Accelerated Policy Evaluation With Adaptive Importance Sampling

ICLR 2021 Workshop on Security and Safety in Machine Learning Systems

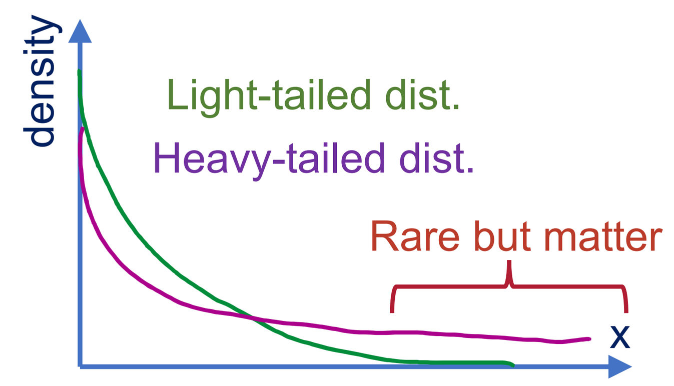
Mengdi Xu[1], Peide Huang[1], Fengpei Li[2,3], Jiacheng Zhu[1], Xuewei Qi[4], Kentaro Oguchi[4], Zhiyuan Huang[5], Henry Lam[3], Ding Zhao[1]

1. Carnegie Mellon University 2. Columbia University 3. Morgan Stanley AI CoE 4. Toyota Motor North America R&D 5. Tongji University. Contact: mengdixu@andrew.cmu.edu

Motivation

How to evaluate a policy in the presence of rare events?

Rare events are important in safety-critical applications, such as medical treatments, marketing and finance, autonomous driving, and healthcare robotics.



[Stock Market Crash, Wiki]

[Tesla Autopilot, 2019]

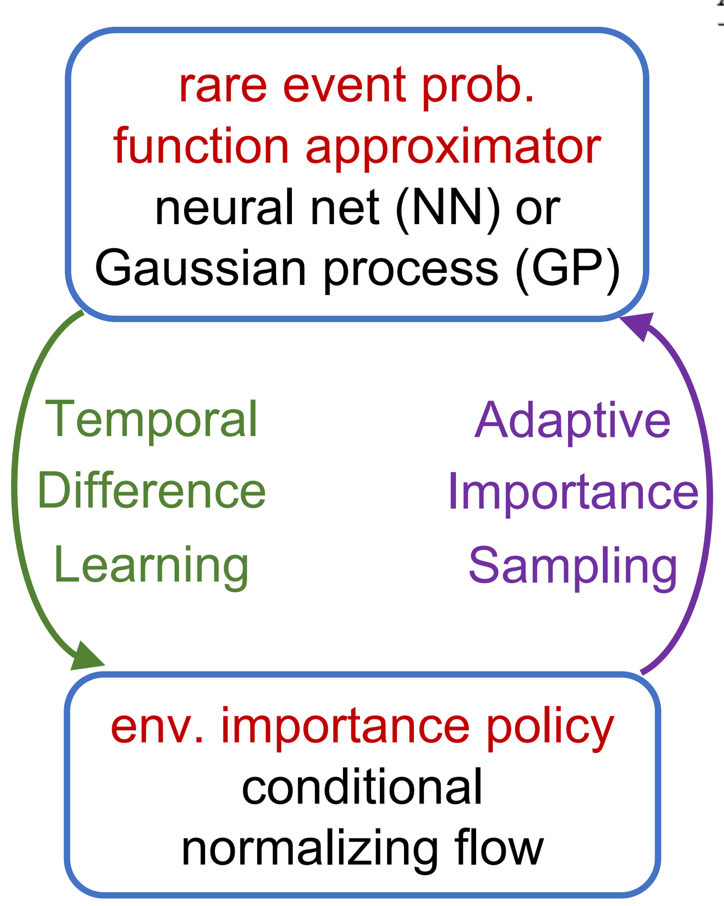
Existing Methods in estimating rare event-related probabilities in finite state-space Markov chains:

- cross-entropy methods [De Boer, 2001]
- adaptive Monte Carlo [Desai, 2001]
- adaptive importance sampling [Ahamed, 2006]

Limitations:

- Small discrete state and action spaces (limited scalability);
- Rely on discretization when continuous spaces (drops the environment or action structure information).

