

On the Convergence Rate of Density-Ratio Based Off-Policy Policy Gradient Methods

Jiawei Huang (UIUC), Nan Jiang (UIUC)

1. Preliminary

Density-Ratio Based Off-Policy Evaluation [1-3]

$$J(\pi) \sim \max_{w} \min_{Q} \mathcal{L}(\pi, w, Q)$$

- Evaluate $J(\pi) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t | \pi]$ of target policy π , given data generated from policy d^{μ} ; where d^{μ} denotes the normalized discounted state-action occupancy of μ
- Formulated as a mini-max game between w and Q; \mathscr{L} is the objective function built with d^{μ} ; w takes the role like density ratio d^{π}/d^{μ} ; Q takes the role like Q^{π} .
- A Natural Extension: Density ratio based off-policy policy improvement

$$\max_{\pi} J(\pi) \sim \max_{\pi} \max_{w} \min_{Q} \mathcal{L}(\pi, w, Q)$$

- Convergence? Any guarantee for the solution?
- Contribution:
 - Focus on $\mathscr{L}^D(\pi_\theta, w_\zeta, Q_\xi)$ built with d^D induced from the dataset $D \sim d^\mu$, which is a practical version of \mathscr{L} .

$$\max_{\theta \in \Theta} J(\pi_{\theta}) \sim \max_{\theta \in \Theta} \max_{\zeta \in Z} \min_{\xi \in \Xi} \mathcal{L}^{D}(\pi_{\theta}, w_{\zeta}, Q_{\xi})$$

 Two strategies: Convergence rate analysis Q; Guarantee of the quality of the solution \mathbf{V} ;

2. Bias (Lower Bound) of the Solution

Lower bound of the Performance Resulting from Biases:

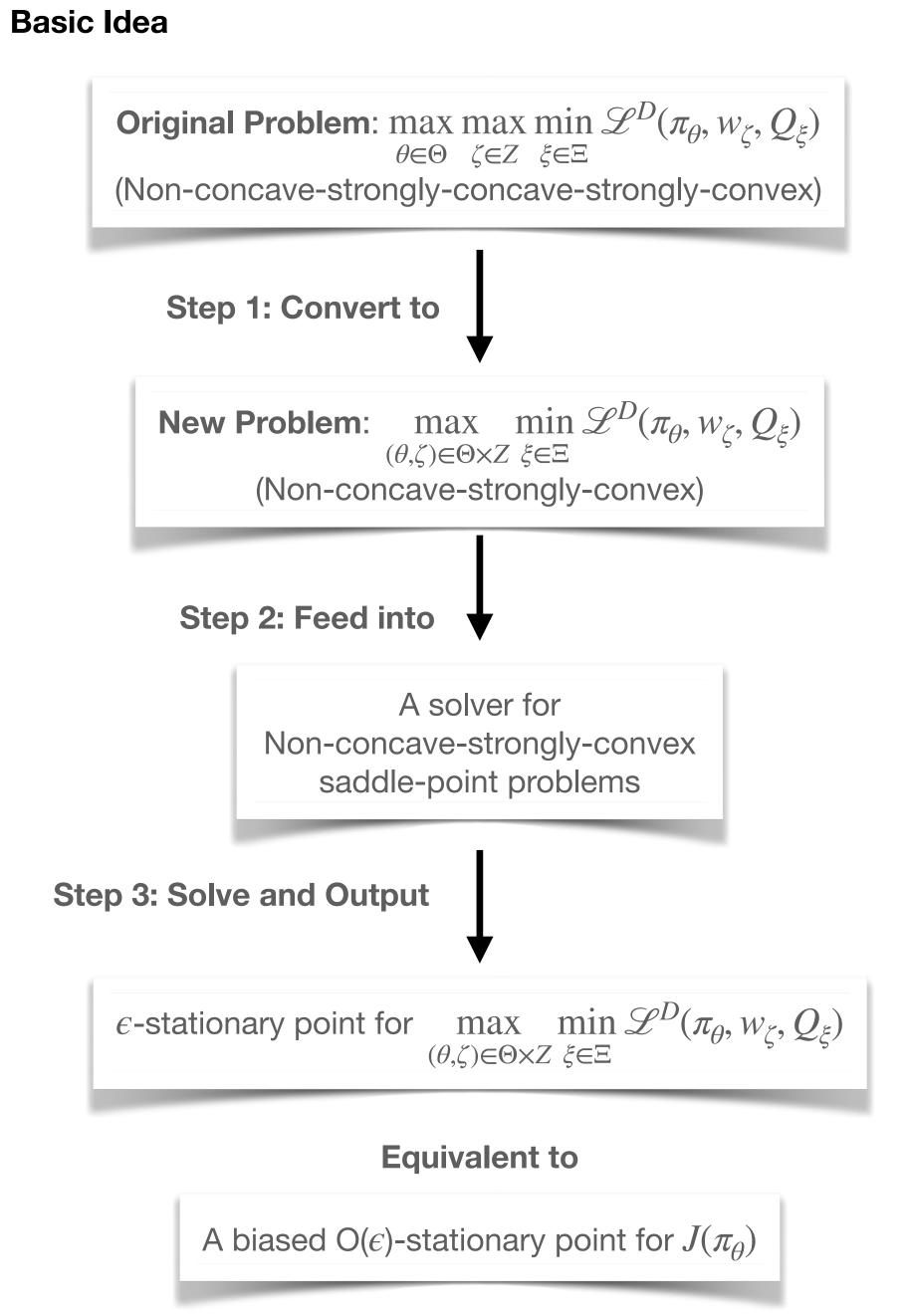
$$\|\nabla_{\theta} J(\pi_{\theta})\| \leq \|\nabla_{\theta} \max_{\zeta \in Z} \min_{\xi \in \Xi} \mathcal{L}^{D}(\pi_{\theta}, w_{\zeta}, Q_{\xi})\| + \underbrace{\epsilon_{reg} + \epsilon_{func} + \epsilon_{data}}_{bias}$$

