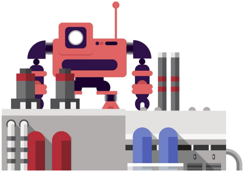
Variance Penalized On-Policy and Off-Policy Actor-Critic

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

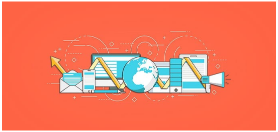
Motivation

Sequential decision-making under uncertainty

(a) Medical diagnosis



(b) industrial automation



(c) Portfolio management

Notation

Markov Decision Process

- MDP is tuple of $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, P, \gamma \rangle$
- Infinite horizon discounted setting
- $\mathcal{S}$: set of states
- $\mathcal{A}$: set of actions
- R_{t+1}: reward at t time step
- $P(\cdot|s, a)$: transition probability distribution
- γ: discount factor
- $G_t = \sum_{l=t}^{\infty} \gamma^{l-t} R_{l+1}$: discounted return

Notation

Let policy be parameterized by θ: π_θ.

$$V_{\pi_\theta}(s) = \mathbb{E}_{\pi_\theta}\left[G_t | S_t = s\right]$$
$$Q_{\pi_\theta}(s, a) = \mathbb{E}_{\pi_\theta}\left[G_t | S_t = s, A_t = a\right]$$

Risk Neutral Objective

$$J_{d_0}(\theta) = \sum_s d_0(s) V_{\pi_\theta}(s)$$

- $d_0(s)$: initial state distribution

Risk-Sensitive Criteria

Looking at a criteria where variability is penalized.

1. Stochasticity in Reward & Transition - Risk-Sensitive MDPs.
2. Imperfect knowledge about model - Robust MDPs.

$$G_t = \sum_{l=t}^{\infty} \gamma^{l-t} R_{l+1}$$

- Return: a random variable

Here, we use **variance** as way to measure variability.

Risk-Sensitive Criteria

Indirect Variance (Sobel, 1982)

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

Uses second moment of return.

Risk-Sensitive Criteria

Indirect Variance (Sobel, 1982)

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

Uses second moment of return.

Direct Variance (Sherstan et al., 2018)

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

Skips calculation of second moment of return.

Direct Variance

Sherstan et al. 2018, established benefits of direct variance in policy evaluation setting

- Noisy value estimates
- Eligibility traces with value estimation
- Variance estimated from off-policy samples
- Direct variance estimation simpler than Indirect variance

Our Contribution

For discounted reward setting,

1. modify policy gradient objective to include direct variance estimator to learn variance-penalized policy

Our Contribution

For discounted reward setting,

1. modify policy gradient objective to include direct variance estimator to learn variance-penalized policy
2. develop three-timescale VPAC **actor-critic** algorithm by deriving gradient of direct variance for
   - on-policy
   - off-policy

Our Contribution

For discounted reward setting,

1. modify policy gradient objective to include direct variance estimator to learn variance-penalized policy
2. develop three-timescale VPAC **actor-critic** algorithm by deriving gradient of direct variance for
   - on-policy
   - off-policy
3. establish convergence for on-policy setting

Our Contribution

For discounted reward setting,

1. modify policy gradient objective to include direct variance estimator to learn variance-penalized policy
2. develop three-timescale VPAC **actor-critic** algorithm by deriving gradient of direct variance for
   - on-policy
   - off-policy
3. establish convergence for on-policy setting
4. demonstrate the usefulness of on- and off- policy VPAC in tabular, linear and **Mujoco** environments

On-Policy VPAC

Mila

Variance in Return

For a given policy π,

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

On-Policy VPAC

Variance in Return

For a given policy π,

$$\sigma_\pi(s,a) = \mathbb{E}_\pi\Big[\underbrace{\delta_{t,\pi}^2}_{\text{meta-reward}} + \underbrace{\bar{\gamma}}_{\bar{\gamma}=\gamma^2} \sigma_\pi(S_{t+1}, A_{t+1}) \Big| S_t = s, A_t = a \Big]$$

$\delta_{t,\pi} = R_{t+1} + \gamma Q_\pi(S_{t+1}, A_{t+1}) - Q_\pi(S_t, A_t)$ [TD Error]

On-Policy VPAC

Variance in Return

For a given policy π,

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

$$\delta_{t,\pi} = R_{t+1} + \gamma Q_\pi(S_{t+1},A_{t+1}) - Q_\pi(S_t,A_t) \quad \text{[TD Error]}$$

Optimization Problem

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

Need to evaluate $\nabla_\theta V_\pi(s)$ and $\nabla_\theta \sigma_\pi(s)$ to tune θ.

