# Risk-averse Total-reward MDPs with ERM and EVaR

**Xihong Su**[1], **Marek Petrik**[1], **Julien Grand-Clément**[2]

[1]University of New Hampshire, 33 Academic Way, Durham, NH, 03824 USA
[2] HEC Paris, 1 Rue de la Libération, Jouy-en-Josas, 78350 France
xihong.su@unh.edu, mpetrik@cs.unh.edu, grand-clement@hec.fr

**Abstract**

Optimizing risk-averse objectives in discounted MDPs is challenging because most models do not admit direct dynamic programming equations and require complex history-dependent policies. In this paper, we show that the risk-averse *total reward criterion*, under the Entropic Risk Measure (ERM) and Entropic Value at Risk (EVaR) risk measures, can be optimized by a stationary policy, making it simple to analyze, interpret, and deploy. We propose exponential value iteration, policy iteration, and linear programming to compute optimal policies. Compared with prior work, our results only require the relatively mild condition of transient MDPs and allow for *both* positive and negative rewards. Our results indicate that the total reward criterion may be preferable to the discounted criterion in a broad range of risk-averse reinforcement learning domains.

## 1 Introduction

Risk-averse Markov decision processes (MDP) (Puterman 2005) that use *monetary risk measures* as their objective have been gaining in popularity in recent years (Kastner, Erdogdu, and Farahmand 2023; Marthe, Garivier, and Vernade 2023; Lam et al. 2022; Li, Zhong, and Brandeau 2022; Bäuerle and Glauner 2022; Hau, Petrik, and Ghavamzadeh 2023; Hau et al. 2023; Su, Petrik, and Grand-Clément 2024a,b). Risk-averse objectives, such as Value at Risk (VaR), Conditional Value at Risk (CVaR), Entropic Risk Measure (ERM), or Entropic Value at Risk (EVaR), penalize the variability of returns (Follmer and Schied 2016). As a result, these risk measures yield policies with stronger guarantees on the probability of catastrophic losses, which is important in domains like healthcare or finance.

In this paper, we target the *total reward criterion* (TRC) (Kallenberg 2021; Puterman 2005) instead of the common discounted criterion. TRC also assumes an infinite horizon but does not discount future rewards. To control for infinite returns, we assume that the MDP is *transient*, i.e. that there is a positive probability that the process terminates after a finite number of steps, an assumption commonly used in the TRC literature (Filar and Vrieze 2012). We consider the TRC with both positive and negative rewards. When the rewards are non-positive, the TRC is equivalent to the *stochastic shortest path* problem, and when they are non-negative, it is equivalent to the *stochastic longest path* (Dann, Wei, and Zimmert 2023).

Two reasons motivate our departure from discounted objectives in risk-averse MDPs. First, considering risk affects discounted objectives significantly. It is common to use discounted objectives because they admit optimal stationary policies and value functions that can be computed using dynamic programs. However, most risk-averse discount objectives, such as VaR, CVaR, or EVaR, require that optimal policies are *history-dependent* (Bäuerle and Ott 2011; Hau et al. 2023; Hau, Petrik, and Ghavamzadeh 2023) and do not admit standard dynamic programming optimality equations.

Second, TRC captures the concept of stochastic termination, which is common in reinforcement learning (Sutton and Barto 2018). In risk-neutral objectives, discounting can serve well to model the probability of termination because it guarantees the same optimal policies (Puterman 2005; Su and Petrik 2023). However, as we show in this work, no such correspondence exists with risk-averse objectives, and the difference between them may be arbitrarily significant. Modeling stochastic termination using a discount factor in *risk-averse* objectives is inappropriate and leads to dramatically different optimal policies.

As our main contribution, we show that the risk-averse TRC with ERM and EVaR risk measures admit optimal stationary policies and optimal value functions in transient MDPs. We also show that the optimal value function satisfies dynamic programming equations and can be computed with exponential value iteration, policy iteration, or linear programming algorithms. These algorithms are simple and closely resemble the algorithms for solving MDPs.

Our results indicate that EVaR is a particularly interesting risk measure in reinforcement learning. ERM and the closely related exponential utility functions have been popular in sequential decision-making problems because they admit dynamic programming decompositions (Patek and Bertsekas 1999; de Freitas, Freire, and Delgado 2020; Smith and Chapman 2023; Denardo and Rothblum 1979; Hau, Petrik, and Ghavamzadeh 2023; Hau et al. 2023). Unfortunately, ERM is difficult to interpret; it is scale-dependent; and it is incomparable with popular risk measures like VaR and CVaR. Because EVaR reduces to an optimization over ERM, it preserves most of the computational advantages of ERM, and since

| | Risk properties | | Optimal policy | |
|---|---|---|---|---|
| Risk measure | Coherent | Law inv. | Disc. | TRC |
| $\mathbb{E}$ | yes | yes | S | S |
| EVaR | yes | yes | M | **S** |
| ERM | no | yes | M | **S** |
| NCVaR | yes | no | S | S |
| VaR | yes | yes | H | H |
| CVaR | yes | yes | H | H |

Table 1: Structure of optimal policies in risk-averse MDPs: "S", "M" and "H" refer to Stationary, Markov and History-dependent policies respectively.

EVaR closely approximates CVaR and VaR at the same risk level, its value is also much easier to interpret. Finally, EVaR is also a coherent risk measure, unlike ERM (Ahmadi-Javid 2012; Ahmadi-Javid and Pichler 2017).

Table 1 puts our contribution in the context of other work on risk-averse MDP objectives. Optimal policies for VaR and CVaR are known to be history-dependent in the discounted objective (Bäuerle and Ott 2011; Hau et al. 2023) and must be history-dependent in TRC because TRC generalizes the finite-horizon objective. The TRC with Nested risk measures, such as Nested CVaR (NCVaR), applies the risk measure in each level of the dynamic program independently and preserves most of the favorable computational properties of risk-neutral MDPs (Ahmadi et al. 2021a). Unfortunately, nested risk measures are difficult to interpret; their value depends on the sequence in which the rewards are obtained in a complex and unpredictable way (Kupper and Schachermayer 2006) and may be unbounded even if MDPs are transient.

While we are unaware of prior work on the TRC objective with ERM or EVaR risk-aversion *allowing both positive and negative rewards*, the ERM risk measure is closely related to exponential utility functions. Prior work on TRC with exponential utility functions also imposes constraints on the sign of the instantaneous rewards, such as all positive rewards (Blackwell 1967) or all negative rewards (Bertsekas and Tsitsiklis 1991; Freire and Delgado 2016; Carpin, Chow, and Pavone 2016; de Freitas, Freire, and Delgado 2020; Fei et al. 2021; Fei, Yang, and Wang 2021; Ahmadi et al. 2021a; Cohen et al. 2021; Meggendorfer 2022). Disallowing a mix of positive and negative rewards limits the modeling power of prior work because it requires that either all states are more desirable or all states are less desirable than the terminal state. Allowing rewards with mixed signs raises some technical challenges, which we address by employing a squeeze argument that takes advantage of MDP's transience.

