Variance Penalized On-Policy and Off-Policy Actor-Critic

Arushi Jain, *Gandharv Patil, Ayush Jain, Khimya Khetarpal*, **Doina Precup**

Mila and McGill University, Montreal, Canada

Correspondence to Arushi Jain: <arushi.jain@mail.mcgill.ca>

Objectives

Problem: In risk-sensitive applications, standard RL objective can't ensure *reliability* of algorithm, which is often required to deploy RL.
Solution: We represent reliability of algorithm by measuring *variability* in performance. We propose *On-Policy* and *Off-Policy* Variance Penalized Actor-Critic (`VPAC`) with,

- penalty using simpler **direct variance** operator,
- multi-timescale actor-critic updates,
- incremental *TD style* updates,
- convergence analysis for on-policy setting,
- experimental demonstrations in tabular and *MuJoCo* environments with comparison to baseline `VAAC` - indirect variance penalization.

Variance Estimators

G: discounted return

- **Indirect Variance** [1]

$$Var_\pi(G) = \mathbb{E}_\pi[G^2] - \mathbb{E}_\pi[G]^2, \qquad (1)$$

requires *second moment of return* operator to calculate variance.

- **Direct Variance** [2]

$$Var_\pi(G) = \mathbb{E}_\pi\left[\left(G - \mathbb{E}_\pi[G]\right)^2\right], \qquad (2)$$

skips calculation of second moment of return.
Direct is better than Indirect variance estimator when -

- value estimates are noisy,
- traces are used with value estimation,
- off-policy samples are used to estimate variance.

Notation

- $\sigma(s,a)$: variance in return
- d_0: initial state distribution
- ψ: mean-variance trade-off
- π_θ: policy parameterized by θ

Optimization problem

$$J_{d_0}(\theta) = \mathbb{E}_{s\sim d_0}\Big[\sum_a \pi_\theta(a|s)\Big(\overbrace{Q_{\pi_\theta}(s,a)}^{\text{value func}} - \underbrace{\psi}_{\text{tradeoff}} \overbrace{\sigma_{\pi_\theta}(s,a)}^{\text{variance func}}\Big)\Big], \qquad (3)$$

Direct Variance in Return

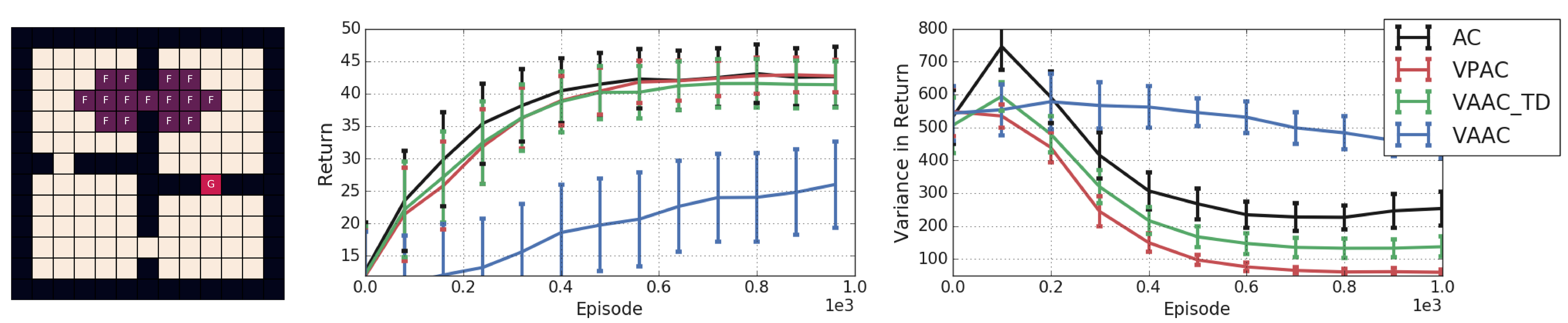
$$\sigma_{\pi_\theta}(s,a) = \mathbb{E}_{\pi_\theta}\big[\underbrace{\delta^2_{t,\pi_\theta}}_{\text{meta-reward}} + \underbrace{\bar{\gamma}}_{\bar{\gamma}=\gamma^2}\sigma_{\pi_\theta}(S_{t+1},A_{t+1})\big|S_t=s, A_t=a\big], \qquad (4)$$

where, $\delta_{t,\pi_\theta} = R_{t+1} + \gamma Q_{\pi_\theta}(S_{t+1},A_{t+1}) - Q_{\pi_\theta}(S_t,A_t)$ is the TD error.

