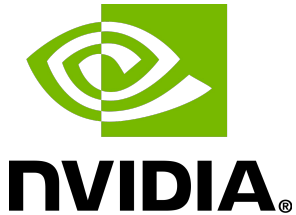
Spectral Bellman Method

Ofir Nabati, Bo Dai, Shie Mannor, Guy Tennenholtz



Motivation

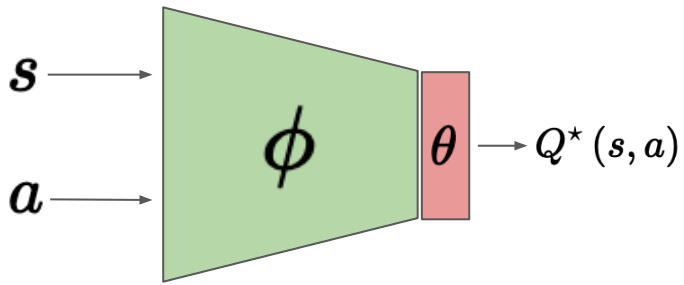
Linear representation works great in a contextual bandit setting for **reward models.**

Why not using it in general RL for the **optimal Q-function?**

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t) | s_0 = s, a_0 = a\right]$$

$$Q^{\star}(s,a) = \max_{\pi} Q^{\pi}(s,a) = \phi(s,a)^{\top}\theta$$

$$\phi(s,a) \in \mathbb{R}^d$$



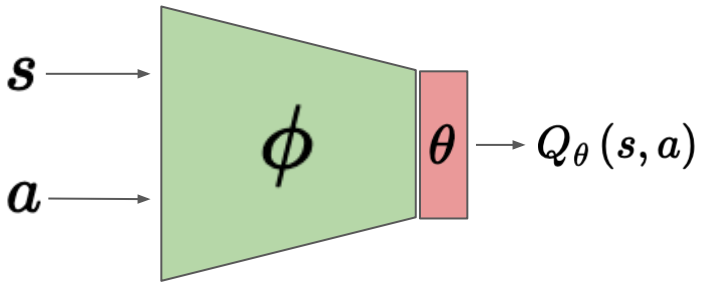
Motivation

Linear representation works great in a contextual bandit setting for **reward models.**

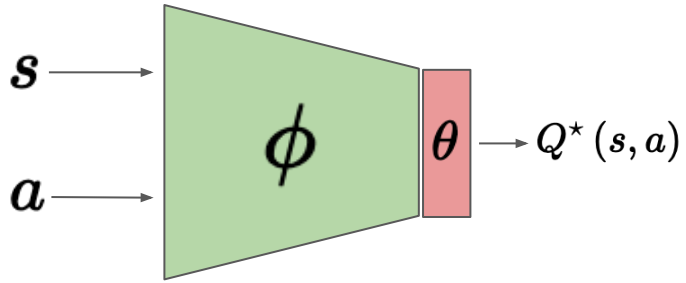
Why not using it in general RL for the **optimal Q-function?**

$$Q^{\star}(s,a) = \max_{\pi} Q^{\pi}(s,a) = \phi(s,a)^{\top}\theta$$

$$\phi(s,a) \in \mathbb{R}^{d}$$



Motivation



It is not enough...

Even with a perfect representation (up tp scale) w.r.t optimal Q-function:

$$Q^\star(s,a) \propto \phi(s,a)^\top \theta$$

We will still need an **exponential number** of trajectories to find a near-optimal policy!

"Is a Good Representation Sufficient for Sample Efficient Reinforcement Learning?", Du et al. (ICLR 2020)

Motivation

The optimal Bellman operator:

$$\mathcal{T}Q(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(s'|s,a)}[\max_{a' \in \mathcal{A}} Q(s',a')]$$

Define a linear Q-function: $Q_\theta(s,a) = \phi(s,a)^\top \theta$

The function space of linear Q-functions:

$$\mathcal{Q}_\phi = \{Q_\theta(s,a) = \phi(s,a)^\top \theta \mid \theta \in \mathcal{B}_\phi\}, \quad \mathcal{B}_\phi = \{\theta \in \mathbb{R}^d \mid \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} |\phi(s,a)^\top \theta| \leq D\}$$

Motivation

The Inherent Bellman Error (IBE):

$$\mathcal{I}_\phi := \sup_{Q \in \mathcal{Q}_\phi} \inf_{\tilde{Q} \in \mathcal{Q}_\phi} \left\| \mathcal{T}Q - \tilde{Q} \right\|_\infty = \sup_{\theta \in \mathcal{B}_\phi} \inf_{\tilde{\theta} \in \mathcal{B}_\phi} \left\| \mathcal{T}Q_\theta - Q_{\tilde{\theta}} \right\|_\infty$$

When IBE=0, $\mathcal{Q}_\phi$ is closed under the Bellman operator.

