EVaR Optimization in MDPs with Total Reward Criterion

Xihong Su
Department of Computer Science
University of New Hampshire
Durham, NH 03824
xihong.su@unh.edu

Marek Petrik
Department of Computer Science
University of New Hampshire
Durham, NH 03824
mpetrik@cs.unh.edu

Julien Grand-Clément
Information Systems and Operations Management Department
HEC Paris
Jouy-en-Josas, France, 78350
rand-clement@hec.fr

Abstract

The infinite-horizon discounted objective is popular in reinforcement learning, partly due to stationary optimal policies and convenient analysis based on contracting Bellman operators. Unfortunately, optimal policies must be history-dependent for most common coherent risk-averse discounted objectives, such as Value at Risk (VaR) and Conditional Value at Risk (CVaR). They also must be computed using complex state augmentation schemes. In this paper, we show that the *total reward* objective, under the Entropic Risk Measure (ERM) and Entropic Value at Risk (EVaR), can be optimized by a stationary policy, an essential property for practical implementations. In addition, an optimal policy can be efficiently computed using linear programming. Importantly, our results only require the relatively mild condition of transient MDPs and allow for *both* positive and negative rewards, unlike prior work requiring assumptions on the sign of the rewards. Our results suggest that the total reward criterion may be preferable to the discounted criterion in a broad range of risk-averse reinforcement learning problems.

1 Introduction

The literature on Markov decision processes (MDP) [Puterman, 2005] has seen a growing interest in risk-averse objectives [Kastner et al., 2023, Marthe et al., 2023, Lam et al., 2022, Li et al., 2022, Bäuerle and Glauner, 2022, Hau et al., 2023b,a, Su et al., 2024a,b]. Risk-averse objectives penalize the variability of returns and prefer policies with stronger guarantees on the probability of catastrophic losses. As a result, risk-averse objectives are important in critical applications, such as healthcare, autonomous driving, or finance, where avoiding disastrous failures is essential. In modern work, the most common metric in risk-averse objectives is to use a *monetary risk measure*, which generalizes the expectation operator and assigns a real value to any random variable [Follmer and Schied, 2016].

Most reinforcement learning (RL) algorithms, risk-neutral and risk-averse alike, are designed for the *discounted* objective, which computes a weighted sum of rewards over an infinite time horizon with weights that decrease geometrically with time according to a known discount rate [Puterman, 2005, Su and Petrik, 2023]. In financial applications of RL, discounting future rewards accounts for inflation or the option to invest gains. In non-financial applications, the justification for discounting is

17th European Workshop on Reinforcement Learning (ewrl 2024).

more complex and often driven by its algorithmic convenience—discounting guarantees the Bellman operator is a contraction.

The *total reward criterion* (TRC), also known as the stochastic shortest path, is an alternative objective to discounting [Puterman, 2005, Kallenberg, 2021]. In TRC, the horizon is infinite, and future rewards are undiscounted. While the undiscounted sum of rewards may be unbounded in general, a common assumption on the model is that the model is transient. In transient MDPs, there is some positive probability that the process terminates in a bounded number of steps and reaches an absorbing *sink state*. In the risk-neutral settings, transience guarantees that the total sum of rewards remains finite under any policy and, consequently, that an optimal policy exists [Kallenberg, 2021, Filar and Vrieze, 2012].

In this paper, we analyze the foundations of *risk-averse MDPs under the TRC objective* and propose algorithms for solving it. We focus on risk aversion modeled by the Entropic Value-at-Risk (EVaR) and Entropic Risk Measures (ERM) risk measures. As our *main contribution*, we show that stationary deterministic optimal policies always exist for TRC with EVaR and ERM risk-averse objectives. We also show that these stationary policies and value functions can be computed using linear programming. Implementing these algorithms is simple and closely resembles the algorithms for solving MDPs.

Transient MDPs with the TRC criterion are a particularly salient model in risk-averse reinforcement learning. In reinforcement learning, it is common to adopt discounted objectives to account for a probability of termination (transition to a sink state) [Sutton and Barto, 2018]. In risk-neutral settings, there is an equivalence between the probability of terminating and the use of a discount factor. However, as our previous work [Su et al., 2024a] shows, no such correspondence exists with risk-averse objectives, and the difference between them may be arbitrarily large.

Our results also show 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 et al., 2020, Smith and Chapman, 2023, Denardo and Rothblum, 1979, Hau et al., 2023b,a]. Unfortunately, ERM is difficult to interpret; its risk level is scale-dependent, and it is difficult to relate it to popular risk measures like VaR and CVaR. Because EVaR reduces to an optimization over ERM, it preserves most of the computational advantages of ERM. Because EVaR closely approximates CVaR and VaR at the same risk level, its value is much easier to interpret. Finally, EVaR is also a coherent risk measure, unlike ERM [Ahmadi-Javid, 2012, Ahmadi-Javid and Pichler, 2017].

While we are unaware of prior work on the TRC objective with ERM or EVaR risk-aversion, the ERM risk measure is closely related to exponential utility functions. All prior works on TRC with exponential utility functions impose some constraints on the sign of the instantaneous rewards, such as positive rewards [Blackwell, 1967] or negative rewards [Bertsekas and Tsitsiklis, 1991, Freire and Delgado, 2016, de Freitas et al., 2020, Fei et al., 2021a,b]. Note that this significantly limits the modeling power of prior approaches since assuming positive (resp. negative) rewards means that all states are more desirable (resp. detrimental) than the sink state. Our analysis allows rewards that are negative as well as positive.

The remainder of the paper is organized as follows. Section 2 describes basic properties of transient MDPs and common risk measures. Section 3 establishes the main properties of ERM-TRC and computes its optimal stationary policy using linear programming. Section 4 establishes the main properties of EVaR-TRC by reducing the EVaR-TRC to a sequence of ERM-TRC problems and shows that the optimal EVaR-TRC policy is stationary. Section 5 evaluates our algorithm on a tabular transient MDP that includes positive and negative rewards.

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 $z\colon \{1, 2, \ldots, N\} \to \mathbb{R}$, we also treat it interchangeably as a vector $\boldsymbol{z} \in \mathbb{R}^n$ such that $z_i = z(i), \forall i = 1, \ldots, n$.

2 Background on risk-averse MDPs

Markov Decision Processes We focus on solving Markov decision processes (MDPs) [Puterman, 2005], modeled by a tuple $(\bar{\mathcal{S}}, \mathcal{A}, p, r, \mu)$, where $\bar{\mathcal{S}} = \{1, 2, \ldots, S, S+1\}$ is the finite set of states and $\mathcal{A} = \{1, 2, \ldots, A\}$ is the finite set of actions. The transition function $p\colon \bar{\mathcal{S}} \times \mathcal{A} \to \Delta_{\bar{\mathcal{S}}}$ represents the probability $p(s, a, s')$ of transitioning to $s' \in \bar{\mathcal{S}}$ after taking $a \in \mathcal{A}$ in $s \in \bar{\mathcal{S}}$ and $\boldsymbol{p}_{sa} \in \Delta_{\bar{\mathcal{S}}}$ is such that $(\boldsymbol{p}_{sa})_{s'} = p(s, a, s')$. The function $r\colon \bar{\mathcal{S}} \times \mathcal{A} \times \bar{\mathcal{S}} \to \mathbb{R}$ represents the reward $r(s, a, s') \in \mathbb{R}$ associated with transitioning from $s \in \bar{\mathcal{S}}$ and $a \in \mathcal{A}$ to $s' \in \bar{\mathcal{S}}$. The vector $\boldsymbol{\mu} \in \Delta_{\bar{\mathcal{S}}}$ is the initial state distribution. We designate the state $e := S + 1$ as a *sink state* and use $\mathcal{S} = \{1, \ldots, S\}$ to denote the set of all non-sink states. The sink state e must satisfy that $p(e, a, e) = 1$ and $r(e, a, e) = 0$ for each $a \in \mathcal{A}$, and $\mu_e = 0$.

