Convergence of an actor-critic gradient flow for entropy regularised MDPs in general spaces

Denis Zorba, David Siska, Lukasz Szpruch

University of Edinburgh

ICLR 2026

Entropy regularised MDPs

For some regularisation parameter $\tau > 0$, initial state distribution $\rho \in \mathcal{P}(S)$, the value function is defined as

$$V_\tau^\pi(\rho) = \mathbb{E}_{s_0 \sim \rho}^\pi \left(\sum_{n=0}^{\infty} \gamma^n \left(c(s_n, a_n) + \tau \, \mathrm{KL}(\pi(\cdot|s_n)|\mu) \right) \right).$$

Entropy regularised MDPs

For some regularisation parameter $\tau > 0$, initial state distribution $\rho \in \mathcal{P}(S)$, the value function is defined as

$$V_\tau^\pi(\rho) = \mathbb{E}_{s_0 \sim \rho}^\pi \left(\sum_{n=0}^{\infty} \gamma^n \left(c(s_n, a_n) + \tau \, \mathrm{KL}(\pi(\cdot|s_n)|\mu) \right) \right).$$

The state-action value function is defined as

$$Q_\tau^\pi(s, a) = c(s, a) + \gamma \int_S V_\tau^\pi(s') P(ds'|s, a),$$

and satisfies $T_\tau^\pi Q_\tau^\pi = Q_\tau^\pi$ with

$$T_\tau^\pi f(s, a) = c(s, a) + \gamma \int_{S \times A} f(s', a') \pi(da'|s') P(ds'|s, a) + \tau\gamma \int_S \mathrm{KL}(\pi(\cdot|s')|\mu) P(ds'|s, a).$$

Entropy regularised MDPs

For some regularisation parameter $\tau > 0$, initial state distribution $\rho \in \mathcal{P}(S)$, the value function is defined as

$$V_\tau^\pi(\rho) = \mathbb{E}_{s_0 \sim \rho}^\pi \left(\sum_{n=0}^{\infty} \gamma^n \left(c(s_n, a_n) + \tau \, \mathrm{KL}(\pi(\cdot|s_n)|\mu) \right) \right).$$

The state-action value function is defined as

$$Q_\tau^\pi(s, a) = c(s, a) + \gamma \int_S V_\tau^\pi(s') P(ds'|s, a),$$

and satisfies $T_\tau^\pi Q_\tau^\pi = Q_\tau^\pi$ with

$$T_\tau^\pi f(s, a) = c(s, a) + \gamma \int_{S \times A} f(s', a') \pi(da'|s') P(ds'|s, a) + \tau\gamma \int_S \mathrm{KL}(\pi(\cdot|s')|\mu) P(ds'|s, a).$$

Goal: compute the minimiser $\pi^* \in \mathcal{P}(A|S)$ such that $\pi^* = \arg\min V_\tau^\pi(\rho)$.

Temporal difference and Mirror descent

Let $Q(s,a;\theta) = \langle \theta, \phi(s,a) \rangle$ for some $\phi : S \times A \to \mathbb{R}^N$. Directly from this, we define the approximate advantage function for a policy π as

$$A(s,a;\theta) = Q(s,a;\theta) + \tau \log \frac{d\pi}{d\mu}(s,a) - V(s;\theta),$$

with $V(s;\theta) = \int_A \left(Q(s,a;\theta) + \tau \log \frac{d\pi}{d\mu}(s,a) \right) \pi(da|s)$.

Temporal difference and Mirror descent

Let $Q(s,a;\theta) = \langle \theta, \phi(s,a) \rangle$ for some $\phi : S \times A \to \mathbb{R}^N$. Directly from this, we define the approximate advantage function for a policy π as

$$A(s,a;\theta) = Q(s,a;\theta) + \tau \log \frac{d\pi}{d\mu}(s,a) - V(s;\theta),$$

with $V(s;\theta) = \int_A \left(Q(s,a;\theta) + \tau \log \frac{d\pi}{d\mu}(s,a) \right) \pi(da|s)$. How to update θ and π ?

