Survey on Risk-Averse Reinforcement Learning

Gersi Doko
Gersi.Doko@unh.edu

1 Introduction

Reinforcement learning (RL) is a powerful framework for developing agents that can learn to make decisions in complex, uncertain environments (Sutton and Barto, 2018; Puterman, 1994; Mehta et al., 2007; Hall et al., 2018). The core goal of RL is to tractably compute an optimal policy for acting in such an environment. A key challenge in RL is that the environment is often stochastic, and decisions made now can affect state distributions encountered in the future. Choosing an objective function that accurately captures the desired behavior of the agent is a difficult task.

We would hope the objective that an RL agent aims to optimize would consider the 'risk' of the actions it takes. Before even phrasing such an objective, one first needs to consider how to quantify risk into a numerical value. A value that encapsulates the uncertainty of the outcome of an action, so that an agent can distinguish between actions that are risky and those that are safe.

There is a vast field of research that focuses on developing risk measures (Föllmer and Schied, 2016; Hau et al., 2023; Bäuerle and Ott, 2011; Howard and Matheson, 1972). The goal of a risk measure is to quantify the uncertainty of a random variable, often while being interpretable and computationally tractable. The intersections of risk and reinforcement learning is the topic of this review, where we will explore the field of risk-averse reinforcement learning.

One must be cautious when optimizing a risk sensitive objective due to the common pitfalls that arise even in the published literature (Hau et al., 2023). Many risk measures do not satisfy the properties common to expectation. And therefore many blunders found in the literature can be attributed to either misunderstanding of the optimization literature or the risk measure itself.

2 Preliminaries

In order to introduce the concepts of risk-averse reinforcement learning, we first need to define some basic concepts in probability theory and reinforcement learning.

Unless explicitly stated we assume all variables are discrete, and finite. We adopt the notation that all random variables are adorned with a $\tilde{\cdot}$ unless obvious from context, all matrices are capitalized, and all sets are calligraphic.

We adopt the common notation that Δ^n denotes the n-dimensional simplex. For two sets A and B, A^B denotes the set of all functions from B to A. For a set A, $|A|$ denotes the cardinality of A.

The environment is a Markov Decision Process (MDP) defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$ where

- $\mathcal{S}$ is the state space,
- $\mathcal{A}$ is the action space,
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta^{\mathcal{S}}$ is the transition probability function,
- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function,
- $\gamma \in [0, 1$ is the discount factor.

Finally, we define the set of all deterministic policies $\Pi_D = \mathcal{A}^{\mathcal{S}}$. And the set of all randomized policies $\Pi_R = [\Delta^{\mathcal{A}}]^{\mathcal{S}}$. We encourage the reader to refer to (Puterman, 1994) for a more detailed introduction to MDPs.

3 Risk and Reward

In this section we introduce the concept of risk measures, and how they are used to quantify the uncertainty of a random variable. We do not delve into the mathematical consequences of risk measures, but rather focus on their interpretation and use in reinforcement learning. For a more detailed approach, we refer the reader to the work of (Föllmer and Schied, 2016).

Definition 1 (Risk Measure)**.** *A Risk Measure is a mapping from a random variable $\tilde{x}$ to a real number. Quantifying the uncertainty of an outcome. Given a metric space $(\Omega, \mathcal{F}, P)$, a risk measure is a function $\psi : \mathcal{X} \to \mathbb{R}$. Satisfying*

1. *Monotonic: $\tilde{x} \leq \tilde{y} \implies \psi(\tilde{x}) \leq \psi(\tilde{y})$*
2. *Translation Invariance: $\psi(\tilde{x} + c) = \psi(\tilde{x}) + c \quad \forall c \in \mathbb{R}$*

Definition 2 (Coherent Risk Measure)**.** *Given a metric space $(\Omega, \mathcal{F}, P)$, a coherent risk measure is a function $\psi : \mathcal{X} \to \mathbb{R}$. Satisfying*

1. *All items in definition 1*
2. *Sub-additivity: $\psi(\tilde{x} + \tilde{y}) \leq \psi(\tilde{x}) + \psi(\tilde{y})$*
3. *Positive Homogeneity: $\psi(\lambda\tilde{x}) = \lambda\psi(\tilde{x})$ for $\lambda \in \mathbb{R}_+$*

Proposition 1. *All coherent risk measures are convex risk measures.*

Proof. Let ψ be a coherent risk measure. Then for $\tilde{x}, \tilde{y} \in \mathcal{X}$ and $\lambda \in [0, 1]$ we have

$$\begin{aligned} \psi(\lambda\tilde{x} + (1-\lambda)\tilde{y}) &\leq \psi(\lambda\tilde{x}) + \psi((1-\lambda)\tilde{y}) && \text{Sub-Additivity} \\ &= \lambda\psi(\tilde{x}) + (1-\lambda)\psi(\tilde{y}) && \text{Positive Homogeneity} \end{aligned}$$

□

It is important to note the converse of Proposition 1 is not true. That is, not all convex risk measures are coherent. This is due to positive homogeneity.