The solution to an MDP is a *policy*. Given a horizon $t \in \mathbb{N}$, a history-dependent policy in the set Π^t_{HR} maps the history of states and actions to a distribution over actions. A *Markov policy* $\pi \in \Pi^t_{\mathrm{MR}}$ is a sequence of decision rules $\pi = (\boldsymbol{d}_0, \boldsymbol{d}_1, \ldots, \boldsymbol{d}_{t-1})$ 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* Π_{SR} are of Markov policies with $\pi = (\boldsymbol{d}, \boldsymbol{d}, \ldots) := (\boldsymbol{d})_\infty$ 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 Π^t_{MD} and Π_{SD}. Finally, we omit the subscript t to indicate infinite horizon definitions of policies for histories of any length.

Optimizing the risk-neutral Total Reward Criterion (TRC) involves solving for

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

where the random variables are denoted by a tilde and $\tilde{s}_t$ and $\tilde{a}_t$ represent the state from $\bar{\mathcal{S}}$ and action from $\mathcal{A}$ at time t. The superscript π denotes the policy that governs the actions $\tilde{a}_t$ when visiting $\tilde{s}_t$ 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.

In risk-neutral objectives, TRC is more challenging to optimize than the discounted criterion. Without any additional assumptions, it is known that TRC may be unbounded, optimal policies may not exist, or may be non-stationary [Bertsekas and Yu, 2013, James and Collins, 2006]. A common assumption that guarantees that the total return is well-behaved is that all policies have a positive probability of eventually transitioning to the sink state. Such MDPs are referred to as being transient [Kallenberg, 2021].

Definition 2.1 (Transient MDP)**.** An MDP is *transient* if for any $\pi \in \Pi_{\mathrm{SD}}$:

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

Transient MDPs are important because their optimal policies exist and can be chosen stationary and deterministic [Kallenberg, 2021, theorem 4.12]. 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: $\rho(\boldsymbol{A}) := \max_{i=1,\ldots,n} |\lambda_i|$ where λ_i is the i-th eigenvalue [Horn and Johnson, 2013].

Lemma 2.2 (Theorem 4.8 in Kallenberg [2021])**.** *An MDP is transient if and only if* $\rho(\boldsymbol{P}^\pi) < 1$ *for all* $\pi \in \Pi_{\mathrm{SD}}$.

One can verify if an MDP is transient in polynomial time without enumerating all policies by solving a linear programming [Kallenberg, 2021, Algorithm 4.1].

Now, let us understand the basic setting 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 are one non-sink state s and one action a. A triple tuple represents an action, transition probability, and an immediate reward separately. Note that every discounted MDP can be converted to a transient MDP by (19) in Appendix B. For the discounted MDP, the discount factor is γ. For the transient MDP, e is the sink state, and there is the 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 γ equals ϵ, their value functions have identical values. Please see Proposition B.1 in Appendix B for details.

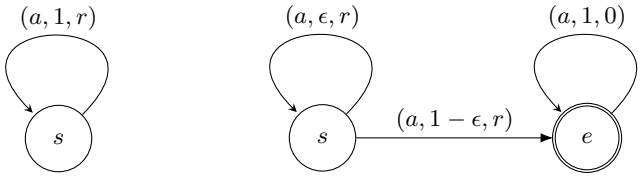

Figure 1: left: a discounted MDP, right: a transient MDP

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]

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

and extended to $\beta \in [0, \infty]$ as $\mathrm{ERM}_0[\tilde{x}] = \lim_{\beta \to 0^+} \mathrm{ERM}_\beta[\tilde{x}] = \mathbb{E}[\tilde{x}]$ and $\mathrm{ERM}_\infty[\tilde{x}] = \lim_{\beta \to \infty} \mathrm{ERM}_\beta[\tilde{x}] = \operatorname{ess\,inf}[\tilde{x}]$. ERM plays a special role in sequential decision-making because it is the only law-invariant risk measure that satisfies the tower property shown in Proposition A.1 [Kupper and Schachermayer, 2006, Marthe et al., 2023], which is essential in constructing dynamic programs [Hau et al., 2023b].

Unfortunately, two significant limitations of ERM hinder its practical applications. First, it 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 et al., 2023b, Ahmadi-Javid, 2012]. Second, the risk parameter β is challenging to interpret and does not relate well to other common risk measures, like VaR or CVaR.

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

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

and is extended to $\mathrm{EVaR}_0[\tilde{x}] = \operatorname{ess\,inf}[\tilde{x}]$ and $\mathrm{EVaR}_1[\tilde{x}] = \mathbb{E}[\tilde{x}]$ [Ahmadi-Javid, 2012]. EVaR addresses the limitations of ERM while preserving its main benefits. First, EVaR is coherent and, therefore, positively homogenous. Second, EVaR is a good approximation to interpretable quantile-based risk measures, like VaR and CVaR [Ahmadi-Javid, 2012, Hau et al., 2023b].

3 Analysis of ERM Total Reward Criterion

In this section, we analyze the ERM-TRC problem. We show that an optimal stationary policy exists for this criterion, and we describe linear programming algorithm for computing it. As discussed previously, the main innovation in this result is that we do not need to assume that the rewards are positive or negative, and we make no additional assumptions on the value of β unlike prior work [Patek, 2001, Denardo and Rothblum, 1979, de Freitas et al., 2020].

Our objective $l \colon \mathbb{R}_{++} \to \bar{\mathbb{R}}$ in this section is to maximize the ERM of the infinite-horizon total sum of rewards, which is formally defined as

$$l(\beta) \;:=\; \sup_{\pi \in \Pi_{\mathrm{HR}}} \liminf_{T \to \infty} \mathrm{ERM}_\beta^{\pi, \boldsymbol{\mu}} \left[\sum_{t=0}^{T} r(\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1}) \right], \tag{5}$$

where $\mathbb{R}_{++}$ is the set of positive real numbers. Analogously to the expectation operator in (1), we use the superscript in the risk measure to indicate the policy and initial distribution governing the distribution of $\tilde{s}_t$ and $\tilde{a}_t$.

Surprisingly, adding risk aversion to the TRC can result in unbounded returns even when the MDP is transient, as we show below. In the remainder of the section, we generally assume that the risk level $\beta > 0$ is fixed and omit it in notations when its value is unambiguous from the context.

3.1 Existence of Optimal Value Functions and Stationary Policies

To prove our results, we first study the finite-horizon ERM objective and then treat the total reward criterion as a limiting case as the horizon tends to infinity.

For the finite-horizon ERM objective, there always exists a Markov deterministic optimal policy (see Appendix C.1), and we therefore define finite-horizon time-dependent value functions for such policies. The finite-horizon value and optimal value functions, $\boldsymbol{v}^t(\pi) \in \mathbb{R}^{\bar{\mathcal{S}}}$ and $\boldsymbol{v}^{t,\star} \in \mathbb{R}^{\bar{\mathcal{S}}}$ respectively, are defined for each horizon $t = 0, \dots$ and policy $\pi \in \Pi_{\mathrm{MD}}$, $s \in \bar{\mathcal{S}}$ as

$$v_s^t(\pi) \;:=\; \mathrm{ERM}_\beta^{\pi,s}\left[\sum_{k=0}^{t-1} r(\tilde{s}_k, \tilde{a}_k, \tilde{s}_{k+1})\right], \qquad v_s^{t,\star} \;=\; \max_{\pi \in \Pi_{\mathrm{MD}}} v_s^t(\pi). \tag{6}$$

Instead of value functions, it will be convenient to consider their exponential transformation that will linearize the corresponding Bellman operators. The *exponential value function* $w_s^\pi \in \mathbb{R}^{\bar{\mathcal{S}}}$ for $\pi \in \Pi_{\mathrm{MD}}$, $t = 0, 1, \dots$, and $s \in \bar{\mathcal{S}}$ is defined as

$$w_s^t(\pi) \;:=\; -\exp\left(-\beta \cdot v_s^t(\pi)\right) \;=\; -\mathbb{E}^{\pi,s}\left[\exp\left(-\beta \cdot \sum_{k=0}^{t-1} r(\tilde{s}_k, \tilde{a}_k, \tilde{s}_{k+1})\right)\right]. \tag{7}$$

