Analysis of a Target-Based Actor-Critic Algorithm with Linear Function Approximation

Anas Barakat[1*], Pascal Bianchi[2], Julien Lehmann[2]

[1] ETH Zurich, [2] LTCI, Télécom Paris, Institut Polytechnique de Paris, France

[*] Work was done when the first author was at Télécom Paris

anas.barakat@inf.ethz.ch

Motivation

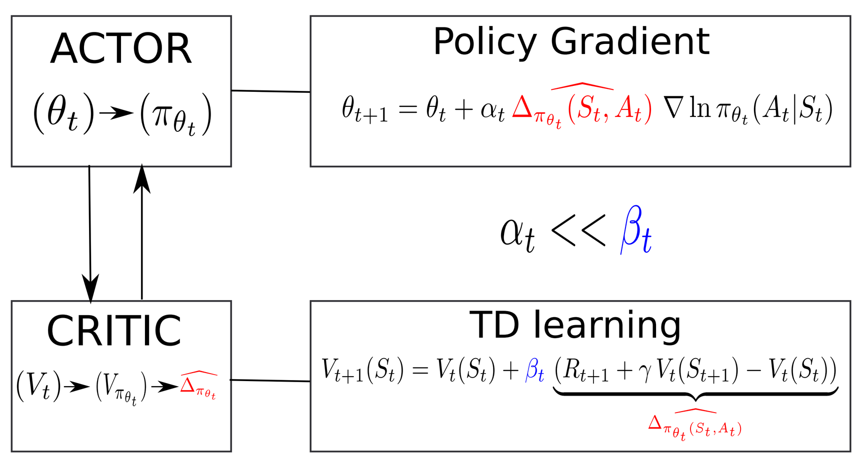
- Target network mechanism was proposed for DQN. Several deep RL **actor-critic** use this trick. Is this theoretically sound?
- Here, we look at linear FA to pave the way for nonlinear FA.
- Even in this setting, not understood for AC. Some related works:
  [1] single timescale target-TD,
  [2] value-based methods.

MDP and RL problem

- MDP $(\mathcal{S}, \mathcal{A}, p, R, \rho, \gamma)$.
- Policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$.
- "regular" policy parametrization π_θ (e.g. softmax).

$$\max_{\theta \in \mathbb{R}^d} J(\pi_\theta) := \mathbb{E}_{\rho,\pi_\theta}\left[\sum_{t=0}^{+\infty} \gamma^t R_{t+1}\right]$$

