PER-ETD: A Polynomially Efficient Emphatic Temporal Difference Learning Method

Ziwei Guan, Tengyu Xu, Yingbin Liang

The Ohio State University

Mar, 2022

Introduction

- Value function of target policy π

$$V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s\right],$$

with $s \in \mathcal{S}$ and $|\mathcal{S}| = N$.

Introduction

- Value function of target policy π

$$V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s\right],$$

with $s \in \mathcal{S}$ and $|\mathcal{S}| = N$.

- Linear function approximation

$$V_\pi(s) \approx V_\theta(s) = \theta^\top \phi(s), \phi(s) \text{ and } \theta \in \mathbb{R}^d (d \ll N).$$

Introduction

On-policy: $\mathcal{D} = \{(s_t, a_t, r_t, s_{t+1})\}_{t=0}^{\infty}$ with $a_t \sim \pi(\cdot|s)$,

Introduction

On-policy: $\mathcal{D} = \{(s_t, a_t, r_t, s_{t+1})\}_{t=0}^{\infty}$ with $a_t \sim \pi(\cdot|s)$,

- Temporal-Difference Learning

$$\theta_{t+1} = \theta_t + \eta(r_t + \gamma\theta_t^\top \phi_{t+1} - \theta_t^\top \phi_t)\phi_t,$$

where $\phi_t = \phi(s_t)$, $r_t = r(s_t, a_t)$ and $a_t \sim \pi(\cdot|s_t)$.

Introduction

On-policy: $\mathcal{D} = \{(s_t, a_t, r_t, s_{t+1})\}_{t=0}^{\infty}$ with $a_t \sim \pi(\cdot|s)$,

- Temporal-Difference Learning

$$\theta_{t+1} = \theta_t + \eta(r_t + \gamma\theta_t^\top \phi_{t+1} - \theta_t^\top \phi_t)\phi_t,$$

where $\phi_t = \phi(s_t)$, $r_t = r(s_t, a_t)$ and $a_t \sim \pi(\cdot|s_t)$.

  - Converges asymptotically (Sutton, 1988; Jaakkola et.al., 1994)

$$\theta_t \to \theta^*, \quad t \to \infty.$$

Introduction

On-policy: $\mathcal{D} = \{(s_t, a_t, r_t, s_{t+1})\}_{t=0}^{\infty}$ with $a_t \sim \pi(\cdot|s)$,

- Temporal-Difference Learning

$$\theta_{t+1} = \theta_t + \eta(r_t + \gamma\theta_t^\top \phi_{t+1} - \theta_t^\top \phi_t)\phi_t,$$

where $\phi_t = \phi(s_t)$, $r_t = r(s_t, a_t)$ and $a_t \sim \pi(\cdot|s_t)$.

  - Converges asymptotically (Sutton, 1988; Jaakkola et.al., 1994)

$$\theta_t \to \theta^*, \quad t \to \infty.$$

  - Finite-sample Guarantee (Dalal et al., 2018; Bhandari et al., 2018)

$$\mathbb{E}\left[\|\theta_t - \theta^*\|^2\right] \leq \tilde{\mathcal{O}}\left(\frac{1}{t}\right).$$

Introduction

Off-policy: $\mathcal{D} = \{(s_t, a_t, r_t, s_{t+1})\}_{t=0}^{\infty}$ with $a_t \sim \mu(\cdot|s_t)$

Introduction

Off-policy: $\mathcal{D} = \{(s_t, a_t, r_t, s_{t+1})\}_{t=0}^{\infty}$ with $a_t \sim \mu(\cdot|s_t)$

- Temporal-Difference Learning

$$\theta_{t+1} = \theta_t + \eta\rho_t(r_t + \gamma\theta_t^\top\phi_{t+1} - \theta_t^\top\phi_t)\phi_t,$$

where $\rho_t = \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}$.

Introduction

Off-policy: $\mathcal{D} = \{(s_t, a_t, r_t, s_{t+1})\}_{t=0}^{\infty}$ with $a_t \sim \mu(\cdot|s_t)$

- Temporal-Difference Learning

$$\theta_{t+1} = \theta_t + \eta\rho_t(r_t + \gamma\theta_t^\top\phi_{t+1} - \theta_t^\top\phi_t)\phi_t,$$

where $\rho_t = \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}$.

  - TD learning diverges (Sutton, 2018)

Introduction

Off-policy: $\mathcal{D} = \{(s_t, a_t, r_t, s_{t+1})\}_{t=0}^{\infty}$ with $a_t \sim \mu(\cdot|s_t)$

- Temporal-Difference Learning

$$\theta_{t+1} = \theta_t + \eta\rho_t(r_t + \gamma\theta_t^\top\phi_{t+1} - \theta_t^\top\phi_t)\phi_t,$$

where $\rho_t = \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}$.

