
&= \frac{1}{k} (I - \gamma P_{\pi_k})^{-1} M_{\pi_k} \left(\bar{q}_{k-1} - kr - \gamma P_{\pi_k} \bar{q}_{k-1}\right)
\end{aligned}
$$

Recall that $\bar{q}_{k-1} \approx \sum_{j=0}^{k-1} q^{\pi_j}$, so

$$
\begin{aligned}
&\bar{q}_{k-1} - kr - \gamma P_{\pi_k} \bar{q}_{k-1} \\
&\qquad \approx \gamma P \sum_{j=0}^{k-1} \left(M_{\pi_j} - M_{\pi_k}\right) \hat{q}_j \leq 0
\end{aligned}
$$

STATE AGGREGATION AND EXTRAPOLATION FRIENDLINESS

Recall LSPI-G's performance upper-bound.

$$\delta \leq \frac{2(1+\sqrt{d})}{(1-\gamma)^2}\tilde{\varepsilon}(M,\phi)+\varepsilon' \,.$$

What's the source of this $\sqrt{d}$?

BOUNDING EXTRAPOLATION ERROR

$$\begin{aligned}&\left|\phi(z)^\top \hat{\theta} - \phi(z)^\top \theta\right| \\ &\leq \left(\max_{z' \in C} |\varepsilon(z')|\right) \underbrace{\sum_{z' \in C} \varrho(z') |\phi(z)^\top G_\varrho^{-1} \phi(z')|}_{\leq \sqrt{d}}.\end{aligned}$$

BETTER DESIGN FOR STATE-AGGREGATION

Pick up one element s_i from each $\mathcal{S}_i$ and let

$$\rho(s,a) = \frac{1}{Ad}\sum_{i=1}^{d} \mathbb{I}\{s = s_i\}$$

$$\mathcal{C} = \{(s,a) : s \in \{s_1, \ldots, s_d\}, a \in \mathcal{A}\}$$

Then, $G_\varrho = \dfrac{1}{Ad} I,$

and $\phi(s,a)^\top \phi(s',a') = 1$ iff $s, s' \in \mathcal{S}_i$ and $a = a'$.

$$
\begin{aligned}
\implies \sum_{z' \in C} \varrho(z') |\phi(z)^\top G_\varrho^{-1} \phi(z')| \\
= \sum_{i=1}^{d} \sum_{a' \in \mathcal{A}} \mathbb{I}\{s' = s_i\} |\phi(s, a)^\top \phi(s', a')| = 1
\end{aligned}
$$

So, we don't have $\sqrt{d}$ anymore!

LSVI-G

For least-squares value iteration with G-optimal design (LSVI-G), a result performance guarantee holds.

Concretely, it can produce a policy π such that the suboptimality gap δ of π satisfies

$$\delta \leq \frac{4(1+\sqrt{d})}{(1-\gamma)^{2}} \varepsilon_{\mathrm{BOO}} + \varepsilon' ,$$

where

$$\varepsilon_{\mathrm{BOO}} := \sup_{\theta} \inf_{\theta'} \|\Phi\theta' - T\Pi\Phi\theta\|_{\infty} .$$

LINEAR MDPS

An MDP-feature-map pair (M, ϕ) is said to be approximately linear if $\exists \zeta_r, \zeta_P \in \mathbb{R}_+$ s.t.

$$\zeta_r = \inf_\theta \|\Phi\theta_r - r\|_\infty \text{ and } \zeta_P = \inf_W \|\Phi W - P\|_\infty .$$

If these hold, for any π and $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,

$$\begin{aligned}
&\inf_\theta \|r + \gamma P M_\pi f - \Phi\theta\|_\infty \\
&\leq \inf_{\theta_r} \|r - \Phi\theta_r\|_\infty + \gamma \inf_W \|P - \Phi W\|_\infty \|f\|_\infty \\
&\leq \zeta_r + \gamma \zeta_P \|f\|_\infty .
\end{aligned}$$

$$\implies \varepsilon_{\text{BOO}} \leq \zeta_r + \frac{\gamma \zeta_P}{1-\gamma}$$
$$\varepsilon \leq \zeta_r + \frac{\gamma \zeta_P}{1-\gamma}$$

That's all! Any question?