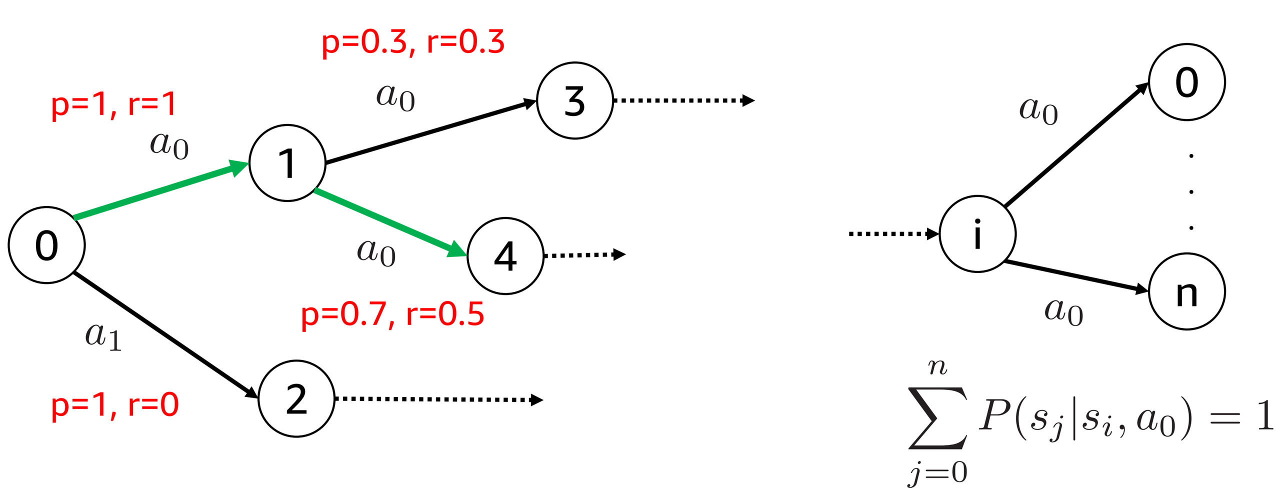
Stein's Method for Policy Gradients Methods in Deep Reinforcement Learning

Sahika Genc, Ph. D.
Amazon Artificial Intelligence, Seattle, WA

Policy Optimization and Surrogate Loss Using on Advantage Function

Markov Decision Process (MDP)

$$\begin{aligned} \mathcal{S} \quad & \text{Set of states} \\ \mathcal{A} \quad & \text{Set of actions} \\ P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R} \quad & \text{Transition probability distribution} \\ c : \mathcal{S} \to \mathbb{R} \quad & \text{Cost function} \\ \rho_0 : \mathcal{S} \to \mathbb{R} \quad & \text{Initial state distribution} \end{aligned}$$



$$\sum_{j=0}^{n} P(s_j | s_i, a_0) = 1$$

Stochastic Policy

Goal: Obtain a policy to restrict the behavior of the MDP to obtain a desired behavior.

$$\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$$

Desired Behavior: Minimize (maximize) expected discounted cumulative cost (reward).

$$\eta(\pi) = \mathbb{E}_{\mathcal{S},\mathcal{A}} \left[\sum_{t=0}^{\infty} \gamma^t c(s_t) \right]$$

Policy Optimization and Advantage Function

Express the expected reward of another policy in terms of the advantage over the current policy, accumulated over timesteps.

$$A_\pi(s, a) = Q_{s,a} - V_\pi(s)$$

Q-function is the future cumulative discounted reward is from s take action a

Value function is the future cumulative discounted reward is from s take any action

Any policy update that has non-positive expected advantage at every state will reduce the expected cumulative reward or keep it constant when the advantage is zero at every state.

$$\pi \to \tilde{\pi} \Rightarrow \eta(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\tilde{\pi}} \sum_a \tilde{\pi}(a|s) A_\pi(s, a)$$

Difficult to optimize directly → Local approximation using visitation frequency from the old policy to update to the new policy.

$$L(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_\pi \sum_a \tilde{\pi}(a|s) A_\pi(s, a)$$