The optimal exponential value function $\boldsymbol{w}^{t,\star} \in \mathbb{R}^{\bar{\mathcal{S}}}$ is defined analogously. Note that the exponential value functions satisfy $\boldsymbol{w}^t < \mathbf{0}$ (componentwise) and $\boldsymbol{w}^0(\pi) = \boldsymbol{w}^{t,\star} = -\mathbf{1} = -(1, ..., 1)$ for any $\pi \in \Pi_{\mathrm{MD}}$. The value function can be recovered as

$$v_s^t(\pi) \;=\; -\beta^{-1} \log(-w_s^t(\pi)), \qquad \forall s \in \bar{\mathcal{S}}, t = 0, 1, \dots. \tag{8}$$

As is usual in MDPs, we employ dynamic programming to compute exponential value functions. The *exponential Bellman operator* for $\boldsymbol{w} \in \mathbb{R}^{\mathcal{S}}$ is defined as

$$L^{\boldsymbol{d}}\boldsymbol{w} \;:=\; \boldsymbol{B}^{\boldsymbol{d}}\boldsymbol{w} - \boldsymbol{b}^{\boldsymbol{d}}, \qquad L^\star \boldsymbol{w} \;:=\; \max_{\boldsymbol{d} \in \mathcal{D}} L^{\boldsymbol{d}}\boldsymbol{w} = \max_{\boldsymbol{d} \in \operatorname{ext} \mathcal{D}} L^{\boldsymbol{d}}, \tag{9}$$

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

$$B_{s,s'}^{\boldsymbol{d}} \;:=\; \sum_{a \in \mathcal{A}} p(s, a, s') \cdot d_a(s) \cdot \exp\left(-\beta \cdot r(s, a, s')\right), \tag{10a}$$

$$b_s^{\boldsymbol{d}} \;:=\; \sum_{a \in \mathcal{A}} p(s, a, e) \cdot d_a(s) \cdot \exp\left(-\beta \cdot r(s, a, e)\right). \tag{10b}$$

The following theorem builds on previous results for MDPs with ERM [Hau et al., 2023b] and exponential utility functions [Patek, 1997] to show that exponential value functions can be computed by applying the exponential Bellman operator iteratively. We use the shorthand notation $\pi_{1:t-1} = (\boldsymbol{d}_1, \dots, \boldsymbol{d}_{t-1}) \in \Pi_{\mathrm{MR}}^{t-1}$ to denote the tail of π that starts with $\boldsymbol{d}_1$ instead of $\boldsymbol{d}_0$.

Theorem 3.1. *The exponential value functions $\boldsymbol{w}^t(\pi)$ in (7) for $\pi = (\boldsymbol{d}_0, \dots, \boldsymbol{d}_{t-1}) \in \Pi_{\mathrm{MR}}^t$ and $\boldsymbol{w}^{t,\star}$ can be computed from $\boldsymbol{w}^t(\pi_{1:t-1})$ and $\boldsymbol{w}^{t,\star}$ respectively for $t = 1, \dots$ as:*

$$\boldsymbol{w}^t = L^{\boldsymbol{d}_t}\boldsymbol{w}^{t-1}(\pi_{1:t-1}), \qquad \boldsymbol{w}^{t,\star} = L^\star \boldsymbol{w}^{t-1,\star},$$

and $\boldsymbol{w}^0(\pi) = \boldsymbol{w}^{0,\star} = -\mathbf{1}$. Moreover, there exists $\pi_t^\star \in \Pi_{\mathrm{MD}}^t$ such that $\boldsymbol{w}^t(\pi_t^\star) = \boldsymbol{w}^{t,\star}$ for each $t = 0, \dots$.

The proof of Theorem 3.1 is shown in Appendix C.3.

Using the notation and results above, we now turn to constructing infinite-horizon optimal policies as a limiting case of the finite horizon. The ERM-TRC objective defined in (5) can be expressed for the initial distribution $\boldsymbol{\mu} \in \Delta_{\bar{\mathcal{S}}}$ as

$$\sup_{\pi \in \Pi_{\mathrm{HR}}} \liminf_{t \to \infty} \boldsymbol{\mu}^\top \boldsymbol{v}^t(\pi). \tag{11}$$

As in the finite-horizon case, it will be beneficial to define an exponential transformation of the value function for each $\pi \in \Pi_{\mathrm{MR}}$ as elementwise limits:

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

The following theorem shows the main results of this section. It establishes the existence of an optimal exponential value function, attained by a stationary deterministic policy, and shows that it is the fixed point of the exponential Bellman operator.

Theorem 3.2. *Assume the MDP is transient, $\boldsymbol{\mu} > 0$, and $\boldsymbol{\mu}^\top \boldsymbol{w}^{\infty,\star} > -\infty$. Then there exists $\pi^\star = (\boldsymbol{d}^\star)_\infty \in \Pi_{\text{SD}}$ such that*

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

Moreover, $\boldsymbol{w}^{\infty,\star}$ is the unique fixed point of $L^{\boldsymbol{d}^\star}$.

Before discussing the proof of Theorem 3.2, we state its immediate corollary. That is, there exists an optimal stationary policy that solves the ERM-TRC objective.

Corollary 3.3. *Assume a transient MDP and $\boldsymbol{\mu} > \mathbf{0}$. Then:*

$$\boldsymbol{\mu}^\top \boldsymbol{v}^{\infty,\star} \quad = \quad \max_{\pi \in \Pi_{\text{SD}}} \boldsymbol{\mu}^\top \boldsymbol{v}^{\infty}(\pi).$$

The proof of Corollary 3.3 is shown in Appendix C.5.

Our results are somewhat stronger than purely showing the existence of optimal stationary policies in infinite-horizon objectives. Our results show an optimal stationary policy exists whenever the planning horizon t is sufficiently large. This property mirrors *turnpikes* in discounted MDPs [Puterman, 2005].

Finally, we use the properties above to discuss the impact of β on the objective function.

Proposition 3.4. *There exists a transient MDP and a risk level $\beta > 0$ such that $l(\beta) = -\infty$.*

The proof of Proposition 3.4 is shown in Appendix C.6.

Although the TRC may be unbounded, for each transient MDP, there exists a β such that the TRC is bounded. This result will be important in the analysis of the EVaR objective.

Lemma 3.5. *Assume that the MDP is transient. Then there exists $\beta > 0$ such that $\infty > l(\beta) > -\infty$.*

The proof of Lemma 3.5 is shown in Appendix C.7.

3.2 Outline of the Convergence Proof of Theorem 3.2

We now outline the proof of Theorem 3.2; please see Appendix C.4 for details. To establish Theorem 3.2, we show that $\boldsymbol{w}^{t,\star}$ converges to a fixed point as $t \to \infty$.

Note that standard discounted infinite-horizon arguments do not apply to our ERM-TRC setting: in discounted objectives, one would usually use the contraction property of the Bellman operator under the L_∞ norm to establish the existence of a single fixed point. However, under the TRC, the Bellman operator is not a L_∞-contraction. There are two common techniques in TRC with transient MDPs. The first one is to argue that the Bellman operator is a contraction under specially weighted L_∞ norm [Bertsekas, 2018]. The second one is to argue that $(T^{\boldsymbol{d}})^k$ is an L_∞ contraction for a sufficiently large k where T is the Bellman operator [Bertsekas, 2017].

The risk-neutral TRC proof techniques rely on the linearity of the Bellman evaluation operator [Kallenberg, 2021] and cannot be applied to the nonlinear ERM Bellman operator. To overcome this nonlinearity, we consider the exponential Bellman operator $L^{\boldsymbol{d}}$, which is linear for each $\boldsymbol{d} \in \mathcal{D}$. Although $\boldsymbol{B}^{\boldsymbol{d}}$ is linear, it may be a non-contraction with $\rho(\boldsymbol{B}^{\boldsymbol{d}}) \geq 1$ when the MDP is a transient. This is because the transformation in (10) can increase the transition probabilities leading to rows sums greater than 1 [Horn and Johnson, 2013, theorem 8.1.22]. This precludes us from using standard fixed-point arguments to argue that the limit exists (does not oscillate with time) and is bounded.

