Provably Robust Temporal Difference Learning for Heavy-Tailed Rewards

Semih Cayci[1] Atilla Eryilmaz[2]

[1]RWTH Aachen University

[2]The Ohio State University

Reinforcement Learning

Reinforcement Learning under Stochastic Rewards

- Existing RL methods assume either deterministic or light-tailed stochastic rewards.
- In many applications, rewards have heavy-tailed distributions with infinite variance.

Observation

Existing TD learning methods are not robust to heavy tails: they may not converge.

Main Question

How can we design new

- temporal difference learning (for policy evaluation),
- natural actor-critic (for policy optimization),

that achieve global optimality under stochastic rewards with heavy-tailed distributions?

Policy Evaluation Problem

Markov reward process

- $(X_t, R_t)_{t\geq 0}$ with finite but arbitrarily large state space $\mathbb{X}$,
- Value function

$$\mathcal{V}(x) = \mathbb{E}\Big[\sum_{t=1}^{\infty} \gamma^{t-1} R_t(X_t)\Big| X_0 = x\Big],$$

- $\mathbb{E}\big[|R_t(X_t)|^{1+p}\big|\sigma(X_t)\big] \leq u_0 < \infty$, for some $p \in (0, 1]$ for every $t \geq 0$.

TD learning with norm-control (Sutton, 1988; Bhandari et al., 2018)

- Let $f_\Theta(x) = \langle \Theta, \Phi(x)\rangle$. To learn $\Theta^\star \in \arg\min_{\Theta\in\mathbb{R}^d} \underset{x\sim\mu}{\mathbb{E}}\Big[\Big(\mathcal{V}(x) - f_\Theta(x)\Big)^2\Big]$, use

$$\Theta(t+1) = \Pi_{B_2(0,\rho)}\Big\{\Theta(t) + \eta \cdot g_t\Big\},$$

where $g_t = \Big(R_t(X_t) + \gamma f_{\Theta(t)}(X_{t+1}) - f_{\Theta(t)}(X_t)\Big)\Phi(X_t)$.

TD Learning under Heavy Tails

Fact: $\mathbb{E}[|R_t|^{1+p}|\sigma(X_t)] < \infty$ for $p \in (0,1]$ implies that $\mathbb{E}[\|g_t\|_2^{1+p}|\sigma(X_t,\Theta(t))] < \infty$.

Existing analyses[a] assume

$$\mathbb{E}\left[\|g_t\|_2^2 \middle| \sigma(\Theta(t), X_t)\right] < \infty,$$

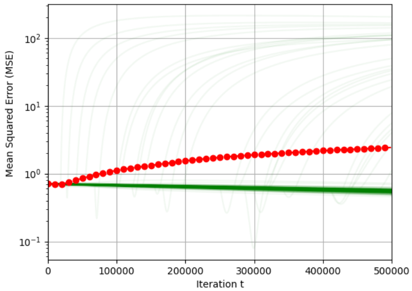
for all $t \geq 1$.

Question: Does TD learning converge if

$$\mathbb{E}[\|g_t\|^{1+p}|\sigma(X_t,\Theta(t))] < \infty,$$

for $p < 1$?

[a] See (van Roy and Tsitsiklis, 1997; Bhandari et al., 2018; Srikant and Ying, 2019).



Observation

TD learning does not converge under heavy-tailed reward – even with projection.

Robust TD Learning

Algorithm: Robust TD learning

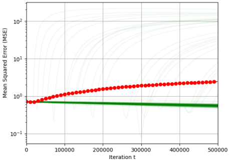
$$\widetilde{\Theta}(t+1) = \Theta(t) + \eta_t \cdot g_t \cdot \mathbb{1}\{\|g_t\|_2 \le b_t\} \qquad \texttt{(Dynamic gradient clipping)}$$

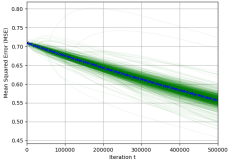
$$\Theta(t+1) = \Pi_{B_2(0,\rho)}\left\{\widetilde{\Theta}(t+1)\right\},$$

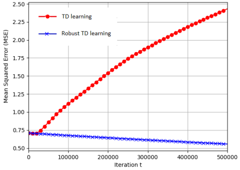
Theorem

- $b_t = \mathcal{O}(t^{\frac{1}{1+p}})$ *yields* $|\mathcal{V}(x) - f_{\bar{\Theta}(T)}(x)|^2 = \mathcal{O}\left(\frac{1}{T^{\frac{1}{1+p}}}\right)$.
- $b_t = t$ *implies* $|\mathcal{V}(x) - f_{\bar{\Theta}(T)}(x)|^2 = \tilde{\mathcal{O}}\left(\frac{1}{T^p}\right)$ *if* $\lambda_{min}\left(\sum_{x\in\mathbb{X}} \mu(x)\Phi(x)\Phi^\top(x)\right) > 0$.

Light-tailed case ($p = 1$): the bounds match the existing bounds (Bhandari et al., 2018).







Robust Natural Actor-Critic for Policy Optimization

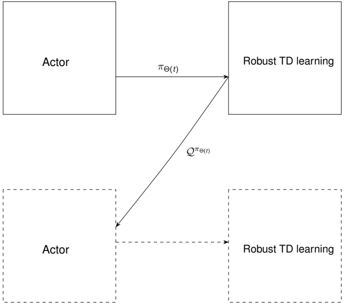
We can extend Robust TD learning to policy optimization under heavy tails.

Log-linear policy parameterization:

$$\pi_\Theta(a|s) = \frac{e^{\Theta^\top \Phi(s,a)}}{\sum_{a' \in \mathbb{A}} e^{\Theta^\top \Phi(s,a')}}.$$

Policy optimization:

$$\mathcal{V}^{\pi_\Theta}(\lambda) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R_t(S_t, A_t) \middle| S_0 \sim \lambda\right].$$



Theorem

Assume $\mathbb{E}[|R_t|^{1+p}|S_t, A_t] < \infty,\ \forall t \geq 0$ for some $p \in (0,1]$. Then, Robust NAC achieves ϵ-optimality[1] with $\mathcal{O}(\epsilon^{-4-2/p})$ samples.

[1] up to a function approximation error