Mean Field Langevin Actor-Critic:

Faster Convergence and Global Optimality beyond Lazy Learning

Kakei Yamamoto[1], Kazusato Oko[2,3], Zhuoran Yang[4], Taiji Suzuki[2,3]

[1] MIT [2] UTokyo [3] AIP [4] Yale

ICML International Conference On Machine Learning

OpenReview.net

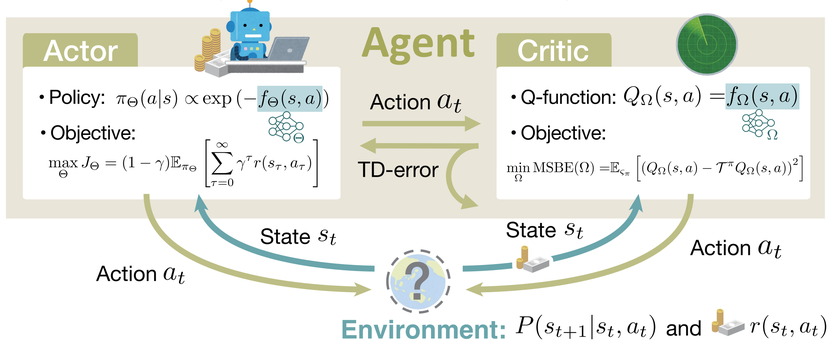
Question:

Does neural actor-critic (AC) provably learn features on the way to the global optima?

- Introduce new AC alg. = **MFLTD** + **MFLPG**
- Theoretical analysis based on Wasserstein gradient flow

Reinforcement Learning Setup

- Markov decision model (Puterman '14): $(\mathcal{S}, \mathcal{A}, \gamma, P, r)$
- Actor-critic (**AC**; Konda & Tsitsiklis '00)



NN Parameterization for Neural AC

- **Linear func. approx.** (Konda+ '00)
- **Lazy-training/NTK** depends on initialization (Wang+ '19)

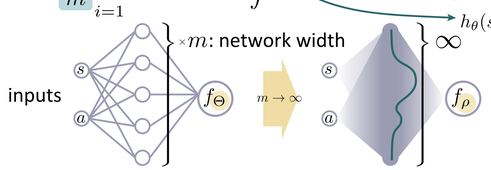
$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m} h_{\theta^i}(s,a) \overset{m\to\infty}{\to} f^\circ(s,a) + \phi_{\theta^\circ}(s,a)^\top\theta$$

✓**Mean-field networks** learn features from trainig data

$$\frac{1}{m}\sum_{i=1}^{m} h_{\theta^i}(s,a) \overset{m\to\infty}{\to} \int h_\theta(s,a)\,\rho(\mathrm{d}\theta) = f_\rho(s,a)$$

weight parameters $\theta = (w,b)$

$h_\theta(s,a) = R\cdot\beta(b)\cdot\sigma(w^\top(s,a,1))$ — activation function



Mean-field Langevin TD alg. (MFLTD)

- **Outer-loop** ($\times T_{\mathrm{TD}}$)
  - **Inner-loop** (Proximal Noisy GD; $\times S$) ◀ Convex on measure space!

$$L_l[q] = \mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - \mathcal{T}^\pi Q^{(l)})\cdot(Q_q - Q_\pi)] + \frac{1}{2(1-\gamma)}\mathbb{E}_{\varsigma_\pi}[(Q^{(l)} - Q_q)^2] + \frac{\lambda_{\mathrm{TD}}}{2}\mathbb{E}_q[\|\omega\|_2^2]$$

Objective — TD error — Weighted proximal norm — l2-norm

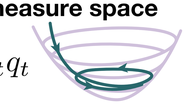
(1) Compute Average Q-function: $\overline{Q}_n^{(l)} = \frac{1}{1-\gamma}(Q_{\Omega_n} - \gamma Q^{(l)})$

(2) Update w/ iid noise $\xi_n^j \sim \mathcal{N}(0, I_d)$ for all $j \in [M]$:

$$\omega_{n+1}^j \leftarrow (1-\eta_{\mathrm{TD}}\lambda_{\mathrm{TD}})\omega_n^j - \eta_{\mathrm{TD}}\,\mathbb{E}_{\varsigma_\pi}\left[\left(\overline{Q}_n^{(l)} - \mathcal{T}^\pi Q^{(l)}\right)\nabla h_{\omega_n^j}(x)\right] + \sqrt{2\lambda\eta}\xi_n^j$$

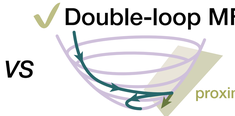
✗ **Single-loop TD(1) on measure space**

→ (Wasserstein gradient) $\nexists\ \partial_t q_t$



vs

✓ **Double-loop MFLTD**



proximally linearize

Assum. 1 β and σ are bounded, Lipschits, and smooth

Assum. 2 $Q_\pi, A_\pi \in \mathcal{F}_{R,M}$

→ Our NN class can capture Q_π and $A_\pi = Q_\pi - \int \mathrm{d}\pi Q_\pi$

$$\mathcal{F}_{R,M} = \left\{\int \beta'\cdot\sigma(w^\top(s,a,1))\cdot\rho'(\beta',w) : \mathrm{KL}(\rho'\|\nu) \le M,\ \rho' \in \mathcal{P}((-R,R)\times\mathbb{R}^{d-1})\right\}$$

◀ sub-Barron class!

