Finite-Sample Analysis of Policy Evaluation for Robust Average Reward Reinforcement Learning

Yang Xu[1], Washim Uddin Mondal[2], Vaneet Aggarwal[1]
[1]Purdue University; [2]Indian Institute of Technology Kanpur

Problem Setup

Motivating question: Can a non-asymptotic result be obtained for model-free algorithms in distributionally robust RL with average-reward objectives?

Setting:

- Distributionally robust average-reward MDP:

$$g_{\mathcal{P}}^{\pi}(s)=\min_{\mathsf{P}\in\mathcal{P}}\lim_{T\to\infty}\mathbb{E}_{\pi,\mathsf{P}}\left[\frac{1}{T}\sum_{t=0}^{T-1}r_t|S_0=s\right]$$

$$V_{\mathcal{P}}^{\pi}(s):=\mathbb{E}_{\pi,P^\star}\left[\sum_{t=0}^{\infty}(r_t-g_{\mathcal{P}}^{\pi})\middle|S_0=s\right],\qquad g_{P^\star}^{\pi}=g_{\mathcal{P}}^{\pi}$$

- Contamination $\mathcal{P}_s^a=\{(1-\delta)\tilde{\mathsf{P}}_s^a+\delta q: q\in\Delta(\mathcal{S})\}$
- Total variation (TV) $\mathcal{P}_s^a=\{q\in\Delta(|\mathcal{S}|):\frac{1}{2}\|q-\tilde{\mathsf{P}}_s^a\|_1\le\delta\}$
- Wasserstein $\mathcal{P}_s^a=\{q\in\Delta(\mathcal{S}):W_l(\tilde{\mathsf{P}}_s^a,q)\le\delta\}$

Robust average-reward Bellman operator [1]:

$$V(s)=\underbrace{\sum_a\pi(a\mid s)\Big(r(s,a)-g+\sigma_{\mathcal{P}_s^a}(V)\Big)}_{\mathbf{T}_g(V)(s)}$$

$$\sigma_{\mathcal{P}_s^a}(V)\triangleq\min_{p\in\mathcal{P}_s^a}p^\top V$$

Goal: Estimating the robust value function and robust average reward for a given policy by only accessing the **ergodic** nominal model. (Under some radius restrictions for the nominal model, all kernels in the set are ergodic.)

Challenges

- Off-policy sampling nature; hindering directly sampling approaches used in non-robust average-reward RL.
- Non-linearity in robust Bellman operator and the absence of a discount factor; complicating the process of establishing some form of negative drift.