**Notation.** We use a tilde to mark random variables, e.g. $\tilde{x}$. Bold lower-case letters represent vectors, and upper-case bold letters represent matrices. Sets are either calligraphic or upper-case Greek letters. The symbol $\mathbb{X}$ represents the space of real-valued random variables. When a function is defined over an index set, such as $z\colon \{1, 2, \dots, N\} \to \mathbb{R}$, we also treat it interchangeably as a vector $\boldsymbol{z} \in \mathbb{R}^n$ such that $z_i = z(i), \forall i = 1, \dots, n$. Finally, $\mathbb{R}, \mathbb{R}_+, \mathbb{R}_{++}$ denote real, non-negative real, and positive real numbers, respectively. $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. Given a finite set $\mathcal{Y}$, the probability simplex is $\Delta_{\mathcal{Y}} := \{x \in \mathbb{R}_+^{\mathcal{Y}} \mid \mathbf{1}^{\mathrm{T}} x = 1\}$.

## 2 Background on Risk-averse MDPs

**Markov Decision Processes** We focus on solving Markov decision processes (MDPs) (Puterman 2005), modeled by a tuple $(\bar{\mathcal{S}}, \mathcal{A}, \bar{p}, \bar{r}, \bar{\boldsymbol{\mu}})$, where $\bar{\mathcal{S}} = \{1, 2, \dots, S, S+1\}$ is the finite set of states and $\mathcal{A} = \{1, 2, \dots, A\}$ is the finite set of actions. The transition function $\bar{p}\colon \bar{\mathcal{S}} \times \mathcal{A} \to \Delta_{\bar{\mathcal{S}}}$ represents the probability $\bar{p}(s, a, s')$ of transitioning to $s' \in \bar{\mathcal{S}}$ after taking $a \in \mathcal{A}$ in $s \in \bar{\mathcal{S}}$ and $\bar{\boldsymbol{p}}_{sa} \in \Delta_{\bar{\mathcal{S}}}$ is such that $(\bar{\boldsymbol{p}}_{sa})_{s'} = \bar{p}(s, a, s')$. The function $\bar{r}\colon \bar{\mathcal{S}} \times \mathcal{A} \times \bar{\mathcal{S}} \to \mathbb{R}$ represents the reward $\bar{r}(s, a, s') \in \mathbb{R}$ associated with transitioning from $s \in \bar{\mathcal{S}}$ and $a \in \mathcal{A}$ to $s' \in \bar{\mathcal{S}}$. The vector $\bar{\boldsymbol{\mu}} \in \Delta_{\bar{\mathcal{S}}}$ is the initial state distribution.

We designate the state $e := S + 1$ as a *sink state* and use $\mathcal{S} = \{1, \dots, S\}$ to denote the set of all non-sink states. The sink state $e$ must satisfy that $\bar{p}(e, a, e) = 1$ and $\bar{r}(e, a, e) = 0$ for each $a \in \mathcal{A}$, and $\bar{\mu}_e = 0$. Throughout the paper, we use a bar to indicate whether the quantity involves the sink state $e$. Note that the sink state can indicate a goal when all rewards are negative and an undesirable terminal state when all rewards are positive.

The following technical assumption is needed to simplify the derivation. To lift the assumption, one needs to carefully account for infinite values, which adds complexity to the results and distracts from the main ideas.

**Assumption 2.1.** The initial distribution $\boldsymbol{\mu}$ satisfies that

$$\boldsymbol{\mu} > \mathbf{0}.$$

The solution to an MDP is a *policy*. Given a horizon $t \in \mathbb{N}$, a history-dependent policy in the set $\Pi_{\mathrm{HR}}^t$ maps the history of states and actions to a distribution over actions. A *Markov policy* $\pi \in \Pi_{\mathrm{MR}}^t$ is a sequence of decision rules $\pi = (\boldsymbol{d}_0, \boldsymbol{d}_1, \dots, \boldsymbol{d}_{t-1})$ with $\boldsymbol{d}_k\colon \mathcal{S} \to \Delta_{\mathcal{A}}$ the decision rule for taking actions at time $k$. The set of all *randomized decision rules* is $\mathcal{D} = (\Delta_{\mathcal{A}})^{\mathcal{S}}$. *Stationary policies* $\Pi_{\mathrm{SR}}$ are Markov policies with $\pi := (\boldsymbol{d})_\infty := (\boldsymbol{d}, \boldsymbol{d}, \dots)$ with the identical decision rule in every timestep. We treat decision rules and stationary policies interchangeably. The sets of *deterministic* Markov and stationary policies are denoted by $\Pi_{\mathrm{MD}}^t$ and $\Pi_{\mathrm{SD}}$. Finally, we omit the superscript $t$ to indicate infinite horizon definitions of policies.

The risk-neutral Total Reward Criterion (TRC) objective is:

$$\sup_{\pi \in \Pi_{\mathrm{HR}}} \liminf_{t \to \infty} \mathbb{E}^{\pi, \boldsymbol{\mu}} \left[ \sum_{k=0}^{t-1} r(\tilde{s}_k, \tilde{a}_k, \tilde{s}_{k+1}) \right], \qquad (1)$$

where the random variables are denoted by a tilde and $\tilde{s}_k$ and $\tilde{a}_k$ represent the state from $\bar{\mathcal{S}}$ and action at time $k$. The superscript $\pi$ denotes the policy that governs the actions $\tilde{a}_k$ when visiting $\tilde{s}_k$ and $\boldsymbol{\mu}$ denotes the initial distribution. Finally, note that $\liminf$ gives a conservative estimate of a policy's return since the limit does not necessarily exist for non-stationary policies.

Unlike the discounted criterion, the risk-neutral TRC may be unbounded, optimal policies may not exist, or may be non-stationary (Bertsekas and Yu 2013; James and Collins 2006).

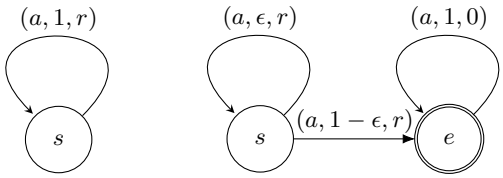


Figure 1: left: a discounted MDP, right: a transient MDP

To circumvent these issues, we assume that all policies have a positive probability of eventually transitioning to the sink state.

**Assumption 2.2.** The MDP is *transient* for any $\pi \in \Pi_{\text{SD}}$:

$$\sum_{t=0}^{\infty} \mathbb{P}^{\pi,s}\left[\tilde{s}_t = s'\right] \;<\; \infty, \qquad \forall s, s' \in \mathcal{S}. \tag{2}$$

Assumption 2.2 underlies most of our results. Transient MDPs are important because their optimal policies exist and can be chosen to be stationary deterministic (Kallenberg 2021, theorem 4.12). Transient MDPs are also common in stochastic games (Filar and Vrieze 2012) and generalize the stochastic shortest path problem (Bertsekas and Yu 2013).

An important tool in their analysis is the *spectral radius* $\rho\colon \mathbb{R}^{n\times n} \to \mathbb{R}$ which is defined for each $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ as the maximum absolute eigenvalue: $\rho(\boldsymbol{A}) := \max_{i=1,\dots,n} |\lambda_i|$ where $\lambda_i$ is the $i$-th eigenvalue (Horn and Johnson 2013).

