Safe Hierarchical Policy Optimization using Constrained Return Variance in Options

Arushi Jain
School of Computer Science
Mila - McGill University
Montreal, Canada
arushi.jain@mail.mcgill.ca

Doina Precup
School of Computer Science
McGill University, Google Deepmind
Montreal, Canada
dprecup@cs.mcgill.ca

Abstract

The standard setting in reinforcement learning (RL) to maximize the mean return does not assure a reliable and repeatable behavior of an agent in safety-critical applications like autonomous driving, robotics, and so forth. Often, penalization of the objective function with the variance in return is used to limit the unexpected behavior of the agent shown in the environment. While learning the end-to-end options have been accomplished, in this work, we introduce a novel Bellman style direct approach to estimate the variance in return in hierarchical policies using the option-critic architecture (Bacon et al., 2017). The penalization of the mean return with the variance enables learning safer trajectories, which avoids inconsistently behaving regions. Here, we present the derivation in the policy gradient style method with the new safe objective function which would provide the updates for the option parameters in an online fashion.

Keywords: option-critic,
safety,
constrained variance in return,
policy gradient

Acknowledgements

This research has received funding from Center of Research Institute in Montreal (CRIM).

1 Introduction

The objective function of maximizing the mean return does not offer any constraint on the distribution of the return, making it a vulnerable strategy for the risk-sensitive domains. The notion of avoiding risks arising from the stochastic nature of the environment (*inherent uncertainty*) using the constraint on the variance in return has been studied for a long time by the research community. Prashanth and Ghavamzadeh (2013); Sato et al. (2001); Tamar et al. (2012, 2016, 2013) constraint the indirect estimate of the variance using the second-order moment methods or directly estimated the cost-to-go returns with the updates provided after completing the entire trajectory. Sherstan et al. (2018) came up with a direct estimation of the variance in the λ-return using a Bellman operator in the policy evaluation methods. This work demonstrated the superiority of the direct estimator over the indirect approaches to estimate the variance.

Temporal abstraction provides a way to learn the policies in a hierarchical fashion which has been shown to improve exploration, robustness against model misspecification and increases the learning speed in transfer learning. Recently, option-critic architecture (Bacon et al., 2017) introduced an end-to-end style of learning the options. Jain et al. (2018) used the variance in the temporal difference (TD) error over the initial state-option pair distribution to estimate the controllable states in the option-critic.

In this work, we came up with a novel hierarchical safe policy learning approach in the option-critic architecture where the hierarchical policies are learned by penalizing the direct estimate of the variance in return extending from Sherstan et al. (2018) in a control setting. We seek to maximize the mean return and minimize the direct estimate of the variance in return given an initial state-option pair distribution in the policy gradient style.

2 Background

In a Markov Decision Process (MDP), an agent takes an action $a \in A$, transitions from state S_t to state S_{t+1}, and receives an immediate reward R_{t+1} from the environment. The expected reward is $r(S_t, A_t) = \sum_{r \in \mathbb{R}} r \sum_{s'} P(s', r | S_t, A_t)$ where $r : S \times A \to \mathbb{R}$. The environment dynamics is modeled by $P(S_{t+1}|S_t, A_t)$, where $P : S \times A \times S \to [0, 1]$. A stochastic policy $\pi(A_t|S_t)$ determines the probability of taking an action in a given state. The MDP is represented by a tuple $\langle S, A, P, r, \gamma \rangle$, where $\gamma \in [0, 1]$ is a factor discounting the future rewards.

2.1 Option-Critic

The option-critic architecture (Bacon et al., 2017), an option $w \in W$ is defined as a tuple of $\langle I_w, \pi_w, \beta_w \rangle$; where I_w contains the initial set of states where an option can start, π_w is the option policy defining a distribution over action space and β_w determines the termination probability of an option in a state. The policy over the options is denoted by $\mu(w|s)$ describing the distribution over options given a state. Let $\Theta = [\theta, \nu, \kappa]$ be the parameters of intra-option policy π_w, termination condition β_w and policy over options μ respectively. $J_{\pi,\mu}$ denotes the objective function of maximizing the mean return. The intra-option policy gradient (Bacon et al., 2017) update is:

$$\nabla_\theta J_{\pi,\mu}(\Theta) = \mathbb{E}_{\pi,\mu}[\nabla_\theta \log \pi_\theta(A_t|S_t, W_t) Q_{\pi,\mu}(S_t, W_t, A_t)],$$

and the termination gradient (Bacon et al., 2017) is given by:

$$\nabla_\nu J_{\pi,\mu}(\Theta) = \mathbb{E}_{\pi,\mu}[-\nabla_\nu \beta_\nu(S_{t+1}, W_t) A_{\Theta,Q}(S_{t+1}, W_t)]$$

where, $A_{\Theta,Q}(S_{t+1}, W_t) = Q(S_{t+1}, W_t) - V(S_{t+1})$ is the advantage function describing the importance of an option value over the mean value. In the following work we assume that all the options can be started from any state ($I_w \in S \,\forall\, w \in W$).

