Quasi-optimal Reinforcement Learning with Continuous Actions

Yuhan Li*, Wenzhuo Zhou*, Ruoqing Zhu

Mar 2023

Yuhan Li
yuhanli8@illinois.edu

RL on Continuous Action Space

- Policy learning in continuous action space is important for many real-world RL applications (e.g. precision medicine, autonomous driving).
- Discretize action space
  - Large bias for coarse discretization (Lee et al., 2018a)
  - Curse of dimensionality for fine-grid (Chou et al., 2017)
- Gaussian distribution policy representation
  - Infinite support policy may assign risky actions (Fatemi et al., 2021)
  - Off-support bias in bounded action space (Chou et al., 2017)
- Motivation
  - Policy class with bounded support
  - Identify near-optimal action regions

- Revisit the Bellman optimality equation via a policy explicit view,

$$\mathcal{B}V^*(s) := \max_{\pi} \mathbb{E}_{a\sim\pi(\cdot|s),\ S^{t+1}|s,a}\left[R(S^{t+1}, s, a) + \gamma V^*(S^{t+1})\right] = V^*(s).$$

- Define a quasi-optimal counterpart for Bellman operator $\mathcal{B}_\mu$

$$\mathcal{B}_\mu V_\mu^*(s) = \max_{\pi\in\Delta_{\text{convex}}(\mathcal{A})} \int_{a\in\mathcal{A}} \left[Q_\mu^*(s,a)\pi(a|s) + \mu\text{prox}(\pi(a|s))\right] da,$$

where $\text{prox}(x) = x(1-x)$.

- $\mathcal{B}_\mu$ is a proximal approximation to $\mathcal{B}$
- $\mathcal{B}_\mu$ is a smoothed substitute for $\mathcal{B}$

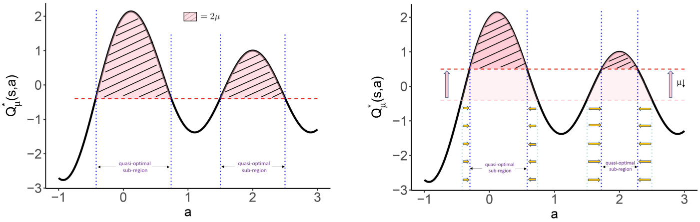
Quasi-optimal Support Region

- The induced optimal policy π_μ^* has closed-form solution.

$$\pi_\mu^*(a|s) = \left(\frac{Q_\mu^*(s,a)}{2\mu} - \frac{\int_{a\in\mathcal{W}_s} Q_\mu^*(s,a)da}{2\mu\sigma(\mathcal{W}_s)} + \frac{1}{\sigma(\mathcal{W}_s)} \right)^+,$$

- The threshold parameter μ controls the screening intensity.

$\mu \to 0, \pi \to$ Point Mass; $\mu \to \infty, \pi \to$ Uniform Distribution.



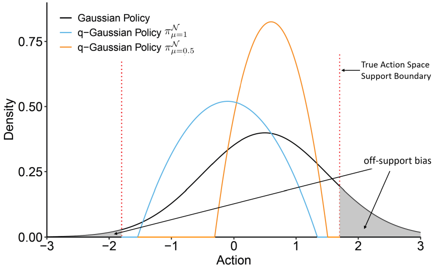
q-Gaussian Policy Distribution

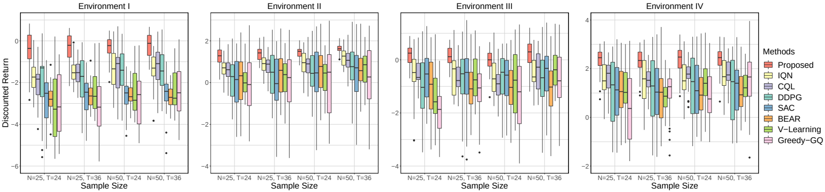
- Model $Q^*_\mu(s,a)$ as a concavely quadratic function:

$$Q^*_\mu(s,a) = -\alpha_1(s)a^2 + \alpha_2(s)a + \alpha_3(s).$$

- The optimal policy π^*_μ follows q-Gaussian distribution

$$\pi^*_\mu(a|s) = \left(\frac{\alpha_1(s)}{2\mu} \left(a + \frac{\alpha_2(s)}{2\alpha_1(s)} \right)^2 - \frac{3}{2} \left(\frac{\alpha_1(s)}{12\mu} \right)^{\frac{1}{3}} \right)^+ .$$





- Boxplots of the discounted return over 50 repeated experiments in 4 different environments with varying sample size.
- Environment I and II: Bounded action space to evaluate the potential of quasi-optimal learning for addressing off-support bias.
- Environment III and IV: Unbounded action space and more complex environments.

Summary

- Construct a novel quasi-optimal Bellman operator which is able to identify near-optimal action regions.
- Formalize an unbiased learning framework for estimating the designed quasi-optimal policy.
- Investigate the theoretical properties of the quasi-optimal learning algorithm, including the loss consistency, convergence analysis and the finite-sample bound for performance error.
- Empirical analyses in simulated experiments and a dose suggestion real application to Ohio Type 1 diabetes dataset.

Thank you!