**Lemma 2.3** (Theorem 4.8 in Kallenberg (2021))**.** *An MDP is transient if and only if $\rho(\boldsymbol{P}^\pi) < 1$ for all $\pi \in \Pi_{\text{SR}}$.*

Now, let us understand the differences between a discounted MDP and a transient MDP, which are useful in demonstrating the behavior of risk-averse objectives. Consider the MDPs in Figure 1. There is one non-sink state $s$ and one action $a$. A triple tuple represents an action, transition probability, and a reward separately. Note that every discounted MDP can be converted to a transient MDP as described in Su, Grand-Clément, and Petrik (2024, appendix B). For the discounted MDP, the discount factor is $\gamma$. For the transient MDP, $e$ is the sink state, and there is a positive probability $1-\epsilon$ of transiting from state $s$ to state $e$. Once the agent reaches the state $e$, it stays in $e$. For the risk-neutral objective, if $\gamma$ equals $\epsilon$, their value functions have identical values. However, for risk-aversion objectives, such as ERM, we show that the value functions in a discounted MDP can diverge from those in a transient MDP in Section 5.

**Monetary risk measures** Monetary risk measures aim to generalize the expectation operator to account for the spread of the random variable. *Entropic risk measure* (ERM) is a popular risk measure, defined for any risk level $\beta > 0$ and $\tilde{x} \in \mathbb{X}$ as (Follmer and Schied 2016)

$$\text{ERM}_\beta\left[\tilde{x}\right] \;=\; -\beta^{-1} \cdot \log \mathbb{E} \exp\left(-\beta \cdot \tilde{x}\right). \tag{3}$$

and extended to $\beta \in [0,\infty]$ as $\text{ERM}_0[\tilde{x}] = \lim_{\beta\to 0^+} \text{ERM}_\beta\left[\tilde{x}\right] = \mathbb{E}[\tilde{x}]$ and $\text{ERM}_\infty[\tilde{x}] = \lim_{\beta\to\infty} \text{ERM}_\beta\left[\tilde{x}\right] = \operatorname{ess\,inf}[\tilde{x}]$. ERM plays a unique role in sequential decision-making because it is the only law-invariant risk measure that satisfies the tower property (e.g., Su, Grand-Clément, and Petrik (2024, proposition A.1)), which is essential in constructing dynamic programs (Hau, Petrik, and Ghavamzadeh 2023). Unfortunately, two significant limitations of ERM hinder its practical applications. First, ERM is not positively homogenous and, therefore, the risk value depends on the scale of the rewards, and ERM is not coherent (Follmer and Schied 2016; Hau, Petrik, and Ghavamzadeh 2023; Ahmadi-Javid 2012). Second, the risk parameter $\beta$ is challenging to interpret and does not relate well to other standard risk measures, like VaR or CVaR.

For these reasons, we focus on the *Entropic Value at Risk* (EVaR), defined as, for a given $\alpha \in (0,1)$,

$$\begin{aligned} \text{EVaR}_\alpha\left[\tilde{x}\right] \;&=\; \sup_{\beta>0} -\beta^{-1} \log\left(\alpha^{-1}\mathbb{E}\exp\left(-\beta\tilde{x}\right)\right) \\ &=\; \sup_{\beta>0} \text{ERM}_\beta\left[\tilde{x}\right] + \beta^{-1}\log\alpha, \end{aligned} \tag{4}$$

and extended to $\text{EVaR}_0\left[\tilde{x}\right] = \operatorname{ess\,inf}[\tilde{x}]$ and $\text{EVaR}_1\left[\tilde{x}\right] = \mathbb{E}\left[\tilde{x}\right]$ (Ahmadi-Javid 2012). It is important to note that the supremum in (4) may not be attained even when $\tilde{x}$ is a finite discrete random variable (Ahmadi-Javid and Pichler 2017).

EVaR addresses the limitations of ERM while preserving its benefits. EVaR is coherent and positively homogenous. EVaR is also a good approximation to interpretable quantile-based risk measures, like VaR and CVaR (Ahmadi-Javid 2012; Hau, Petrik, and Ghavamzadeh 2023).

**Risk-averse MDPs.** Risk-averse MDPs, using static VaR and CVaR risk measures, under the discounted criterion received abundant attention (Hau et al. 2023; Bäuerle and Ott 2011; Bäuerle and Glauner 2022; Pflug and Pichler 2016; Li, Zhong, and Brandeau 2022), showing that these objectives require history-dependent optimal policies. In contrast, nested risk measures under the TRC may admit stationary policies that can be computed using dynamic programming (Ahmadi et al. 2021a; Meggendorfer 2022; de Freitas, Freire, and Delgado 2020; Gavriel, Hanasusanto, and Kuhn 2012). However, the TRC with nested CVaR can be unbounded (Su, Grand-Clément, and Petrik 2024, proposition C.1). Recent work has shown that optimal Markov policies exist for EVaR discounted objectives, and they can be computed via dynamic programming (Hau, Petrik, and Ghavamzadeh 2023), building upon similar results established for ERM (Chung and Sobel 1987). However, in TRC with ERM, the value functions may also be unbounded (Su, Grand-Clément, and Petrik 2024, proposition D.1).

## 3 Solving ERM Total Reward Criterion

This section shows that an optimal stationary policy exists for ERM-TRC and that the value function satisfies dynamic programming equations. We then outline algorithms for computing it.

Our objective in this section is to maximize the ERM-TRC objective for some given $\beta > 0$ defined as

$$\sup_{\pi\in\Pi_{\text{HR}}} \liminf_{t\to\infty} \text{ERM}_\beta^{\pi,\boldsymbol{\mu}}\left[\sum_{k=0}^{t-1} r(\tilde{s}_k, \tilde{a}_k, \tilde{s}_{k+1})\right]. \tag{5}$$

The definition employs limit inferior because the limit may not exist for non-stationary policies. Return functions $g_t \colon \Pi_{\mathrm{HR}} \times \mathbb{R}_{++} \to \mathbb{R}$ and $g_t^\star \colon \mathbb{R}_{++} \to \mathbb{R}$ for a horizon $t \in \mathbb{N}$ and the infinite-horizon versions $g_t \colon \Pi_{\mathrm{HR}} \times \mathbb{R}_{++} \to \bar{\mathbb{R}}$ and $g_t^\star \colon \mathbb{R}_{++} \to \bar{\mathbb{R}}$ are defined

$$\begin{aligned} g_t(\pi, \beta) &:= \mathrm{ERM}_\beta^{\pi, \boldsymbol{\mu}} \left[ \sum_{k=0}^{t-1} r(\tilde{s}_k, \tilde{a}_k, \tilde{s}_{k+1}) \right], \\ g_t^\star(\beta) &:= \sup_{\pi \in \Pi_{\mathrm{HR}}} g_t(\pi, \beta), \\ g_\infty(\pi, \beta) &:= \liminf_{t \to \infty} g_t(\pi, \beta), \\ g_\infty^\star(\beta) &:= \liminf_{t \to \infty} g_t^\star(\beta). \end{aligned} \tag{6}$$

Note that the functions $g_\infty$ and $g_\infty^\star$ can return infinite values and that (5) differs from $g_\infty^\star$ in the order of the limit and the supremum. Finally, when $\beta = 0$, we assume that all $g$ functions are defined as the expectation. In the remainder of the section, we assume that the risk level $\beta > 0$ is fixed and omit it in notations when its value is unambiguous from the context.

## 3.1 Finite Horizon

We commence the analysis with definitions and basic properties for the finite horizon criterion. To the best of our knowledge, this analysis is original in the context of the ERM but builds on similar approaches employed in the study of exponential utility functions.

