Entropic Risk Measure in Policy Search

David Nass, Boris Belousov, and Jan Peters
Intelligent Autonomous Systems Lab, Technische Universität Darmstadt, Germany

I. Introduction

Applying reinforcement learning (RL) to robotics is notoriously hard due to the curse of dimensionality [1]. Robots operate in continuous state-action spaces and visiting every state quickly becomes infeasible. Therefore, function approximation has become essential to limit the number of parameters that need to be learned. Policy search methods, that employ pre-structured parameterized policies to deal with continuous action spaces, have been successfully applied in robotics [2]. These methods include policy gradient [3], [4], natural policy gradient [5], expectation maximization (EM) policy search [6], [7], and information theoretic approaches [8].

A common feature of the aforementioned policy search methods is that they all aim to maximize the expected reward. Therefore, they do not take into account the variability and uncertainty of the performance measure. However, robotic systems need to act in stochastic, non-stationary, partially observable environments. To account for these challenges, the objective function should include some variance related criteria in addition to the standard expected reward maximization objective.

II. Risk Sensitive Policy Search

In the recent years, there have been some advances in risk-sensitive policy search using policy gradients. In [14], a policy gradient algorithm was developed that accounted for the variance in the objective either through a penalty or as a constraint. The Conditional value at risk (CVaR) criterion was combined with policy gradients in [15], [16]. In this paper, we study properties of policy gradient methods with the entropic risk measure [9] in the objective, defined as follows

$$J_{\text{risk}}(R) = -\frac{1}{\gamma}\log \mathbb{E}[\exp(-\gamma R)], \tag{1}$$

where the risk-sensitivity depends on γ. Positive $\gamma > 0$ result in pessimistic, *risk-averse* behavior, while $\gamma < 0$ favors high-variance rewards and is called *risk-seeking*.

To introduce risk-sensitivity into policy search, we propose to optimize the entropic risk-measure (1) instead of optimizing for average performance. Rewriting it for the parameters ω of an upper-level policy $\pi_\omega(\theta) = \pi(\theta|\omega)$ yields

$$J_\gamma(\omega) = -\frac{1}{\gamma}\log \mathbb{E}_{\theta\sim\pi_\omega}[\exp(-\gamma R(\theta))]. \tag{2}$$

As the name suggests [2], policy gradient methods aim to maximize the objective $J(\omega)$ by gradient ascent on the policy parameters.

The likelihood ratio trick is commonly invoked to derive an estimate of the gradient. For the risk-sensitive objective (2), the likelihood ratio gradient yields

$$\nabla J_\gamma = \mathbb{E}_{\theta\sim\pi_\omega}\left[\nabla \log \pi_\omega(\theta)\left\{-\frac{1}{\gamma}\mathrm{e}^{-\gamma(R(\theta)-\psi_\gamma(\pi_\omega))}\right\}\right], \tag{3}$$

where $\psi_\gamma(\pi_\omega) = -\gamma^{-1}\log \mathbb{E}_{\pi_\omega}[\exp(-\gamma R)]$ is the log-partition function [21].

The first point to make about (3) is the relation between the risk-sensitive policy gradient and the standard, risk-neutral one. Observe from (2) that the risk-sensitive objective $J_\gamma(\omega)$ becomes risk-neutral for $\gamma \to 0$. Surprisingly, however, *the gradient of the risk-sensitive objective does not correspond to the vanilla policy gradient (PG)* $\nabla J = \mathbb{E}[\nabla \log \pi \cdot R]$ but instead to the PG with an average reward baseline

$$\nabla J_\gamma \xrightarrow[\gamma\to 0]{} \mathbb{E}[\nabla \log \pi \cdot (R - \mathbb{E}[R])]. \tag{4}$$

The log-partition function $\psi_\gamma(\pi_\omega)$ plays the role of the risk-sensitive baseline, since $\psi_\gamma \to \mathbb{E}[R]$ for $\gamma \to 0$. Therefore, risk-sensitive PG (3) automatically has lower variance compared to vanilla PG due to the presence of the baseline.

