Journal of Machine Learning Research () Submitted ; Published

Policy Gradient for Coherent Risk Measures

Aviv Tamar AVIVT@TX.TECHNION.AC.IL
Electrical Engineering Department
The Technion - Israel Institute of Technology

Yinlam Chow YCHOW@STANFORD.EDU
Institute for Computational & Mathematical Engineering (ICME)
Stanford University

Mohammad Ghavamzadeh GHAVAMZA@ADOBE.COM
Adobe Research & INRIA

Shie Mannor SHIE@EE.TECHNION.AC.IL
Electrical Engineering Department
The Technion - Israel Institute of Technology

Editor: TBD

1. Introduction

Risk-sensitive optimization considers problems in which the objective involves a *risk measure* of the random cost, in contrast to the typical *expected* cost objective. Such problems are important when the decision-maker wishes to manage the *variability* of the cost, in addition to its expected outcome, and are standard in various applications in finance and operations research. In reinforcement learning, risk-sensitive objectives have been gaining popularity as a means to regularize the variability of the total (discounted) reward in an MDP.

Many risk objectives have been investigated in the literature, and applied to reinforcement learning (RL), such as mean-variance (Tamar et al., 2012; Prashanth and Ghavamzadeh, 2013), and Conditional Value at Risk (CVaR) (Chow and Ghavamzadeh, 2014; Tamar et al., 2015b). The preference of one risk measure over another depends on factors such as sensitivity to rare events, ease of estimation from data, and computational tractability of the optimization problem, and in general, there is no single choice that dominates over the rest. However, the highly influential paper of Artzner et al. (1999) identified a set of natural properties that are desirable for a risk measure to satisfy. Risk measures that satisfy these properties are termed *coherent* and have obtained widespread acceptance in financial applications, among others. In this work, we present algorithms for RL with a coherent risk. Our approach is general and applies to *the whole class* of coherent risk measures, thereby generalizing and unifying previous approaches that focused on individual risk measures.

2. Formulation and Results

Consider a random variable Z (corresponding to a random cost) on a sample space Ω, with a distribution P_θ parameterized by θ. A fundamental property of coherent risk is that every coherent risk measure ρ can be written as a ξ-weighted expectation, where the weights ξ are chosen adversarially from a suitable convex set $\mathcal{U}(P_\theta)$, referred to as the *risk envelope*:

$$\rho(Z;\theta) = \max_{\xi\,:\,\xi P_\theta \in \mathcal{U}(P_\theta)} \mathbb{E}_\xi[Z]. \tag{1}$$

In this work we are interested in solving general problems of the form

$$\min_\theta \ \rho(Z;\theta), \tag{2}$$

where ρ is a coherent risk measure. For example, in an RL setting, Z may correspond to the cumulative discounted cost $Z = C(x_0, a_0) + \gamma C(x_1, a_1) + \cdots + \gamma^T C(x_T, a_T)$ of a trajectory from an MDP with a policy parameterized by θ.

Another interesting risk measure for sequential decision problems is *dynamic* coherent risk, which for MDPs is given by

$$\rho_{\text{dynamic}} = C(x_0, a_0) + \gamma\rho\left(C(x_1, a_1) + \gamma\rho\left(C(x_2, a_2) + \ldots + \gamma\rho\left(C(x_T, a_T)\right)\ldots\right)\right). \tag{3}$$

From the arguments in Roorda et al. (2005), the dynamic risk is often more conservative than the static risk in Eq. (2), but from Ruszczyński (2010), this risk has the advantage of being time-consistent, meaning that it satisfies a "dynamic programming" style property: if a strategy is risk-optimal for an n-stage problem, then the component of the policy from the t-th time until the end (where $t < n$) is also risk-optimal.

In this work, we develop sampling-based algorithms for estimating the gradient $\nabla_\theta \rho(C;\theta)$, when ρ is either a static (2) or a dynamic (3) coherent risk measure. The optimization of θ may then be carried out using standard stochastic gradient descent techniques.