Key Methodology

Reformulation: Solving a fixed-point problem in a quotient space

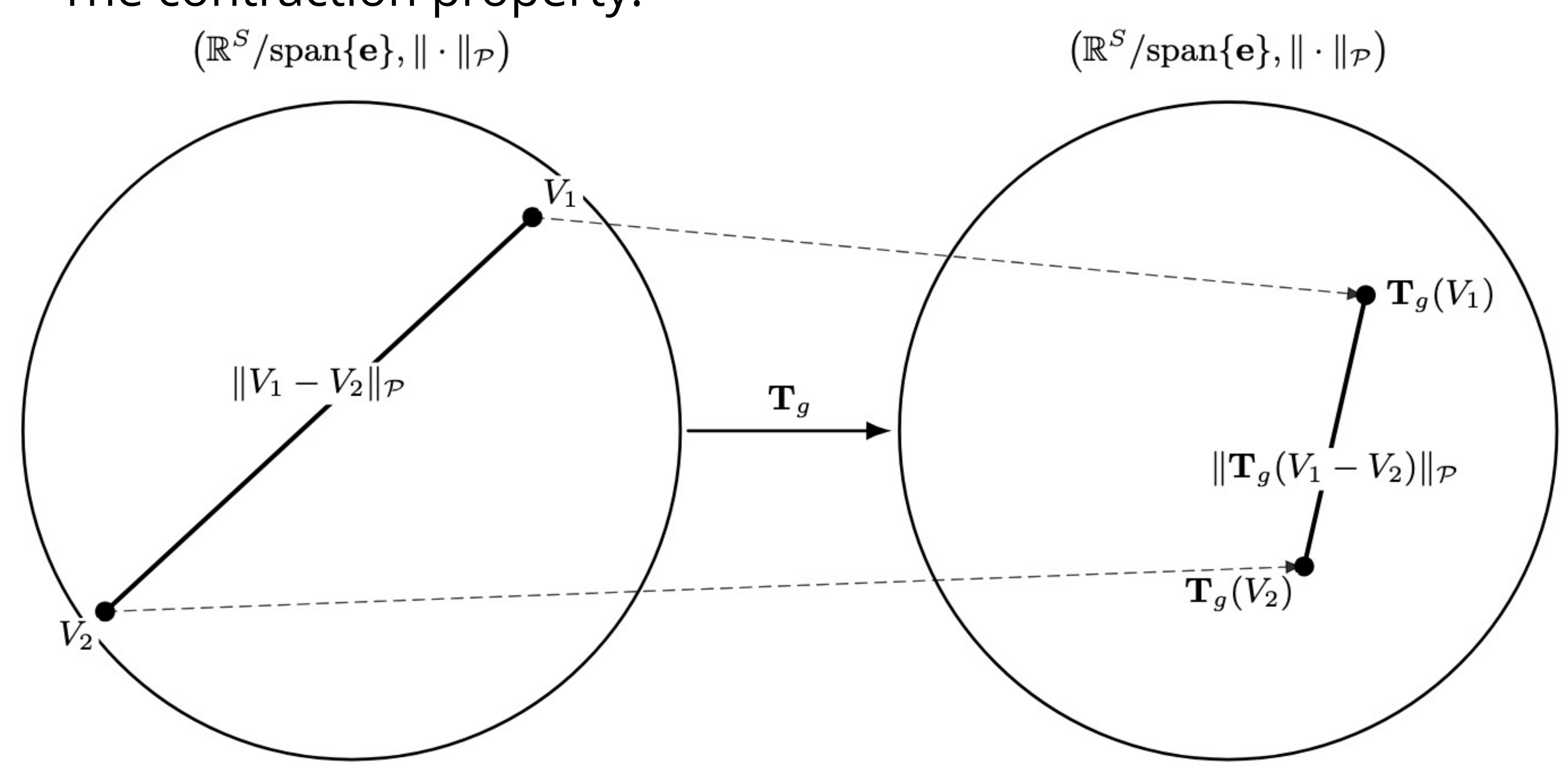
$$H(x)-x\in\bar{E}\text{ where }\bar{E}=\{c\mathbf{e}:c\in\mathbb{R}\}\text{ and }H=\mathbf{T}_g$$

TD-style update: $x^{t+1}\leftarrow x^t+\eta_t\big(\widehat{H}(x^t)-x^t\big)$

[2] states the following assumptions needed to solve the above:

- Unique solution in the quotient space **Done**
- Unbiased noise with bounded variance **Handleable**
- Contraction property for H **Challenging**

The contraction property:



$$\|(V_1-V_2)\|_{\mathcal{P}}<\gamma\|\mathbf{T}_g(V_1-V_2)\|_{\mathcal{P}}\quad\text{for some }0<\gamma<1.$$

The construction of semi-norm $\|\cdot\|_{\mathcal{P}}$:

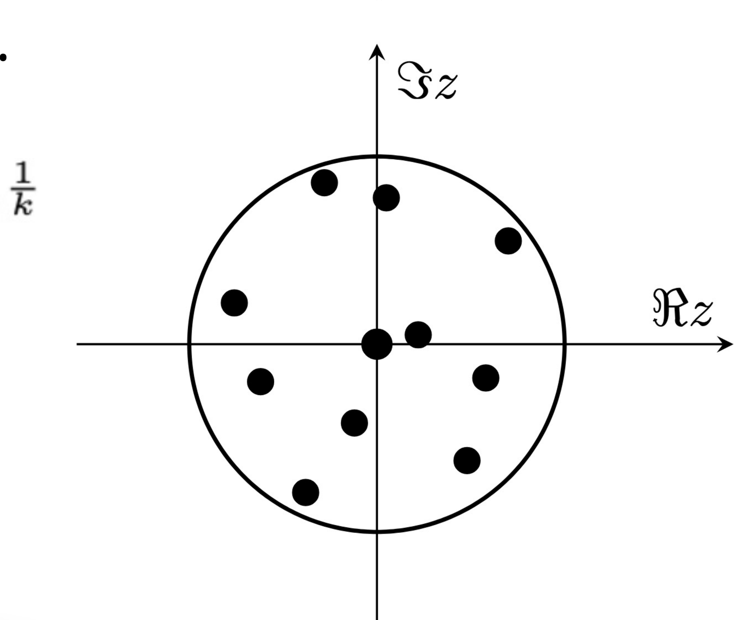
$$\|x\|_{\mathcal{P}}:=\sup_{Q\in\mathcal{Q}}\big\|Qx\big\|_{\mathrm{ext}}+\epsilon\inf_{c\in\mathbb{R}}\big\|x-c\mathbf{e}\big\|_{\mathrm{ext}}$$

$$\|x\|_{\mathrm{ext}}:=\sup_{k\ge0}\sup_{Q_1,\dots,Q_k\in\mathcal{Q}}\alpha^{-k}\|Q_kQ_{k-1}\dots Q_1x\|_2$$

$$\mathcal{Q}:=\{\mathsf{P}^\pi-\mathbf{e}(d_{\mathsf{P}}^{\pi})^\top:\mathsf{P}\in\mathcal{P}\}$$