With a low IBE the regret :

$$\tilde{\mathcal{O}}(dH^{1.5}\sqrt{T} + \sqrt{d}HT\mathcal{I}_\phi)$$

and a low number of episodes for a near-optimal policy!

"Provably Efficient Reward-Agnostic Navigation with Linear Value Iteration", Zanette et al. (NeurIPS 2020)
"Learning Near Optimal Policies with Low Inherent Bellman Error", Zanette et al. (ICML 2020)

Efficient Exploration

- Given:
  - dataset $\mathcal{D} = \{s_i, a_i, r_i, s_i'\}_{i=1}^N$
  - covariance $\Sigma = \lambda I + \sum_{(s,a)\in\mathcal{D}} \phi(s,a)\phi(s,a)^\top$
- Alternate between:
  - Solve: $\hat{\theta}_{LS} \in \arg\min_{\theta\in\mathcal{B}_\phi} \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[(\mathcal{T}Q_\theta(s,a) - \phi(s,a)^\top\theta)^2\right]$
  - Collect rollouts with: $\hat{\theta}_{TS} \sim \mathcal{N}(\hat{\theta}_{LS}, \sigma_{exp}\Sigma^{-1})$ $\pi_{\hat{\theta}_{TS}}(s) = \arg\max_{a\in\mathcal{A}} \phi(s,a)^\top\hat{\theta}_{TS}$
  - Add data to $\mathcal{D}$ and update the covariance.

Naive Approach

How to learn such representation s.t $\mathcal{I}_\phi \approx 0$?

Bellman parameter mapping: $\tilde{\theta}(\theta) := \arg\min_{\tilde{\theta}} \|\mathcal{T}Q_\theta - Q_{\tilde{\theta}}\|_\infty$

Naive objective:

$$\mathcal{L}_{MSE}(\phi, \tilde{\theta}) = \mathbb{E}_{(s,a)\sim\rho(s,a),\theta\sim\nu(\theta)}\left[\left\|\mathcal{T}Q_\theta(s,a) - \phi(s,a)^\top \tilde{\theta}(\theta)\right\|_2^2\right]$$

$\rho(s,a)$ - Distribution over state-action pairs (replay buffer)

$\nu(\theta)$ - Distribution over (linear) Q-functions (over $\mathcal{B}_\phi$)

The Spectral Bellman Method

Consider finite state-action and parameter spaces $|\mathcal{B}_\phi| = m,\ |\mathcal{S} \times \mathcal{A}| = n$

The covariance matrices:

$$\Lambda_1 = \mathbb{E}_{(s,a)\sim\rho}[\phi(s,a)\phi(s,a)^\top] \qquad \Lambda_2 = \mathbb{E}_{\theta\sim\nu}[\tilde{\theta}(\theta)\tilde{\theta}(\theta)^\top]$$

The feature and parameter matrices:

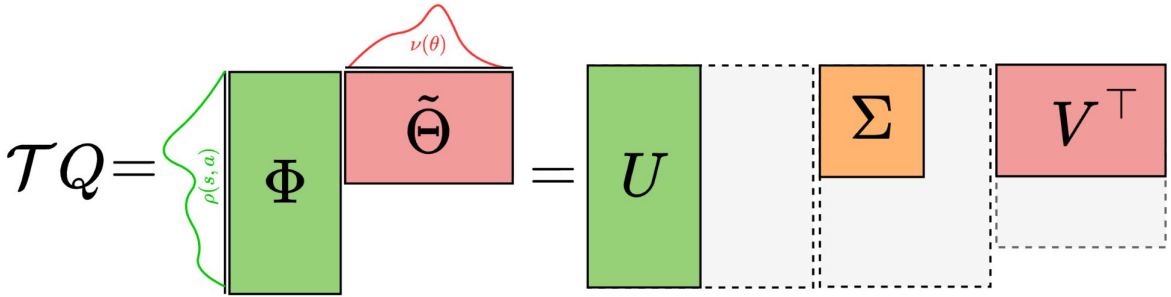
$$\Phi = [\phi_1^\top, \ldots, \phi_n^\top]^\top \in \mathbb{R}^{n\times d}$$

$$\tilde{\Theta} = [\tilde{\theta}_1, \ldots, \tilde{\theta}_m] \in \mathbb{R}^{d\times m}$$

The Spectral Bellman Method

Under the assumption of zero IBE,

the Bellman matrix:



The spectral Bellman method enforce features and parameters to be aligned

with the singular-vectors and singular values: $\Phi = U\Sigma \quad \tilde{\Theta} = V^\top$

The Spectral Bellman Method

We can use the power iteration* to extract Φ !