3 Safety in Option-Critic

Taking inspiration from the notion of safety in the actor-critic framework using the constraint variance in return (Jain et al., 2019), we similarly derive the safe framework in the option-critic. Our notion of *safety* emphasizes minimizing the *erratic* or the *harmful* behavior of an agent in the environment (Amodei et al., 2016). The higher is the variance in return from a state; the higher would be the uncertainty in the value estimate of that state. Uncertainty in the value estimate of a state reflects an inconsistent behavior of the agent in that particular state. Considering that the irregular or sudden behavior is classified as unsafe, potentially, the unsafe states would exhibit higher variance in return.

Let the return be denoted by

$$G_{t,\pi,\mu} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = R_{t+1} + \gamma G_{t+1,\pi,\mu}.$$

We consider $Z_t = (S_t, W_t)$ as an augmented state space - a space of state-option pair. Here, the transition matrix over the augmented state space is given by:

$$P(z'|z, a) = P(s'|s, a)[(1 - \beta_\nu(s', w)) \mathbb{1}_{w=w'} + \beta_\nu(s', w) \mu_\kappa(w'|s')] \tag{1}$$

The rewards are coming from a base MDP, where we write $r(z, a, z') = r(s, a)$. Since, $\sum_{z'} P(z'|z, a) = 1$, therefore, the reward model is defined as:

$$r(s, a) = \mathbb{E}_{\pi,\mu}[R_{t+1}|S_t = s, A_t = a] = \sum_{z'} P(z'|z, a)r(z, a, z')$$

Lemma 1. $\mathbb{E}_b[\gamma\lambda\delta_{t,\pi}\big(\rho_{t+1}G^{\lambda}_{t+1,\pi} - \mathbb{E}_b[\rho_{t+1}Q_\pi(S_{t+1}, A_{t+1})|S_t = s, A_t = a]\big)\big|S_t = s, A_t = a] = 0.$

Proof. The proof of the lemma is given in Jain et al. (2019). Here λ is trace decay parameter and $\delta_{t,\pi}$ is the 1-step TD error. □

Theorem 1. *The Bellman equation for the variance in the return from a given augmented state-action pair is:*

$$\sigma_{\pi,\mu}(z, a) = \mathbb{E}_{\pi,\mu}\big[\delta^2_{t,\pi} + \bar{\gamma}\sigma_{\pi,\mu}(Z_{t+1}, A_{t+1})\big|Z_t = z, A_t = a\big] \tag{2}$$

where $\bar{\gamma} = \gamma^2$ and δ_t is the 1-step TD error.

Proof. On expanding $G_{t,\pi,\mu} - Q_{\pi,\mu}(z, a)$,

$$\begin{aligned}
G_{t,\pi,\mu} - Q_{\pi,\mu}(z, a) =& R_{t+1} + \gamma G_{t+1,\pi,\mu} - Q_{\pi,\mu}(z, a) \\
=& R_{t+1} + \gamma \sum_{z',a'} P(z'|Z_t, A_t)\pi_\theta(a'|z')Q_{\pi,\mu}(z', a') - Q_{\pi,\mu}(z, a) \\
&+ \gamma\{G_{t+1,\pi,\mu} - \sum_{z',a'} P(z'|Z_t, A_t)\pi_\theta(a'|z')Q_{\pi,\mu}(z', a')\} \\
=& \delta_t + \gamma\Big(G_{t+1,\pi,\mu} - \mathbb{E}_{\pi,\mu}[Q_{\pi,\mu}(Z_{t+1}, A_{t+1})|Z_t = z, A_t = a]\Big)
\end{aligned} \tag{3}$$