Finite-horizon functions $\boldsymbol{v}^t(\pi) \in \mathbb{R}^S$ and $\boldsymbol{v}^{t,\star} \in \mathbb{R}^S$ are defined for each horizon $t \in \mathbb{N}$ and policy $\pi \in \Pi_{\mathrm{MD}}$, $s \in \mathcal{S}$ as

$$\begin{aligned} v_s^t(\pi) &:= \mathrm{ERM}_\beta^{\pi, s} \left[ \sum_{k=0}^{t-1} r(\tilde{s}_k, \tilde{a}_k, \tilde{s}_{k+1}) \right], \\ v_s^{t,\star} &:= \max_{\pi \in \Pi_{\mathrm{MD}}} v_s^t(\pi), \end{aligned} \tag{7}$$

and $v_e^t(\pi) := 0$.

Because the nonlinearity of ERM complicates the analysis, it will be convenient to instead rely on *exponential value function* $\boldsymbol{w}^t(\pi) \in \mathbb{R}^S$ for $\pi \in \Pi_{\mathrm{MD}}$, $t \in \mathbb{N}$, and $s \in \mathcal{S}$ that satisfy

$$w_s^t(\pi) := -\exp\left(-\beta \cdot v_s^t(\pi)\right), \tag{8}$$

$$v_s^t(\pi) = -\beta^{-1} \log(-w_s^t(\pi)). \tag{9}$$

The optimal $\boldsymbol{w}^{t,\star} \in \mathbb{R}^S$ is defined analogously from $\boldsymbol{v}^{t,\star}$. Note that $\boldsymbol{w}^t < \mathbf{0}$ (componentwise) and $\boldsymbol{w}^0(\pi) = \boldsymbol{w}^{0,\star} = -\mathbf{1}$ for any $\pi \in \Pi_{\mathrm{MD}}$. Similar exponential value functions have been used previously in exponential utility function objectives (Denardo and Rothblum 1979; Patek 2001), in the analysis of robust MDPs, and even in regularized MDPs (see Grand-Clément and Petrik (2022) and references therein).

One can define a corresponding *exponential Bellman operator* for any $\boldsymbol{w} \in \mathbb{R}^S$ as

$$\begin{aligned} L^{\boldsymbol{d}} \boldsymbol{w} &:= \boldsymbol{B}^{\boldsymbol{d}} \boldsymbol{w} - \boldsymbol{b}^{\boldsymbol{d}}, \\ L^\star \boldsymbol{w} &:= \max_{\boldsymbol{d} \in \mathcal{D}} L^{\boldsymbol{d}} \boldsymbol{w} = \max_{\boldsymbol{d} \in \operatorname{ext} \mathcal{D}} L^{\boldsymbol{d}} \boldsymbol{w}, \end{aligned} \tag{10}$$

where $\operatorname{ext} \mathcal{D}$ is the set of extreme points of $\mathcal{D}$ corresponding to deterministic decision rules and $\boldsymbol{B}^{\boldsymbol{d}} \in \mathbb{R}_+^{S \times S}$ and $\boldsymbol{b}^{\boldsymbol{d}} \in \mathbb{R}_+^S$ are defined for $s, s' \in \mathcal{S}$ and $\boldsymbol{d} \in \mathcal{D}$ as

$$B_{s,s'}^{\boldsymbol{d}} := \sum_{a \in \mathcal{A}} p(s, a, s') \cdot d_a(s) \cdot e^{-\beta \cdot r(s,a,s')}, \tag{11a}$$

$$b_s^{\boldsymbol{d}} := \sum_{a \in \mathcal{A}} p(s, a, e) \cdot d_a(s) \cdot e^{-\beta \cdot r(s,a,e)}. \tag{11b}$$

The following theorem shows that $L$ can be used to compute $\boldsymbol{w}$. We use the shorthand notation $\pi_{1:t-1} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_{t-1}) \in \Pi_{\mathrm{MR}}^{t-1}$ to denote the tail of $\pi$ that starts with $\boldsymbol{d}_1$ instead of $\boldsymbol{d}_0$.

**Theorem 3.1.** *For each $t = 1, \ldots,$ and $\pi = (\boldsymbol{d}_0, \ldots, \boldsymbol{d}_{t-1}) \in \Pi_{\mathrm{MR}}^t$, the exponential values satisfy that*

$$\begin{aligned} \boldsymbol{w}^t(\pi) &= L^{\boldsymbol{d}_t} \boldsymbol{w}^{t-1}(\pi_{1:t-1}), & \boldsymbol{w}^0(\pi) &= -\mathbf{1}, \\ \boldsymbol{w}^{t,\star} &= L^\star \boldsymbol{w}^{t-1,\star} = \boldsymbol{w}^t(\pi^\star) \geq \boldsymbol{w}^t(\pi), & \boldsymbol{w}^{0,\star} &= -\mathbf{1}, \end{aligned}$$

*for some $\pi^\star \in \Pi_{\mathrm{MD}}^t$.*

The proof of Theorem 3.1 is standard and has been established both in the context of ERMs (Hau, Petrik, and Ghavamzadeh 2023) and utility functions (Patek 1997).

The following corollary follows directly from Theorem 3.1 by algebraic manipulation and by the monotonicity of exponential value function transformation and the ERM.

**Corollary 3.2.** *We have that*

$$\begin{aligned} g_t(\pi, \beta) &= \mathrm{ERM}_\beta^{\boldsymbol{\mu}} \left[ v_{\tilde{s}_0}^t(\pi) \right], \\ g_t^\star(\beta) &= \mathrm{ERM}_\beta^{\boldsymbol{\mu}} \left[ v_{\tilde{s}_0}^{t,\star} \right] = \max_{\pi \in \Pi_{\mathrm{MD}}} \mathrm{ERM}_\beta^{\boldsymbol{\mu}} \left[ v_{\tilde{s}_0}^t(\pi) \right]. \end{aligned}$$

## 3.2 Infinite Horizon

We now turn to construct infinite-horizon optimal policies as a limiting case of the finite horizon. An important quantity is the infinite-horizon exponential value function defined for each $\pi \in \Pi_{\mathrm{HR}}$ as

$$\boldsymbol{w}^\infty(\pi) := \liminf_{t \to \infty} \boldsymbol{w}^t(\pi), \quad \boldsymbol{w}^{\infty,\star} := \liminf_{t \to \infty} \boldsymbol{w}^{t,\star}.$$

Note again that we use the inferior limit because the limit may not be defined for non-stationary policies. The limiting infinite-horizon value functions $\boldsymbol{w}^\infty(\pi)$ and $\boldsymbol{w}^{\infty,\star}$ are defined analogously from $\boldsymbol{v}^t(\pi)$ and $\boldsymbol{v}^{t,\star}$ using the inferior limit. The following theorem is the main result of this section. It shows that for an infinite horizon, the optimal exponential value function is attained by a stationary deterministic policy and is a fixed point of the exponential Bellman operator.