Definition 3 (Variance)**.** *The variance of a random variable X is defined as*

$$Var(\tilde{x}) = \mathbb{E}[(\tilde{x} - \mathbb{E}[\tilde{x}])^2].$$

Definition 4 (Expectation)**.** *A common risk measure is the expectation of a random variable. Given a random variable $X : \Omega \to \mathbb{R}$, the expectation of X is defined as*

$$\mathbb{E}[\tilde{x}] = \sum_{\omega \in \Omega} \tilde{x}(\omega)\mathbb{P}(\omega).$$

Definition 5 (Value-at-Risk — VaR_α)**.** *For $\alpha \in [0, 1]$, the Value-at-Risk of a random variable $\tilde{x}$ at risk-level α is defined as*

$$\text{VaR}_\alpha(\tilde{x}) = \sup_{t \in \mathbb{R}} \{\mathbb{P}(\tilde{x} < t) \leq \alpha\} = \inf_{t \in \mathbb{R}} \{\mathbb{P}(\tilde{x} \leq t) > \alpha\}.$$

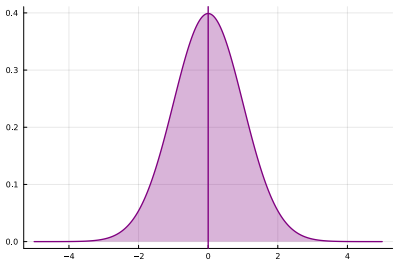
Value at Risk can be interpreted as the maximum loss that will not be exceeded with probability α. In other words it is a *guarantee* that in the worst α percent of cases the loss will not exceed $\text{VaR}_\alpha(\tilde{x})$.

Definition 6 (Conditional Value-at-Risk — CVaR_α)**.** *For $\alpha \in [0, 1]$, the Conditional-Value-at-Risk of a random variable $\tilde{x}$ with pdf p at risk-level α is defined as*

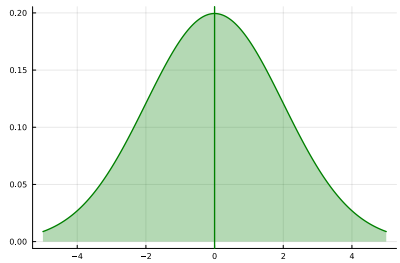
$$\text{CVaR}_\alpha(\tilde{x}) = \sup_{t \in \mathbb{R}} \left\{ t - \frac{1}{\alpha}\mathbb{E}[t - \tilde{x}]_+ \right\} = \inf_{\xi \in \Delta^n} \left\{ \xi^\mathsf{T} x \mid \alpha \cdot \xi \leq p \right\}.$$

Conditional Value at Risk can be interpreted as the expected loss given that the loss exceeds the α percentile. In other words it is a *guarantee* that in the worst α percent of cases the expected loss will not exceed $\text{CVaR}_\alpha(\tilde{x})$. This is a subtle difference from Value at Risk, but an important one to make because CVaR is convex, while VaR is not.

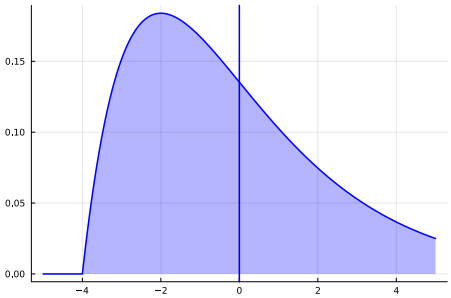
From the definitions of both VaR and CVaR we can see that $\text{VaR}_1[\tilde{x}] = \infty$ and $\text{CVaR}_1[\tilde{x}] = \mathbb{E}[\tilde{x}]$. $\text{CVaR}_0[\tilde{x}] = \operatorname{ess\,inf}[\tilde{x}]$ where $\operatorname{ess\,inf}[\tilde{x}]$ is the essential infimum of $\tilde{x}$ with pdf p (smallest value with nonzero probability).



(a) $\mathcal{N}(0, 1)$



(b) $\mathcal{N}(0, 2)$



(c) $\Gamma(2, 2) - 4$

Figure 1: Probability density functions (PDFs) of various distributions. All with a shared mean of 0 denoted by the vertical line.

