PARTICLE VALUE FUNCTIONS

Chris J. Maddison[1,2], Dieterich Lawson[3], George Tucker[3],
Nicolas Heess[2], Arnaud Doucet[1], Andriy Mnih[2], Yee Whye Teh[1,2]
[1]University of Oxford, [2]DeepMind, [3]Google Brain
cmaddis@stats.ox.ac.uk

ABSTRACT

The policy gradients of the expected return objective can react slowly to rare rewards. Yet, in some cases agents may wish to emphasize the low or high returns regardless of their probability. Borrowing from the economics and control literature, we review the risk-sensitive value function that arises from an exponential utility and illustrate its effects on an example. This risk-sensitive value function is not always applicable to reinforcement learning problems, so we introduce the *particle value function* defined by a particle filter over the distributions of an agent's experience, which bounds the risk-sensitive one. We illustrate the benefit of the policy gradients of this objective in Cliffworld.

1 INTRODUCTION

The expected return objective dominates the field of reinforcement learning, but makes it difficult to express a tolerance for unlikely rewards. This kind of risk sensitivity is desirable, e.g., in real-world settings such as financial trading or safety-critical applications where the risk required to achieve a specific return matters greatly. Even if we ultimately care about the expected return, it may be beneficial during training to tolerate high variance in order to discover high reward strategies.

In this paper we introduce a risk-sensitive value function based on a system of interacting trajectories called a *particle value function* (PVF). This value function is amenable to large-scale reinforcement learning problems with nonlinear function approximation. The idea is inspired by recent advances in variational inference which bound the log marginal likelihood via importance sampling estimators (Burda et al., 2016; Mnih & Rezende, 2016), but takes an orthogonal approach to reward modifications, e.g. (Schmidhuber, 1991; Ng et al., 1999). In Section 2, we review risk sensitivity and a simple decision problem where risk is a consideration. In Section 3, we introduce a particle value function. In Section 4, we highlight its benefits on Cliffworld trained with policy gradients.

2 RISK SENSITIVITY AND EXPONENTIAL UTILITY

We look at a finite horizon Markov Decision Process (MDP) setting where R_t is the instantaneous reward generated by an agent following a non-stationary policy π, see Appendix A. A utility function $u : \mathbb{R} \to \mathbb{R}$ is an invertible non-decreasing function, which specifies a ranking over possible returns $\sum_{t=0}^{T} R_t$. The expected utility $\mathbb{E}[u(\sum_{t=0}^{T} R_t)|S_0 = s]$ specifies a ranking over policies (Von Neumann & Morgenstern, 1953). For an agent following u, a natural definition of the "value" of a state is the real number $V_T^\pi(s, u)$ whose utility is the expected utility:

$$V_T^\pi(s, u) = u^{-1}\left(\mathbb{E}\left[u\left(\sum_{t=1}^{T} R_t\right)\,\middle|\, S_0 = s\right]\right). \tag{1}$$

Note, when u is the identity we recover the expected return. We consider exponential utilities $u(x) = \operatorname{sgn}(\beta)\exp(\beta x)$ where $\beta \in \mathbb{R}$. This choice is well-studied, and it is implied by the assumption that $V_T^\pi(s, u)$ is additive for deterministic translations of the reward function (Pratt, 1964; Howard & Matheson, 1972; Coraluppi, 1997). The corresponding value function is

$$V_T^\pi(s, \beta) = \frac{1}{\beta}\log\mathbb{E}\left[\exp\left(\beta\sum_{t=0}^{T} R_t\right)\,\middle|\, S_0 = s\right], \tag{2}$$

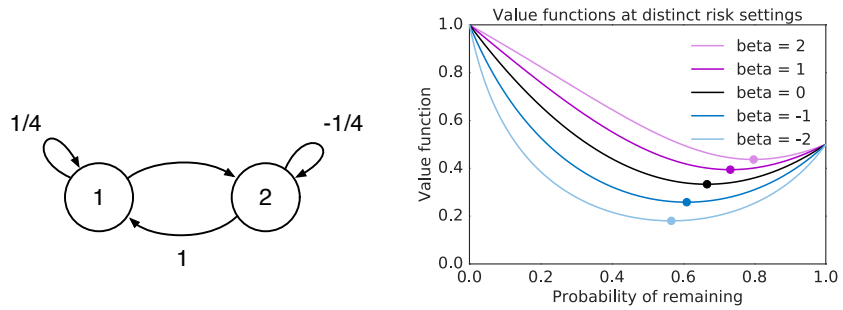

Figure 1: A two state MDP. The plot shows $V_2^p(1, \beta)$ for distinct β, assuming aliased states and a policy parameterized simply by the probability $p \in [0, 1]$ of remaining in a state.

and as $\beta \to 0$ we recover the expected return. See Appendix B for details. One way to interpret this value is through the following thought experiment. If the agent is given a choice between a single interaction with the environment and an immediate deterministic return, then $V_T^\pi(s, \beta)$ represents the minimum return that our agent would take in exchange for forgoing an interaction. If $\beta < 0$, then $V_T^\pi(s, \beta) \leq V_T^\pi(s, 0)$, meaning that the agent is willing to to take a loss relative to the expected return in exchange for certainty. This is a risk-avoiding attitude, which emphasizes low returns. If $\beta > 0$, then $V_T^\pi(s, \beta) \geq V_T^\pi(s, 0)$, and the agent would only forgo an interaction for more than it can expect to receive. This is risk-seeking behavior, which emphasizes high returns.