Algorithm



Iterative Updating Scheme

Check the paper for more info!

Algorithm 1 ASA for continuous state action space MDPs

Input: Evaluation policy π_A, Simulation envirionment E, Ground truth transition prob $\pi_E(a_E|a_A, x)$, Horizon N, initial value function parameters ψ, value function learning rate α_J, initial NF parameters $\theta = \{\theta, \phi\}$, NF learning rate α_π

Initialization: $n = 0$

Pretrain $\pi_{E,\theta}(\cdot|a_A, x)$ with target conditional probability $\pi_E(\cdot|a_A, x)$, $\forall a_A \in \mathcal{A}_A, x \in \mathcal{X}$

for $n = 0$ **to** $N-1$ **do**
  Reset environment
  repeat
    Sample agent action $a_A \sim \pi_A(\cdot|x)$
    Sample env action $a_E \sim \pi_{E,\theta}(\cdot|a_A, x)$
    Execute a_A and a_E and observe cost g, next state x'
    Calculate importance weight ρ
    Add data pair $d = (x, a_A, a_E, x', g, \rho)$ to buffer $\mathcal{D}$
    Update value function parameters ψ with gradient descent
    Update transition target based on d
    Train $\pi_{E,\theta}(\cdot|a_A, x)$ with gradient descent
    $x \leftarrow x'$
    $n \leftarrow n+1$
  until episode finish
end for

Method

An accelerated and scalable policy evaluation method suitable for Markov Decision Processes (MDPs) with large discrete or continuous state and action spaces.

Estimate expected costs till termination, $J^*(s) = \mathbb{E}_{x_0=s}\left[\sum_{n=0}^{\infty} g(x_n, x_{n+1})\right]$

e.g., rare event probability.

the prob. of hitting rare termination set before hitting other termination sets. $g(x,y) = \mathbf{I}_{y \in R}$

Adaptive IS for discrete Markov chains [Ahamed, 2006]

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

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

Adaptive stochastic approximation for discrete MDPs

- Contribution 1: Extend to MDPs by treating environment nature as an agent with its policy as the importance distribution

$$p(x_{n+1}|a_{A,n}, x_n) = \pi_E(a_{E,n}|a_{A,n}, x_n)$$
$$x_{n+1} = f_E(x_n, a_{A,n}, a_{E,n})$$
$$a_{E,n} = f_E^{-1}(x_n, a_{A,n}, x_{n+1})$$

- stochastic approximation

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

- Iterative update rule:

Importance weight: $\boxed{\rho_n}$

$$J^{(n+1)}(x_n) = J^{(n)}(x_n) + a\Big[-J^{(n)}(x_n) + \Big(g(x_n, x_{n+1}) + J^{(n)}(x_{n+1})\Big) \cdot \boxed{\frac{\pi_E(a_E|a_{A,n}, x_n)}{\pi^{(n)}_E(a_E|a_{A,n}, x_n)}}\Big]$$

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

Adaptive stochastic approximation for continuous MDPs

- Contribution 2: Integrate adaptive IS with function approximations
- value function approx.: GP or NN; batch gradient descent (GD)

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

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

- importance policy approx.: cMAF [Papamakarios, 2018]; GD

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

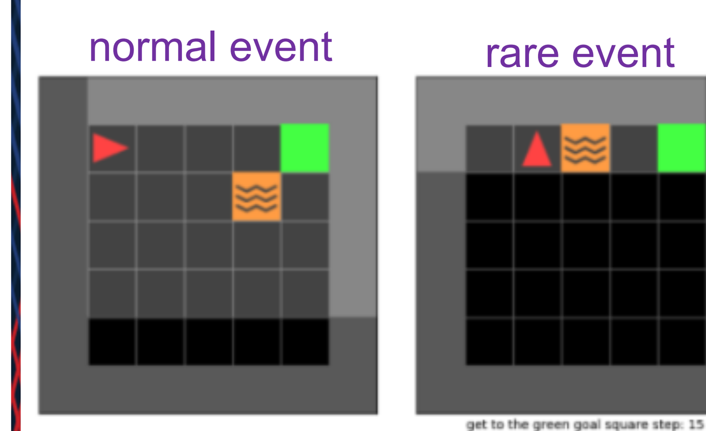
Importance policy target density: $p_E(a_E|C_n) = \gamma(\pi_{E,\theta}(a_E|C_n) + \beta \cdot \mathcal{N}(a_{E,n}, \sigma))$