  - TD learning diverges (Sutton, 2018)

- Emphatic TD Learning (Sutton et al., 2016)

$$\begin{aligned}\theta_{t+1} &= \theta_t + \eta F_t\rho_t(r_t + \gamma\theta_t^\top\phi_{t+1} - \theta_t^\top\phi_t)\phi_t \\ F_t &= \gamma\rho_t F_{t-1} + 1,\end{aligned}$$

where F_t is the emphatic weight and is initialized to be $F_0 = 1$.

Introduction

Emphatic TD Learning (Sutton et al., 2016)

$$\theta_{t+1} = \theta_t + \eta F_t \rho_t (r_t + \gamma \theta_t^\top \phi_{t+1} - \theta_t^\top \phi_t)\phi_t$$
$$F_t = \gamma \rho_t F_{t-1} + 1$$

- Converges asymptotically (Yu, 2015)

Introduction

Emphatic TD Learning (Sutton et al., 2016)

$$\theta_{t+1} = \theta_t + \eta F_t \rho_t (r_t + \gamma \theta_t^\top \phi_{t+1} - \theta_t^\top \phi_t)\phi_t$$
$$F_t = \gamma \rho_t F_{t-1} + 1$$

- Converges asymptotically (Yu, 2015)
- Variance of F_t explodes exponentially (Hallak et al., 2016)

Question:

Can we design an ETD algorithm that could converge to its fixed point in a polynomial rate?

Main results

PEriodically Restarted-ETD (PER-ETD)

- Restart F with period length b (details in Alg. 1& 2)

Main results

PEriodically Restarted-ETD (PER-ETD)

- Restart F with period length b (details in Alg. 1& 2)
- With $b = \mathcal{O}(\log(T))$ and $\eta_t = \mathcal{O}(\frac{1}{t})$, we prove

$$\mathbb{E}\left[\|\theta_T - \theta^*\|_2^2\right] \leq \mathcal{O}\left(\frac{1}{T^a}\right),$$

where a is a constant depending on $\max_{s,a}\left\{\frac{\pi(a|s)}{\mu(a|s)}\right\}$.

Main results

PEriodically Restarted-ETD (PER-ETD)

- Restart F with period length b (details in Alg. 1& 2)
- With $b = \mathcal{O}(\log(T))$ and $\eta_t = \mathcal{O}(\frac{1}{t})$, we prove

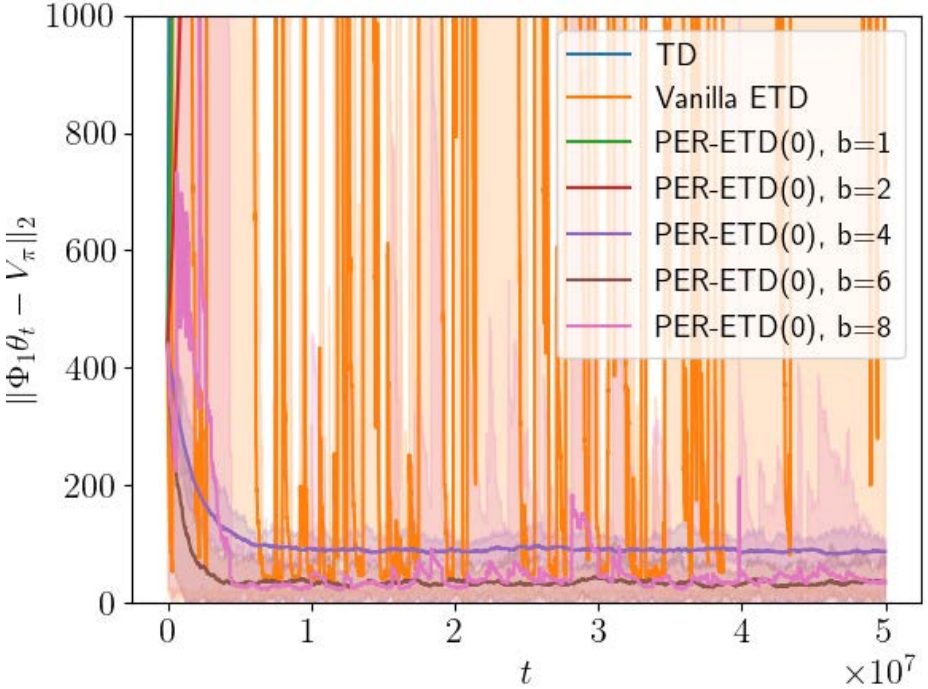
$$\mathbb{E}\left[\|\theta_T - \theta^*\|_2^2\right] \leq \mathcal{O}\left(\frac{1}{T^a}\right),$$