Furthermore, we can view (3) as a risk-neutral PG for an exponentially transformed reward function given by the expression in curly braces in (3). Therefore, along with the multiplicative baseline $\psi_\gamma(\omega)$, the usual additive baseline can also be subtracted to further reduce variance. Moreover, standard algorithms, such as natural policy gradient (NPG) [5] and proximal policy optimization (PPO) [22], can be directly applied to optimize the risk-sensitive objective (2) thanks to the form (3) of the risk-sensitive policy gradient.

It turns out, another important property of the gradient estimator (3) can be revealed by recognizing it as the gradient of the maximum likelihood policy update in Relative entropy policy search (REPS) [8]. REPS belongs to the category of information-theoretic policy search approaches [2]. This class of methods follows the idea of limiting the loss of information in-between policy updates.

There exists a closed form solution to the optimization problem of REPS that can be estimated by samples [8]. Fitting a parametric policy $\pi_\omega(\theta)$ to the said solution by moment projection [2] yields

$$\underset{\omega}{\text{maximize}}\ \mathbb{E}_{\theta\sim q}\left[\log \pi_\omega(\theta)\exp\left(\frac{R(\theta)-\psi_{-1/\eta}(q)}{\eta}\right)\right], \tag{5}$$

where η is a Lagrangian multiplier which corresponds to the bound of information loss between policy updates and θ is sampled from the distribution q.

The correspondence between the gradient of (5) and (3) is established by identifying $\gamma = -1/\eta$. Thus, the policy update of REPS (5) can be identified with the risk-sensitive update (3) under the assumption that the information loss bound is small, such that $q \approx \pi_{\omega}$ and one step in the direction of the gradient solves (5). Importantly, though, the temperature parameter $\eta = -1/\gamma$ gets optimized in REPS and thus changes with iterations, whereas when applying (3), it has to be scheduled manually.

Another interesting distinction between risk-sensitive optimization and REPS stems from the fact that the temperature parameter η must be positive in REPS. This means $\gamma < 0$, or risk-seeking optimization. Thus, REPS is risk-seeking by construction, unlike the risk-sensitive PG (3) which can also be risk-averse.

III. Experiments

To analyze the properties of the risk-sensitive policy gradient algorithm, we first consider a prototypical risk-sensitive portfolio optimization problem to establish the validity of our approach, then we proceed to apply the risk-sensitive policy gradient method to a toy robot badminton setup, and finally, we report the results obtained by applying the algorithm to a real-robot task of learning to return a shuttlecock in the game of badminton with the Barrett WAM robot.

A basic problem of portfolio optimization [27] consists of an individual who wants to invest a unit of capital in N assets with the goal of making profit. The distribution of capital over assets $\mathbf{x}$ is called portfolio. Returns of various assets are random variables and assumed to be Gaussian distributed, $\mathbf{r} \sim \mathcal{N}(\mu_{\mathbf{r}}, \Sigma_{\mathbf{r}})$. Then, return of a portfolio $\mathbf{x}$ is a random variable $R \sim \mathcal{N}(\mu_{\mathbf{r}}^{\mathrm{T}}\mathbf{x}, \mathbf{x}^{\mathrm{T}}\Sigma_{\mathbf{r}}\mathbf{x})$. Returns with a high expected value are accompanied with higher risks, whereas lower risk returns yield lower but more consistent reward. When comparing two policies π_1 and π_2 corresponding to risk factors $\gamma_1 > \gamma_2$, policy π_1 will prefer lower risk assets and yield lower return on average than π_2. Results obtained by our risk sensitive policy gradient algorithm can confirm the theory and imply that our algorithm works as intended.

