Global Optimality and Finite Sample Analysis of Softmax Off-Policy Actor Critic under State Distribution Mismatch

Shangtong Zhang, University of Virginia
Remi Tachet des Combes, Wayve
Romain Laroche

November 8, 2023

Practitioners do not implement what theorists want

What policy gradient theorem (Sutton et al., 1999) suggests:

$$\theta_{t+1} \doteq \theta_t + \beta_t \gamma^t q_{\pi,\gamma}(S_t, A_t) \nabla \log \pi(A_t | S_t)$$

What practitioners implement:

$$\theta_{t+1} \doteq \theta_t + \beta_t q_{\pi,\gamma}(S_t, A_t) \nabla \log \pi(A_t | S_t)$$

Theorists are fans of γ^t

- Asymptotic convergence (Konda, 2002; Zhang et al., 2020a)
- Finite sample analysis (Wu et al., 2020)
- Global optimality (Agarwal et al., 2020; Mei et al., 2020)
- TRPO (Schulman et al., 2015)
- Option-critic (Bacon et al., 2017)
- . . .

Practitioners do not like γ^t

- A3C (Mnih et al., 2016)
- PPO (Schulman et al., 2017)
- Option-critic (Bacon et al., 2017)
- ...

The distribution mismatch in off-policy actor critic

What theorists analyze:

$$\theta_{t+1} \doteq \theta_t + \beta_t \frac{d_{\pi,\gamma}(S_t)}{d_\mu(S_t)} \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)} q_{\pi,\gamma}(S_t, A_t) \nabla \log \pi(A_t|S_t)$$

What practitioners implement:

$$\theta_{t+1} \doteq \theta_t + \beta_t \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)} q_{\pi,\gamma}(S_t, A_t) \nabla \log \pi(A_t|S_t)$$

Theorists are fans of this ratio

Density ratio / marginalized importance sampling / inverse propensity score

- Asymptotic convergence (Liu et al., 2019; Zhang et al., 2020b)
- Finite sample analysis (Huang and Jiang, 2021; Xu et al., 2021)

Practitioners rarely implement this ratio

- DDPG (Lillicrap et al., 2016)
- ACER (Wang et al., 2017)
- IMPALA (Espeholt et al., 2018)
- AlphaStar (Vinyals et al., 2019)
- Schmitt et al. (2020); Zahavy et al. (2020) . . .

Can we prove the effectiveness of off-policy actor critic without correcting state distribution?

We prove the optimality of an asynchronous and stochastic off-policy actor critic without correcting state distribution

An off-policy actor critic with state distribution mismatch

At time step t,

1. Sample $A_t \sim \mu_{\theta_t}(\cdot|S_t)$
2. Execute A_t, get $R_{t+1} \doteq r(S_t, A_t), S_{t+1} \sim p(\cdot|S_t, A_t)$
3. Update critic with off-policy expected SARSA

$$\delta_t \doteq R_{t+1} + \gamma \sum_{a'} \pi_{\theta_t}(a'|S_{t+1}) q_t(S_{t+1}, a') - q_t(S_t, A_t)$$

$$q_{t+1}(S_t, A_t) \doteq q_t(S_t, A_t) + \alpha_t \delta_t$$

An off-policy actor critic with state distribution mismatch

4. Update actor with KL regularization

$$\theta_{t+1} \doteq \theta_t + \beta_t \rho_t \prod (q_t(S_t, A_t)) \nabla \log \pi_{\theta_t}(A_t|S_t) - \beta_t \lambda_t \nabla \mathrm{KL}(\mathcal{U}_{\mathcal{A}} || \pi_{\theta_t}(\cdot|S_t))$$

$\prod$: a projection onto the ball with radius $\frac{r_{max}}{1-\gamma}$

Sub-optimality decreases and success probability increases

- For any t, select a k uniformly randomly from the set $\left\{\frac{t}{2}, \frac{t}{2}+1, \ldots, t\right\}$, then

$$J(\pi_{\theta_k}) \geq J(\pi_*) - \mathcal{O}\left(k^{-\epsilon_\lambda}\right)$$

holds with probability at least

$$1 - \mathcal{O}\left(\frac{1}{t^{\epsilon_{\alpha,\beta,\lambda}}}\right)$$

Finite sample analysis of stochastic approximations with

1. Asynchronous updates
2. Markovian and Martingale difference noise
3. Time-inhomogenous Markov chain
4. Time-inhomogenous operator

$$
\begin{aligned}
v_{t+1} &\doteq v_t + \alpha_t \left(F_{\theta_t}(v, S_t) - v_t + \epsilon_t \right) \\
S_{t+1} &\sim P_{\theta_{t+1}}(S_t, \cdot)
\end{aligned}
$$

Let v_θ be the fixed point of $\bar{F}_\theta(v)$, then

$$
\lim_{t \to \infty} \mathbb{E}\left[\| v_t - v_{\theta_t} \|^2 \right] = \mathcal{O}\left(\frac{1}{t^\epsilon} \right)
$$