**Theorem 3.3.** *Whenever $\boldsymbol{w}^{\infty,\star} > -\boldsymbol{\infty}$ there exists $\pi^\star = (\boldsymbol{d}^\star)_\infty \in \Pi_{\mathrm{SD}}$ such that*

$$\boldsymbol{w}^{\infty,\star} = \boldsymbol{w}^\infty(\pi^\star) = L^{\boldsymbol{d}^\star} \boldsymbol{w}^{\infty,\star},$$

*and $\boldsymbol{w}^{\infty,\star}$ is the unique value that satisfies this equation.*

**Corollary 3.4.** *Asuming the hypothesis of Theorem 3.3, we have that $\boldsymbol{v}^{\infty,\star} = \boldsymbol{v}^\infty(\pi^\star)$ and*

$$g_\infty^\star(\beta) = \mathrm{ERM}_\beta^{\boldsymbol{\mu}} \left[ v_{\tilde{s}_0}^{\infty,\star} \right] = \max_{\pi \in \Pi_{\mathrm{SD}}} \mathrm{ERM}_\beta^{\boldsymbol{\mu}} \left[ v_{\tilde{s}_0}^\infty(\pi) \right].$$

We now outline the proof of Theorem 3.3; see Su, Grand-Clément, and Petrik (2024, appendix D) for details. To establish Theorem 3.3, we show that $\boldsymbol{w}^{t,\star}$ converges to a fixed point as $t \to \infty$. Standard arguments do not apply to our setting (Puterman 2005; Kallenberg 2021; Patek 2001) because the ERM-TRC Bellman operator is not an $L_\infty$-contraction, it is not linear, and the values in value iteration do not increase or decrease monotonically. Although the exponential Bellman operator $L^{\boldsymbol{d}}$ is linear, it may not be a contraction.

The main idea of the proof is to show that whenever the exponential value functions are bounded, the exponential Bellman operator must be *weighted-norm* contraction with a unique fixed point. To facilitate the analysis, we define $\boldsymbol{w}^t \colon \Pi^t_{\mathrm{SR}} \times \mathbb{R}^S \to \mathbb{R}^S, t \in \mathbb{N}$ for $\boldsymbol{z} \in \mathbb{R}^S$, $\pi \in \Pi^t_{\mathrm{SR}}$, as

$$\begin{aligned} \boldsymbol{w}^t(\pi, \boldsymbol{z}) \; &= L^{\boldsymbol{d}} \boldsymbol{w}(\pi_{1:t-1}) \; = \; L^{\boldsymbol{d}} L^{\boldsymbol{d}} \dots L^{\boldsymbol{d}}(-\boldsymbol{z}) \\ &= \; -(\boldsymbol{B}^{\boldsymbol{d}})^t \boldsymbol{z} - \sum_{k=0}^{t-1} (\boldsymbol{B}^{\boldsymbol{d}})^k \boldsymbol{b}^{\boldsymbol{d}}. \end{aligned} \tag{12}$$

The value $\boldsymbol{z}$ can be interpreted as the exponential value function at the termination of the process following $\pi$ for $t$ periods. Note that $\boldsymbol{w}^t(\pi) = \boldsymbol{w}^t(\pi, \mathbf{1})$, $\forall \pi \in \Pi_{\mathrm{MR}}, t \in \mathbb{N}$.

An important technical result we show is that the only way a *stationary* policy's return can be bounded is if the policy's matrix has a spectral radius strictly less than 1.

**Lemma 3.5.** *For each $\pi = (\boldsymbol{d})_\infty \in \Pi_{\mathrm{SR}}$ and $\boldsymbol{z} \geq \mathbf{0}$:*

$$\boldsymbol{w}^\infty(\pi, \boldsymbol{z}) > -\boldsymbol{\infty} \quad \Rightarrow \quad \rho(\boldsymbol{B}^{\boldsymbol{d}}) < 1.$$

Lemma 3.5 uses the transience property to show that the Perron vector (with the maximum absolute eigenvalue) $\boldsymbol{f}$ of $\boldsymbol{B}^{\boldsymbol{d}}$ satisfies that $\boldsymbol{f}^\top \boldsymbol{b}^{\boldsymbol{d}} > 0$. Therefore, $\rho(\boldsymbol{B}^{\boldsymbol{d}}) < 1$ is necessary for the series in (12) to be bounded.

The limitation of Lemma 3.5 is that it only applies to stationary policies. The lemma does not preclude the possibility that all stationary policies have unbounded returns, but a Markov policy with a bounded return exists. We construct an upper bound on $\boldsymbol{w}^{t,\star}$ that decreases monotonically in $t$ and converges to show this is impossible. The proof then concludes by squeezing $\boldsymbol{w}^{t,\star}$ between a lower and the upper bound with the same limits. This technique allows us to relax the limiting assumptions from prior work (Patek 2001; de Freitas, Freire, and Delgado 2020). Finally, our results imply an optimal stationary policy exists whenever the planning horizon $T$ is sufficiently large. Because the set $\Pi_{\mathrm{SD}}$ is finite, one policy must be optimal for a sufficiently large $T$. This property suggests behavior similar to *turnpikes* in discounted MDPs (Puterman 2005).

### 3.3 Algorithms

We now briefly describe the algorithms we use to compute the optimal ERM-TRC policies. Surprisingly, the main algorithms for discounted MDPs, including value iteration, policy iteration, and linear programming, can be adapted to this risk-averse setting with only minor modifications.

*Value iteration* is the most direct method for computing the optimal value function (Puterman 2005). The value iteration computes a sequence of $\boldsymbol{w}^k, k = 0, \dots$ such that

$$\boldsymbol{w}^{k+1} = L^\star \boldsymbol{w}^k, \quad \boldsymbol{w}^0 = \mathbf{0}.$$

The initialization of $\boldsymbol{w}^0 = \mathbf{0}$ is essential and guarantees convergence directly from the monotonicity argument used to prove Theorem 3.3.

*Policy iteration* (PI) starts by initializing with a stationary policy $\pi_0 = (\boldsymbol{d}^0)_\infty \in \Pi_{\mathrm{SD}}$. Then, for each iteration $k = 0, \dots,$ PI alternates between the policy evaluation step and the policy improvement step:

$$\boldsymbol{w}^k = -(\boldsymbol{I} - \boldsymbol{B}^{\boldsymbol{d}^k})^{-1} \boldsymbol{b}^{\boldsymbol{d}^k}, \; \boldsymbol{d}^{k+1} \in \operatorname*{argmax}_{\boldsymbol{d} \in \mathcal{D}} \boldsymbol{B}^{\boldsymbol{d}} \boldsymbol{w}^k - \boldsymbol{b}^{\boldsymbol{d}}.$$

PI converges because it monotonically improves the value functions when initialized with a policy $\boldsymbol{d}^0$ with bounded return (Patek 2001). However, we lack a practical approach to finding such an initial policy.

Finally, *linear programming* is a fast and convenient method for computing optimal exponential value functions:

$$\min \left\{ \mathbf{1}^\top \boldsymbol{w} \mid \boldsymbol{w} \in \mathbb{R}^S, \boldsymbol{w} \geq -\boldsymbol{b}^a + \boldsymbol{B}^a \boldsymbol{w}, \, \forall a \in \mathcal{A} \right\}. \tag{13}$$

Here, $\boldsymbol{B}^a_{s,\cdot} = (B^a_{s,s_1}, \cdots, B^a_{s,s_S})$, $B^a_{s,s'}$ and $b^a_s$ are constructed as in (11). We use the shorthand $\boldsymbol{B}^a = \boldsymbol{B}^{\boldsymbol{d}}$ and $\boldsymbol{b}^a = \boldsymbol{b}^{\boldsymbol{d}}$ where $d_{a'}(s) = 1$ if $a = a'$ for each $s \in \mathcal{S}, a' \in \mathcal{A}$.