To illustrate one effect of risk, consider the two state MDP shown in Figure 1. The agent begins in state 1 and acts for two time steps, choosing between leaving or remaining. Suppose that the agent's policy is defined by a single parameter $p \in [0, 1]$ that describes the probability of remaining. Then the expected return $V_2^p(1, 0) = \frac{3}{2}p^2 - 2p + 1$ has two local maxima at $p \in \{0, 1\}$ and the solution $p = 1$ is not a global maximum. Any policy gradient trajectory initialized with $p > 2/3$ will converge to the suboptimal solution $p = 1$, but as our risk appetite β grows, the basin of attraction to the global maximum of $p = 0$ expands to the entire unit interval. This sort of state aliasing happens often in reinforcement learning with non-linear function approximation. In these cases, modifying the risk appetite (either towards risk-avoidance or seeking) may favorably modify the convergence of policy gradient algorithms, even if our ultimate objective is the expected return.

The risk-seeking variant may be helpful in deterministic environments, where an agent can exactly reproduce a previously experienced trajectory. Rare rewards are rare only due to our current policy, and it may be better to pursue high yield trajectories more aggressively. Note, however, that $V_T^\pi(s, \beta)$ is non-decreasing in β, so in general risk-seeking is not guaranteed to improve the expected return. Note also that the literature on KL regularized control (Todorov, 2006; Kappen, 2005; Tishby & Polani, 2011) gives a different perspective on risk sensitive control, which mirrors the relationship between variational inference and maximum likelihood. See Appendix C for related work.

3 Particle Value Functions

Algorithms for optimizing $V_T^\pi(s, \beta)$ may suffer from numerical issues or high variance, see Appendix B. Instead we define a value function that bounds $V_T^\pi(s, \beta)$ and approaches it in the infinite sample limit. We call it a particle value function, because it assigns a value to a bootstrap particle filter with K particles representing state-action trajectories. This is distinct, but related to Kantas (2009), which investigates particle filter algorithms for infinite horizon risk-sensitive control.

Briefly, a bootstrap particle filter can be used to estimate normalizing constants in a hidden Markov model (HMM). Let (X_t, Y_t) be the states of an HMM with transitions $X_t \sim p(\cdot|X_{t-1})$ and emissions $Y_t \sim q(\cdot|X_t)$. Given a sample $y_0 \dots y_T$, the probability $p(\{Y_t = y_t\}_{t=0}^T)$ can be computed with the forward algorithm. The bootstrap particle filter is a stochastic procedure for the forward algorithm that avoids integrating over the state space of the latent variables. It does so by propagating a set of K particles $X_t^{(i)}$ with the transition model $X_t^{(i)} \sim p(\cdot|X_{t-1}^{(i)})$ and a resampling step in proportion to the potentials $q(y_t|X_t^{(i)})$. The result is an unbiased estimator $\prod_{t=0}^T (K^{-1} \sum_{i=1}^K q(y_t|X_t^{(i)}))$ of the desired probability (Del Moral, 2004; Pitt et al., 2012). The insight is that if we treat the state-action pairs (S_t, A_t) as the latents of an HMM with emission potentials $\exp(\beta R_t(S_t, A_t))$ (similar to Toussaint & Storkey, 2006; Rawlik et al., 2010), then a bootstrap particle filter returns an unbiased estimate of $\mathbb{E}[\exp(\beta \sum_{t=0}^T R_t)|S_0 = s]$. Algorithm 1 summarizes this approach.

Algorithm 1 An estimator of the PVF $V^{\pi}_{T,K}(s^{(1)}, \ldots, s^{(K)}, \beta)$

for $i = 1 : K$ **do**
  $S_0^{(i)} = s^{(i)}$
  $A_0^{(i)} \sim \pi_T(\cdot|s^{(i)})$
  $W_0^{(i)} = \exp(\beta R_0^{(i)})$
end for
$Z_0 = \frac{1}{K}\sum_{i=1}^{K} W_0^{(i)}$
for $t = 1 : T$ **do**
  for $i = 1 : K$ **do**
    $I \sim \mathbb{P}(I = j) \propto W_{t-1}^{(j)}$ # select random parent
    $S_t^{(i)} \sim p(\cdot|S_{t-1}^{(I)}, A_{t-1}^{(I)})$ # inherit from parent
    $A_t^{(i)} \sim \pi_{T-t}(\cdot|S_t^{(i)})$
    $W_t^{(i)} = \exp(\beta R_t^{(i)})$
  end for
  $Z_t = \frac{1}{K}\sum_{i=1}^{K} W_t^{(i)}$
end for
return $\frac{1}{\beta}\sum_{t=0}^{T} \log Z_t$

Taking an expectation over all of the random variables not conditioned on we define the PVF associated with the bootstrap particle filter dynamics:

$$V^{\pi}_{T}(s^{(1)}, \ldots, s^{(K)}, \beta) = \mathbb{E}\left[\frac{1}{\beta}\sum_{t=0}^{T} \log Z_t \,\middle|\, \left\{S_0^{(i)} = s^{(i)}\right\}_{i=1}^{K}\right]. \tag{3}$$

Note, more sophisticated sampling schemes, see Doucet & Johansen (2011), result in distinct PVFs.

Consider the value if we initialize all particles at s, $V^{\pi}_{T,K}(s, \beta) = V^{\pi}_{T}(s, \ldots, s, \beta)$. If $\beta > 0$, then by Jensen's inequality and the unbiasedness of the estimator we have that $V^{\pi}_{T,K}(s, \beta) \leq V^{\pi}_{T}(s, \beta)$. For $\beta < 0$ the bound is in the opposite direction. It is informative to consider the behaviour of the trajectories for different values of β. For $\beta > 0$ this algorithm greedily prefers trajectories that encounter large rewards, and the aggregate return is a per time step soft-max. For $\beta < 0$ this algorithm prefers trajectories that encounter large negative rewards, and the aggregate return is a per time step soft-min. See Appendix D for the Bellman equation and policy gradient of this PVF.