Similar to Jain et al. (2019), let the variance for the augmented state-action pair be given as:

$$\begin{aligned}
\sigma_{\pi,\mu}(z, a) =& \mathbb{E}_{\pi,\mu}\left[(G_{t,\pi,\mu} - \mathbb{E}_{\pi,\mu}[G_{t,\pi,\mu}|Z_t = z, A_t = a])^2|Z_t = z, A_t = a\right] \\
=& \mathbb{E}_{\pi,\mu}\left[(G_{t,\pi,\mu} - Q_{\pi,\mu}(z, a))^2|Z_t = z, A_t = a\right] \\
=& \mathbb{E}_{\pi,\mu}\left[\Big(\delta_t + \gamma(G_{t+1,\pi,\mu} - \mathbb{E}_{\pi,\mu}[Q_{\pi,\mu}(Z_{t+1}, A_{t+1})|Z_t = z, A_t = a])\Big)^2|Z_t = z, A_t = a\right] \\
=& \mathbb{E}_{\pi,\mu}\left[\delta^2_t|Z_t = z, A_t = a\right] + \gamma^2\,\mathbb{E}_{\pi,\mu}\left[(G_{t+1,\pi,\mu} - \mathbb{E}_{\pi,\mu}[Q_{\pi,\mu}(Z_{t+1}, A_{t+1})|Z_t = z, A_t = a])^2|Z_t = z, A_t = a\right] \\
&+ 2\gamma\,\mathbb{E}_{\pi,\mu}\left[\delta_t(G_{t+1,\pi,\mu} - \mathbb{E}_{\pi,\mu}[Q_{\pi,\mu}(Z_{t+1}, A_{t+1})|Z_t = z, A_t = a])\right]
\end{aligned} \tag{4}$$

Using the Lemma 1, by substituting $\rho, \lambda = 1$ and changing the state S as an augmented state Z, the third term in the above (4) goes to 0. This leads the variance to $\sigma_{\pi,\mu}(z, a) = \mathbb{E}_{\pi,\mu}\big[\delta^2_{t,\pi} + \bar{\gamma}\sigma_{\pi,\mu}(Z_{t+1}, A_{t+1})\big|Z_t = z, A_t = a\big]$. □

The new safe objective function is defined as:

$$\boldsymbol{J(\Theta)} = \mathbb{E}_{z\sim d}\big[Q_\Theta(z) - \psi_z\,\sigma_\Theta(z)\big],$$

where d describes the initial state-option distribution and ψ_z is the regularizer for the variance penalty term which is a function of the augmented state space.

Theorem 2. *(Safe Intra-Option Policy Gradient Theorem) Given Markov options, $\pi_{w,\theta}$ policy differentiable in parameter θ, the gradient of the objective function J w.r.t. θ starting from state s and option w is:*

$$\nabla_\theta J(\Theta) = \mathbb{E}_{d,\Theta}[\sum_a \nabla_\theta\pi_\theta(a|Z_t)\big(Q_{\pi,\mu}(Z_t, a) - \psi_{Z_t}\,\sigma_{\pi,\mu}(Z_t, a)\big)]$$

Proof. The gradient of the $\sigma_\Theta(z)$ w.r.t. θ is calculated in a similar fashion as the gradient of $Q_\Theta(z)$ w.r.t. θ in the *Intra-option Policy Gradient Theorem* (Bacon et al., 2017). □

The gradient update for the intra-option policy moves in the direction to maximize the mean return value and minimize the variance in the return.

Theorem 3. *(Safe Termination Gradient Theorem) Given Markov options, $\beta_{w,\nu}$ termination function differentiable in parameter ν, the gradient of the objective function J w.r.t. ν starting from state s and option w is:*

$$\nabla_\nu J(\Theta) = \mathbb{E}_{d,\Theta}[-\nabla_\nu\beta_\nu(S_{t+1}, W_t)\big(A_{\pi,\mu,Q}(S_{t+1}, W_t) - \psi_z\,A_{\pi,\mu,\sigma}(S_{t+1}, W_t)\big)]$$

where $A_{\pi,\mu,\sigma}(S_{t+1}, W_t) = \sigma_\Theta(S_{t+1}, W_t) - \sigma_\Theta(S_{t+1})$ is the advantage function for the variance similar to the value function.

Proof. The gradient of the $\sigma_\Theta(z)$ w.r.t. ν is calculated in a similar fashion as the gradient of $Q_\Theta(z)$ w.r.t. ν in the *Termination Gradient Theorem* (Bacon et al., 2017). □

Similar to the Option-Critic, when the advantage of the value function is positive for an option, the gradient for the termination descends. On the other hand, the positive advantage function for the variance makes the gradient update for the termination ascent. It matches with the intuition, when the variance of an option is higher than the average variance, it would be desirable to terminate the option and choose a better option with a lower variance.