Our main contribution is to show that whenever the exponential value functions are bounded, they must be contractions, and the limit exists. To facilitate the analysis, we define $\boldsymbol{w}^t \colon \Pi^t_{\text{SR}} \times \mathbb{R}^{\bar{\text{S}}} \to \mathbb{R}^{\bar{\text{S}}}, t = 0, \ldots$ for $\boldsymbol{z} \in \mathbb{R}^{\bar{\text{S}}}$ as, for $\pi \in \Pi^t_{\text{SR}}$,

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

The value $\boldsymbol{z}$ can be interpreted as the exponential value function at the termination of the process following π for t periods, with exponential value at termination $\boldsymbol{z}$. Note that $\boldsymbol{w}^t(\pi) = \boldsymbol{w}^t(\pi, \mathbf{1})$, $\forall \pi \in \Pi_{\text{MR}}, t = 0, \ldots$. 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 contracting spectral radius.

Lemma 3.6. *Assume a transient MDP and* $\pi = (\boldsymbol{d})_\infty \in \Pi_{\mathrm{SR}}$. *Then for each* $\boldsymbol{\mu} > \mathbf{0}$ *and* $\boldsymbol{z} \geq \mathbf{0}$

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

The proof of Lemma 3.6 is shown in Appendix C.8.

Lemma 3.6 uses the transience property to show that Perron vector $\boldsymbol{f}$ of $\boldsymbol{B}^{\boldsymbol{d}}$ satisfies that $\boldsymbol{f}^\top \boldsymbol{b}^{\boldsymbol{d}} > 0$. Recall that the Perron vector of a non-negative matrix is the eigenvector with the maximum absolute eigenvalue [Horn and Johnson, 2013]. Therefore, $\rho(\boldsymbol{B}^{\boldsymbol{d}}) < 1$ is necessary for the series in (12) to be bounded.

The limitation of Lemma 3.6 is that it only applies to stationary policies and does not preclude the possibility that all stationary policies have unbounded returns while there exists a Markov policy that has a bounded and superior return. To show that this is impossible, we construct an upper bound on $\boldsymbol{w}^{t,\star}$ that decreases monotonically with t and converges when bounded. The proof then concludes by squeezing $\boldsymbol{w}^{t,\star}$ between a lower bound that converges to the upper bound.

3.3 Linear Programming for Computing Value Functions

We now describe the linear program to compute the optimal exponential value function, and the regular value function is recovered in (8). The optimal exponential value function can be computed using the following linear program

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^{\bar{\mathcal{S}}}} \quad & \mathbf{1}^\top \boldsymbol{w} \\ \text{subject to} \quad & w_s \geq -b_s^a + \boldsymbol{B}_{s,\cdot}^a \cdot \boldsymbol{w}, \quad \forall s \in \bar{\mathcal{S}}, a \in \mathcal{A} \end{aligned} \tag{13}$$

where $\boldsymbol{w} = (w_1, \cdots, w_{|\bar{\mathcal{S}}|})$ and $\boldsymbol{B}_{s,\cdot}^a = (\boldsymbol{B}_{s,s_1}^a, \cdots, \boldsymbol{B}_{s,s_{|\bar{\mathcal{S}}|}}^a)$, $\boldsymbol{B}_{s,s'}^a$ and b_s^a are constructed in (10).

4 Reduction of EVaR-TRC to ERM-TRC

In this section, we analyze the EVaR-TRC objective and show that it can be reduced to a sequence of ERM-TRC problems. This reduction is inspired by a reduction proposed for discounted MDPs [Hau et al., 2023b]. Using this reduction, we show that the optimal EVaR-TRC policy is stationary.

The objective in this section is to compute a policy that maximizes the EVaR of the random return $\sum_{t=0}^\infty r(\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1})$ at some given risk level $\alpha \in (0, 1)$ as

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

where π denotes the policy that governs the actions $\tilde{a}_t$ when visiting $\tilde{s}_t$ and $\boldsymbol{\mu}$ denotes the initial state distribution. In (14), we interpret the EVaR of the infinite sum as

$$\mathrm{EVaR}_\alpha^{\pi, \boldsymbol{\mu}} \left[\sum_{t=0}^\infty r(\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1}) \right] = \sup_{\beta > 0} \lim_{T \to \infty} \mathrm{ERM}_\beta^{\pi, \boldsymbol{\mu}} \left[\sum_{t=0}^T r(\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1}) \right] + \beta^{-1} \cdot \log \alpha.$$

One could formulate this objective in other ways, such as putting the limit outside of the supremum operator or inside of the ERM. We chose this formulation because of its convenience and leave the study of related objectives for future work.

Note that the objective in (14) differs from prior work [Ahmadi et al., 2021, Hau et al., 2023b] on EVaR in MDPs is that it considers an undiscounted criterion and a static risk measure.

To compute an optimal policy for the EVaR-TRC objective, we define a proxy objective function $h \colon \mathbb{R} \to \bar{\mathbb{R}}$ as

$$h(\beta) \; := \; \max_{\pi \in \Pi_{\mathrm{SD}}} \left(\lim_{T \to \infty} \mathrm{ERM}_\beta^{\pi, \boldsymbol{\mu}} \left[\sum_{t=0}^T r(\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1}) \right] + \beta^{-1} \cdot \log(\alpha) \right). \tag{15}$$

If we can find a β value that maximizes the function h, then we can use ERM-TRC to compute an optimal policy for EVaR-TRC objective. The main result of this section is given by Theorem 4.1.

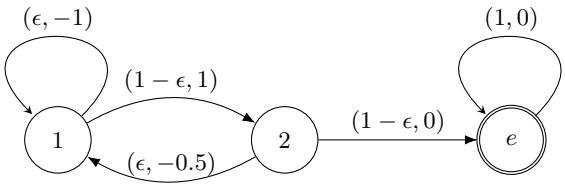

Figure 2: Transient MDP without an equivalent discounted MDP

Theorem 4.1. *Assume a transient MDP. Then:*

1. *if* $\sup_{\beta>0} h(\beta)$ *is attained, then there exists* $\beta^\star > 0$ *and* $\pi^\star \in \Pi_{\text{SD}}$ *such that*

$$h(\beta^\star) = \sup_{\beta>0} h(\beta), \quad \pi^\star \in \arg\max_{\pi\in\Pi_{\text{SD}}} \left(\lim_{T\to\infty} \text{ERM}_{\beta^\star}^{\pi,\boldsymbol{\mu}} \left[\sum_{t=0}^{T} r\left(\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1}\right) \right] \right), \quad (16)$$

   and $\pi^\star$ *is EVaR-TRC optimal in* (14) *and achieves a finite return.*

2. *if* $\sup_{\beta>0} h(\beta)$ *is unattained, then* $\pi^\star$ *is optimal in* (14) *if*

$$\pi^\star \in \arg\max_{\pi\in\Pi_{\text{SD}}} \operatorname{ess\,inf}^{\pi,\boldsymbol{\mu}} \left[\sum_{t=0}^{\infty} r\left(\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1}\right) \right].$$

The proof of Theorem 4.1 is shown in Appendix D.1.

Now, we reduce the EVaR-TRC problem to a specific sequence of ERM-TRC problems. Given an approximation error δ, we use a discrete grid $\beta_0 \cup \{\beta_k\}_{k=1}^K$ to search over the risk level β that can maximize the function h. β_0 is chosen to be a very small number. $K \in \mathbb{N}$ is sufficiently large. Find the $k^\star$ value, $k^\star \in \arg\max_{k=0:K} h(\beta_k)$, and then $\pi^{k^\star}$ is the δ-sub-optimal policy to EVaR-TRC objective in (14). To guarantee δ-sub-optimality of the computed policy, the values $\{\beta_k\}_{k=1}^K$ can be constructed for $k = 1, \ldots, K-2$ as

$$\beta_0, \quad \beta_{k+1} = \frac{\beta_k \log(\alpha)}{\beta_k \delta + \log(\alpha)}. \quad \beta_K = \frac{-\log(\alpha)}{\delta}, \quad (17)$$

Please see Equations (19) and (20) in [Hau et al., 2023b] for more details, derivation, and guarantees.

5 Numerical Illustration

In this section, we first show that there exists a transient MDP that can not be converted to a discounted MDP. Then we illustrate the results of Section 3 and discuss the influence of the risk parameter β on ERM and the influence of termination probability on EVaR on a tabular transient MDP that includes positive rewards and negative rewards.

