Entropy regularized Markov decision processes (MDPs)

Value function:

\[ \pi \in \mathcal{P}_\mu(A|S) \mapsto V^\pi_\tau(\rho) = \mathbb{E}_\rho^{\pi} \sum_{t=0}^{\infty} \gamma^t \left[ r(s_t, a_t) - \tau \ln \frac{d\pi}{d\mu}(a_t|s_t) \right] \]

- \( S \) and \( A \): polish state and action spaces
- \( P \in \mathcal{P}(S|S \times A) \): stochastic transition kernel
- \( \rho \in \mathcal{P}(S) \): arbitrary initial state distribution
- \( r \in B_b(S \times A) \): bounded measurable reward
- \( \gamma \in [0, 1) \): discount factor
- \( \mu \): finite reference measure on \( \mathcal{B}(A) \)
- \( \tau \): reward-based entropy regularization
Denoting $V^\pi_\tau(s) = V^\pi_\tau(\delta_s)$ for $s \in S$, we define

$$Q^\pi_\tau(s, a) = r(s, a) + \gamma \int_S V^\pi_\tau(s') P(ds'|s, a).$$

Let $V^*(s) = \sup_\pi V^\pi(s)$ and define $Q^*$ analogously.

**Theorem (\(\tau\)-entropy regularized DPP)**

*If \(\tau = 0\), the usual Bellman equation holds. If \(\tau > 0\), then for all \(s \in S\),

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

and $V^*_\tau(\rho) = \int_S V^*(s) \rho(ds)$. Moreover, there is a unique optimal policy

$$\pi^*_\tau(da|s) = \exp \left(\frac{(Q^*_\tau(s, a) - V^*_\tau(s))/\tau}{\mu(da)}\right).$$
Unknown dynamics or high dimension

What do you do if you don’t know the dynamics or the dimension too large?

- direct: learn the dynamics and solve Bellman if dimension is low
- indirect: \( Q \)-learning i.e., swap \( \sup \) to \( \mathbb{E} \) sup and use stochastic approx
- indirect: policy gradient i.e., parameterize policy
- indirect: hybrid e.g., actor-critic
- all other approximate dynamic programming [Bertsekas et al., 2011]

If you don’t know the dynamics, you can compare algorithms by their performance on a finite number of “plays or samples” (i.e., regret)
Policy gradient in a nutshell

Parameterize policy:

\[ J^\tau(\theta) = V^\pi_\tau(\rho), \quad \text{where} \quad \pi_\theta(da|s) \sim \exp(f(s, a, \theta))\mu(da) \]

Policy gradient:

\[ \nabla_\theta J^\tau(\theta) = \mathbb{E}_{d^\pi_\theta}(Q^\pi_\tau - \tau \ln \frac{d\pi_\theta}{d\mu}) \nabla_\theta \ln \pi_\theta \]

\[ d^\pi_\theta(ds) = \mathbb{E}_{\rho}[(\text{id} - \gamma P^\pi)^{-1}]: \text{occupancy measure} \]

Estimate gradient using rollouts or stochastic approximation of \( Q^\pi_\tau \)

Policy gradient flow:

\[ \dot{\theta}_t = \nabla_\theta \hat{J}^\tau(\theta) \]

\( \tau = \tau_t \) helps with exploring AND convergence
Softmax mean-field parameterized policy

Softmax parameterized policy:

\[ \nu \in \mathcal{P}(\mathbb{R}^d) \mapsto \pi_{\nu}(da|s) \sim \exp \left( \int_{\mathbb{R}^d} f(s, a, \theta) \nu(d\theta) \right) \mu(da) \]

\[ \circ \ f \in L^\infty(S \times A; C_b^2(\mathbb{R}^d)): \text{ smooth parametric family} \]

Neural network mean-field approximation

Let \( S = \mathbb{R}^{d_s}, A = \mathbb{R}^{d_A}, \psi : \mathbb{R} \to [-1, 1] \) smooth,

\[ f(s, a, (c, w, b)) = \sum_{k=1}^{K} \psi(c_k) \tanh(\langle w_k, (s, a) \rangle + b_k). \]