Theorem 4. *(Safe Policy over Options Gradient Theorem) Given Markov options, μ_κ policy over options differentiable in parameter κ, the gradient of the objective function J w.r.t. κ starting from state s and option w is:*

$$\nabla_\kappa J(\Theta) = \mathbb{E}_{d,\Theta}[\beta_\nu(S_{t+1}, W_t) \sum_{w'} \nabla_\kappa \mu_\kappa(w'|S_{t+1}) \big(Q_\Theta(z') - \psi_{z'}\, \sigma_\Theta(z')\big)]$$

Proof. Let 1-step augmented state transition using (1) be: $P^{(1)}_{\bar\gamma}(Z_{t+1}|Z_t) \overset{\text{def}}{=} \bar\gamma \sum_a \pi_\theta(a|Z_t) P(Z_{t+1}|Z_t, a)$. Similarly, the k-step transition would be defined as: $P^{(k)}_{\bar\gamma}(Z_{t+k}|Z_t) \overset{\text{def}}{=} P^{(1)}_{\bar\gamma}(Z_{t+k}|Z_{t+k-1}) \times P^{(k-1)}_{\bar\gamma}(Z_{t+k-1}|Z_t)$. The gradient of the variance w.r.t. κ parameter following (2),

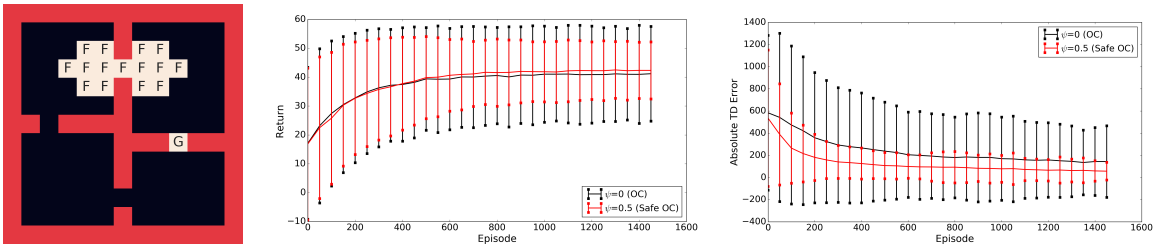
$$\begin{aligned}
\nabla_\kappa \sigma_\Theta(z) =& \nabla_\kappa \Big[\sum_a \pi_\theta(a|z) \bar\gamma \sum_{s'} P(s'|s,a) \big[(1-\beta_\nu(s',w))\sigma_\Theta(s',w) + \beta_\nu(s',w) \sum_{w'} \mu_\kappa(w'|s')\sigma_\Theta(s',w')\big]\Big] \\
=& \sum_a \pi_\theta(a|z) \bar\gamma \sum_{s',w'} P(s'|s,a) \big[(1-\beta_\nu(s',w)) \mathbb{1}_{w=w'} + \beta_\nu(s',w)\mu_\kappa(w'|s')\big] \nabla_\kappa \sigma_\Theta(s',w') \\
&+ \sum_a \pi_\theta(a|z) \bar\gamma \sum_{s',w'} P(s'|s,a)\beta_\nu(s',w) \sum_{w'} \nabla_\kappa \mu_\kappa(w'|s') \sigma_\Theta(s',w') \\
=& \sum_{k=0}^{\infty} \sum_{z'} P^{(k)}_{\bar\gamma}(z'|z) \sum_{a'} \pi_\theta(a'|z') \sum_{s''} \bar\gamma P(s''|s',a') \beta_\nu(s'',w') \sum_{w''} \nabla_\kappa \mu_\kappa(w''|s'') \sigma_\Theta(s'',w'')
\end{aligned}$$

Similarly, the gradient of the $Q_\Theta(z)$ value function can be derived similarly, leading to the proof. □

The above theorem states that the gradient of the policy over the options is updated in the direction of maximizing the expected Q-value and minimizing the variance function achieved from all other possible options after termination of the current option.

4 Experiment

Grid-World: We experiment in the classic grid-world four rooms (FR) environment (Bacon et al., 2017). To test safety, we created a variable reward frozen patch (F) in one of the hallway generated from $\mathcal{N}(0, 8)$ distribution. The rest of the states are given a 0 reward. Agent receives a reward of 50 on reaching the goal (G) (See Fig. 1a). γ is kept as 0.99. The step size of value function, variance function, intra-option policy, termination, policy over options are $1.0, 2e-3, 1e-3, 5e-3, 1e-4$ respectively for both option-critic ($\psi_z = 0$) and safe option-critic ($\psi_z = 0.5$) $\forall z \in Z$. Fig. 1b and Fig. 1c depict the performance in the FR environment.