We describe how to construct a transient MDP from a discounted MDP by Remark 1 in Appendix B. Note that for a discounted MDP, there always exists an equivalent transient MDP constructed by Remark 1. However, there exists a transient MDP shown in Figure 2 that can not be converted to a discounted MDP. For the transient MDP in Figure 2, the state space is $\bar{\mathbb{S}} = \{1, 2, e\}$. There is only one available action a at each state. The initial state distribution $\mu(s_1) = 0.5, \mu(s_2) = 0.5$ and $\mu(e) = 0$. The tuple represents a transition probability and an immediate reward separately. $\epsilon \in [0, 1]$ is used to show the connection between a transient MDP and a discounted MDP. In state s_1, the probability of transitioning from state s_1 to e is $(1-\epsilon)^2$, so the discount factor can be considered as $1-(1-\epsilon)^2$. In state s_2, the probability of transitioning from state s_2 to e is $1-\epsilon$, so the discounted factor can be considered as ϵ. When $1-(1-\epsilon)^2 \neq \epsilon$, the future rewards in states s_1 and s_2 are discounted differently, so there is no corresponding discounted MDP.

Let us illustrate the results of Section 3 and the influence of the risk parameter β on ERM. We set ϵ to 0.5. Compute the exponential transition matrix $\boldsymbol{B}^{\boldsymbol{d}}$ in (10). Because there is only one action a available at each state, the only policy is to take the action a at each state and $d_a(s) = 1, \forall s \in \bar{\mathbb{S}}$. We

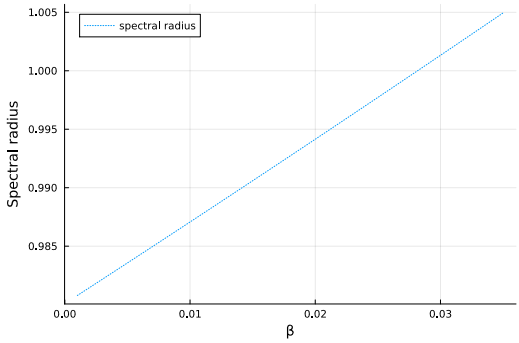

Figure 3: Spectral radius $\rho(\boldsymbol{B^d})$ with $\epsilon = 0.85$

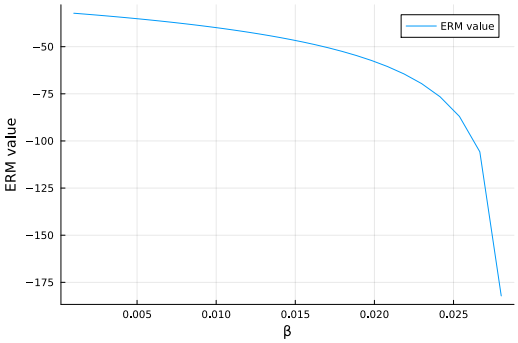

Figure 4: ERM value with $\epsilon = 0.85$

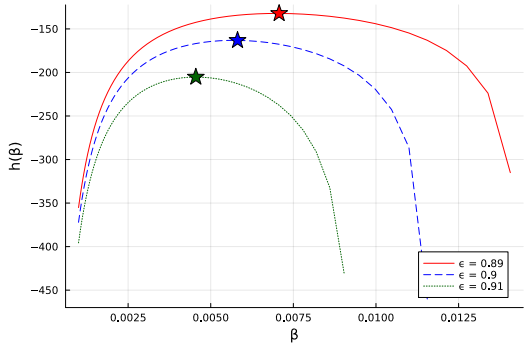

Figure 5: $h(\beta)$ with the risk level $\alpha = 0.75$, EVaR values are labelled by stars

use linear program in (13) to compute the exponential ERM value and recover the regular ERM value by (8). Figures 3 and 4 show the relationship between the spectral radius of $\boldsymbol{B^d}$ and the ERM value. As the spectral radius $\rho(\boldsymbol{B^d})$ approaches to 1, the ERM value dramatically decreases. It is obvious that when $\rho(\boldsymbol{B^d}) \geq 1$, the ERM value will be unbounded. Therefore, $\rho(\boldsymbol{B^d}) < 1$ is the necessary condition for the ERM return to be bounded.

Now we use the transient MDP in Figure 2 to discuss the influence of termination probability on EVaR. The first case is $1 - (1 - \epsilon)^2 = \epsilon$. That is, ϵ is 1 or 0. When $\epsilon = 1$, the probability of transitioning from state s_2 to e is 0, then the MDP is not transient, and the accumulated reward will be $-\infty$. When $\epsilon = 0$,in a discounted criterion, it is a one-step discounted MDP, and the return is 0.5. When $\epsilon = 0$, in a TRC criterion, the agent has the probability 1 of entering the sink state at most 2 steps, and the return is 0.5. For this special case, the EVaR values are identical in discounted and TRC criteria. The second case is $1 - (1 - \epsilon)^2 \neq \epsilon$. That is, $\forall \epsilon \in (0, 1)$, the transient MDP has no corresponding discounted MDP. We set the risk level α to 0.75, δ to 0.01 and δ to $2e - 7$. The optimal EVaR-TRC value is computed by (16) in Theorem 4.1. Figure 5 shows how ϵ values affect $h(\beta)$ defined in (15), EVaR values, and the optimal β values. In general, $h(\beta)$ is not a concave function. For this transient MDP, $h(\beta)$ is a concave function with respect to β. As ϵ increases, the optimal EVaR-TRC value and the optimal β value decrease. Note that β_K defined in (17) is equal to $\frac{-\log(\alpha)}{\delta} = 28.76$, but the optimal β values in Figure 5 are much smaller than β_K.

6 Conclusion

We analyze transient MDPs with two risk measures: ERM and EVaR. We establish the necessary and sufficient conditions for the existence of stationary optimal policies. We allow negative and positive rewards. We prove the convergence of value iteration and show that the optimal stationary policy can be computed using linear programming. Our numerical illustration shows that TRC may be preferable to the discounted criterion under the ERM and EVaR.

References

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

A. Ahmadi-Javid. Entropic Value-at-Risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, 2012.

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

E. Altman. *Constrained Markov Decision Processes*. Routledge, 1998.

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

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

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

P. D. Bertsekas. *Dynamic programming and optimal control 4th edition, volume i*. Athena Scientific, 2017.

P. D. Bertsekas. *Dynamic programming and optimal control 4th edition, volume ii*. Athena Scientific, 2018.

D. Blackwell. Positive dynamic programming. In *Proceedings of the 5th Berkeley symposium on Mathematical Statistics and Probability*, volume 1, pages 415–418. University of California Press Berkeley, 1967.

K.-J. Chung and M. J. Sobel. Discounted MDP's: Distribution functions and exponential utility maximization. *SIAM Journal on Control and Optimization*, 25(1):49–62, 1987.

E. M. de Freitas, V. Freire, and K. V. Delgado. Risk sensitive stochastic shortest path and logsumexp: From theory to practice. In R. Cerri and R. C. Prati, editors, *Intelligent Systems*, Lecture Notes in Computer Science, pages 123–139, 2020.

E. Delage, D. Kuhn, and W. Wiesemann. "Dice"-sion-making under uncertainty: When can a random decision reduce risk? *Management Science*, 65(7):3282–3301, 2019.

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

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

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

E. A. Feinberg and J. Huang. On the reduction of total-cost and average-cost MDPs to discounted MDPs. *Naval Research Logistics (NRL)*, 66(1):38–56, 2019.

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

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

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

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 (NeurIPS)*, 2023a.

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

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

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

R. Johnsonbaugh and W. E. Pfaffenberger. *Foundations of Mathematical Analysis*. Dover Publications, 1981.

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

T. Kastner, M. A. Erdogdu, and A.-m. Farahmand. Distributional model equivalence for risk-sensitive reinforcement learning. In *Conference on Neural Information Processing Systems*, 2023.

M. Kupper and W. Schachermayer. Representation results for law invariant time consistent functions. *Mathematics and Financial Economics*, 16(2):419–441, 2006.

T. Lam, A. Verma, B. K. H. Low, and P. Jaillet. Risk-aware reinforcement learning with coherent risk measures and non-linear function approximation. In *International Conference on Learning Representations*, 2022.