with $\beta = \sqrt{2\pi}\sigma(\tilde{\pi}_E(a_{E,n}|C_n) - \pi_{E,\theta}(a_{E,n}|C_n))$, $\gamma = 1/(1+\beta)$

Experiments

Validation of Our MDP Evaluation Scheme

- Ours requires an order of magnitude fewer data than MC.
- Ours has a smaller variance than MC.



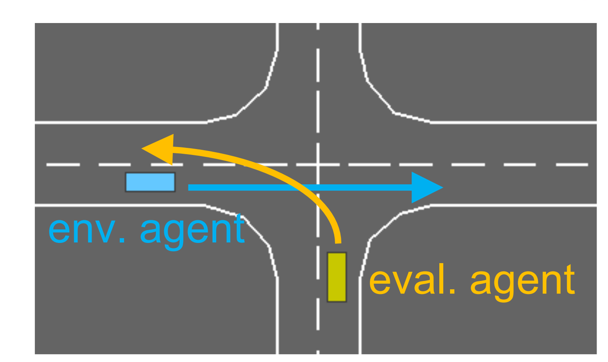
[gym-minigrid, 2018]

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	$n=15$		$n=6$	

Performance of Function Approximation

- In intersection-v0, ours is more stable than discretization.
- In intersection-v1, ours has smaller variance than MC. indicating better performance with smaller rare prob.
- Ours samples more rare events → closer to zero-variance dist..

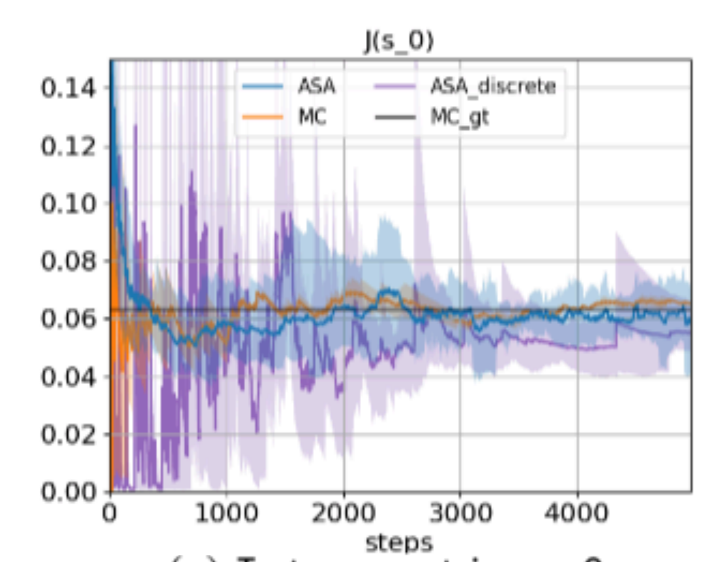
GP as J function approximator; rare event defined as crash.