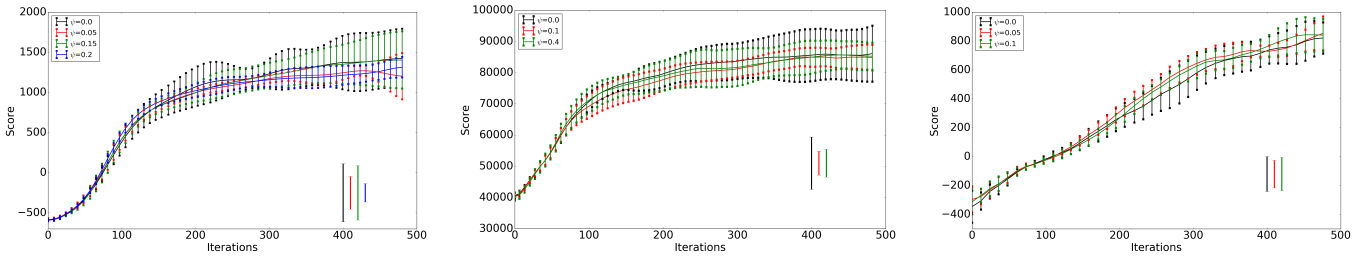


(a) Environment (b) Learning curve (c) Absolute TD error

Figure 1: **Performance in the FR**: Shows the performance averaged over 50 trials where the vertical bands depict the std. dev.. Shows b) the return, and c) sum of the absolute TD error. The safe policy (red) has a smaller standard deviation as compared to the baseline (black) signifying safety helps an agent to avoid variance inducing regions.

Continuous State-Action Space: Here we performed the experiments in Mujoco environments to test the real-world use case of introducing safe trajectories while learning in an environment. We implemented our safe algorithm over existing

proximal policy option-critic (PPOC) (Klissarov et al., 2017). We compare the performance of the agent using both the baseline PPOC and Safe-PPOC in Fig. 2. The videos[1] compare the performance of the agent using both the algorithms in the Mujoco environments.



(a) Half Cheetah (b) Humanoid Standup (c) Ant

Figure 2: **Performance in Mujoco**: Learning curve average over 10 runs where vertical bands depict the std. dev.. The vertical bars at right most corner display the std. dev. in performance over the last 50 iterations. The variance regularized PPOC ($\psi > 0$) helps in reducing the variation across multiple seed values leading to a more consistent performance.

5 Conclusion & Future Work

This work aims to introduce the constraint over the variance in return to the existing option-critic architecture in order to incorporate responsible behavior in the risk-sensitive domains. Firstly, we propose a direct estimator of the variance in the hierarchical policy framework. Then, we establish a method to learn a safe and reliable policy in option-critic, which uses the above direct estimator of the variance to avoid unpredictably behaving regions. The above framework is generic, which makes no assumption about the environment, making it a simple strategy to combine with the current policy gradient techniques. The *future work* is to experiment with more different environments like Atari to understand the scalability of the safe algorithm.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Bacon, P.-L., Harb, J., and Precup, D. (2017). The option-critic architecture. In *AAAI*, pages 1726–1734.

Jain, A., Jain, A., Khetarpal, K., , Aboutalebi, H., and Precup, D. (2019). On-policy and off-policy actor-critic with constrained return variance. In *Under Submission*.

Jain, A., Khetarpal, K., and Precup, D. (2018). Safe option-critic: Learning safety in the option-critic architecture. *arXiv preprint arXiv:1807.08060*.

Klissarov, M., Bacon, P.-L., Harb, J., and Precup, D. (2017). Learnings options end-to-end for continuous action tasks. *arXiv preprint arXiv:1712.00004*.

Prashanth, L. and Ghavamzadeh, M. (2013). Actor-critic algorithms for risk-sensitive MDPs. In *Advances in neural information processing systems*, pages 252–260.

Sato, M., Kimura, H., and Kobayashi, S. (2001). TD algorithm for the variance of return and mean-variance reinforcement learning. *Transactions of the Japanese Society for Artificial Intelligence*, 16(3):353–362.

Sherstan, C., Ashley, D. R., Bennett, B., Young, K., White, A., White, M., and Sutton, R. S. (2018). Comparing direct and indirect temporal-difference methods for estimating the variance of the return. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 63–72.

Tamar, A., Di Castro, D., and Mannor, S. (2012). Policy gradients with variance related risk criteria. In *Proceedings of the twenty-ninth international conference on machine learning*, pages 387–396.

Tamar, A., Di Castro, D., and Mannor, S. (2016). Learning the variance of the reward-to-go. *Journal of Machine Learning Research*, 17(13):1–36.

Tamar, A., Xu, H., and Mannor, S. (2013). Scaling up robust MDPs by reinforcement learning. *arXiv preprint arXiv:1306.6189*.

[1]The performance videos of the agent in PPOC and Safe PPOC in Mujoco domains is here.