It is important to note that the value functions, as well as the coefficients of $\boldsymbol{B}^{\boldsymbol{d}}$ may be irrational. It is, therefore, essential to study the sensitivity of the algorithms to errors in the input. However, this question is beyond the scope of the present paper, and we leave it for future work.

## 4 Solving EVaR Total Reward Criterion

This section shows that the EVaR-TRC objective can be reduced to a sequence of ERM-TRC problems, similarly to the discounted case (Hau, Petrik, and Ghavamzadeh 2023). As a result, an optimal stationary EVaR-TRC policy exists and can be computed using the methods described in Section 3.

Formally, we aim to compute a policy that maximizes the EVaR of the random return at some given fixed risk level $\alpha \in (0, 1)$ defined as

$$\sup_{\pi \in \Pi_{\mathrm{HR}}} \liminf_{t \to \infty} \mathrm{EVaR}^{\pi, \boldsymbol{\mu}}_\alpha \left[ \sum_{k=0}^{t-1} r(\tilde{s}_k, \tilde{a}_k, \tilde{s}_{k+1}) \right]. \tag{14}$$

In contrast with Ahmadi et al. (2021b), the objective in (14) optimizes EVaR rather than Nested EVaR.

### 4.1 Reduction to ERM-TRC

To solve (14), we exploit that EVaR can be defined in terms of ERM as shown in (4). To that end, define a function $h_t \colon \Pi_{\mathrm{HR}} \times \mathbb{R} \to \bar{\mathbb{R}}$ for $t \in \mathbb{N}$ as

$$h_t(\pi, \beta) \; := \; g_t(\pi, \beta) + \beta^{-1} \log(\alpha), \tag{15}$$

where $g_t$ is the ERM value of the policy defined in (6). Also, $h^\star_t$, $h_\infty$, $h^\star_\infty$ are defined analogously in terms of $g^\star_t$, $g_\infty$, and $g^\star_\infty$ respectively. The functions $h$ are useful, because by (4):

$$\mathrm{EVaR}^{\pi, \boldsymbol{\mu}}_\alpha \left[ \sum_{k=0}^{t-1} r(\tilde{s}_k, \tilde{a}_k, \tilde{s}_{k+1}) \right] \; = \; \sup_{\beta > 0} h_t(\pi, \beta), \tag{16}$$

for each $\pi \in \Pi_{\mathrm{HR}}$ and $t \in \mathbb{N}$. However, note that the limit in the definition of $\sup_{\beta>0} h^\star_\infty(\beta)$ is inside the supremum unlike in the objective in (14).

There are two challenges with solving (14) by reducing it to (16). First, the supremum in the definition of EVaR in (4) may not be attained, as mentioned previously. Second, the functions $g_t^\star$ and $h_t^\star$ may not converge *uniformly* to $g_\infty^\star$ and $h_\infty^\star$. Note that Theorem 3.3 only shows *pointwise* convergence when the functions are bounded.

To circumvent the challenges described above, we replace the supremum in (16) with a maximum over a *finite* set $\mathcal{B}(\beta_0, \delta)$ of discretized $\beta$ values:

$$\mathcal{B}(\beta_0, \delta) \;:=\; \{\beta_0, \beta_1, \ldots, \beta_K\}, \tag{17a}$$

where $\delta > 0$, $0 < \beta_0 < \beta_1 < \cdots < \beta_K$, and

$$\beta_{k+1} \;:=\; \frac{\beta_k \log \frac{1}{\alpha}}{\log \frac{1}{\alpha} - \beta_k \delta}, \quad \beta_K \;\geq\; \frac{\log \frac{1}{\alpha}}{\delta}, \tag{17b}$$

for an appropriately chosen value $K$ for each $\beta_0$ and $\delta$. We assume that the denominator in the expression for $\beta_{k+1}$ in Equation (17b) is positive; otherwise $\beta_{k+1} = \infty$ and $\beta_k$ is sufficiently large.

The construction in (17) resembles equations (19) and (20) in Hau, Petrik, and Ghavamzadeh (2023) but differs in the choice of $\beta_0$ because Hoeffding's lemma does not readily bound the TRC criterion.

The following proposition upper-bounds the value of $K$; see (Hau, Petrik, and Ghavamzadeh 2023, theorem 4.3) for a proof that $K$ is polynomial in $\delta$.

**Proposition 4.1.** *Assume a given $\beta_0 > 0$ and $\delta \in (0, 1)$ such that $\beta_0 \delta < \log \frac{1}{\alpha}$. Then, to satisfy the condition in* (17b)*, it is sufficient to choose $K$ as*

$$K := \frac{\log z}{\log(1 - z)}, \quad \textit{where} \quad z := \frac{\beta_0 \delta}{\log \frac{1}{\alpha}}. \tag{18}$$

The following theorem shows that one can obtain an optimal ERM policy for an appropriately chosen $\beta$ that approximates an optimal EVaR policy arbitrarily closely.

**Theorem 4.2.** *For any $\delta > 0$, let*

$$(\pi^\star, \beta^\star) \in \operatorname*{argmax}_{(\pi,\beta) \in \Pi_{\mathrm{SD}} \times \mathcal{B}(\beta_0, \delta)} h_\infty(\pi, \beta),$$

*where $\beta_0 > 0$ is chosen such that $g_\infty^\star(0) \leq g_\infty^\star(\beta_0) - \delta$. Then the limits below exist and satisfy:*

$$\begin{aligned} \lim_{t \to \infty} \mathrm{EVaR}_\alpha^{\pi^\star, \boldsymbol{\mu}} \left[ \sum_{k=0}^{t-1} r(\tilde{s}_k, \tilde{a}_k, \tilde{s}_{k+1}) \right] \\ \geq \sup_{\pi \in \Pi_{\mathrm{HR}}} \lim_{t \to \infty} \sup_{\beta > 0} h(\pi, \beta) - \delta. \end{aligned} \tag{19}$$

Note that the right-hand side in (19) is the $\delta$-optimal objective in (14).

The first implication of Theorem 4.2 is that there exists an optimal stationary deterministic policy.

**Corollary 4.3.** *There exists an optimal stationary deterministic policy $\pi^\star \in \Pi_{\mathrm{SD}}$ that attains the supremum in* (14)*.*

The second implication of Theorem 4.2 is that it suggests an algorithm for computing the optimal, or near-optimal, stationary policy. We summarize it in Section 4.2.

**Algorithm 1:** Simple EVaR algorithm

**Data:** MDP and desired precision $\delta > 0$
**Result:** $\delta$-optimal policy $\pi^\star \in \Pi_{\mathrm{SD}}$
**while** $g_\infty^\star(0) - g_\infty^\star(\beta_0) > \delta$ **do**
  $\beta_0 \leftarrow \beta_0 / 2$ ;
Construct $\mathcal{B}(\beta_0, \delta)$ as described in (17a) ;
Compute
  $\pi^\star \in \operatorname{argmax}_{\pi \in \Pi_{\mathrm{SD}}} \max_{\beta \in \mathcal{B}(\beta_0, \delta)} h_\infty(\pi, \beta)$ by
  solving a linear program in (13) ;

## 4.2 Algorithms

We now propose a simple algorithm for computing a $\delta$-optimal EVaR policy described in Algorithm 1. The algorithm reduces finding optimal EVaR-TRC policies to solving a sequence of ERM-TRC problems in (5). As Theorem 4.2 shows, there exists a $\delta$-optimal policy such that it is ERM-TRC optimal for some $\beta \in \mathcal{B}(\beta_0, \delta)$. It is, therefore, sufficient to compute an ERM-TRC optimal policy for one of those $\beta$ values.

