Provably Convergent Two-Timescale Off-Policy Actor-Critic with Function Approximation

Shangtong Zhang[1], Bo Liu[2], Hengshuai Yao[3], Shimon Whiteson[1]

[1] University of Oxford
[2] Auburn University
[3] Huawei

Preview

- Off-policy control under the excursion objective $\sum_s d_\mu(s) v_\pi(s)$
- The first provably convergent two-timescale off-policy actor-critic algorithm with function approximation
- New perspective for Emphatic TD (Sutton et al, 2016)
- Convergence of Regularized GTD-style algorithms under a changing target policy

The excursion objective is commonly used for off-policy control

- $J(\pi) = \sum_s d_\mu(s) i(s) v_\pi(s)$
  d_μ: stationary distribution of the behaviour policy
  v_π: value function of the target policy
  $i : \mathcal{S} \rightarrow [0, \infty)$, the interest function (Sutton et al, 2016)

Off-policy policy gradient theorem gives the exact the gradient (Imani et al, 2018)

- $\nabla J(\pi) = \sum_s \bar{m}(s) \sum_a q_\pi(s, a) \nabla \pi(a \mid s)$
  $\bar{m} \doteq (I - \gamma P_\pi^\top)^{-1} D i \in \mathbb{R}^{N_s}$
  $D = \text{diag}(d_\mu)$

Rewriting the gradients gives a taxonomy of previous algorithms

- $\nabla_\theta J(\pi) = \mathbb{E}_{s \sim d_\mu, a \sim \mu(\cdot|s)}[m_\pi(s)\rho_\pi(s,a)q_\pi(s,a)\nabla_\theta \log \pi(a\,|\,s)]$
  $m_\pi \doteq D^{-1}(I - \gamma P_\pi^\top)^{-1} D i$ (emphasis)

1. Ignoring $m_\pi(s)$ (Degris et al, 2012)

2. Use followon trace to approximate $m_\pi(s)$ (Imani et al, 2018)

3. Learn $m_\pi(s)$ with function approximation (Ours)

Ignoring emphasis is theoretically justified only in tabular setting

- Gradient Estimator (Degris et al, 2012):
  $\rho_\pi(S_t, A_t) q_\pi(S_t, A_t) \nabla_\theta \log \pi(A_t \mid S_t)$

- Off-Policy Actor Critic (Off-PAC)
  Extensions: Off-policy DPG, DDPG, ACER, Off-policy EPG, TD3, IMPALA

- Off-PAC is biased even with linear function approximation (Degris et al, 2012, Imani et al, 2018, Maei et al, 2018, Liu et al, 2019)

Followon trace is unbiased only in a limiting sense

- Gradient Estimator (Imani et al, 2018):
  $M_t \rho_\pi(S_t, A_t) q_\pi(S_t, A_t) \nabla_\theta \log \pi(A_t \mid S_t)$
  $M_t \doteq i(S_t) + \gamma \rho_{t-1} M_{t-1}$ (followon trace)
  Assuming π is FIXED, $\lim_{t \to \infty} \mathbb{E}_\mu [M_t \mid S_t = s] = m_\pi(s)$

- M_t is a scalar, but m_π is a vector!

Emphasis is the fixed point of a Bellman-like operator

- $\hat{\mathbb{T}} y \doteq i + \gamma D^{-1} P_\pi^\top D y$
  - $\hat{\mathbb{T}}$ is a contraction mapping w.r.t. some weighted maximum norm (for any $\gamma < 1$)
  - The emphasis m_π is its fixed point

We propose to learn emphasis based on $\hat{\mathbb{T}}$

- A semi-gradient update based on $\hat{\mathbb{T}}$
- Gradient Temporal Difference Learning (GTD)
MSPBE: $L(\nu) \doteq ||\Pi \mathbb{T} v - v||_D^2 \quad (v = X\nu)$
- Gradient Emphasis Learning (GEM)
$L(w) \doteq ||\Pi \hat{\mathbb{T}} m - m||_D^2 \quad (m = Xw)$
- $\nabla_\theta J(\pi) = \mathbb{E}_{s\sim d_\mu, a\sim\mu(\cdot|s)}[m_\pi(s)\rho_\pi(s,a)q_\pi(s,a)\nabla_\theta \log \pi(a|s)]$