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

A. Marthe, A. Garivier, and C. Vernade. Beyond average return in Markov decision processes. In *Conference on Neural Information Processing Systems*, 2023.

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

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

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

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

R. T. Rockafellar and R. J. Wets. *Variational Analysis*. Springer, 2009.

S. M. Ross and E. A. Peköz. *A second course in probability*. www.ProbabilityBookstore.com, 2023.

K. M. Smith and M. P. Chapman. On exponential utility and conditional value-at-risk as risk-averse performance criteria. *IEEE Transactions on Control Systems Technology*, 2023.

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

X. Su, M. Petrik, and J. Grand-Clément. 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*, 2024a.

X. Su, M. Petrik, and J. Grand-Clément. Stationary policies are optimal in risk-averse total-reward mdps with evar. *arXiv preprint arXiv:2408.17286*, 2024b.

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

A Background

Proposition A.1. *Tower Property for Expectation(Proposition 3.4 in [Ross and Peköz, 2023],Proposition B.1. in [Hau et al., 2023b]) Any two random variables $X_1, X_2 \in \mathbb{X}$, we have*

$$\mathbb{E}[X_1] = \mathbb{E}[\mathbb{E}[X_1|X_2]]$$

B Transient MDP Construction

In the risk-neutral setting, it is well-known that the discounted objective can be interpreted as TRC. The discounted infinite-horizon objective for a factor $\gamma \in (0,1)$ is [Puterman, 2005]

$$\max_{\pi \in \Pi_{\mathrm{SD}}} \rho_\gamma(\pi) := \mathbb{E}^{\pi,\mu}\left[\sum_{t=0}^{\infty} \gamma^t \cdot r(\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1})\right]. \tag{18}$$

Here, $\tilde{s}_t$ and $\tilde{a}_t$ are random variables for the state $\tilde{s}_t$ and action $\tilde{a}_t$ at time t distributed according to the transition probabilities p. The superscript of $\mathbb{E}$ denotes the policy that governs the distribution of $\tilde{a}_t$ and the distribution over the initial state $\tilde{s}_0$. We replace the distribution by a specific state $s \in \mathcal{S}$ when $\mu_s = 1$.

We describe how to construct a transient MDP from a discounted MDP as follows.

Remark 1 (Transient MDP construction). Given an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \mu)$ and a discount factor $\gamma \in [0,1)$, construct an MDP $\bar{\mathcal{M}}_\gamma = (\bar{\mathcal{S}}, \bar{\mathcal{A}}, \bar{p}, \bar{r}, \bar{\mu})$ such that $\bar{\mathcal{S}} = \mathcal{S} \cup \{g\}$, $\bar{\mathcal{A}} = \mathcal{A}$, and $\bar{\mu}(s) = \mu(s), \forall s \in \mathcal{S}$ and $\mu(g) = 0$. The transition function is defined as

$$\bar{p}(s,a,s') = \begin{cases} \gamma \cdot p(s,a,s') & \text{if} \quad s,s' \in \mathcal{S}, a \in \mathcal{A}, \\ 1-\gamma & \text{if} \quad s \in \mathcal{S}, s' = g, a \in \mathcal{A}, \\ 1 & \text{if} \quad s = s' = g, a \in \mathcal{A}, \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

When the rewards $r\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are independent of the next state, then $\bar{r}\colon \bar{\mathcal{S}} \times \bar{\mathcal{A}} \to \mathbb{R}$ are defined as $\bar{r}(s,a) = r(s,a)$ when $s \in \mathcal{S}, a \in \mathcal{A}$ and $r(s,a) = 0$ otherwise. The model can be readily extended to account for the target state dependence by constructing an $\bar{\mathcal{M}}_\gamma$ with a random reward function.

It is well-known that discounted MDPs reduce to TRCs [Altman, 1998, Section 1.10]. The construction can readily be shown using standard dynamic programming techniques to satisfy the following property [Feinberg and Huang, 2019].

Proposition B.1. *For each MDP $\mathcal{M}$, discount factor $\gamma \in [0,1)$, and $\pi \in \Pi_{\mathrm{SD}}$ we have that*

$$\rho_\gamma(\pi, \mathcal{M}) \quad = \quad \rho(\bar{\pi}, \bar{\mathcal{M}}_\gamma),$$

where ρ_γ, ρ are model-dependent and $\bar{\pi}$ extends π to $\bar{\mathcal{S}}$.

C Proofs for Section 3

C.1 Optimality of Markov Policies

The equivalence to solving finite-horizon MDPs with exponential utility functions gives us the following result.

Theorem C.1. *For each $\beta > 0$, there exists an optimal deterministic Markov policy $\pi^{t,\star} \in \Pi_{\mathrm{MD}}$ for each horizon $t = 0, \ldots$:*

$$\max_{\pi \in \Pi_{\mathrm{MR}}} \mathrm{ERM}_\beta^{\pi,s}\left[\sum_{k=0}^{t-1} r(\tilde{s}_k, \tilde{a}_k, \tilde{s}_{k+1})\right] = \max_{\pi \in \Pi_{\mathrm{HR}}} \mathrm{ERM}_\beta^{\pi,s}\left[\sum_{k=0}^{t-1} r(\tilde{s}_k, \tilde{a}_k, \tilde{s}_{k+1})\right].$$

See [Hau et al., 2023b, Corollary 4.2] for a proof. The result can also be derived from the optimality of Markov deterministic policies in MDPs with exponential utility functions [Chung and Sobel, 1987, Patek, 2001].

C.2 Bellman Operator

Lemma C.2. *The exponential Bellman operator is monotone. That is for* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^S$

$$\boldsymbol{x} \geq \boldsymbol{y} \quad \Rightarrow \quad L^{\boldsymbol{d}}\boldsymbol{x} \geq L^{\boldsymbol{d}}\boldsymbol{y}, \qquad \forall \boldsymbol{d} \in \mathcal{D} \tag{20}$$

$$\boldsymbol{x} \geq \boldsymbol{y} \quad \Rightarrow \quad L^{\star}\boldsymbol{x} \geq L^{\star}\boldsymbol{y}. \tag{21}$$

Proof. The property in (20) follows immediately from non-negattivity of $\boldsymbol{B}^{\boldsymbol{d}}$. The property in (21) then follows from the monotonicity of the $\max$ operator. □

Lemma C.3. *The exponential Bellman operators* $L^{\boldsymbol{d}}$, $\forall \boldsymbol{d} \in \mathcal{D}$ *and* $L^{\star}$ *are continuous.*

Proof. The lemma follows directly from the continuity of linear operators and from the fact that the pointwise maximum of a finite number of continuous functions is continuous. See also [Patek, 2001, lemma 5] □

C.3 Proof of Theorem 3.1

Proof of Theorem 3.1. To construct the value function, we can define a Bellman operator $T^{\boldsymbol{d}}\colon \mathbb{R}^S \to \mathbb{R}^S$ for any decision rule $\boldsymbol{d}\colon \mathcal{S} \to \Delta_A$ and the optimal Bellman operator $T^{\star}\colon \mathbb{R}^S \to \mathbb{R}^S$ for a *value vector* $\boldsymbol{v} \in \mathbb{R}^S$ as

$$\begin{aligned} (T^{\boldsymbol{d}}\boldsymbol{v})_s &:= \mathrm{ERM}_{\beta}^{\boldsymbol{d},s}\left[r(s, \tilde{a}_0, \tilde{s}_1) + v_{\tilde{s}_1}\right], \\ T^{\star}\boldsymbol{v} &:= \max_{\boldsymbol{d}\in\mathcal{D}} T^{\boldsymbol{d}}\boldsymbol{v} = \max_{\boldsymbol{d}\in\operatorname{ext}\mathcal{D}} T^{\boldsymbol{d}}\boldsymbol{v}. \end{aligned} \tag{22}$$

It is easy to see that $\boldsymbol{d}$ can be chosen independently for each state to maximize $\boldsymbol{v}$ uniformly across states. The optimality of deterministic decision rules, $\boldsymbol{d} \in \operatorname{ext}\mathcal{D}$, follows because ERM is a mixture quasi-convex function [Delage et al., 2019].

The existence of value function for finite-horizon problem under the ERM objective has been analyzed previously [Hau et al., 2023b] including in the context of exponential utility functions [Chung and Sobel, 1987].

