ACCELERATED POLICY EVALUATION WITH ADAPTIVE IMPORTANCE SAMPLING

Mengdi Xu[1], **Peide Huang**[1], **Fengpei Li**[2,3], **Jiacheng Zhu**[1], **Xuewei Qi**[4], **Kentaro Oguchi**[4], **Zhiyuan Huang**[5], **Henry Lam**[2], **Ding Zhao**[1]
1. Carnegie Mellon University 2. Columbia University 3. Morgan Stanley AI CoE
4. Toyota Motor North America R&D 5. Tongji University

ABSTRACT

Policy evaluation with rare events, in general, requires a large amount of data. Most current acceleration methods for Markov chains, such as cross-entropy methods, and adaptive importance sampling methods, focus on small finite spaces. They become unaffordable in large finite spaces and suffer from the curse of dimensionality due to the discretization when continuous spaces. In this paper, we propose an accelerated policy evaluation method with adaptive importance sampling, that is scalable to Markov decision processes with large discrete or continuous spaces, by treating environment nature as an agent and integrating with function approximators for both value function and importance distribution. The experiment results in minigrid and highway environments show that when the rare probability gets smaller, our method is better in terms of the low variance at convergence and the high number of sampled rare events.

1 INTRODUCTION

Machine learning-based sequential decision-making algorithms including reinforcement learning methods have made impressive success in various areas, but most still lack interpretability. One way to ensure safety is to do a reliable policy evaluation before deployment, especially when the applications are safety-critical or expensive, such as medical treatments, bankruptcy, autonomous driving, and healthcare robotics. Rare failure cases in safety-critical applications make the evaluation more challenging due to the large amount of data required. Most current methods in estimating rare event-related probabilities focuses on finite state-space Markov chains (Heidelberger, 1995; Ahamed et al., 2006) with small discrete state and action spaces, such as cross-entropy methods, adaptive monte carlo methods (Desai & Glynn, 2001), and adaptive importance sampling methods (Ahamed et al., 2006). They rely on discretization when continuous spaces suffering from the curse of dimensionality. Discretization also drops the environment or action structure information and thus biases evaluation results.

In this paper, we propose an efficient policy evaluation method with adaptive importance sampling, that is scalable to Markov Decision Processes (MDPs) with large discrete or continuous state and action spaces. We aim to estimate the expected costs till termination, which can be rare-event related probabilities or expected rewards. We assume that both the environment transition probability and the evaluation policy are known. To handle MDPs, we treat uncertainties in environment transition probabilities as decisions made by the environment nature. Therefore, the importance weights are related to the policies of the environment. We propose to use the conditional normalizing flow to represent the environment policy (analogous to the importance distribution). Instead of using tables to store estimated values, we use a function approximator with high flexibility and capacity as the representation. The experiments show that our method achieves better performance than other baselines in terms of the low variance at convergence and the high number of sampled rare events.

2 ADAPTIVE IMPORTANCE SAMPLING FOR DISCRETE MDP

In this section, we introduce the adaptive importance sampling method in MDP settings by treating uncertainties in environment transition probabilities as decisions made by the environment nature.

Our method stems from the adaptive stochastic approximation (ASA) (Ahamed et al., 2006) for the Markov chain. ASA is an online algorithm with the value function and importance distribution iteratively updating based on recently collected data pairs using dynamic programming paradigm. At step n, the value function $J^{(n+1)}(x_n)$ of current state x_n is updated as

$$J^{(n+1)}(x_n) = J^{(n)}(x_n) + a\Big[-J^{(n)}(x_n) + \Big(g(x_n, x_{n+1}) + J^{(n)}(x_{n+1})\Big) \cdot \frac{p_{x_n x_{n+1}}}{p^{(n)}_{x_n x_{n+1}}}\Big], \tag{1}$$

where $g(x_n, x_{n+1})$ is the current cost. $p_{x_n x_{n+1}}$ is the true transition probability and $p^{(n)}_{x_n x_{n+1}}$ is the importance distribution at step n. The un-normalized importance distribution is then updated as

$$\tilde{p}^{n+1}_{x_n x_{n+1}} = \max\Big(\delta,\ p_{x_n x_{n+1}} \cdot \frac{g(x_n, x_{n+1}) + J^{n+1}(x_{n+1})}{J^{n+1}(x_n)}\Big), \tag{2}$$

where δ is a positive value much smaller than true transition probability. ASA is proven to find the zero-variance importance distribution and converge to the true value J^* with diminishing step size a asymptotically.

