SPSA based Actor-Critic Algorithm for Risk Sensitive Control

Prashanth L.A. Mohammad Ghavamzadeh
INRIA Lille - Nord Europe, Team SequeL
{prashanth.la,mohammad.ghavamzadeh}@inria.fr

1 Introduction

In a discounted reward Markov decision process (MDP) setting, we consider the problem of minimizing the variability in rewards in addition to maximizing the expected return. We first define a measure of variability for a policy similar to [2] and then develop an actor-critic algorithm that incorporates a simultaneous perturbation based gradient estimate of the risk-sensitive performance measure (see (1) below). The proposed algorithm can be shown to converge to a locally risk sensitive optimal policy.

1.1 The Problem

Consider an MDP with state space $\mathcal{X} = \{1, \ldots, n\}$ and action space $\mathcal{A} = \{1, \ldots, m\}$. Let $r(x, a)$ denote the expected reward when $a \in \mathcal{A}$ is chosen in state $x \in \mathcal{X}$. A *stationary policy* $\mu(\cdot|x)$ is a probability distribution over actions, conditioned on the current state. Let $D^\mu(x)$ denote the discounted return for a given policy μ and is defined as $D^\mu(x) = \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu$. The value and square value functions $V^\mu(x)$ and $U^\mu(x)$ are defined as the expected value of $D^\mu(x)$ and $D^\mu(x)^2$, respectively.

The aim here is to find a stationary policy (for a given starting state x^0) that maximizes the expected return subject to a constraint on the variability of the return, i.e.,

$$\max_{\mu \in \mathcal{U}} V^\mu(x^0) \quad \text{subject to} \quad \Lambda^\mu(x^0) \leq \alpha \quad \Longleftrightarrow \quad \max_{\lambda} \min_{\theta} L(\theta, \lambda) \stackrel{\triangle}{=} -V^\mu(x^0) + \lambda\big(\Lambda^\mu(x^0) - \alpha\big). \tag{1}$$

It is challenging to devise an efficient method to estimate the gradient of the Lagrangian $L(\theta, \lambda)$ because **1)** two different sampling distributions are used for ∇V^μ and ∇U^μ, and **2)** $\nabla V^\mu(x')$ is required for all $x' \in \mathcal{X}$ to compute ∇U^μ. To alleviate these issues, we employ the technique of simultaneous perturbation to estimate the gradient of $L(\theta, \lambda)$.

2 The Algorithm

The idea is to estimate the gradients $\nabla V^{\mu_t}(x^0)$ and $\nabla U^{\mu_t}(x^0)$ using two simulated trajectories of the system corresponding to policies μ_t with parameter θ_t and μ_t^+ with parameter $\theta_t^+ = \theta_t + \beta\Delta_t$, respectively. The *simultaneous perturbation stochastic approximation* (SPSA) estimate of the gradient of $V^{\theta_t}(x^0)$ is given by:

$$\nabla V^{\mu_t, i}(x^0) \quad \approx \quad \frac{\widehat{V}(\theta + \beta\Delta) - \widehat{V}(\theta)}{\beta\Delta^i}, \quad i = 1, \ldots, \kappa_1,$$

where $\widehat{V}(\theta)$ is an estimate of the value function corresponding to a policy with parameter θ. Note that we use a linear function to approximate the value and square value functions,

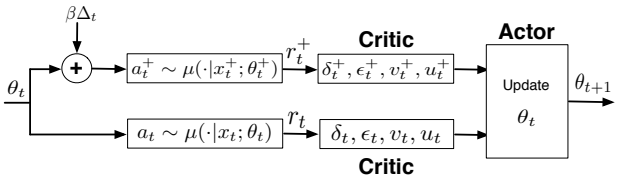

Figure 1: Overall flow of simultaneous perturbation algorithms.

i.e., $\widehat{V}(x) \approx v^\top f(x)$ and $\widehat{U}(x) \approx u^\top g(x)$, where the features $f(\cdot)$ and $g(\cdot)$ are from low-dimensional spaces in $\mathbb{R}^{\kappa_1}$ and $\mathbb{R}^{\kappa_2}$, respectively. Further, $\beta > 0$ is a small positive constant and $\Delta_t = (\Delta_t^1, \ldots, \Delta_t^{\kappa_1})^\top$ is a perturbation random variable. A popular choice for Δ_t^i is the symmetric, ± 1-valued Bernoulli random variable. The estimate of $\nabla_\theta U^{\mu_t}(x^0)$ is along similar lines. Note that $\nabla \Lambda^\mu(x^0) = \nabla U^\mu(x^0) - 2V^\mu(x^0)\nabla V^\mu(x^0)$.

Critic Update: At each time step t, the critic updates the parameters of the value (v_t, v_t^+) and square value (u_t, u_t^+) functions for policies μ_t and μ_t^+ using TD as follows:

$$\begin{aligned} v_{t+1} &= v_t + c(t)\delta_t f(x_t), u_{t+1} = u_t + c(t)\epsilon_t g(x_t), \\ \text{where } \delta_t &= r(x_t, a_t) + \gamma v_t^\top f(x_{t+1}) - v_t^\top f(x_t), \\ \epsilon_t &= r(x_t, a_t)^2 + 2\gamma r(x_t, a_t) v_t^\top f(x_{t+1}) + \gamma^2 u_t^\top g(x_{t+1}) - u_t^\top g(x_t), \end{aligned}$$

The TD updates for v_t^+, u_t^+ using the perturbed simulation x_t^+ are similar to the above.

Actor Update: The actor updates according to:

$$\theta_{t+1}^{(i)} = \Gamma_i\Big(\theta_t^{(i)} - b(t)\Big(- \big(1 + 2\lambda v_t^\top f(x^0)\big)\frac{(v_t^+ - v_t)^\top f(x^0)}{\beta\Delta_t^{(i)}} + \lambda\frac{(u_t^+ - u_t)^\top g(x^0)}{\beta\Delta_t^{(i)}}\Big)\Big), \quad (2)$$

$$\lambda_{t+1} = \Gamma_\lambda\Big[\lambda_t + a(t)\Big(u_t^\top g(x^0) - \big(v_t^\top f(x^0)\big)^2 - \alpha\Big)\Big], \quad (3)$$

In the above, Γ and Γ_λ are projection operators that are used to keep the iterates bounded (a requirement to guarantee convergence) and the step-sizes $\{c(t), b(t), a(t)\}$ are chosen such that the critic updates are on the fastest time-scale, the policy parameters on the intermediate time-scale, and the Lagrange multiplier update on the slowest. The above algorithm can be shown to converge to a (local) saddle point of $\widehat{L}(\theta, \lambda) = -\widehat{V}(\theta) + \lambda\big(\widehat{U}(\theta) - \widehat{V}^2(\theta) - \alpha\big)$, i.e., to a pair (θ^*, λ^*) that are a local minimum w.r.t. θ and a local maximum w.r.t. λ of $\widehat{L}$.

We also developed another actor-critic algorithm that incorporates the smoothed functional (SF) scheme for estimating the gradients, instead of SPSA. SF schemes are simultaneous perturbation methods as well and use Gaussian random variable for the perturbation. Further, both our algorithms can be easily extended to other risk-sensitive measures such as Sharpe ratio. The interested reader is referred to [1] for further information.

References

[1] Prashanth L.A. and Mohammad Ghavamzadeh. Actor-Critic Algorithms for Risk-Sensitive MDPs. Technical report inria-00794721, INRIA, 2013.

[2] M. Sobel. The variance of discounted Markov decision processes. *Applied Probability*, pages 794–802, 1982.