The simplest, and most accessible coherent risk measure is the expectation operator $\mathbb{E}$, see definition 4. Expectation often ignores tail risk, and reduces a random variable to its mean as seen in Figure 1. Expectation also cannot easily be parameterized to capture the risk preferences of the agent. As such other risk measures have been developed to capture the risk preferences of the agent, such as VaR and CVaR.

Our introduction of the concept of risk is largely inspired by the influential work of (Föllmer and Schied, 2016). We refer the reader there for a more delicate and detailed treatment of risk measures and their subsequent properties.

4 Vanilla Reinforcement Learning

In order to make decisions one must adopt a policy which prescribes the agent's behavior. A policy is a mapping from states to actions, and the goal of reinforcement learning is to find an optimal policy that maximizes the expected return. Policies may depend on time, and the agent may have a horizon of T time steps to act. In this paper we assume a finite horizon, with each $s \in S$ having a time component, thus we omit the subscript t on all variables for simplicity.

Definition 7 (Finite Horizon Expected Discounted Return)**.** *Given a policy $\pi : \mathcal{S} \to \mathcal{A}$ the expected finite horizon discounted return $\rho(\pi)$ can be defined as*

$$\rho(\pi) = \mathbb{E}_{\pi,\mathcal{P}}\left[\sum_{t=0}^{T} \gamma^t r(\tilde{s}_t, \pi(\tilde{s}_t))\right].$$

Where $\tilde{s}_0 \sim p_0$ and $\tilde{s}_{t+1} \sim \mathcal{P}(s_t, \pi(s_t))$, γ is the discount factor, and T is the time horizon.

Definition 8 (Objective of Reinforcement Learning)**.** *The objective of reinforcement learning (RL) is as follows*

$$\max_{\pi \in \Pi} \rho(\pi).$$

Definition 9 (Optimal Policy)**.** *An optimal policy $\pi^\star \in \arg\max_{\pi \in \Pi} \rho(\pi)$.*

Definition 10 (Optimal Value Function)**.** *Given an optimal policy $\pi^\star$, the optimal value function starting at time $l \in [0 \dots T-1]$ $V_l^\star = V_l^{\pi^\star} : \mathcal{S} \to \mathbb{R}]$ is defined as*

$$V_l^\star(s) = r(s, \pi(s)) + \mathbb{E}_{\mathcal{P}}\left[\sum_{t=l+1}^{T} \gamma^t r(\tilde{s}_t, \pi^\star(\tilde{s}_t))\right].$$

Where $\tilde{s}_{l+1} \sim \mathcal{P}(s, \pi(s))$ and

$$V_T^\star(s) = r(s).$$

Notice in 9 that the optimal policy may not be unique, as there may be multiple policies that achieve the same expected return. Notice also that our definition of the value function presupposes the existence of an optimal policy and requires that policy in order to derive the value function. For a general MDP this need not be the case, and we refer the reader to (Puterman, 1994) for a description on the deep and rich relationship between value functions and policies which we do not embark on here.

4.1 Dynamic Programming

A common approach to solving the objective of reinforcement learning is to use dynamic programming. Dynamic programming is a method for solving complex problems by breaking them down into simpler sub-problems. In the context of reinforcement learning, dynamic programming is used to compute the optimal value function and policy.

Definition 11 (Bellman Equation)**.** *Given a policy $\pi : \mathcal{S} \to \mathcal{A}$, the Bellman equation for the value function is defined as*

$$V_l^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(s,\pi(s))}[V_{l+1}^\pi(s')].$$

The Bellman equation is a recursive equation that decomposes the value function into two parts: the immediate reward and the value of the next state. While we consider deterministic policies in our definition here, the Bellman equation can be extended to randomized policies as well through the use of an expectation operator.

Definition 12 (Bellman Optimality Equation)**.** *Given a policy $\pi : \mathcal{S} \to \mathcal{A}$, the Bellman optimality equation for the value function is defined as*

$$V_l^\star(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(s,a)}[V_{l+1}^\star(s')] \right\}.$$

Proposition 2. *There exists an optimal policy $\pi^\star$ such that $V_l^\star(s) = V_l^{\pi^\star}(s)$ for all $s \in \mathcal{S}$ and for all $l \in [1 \dots T]$.*

While Proposition 2 seems obvious, and indeed the proof is trivial by contradiction. We include it here to illustrate a key idea which allows our goal of finding an optimal policy that maximizes the expected return by dynamic programming.

We urge the reader to consider the work of (Sutton and Barto, 2018) for a more detailed introduction to reinforcement learning beyond dynamic programming. For a more detailed account of the theoretical results glossed over in this summary, please see (Puterman, 1994).

5 Risk-Averse Reinforcement Learning

Consider the return function $\rho(\pi)$, which is of key importance in reinforcement learning 8. The inner expectation of $\rho(\pi)$ is a risk measure, as such one can imagine replacing the expectation with a different risk measure ψ.