4 Experiments

To highlight the benefits of using PVFs we apply them to a variant of the Gridworld task called Cliffworld, see Appendix E for comparison to other methods and more details. We trained time dependent tabular policies using policy gradients from distinct PVFs for $\beta \in \{-1, -0.5, 0, 0.5, 1, 2\}$. We tried $K \in \{1, \ldots, 8\}$ and learning rates $\epsilon \in \{1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}, 5 \times 10^{-5}\}$. For the $\beta = 0$ case we ran K independent non-interacting trajectories and averaged over a policy gradient with estimated baselines. Figure 2 shows the density over the final state of the trained MDP under varying β treatments but $K = 4$. Notice that the higher the risk parameter, the broader the policy, with the agent eventually solving the task. No $\beta = 0$, corresponding to standard REINFORCE, runs solved this task, even after increasing the number of agents to 64.

5 Conclusion

We introduced the particle value function, which approximates a risk-sensitive value function for a given MDP. We will seek to address theoretical questions, such as whether the PVF is increasing in β and monotonic in the number of particles. Also, the PVF does not have an efficient tabular representation, so understanding the effect of efficient approximations would be valuable. Experimentally, we hope to explore these ideas for complex sequential tasks with non-linear function approximators. One obvious example of such tasks is variational inference over a sequential model.

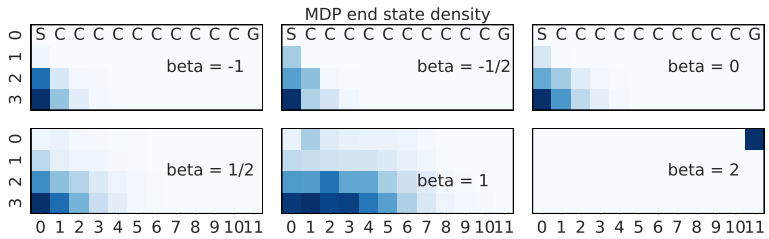

Figure 2: Last state distribution under policies trained with PVFs with distinct β.

ACKNOWLEDGEMENTS

We thank Rémi Munos, Theophane Weber, David Silver, Marc G. Bellemare, and Danilo J. Rezende for helpful discussion and support in this project.

REFERENCES

Francesca Albertini and Wolfgang J Runggaldier. Logarithmic transformations for discrete-time, finite-horizon stochastic control problems. *Applied mathematics & optimization*, 18(1):143–161, 1988.

Kenneth Joseph Arrow. Essays in the theory of risk-bearing. 1974.

Nicole Bäuerle and Ulrich Rieder. More risk-sensitive markov decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2013.

Vivek S Borkar and Sean P Meyn. Risk-sensitive optimal control for markov decision processes with monotone cost. *Mathematics of Operations Research*, 27(1):192–209, 2002.

Bart van den Broek, Wim Wiegerinck, and Hilbert Kappen. Risk sensitive path integral control. *arXiv preprint arXiv:1203.3523*, 2012.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance Weighted Autoencoder. In *ICLR*, 2016.

Stefano Coraluppi. *Optimal control of Markov decision processes for performance and robustness*. PhD thesis, University of Maryland, 1997.

Peter Dayan and Geoffrey E Hinton. Using expectation-maximization for reinforcement learning. *Neural Computation*, 9(2):271–278, 1997.

Pierre Del Moral. Feynman-kac formulae. In *Feynman-Kac Formulae*, pp. 47–93. Springer, 2004.

Arnaud Doucet and Adam M Johansen. A tutorial on particle filtering and smoothing: fiteen years later. 2011.

Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. *arXiv preprint arXiv:1512.08562*, 2015.

Matt Hoffman, Arnaud Doucet, Nando De Freitas, and Ajay Jasra. On solving general state-space sequential decision problems using inference algorithms. Technical report, Technical Report TR-2007-04, University of British Columbia, Computer Science, 2007.

Ronald A Howard and James E Matheson. Risk-sensitive Markov decision processes. *Management science*, 18(7):356–369, 1972.

Nikolas Kantas. *Sequential decision making in general state space models*. PhD thesis, Citeseer, 2009.

Nikolas Kantas, Arnaud Doucet, Sumeetpal S Singh, Jan Maciejowski, Nicolas Chopin, et al. On particle methods for parameter estimation in state-space models. *Statistical science*, 30(3):328–351, 2015.

Hilbert J Kappen. Linear theory for control of nonlinear stochastic systems. *Physical review letters*, 95(20):200201, 2005.

Hilbert J Kappen, Vicenç Gómez, and Manfred Opper. Optimal control as a graphical model inference problem. *Machine learning*, 87(2):159–182, 2012.

Sven Koenig and Reid G Simmons. Risk-sensitive planning with probabilistic decision graphs. In *Proceedings of the 4th international conference on principles of knowledge representation and reasoning*, pp. 363, 1994.

Steven I Marcus, Emmanual Fernández-Gaucherand, Daniel Hernández-Hernandez, Stefano Coraluppi, and Pedram Fard. Risk sensitive markov decision processes. In *Systems and control in the twenty-first century*, pp. 263–279. Springer, 1997.

Oliver Mihatsch and Ralph Neuneier. Risk-sensitive reinforcement learning. *Machine learning*, 49 (2-3):267–290, 2002.