On-Policy VPAC

AC Update

$$\theta_{t+1} = \theta_t + \alpha \nabla \log \pi_{\theta_t}(A_t|S_t)\Big(\gamma^t Q_{\pi_{\theta_t}}(S_t, A_t)\Big)$$

On-Policy VPAC

AC Update

$$\theta_{t+1} = \theta_t + \alpha \nabla \log \pi_{\theta_t}(A_t|S_t)\Big(\gamma^t Q_{\pi_{\theta_t}}(S_t, A_t)\Big)$$

VPAC update

$$\theta_{t+1} = \theta_t + \alpha \nabla \log \pi_{\theta_t}(A_t|S_t)\Big(\gamma^t Q_{\pi_{\theta_t}}(S_t, A_t) - \underbrace{\psi \gamma^{2t} \sigma_{\pi_{\theta_t}}(S_t, A_t)}_{\text{Variance Penalization}}\Big)$$

Multi Timescale Actor-Critic Update

$I_Q, I_\sigma = 1, 1.$
At every time step t,

1. **Critic Update**
   1. Q value Update: $w \leftarrow w + \alpha_Q \delta \nabla_w \hat{Q}(S, A, w)$
   2. σ value update: $z \leftarrow z + \alpha_\sigma \bar{\delta} \nabla_z \hat{\sigma}(S, A, z)$,
      where $\bar{\delta} \leftarrow \delta^2 + \gamma^2 \hat{\sigma}(S', A', z) - \hat{\sigma}(S, A, z)$
2. **Actor Update**
   1. $\theta \leftarrow \theta + \alpha_\theta \nabla_\theta \log(\pi_\theta(A|S)) \Big(I_Q \hat{Q}(S, A, w) - \psi I_\sigma \hat{\sigma}(S, A, z) \Big)$
   2. $I_Q * = \gamma$
   3. $I_\sigma * = \gamma^2$

Learning Rate Speed

$\alpha_Q > \alpha_\sigma > \alpha_\theta$

Off-Policy VPAC

Variance in Return

For a given target policy π and behavior policy b,

$$\sigma_\pi(s,a) = \mathbb{E}_b\big[\delta^2_{t,\pi} + \gamma^2\rho^2_{t+1}\sigma_\pi(S_{t+1},A_{t+1})\big|S_t = s, A_t = a\big]$$

$$\delta_{t,\pi} = R_{t+1} + \gamma\rho_{t+1}Q_\pi(S_{t+1},A_{t+1}) - Q_\pi(S_t,A_t) \quad \text{[TD Error]}$$

Off-Policy VPAC

Mila

Variance in Return

For a given target policy π and behavior policy b,

$$\sigma_\pi(s,a) = \mathbb{E}_b\left[\delta^2_{t,\pi} + \gamma^2 \rho^2_{t+1} \sigma_\pi(S_{t+1}, A_{t+1}) \middle| S_t = s, A_t = a\right]$$

$\delta_{t,\pi} = R_{t+1} + \gamma \rho_{t+1} Q_\pi(S_{t+1}, A_{t+1}) - Q_\pi(S_t, A_t)$ [TD Error]

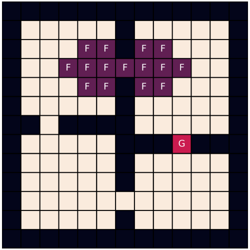
Optimization Problem

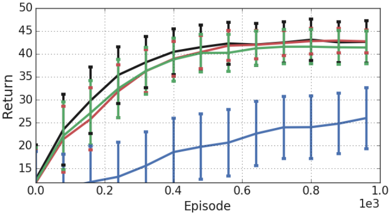
$$J_{d_0}(\theta) = \mathbb{E}_{s\sim d_0, a\sim b}\Big[\rho(s,a)\big(Q_\pi(s,a) - \psi\sigma_\pi(s,a)\big)\Big]$$

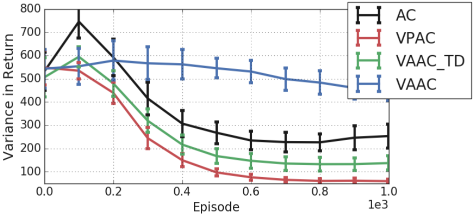
$\rho(s,a) = \frac{\pi(s,a)}{b(s,a)}$ importance sampling correction

Experiments

Experiments: Tabular





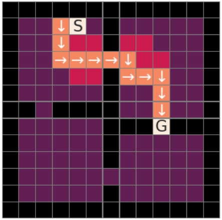


VAAC (Tamar et al. 2013)