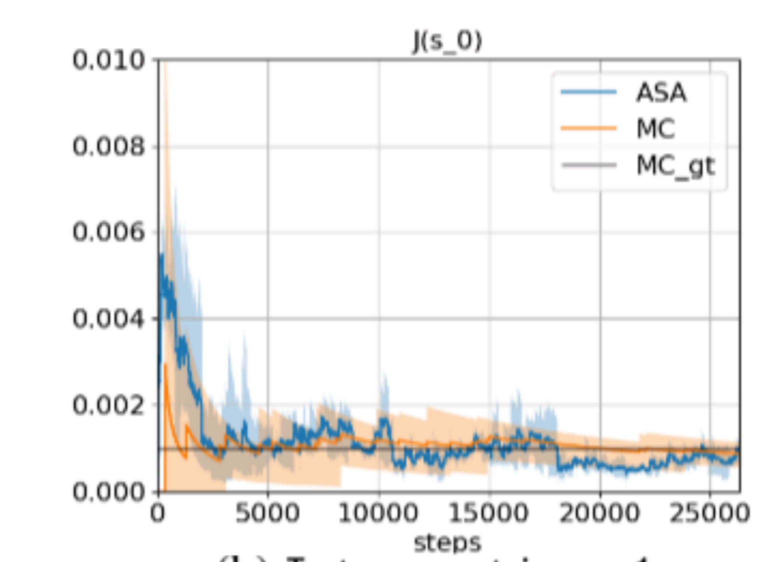
[highway-env, 2018]

methods	Intersection-v0	Intersection-v1
MC	0.06	0.001
ASA_discrete	0.12 (2 times)	x
ASA (ours)	0.44 (7.5 times)	0.09 (90 times)

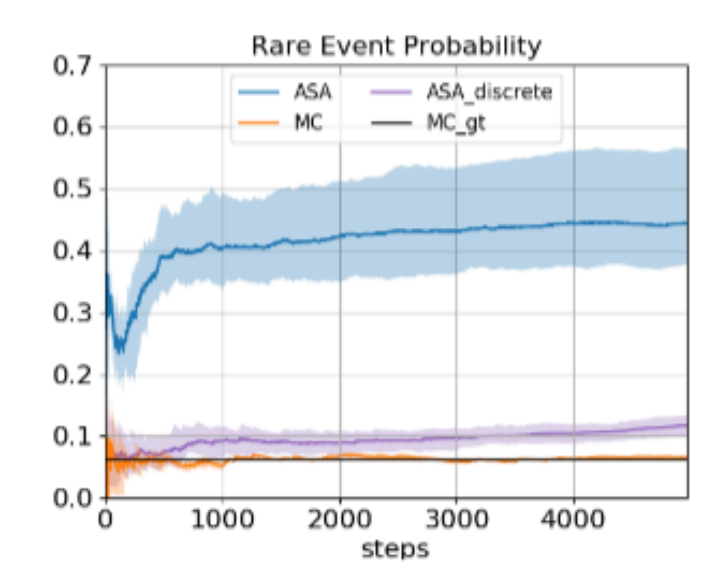
Sampled rare event probability



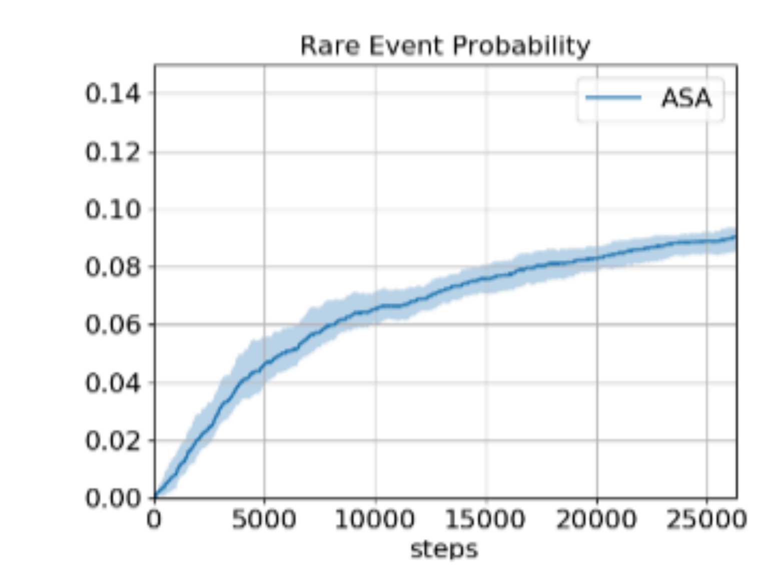
(a) Intersection-v0



(b) Intersection-v1



(c) Intersection-v0



(d) Intersection-v1