Next, we considered a simplified scenario of a robot learning to return a shuttlecock in the game of badminton. We assume a two dimensional world and a ball following a parabolic flight trajectory. The goal is to determine the hitting velocity of the racket which results in the ball arriving at a desired target location. The hypothesis is that for different values of the risk-aversion factor γ, the agent will learn different strategies: either aggressive hits but with high variability, or safe returns however with smaller expected reward. The problem is specified as follows

$$\underset{\omega}{\text{minimize}} \quad \frac{1}{\gamma} \log \mathbb{E}\left[\exp(\gamma |x_{\text{des}} - x_1(v_{x,0} \;\; v_{y,0})|)\right] \qquad (6)$$

The final shuttlecock position x_1 is constrained by the equations of motion. We treat the initial ball velocity as the control variable and add a bit of noise, such that $(v_{x,0} \;\; v_{y,0}) = \mathbf{v}_0 \sim \mathcal{N}(\mathbf{u}, \Sigma_{\mathbf{v}_0})$. As usually, we employ a Gaussian policy $\mathbf{u} \sim \pi_{\omega}(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mu_{\mathbf{u}}, \Sigma_{\mathbf{u}})$.

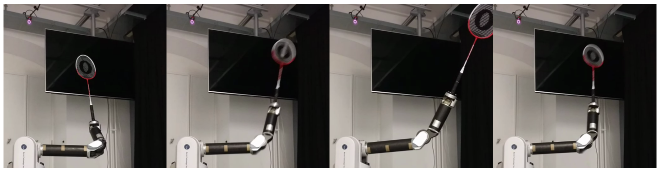
We evaluate the problem using the risk-sensitive policy gradient for various values of γ. The central observation was that both risk-seeking and risk-averse policies corresponding to extreme values of γ fail at returning the ball to the desired target. This effect is due to the dual nature of the objective function which trades mean performance against variability. Extreme risk-averse policies tend to undershoot the target, while extreme risk-seeking ones tend to overshoot it. The same conclusion can be made based on obtained initial velocities. Risk-averse, pessimistic policies favor smaller initial velocities. In contrast, risk-seeking, optimistic policies chose larger initial velocities. Further, we examined that large negative values of γ negatively affect optimization, due to objective (6) becoming very sharp, close to a delta function.

Finally we apply our algorithm to a real-robot badminton task. Unfortunately, with our current setup, we were not able to achieve the goal of training robot-badminton skills of varying degree of riskiness, due to hardware constraints. Returning a shuttlecock in badminton to a desired location requires a high degree of precision. In our experiments, we had to relax this requirement and only optimize for returning the shuttlecock at all. To test the limits of achievable performance in the badminton task, we carried out an extended learning trial in which 800 iterations of policy improvement were performed with 100 roll-outs per iteration. The best risk-neutral controller could return 95% of the served balls. An example successful hitting movement is shown in Fig. 1.

IV. Conclusion

The entropic risk measure was considered as the optimization objective for policy gradient methods. By analyzing the exact form of its gradient, we found that it is related to the standard policy gradient but inherently incorporates a baseline. Furthermore, risk-sensitive policy update was shown to correspond to a certain limiting case of the policy update in REPS. Exploring this connection to information-theoretic methods appears to be a fruitful direction for future work. Entanglement between exploration variance and inherent system variability was found to be a strong limiting factor. Approaches for separating these two sources of uncertainty need to be searched for.

To reveal strengths and weaknesses of risk-sensitive optimization in a real robotic context, we applied our policy gradient method to the problem of learning risk-sensitive movement primitives in a badminton task. In a simplified model, we observed that policies optimized for different values of risk aversion demonstrate qualitatively different behaviors. Namely, risk-averse policies hit the shuttlecock with smaller velocity and tended to undershoot, whereas risk-seeking policies favored higher velocities and typically overshot the target. Finally, we carried out experiments on the real robot, which showed that moderate values of risk aversion can help finding better solutions for the original, risk-neutral problem. However, our attempt at learning risk-sensitive movement primitives on the real robot had limited success due to limitations of the hardware platform and the entanglement of sources of variability.



(a) rest (b) hit (c) swing (d) retract

Fig. 1: Phases of the hitting movement of the Barrett WAM.

References

[1] R. E. Bellman, *Dynamic programming*. Princeton Univ. Press, 1957.

[2] M. P. Deisenroth, G. Neumann, J. Peters, *et al.*, "A survey on policy search for robotics," *Foundations and Trends® in Robotics*, vol. 2, no. 1–2, pp. 1–142, 2013.