Andriy Mnih and Danilo Rezende. Variational Inference for Monte Carlo Objectives. In *ICML*, 2016.

Ralph Neuneier and Oliver Mihatsch. Risk sensitive reinforcement learning. In *Proceedings of the 11th International Conference on Neural Information Processing Systems*, pp. 1031–1037. MIT Press, 1998.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, pp. 278–287, 1999.

Michael K Pitt, Ralph dos Santos Silva, Paolo Giordani, and Robert Kohn. On some properties of markov chain monte carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134–151, 2012.

John W Pratt. Risk aversion in the small and in the large. *Econometrica*, pp. 122–136, 1964.

Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. An approximate inference approach to temporal optimization in optimal control. In *Advances in neural information processing systems*, pp. 2011–2019, 2010.

H-Ch Ruiz and HJ Kappen. Particle smoothing for hidden diffusion processes: Adaptive path integral smoother. *arXiv preprint arXiv:1605.00278*, 2016.

Jürgen Schmidhuber. Curious model-building control systems. In *Neural Networks, 1991. 1991 IEEE International Joint Conference on*, pp. 1458–1463. IEEE, 1991.

Yun Shen, Michael J Tobia, Tobias Sommer, and Klaus Obermayer. Risk-sensitive reinforcement learning. *Neural computation*, 26(7):1298–1328, 2014.

Naftali Tishby and Daniel Polani. Information theory of decisions and actions. In *Perception-action cycle*, pp. 601–636. Springer, 2011.

Emanuel Todorov. Linearly-solvable markov decision problems. In *NIPS*, pp. 1369–1376, 2006.

Marc Toussaint and Amos Storkey. Probabilistic inference for solving discrete and continuous state markov decision processes. In *Proceedings of the 23rd international conference on Machine learning*, pp. 945–952. ACM, 2006.

John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1953.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

A MARKOV DECISION PROCESSES

We consider decision problems in which an agent selects actions and receives rewards in a stochastic environment. For the sake of exposition, we consider a finite horizon MDP, which consists of: a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, a stationary environmental transition kernel satisfying the Markov property $p(\cdot|S_t, A_t, \ldots, S_0, A_0) = p(\cdot|S_t, A_t)$, and reward functions $r_{T-t} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. At each time step the agent chooses actions according to a policy $\pi_{T-t}(\cdot|S_t)$ given the current state. $\pi_{T-t}(\cdot|S_t)$ is the action distribution and r_{T-t} the reward function when there are $T-t$ steps remaining. All together the MDP proceeds stochastically producing a sequence of random variables (S_t, A_t) according to the following dynamics for $T \in \mathbb{N}$ time steps. Let $t \in \{0, \ldots, T\}$,

$$S_0 = s \tag{4}$$

$$A_t \sim \pi_{T-t}(\cdot|S_t) \tag{5}$$

$$S_{t+1} \sim p(\cdot|S_t, A_t) \tag{6}$$

The agent receives a reward $R_t = r_{T-t}(S_t, A_t)$ at each time step. We will call a single realization of the MDP a trajectory. The objective in classical reinforcement learning is to discover the policies $\pi = \{\pi_t(\cdot|s)\}_{t=0}^T$ that maximize the value function,

$$V_T^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{T} R_t \,\middle|\, S_0 = s\right]. \tag{7}$$

where the expectation is taken with respect to all the stochastic elements not conditioned on.

All of the results we present can be simply extended to the infinite horizon case with discounted or episodic returns as well as more general uncountable state and action spaces.

B RISK-SENSITIVE VALUE FUNCTION DETAILS

Utility theory gives us a language for describing the relative importance of high or low returns. A utility function $u : \mathbb{R} \to \mathbb{R}$ is an invertible non-decreasing function, which specifies a ranking over possible returns $\sum_{t=0}^T R_t$. The expected utility $\mathbb{E}[u(\sum_{t=0}^T R_t)|S_0 = s]$ specifies a ranking over policies (Von Neumann & Morgenstern, 1953). The expected utility does not necessarily have an interpretable scale, because any affine transformation of the utility function results in the same relative ordering of policies or return outcomes. Therefore we define the value associated with a utility u by returning it to the scale of the rewards defined by the MDP. For an agent following u, the "value" of a state is the real number $V_T^\pi(s, u)$ whose utility is the expected utility:

$$V_T^\pi(s, u) = u^{-1}\left(\mathbb{E}\left[u\left(\sum_{t=1}^{T} R_t\right) \,\middle|\, S_0 = s\right]\right). \tag{8}$$

Note that when u is the identity we recover the expected return. Of course for non-decreasing invertible utilities, the value gives the same ranking over policies. One way to interpret this value is through the following thought experiment. If the agent is given a choice between a single interaction with the environment or an immediate deterministic return, then $V_T^\pi(s, u)$ represents the minimum return that our agent would take in exchange for forgoing an interaction. If

$$V_T^\pi(s, u) \leq \mathbb{E}\left[\sum_{t=0}^{T} R_t \middle| S_0 = s\right]$$

then our agent is willing to take a loss relative to the expected return in exchange for certainty. This is a risk-avoiding attitude, which emphasizes the smallest returns, and one can show that this occurs iff u is concave. If

$$V_T^\pi(s, u) \geq \mathbb{E}\left[\sum_{t=0}^{T} R_t \middle| S_0 = s\right]$$

then the agent would only forgo an interaction for more than it can expect to receive. This is risk-seeking behavior, which emphasizes the largest returns, and one can show that this occurs iff u is

convex. The case when $u(x)$ is linear is the risk-neutral case. For these reasons, $V_T^\pi(s,u)$ is also known as the *certain equivalent* in economics (Howard & Matheson, 1972).

