

CONVERGENCE OF POLICY GRADIENT FOR ENTROPY REGULARIZED MDPs

with Neural Network Approximation
in the Mean-Field Regime

James-Michael Leahy*
Imperial College London

B. Kerimkulov*, D. Siska*, and L. Szpruch*
University of Edinburgh

ICML 2022

* Equal contribution

Entropy regularized MDPs

Value function: continuous state and action space

$$V_{\tau}^{\pi}(\rho) = \mathbb{E}_{s_0 \sim \rho}^{\pi} \sum_{t=0}^{\infty} \gamma^t \left[r(s_t, a_t) - \tau \ln \frac{d\pi}{d\mu}(a_t | s_t) \right]$$

τ : reward-based entropy regularization strength

μ : finite reference measure on action space A

Goal: compute maximizer $\pi_{\tau}^* : S \rightarrow \mathcal{P}_{\mu}(A)$

Consequences of entropy regularization

Optimal value and policy of softmax form

$$V_{\tau}^*(s) = \tau \ln \int_A \exp\left(\frac{Q_{\tau}^*(s, a)}{\tau}\right) \mu(da)$$

$$\pi_{\tau}^*(da|s) = \exp\left(\frac{Q_{\tau}^*(s, a) - V_{\tau}^*(s)}{\tau}\right) \mu(da),$$

where

$$Q_{\tau}^*(s, a) = r(s, a) + \gamma \int_S V_{\tau}^*(s') P(ds'|s, a)$$

Neural network softmax policy

Same functional form as optimal policy

$$\pi_\nu(da|s) \sim \exp\left(\int_{\mathbb{R}^d} f(s, a, \theta) \nu(d\theta)\right) \mu(da)$$

f : e.g., one-hidden layer neural network

ν : distribution on parameter space \mathbb{R}^d

Mean-field \rightarrow infinite width NN and universal approximator

$$\int_{\mathbb{R}^d} f(s, a, \theta) \nu(d\theta) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N c_k^{(n)} \sigma(\langle w_k^{(n)}, (s, a) \rangle)$$

Convergence of softmax policy gradient

Tabular: $\pi_{\theta}(s|a) = \text{softmax}(\theta(s, a))$

- ⊙ $O(e^{-ct})$ -convergence of entropy-regularized PG [Mei et al., 2020]

Continuous state and action: softmax mean-field π_{ν}

- ⊙ if PG flow ν_t converges to ν^* with full support, then $\pi_{\nu^*} = \pi_{\tau}^*$ [Agazzi and Lu, 2021]

But does it converge?

Entropy regularized objective

Introduce entropy in parameter measure

$$J^{\tau, \sigma}(v) = V_{\tau}^{\pi_v}(\rho) - \frac{\sigma^2}{2} \text{KL}(v|e^{-U})$$

σ : parameter-based entropy regularization strength

U : strong convex potential $\nabla^2 U \succ \kappa$

Goal: compute maximizer $v^* = v_{\sigma}^*$ and quantify bias

Policy gradient: the Lions derivative

What does the gradient look like?

$$\nabla \frac{\delta J^{\tau, \sigma}}{\delta v}(v, \theta) = \nabla \frac{\delta V_{\tau}^{\pi_v}(\rho)}{\delta v}(v, \theta) - \frac{\sigma^2}{2} (\nabla U(\theta) + \nabla \ln v(\theta)) ,$$

where

$$\nabla \frac{\delta V_{\tau}^{\pi_v}(\rho)}{\delta v}(v, \theta) = \frac{1}{1-\gamma} \mathbb{E}_{d_{\rho}^{\pi}} \text{cov}_{\pi_v} \left(Q_{\tau}^{\pi_v} - \tau \ln \frac{d\pi_v}{d\mu}, \nabla f(\theta) \right)$$

$d_{\rho}^{\pi}(ds)$: occupancy measure

Policy gradient flow

Non-linear Fokker-Planck equation:

$$\partial_t \nu_t = -\nabla \cdot \left(\left(\nabla \frac{\delta V_\tau^{\pi_\nu}(\rho)}{\delta \nu}(\nu) - \frac{\sigma^2}{2} \nabla U \right) \nu_t \right) + \frac{\sigma^2}{2} \Delta \nu_t$$

McKean-Vlasov SDE representation: $\nu = \text{Law}(\theta)$

$$d\theta_t = \left(\nabla \frac{\delta V_\tau^{\pi_\nu}(\rho)}{\delta \nu}(\nu_t, \theta_t) - \frac{\sigma^2}{2} \nabla U(\theta_t) \right) dt + \sigma dW_t$$

σ controls dissipation and convexity strength

Policy gradient flow approximation

Particle approximation: noisy gradient ascent

$$\theta_{k+1}^{(n)} = \theta_k^{(n)} + \eta \left(\nabla \frac{\delta V_{\tau}^{\pi_v}(\rho)}{\delta v}(v_k^{(N)}, \theta_k^{(n)}) - \frac{\sigma^2}{2} \nabla U(\theta_k^{(n)}) \right) + \sqrt{\eta} \sigma \zeta_{k+1}^{(n)},$$

$v_t^{(N)} = \frac{1}{N} \sum_{n=1}^N \delta_{\theta_t^{(n)}}: \text{empirical measure}$

$\eta: \text{learning rate}$

$\zeta_k^{(n)} \stackrel{\text{i.i.d.}}{\sim} N(0, I_{d \times d}): \text{noise}$

Policy improvement

Entropy regularized value increases in time

$$\frac{d}{dt} J^{\tau, \sigma}(v_t) = \int_{\mathbb{R}^d} \left| \nabla \frac{\delta J^{\tau, \sigma}}{\delta v}(v_t) \right|^2 v_t(d\theta) \geq 0$$

Argument: integration by parts

Key difficulty: entropy non-differentiable

Theorem: convergence of policy gradient

Assuming regularization strong enough, we prove

- ⊙ there exists a unique global maximizer v^* of $J^{\tau,\sigma}$,
- ⊙ v^* is unique solution of



$$\nabla \cdot \left(\left(\nabla \frac{\delta J^{\tau,0}}{\delta v}(v^*) - \frac{\sigma^2}{2} \nabla U \right) v^* \right) + \frac{\sigma^2}{2} \Delta v^* = 0 ,$$

- ⊙ exponential Wasserstein convergence

$$W_2(v_t, v^*) \leq e^{-\beta t} W_2(v_0, v^*) .$$

We also quantify the bias introduced by τ, σ -regularization

References

-  Agazzi, A. and Lu, J. (2021).
Global optimality of softmax policy gradient with single hidden layer neural networks in the mean-field regime.
In International Conference on Learning Representations.
-  Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. (2020).
On the global convergence rates of softmax policy gradient methods.
In International Conference on Machine Learning, pages 6820–6829. PMLR.