Conservative policy iteration update provides explicit lower bound on the improvement from the old policy to the new one.

$$\pi_{new}(a|s) = (1 - \alpha)\pi_{old}(a|s) + \alpha \pi'(a|s), \quad \pi' = arg \min_{\pi'} L_{\pi_{old}}(\pi')$$

$$\eta(\pi_{new}) \leq L_{\pi_{old}}(\pi_{new}) + \frac{2\epsilon\gamma}{(1-\gamma)^2}\alpha^2, \quad \alpha, \gamma \in [0, 1] \qquad (1)$$

Stein's Method for Policy Optimization

Goal: Kernelized Stein Discrepancy (KSD) to bound step size in policy update to robustly allow large policy updates.

Stein's method is a general theoretical tool for bounding differences between distributions. The score function is independent of normalization.

$$s_p(x) = \nabla x \log p(x) = \frac{\nabla_x p(x)}{p(x)}$$

KSD for policy update given a positive definite kernel and reproducing kernel Hilbert space (RKHS) has a closed form solution.

$$\mathbb{S}(\pi_{new}, \pi_{old}) = \mathbb{E}_{x,x' \ \pi_{new}} \approx \frac{1}{n(n-1)} \sum_{i \neq j} \kappa_{\pi_{old}} = trace(\mathcal{A}^x_{\pi_{old}} \mathcal{A}^x_{\pi_{old}} k(x, x'))$$

Policy Optimization and KL divergence

Lower bound on the improvement update shown in Equation (1) can be written in terms of total variation divergence.

$$\alpha = D^{max}_{TV}(\pi_{old}, \pi_{new}) \qquad D_{TV}(p||q) = \frac{1}{2}\sum_i |p_i - q_i|$$

Using the relation between total variation divergence and KL divergence, obtain a bound in terms of maximum KL divergence.

$$\eta(\pi_{new}) \leq L_{\pi_{old}}(\pi_{new}) + \frac{2\epsilon\gamma}{(1-\gamma)^2} \underbrace{D^{max}_{KL}(\pi_{old}) || \pi_{new})}_{\text{Replaced } \alpha \text{ in Equation (1)}}$$

Fisher divergence (FD) vs Stein:

- The total variance divergence squared is the FD given the score function.
- KSD is a smoothed version of Fisher divergence with a kernel.

KL divergence vs Stein:

- FD is the derivative of KL when variables are perturbed by i.i.d. Gaussian.
- KSD is the derivative of KL when variables are perturbed by smooth functions in RKHS.

Observations:

- Policy optimization methods using REINFORCE and Fisher or KL divergence metrics have been shown to converge on complex problems.
- When policy parameters are perturbed by smooth functions in RKHS as in REINFORCE, then policy optimization problem can be solved without the need to find the correct normalization factor.

References

- Qiang Liu, Jason D. Lee, and Michael Jordan. 2016. A kernelized stein discrepancy for goodness-of-fit tests. In Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16), Maria Florina Balcan and Kilian Q. Weinberger (Eds.), Vol. 48. JMLR.org 276-284
- Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. 2016. A kernel test of goodness of fit. In Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16), Maria Florina Balcan and Kilian Q. Weinberger (Eds.), Vol. 48. JMLR.org 2606-2615.
- Dhariwal, P., Hesse, C., Klimov, O., Nichol, A., Plappert, M., Radford, A., Schulman, J., Sidor, S., Wu, Y., and Zhokhov, P. OpenAI Baselines. https://github. com/openai/baselines, 2017.
- Jie, T. and Abbeel, P. On a connection between importance sampling and the likelihood ratio policy gradient. In Laf- ferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds.), *Advances in Neural Infor- mation Processing Systems 23*, pp. 1000–1008. Curran Associates, Inc., 2010.
- Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *NIPS*, 2016.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv. org/abs/1707.06347.
- Wang, Z. M., Feng, Y., and Liu, Q. Learning to sample using stein discrepancy. 2016.

ICML Workshop, Stein's Method, February 25-March 1, 2019, Los Angeles, CA USA