To derive the exponential Bellman operator for the exponential value function for $\boldsymbol{d}\colon \mathcal{S} \to \Delta_A$ we concatenate the Bellman operator with the transformations to and from the exponential value function:

$$\begin{aligned} (L^{\boldsymbol{d}}\boldsymbol{w})_s &:= -\exp\left(-\beta \cdot T^{\boldsymbol{d}}(-\beta^{-1}\log(-\boldsymbol{w}))\right) \\ &= -\mathbb{E}^{\boldsymbol{d},s}\left[\exp\left(-\beta \cdot r(s, \tilde{a}_0, \tilde{s}_1) + \log(-w_{\tilde{s}_1})\right)\right] \\ &= -\mathbb{E}^{\boldsymbol{d},s}\left[\exp\left(-\beta \cdot r(s, \tilde{a}_0, \tilde{s}_1)\right) \cdot (-w_{\tilde{s}_1})\right] \\ &= \sum_{s'\in\bar{\mathcal{S}}}\sum_{a\in\mathcal{A}} p(s,a,s') \cdot d_a(s) \cdot \exp\left(-\beta \cdot r(s,a,s')\right) \cdot w_{s'} \\ &= \sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} p(s,a,s') \cdot d_a(s) \cdot \exp\left(-\beta \cdot r(s,a,s')\right) \cdot w_{s'} \\ &\qquad - \sum_{a\in\mathcal{A}} p(s,a,e) \cdot d_a(s) \cdot \exp\left(-\beta \cdot r(s,a,e)\right). \end{aligned} \tag{23}$$

The derivation above uses the fact that $w_e = -1$ since $v_e = 0$ by definition. The statement of the theorem then follows by algebraic manipulation of $\boldsymbol{B}^{\boldsymbol{d}}, \boldsymbol{b}^{\boldsymbol{d}}$ and by induction on t. The base case hold by the definition of $\boldsymbol{w}^0(\pi) = \boldsymbol{w}^{0,\star} = -\mathbf{1}$.

The existence of an optimal $\pi^{\star}$ follows by choosing the maximum in the definition of $L^{\star}$, which is attained by compactness and continuity of the objective. □

C.4 Proof of Theorem 3.2

Lemma C.4. *Assume some* $\pi = (\boldsymbol{d})_{\infty} \in \Pi_{\mathrm{SR}}$ *such that* $\rho(\boldsymbol{B}^{\boldsymbol{d}}) < 1$. *Then for all* $\boldsymbol{z} \in \mathbb{R}^S$

$$\boldsymbol{w}^{\infty}(\pi) = \boldsymbol{w}^{\infty}(\pi, \boldsymbol{z}) = L^{\boldsymbol{d}}\boldsymbol{w} > -\boldsymbol{\infty}.$$

Proof. The result follows by algebraic manipulation from (12) and basic matrix analysis. When $\rho(\boldsymbol{B}^{\boldsymbol{d}}) < 1$, we get from Neumann series [Horn and Johnson, 2013, problem 5.6.P26]

$$\sum_{k=0}^{t-1} (\boldsymbol{B}^{\boldsymbol{d}})^k \boldsymbol{b}^{\boldsymbol{d}} = (\boldsymbol{I} - \boldsymbol{B}^{\boldsymbol{d}})^{-1} \boldsymbol{b}^{\boldsymbol{d}}$$

and a consequence of Gelfand's formula [Kallenberg, 2021, theorem 4.5]

$$\lim_{k \to \infty} (\boldsymbol{B}^{\boldsymbol{d}})^k \boldsymbol{z} = \boldsymbol{0}.$$

□

Proof of Theorem 3.2. When $\boldsymbol{\mu}^\top \boldsymbol{v}^{\infty,\star} = -\infty$ then the result follows immediately because $-\infty = \sup_{\pi \in \Pi_{\mathrm{MR}}} \boldsymbol{\mu}^\top \boldsymbol{v}^\infty(\pi) \geq \max_{\pi \in \Pi_{\mathrm{SR}}} \boldsymbol{\mu}^\top \boldsymbol{v}^\infty(\pi) \geq -\infty$.

For the remainder of the proof, suppose that $\boldsymbol{\mu}^\top \boldsymbol{v}^\infty(\pi_{\mathrm{M}}^\star) > -\infty$. Then, the exponential value function $\boldsymbol{w}^{t,\star} = \boldsymbol{w}^t(\pi_{\mathrm{M}}) \in \mathbb{R}^S$ of π_{M} satisfies by Theorem 3.1 that

$$\boldsymbol{w}^{0,\star} = -\boldsymbol{1}, \qquad \boldsymbol{w}^t(\pi_{\mathrm{M}}^\star) = L^\star \boldsymbol{w}^{t-1}(\pi_{\mathrm{M}}^\star), \quad t = 1, \dots.$$

We show that $\lim_{t \to \infty} \boldsymbol{w}^{t,\star}$ exists and that it is attained by a stationary policy. We construct a sequence $\boldsymbol{w}_{\mathrm{u}}^t \in \mathbb{R}^S, t = 0, \dots$ as

$$\boldsymbol{w}_{\mathrm{u}}^0 = \boldsymbol{0}, \qquad \boldsymbol{w}_{\mathrm{u}}^t = L^\star \boldsymbol{w}_{\mathrm{u}}^{t-1}, \quad t = 1, \dots.$$

First, we show by induction that

$$\boldsymbol{w}_{\mathrm{u}}^t \quad \geq \quad \boldsymbol{w}^{t,\star}, \qquad t = 0, \dots. \tag{24}$$

The base case $t = 0$ follows immediately from the definitions of $\boldsymbol{w}_{\mathrm{u}}^0$ and $\boldsymbol{w}^{0,\star}$. Next, suppose that (24) holds for some $t > 0$, then it also holds for $t + 1$:

$$\boldsymbol{w}_{\mathrm{u}}^{t+1} \;=\; L^\star \boldsymbol{w}_{\mathrm{u}}^t \;\geq\; L^\star \boldsymbol{w}^{t,\star} \;=\; \boldsymbol{w}^{t+1,\star},$$

where the inequality follows from the inductive assumption and from Lemma C.2. Second, we show by induction that

$$\boldsymbol{w}_{\mathrm{u}}^{t+1} \quad \leq \quad \boldsymbol{w}_{\mathrm{u}}^t, \qquad t = 0, \dots. \tag{25}$$

The base case for $t = 0$ holds as

$$\boldsymbol{w}_{\mathrm{u}}^1 \;=\; L^\star \boldsymbol{w}_{\mathrm{u}}^0 \;=\; L^\star \boldsymbol{0} \;=\; \max_{\boldsymbol{d} \in \mathcal{D}} -\boldsymbol{b}^{\boldsymbol{d}} \;\leq\; \boldsymbol{0} \;=\; \boldsymbol{w}_{\mathrm{u}}^0,$$

where the inequality holds because $\boldsymbol{b}^{\boldsymbol{d}} \geq 0$ from its construction. To prove the inductive step, assume that (25) holds for $t > 0$ and prove it for $t + 1$:

$$\boldsymbol{w}_{\mathrm{u}}^{t+1} \;=\; L^\star \boldsymbol{w}_{\mathrm{u}}^t \;\leq\; L^\star \boldsymbol{w}_{\mathrm{u}}^{t-1} \;=\; \boldsymbol{w}_{\mathrm{u}}^t,$$

where the inequality follows from the inductive assumption and Lemma C.2.