We focus on exponential utilities of the form $u(x) = \operatorname{sgn}(\beta)\exp(\beta x)$ where $\beta \in \mathbb{R}$. This is a broadly studied choice that is implied by the assumption that the value function $V_T^\pi(s,u)$ is additive for deterministic translations of the return (Pratt, 1964; Howard & Matheson, 1972; Coraluppi, 1997). This assumption is nice, because it preserves the Markov nature of the decision process: if the agent were given a choice at *every* time step t between continuing the interaction or terminating and taking its value as a deterministic return, then additivity in the value function means that the same decision is made regardless of the return accumulated so far (Howard & Matheson, 1972). The value function corresponding to an exponential utility is

$$V_T^\pi(s,\beta) = \frac{1}{\beta}\log\mathbb{E}\left[\exp\left(\beta\sum_{t=0}^{T} R_t\right)\,\middle|\, S_0 = s\right], \tag{9}$$

and as $\beta \to 0$ we recover the expected return. We list a few of its properties.

1. For β near 0
$$V_T^\pi(s,\beta) = \mathbb{E}\left[\sum_{t=0}^{T} R_t \,\middle|\, S_0 = s\right] + \frac{\beta}{2}\operatorname*{var}_\pi\left[\sum_{t=0}^{T} R_t \,\middle|\, S_0 = s\right] + o(\beta^2) \tag{10}$$
2. $\lim_{\beta\to\infty} V_T^\pi(s,\beta) = \sup\{r|\,\mathbb{P}(\sum_{t=0}^{T} R_t = r) > 0\}$
3. $\lim_{\beta\to-\infty} V_T^\pi(s,\beta) = \inf\{r|\,\mathbb{P}(\sum_{t=0}^{T} R_t = r) > 0\}$
4. $V_T^\pi(s,\beta)$ is continous and non-decreasing in β.
5. $V_T^\pi(s,\beta)$ is risk-seeking for $\beta > 0$, risk-avoiding for $\beta = 0$, and risk-neutral for $\beta = 0$

For proofs,

1. From Coraluppi (1997).
2. From Coraluppi (1997).
3. From Coraluppi (1997).
4. $V_T^\pi(s,\beta)$ is clearly continuous for all $\beta \neq 0$. If we extend $V_T^\pi(s,0) \equiv \mathbb{E}\left[\sum_{t=0}^{T} R_t \,\middle|\, S_0 = s\right]$, then 1. gives us the continuity everywhere. For non-decreasing let $\alpha,\beta \in \mathbb{R}$ and $\alpha \neq 0$ and $\beta \neq 0$. Furthermore assume $\beta \geq \alpha$. Then $\beta/\alpha > 0$ or $\beta/\alpha \geq 1$. Now,
$$\begin{aligned}\mathbb{E}\left[\exp\left(\beta\sum_{t=0}^{T} R_t\right)\,\middle|\, S_0 = s\right] &= \mathbb{E}\left[\exp\left(\alpha\sum_{t=0}^{T} R_t\right)^{\beta/\alpha}\,\middle|\, S_0 = s\right] \\ &\geq \mathbb{E}\left[\exp\left(\alpha\sum_{t=0}^{T} R_t\right)\,\middle|\, S_0 = s\right]^{\beta/\alpha}\end{aligned}$$
since x^p is convex on $x > 0$ for $p \geq 1$ or $p < 0$, Jensen's inequality gives us the result. Taking log of both sides gives us the result in that case. In the case that $\alpha = 0$ or $\beta = 0$, 4. and Jensen's inequality gives us the result by the concavity of log.

From a practical point of view the value function $V_T^\pi(s,\beta)$ behaves like a soft-max or soft-min depending on the sign of β, emphasizing the avoidance of low returns when $\beta < 0$ and the pursuit of high returns when $\beta > 0$. As $\beta \to \infty$ the value $V_T^\pi(s,\beta)$ approaches the supremum of the returns over all trajectories with positive probability, a best-case penalty. As $\beta \to -\infty$ it approaches the infimum, a worst-case value (Coraluppi, 1997). Thus for large positive β this value is tolerant of high variance if it can lead to high returns. For large negative β it is very intolerant of rare low returns.

Despite having attractive properties the risk-sensitive value function is not always applicable to reinforcement learning tasks (see also Mihatsch & Neuneier (2002)). The value function satisfies the multiplicative Bellman equation

$$\exp(\beta V_T^\pi(s,\beta)) = \sum_{a,s'} \pi_T(a|s) p(s'|s,a) \exp(\beta r_T(s,a) + \beta V_{T-1}^\pi(s,\beta)). \tag{11}$$

Operating in $\log$-space breaks the ability to exploit this recurrence from Monte Carlo returns generated by a single trajectory, because expectations do not exchange with $\log$. Operating in $\exp$-space is possible for TD learning algorithms, but we must cap the minimum/maximum possible return so that $\exp(\beta V_T^\pi(s,\beta))$ does not underflow/overflow. This can be an issue when the rewards represent log probabilities as is often the case in variational inference. The policy gradient of $V_T^\pi(s,\beta)$ is

$$\nabla_\pi V_T^\pi(s,\beta) = \mathbb{E}\left[\frac{1}{\beta}\sum_{t=0}^{T} \exp(\beta Q_{T-t}^\pi(S_t,A_t,\beta) - \beta V_T^\pi(S_t,\beta)) \nabla \log \pi_{T-t}(A_t|S_t) \middle| S_0 = s\right] \tag{12}$$

where

$$Q_T^\pi(s,a,\beta) = \frac{1}{\beta}\log \mathbb{E}\left[\exp\left(\beta \sum_{t=0}^{T} R_t\right) \middle| S_0 = s, A_0 = a\right] \tag{13}$$