The analysis above shows that Algorithm 1 is correct.

**Corollary 4.4.** *Algorithm 1 computes the $\delta$-optimal policy $\pi^\star \in \Pi_{\mathrm{SD}}$ that satifies the condition* (19)*.*

Corollary 4.4 follows directly from Theorem 4.2 and from the existence of a sufficiently small $\beta_0$ from the continuity of $g_\infty^\star(\beta)$ for positive $\beta$ around 0.

Algorithm 1 prioritizes simplicity over computational complexity and could be accelerated significantly. Evaluating each $h_\infty^\star(\beta)$ requires computing an optimal ERM-TRC solution which involves solving a linear program. One could reduce the number of evaluations of $h_\infty^\star$ needed by employing a branch-and-bound strategy that takes advantage of the monotonicity of $g_\infty^\star$.

An additional advantage of Algorithm 1 is that the overhead of computing optimal solutions for multiple risk levels $\alpha$ can be small if one selects an appropriate set $\mathcal{B}$.

# 5 Numerical Evaluation

In this section, we illustrate our algorithms and formulations on tabular MDPs that include positive and negative rewards.

The ERM returns for the discounted and transient MDPs in Figure 1 with parameters $r = -0.2$, $\gamma = 0.9$, $\epsilon = 0.9$ are shown in Figure 2. The figure shows that, as expected, the returns are identical in the risk-neutral objective (when $\beta = 0$). However, for $\beta > 0$, the discounted and TRC returns differ significantly. The discounted return is unaffected by $\beta$ while the ERM-TRC return decreases with an increasing $\beta$. Please see Su, Grand-Clément, and Petrik (2024, appendix B) for more details.

To evaluate the effect of risk-aversion on the structure of the optimal policy, we use the *gambler's ruin* problem (Hau, Petrik, and Ghavamzadeh 2023; Bäuerle and Ott 2011). In this problem, a gambler starts with a given amount of capital and seeks to increase it up to a cap $K$. In each turn, the gambler decides how much capital to bet. The bet doubles or is lost with a probability $q$ and $1 - q$, respectively. The

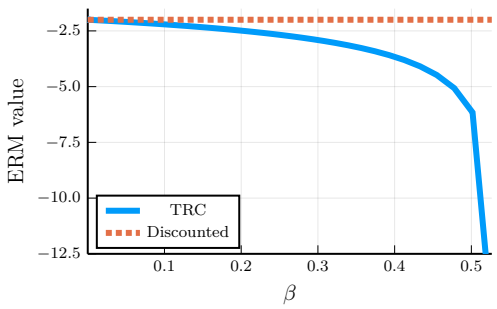


Figure 2: ERM values with TRC and discounted criteria.

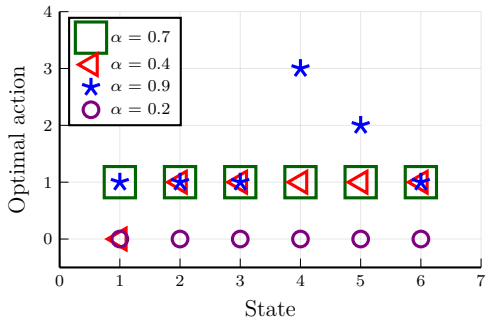


Figure 3: The optimal EVaR-TRC policies.

gambler can quit and keep the current wealth; the game also ends when the gambler goes broke or achieves the cap $K$. The reward equals the final capital, except it is -1 when the gambler is broke. The initial state is chosen uniformly. In the formulation, we use $q = 0.68$, and a cap is $K = 7$. The algorithm was implemented in Julia 1.10, and is available at https://github.com/suxh2019/ERMLP. Please see Su, Grand-Clément, and Petrik (2024, appendix F) for more details.

Figure 3 shows optimal policies for four different EVaR risk levels $\alpha$ computed by Algorithm 1. The state represents how much capital the gambler holds. The optimal action indicates the amount of capital invested. The action 0 means quitting the game. Note that there is only one action when the capital is 0 and 7 for all policies so that action is neglected in Figure 3. Because the optimal policy is stationary, we can interpret and analyze it. The policies become notably less risk-averse as $\alpha$ increases. For example, when $\alpha = 0.2$, the gambler is very risk-averse and always quits with the current capital. When $\alpha = 0.4$, the gambler invests 1 when capital is greater than 1 and quits otherwise to avoid losing it all. When $\alpha = 0.9$, the gambler makes bigger bets, increasing the probability of reaching the cap and losing all capital.

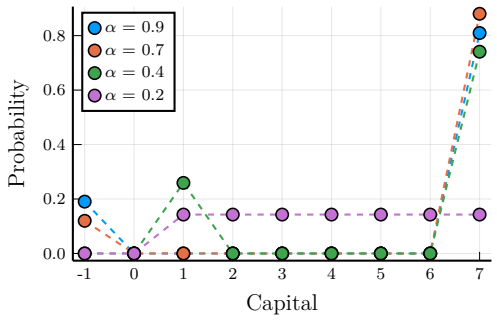


Figure 4: Distribution of the final capital for EVaR optimal policies.

To understand the impact of risk-aversion on the distribution of returns, we simulate the resulting policies over 7,000 episodes and show the distribution of capitals in Figure 4. When $\alpha = 0.2$, the return follows a uniform distribution on $[1, 7]$. When $\alpha = 0.4$, the returns are 1 and 7. When $\alpha = 0.7$ or 0.9, the returns are $-1$ and 7. Overall, the figure shows that for lower values of $\alpha$, the gambler gives up some probability of reaching the cap in exchange for a lower probability of losing all capital.

## 6 Conclusion and Future Work

We analyze transient MDPs with two risk measures: ERM and EVaR. We establish the existence of stationary deterministic optimal policies without any assumptions on the sign of the rewards, a significant departure from past work. Our results also provide algorithms based on value iteration, policy iteration, and linear programming for computing optimal policies.

Future directions include extensions to infinite-state TRC problems, risk-averse MDPs with average rewards, and partial-state observations.

## Acknowledgments

We thank the anonymous reviewers for their detailed reviews and thoughtful comments, which significantly improved the paper's clarity. This work was supported, in part, by NSF grants 2144601 and 2218063. Julien Grand-Clément was supported by Hi! Paris and Agence Nationale de la Recherche (Grant 11-LABX-0047).

## References

Ahmadi, M.; Dixit, A.; Burdick, J. W.; and Ames, A. D. 2021a. Risk-averse stochastic shortest path planning. In *IEEE Conference on Decision and Control (CDC)*, 5199–5204.

Ahmadi, M.; Rosolia, U.; Ingham, M. D.; Murray, R. M.; and Ames, A. D. 2021b. Constrained risk-averse Markov decision processes. In *AAAI Conference on Artificial Intelligence*, volume 35, 11718–11725.

Ahmadi-Javid, A. 2012. Entropic Value-at-Risk: A New Coherent Risk Measure. *Journal of Optimization Theory and Applications*, 155(3): 1105–1123.