[3] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[4] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *NIPS*, 2000, pp. 1057–1063.

[5] S. M. Kakade, "A natural policy gradient," in *Advances in neural information processing systems*, 2002, pp. 1531–1538.

[6] J. Kober and J. Peters, "Learning motor primitives for robotics," in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE, 2009, pp. 2112–2118.

[7] J. Peters and S. Schaal, "Reinforcement learning by reward-weighted regression for operational space control," in *ICML*, 2007, pp. 745–750.

[8] J. Peters, K. Mülling, and Y. Altun, "Relative entropy policy search." in *AAAI*. Atlanta, 2010, pp. 1607–1612.

[9] H. Föllmer and T. Knispel, "Entropic risk measures: Coherence vs. convexity, model ambiguity and robust large deviations," *Stochastics and Dynamics*, vol. 11, no. 02n03, pp. 333–351, 2011.

[10] R. A. Howard and J. E. Matheson, "Risk-sensitive markov decision processes," *Management science*, vol. 18, no. 7, pp. 356–369, 1972.

[11] D. Jacobson, "Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games," *IEEE Transactions on Automatic control*, vol. 18, no. 2, pp. 124–131, 1973.

[12] P. Whittle, "A risk-sensitive maximum principle," *Systems & Control Letters*, vol. 15, no. 3, pp. 183–192, 1990.

[13] W. H. Fleming and W. M. McEneaney, "Risk sensitive optimal control and differential games," in *Stochastic theory and adaptive control*. Springer, 1992, pp. 185–197.

[14] A. Tamar, D. Di Castro, and S. Mannor, "Policy gradients with variance related risk criteria," in *ICML*, 2012, pp. 1651–1658.

[15] L. Prashanth, "Policy gradients for cvar-constrained mdps," in *ALT*, 2014, pp. 155–169.

[16] A. Tamar, "Risk-sensitive and efficient reinforcement learning algorithms," Ph.D. dissertation, Israel Institute of Technology, 2015.

[17] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[18] P. Whittle, "Risk sensitivity, a strangely pervasive concept," *Macroeconomic Dynamics*, vol. 6, no. 1, pp. 5–18, 2002.

[19] ——, "Risk-sensitive linear/quadratic/gaussian control," *Advances in Applied Probability*, vol. 13, no. 4, pp. 764–777, 1981.

[20] J. Kober, A. Wilhelm, E. Oztop, and J. Peters, "Reinforcement learning to adjust parametrized motor primitives to new situations," *Autonomous Robots*, vol. 33, no. 4, pp. 361–379, 2012.

[21] M. J. Wainwright, M. I. Jordan, *et al.*, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[22] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[23] D. Wierstra, T. Schaul, J. Peters, and J. Schmidhuber, "Natural evolution strategies," in *Congress on Evolutionary Computation*. IEEE, 2008, pp. 3381–3387.

[24] G. Neumann, "Variational inference for policy search in changing situations," in *ICML*, 2011, pp. 817–824.

[25] A. Abdolmaleki, B. Price, N. Lau, L. P. Reis, and G. Neumann, "Deriving and improving cma-es with information geometric trust regions," in *GECCO*, 2017, pp. 657–664.

[26] A. Abdolmaleki, J. T. Springenberg, Y. Tassa, R. Munos, N. Heess, and M. Riedmiller, "Maximum a posteriori policy optimisation," *arXiv preprint arXiv:1806.06920*, 2018.

[27] S. Boyd, E. Busseti, S. Diamond, R. N. Kahn, K. Koh, P. Nystrup, J. Speth, *et al.*, "Multi-period trading via convex optimization," *Foundations and Trends® in Optimization*, vol. 3, no. 1, pp. 1–76, 2017.

[28] A. Paraschos, C. Daniel, J. R. Peters, and G. Neumann, "Probabilistic movement primitives," in *NIPS*, 2013, pp. 2616–2624.

[29] A. Rajeswaran, K. Lowrey, E. V. Todorov, and S. M. Kakade, "Towards generalization and simplicity in continuous control," in *NIPS*, 2017, pp. 6550–6561.