Even ignoring underflow/overflow issues, REINFORCE (Williams, 1992) style algorithms would find difficulties, because deriving unbiased estimators of the ratio $\exp(\beta Q_{T-t}^\pi(S_t,A_t,\beta) - \beta V_T^\pi(s,\beta))$ from single trajectories of the MDP may be hard. Lastly, the policy gradient of $\mathbb{E}\left[\exp\left(\beta \sum_{t=0}^{T} R_t\right) \middle| S_0 = s\right]$

$$\nabla_\pi \mathbb{E}\left[\exp\left(\beta \sum_{t=0}^{T} R_t\right) \middle| S_0 = s\right] = \mathbb{E}\left[\sum_{t=0}^{T} \exp\left(\beta \sum_{t=0}^{T} R_t\right) \nabla \log \pi_{T-t}(A_t|S_t) \middle| S_0 = s\right], \tag{14}$$

There are particle methods that would address the estimation of this score, e.g. (Kantas et al., 2015), but for large T the estimate suffers from high mean squared errors.

C Related Work

Risk sensitivity originates in the study of utility and choice in economics (Von Neumann & Morgenstern, 1953; Pratt, 1964; Arrow, 1974). It has been extensively studied for the control of MDPs (Howard & Matheson, 1972; Coraluppi, 1997; Marcus et al., 1997; Borkar & Meyn, 2002; Mihatsch & Neuneier, 2002; Bäuerle & Rieder, 2013). In reinforcement learning, risk sensitivity has been studied (Koenig & Simmons, 1994; Neuneier & Mihatsch, 1998; Shen et al., 2014), although none of these consider the direct policy gradient approach considered in this work. Most of the methods considered are variants of a Q learning approach or policy iteration. As well, the idea of treating rewards as emissions of an HMM is not a new idea (Toussaint & Storkey, 2006; Rawlik et al., 2010).

The idea of treating reinforcement learning as an inference problem is not a new idea (see e.g. Albertini & Runggaldier, 1988; Dayan & Hinton, 1997; Kappen, 2005; Toussaint & Storkey, 2006; Hoffman et al., 2007; Tishby & Polani, 2011; Kappen et al., 2012). Broadly speaking, all of these works still optimize the expected reward objective $V_T^\pi(s,0) = \mathbb{E}[\sum_{t=0}^{T} R_t \mid S_0 = s]$ with or without some regularization penalties on the policy. The ones that share the closest connection to the risk sensitive objective $V_T^\pi(s,\beta)$ studied here, are the KL regularized objectives of the form

$$\hat{V}_T^{\pi,\pi'}(s,\beta) = \mathop{\mathbb{E}}_{\pi'}\left[\sum_{t=0}^{T} R_t + \frac{1}{\beta}\log\frac{\pi(A_t|S_t)}{\pi'(A_t|S_t)} \middle| S_0 = s\right] \tag{15}$$

where the MDP dynamics are sampled from π'. These are studied for example in Albertini & Runggaldier (1988); Kappen (2005); Tishby & Polani (2011); Kappen et al. (2012); Fox et al. (2015);

Ruiz & Kappen (2016). The observation is that in an MDP with *fully controllable* transition dynamics, optimizing a policy π', which completely specifies the transition dynamics, achieves the risk sensitive value at π:

$$\max_{\pi'} \hat{V}_T^{\pi,\pi'}(s,\beta) = V_T^{\pi}(s,\beta) \tag{16}$$

Note that this has an interesting connection to Bayesian inference. Here, π plays the role of the prior, π' the role of the variational posterior, $\hat{V}_T^{\pi,\pi'}(s,\beta)$ the role of the variational lower bound, and $V_T^{\pi}(s,\beta)$ the role of the marginal likelihood. In effect, KL regularized control is like variational inference, where risk sensitive control is like maximum likelihood. Finally, when the environmental dynamics $p(\cdot|s,a)$ are stochastic, (16) does not *necessarily* hold, therefore the risk sensitive value is distinct in this case. Yet, in certain special cases, risk sensitive objectives can also be cast as solutions to path integral control problems (Broek et al., 2012).

To our knowledge no work has considered using particle filters for risk sensitive control by treating the particle filter's estimator of the log partition function as a return whose expectation bounds the risk sensitive value and whose policy gradients are cheap to compute.

D Particle Value Function Details

Recalling Algorithm 1 and the definition of the MDP in Appendix A, define

$$R_t^{(i)} = r_{T-t}(S_t^{(i)}, A_t^{(i)}) \tag{17}$$

the particle value function associated with the bootstrap particle filter dynamics:

$$V_T^{\pi}(s^{(1)},\dots,s^{(K)},\beta) = \mathbb{E}\left[\frac{1}{\beta}\sum_{t=0}^{T}\log Z_t \;\middle|\; \left\{S_0^{(i)} = s^{(i)}\right\}_{i=1}^{K}\right]. \tag{18}$$