Simple **On-Policy VPAC** update -

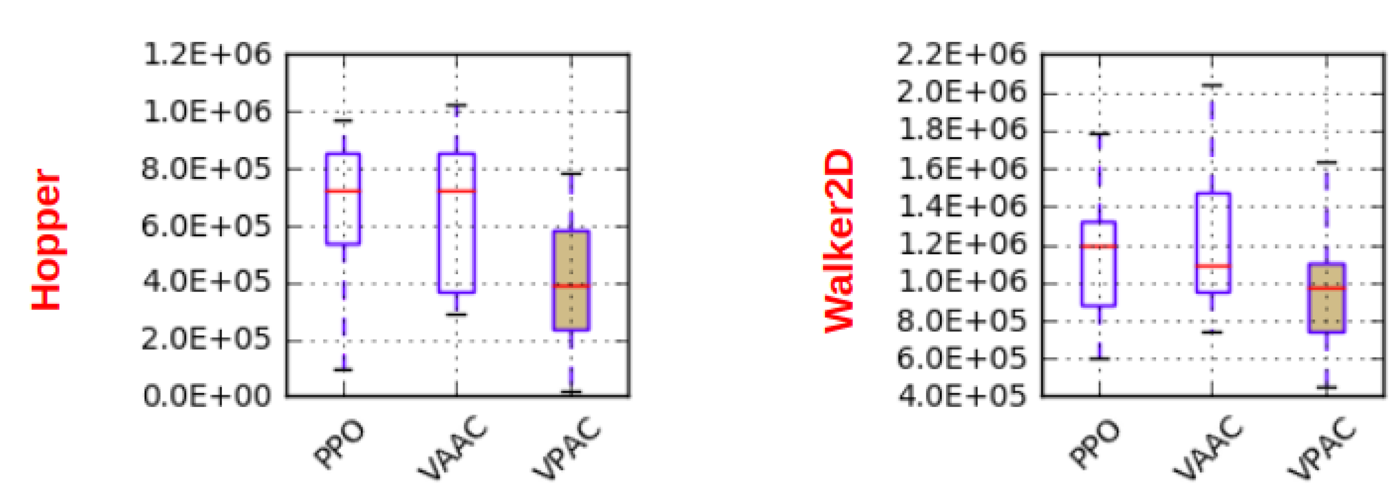
$$\theta_{t+1} = \theta_t + \alpha\nabla\log\pi_{\theta_t}(A_t|S_t)\Big(\gamma^t Q_{\pi_{\theta_t}}(S_t,A_t) - \psi\gamma^{2t}\sigma_{\pi_{\theta_t}}(S_t,A_t)\Big). \qquad (5)$$