Two strong assumptions in previous works are removed

(p_0 is the initial distribution)

	$p_0(s) > 0$ for all s	π_* is unique
this work	X	X
Agarwal et al. (2020)	✓	X
Laroche and Tachet (2021)	X	✓

Expected SAC with a decaying temperature

At time step t,

1. Sample $A_t \sim \mu_{\theta_t}(\cdot|S_t)$
2. Execute A_t, get $R_{t+1} \doteq r(S_t, A_t), S_{t+1} \sim p(\cdot|S_t, A_t)$
3. Update critic with expected soft-SARSA

$$y_t \doteq R_{t+1} + \gamma \sum_{a'} \pi_{\theta_t}(a'|S_{t+1}) \left(q_t(S_{t+1}, a') - \lambda_t \log \pi_{\theta_t}(a'|S_{t+1})\right)$$

$$q_{t+1}(S_t, A_t) \doteq q_t(S_t, A_t) + \alpha_t \left(y_t - q_t(S_t, A_t)\right)$$

Expected SAC with a decaying temperature

- Stochastic actor update?

$$\theta_{t+1} \doteq \theta_t + \beta_t \times$$
$$\rho_t \nabla \log \pi_{\theta_t}(A_t|S_t) \left(\prod(q_t(S_t, A_t)) - \lambda_t \log \pi_{\theta_t}(A_t|S_t) \right)$$

4. Expected actor update (Ciosek and Whiteson, 2020)

$$\theta_{t+1} \doteq \theta_t + \beta_t \times$$
$$\sum_a \pi_{\theta_t}(a|S_t) \nabla \log \pi_{\theta_t}(a|S_t) \left(\prod(q_t(S_t, a)) - \lambda_t \log \pi_{\theta_t}(a|S_t) \right)$$

Thanks!

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2020). Optimality and approximation with policy gradient methods in markov decision processes. In Proceedings of the Conference on Learning Theory.

Bacon, P., Harb, J., and Precup, D. (2017). The option-critic architecture. In Proceedings of the AAAI Conference on Artificial Intelligence.

Ciosek, K. and Whiteson, S. (2020). Expected policy gradients for reinforcement learning. Journal of Machine Learning Research.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., Legg, S., and Kavukcuoglu, K. (2018). IMPALA: scalable distributed deep-rl with importance weighted actor-learner architectures. In Proceedings of the International Conference on Machine Learning.

Huang, J. and Jiang, N. (2021). On the convergence rate of off-policy policy optimization methods with density-ratio correction. arXiv preprint arXiv:2106.00993.

Konda, V. R. (2002). Actor-Critic Algorithms. PhD thesis, Massachusetts Institute of Technology.

Laroche, R. and Tachet, R. (2021). Dr Jekyll and Mr Hyde: the strange case of off-policy policy updates. In Advances in Neural Information Processing Systems.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2016). Continuous control with deep reinforcement learning. In Proceedings of the International Conference on Learning Representations.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. (2019). Off-policy policy gradient with state distribution correction. arXiv preprint arXiv:1904.08473.

Mei, J., Xiao, C., Szepesvári, C., and Schuurmans, D. (2020). On the global convergence rates of softmax policy gradient methods. In Proceedings of the International Conference on Machine Learning.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In

Proceedings of the International Conference on Machine Learning.

Schmitt, S., Hessel, M., and Simonyan, K. (2020). Off-policy actor-critic with shared experience replay. In *Proceedings of the International Conference on Machine Learning*.

Schulman, J., Levine, S., Abbeel, P., Jordan, M. I., and Moritz, P. (2015). Trust region policy optimization. In *Proceedings of the International Conference on Machine Learning*.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., Vezhnevets, A. S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T. L., Gülçehre, Ç., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T. P., Kavukcuoglu, K., Hassabis, D., Apps, C., and Silver, D. (2019). Grandmaster level in starcraft II using multi-agent reinforcement learning. Nature.

Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K., and de Freitas, N. (2017). Sample efficient actor-critic with experience replay. In Proceedings of the International Conference on Learning Representations.

Wu, Y., Zhang, W., Xu, P., and Gu, Q. (2020). A finite-time analysis of two time-scale actor-critic methods. In Advances in Neural Information Processing Systems.

Xu, T., Yang, Z., Wang, Z., and Liang, Y. (2021). Doubly robust off-policy actor-critic: Convergence and optimality. arXiv preprint arXiv:2102.11866.

Zahavy, T., Xu, Z., Veeriah, V., Hessel, M., Oh, J., van Hasselt, H. P., Silver, D., and Singh, S. (2020). A self-tuning actor-critic algorithm. In Advances in Neural Information Processing Systems.

Zhang, K., Koppel, A., Zhu, H., and Basar, T. (2020a). Global convergence of policy gradient methods to (almost) locally optimal policies. SIAM Journal on Control and Optimization.

Zhang, S., Liu, B., Yao, H., and Whiteson, S. (2020b). Provably convergent two-timescale off-policy actor-critic with function approximation. In Proceedings of the International Conference on Machine Learning.