- Initialize $\Phi_0,\ \tilde{\Theta}_0$
- Repeat for k =1,2,3,...:
  - $\bar{\Theta}_k = \Phi_k^\top \mathcal{T}Q \qquad \bar{\Phi}_k = \mathcal{T}Q\ \tilde{\Theta}_k^\top$
  - Project: $\Phi_{k+1} = Proj\left(\bar{\Phi}_k\right)\ \Theta_{k+1} = Proj\left(\bar{\Theta}_k\right)$

* **Actually its subspace iteration**

The Spectral Bellman Method - Practical

The equivalent objective for solving the power method:

orthogonal constraint

$$\mathcal{L}(\phi, \tilde{\theta}; \nu, \rho) = \mathcal{L}_1(\phi) + \mathcal{L}_2(\tilde{\theta}) \quad s.t \quad \phi \in \mathcal{M}^{\rho}_{\mathcal{S}\times\mathcal{A}}, \tilde{\theta} \in \mathcal{M}^{\nu}_{\mathcal{B}_\phi}$$

$$\mathcal{L}_1(\phi) = \mathbb{E}_{\nu(\theta)\rho(s,a)}\left[\|\phi(s,a)\|^2_{\Lambda_{2,t}} - 2\mathcal{T}Q_{\theta,t}(s,a)\phi(s,a)^\top\tilde{\theta}_t(\theta)\right]$$

$$\mathcal{L}_2(\tilde{\theta}) = \mathbb{E}_{\nu(\theta)\rho(s,a)}\left[\|\tilde{\theta}(\theta)\|^2_{\Lambda_{1,t}} - 2\mathcal{T}Q_{\theta,t}(s,a)\tilde{\theta}(\theta)^\top\phi_t(s,a)\right]$$

Regularization terms

The Spectral Bellman Method - Practical

Compare to the “naive” objective:

$$\mathcal{L}_1(\phi) = \mathbb{E}_{\nu(\theta)\rho(s,a)}\left[\|\phi(s,a)\|^2_{\hat{\Lambda}} - 2\mathcal{T}Q_\theta(s,a)\phi(s,a)^\top\tilde{\theta}(\theta)\right]$$

SBM objective:

$$\mathcal{L}_1(\phi) = \mathbb{E}_{\nu(\theta)\rho(s,a)}\left[\|\phi(s,a)\|^2_{\Lambda_{2,t}} - 2\mathcal{T}Q_{\theta,t}(s,a)\phi(s,a)^\top\tilde{\theta}_t(\theta)\right]$$

The naive objective objective is regularized with a noisy single-sample of the covariance

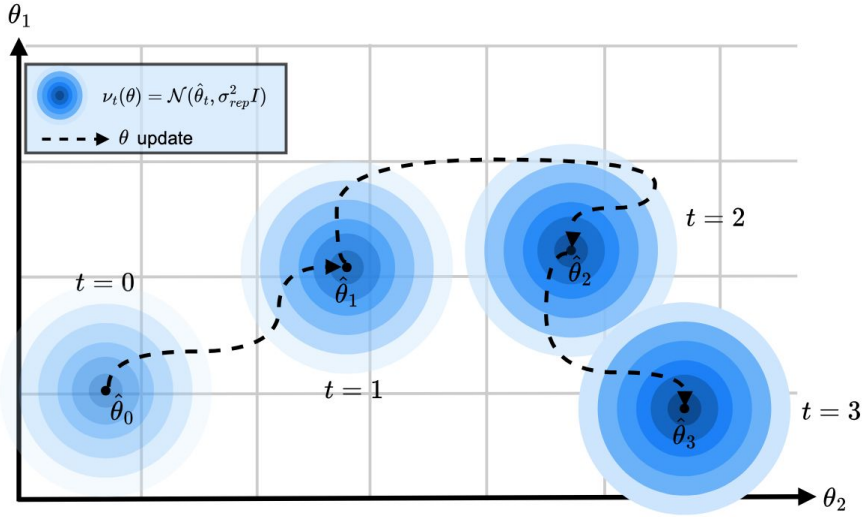
SBM using an average covariance over samples which is robust and less noisy

How to choose ν ?