Four-Rooms Experiment

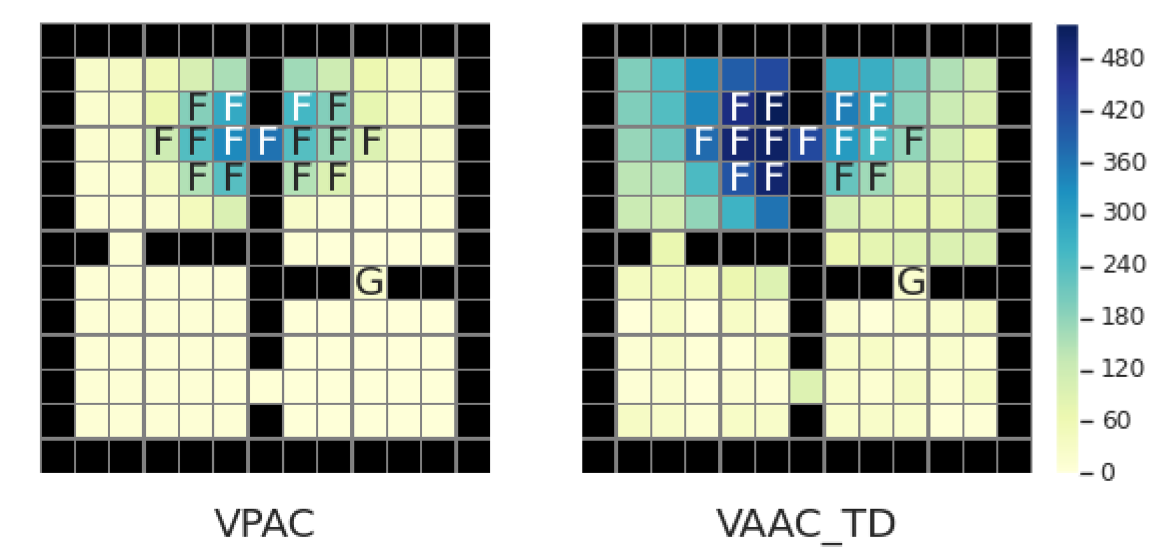


Mujoco Environments

	PPO		VAAC		VPAC	
Environment	Mean	Var(1e5)	Mean	Var(1e5)	Mean	Var(1e5)
HalfCheetah	1557	1.6	1525	0.8 (50%)	1373	**0.1** (93%)
Hopper	1944	6.6	1991	6.5 (1.5%)	1624	**4.0** (39.4%)
Walker2d	3058	12.1	3102	12.5 (-3.3%)	2625	**9.2** (23.9%)

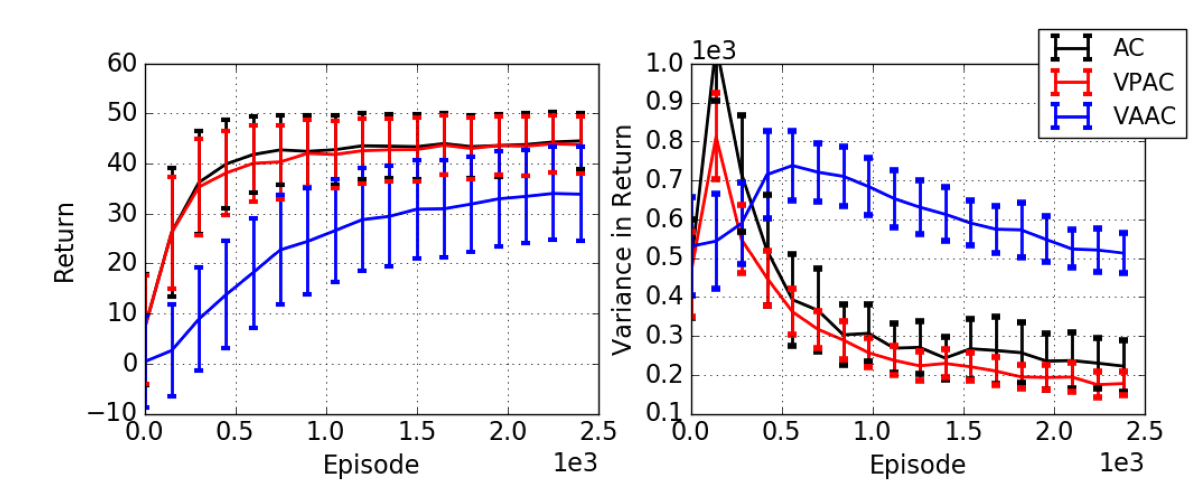


Direct vs Indirect Variance



Compares variance in return from initial state distribution of `VPAC` (direct) and `VAAC_TD` (indirect).

Off-Policy VPAC



Performance in discrete Puddle-World environment.

Conclusion

1. Proposed a **direct variance** risk-sensitive criteria for **control**.
2. Proposed **on-** and **off-policy** actor-critic **variance penalized** algorithm resulting into **lower variance**(reliable) trajectories compared to risk-neutral and indirect variance baseline.

References

[1] Matthew J Sobel.
The variance of discounted Markov decision processes.
Journal of Applied Probability, 19(4), 1982.

[2] Craig Sherstan, Dylan R Ashley, Brendan Bennett, Kenny Young, Adam White, Martha White, and Richard S Sutton.
Comparing direct and indirect temporal-difference methods for estimating the variance of the return.
In *Proceedings of UAI*, 2018.