Intuition: under certain restrictions of the uncertainty radius, the joint spectral radius of $\mathcal{Q}$ is less than 1.

$$\hat{\rho}(\mathcal{Q}):=\lim_{k\to\infty}\sup_{Q_i\in\mathcal{Q}}\rho(Q_k\dots Q_1)^{\frac{1}{k}}$$



Sampling Scheme

Contamination: unbiased estimator

$$\hat{\sigma}_{\mathcal{P}_s^a}(V)\triangleq(1-\delta)V(s')+\delta\min_x V(x)$$

TV and Wasserstein: exponentially decaying bias

$$\sigma_{\mathcal{P}_s^a}(V)=\max_{\mu\ge\mathbf{0}}\left((\tilde{\mathsf{P}}_s^a)^\top(V-\mu)-\delta\|V-\mu\|_{\mathrm{sp}}\right)$$

Wasserstein: exponentially decaying bias

$$\sigma_{\mathcal{P}_s^a}(V)=\sup_{\lambda\ge0}\left(-\lambda\delta^l+\mathbb{E}_{\tilde{\mathsf{P}}_s^a}\Big[\inf_y\big(V(y)+\lambda d(S,y)^l\big)\Big]\right)$$

At most exponentially decaying bias is achievable

Algorithm 1 Truncated MLMC Estimator for TV and Wasserstein Sets

Input: $s\in\mathcal{S}$, $a\in\mathcal{A}$, Max level $N_{\max}$, Value function V

Sample $N\sim\mathrm{Geom}(0.5)$
$N'\leftarrow\min\{N,N_{\max}\}$
Collect $2^{N'+1}$ i.i.d. samples of $\{s_i'\}_{i=1}^{2^{N'+1}}$ with $s_i'\sim\tilde{\mathsf{P}}_s^a$ for each i
$\hat{\mathsf{P}}_{s,N'+1}^{a,E}\leftarrow\frac{1}{2^{N'}}\sum_{i=1}^{2^{N'}}\mathbb{1}_{\{s_{2i}'\}}$
$\hat{\mathsf{P}}_{s,N'+1}^{a,O}\leftarrow\frac{1}{2^{N'}}\sum_{i=1}^{2^{N'}}\mathbb{1}_{\{s_{2i-1}'\}}$
$\hat{\mathsf{P}}_{s,N'+1}^{a}\leftarrow\frac{1}{2^{N'+1}}\sum_{i=1}^{2^{N'+1}}\mathbb{1}_{\{s_i'\}}$
$\hat{\mathsf{P}}_{s,N'+1}^{a,1}\leftarrow\mathbb{1}_{\{s_1'\}}$
Obtain $\sigma_{\hat{\mathsf{P}}_{s,N'+1}^{a,1}}(V),\sigma_{\hat{\mathsf{P}}_{s,N'+1}^{a}}(V),\sigma_{\hat{\mathsf{P}}_{s,N'+1}^{a,E}}(V),\sigma_{\hat{\mathsf{P}}_{s,N'+1}^{a,O}}(V)$
$\Delta_{N'}(V)\leftarrow\sigma_{\hat{\mathsf{P}}_{s,N'+1}^{a}}(V)-\frac{1}{2}\Big[\sigma_{\hat{\mathsf{P}}_{s,N'+1}^{a,E}}(V)+\sigma_{\hat{\mathsf{P}}_{s,N'+1}^{a,O}}(V)\Big]$
$\hat{\sigma}_{\mathcal{P}_s^a}(V)\leftarrow\sigma_{\hat{\mathsf{P}}_{s,N'+1}^{a,1}}(V)+\frac{\Delta_{N'}(V)}{\mathbb{P}(N'=n)}$, where $p'(n)=\mathbb{P}(N'=n)$
Return $\hat{\sigma}_{\mathcal{P}_s^a}(V)$

Results and Contributions

- One-step semi-norm contraction property for both the robust and non-robust Bellman operators under ergodicity.
- Order-optimal $\tilde{\mathcal{O}}(\epsilon^{-2})$ convergence of the TD learning algorithm for policy evaluation studied in this setting.

References

[1] Model-Free Robust Average-Reward Reinforcement Learning, ICML 2023

[2] Finite Sample Analysis of Average-Reward TD Learning and Q-Learning, NeurIPS 2021