Theorem 1 (Global conv. of MFLTD)

$$\frac{1}{T_{\mathrm{TD}}}\sum_{l=1}^{T_{\mathrm{TD}}}\varsigma_\pi[(Q^{(l)} - Q_\pi)^2] \le \frac{4\gamma(2-\gamma)R^2}{(1-\gamma)^2 T_{\mathrm{TD}}}$$ ◀ Outer convergence $\mathcal{O}(T_{TD}^{-1})$

$$+ C_1\lambda_{\mathrm{TD}}^{-\frac{1}{2}}e^{(-\alpha\lambda_{\mathrm{TD}}S)} + C_2 e^{(-2\alpha\lambda_{\mathrm{TD}}S)} + C_3\lambda_{\mathrm{TD}}$$

Inner-loop error — Regularization

Mean-field Langevin Policy Gradient (MFLPG)

- **Objective:** $\mathcal{F}[\rho] = -J[\rho] + \frac{\lambda}{2}\mathbb{E}_\rho[\|\theta\|_2^2] + \lambda\mathrm{Ent}[\rho] + Z_\lambda$

$= \lambda\mathrm{KL}(\rho\|\nu)$ → induce like-strongly-convex

- **Update:**

$$\theta_{k+1}^i \leftarrow (1-\eta\lambda)\,\theta_k^i - \eta\,\mathbb{E}_{\sigma_{\pi_k}}[A_k\nabla h_{\theta_k^i}] + \sqrt{2\lambda\eta}\xi_k^i$$

Advantage func. — Gaussian noise

Assum. 3 (i) $\left\|\mathrm{d}\sigma_t/\mathrm{d}\varsigma_t\right\|_{\varsigma_t,2} \le \iota$, (ii) $\left\|\mathrm{d}\sigma^*/\mathrm{d}\sigma_t\right\|_{\sigma_t,2} \le \kappa$ (visitation, stationary; optimal, visitation)

→ Sampling density moment is bounded

Theorem 2 (Global conv. of MFLPG)

Under uniformly $\mathrm{KL}(\widehat{\rho}_t\|\nu) \le M$

$$\max_\pi J_\pi - J[\rho_t] \le 2R\exp(-\alpha\lambda t) + \mathcal{O}(\lambda^{1/2})$$

Theorem 3 (Time and space discretization)

$$\max_\pi J_\pi - [J_{\Theta_T}] \le 2R\exp(-\alpha\lambda\eta T/2)$$ ◀ Linear convergence

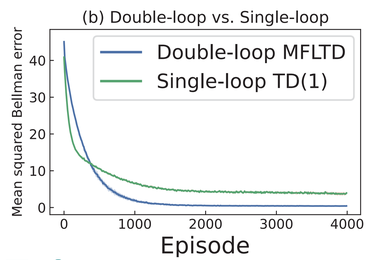
$$+ \mathcal{O}\left(\frac{\eta+\lambda}{m} \vee \frac{\eta(\eta+\lambda)^2}{\lambda} \vee \alpha\lambda^{\frac{3}{2}} \vee \delta_{\mathrm{TD}}^2\right)$$

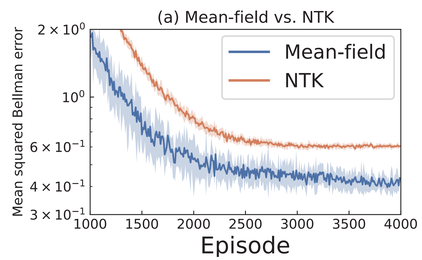
approximation — time discretization — regularization — critic

m: neuron #
η: step size
λ: regularization param.

Numerical Experiments (MFLTD)

- **Double-loops** do not require a "double" calculation
  - → Better accuracy with the same number of **episodes**
- **Mean-fields** may capture a richer reperesentation power
  - → Mean-field nets overwhealm NTK in test accuracy





Takeaway

- Introduced new provaly guaranteed Actor-critic, **MFLTD+PG.**
- To the best of our knowledge, the 1st theoretical analysis on fully **neural Acter-critic** global opt. beyond **lazy training.**
- Gave experimental substantiation (**higher accuracy than TD(1).**)