Temporal difference and Mirror descent

Given some $\theta^0 \in \mathbb{R}^N$ and initial policy π^0, consider the following algorithm

$$\theta^{n+1} = \theta^n - h_n g(\theta^n, \pi^n)$$

Temporal difference and Mirror descent

Given some $\theta^0 \in \mathbb{R}^N$ and initial policy π^0, consider the following algorithm

$$\theta^{n+1} = \theta^n - h_n g(\theta^n, \pi^n)$$

$$\pi^{n+1} = \arg\min_{\pi} \left\{ \int_S \left(\int_A A(s, a; \theta^{n+1}) \pi(\cdot|s) + \frac{1}{\lambda_n} \mathrm{KL}(\pi(\cdot|s)|\pi^n(\cdot|s)) \right) d_\rho^\pi(ds) \right\},$$

where $g : \mathbb{R}^N \times \mathcal{P}(A|S)$ is the semi gradient of the Mean Squared Bellman Error (MSBE), defined as

$$g(\theta, \pi) = \int_{S \times A} \left(Q(s, a; \theta) - \mathrm{T}_\tau^\pi Q(s, a; \theta) \right) \phi(s, a) d_\beta^\pi(da, ds).$$

Dynamics

We consider the continuous time dynamics of the updates:

$$\frac{d}{dt}\theta_t = -\eta_t g(\theta_t, \pi_t),$$

$$\partial_t \pi_t(da|s) = -A(s, a; \theta_t)\pi_t(da|s),$$

where $\eta_t : [0, \infty) \to [1, \infty)$ be a non-decreasing, continuous function. [Zha+21] study a similar flow in the unregularised case ($\tau = 0$) however convergence was only established to a neighbourhood of the optimal policy, and a restarting mechanism was required.

Problems and Challenges

- Due to the KL regularisation and the fact that the state and action space is Polish, there is no a priori upper bound on the value functions.

Problems and Challenges

- Due to the KL regularisation and the fact that the state and action space is Polish, there is no a priori upper bound on the value functions.
- Thus, before addressing the convergence of the coupled flow to the optimal policy π^*, we must first address its stability.

Assumptions

1. A1: For all $(s, a) \in S \times A$ it holds that $|\phi(s, a)|_2 \leq 1$

Assumptions

1. A1: For all $(s, a) \in S \times A$ it holds that $|\phi(s, a)|_2 \leq 1$
2. A2: $\lambda_\beta = \lambda_{\min}\Big(\int_{S \times A} \phi(s, a)\phi(s, a)^\top \beta(da, ds)\Big) > 0$

Stability

To ease notation, define $\Gamma = \lambda_\beta(1-\gamma)(1-\sqrt{\gamma})$, $\mathrm{K}_t = \sup_{s\in S} \mathrm{KL}(\pi_t(\cdot|s)|\mu)$.

Stability

To ease notation, define $\Gamma = \lambda_\beta(1-\gamma)(1-\sqrt{\gamma})$, $\mathrm{K}_t = \sup_{s\in S} \mathrm{KL}(\pi_t(\cdot|s)|\mu)$.

Theorem

Let A1, A2 hold and let $\eta_0 > \frac{\tau}{\Gamma}$. Then there exists constants $a_1, a_2 > 0$ such that for all $\gamma \in (0,1)$, $s \in S$ and $t \geq 0$ it holds that

$$\mathrm{K}_t^2 \leq a_1 + a_2 \int_0^t e^{-\tau(t-r)} \mathrm{K}_r^2 \, dr.$$

A direct corollary then shows that the coupled flow does not blow up in finite time. Uniform boundedness under more restrictive conditions.

Convergence

3. A3: For all $t \geq 0$, there exists $\theta_{\pi_t} \in \mathbb{R}^N$ such that $Q_\tau^{\pi_t}(s,a) = \langle \theta_{\pi_t}, \phi(s,a) \rangle$ for all $s \in S$ and $a \in A$.

Convergence

3. A3: For all $t \geq 0$, there exists $\theta_{\pi_t} \in \mathbb{R}^N$ such that $Q_\tau^{\pi_t}(s,a) = \langle \theta_{\pi_t}, \phi(s,a) \rangle$ for all $s \in S$ and $a \in A$.

Theorem

Let A1, A2 and A3 hold. Then there exists $k_1 > 0$ with $\eta_t = \eta_0 e^{k_1 t}$ and $k_2 > 0$ such that for all $\gamma \in (0,1)$ and $t > 0$ it holds that

$$\min_{r \in [0,t]} V_\tau^{\pi_r}(\rho) - V_\tau^{\pi^*}(\rho)$$
$$\leq \frac{\tau e^{-\frac{\tau}{2}t}}{2(1-\gamma)(1-e^{-\frac{\tau}{2}t})} \left(\int_S \mathrm{KL}(\pi^*(\cdot|s)|\pi_0(\cdot|s)) d_\rho^{\pi^*}(ds) + \frac{k_2}{2\tau} \right)$$

References I

[Zha+21] Yufeng Zhang et al. "Wasserstein Flow Meets Replicator Dynamics: A Mean-Field Analysis of Representation Learning in Actor-Critic". In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021.