For an i.i.d. sample \( \{\theta^{(n)}\}_{n=1}^{\infty} = \{(c^{(n)}, w^{(n)}, b^{(n)})\}_{n=1}^{\infty} \text{ i.i.d. } \sim \nu, \)

\[ \int_{\mathbb{R}^d} f(s, a, \theta) \nu(d\theta) = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} \psi(c_k^{(n)}) \tanh(\langle w_k^{(n)}, (s, a) \rangle + b_k^{(n)}). \]
Convergence of softmax policy gradient

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

- $O(1/\sqrt{t})$-convergence of policy gradient [Agarwal et al., 2021]
- $O(1/t)$-convergence of softmax policy gradient [Mei et al., 2020]
- $O(e^{-ct})$-convergence of entropy-regularized PG [Mei et al., 2020]

Continuous state and action: softmax mean-field $\pi_\nu$

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

But does it converge?
Entropy regularized objective:

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

- \( U \): potential on \( \mathbb{R}^d \)
  - bounded 2nd derivative
  - \( \kappa \)-strong convex
  - satisfies \( \int_{\mathbb{R}^d} e^{-U(\theta)} d\theta = 1 \)
  - e.g., \( U(\theta) = \frac{d}{2} \ln(2\pi) + \frac{1}{2} |\theta|^2 \)
- \( \sigma \): strength of parameter-based entropy regularization

Goal: compute \( v^* \in \max_v J_{\tau,\sigma}(v) \)
Lemma (Lion’s derivative)

For all $\nu \in \mathcal{P}(\mathbb{R}^d)$ and $\theta \in \mathbb{R}^d$,

$$
\nabla_{\nu} \frac{\delta J_{\tau,\sigma}}{\delta \nu}(\nu, \theta) = \frac{\delta V_{\tau}^{\pi_v}(\rho)}{\delta \nu}(\nu, \theta) - \frac{\sigma^2}{2} \left( \nabla U(\theta) + \nabla \ln \nu(\theta) \right),
$$

where

$$
\frac{\delta V_{\tau}^{\pi_v}(\rho)}{\delta \nu}(\nu, \theta) = \frac{1}{1 - \gamma} \mathbb{E}_{d_\mu} \text{cov}_{\pi_v} \left( Q_{\tau}^{\pi_v} - \tau \ln \frac{d\pi_v}{d\mu}, \nabla f(\theta) \right).
$$
**Properties of Lion’s derivative**

**Theorem (Boundedness and Lipschitzness)**

There are constants $C_k, k \in \mathbb{N}, L, \text{ and } D$ such that for all $\tau, \tau' \geq 0, \theta \in \mathbb{R}^d, \nu, \nu' \in \mathcal{P}_1(\mathbb{R}^d)$,

\[
\left| \nabla^k \frac{\delta J^{\tau,0}}{\delta \nu}(\nu, \theta) \right| \leq C_k,
\]

\[
|J^{\tau,0}(\nu') - J^{\tau,0}(\nu)| \leq C_1 W_1(\nu', \nu),
\]

\[
\left| \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu', \theta) - \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu, \theta) \right| \leq L W_1(\nu', \nu),
\]

and

\[
\left| \nabla \frac{\delta J^{\tau',0}}{\delta \nu}(\nu, \theta) - \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu, \theta) \right| \leq D |\tau' - \tau|.
\]
Policy gradient flow

Theorem (Well-posedness, MKV representation, policy improvement)

For every $\nu_0 \in \mathcal{P}_1(\mathbb{R}^d)$, there exists a unique solution of the policy gradient flow

$$\partial_t \nu_t = -\nabla \cdot \left( \nabla \frac{\delta J^{\tau,\sigma}}{\delta \nu} (\nu_t) \nu_t \right) = -\nabla \cdot \left( \left( \nabla \frac{\delta V^{\tau,\nu}_\tau (\rho)}{\delta \nu} (\nu_t, \theta_t) - \frac{\sigma^2}{2} \nabla U \right) \nu_t \right) + \frac{\sigma^2}{2} \Delta \nu_t .$$

The solution has a representation $\nu = \text{Law}(\theta)$ as the law of the McKean-Vlasov SDE:

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

Moreover, along the gradient flow, the regularized optimization objective is increasing

$$\frac{d}{dt} J^{\tau,\sigma}(\nu_t) = \int_{\mathbb{R}^d} \frac{\delta J^{\tau,\sigma}}{\delta \nu} (\nu_t) \partial_t \nu_t (d\theta) = \int_{\mathbb{R}^d} \left| \nabla \frac{\delta J^{\tau,\sigma}}{\delta \nu} (\nu_t) \right|^2 \nu_t (d\theta) \geq 0 .$$
Policy gradient flow approximation

Particle approximation

Approximating \( \nu = (\nu_t)_{t \geq 0} \) with an empirical measure \( \nu_t^{(N)} = \frac{1}{N} \sum_{n=1}^{N} \delta_{\theta_t^{(n)}} \) and discretizing in time with a learning rate \( \eta \), we arrive at noisy gradient ascent

\[
\theta_{k+1}^{(n)} = \theta_k^{(n)} + \eta \left( \nabla \frac{\delta V_{\tau}^{\pi_{\nu}}(\rho)}{\delta \nu} (\nu_k^{(N)}, \theta_k^{(n)}) - \frac{\sigma^2}{2} \nabla U(\theta_k^{(n)}) \right) + \sqrt{\eta \sigma} \zeta_{k+1}^{(n)},
\]

where \( \{\zeta_k^{(n)}\}_{1 \leq n \leq N, k \in \mathbb{N}_0} \) i.i.d. \( \sim N(0, 1) \).
Convergence of entropy-regularized policy gradient

**Theorem (Convergence in the regularized regime)**

If $\beta := \frac{\sigma^2}{2}\kappa - C_2 - L > 0$, then there exists a unique solution $v^*$ of

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

that is the global maximizer $v^*$ of $J_{\tau,0}$ in $\mathcal{P}_2(\mathbb{R}^d)$. Moreover, for all $t \geq 0$,

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

where $W_2$ denotes the Wasserstein-2 distance.
**Theorem (Stability of $W_2$)**

Let $(v_t)_{t \geq 0}$ and $(v'_t)_{t \geq 0}$ be the solutions of the PG flow with parameters and initial data $\sigma, \tau, v_0$ and $\sigma', \tau', v'_0$, respectively. Then for all $\ell > 0$ and $t \in \mathbb{R}^+$,

$$W_2^2(v_t, v'_t) \leq e^{-2\beta_\ell t} W_2^2(v_0, v'_0) + \frac{|\sigma^2 - \sigma'^2|}{8\ell} \int_0^t \int_{\mathbb{R}^d} e^{2\beta_\ell (s-t)} |\nabla U(\theta)|^2 v'_s(d\theta) \, ds$$

$$+ \frac{1}{2\beta_\ell} \left( D|\tau - \tau'| + d|\sigma - \sigma'|^2 \right) (1 - e^{-2\beta_\ell t}),$$

where $\beta_\ell := \frac{\sigma^2}{2} \kappa - C_2(\tau) - L(\tau) - \ell |\sigma^2 - \sigma'^2|$. Moreover, if $\beta := \frac{\sigma^2}{2} \kappa - C_2(\tau) - L(\tau) > 0$ and $v^*$ and $v'^*$ are stationary solutions with $\sigma, \tau$ and $\sigma', \tau'$, respectively, then for all $\ell > 0$ such that $\beta_\ell = \beta - \ell |\sigma^2 - \sigma'^2| > 0$, we have

$$W_2^2(v^*, v'^*) \leq \frac{|\sigma^2 - \sigma'^2|}{16\ell \beta_\ell} \int_{\mathbb{R}^d} |\nabla U(\theta)|^2 v'^*(d\theta) + \frac{1}{2\beta_\ell} \left( D|\tau - \tau'| + d|\sigma - \sigma'|^2 \right).$$
Conclusion

We:

- proved the convergence of PG for continuous state and actions provided we add enough regularization
- quantified bias introduced by $\tau, \sigma$-regularization

What is next:

- relaxing regularization strength by establishing non-local Łojasiewicz inequality
- study full learning setting (e.g., actor-critic or reinforce)
Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22(98):1–76.

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*.

Bertsekas, D. P. et al. (2011). Dynamic programming and optimal control 3rd edition, volume ii. *Belmont, MA: Athena Scientific*. 
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.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). *Policy gradient methods for reinforcement learning with function approximation.* *Advances in neural information processing systems*, 12.