(Standard) Actor-Critic



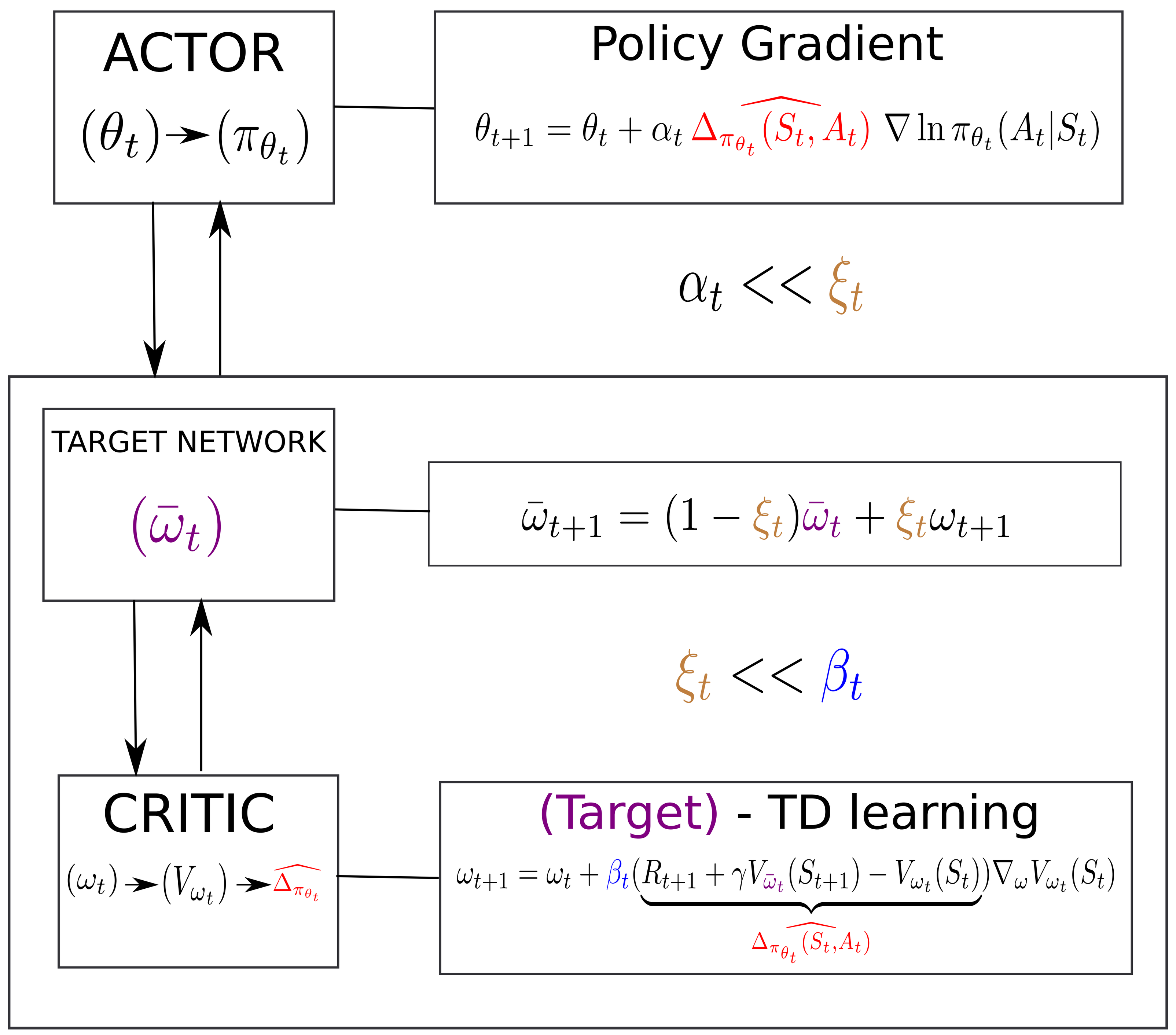
Critic with linear FA

TD learning with FA

$$\omega_{t+1} = \omega_t + \beta_t (R_{t+1} + \gamma V_{\omega_t}(S_{t+1}) - V_{\omega_t}(S_t)) \nabla_\omega V_{\omega_t}(S_t)$$

where $V_{\pi_\theta}(s) \approx V_\omega(s) = \omega^T \phi(s)$ and $\omega \in \mathbb{R}^m$ for $m \ll |\mathcal{S}|$.

Target-based Actor-Critic



Critic convergence analysis

- Multi-timescales SA [3, 4]

Theorem Under standard assumptions (Markov chain ergodicity, stepsizes, independence of the features), if $\frac{\alpha_t}{\xi_t} \to 0$ and $\frac{\xi_t}{\beta_t} \to 0$,

$$\lim_t \|\omega_t - \omega_*(\theta_t)\| = 0 \; w.p.1\,.$$

where $\omega_*(\theta)$ solution to some linear system $\forall\,\theta$.

- same interpretation as TD-like solution with linear FA [5]

Critic finite-time analysis

Let $0 < \beta < \xi < \alpha < 1$. Set $\alpha_t = \frac{c_1}{t^\alpha}$, $\xi_t = \frac{c_2}{t^\xi}$, $\beta_t = \frac{c_3}{t^\beta}$. Then,

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\omega_t - \omega_*(\theta_t)\|^2] = \mathcal{O}\left(\frac{1}{T^{1-\xi}}\right) + \mathcal{O}\left(\frac{\log T}{T^\beta}\right) + \mathcal{O}\left(\frac{1}{T^{2(\alpha-\xi)}}\right) + \mathcal{O}\left(\frac{1}{T^{2(\xi-\beta)}}\right)\,.$$

Actor convergence analysis

Theorem Under same assumptions, if $\frac{\alpha_t}{\xi_t} \to 0$ and $\frac{\xi_t}{\beta_t} \to 0$,

$$\liminf_t \left(\|\nabla J(\theta_t)\| - \underbrace{\|b(\theta_t)\|}_{\text{bias due to linear FA}} \right) \leq 0\,, w.p.1$$

Actor finite-time analysis

Lemma Set $\alpha_t = \frac{c_1}{t^\alpha}$, $\xi_t = \frac{c_2}{t^\xi}$, $\beta_t = \frac{c_3}{t^\beta}$ with $0 < \beta < \xi < \alpha < 1$. Then,

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\nabla J(\theta_t)\|^2] = \mathcal{O}\left(\frac{1}{T^{1-\alpha}}\right) + \mathcal{O}\left(\frac{\log^2 T}{T^\alpha}\right) + \mathcal{O}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\omega_t - \omega_*(\theta_t)\|^2]\right) + \mathcal{O}(\epsilon_{\mathrm{FA}})\,.$$

Theorem (Actor with tuned stepsizes) Set $\alpha_t = \frac{c_1}{t^{2/3}}$, $\xi_t = \frac{c_2}{t^{1/2}}$, $\beta_t = \frac{c_3}{t^{1/3}}$. Then,

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\nabla J(\theta_t)\|^2] = \mathcal{O}\left(\frac{\log T}{T^{1/3}}\right) + \mathcal{O}(\epsilon_{\mathrm{FA}})\,.$$

Conclusion and Perspectives

Contributions: convergence and finite-time analysis: critic (TD-like solution) and actor (gradient norm control and average expected gradient norm).

Perspectives:

- Nonlinear FA for deep RL.
- Off-policy learning.

References

[1] Donghwan Lee and Niao He. Target-based temporal-difference learning. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 3713–3722. PMLR, 09–15 Jun 2019.

[2] Shangtong Zhang, Hengshuai Yao, and Shimon Whiteson. Breaking the deadly triad with a target network. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 12621–12631. PMLR, 18–24 Jul 2021.

[3] Vivek S. Borkar. *Stochastic approximation*. Cambridge University Press, Cambridge; Hindustan Book Agency, New Delhi, 2008. A dynamical systems viewpoint.

[4] Prasenjit Karmakar and Shalabh Bhatnagar. Two time-scale stochastic approximation with controlled Markov noise and off-policy temporal-difference learning. *Math. Oper. Res.*, 43(1):130–151, 2018.

[5] John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE transactions on automatic control*, 42(5):674–690, 1997.