Experiments: Tabular

Trajectory

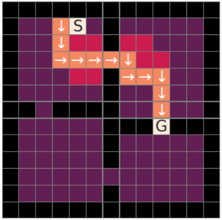
AC



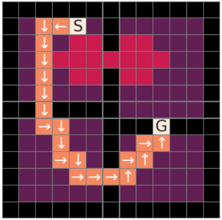
Experiments: Tabular

Trajectory

AC



VPAC

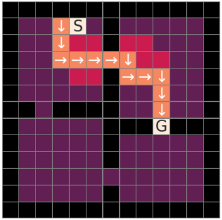


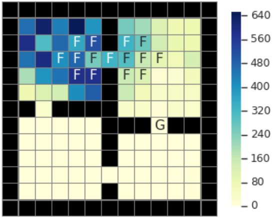
Experiments: Tabular

Trajectory

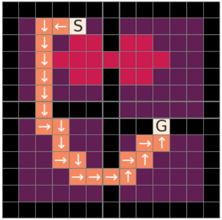
Variance in Return

AC





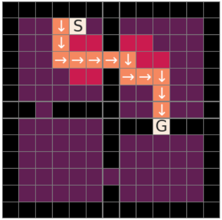
VPAC

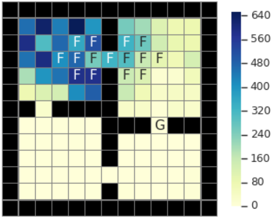


Experiments: Tabular

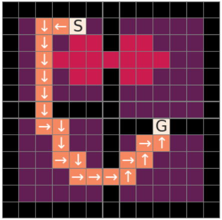
Trajectory | Variance in Return

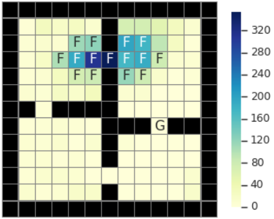
AC





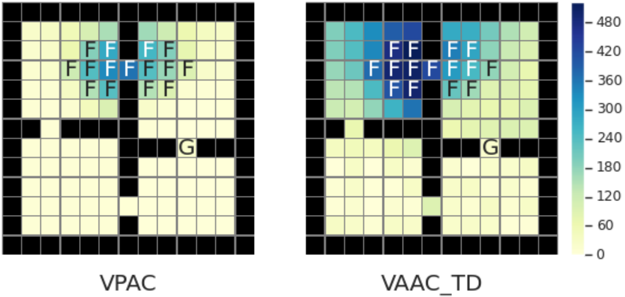
VPAC





Variance Comparison

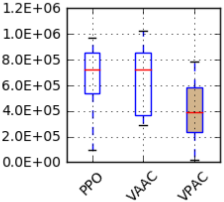
Direct Variance vs Indirect Variance



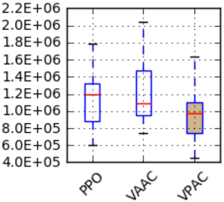
Mujoco Environments

	PPO		VAAC		VPAC	
Environment	Mean	Var	Mean	Var	Mean	Var
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%)

* Var in 1e5



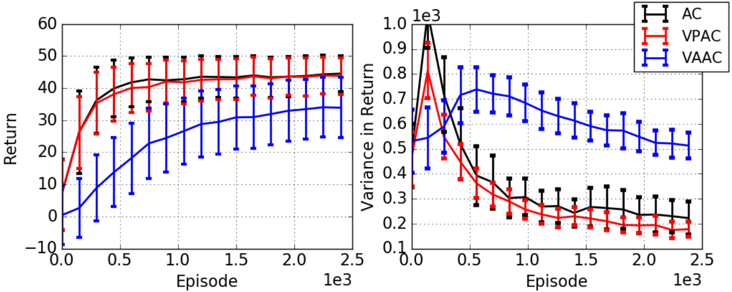
Hopper



Walker2d

Experiments: Off-policy VPAC

- Discrete Puddle World Environment



Conclusion

- Propose a **direct variance** related risk-sensitive criteria for control.
- Direct variance is simpler and better behaved than indirect variance.
- Propose multi-timescale actor-critic approach to learn variance penalized policy for **on-policy** and **off-policy** setting.
- Experiments supports **VPAC** results into lower variance trajectories compared to risk-neutral and indirect variance methods.

Future Work

- Provide convergence analysis for linear function approximation case as well.
- Observe the effects of scheduler on mean-variance tradeoff ψ to provide balance between exploration and variance reduction.