We can also think of this value function as the expected return of an agent whose actions space is the product space $\mathcal{A}^K$, in an environment with state space $\mathcal{S}^K$ whose transition kernel includes the resampling dynamic. Let $s^{(1:K)} = (s^{(1)},\dots,s^{(K)})$, then the PVF satisfies the Bellman equation

$$\beta V_T^{\pi}(s^{(1:K)},\beta) = \sum_{a^{(1:K)}} \Pi_T(a^{(1:K)}|s^{(1:K)}) \log Z_T(a^{(1:K)}, s^{(1:K)}) + \tag{19}$$

$$\sum_{a^{(1:K)}}\sum_{\sigma^{(1:K)}} \Pi_T(a^{(1:K)}|s^{(1:K)}) P_T(\sigma^{(1:K)}|a^{(1:K)}, s^{(1:K)}) \beta V_{T-1}^{\pi}(\sigma^{(1:K)},\beta) \tag{20}$$

where

$$\Pi_T(a^{(1:K)}|s^{(1:K)}) = \prod_{i=1}^{K}\pi_T(a^{(i)}|s^{(i)}) \tag{21}$$

$$\log Z_T(a^{(1:K)}, s^{(1:K)}) = \log\left(\frac{1}{K}\sum_{i=1}^{K}\exp(\beta r_T(s^{(i)}, a^{(i)}))\right) \tag{22}$$

$$P_T(\sigma^{(1:K)}|a^{(1:K)}, s^{(1:K)}) = \prod_{i=1}^{K}\left(\sum_{j=1}^{K}\frac{\exp(\beta r_T(s^{(j)}, a^{(j)}))}{\sum_{k=1}^{K}\exp(\beta r_T(s^{(k)}, a^{(k)}))} p(\sigma^{(i)}|s^{(j)}, a^{(j)})\right) \tag{23}$$

The policy gradient of $V_T^{\pi}(s^{(1:K)},\beta)$ is

$$\nabla_\pi V_T^{\pi}(s^{(1:K)},\beta) = \mathbb{E}\left[\frac{1}{\beta}\sum_{t=0}^{T}\sum_{t'=t}^{T}\sum_{i=1}^{K}\log Z_{t'}\nabla\log\pi_{T-t}(A_t^{(i)}|S_t^{(i)}) \;\middle|\; \{S_0^{(i)} = s^{(i)}\}_{i=1}^{K}\right] \tag{24}$$

In this sense we can think of $\log Z_t/\beta$ as the immediate reward for the whole system of particles and $\sum_{t=0}^{T}\log Z_t/\beta$ as the return.

The key point is that the use of interacting trajectories to generate the Monte Carlo return ensures that this particle value function defines a bound on $V_T^\pi(s,\beta)$. Indeed, consider the particle value function that corresponds to initializing all K trajectories in state $s \in \mathcal{S}$ and define, $V_{T,K}^\pi(s,\beta) = V_T^\pi(s,\ldots,s,\beta)$. Now, for $\beta > 0$ we have, by Jensen's inequality,

$$V_{T,K}^\pi(s,\beta) \leq \frac{1}{\beta}\log \mathbb{E}\left[\prod_{t=0}^{T} Z_t \;\middle|\; \{S_0^{(i)} = s\}_{i=1}^K\right] \tag{25}$$

and since the bootstrap particle filter is unbiased (Del Moral, 2004; Pitt et al., 2012),

$$= V_T^\pi(s,\beta) \tag{26}$$

For $\beta < 0$ we get the reverse inequality, $V_{T,K}^\pi(s,\beta) \geq V_T^\pi(s,\beta)$. As $K \to \infty$ the particle value function converges to $V_T^\pi(s,\beta)$ since the estimator is consistent (Del Moral, 2004). We list this and some other properties:

1. $V_{T,K}^\pi(s,\beta) \leq V_T^\pi(s,\beta)$.
2. $\lim_{K\to\infty} V_{T,K}^\pi(s,\beta) = V_T^\pi(s,\beta)$.
3. $\lim_{\alpha\to 0} V_{T,K}^\pi(s,\alpha) = \mathbb{E}\left[\sum_{t=0}^T R_t \;\middle|\; S_0 = s\right] = V_{T,1}^\pi(s,\beta)$.
4. $V_{T,K}^\pi(s,\beta)$ is continuous in β.

For proofs,

1. $\prod_{t=0}^T Z_t$ is an unbiased estimator of $\mathbb{E}[\exp(\beta\sum_{t=0}^T R_t)|S_0 = s]$ (Del Moral, 2004) and the rest follows from Jensen's inequality.
2. $\prod_{t=0}^T Z_t$ is a consistent estimator (Del Moral, 2004), and the rest follows from exchanging the limit with an expectation.
3. $V_{T,1}^\pi(s,\beta) = \mathbb{E}\left[\sum_{t=0}^T R_t \;\middle|\; S_0 = s\right]$ is clear. Otherwise the limit $\lim_{\beta\to 0} V_{T,K}^\pi(s,\beta)$ approaches the algorithm that resamples uniformly and the value under that sampling strategy is

$$\begin{aligned}
\lim_{\beta\to 0} V_{T,K}^\pi(s,\beta) &= \mathbb{E}\left[\sum_{t=0}^T \lim_{\beta\to 0}\frac{\log Z_t}{\beta} \;\middle|\; \{S_0^{(i)} = s\}_{i=1}^K\right] \\
&= \mathbb{E}\left[\sum_{t=0}^T\sum_{i=1}^K \frac{1}{K} R_t^{(i)} \;\middle|\; \{S_0^{(i)} = s\}_{i=1}^K\right] \\
&= \sum_{t=0}^T\sum_{i=1}^K \frac{1}{K}\,\mathbb{E}\left[R_t^{(i)} \;\middle|\; \{S_0^{(i)} = s\}_{i=1}^K\right]
\end{aligned}$$

   Because each $R_t^{(i)}$ has a genealogy, which is an MDP trajectory

$$\begin{aligned}
&= \sum_{t=0}^T\sum_{i=1}^K \frac{1}{K}\,\mathbb{E}\left[R_t \mid S_0 = s\right] \\
&= \mathbb{E}\left[\sum_{t=0}^T R_t \;\middle|\; S_0 = s\right]
\end{aligned}$$

4. $V_{T,K}^\pi(s,\beta)$ is a finite sum of continuous terms, and if we extend the definition of $V_{T,K}^\pi(s,0) \equiv \lim_{\beta\to 0} V_{T,K}^\pi(s,\beta)$, then we're done.