In MDP, the state transition probability relates to both the agent A's evaluation policy π_A and the environment's true transition probability $p(x_{n+1}|a_A, x_n)$. The corresponding Markov chain transition $p_{x_n x_{n+1}} = \sum_{a_A \in \mathcal{A}_A} \pi_A(a_A|x_n) p(x_{n+1}|a_A, x_n)$. $\mathcal{A}_A$ is the agent action space. Therefore, the importance weight at step n in Eq. 1 changes to

$$\frac{p_{x_n x_{n+1}}}{p^{(n)}_{x_n x_{n+1}}} = \frac{\sum_{a_A \in \mathcal{A}_A} \pi_A(a_A|x_n) p(x_{n+1}|a_A, x_n)}{\sum_{a_A \in \mathcal{A}_A} \pi^{(n)}_A(a_A|x_n) p^{(n)}(x_{n+1}|a_A, x_n)}, \tag{3}$$

where $\pi^{(n)}_A(a_A|x_n)$ and $p^{(n)}(x_{n+1}|a_A, x_n)$ are the importance distributions for the agent policy and the environment transition probability, respectively.

Treating the environment nature as an agent is widely used in robust learning, multi-agent RL and game theoretic approaches (Zhang et al., 2020; Mehta et al., 2020; Pinto et al., 2017). It reduces the computation complexity if the selected environment agent E's action space $\mathcal{A}_E$ has a smaller dimension than that of the state space $\mathcal{X}$. The uncertainty in the environment transition probability transfers to the stochasticity of policy π_E. The simulation then steps based on a deterministic mapping f_E. Formally, at step n,

$$p(x_{n+1}|a_{A,n}, x_n) = \pi_E(a_{E,n}|a_{A,n}, x_n) \tag{4}$$

$$x_{n+1} = f_E(x_n, a_{A,n}, a_{E,n}) \tag{5}$$

$$a_{E,n} = f_E^{-1}(x_n, a_{A,n}, x_{n+1}) \tag{6}$$

For simplicity, we only focus on the adaptive importance sampling over environment transition probability $p(x_{n+1}|a_A, x_n)$ and $\pi^{(n)}_A(a_A|x_n) = \pi_A(a_A|x_n), \forall n$ in Eq. 3. We approximate Eq. 3 with an online stochastic paradigm Eq. 7 to get rid of the integral over $\mathcal{A}_A$.

$$\frac{p_{x_n x_{n+1}}}{p^{(n)}_{x_n x_{n+1}}} \approx \frac{p(x_{n+1}|a_{A,n}, x_n)}{p^{(n)}(x_{n+1}|a_{A,n}, x_n)} = \frac{\pi_E(a_E|a_{A,n}, x_n)}{\pi^{(n)}_E(a_E|a_{A,n}, x_n)} \tag{7}$$

where a_E follows Eq. 6. The approximation Eq. 7 becomes exact if the agent policy π_A is deterministic.

Therefore, the update rules of value function J and environment policy π_E in discrete MDP settings are derived by replacing Markov Chain transition probabilities with environment policy probabilities.

$$J_{TD} = \Big(g(x_n, x_{n+1}) + J^{(n)}(x_{n+1})\Big) \cdot \frac{\pi_E(a_E|a_{A,n}, x_n)}{\pi^{(n)}_E(a_E|a_{A,n}, x_n)} \tag{8}$$

$$J^{(n+1)}(x_n) = J^{(n)}(x_n) + a\Big[-J^{(n)}(x_n) + J_{TD}\Big] \tag{9}$$

$$\tilde{\pi}^{n+1}_E(a_E|a_{A,n}, x_n) = \max\Big(\delta, \pi_E(a_E|a_{A,n}, x_n)\Big(\frac{g(x_n, x_{n+1}) + J^{n+1}(x_{n+1})}{J^{n+1}(x_n)}\Big)\Big) \tag{10}$$

In this paper, we are interested in a specific application, rare-event probability estimation. In this case, $g(\cdot, \cdot)$ is an indicator random variable got from the simulated environment. $J(x)$ represents the probability of entering rare event set starting from state x.

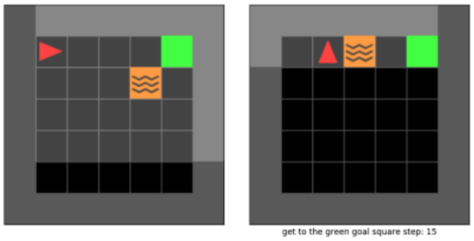


Figure 1: A minigrid environment w/ rare events

methods	MC		ours	
metric (×1e3)	mean	std	mean	std
$J(x_1)$	3.99	0.38	3.90	0.24
$J(x_2)$	3.93	0.47	3.94	0.23
$J(x_3)$	4.00	0.50	3.94	0.17
#transitions	78300	10600	8884	2595
#episodes	19306	2612	552	134
95% CI	$n = 15$		$n = 6$	

Table 1: Estimation results

2.1 A GRIDWORLD TOY EXAMPLE

We first validate the proposed adaptive importance sampling method in discrete MDP settings with a minigrid (Chevalier-Boisvert et al., 2018) toy example as in Fig. 1. The rare event means that the yellow lava moves to the top row and may happen in each step. As shown in Tab. 2, the proposed method requires 10 percent of the data for the monte carlo (MC) method but has a smaller variance.