DQN style learning (targeting only optimal Q-function): $\nu_t(\theta) = \delta(\hat{\theta}_t - \theta)$

SBM theoretical distribution: $\nu_t(\theta) = \mathcal{U}(\mathcal{B}_\phi)$

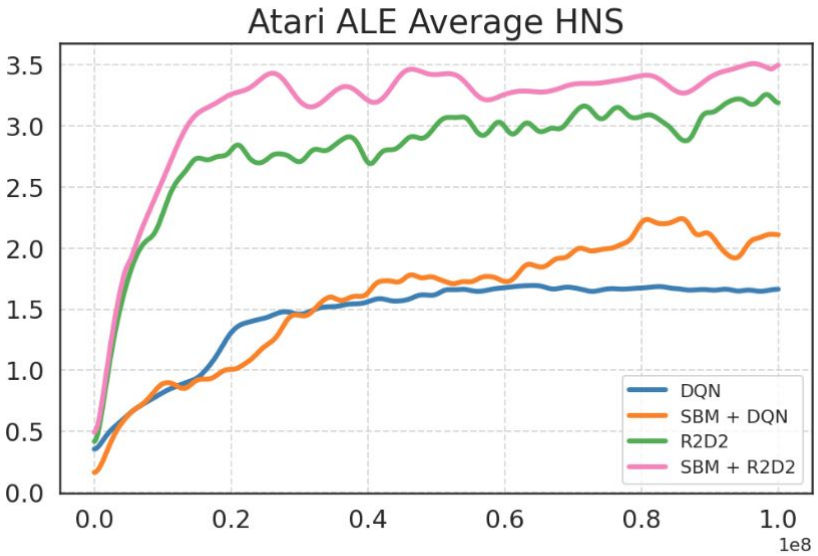
Practical choice: Gaussian distribution around the working point:

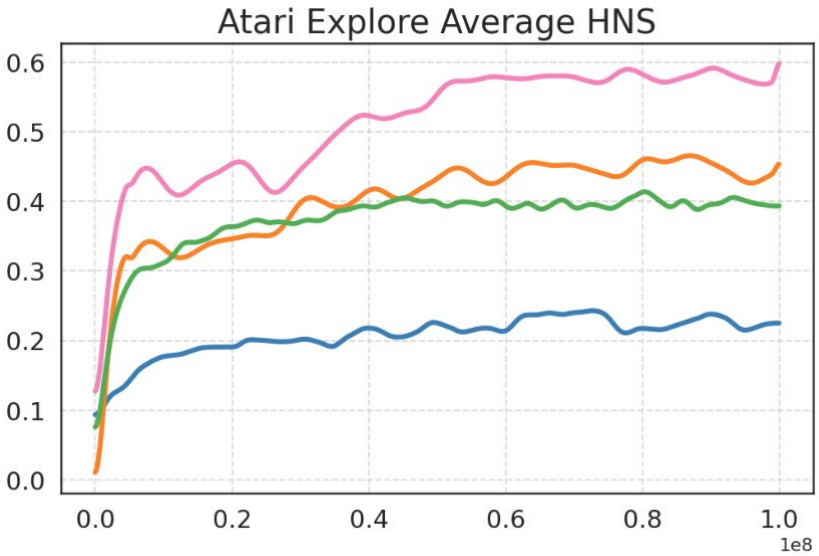


Experiments

- Atari game suite with 100M environment steps.
- Atari Explore: subset of hard exploration games (*Montezuma's Revenge, Pitfall!, Private Eye, Skiing*..)
- SBM + DQN and SBM + R2D2 (extension to multi-step operation)

Experiments





Experiments

Method	Atari ALE		Atari Explore	
	Mean	Median	Mean	Median
DQN (Mnih et al., 2013)	1.62 ± 0.12	0.52	0.24 ± 0.03	0.11
Online PVN (ϵ-greedy)	1.72 ± 0.11	0.60	0.26 ± 0.03	0.21
Online PVN (TS)	1.84 ± 0.15	0.65	0.31 ± 0.04	0.23
SBM + DQN (ϵ-greedy).	1.80 ± 0.13	0.64	0.33 ± 0.03	0.23
SBM + DQN (TS)	$\mathbf{2.23 \pm 0.19}$	**0.85**	$\mathbf{0.45 \pm 0.05}$	**0.24**
R2D2 (Kapturowski et al., 2018)	3.21 ± 0.22	1.14	0.40 ± 0.06	0.22
SBM + R2D2 (ϵ-greedy)	3.3 ± 0.24	1.14	0.45 ± 0.05	0.22
SBM + R2D2 (TS)	$\mathbf{3.53 \pm 0.23}$	**1.37**	$\mathbf{0.61 \pm 0.03}$	**0.30**

Table 1: Aggregated Atari HNS at 100M steps. Our SBM method with TS is in bold.

Method	Atari ALE	Atari Explore
Features from DQN Loss (Azizzadenesheli et al., 2018)	1.73 ± 0.14	0.30 ± 0.03
Features from Naive MSE Loss	1.82 ± 0.12	0.37 ± 0.04
SBM w/o Orthogonality Reg.	2.11 ± 0.11	0.43 ± 0.05
SBM Full	2.23 ± 0.19	0.45 ± 0.05

Table 2: Ablation study on the DQN backbone with TS exploration.