E CLIFFWORLD DETAILS

We considered a finite horizon Cliffworld task, which is a variant on Gridworld. The world is 4 rows by 12 columns, and the agent can occupy any grid location. Each episode begins with the

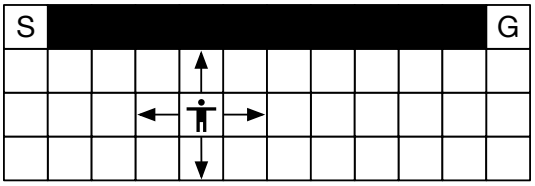

Figure 3: Cliffworld is a $n \times m$ gridworld. S denotes the start state, G the goal state, and the agent is currently in state (4,2). The arrows show the actions available to the agent.

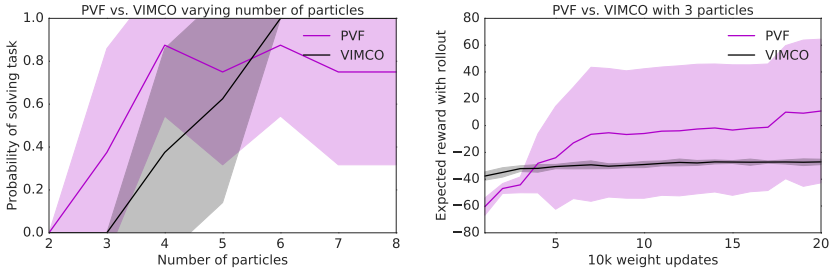

Figure 4: Left plot: probability of solving task with standard deviation, defined as achieving positive average return. Right plot: Average reward during training with standard deviation. Both VIMCO and PVF trained with $\beta = 1.0$ and learning rate $\epsilon = 1 \times 10^{-3}$. Averages are for 8 runs. At 3 particles, some PVF runs began solving Cliffworld, while no VIMCO ones did.

agent in state (0,0) (marked as S in Figure 3) and ends when 24 timesteps have passed. The actions available to the agent are moving north, east, south, and west, but moving off the grid is prohibited. All environmental transitions are deterministic. The 'cliff' occupies all states between the start and the goal (marked G in Figure 3) along the northern edge of the world. All cliff states are absorbing and when the agent enters any cliff state it initially receives a reward of -100 and receives 0 reward each timestep thereafter. The goal state is also absorbing, and the agent receives a +100 reward upon entering it and 0 reward after. The agent receives a -1 reward for every action that does not transition into a cliff or goal state. The optimal policy for Cliffworld is to hug the cliff and proceed from the start to the goal as speedily as possible, but doing so could incur high variance in reward if the agent falls off the cliff. For a uniform random policy most trajectories result in large negative rewards and occasionally a high positive reward. This means that initially for independent trajectories venturing east is high variance and low reward.

We trained non-stationary tabular policies parameterized by parameters θ of size $4 \times 12 \times 4 \times 24$:

$$\pi_{T-t}(a|s) = \frac{\exp(\theta[s_1, s_2, a, T-t])}{\sum_{a=0}^{3} \exp(\theta[s_1, s_2, a, T-t])}$$

The policies were trained using policy gradients from distinct PVFs for $\beta \in \{-1, -0.5, 0, 0.5, 1, 2\}$. We tried $K \in \{1, \ldots, 8\}$ and learning rates $\epsilon \in \{1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}, 5 \times 10^{-5}\}$. For the $\beta = 0$ case we ran K independent non-interacting trajectories and averaged over a policy gradient with estimated baselines. For $\beta = 0$, we used instead a REINFORCE (Williams, 1992) estimator, that was simply estimated from the Monte Carlo returns. For control variates, we used distinct baselines depending on whether $\beta = 0$ or not. For $\beta = 0$, we used a baseline that was an exponential moving average with smoothing factor 0.8. The baselines were also non-stationary, and with dimensionality $4 \times 12 \times 24$. For $\beta \neq 0$ we used no baseline except for VIMCO's control variate (Mnih & Rezende, 2016) for the immediate reward. The VIMCO control variate is not applicable for the whole return as future time steps are correlated with the action through the interaction of trajectories.

We also compared directly to VIMCO (Mnih & Rezende, 2016). Consider VIMCO's value function,

$$\tilde{V}^{\pi}_{T,K}(s, \beta) = \mathbb{E}\left[\frac{1}{\beta} \log\left(\frac{1}{K} \sum_{i=1}^{K} \exp\left(\sum_{t=0}^{T} \beta R_t^{(i)}\right)\right)\right] \tag{27}$$

where $R_t^{(i)}$ is a reward sequence generated by an independent Monte Carlo rollout of the original MDP. VIMCO is also a risk sensitive value function, but it does not decompose over time and so

does not have a temporal Bellman equation. In this case, though, VIMCO policy gradients were able to solve Cliffworld under most of the conditions that the policy gradients of PVF were able to solve. For $K = 3$ and $\beta = 1.0$, PVF occasionally solved Cliffworld while VIMCO did not. See Figure 4. However, once in the regime where VIMCO could solve the task, it did so with more reliability than the PVF variant. Note that in no case did REINFORCE on the expected return solve this variant.