3 ADAPTIVE IMPORTANCE SAMPLING FOR CONTINUOUS MDPS

In this section, we introduce how to extend to large discrete or continuous MDPs. The key part is to select proper function approximators to represent the value function and the environment agent E's importance policy. Before doing that, we first establish some notations. We denote the parameters of J approximators as ψ, and parameters of environment agent E's importance policy as θ. At step i, a data pair $d_i = (x_i, a_{A,i}, a_{E,i}, x_{i+1}, g(x_i, x_{i+1}), \rho_i)$ is appended to a history dataset $\mathcal{D}$. Assume $x \in \mathcal{X} \subset \mathbb{R}^{d_x}$, $a_A \in \mathcal{A}_A \subset \mathbb{R}^{d_a}$, and $a_E \in \mathcal{A}_E \subset \mathbb{R}^{d_e}$. $g(x_i, x_{i+1})$ is the one-step cost. $m = |\mathcal{D}|$ is the cardinality of $\mathcal{D}$. $\rho_i = \pi_E(a_{E,i}|a_{A,i}, x_i)/\pi^{(i)}_{E,\theta}(a_{E,i}|a_{A,i}, x_i)$ is the importance weight.

3.1 VALUE FUNCTION APPROXIMATION

At step n, the target value of state x_n follows the Bellman equations with importance weights as in Eq. 11. Even if $p_{x_n x_{n+1}}$ is accessible, calculating the expectation is still intractable. Therefore, we approximate the target with one-step TD target defined in Eq. 12. At each update iteration, the target of each data pair in buffer $\mathcal{D}$ is calculated based on the most recent updated value function.

$$J_\psi(x_n) = \mathbb{E}_{\pi^{(n)}_{E,\theta}}\left[(g(x_n, x_{n+1}) + J_\psi(x_{n+1})) \cdot \rho_n\right] \tag{11}$$

$$J_{\psi,TD}(x_n) = (g(x_n, x_{n+1}) + J_\psi(x_{n+1})) \cdot \rho_n \tag{12}$$

Deep Neural Network Approximator To be scalable to large discrete or continuous spaces, one way is to regress J with an expressive deep neural network (DNN). DNNs are parametric universal approximators with low prediction complexity. However, DNNs are sensitive to imbalanced data and suffer from the catastrophic forgetting problem (Krawczyk, 2016; Chrysakis & Moens, 2020), especially in online learning settings. Methods that stabilize the online learning process are required.

Gaussian Process Regression Approximator One competitive model to a DNN is a Gaussian Process (GP). The equivalence between GPs and infinitely wide DNNs is derived in (Lee et al., 2017). GPs are data-efficient and flexible in making predictions as a non-parametric model, but sacrifices the prediction complexity (Rasmussen, 2003).

3.2 IMPORTANCE POLICY APPROXIMATION

After fitting the value function approximator, the environment's policy are updated based on the most recent data pair d_n. The unnormalized target density at value $a_{E,n}$ conditioned on $C_n = [a_{A,n}, x_n]$ is defined in Eq. 13. To deal with continuous spaces, we achieve the modification by adding a normal distribution $\mathcal{N}(a_{E,n}, \sigma)$ with σ much smaller than that of the ground truth policy $\pi_E(a_E|x, a_A)$. The target probability density function conditioned on C_n is defined in Eq. 14, where

$\beta = \sqrt{2\pi}\sigma(\tilde{\pi}_E(a_{E,n}|C_n) - \pi_{E,\theta}(a_{E,n}|C_n))$ and $\gamma = 1/(1+\beta)$ is the normalizing constant.