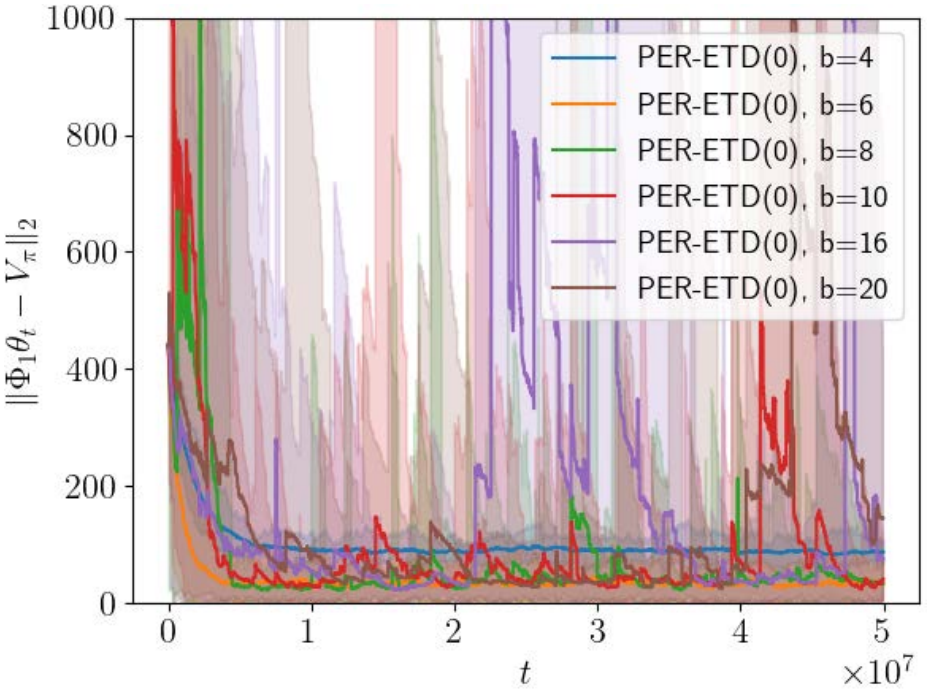
  where a is a constant depending on $\max_{s,a}\left\{\frac{\pi(a|s)}{\mu(a|s)}\right\}$.

- Similar polynomial convergence for the PER-ETD(λ)

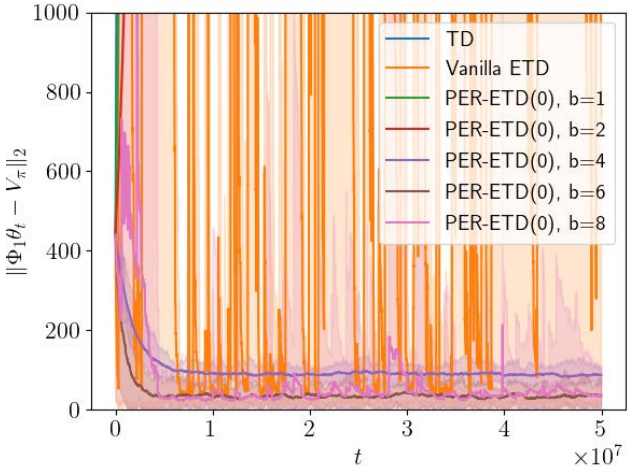
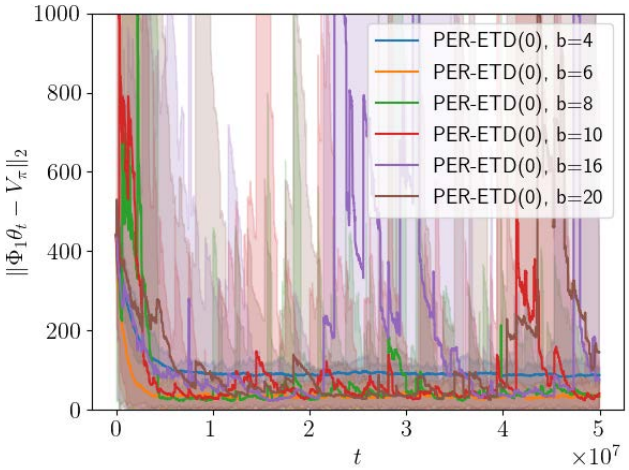
Experiments

Comparison of TD, Vanilla ETD and PER-ETD(0)





Thank you!

PEriodically Restarted-ETD (PER-ETD)

- Key Design
  - Restart follow-on trace with period length b (details in Alg. 1& 2)
- Converges polynomially with properly chosen period length and stepsize
- Numerical Comparison