However, expectation is already a coherent risk measure, so the question arises, why would one want to replace it with a different risk measure? We refer the reader to the discussion in Section 3 of this work and to the work of (Bäuerle and Ott, 2011; Föllmer and Schied, 2016; Howard and Matheson, 1972) for further motivation.

In contrast to Section 4 we now consider the set of all policies to be history dependent randomized policies instead of stationary deterministic policies. That is to say, from now on

$$\Pi = \Pi_{\text{HR}} = \left\{ \pi : [\mathcal{S} \times \mathcal{A}]_0^t \to \Delta^{\mathcal{A}} \right\}.$$

This group of policies is more general than the set of deterministic policies, and allows for a more flexible representation of the agent's behavior. For the particular reader who is interested in why we only considered stationary deterministic policies in Section 4 we refer them to the work of (Puterman, 1994) (Proposition 4.4.3).

There has been prior work on embedding a utility function within the expectation, perhaps one encoding the preferences of an agent who desires higher returns less and less (Howard and Matheson, 1972). This approach while natural, requires the utility function to be known a priori, instead we take a different path and only require a user to specify their risk level $\alpha \in [0, 1]$. This risk level has different understandings between risk measures, however it generally measures how risky an agent is willing to be with 1 being completely risk seeking, and 0 being completely risk averse.

Definition 13 (Risk-Averse Reinforcement Learning)**.** *Given a risk measure ψ, the objective of risk-averse reinforcement learning is as follows*

$$\max_{\pi \in \Pi} \hat{\rho}(\pi) = \max_{\pi \in \Pi} \psi_{\mathcal{P}} \left(\sum_{t=1}^{T} \gamma^t r(\tilde{s}_t, \pi(\tilde{h}_t)) \right).$$

Where $\tilde{s}_1 \sim p_0$ and $\tilde{s}_{t+1} \sim \mathcal{P}(s_t, \pi(\tilde{h}_t))$, $\tilde{h}_t = (\tilde{s}_1, \tilde{a}_1, \dots, \tilde{s}_{t-1}, \tilde{a}_{t-1})$ and $\pi(\tilde{h}_t) \in \Delta^{\mathcal{A}}$.

Definition 14 (Optimal Risk-Averse Policy)**.** *An optimal risk-averse policy $\pi^\star \in \arg\max_{\pi \in \Pi} \hat{\rho}(\pi)$.*

Here however we depart from the standard procedure of defining the value function recursively as would be required for dynamic programming and as was done in Section 4. The reason for this is that the risk measure ψ does not usually satisfy the tower property, which is a key property for the Bellman equation to hold.

An open research question is how to decompose different risk measures as in Section 3 into a recursive form that would allow for dynamic programming to be used for policy optimization. This topic is a difficult one, and we refer the reader to the work of (Hau et al., 2023; Bäuerle and Ott, 2011; Föllmer and Schied, 2016) for a more detailed treatment of this topic.

A goal of my research is to develop a dynamic program for the CVaR (Definition 6) risk measure. This is a difficult task, as CVaR does not satisfy the tower property, and as such cannot easily be decomposed into a recursive form that would allow for dynamic programming to be used for policy optimization. Recently techniques have been developed for VaR (Definition 5) (Hau et al., 2023), and we hope that similar ideas can be extended to CVaR.

References

N. Bäuerle and J. Ott. Markov Decision Processes with Average-Value-at-Risk criteria. *Mathematical Methods of Operations Research*, 74(3):361–379, 2011.

H. Föllmer and A. Schied. *Stochastic Finance: An Introduction in Discrete Time*. De Gruyter, 2016.

K. M. Hall, H. J. Albers, M. Alkaee Taleghan, and T. G. Dietterich. Optimal spatial-dynamic management of stochastic species invasions. *Environmental and Resource Economics*, 70(2), 2018.

J. L. Hau, E. Delage, M. Ghavamzadeh, and M. Petrik. On dynamic programming decompositions of static risk measures in markov decision processes. In *Neural Information Processing Systems*, 2023. URL https://api.semanticscholar.org/CorpusID:259064374.

R. A. Howard and J. E. Matheson. Risk-Sensitive Markov Decision Processes. *Management Science*, 18(7):356–369, 1972.

S. V. Mehta, R. G. Haight, F. R. Homans, S. Polasky, and R. C. Venette. Optimal detection and control strategies for invasive species management. *Ecological Economics*, 61(2):237–245, 2007.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley; Sons, Inc., 1st edition, 1994.

R. S. Sutton and A. G. Barto. *Reinforcement Learning, second edition: An Introduction*. MIT Press, 2018. ISBN 9780262352703.