Regularized GTD-style algorithms converge under a changing policy

- TD converges under a changing policy (Konda's thesis)
  But those arguments can NOT be used to show the convergence of GTD

- Regularization has to be used for GTD-style algorithms
  GEM: $L(m) \doteq ||\Pi\hat{\mathbb{T}}Xw - Xw||_D^2 + ||w||^2$
  GTD: $L(v) \doteq ||\Pi\mathbb{T}X\nu - X\nu||_D^2 + ||\nu||^2$

- Regularization in GTD:

  - Optimization perspective under a fixed π:
    Mahadevan et al. (2014), Liu et al., (2015), Macua et al., (2015), Yu (2017), Du et al. (2017)

  - Stochastic approximation perspective under a changing π

The Convergence Off-Policy Actor-Critic (COF-PAC) algorithm

- $\nabla_\theta J(\pi) = \mathbb{E}_{s \sim d_\mu, a \sim \mu(\cdot|s)}[m_\pi(s)\rho_\pi(s,a)q_\pi(s,a)\nabla_\theta \log \pi(a|s)]$

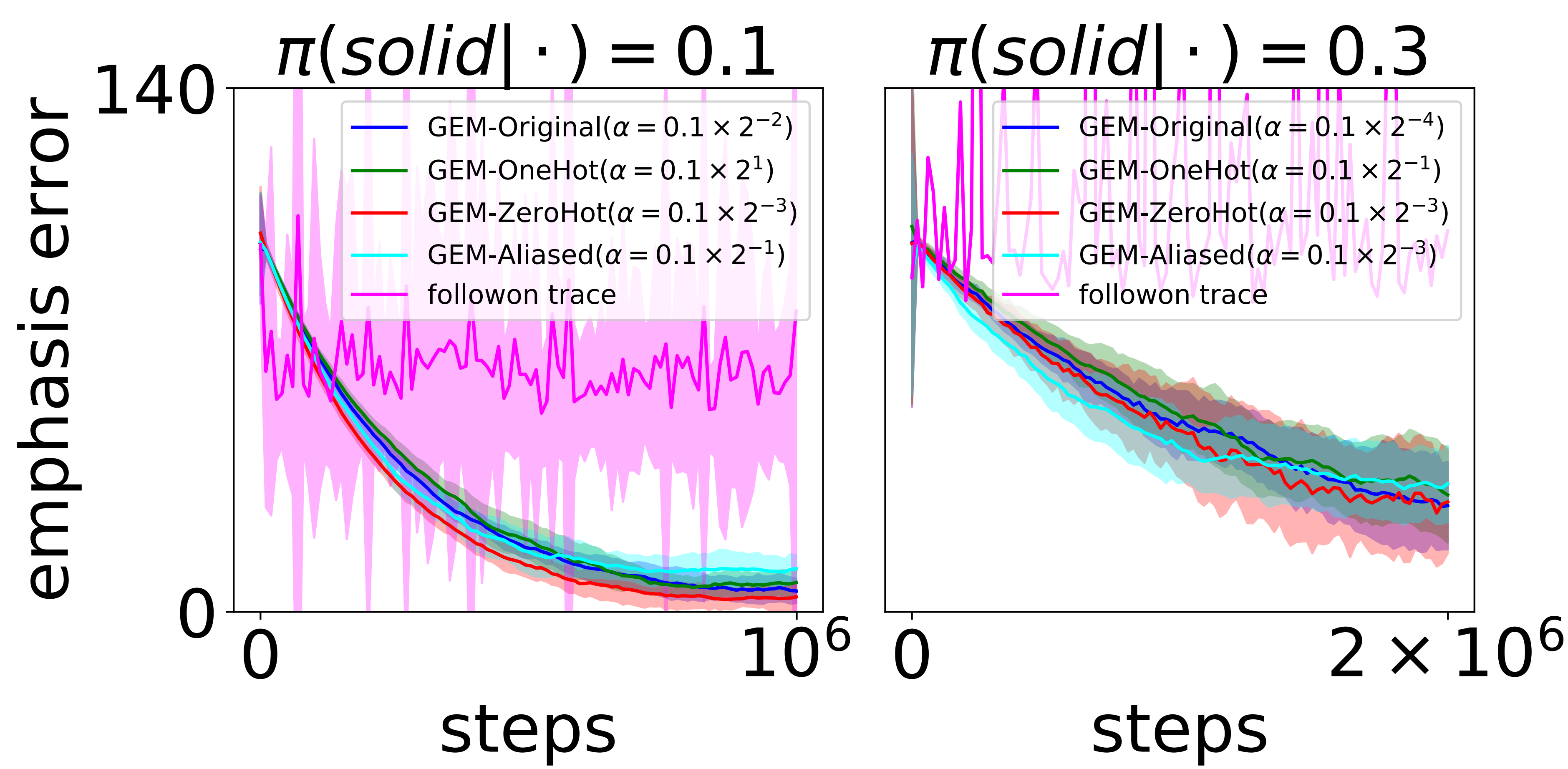
$$L(v) \doteq ||\Pi \mathbb{T} X v - X v||_D + ||v||^2$$