Ahmadi-Javid, A.; and Pichler, A. 2017. An analytical study of norms and Banach spaces induced by the entropic value-at-risk. *Mathematics and Financial Economics*, 11(4): 527–550.

Bäuerle, N.; and Glauner, A. 2022. Markov decision processes with recursive risk measures. *European Journal of Operational Research*, 296(3): 953–966.

Bäuerle, N.; and Ott, J. 2011. Markov decision processes with average-value-at-risk criteria. *Mathematical Methods of Operations Research*, 74: 361–379.

Bertsekas, D. P.; and Tsitsiklis, J. N. 1991. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3): 580–595.

Bertsekas, D. P.; and Yu, H. 2013. Stochastic shortest path problems under weak conditions. *Lab. for Information and Decision Systems Report LIDS-P-2909, MIT*.

Blackwell, D. 1967. Positive dynamic programming. In *Berkeley symposium on Mathematical Statistics and Probability*, volume 1, 415–418. University of California Press Berkeley.

Carpin, S.; Chow, Y.-L.; and Pavone, M. 2016. Risk aversion in finite Markov Decision Processes using total cost criteria and average value at risk. In *IEEE International Conference on Robotics and Automation (ICRA)*, 335–342.

Chung, K.-J.; and Sobel, M. J. 1987. Discounted MDP's: Distribution Functions and Exponential Utility Maximization. *SIAM Journal on Control and Optimization*, 25(1): 49–62.

Cohen, A.; Efroni, Y.; Mansour, Y.; and Rosenberg, A. 2021. Minimax regret for stochastic shortest path. *Advances in neural information processing systems*, 34: 28350–28361.

Dann, C.; Wei, C.-Y.; and Zimmert, J. 2023. A Unified Algorithm for Stochastic Path Problems. In *International Conference on Learning Theory*.

de Freitas, E. M.; Freire, V.; and Delgado, K. V. 2020. Risk Sensitive Stochastic Shortest Path and Logsumexp: From Theory to Practice. In Cerri, R.; and Prati, R. C., eds., *Intelligent Systems*, Lecture Notes in Computer Science, 123–139.

de Freitas, E. M.; Freire, V.; and Delgado, K. V. 2020. Risk Sensitive Stochastic Shortest Path and LogSumExp: From Theory to Practice. In *Intelligent Systems: Brazilian Conference (BRACIS)*, 123–139. Springer.

Denardo, E. V.; and Rothblum, U. G. 1979. Optimal stopping, exponential utility, and linear programming. *Mathematical Programming*, 16(1): 228–244.

Fei, Y.; Yang, Z.; Chen, Y.; and Wang, Z. 2021. Exponential bellman equation and improved regret bounds for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 20436–20446.

Fei, Y.; Yang, Z.; and Wang, Z. 2021. Risk-sensitive reinforcement learning with function approximation: A debiasing approach. In *International Conference on Machine Learning*, 3198–3207. PMLR.

Filar, J.; and Vrieze, K. 2012. *Competitive Markov decision processes*. Springer Science & Business Media.

Follmer, H.; and Schied, A. 2016. *Stochastic finance: an introduction in discrete time*. De Gruyter Graduate, 4th edition.

Freire, V.; and Delgado, K. V. 2016. Extreme risk averse policy for goal-directed risk-sensitive Markov decision process. In *Brazilian Conference on Intelligent Systems (BRACIS)*, 79–84.

Gavriel, C.; Hanasusanto, G.; and Kuhn, D. 2012. Risk-averse shortest path problems. In *IEEE Conference on Decision and Control (CDC)*, 2533–2538.

Grand-Clément, J.; and Petrik, M. 2022. Towards Convex Optimization Formulations for Robust MDPs.

Hau, J. L.; Delage, E.; Ghavamzadeh, M.; and Petrik, M. 2023. On Dynamic Programming Decompositions of Static Risk Measures in Markov Decision Processes. In *Neural Information Processing Systems (NeurIPS)*.

Hau, J. L.; Petrik, M.; and Ghavamzadeh, M. 2023. Entropic risk optimization in discounted MDPs. In *International Conference on Artificial Intelligence and Statistics*, 47–76. PMLR.

Horn, R. A.; and Johnson, C. A. 2013. *Matrix Analysis*. Cambridge University Press, 2nd edition.

James, H. W.; and Collins, E. 2006. An analysis of transient Markov decision processes. *Journal of applied probability*, 43(3): 603–621.

Kallenberg, L. 2021. Markov decision processes. *Lecture Notes. University of Leiden*.

Kastner, T.; Erdogdu, M. A.; and Farahmand, A.-m. 2023. Distributional Model Equivalence for Risk-Sensitive Reinforcement Learning. In *Conference on Neural Information Processing Systems*.

Kupper, M.; and Schachermayer, W. 2006. Representation Results for Law Invariant Time Consistent Functions. *Mathematics and Financial Economics*, 16(2): 419–441.

Lam, T.; Verma, A.; Low, B. K. H.; and Jaillet, P. 2022. Risk-Aware Reinforcement Learning with Coherent Risk Measures and Non-linear Function Approximation. In *International Conference on Learning Representations (ICLR)*.

Li, X.; Zhong, H.; and Brandeau, M. L. 2022. Quantile Markov decision processes. *Operations research*, 70(3): 1428–1447.

Marthe, A.; Garivier, A.; and Vernade, C. 2023. Beyond Average Return in Markov Decision Processes. In *Conference on Neural Information Processing Systems*.

Meggendorfer, T. 2022. Risk-aware stochastic shortest path. In *AAAI Conference on Artificial Intelligence*, volume 36, 9858–9867.

Patek, S. D. 1997. *Stochastic and shortest path games: theory and algorithms*. Ph.D. thesis, Massachusetts Institute of Technology.

Patek, S. D. 2001. On terminating Markov decision processes with a risk-averse objective function. *Automatica*, 37(9): 1379–1386.

Patek, S. D.; and Bertsekas, D. P. 1999. Stochastic shortest path games. *SIAM Journal on Control and Optimization*, 37(3): 804–824.

Pflug, G. C.; and Pichler, A. 2016. Time-consistent decisions and temporal decomposition of coherent risk functionals. *Mathematics of Operations Research*, 41(2): 682–699.

Puterman, M. L. 2005. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Smith, K. M.; and Chapman, M. P. 2023. On Exponential Utility and Conditional Value-at-Risk as risk-averse performance criteria. *IEEE Transactions on Control Systems Technology*.

Su, X.; Grand-Clément, J.; and Petrik, M. 2024. Risk-averse Total-reward MDPs with ERM and EVaR. *arXiv preprint arXiv:2408.17286*.

Su, X.; and Petrik, M. 2023. Solving multi-model MDPs by coordinate ascent and dynamic programming. In *Uncertainty in Artificial Intelligence*, 2016–2025. PMLR.

Su, X.; Petrik, M.; and Grand-Clément, J. 2024a. EVaR Optimization in MDPs with Total Reward Criterion. In *Seventeenth European Workshop on Reinforcement Learning*.

Su, X.; Petrik, M.; and Grand-Clément, J. 2024b. Optimality of Stationary Policies in Risk-averse Total-reward MDPs with EVaR. In *ICML 2024 Workshop: Foundations of Reinforcement Learning and Control–Connections and Perspectives*.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. The MIT Press, 2nd edition.