 ϵ_{reg} : regularization error

 ϵ_{func} : mis-specification error (non-perfect function classes)

 ϵ_{data} : generalization error; $d^D pprox d^\mu$ but $d^D
eq d^\mu$

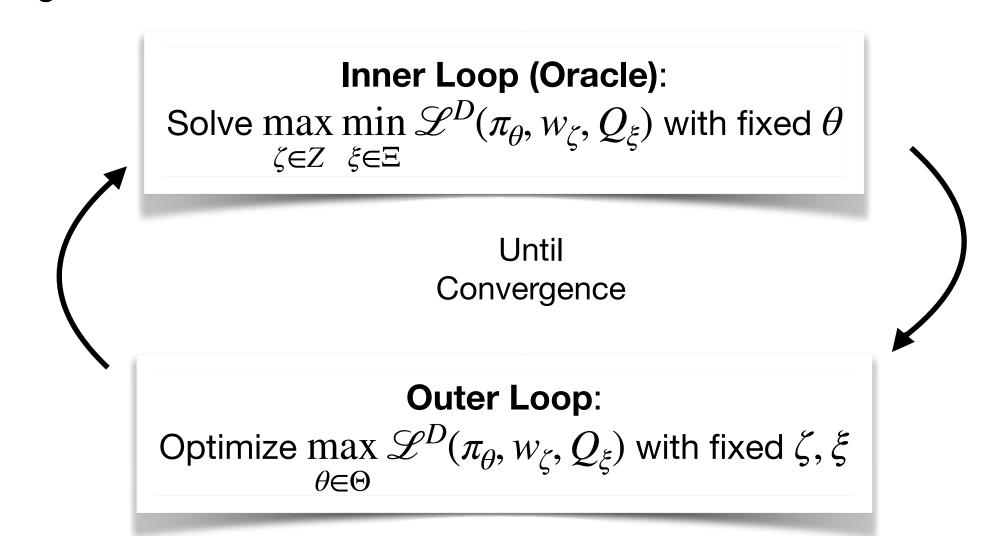
3. Strategy 1: From Max-Max-Min to Max-min



• $O(e^{-3})$ convergence rate can be guaranteed for Strategy 1 by choosing [1] as the solver.

4. Strategy 2: An Off-Policy Actor Critic with Distribution Correction

Algorithm Flow Chart



• Inner Loop: Abstract to an Oracle, s.t. for $\forall 0 < \beta < 1$, given arbitrary (ζ_0, ξ_0) , the oracle can return us ζ, ξ satisfying

$$\mathbb{E}[\|\zeta - \zeta_{\theta}^*\|^2 + \|\xi - \xi_{\theta}^*\|^2] \le \frac{\beta}{2} \mathbb{E}[\|\zeta_0 - \zeta_{\theta}^*\|^2 + \|\xi_0 - \xi_{\theta}^*\|^2] + O(\epsilon^2)$$

where $(\zeta_{\theta}^*, \xi_{\theta}^*)$ is the saddle-point of $\max_{\zeta \in Z} \min_{\xi \in \Xi} \mathscr{L}^D(\pi_{\theta}, w_{\zeta}, Q_{\xi})$

- Concrete example of Oracle: A SVRE (Stochastic Variance-Reduced) Extragradient) Algorithm inspired by [5]
- Complexity of our SVRE is $O(\epsilon^{-2} \log \beta)$, without dependence on the size of dataset D
- Outer Loop: An off-policy SRM (Stochastic Recursive Momentum) algorithm inspired by [6]
- The convergence rate of Strategy 2 is $O(\epsilon^{-4})$

References

- [1] Ofir Nachum et. al. AlgaeDICE: Policy gradient from arbitrary experience. arXiv preprint, 2019
- [2] Nan Jiang, Jiawei Huang. Minimax value interval for off-policy evaluation and policy optimization. NeurlPS 2020
- [3] Mengjiao Yang et. al. Off-policy evaluation via the regularized lagrangian. NeurlPS 2020
- [4] Luo Luo, Ye Haishan and Zhang Tong. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. NeurIPS 2020.
- [5] Tatjana Chavdarova et. al. Reducing noise in gan training with variance reduced extragradient. NeurlPS 2019
- [6] Huizhuo Yuan et. al. Stochastic recursive momentum for policy gradient methods. arXiv preprint 2020