$$L(w) \doteq ||\Pi \hat{\mathbb{T}} X w - X w||_D + ||w||^2$$

- Two-timescale instead of bi-level optimization like SBEED

- COF-PAC visits a neighbourhood of a stationary point of $J(\pi)$ infinitely many times

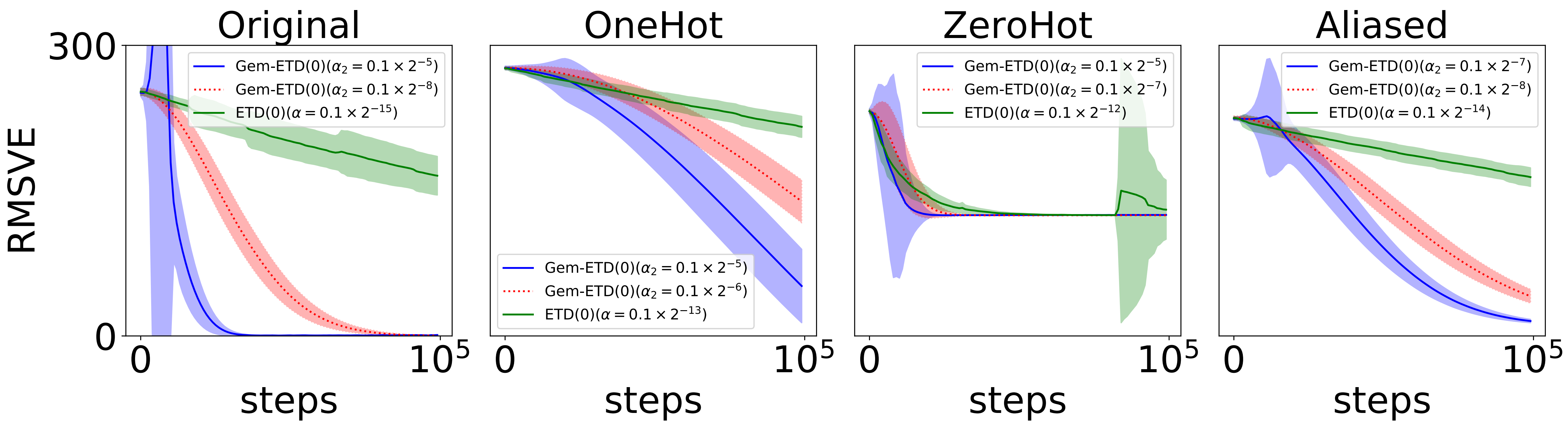
GEM approximates emphasis better than followon trace in Baird's counterexample