The main result driving our approach is a "policy-gradient" style equation for coherent risk. Assume that the risk-envelope is given in a canonical convex programming formulation which consists of affine equality constraints $g_e(\xi, P_\theta)$ (including the equality constraint on ξP_θ to guarantee that it is a probability distribution) and convex inequality constraints $f_i(\xi, P_\theta)$:

$$\mathcal{U}(P_\theta) = \left\{\xi P_\theta : \;\; g_e(\xi, P_\theta) = 0, f_i(\xi, P_\theta) \leq 0, \sum_{\omega\in\Omega} \xi(\omega) P_\theta(\omega) = 1, \; \xi(\omega) \geq 0\right\}. \tag{4}$$

We have the following result of a general policy gradient formula of $\rho(Z)$.

Theorem 1 *For any saddle point $(\xi_\theta^*, \lambda_\theta^e, \lambda_\theta^i, \lambda_\theta^p)$ of the Lagrangian of Eq.* (1) *with $\mathcal{U}(P_\theta)$ given by* (4)*, and where λ_θ^e, λ_θ^i, and λ_θ^p are the Lagrange multipliers for the equality, inequality, and normalization constraints in* (4)*, respectively, we have that*

$$\nabla_\theta \rho(Z) = \mathbb{E}_{\xi_\theta^*}\left[\nabla_\theta \log P(\omega)(Z - \lambda_\theta^p)\right] - \lambda_\theta^e \nabla_\theta g_e(\xi_\theta^*; P_\theta) - \lambda_\theta^i \nabla_\theta f_i(\xi_\theta^*; P_\theta).$$

We use Theorem 1 to devise an estimator for $\nabla_\theta \rho(Z)$ that involves sampling and convex programming, and show that this estimator is consistent. For the dynamic-risk (3), we exploit its dynamic programming property and derive a new formula for the gradient $\nabla_\theta \rho_{\text{dynamic}}$ that involves a *value function* of the risk objective ρ_{dynamic}. This formula uses the result of Theorem 1 and extends the well-known "policy gradient theorem" developed for the expected return (Sutton et al., 2000) to dynamic coherent risk measures. Using this formula, we suggest the following actor-critic style algorithm for estimating $\nabla_\theta \rho_{\text{dynamic}}$:
Critic: For a given policy θ, calculate the *risk-sensitive value function* of ρ_{dynamic}, and
Actor: Using the critic's value function, estimate $\nabla_\theta \rho_{\text{dynamic}}$ by sampling and convex programming.

Further details of both policy gradient (for static risk) and actor critic (for dynamic risk) methods can be found in our extended report (Tamar et al., 2015a).

3. Discussion

While previous works have focused on individual risk measures, in this work we presented a unified framework for the comprehensive class of coherent risk measures (both static and dynamic), thereby granting the decision-maker substantial flexibility in designing her risk preferences.

We believe that these results should steer future work in risk-sensitive RL away from designing more optimization algorithms for additional specific risk measures. Rather, it opens up the possibility of designing *application-dependent* risk measures and exploring the relations between static and dynamic risk.

Acknowledgments

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Program (FP/2007-2013) / ERC Grant Agreement n. 306638.

References

P. Artzner, F. Delbaen, J. Eber, and D. Heath. Coherent measures of risk. *Math. finance*, 1999.

Y. Chow and M. Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In *NIPS*, 2014.

L. Prashanth and M. Ghavamzadeh. Actor-critic algorithms for risk-sensitive MDPs. In *NIPS*, 2013.

B. Roorda, J. M. Schumacher, and J. Engwerda. Coherent acceptability measures in multiperiod models. *Math. Finance*, 15(4):589–612, 2005.

A. Ruszczyński. Risk-averse dynamic programming for Markov decision processes. *Math. Programming*, 125(2):235–261, 2010.

R. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, 2000.

A. Tamar, D. Di Castro, and S. Mannor. Policy gradients with variance related risk criteria. In *ICML*, 2012.

A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor. Policy gradient for coherent risk measures. *arXiv preprint arXiv:1502.03919*, 2015a.

A. Tamar, Y. Glassner, and S. Mannor. Optimizing the CVaR via sampling. In *AAAI*, 2015b.