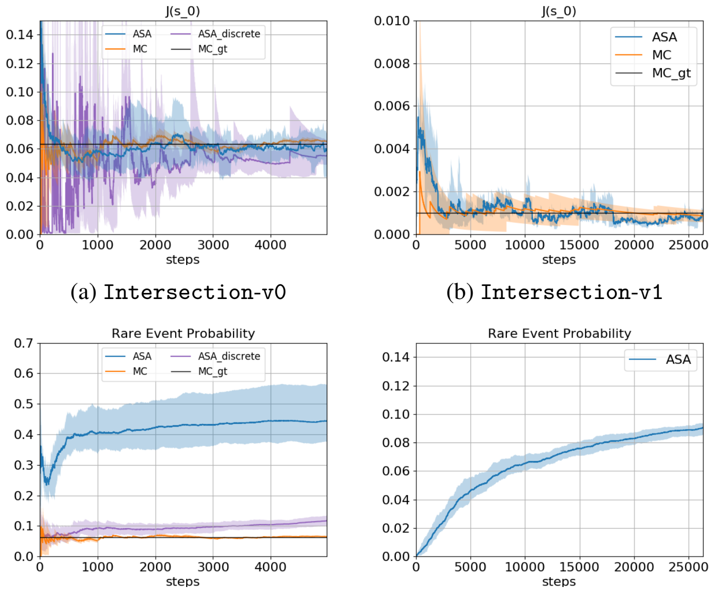
$$\tilde{\pi}_E(a_{E,n}|C_n) = \pi_E(a_{E,n}|C_n)\Big(\frac{g + J_\psi(x_{n+1})}{J_\psi(x_n)}\Big) \tag{13}$$

$$p_E(a_E|C_n) = \gamma(\pi_{E,\theta}(a_E|C_n) + \beta \cdot \mathcal{N}(a_{E,n}, \sigma)) \tag{14}$$

Conditional Normalizing Flow approximator Note that the environment's policy $\pi_E(a_E|a_A, x)$ is a distribution conditioned on continuous random variables. The function approximation $\pi_{E,\theta}$ needs to (1) have low sample complexity, (2) have interpolation/extrapolation ability and (3) be flexible enough to model multi-mode distributions. In this paper, we obtain a close-form parametric representation of $\pi_E(a_E|a_A, x)$ usign a conditional Normalizing Flow (cNF). cNFs are recently proposed in (Papamakarios et al., 2017; Winkler et al., 2019; Oh & Valois, 2020), which are generative models that use inevitable mappings to transform a simple probability distribution into a complex one conditioned on other random variables. Compared with sample-based representation approaches such as MCMC, cNF directly generates one sample by calling one forward path (or inverse path based on implementation), which will dramatically accelerate the evaluation process. cNF is quite expressive and can model distributions that go beyond single-mode Gaussian distributions.

3.3 Experiment in highway-intersection environment

We test the proposed method with GP as the function approximator in highway environments (Leurent, 2018). The rare event is defined as the crash. As in Fig. 2, in `Intersection-v0` with rare probability around 0.06, the proposed method is more stable compared with that using discretization and achieve similar performance with MC. In `Intersection-v1` with rare probability around 0.001, the proposed method has smaller variance than MC. The sampled rare event probability is 7.5 times and 90 times the ground truth probability in `v0` and `v1`, respectively.



(a) `Intersection-v0` (b) `Intersection-v1`

(c) `Intersection-v0` (d) `Intersection-v1`

Figure 2: Experiment results with GP in highway-intersection

4 Conclusion

We propose an efficient policy evaluation method scalable to MDPs with large discrete or continuous spaces in the presence of rare events. The experiment results show that when the rare probability gets smaller, our method is better in terms of the low variance at convergence and the high number of sampled rare events. Future work involves experimenting with even smaller rare event probabilities.

REFERENCES

TP Imthias Ahamed, Vivek S Borkar, and S Juneja. Adaptive importance sampling technique for markov chains using stochastic approximation. *Operations Research*, 54(3):489–504, 2006.

Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for openai gym. `https://github.com/maximecb/gym-minigrid`, 2018.

Aristotelis Chrysakis and Marie-Francine Moens. Online continual learning from imbalanced data. In *International Conference on Machine Learning*, pp. 1952–1961. PMLR, 2020.

Paritosh Y Desai and Peter W Glynn. A markov chain perspective on adaptive monte carlo algorithms. In *Proceeding of the 2001 Winter Simulation Conference (Cat. No. 01CH37304)*, volume 1, pp. 379–384. IEEE, 2001.

Philip Heidelberger. Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 5(1):43–85, 1995.

Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

Edouard Leurent. An environment for autonomous driving decision-making. `https://github.com/eleurent/highway-env`, 2018.

Bhairav Mehta, Manfred Diaz, Florian Golemo, Christopher J Pal, and Liam Paull. Active domain randomization. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 1162–1176. PMLR, 2020.

Geunseob Oh and Jean-Sebastien Valois. Hcnaf: Hyper-conditioned neural autoregressive flow and its application for probabilistic occupancy map forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14550–14559, 2020.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pp. 2338–2347, 2017.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pp. 2817–2826, 2017.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pp. 63–71. Springer, 2003.

Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*, 2019.

Kaiqing Zhang, Tao Sun, Yunzhe Tao, Sahika Genc, Sunil Mallya, and Tamer Basar. Robust multi-agent reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 33, 2020.