Averaged over 30 runs, mean + std

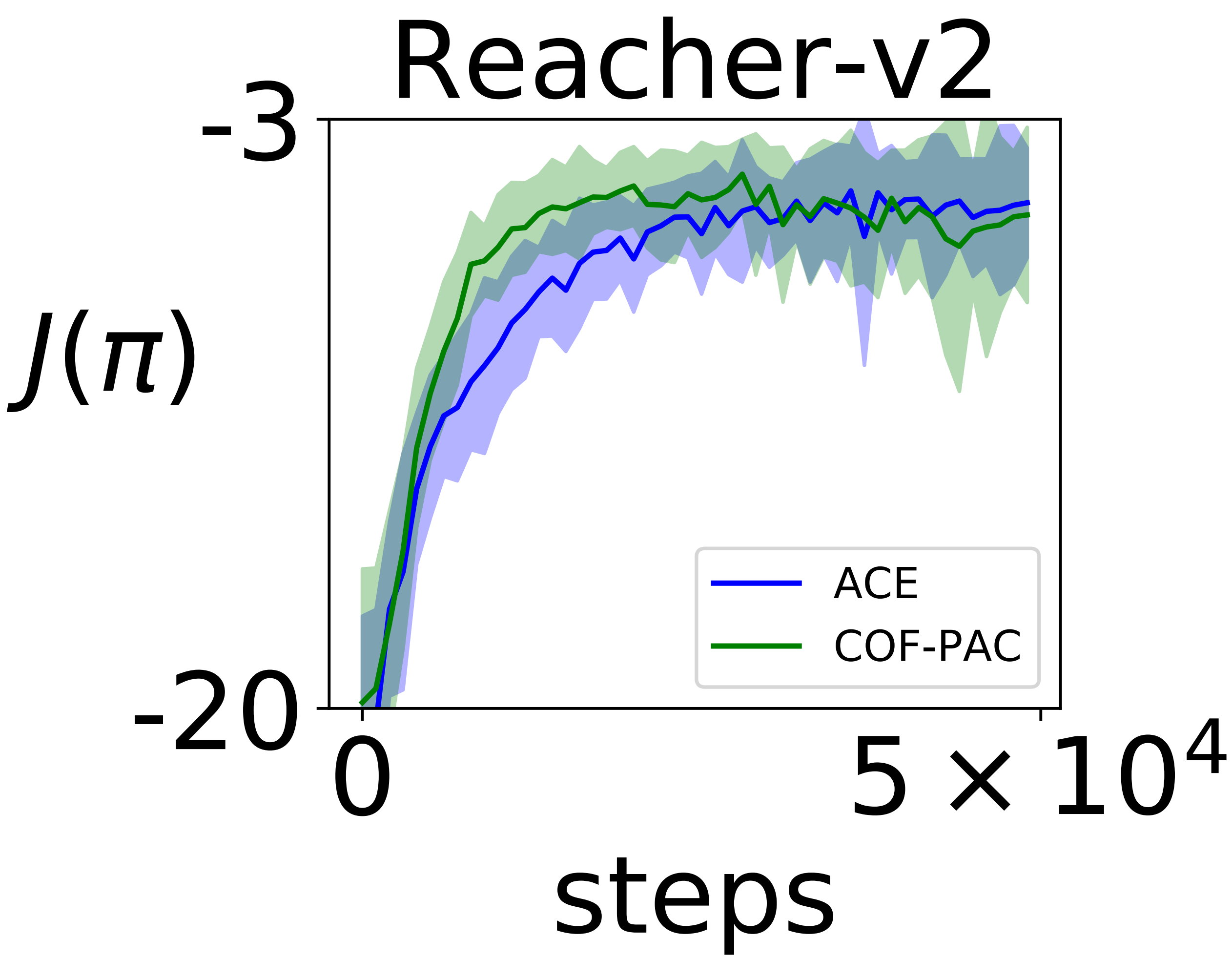
GEM-ETD doss better policy evaluation than ETD in Baird's counterexample

- ETD: $\nu_{t+1} \leftarrow \nu_t + \alpha M_t \rho_t (R_{t+1} + \gamma x_{t+1}^\top \nu_t - x_t^\top \nu_t) x_t^\top$
- GEM-ETD: $\nu_{t+1} \leftarrow \nu_t + \alpha_2 (w_t^\top x_t) \rho_t (R_{t+1} + \gamma x_{t+1}^\top \nu_t - x_t^\top \nu_t) x_t^\top$



Averaged over 30 runs, mean + std

COF-PAC does better control than ACE in Reacher



Averaged over 30 runs, mean + std

Thanks

- Code and Dockerfile are available at https://github.com/ShangtongZhang/DeepRL