Then, using the Monotone Convergence Theorem [Johnsonbaugh and Pfaffenberger, 1981, theorem 16.2], finite $\mathcal{S}$, and $\inf_{t=0,\dots} \boldsymbol{w}_{\mathrm{u}}^t \geq \inf_{t=0,\dots} \boldsymbol{w}^{t,\star} > -\infty$, we get that there exists $\boldsymbol{w}_{\mathrm{u}}^\star \in \mathbb{R}^S$ such that

$$\boldsymbol{w}_{\mathrm{u}}^\star = \lim_{t \to \infty} \boldsymbol{w}_{\mathrm{u}}^t,$$

and the limit exists. Then, taking the limit of both sides of $\boldsymbol{w}_{\mathrm{u}}^t = L^\star \boldsymbol{w}_{\mathrm{u}}^{t-1}$, we have that

$$\begin{aligned}
\lim_{t \to \infty} \boldsymbol{w}_{\mathrm{u}}^t &= \lim_{t \to \infty} L^\star \boldsymbol{w}_{\mathrm{u}}^{t-1} \\
\boldsymbol{w}_{\mathrm{u}}^\star &= \lim_{t \to \infty} L^\star \boldsymbol{w}_{\mathrm{u}}^{t-1} \\
\boldsymbol{w}_{\mathrm{u}}^\star &= L^\star \lim_{t \to \infty} \boldsymbol{w}_{\mathrm{u}}^{t-1} \\
\boldsymbol{w}_{\mathrm{u}}^\star &= L^{\boldsymbol{d}^\star} \boldsymbol{w}_{\mathrm{u}}^\star,
\end{aligned}$$

where $\boldsymbol{d}^\star = \operatorname{argmax}_{\boldsymbol{d} \in \mathcal{D}} L^{\boldsymbol{d}} \boldsymbol{w}_{\mathrm{u}}$. Above, we can exchange the operators $L^\star$ and $\lim$ by the continuity of $L^\star$ (Lemma C.3).

Now, define $\boldsymbol{w}_{\mathrm{l}}^t \in \mathbb{R}^S, t = 0, \ldots$ as

$$\boldsymbol{w}_{\mathrm{l}}^0 = -\mathbf{1}, \qquad \boldsymbol{w}_{\mathrm{l}}^t = L^{\boldsymbol{d}^\star} \boldsymbol{w}_{\mathrm{l}}^{t-1}, \quad t = 1, \ldots.$$

From the definition of $L^\star$ and by induction on t we have that

$$\boldsymbol{w}_{\mathrm{l}}^t \leq \boldsymbol{w}^{t,\star}.$$

By Lemma 3.6 for $\boldsymbol{z} = \mathbf{0}$, we have that $\rho(\boldsymbol{B}^{\boldsymbol{d}^\star}) < 1$ and therefore from Lemma C.4

$$\lim_{t\to\infty} \boldsymbol{w}_{\mathrm{l}}^t = \lim_{t\to\infty} \boldsymbol{w}_{\mathrm{u}}^t = \boldsymbol{w}_{\mathrm{u}}^\star.$$

In addition, because

$$\boldsymbol{w}_{\mathrm{u}}^t \;\geq\; \boldsymbol{w}^{t,\star} \;\geq\; \boldsymbol{w}_{\mathrm{l}}^t, \qquad t = 0, \ldots$$

The Squeeze Theorem [Johnsonbaugh and Pfaffenberger, 1981, theorem 14.3] then shows that

$$\lim_{t\to\infty} \boldsymbol{w}^{t,\star} = \boldsymbol{w}_{\mathrm{u}}^\star,$$

and $\boldsymbol{d}^\star$ is a stationary policy that attains the return of $\pi_{\mathrm{M}}^\star$. □

C.5 Proof of Corollary 3.3

Proof of Corollary 3.3. From the existence of an optimal stationary policy $\pi^\star \in \Pi_{\mathrm{SD}}$ from for a sufficiently large horizon t from Theorem 3.2 and Appendix C.1, we get that

$$\boldsymbol{\mu}^\top \boldsymbol{v}^\infty(\pi^\star) \leq \sup_{\pi\in\Pi_{\mathrm{HR}}} \liminf_{t\to\infty} \boldsymbol{\mu}^\top \boldsymbol{v}^t(\pi) \leq \liminf_{t\to\infty} \sup_{\pi\in\Pi_{\mathrm{HR}}^t} \boldsymbol{\mu}^\top \boldsymbol{v}^t(\pi) \leq \boldsymbol{\mu}^\top \boldsymbol{v}^\infty(\pi^\star),$$

which implies that all inequalities above hold with equality. □

C.6 Proof of Proposition 3.4

Proof of Proposition 3.4. We use the transient MDP in Figure 1 to show this result. Because the returns of this MDP follow a truncated geometric distribution, its risk-averse return for each $\beta > 0$ and $\epsilon \in (0, 1)$ can be expressed analytically for $t \geq 1$ as

$$\begin{aligned} \mathrm{ERM}_\beta^\pi \left[\sum_{k=0}^{t-1} r(\tilde{s}_k, \tilde{a}_k, \tilde{s}_{k+1})\right] &= -\frac{1}{\beta} \log\left(\sum_{k=0}^{t-1} (1-\epsilon)\epsilon^k \cdot \exp\left(-\beta \cdot k \cdot r\right) + \epsilon^t \cdot 0\right) \\ &= -\frac{1}{\beta} \log\left(\sum_{k=0}^{t-1} (1-\epsilon)\epsilon^{k+1} \cdot \exp\left(-\beta \cdot r\right)^k\right). \end{aligned} \tag{26}$$

Here, $(1-\epsilon)\epsilon^k$ is the probability that the process terminates after exacly k steps, and ϵ^t is the probability that the process does not terminate before reaching the horizon. Then, using the fact that a geometric series $\sum_{i=0}^\infty a \cdot q^i$ for $a \neq 0$ is bounded if and only if $|q| < 1$ we get that

$$\begin{aligned} \lim_{t\to\infty} \mathrm{ERM}_\beta^\pi \left[\sum_{k=0}^{t-1} r(\tilde{s}_k, \tilde{a}_k, \tilde{s}_{k+1})\right] > -\infty \\ \Updownarrow \\ \epsilon \cdot \exp\left(-\beta \cdot r\right) < 1. \end{aligned}$$

Note that $\epsilon \cdot \exp\left(-\beta \cdot r\right) \geq 0$ from its definition. Then, setting $r = -1$ and $\beta > -\log \epsilon$ proves the result. □

C.7 Proof of Lemma 3.5

The result follows from an identical argument to [Patek and Bertsekas, 1999, lemma 1].

C.8 Proof of Lemma 3.6

We use $\boldsymbol{p}^{\boldsymbol{d}} \in \mathbb{R}_+^S$ to represent the probability of terminating in any state for each $\boldsymbol{d} \in \mathcal{D}$:

$$p_s^{\boldsymbol{d}} = \sum_{a \in \mathcal{A}} d_a(s) \cdot \bar{p}(s, a, e), \quad \forall s \in \mathcal{S}.$$

The following lemma establishes a convenient representation of the termination probabilities.

Lemma C.5. *Assume a transient MDP and a policy $\pi = (\boldsymbol{d})_\infty \in \Pi_{\mathrm{SR}}$. Then, the probability of termination in $t = 0, 1, \ldots$ or fewer steps is*

$$\sum_{k=0}^{t} \boldsymbol{\mu}^\top (\boldsymbol{P}^{\boldsymbol{d}})^k \boldsymbol{p}^{\boldsymbol{d}} \;=\; \boldsymbol{\mu}^\top (\boldsymbol{I} - (\boldsymbol{P}^{\boldsymbol{d}})^{t+1}) \mathbf{1}.$$

Proof. We have by algebraic manipulation that

$$\boldsymbol{p}^{\boldsymbol{d}} = (\boldsymbol{I} - \boldsymbol{P}) \mathbf{1}.$$

The probability of terminating in step $t \geq 0$ is

$$\boldsymbol{\mu}^\top (\boldsymbol{P}^{\boldsymbol{d}})^t \boldsymbol{p}^{\boldsymbol{d}}.$$

Using algebraic manipulation and recognizing a telescopic sum, we have that the probability of terminating in $k \leq t$ steps is

$$\sum_{k=0}^{t} \boldsymbol{\mu}^\top (\boldsymbol{P}^{\boldsymbol{d}})^k \boldsymbol{p}^{\boldsymbol{d}} = \sum_{k=0}^{t} \boldsymbol{\mu}^\top (\boldsymbol{P}^{\boldsymbol{d}})^k (\boldsymbol{I} - \boldsymbol{P}^{\boldsymbol{d}}) \mathbf{1} = \boldsymbol{\mu}^\top (\boldsymbol{I} - (\boldsymbol{P}^{\boldsymbol{d}})^{t+1}) \mathbf{1}.$$

□

Lemma C.6. *For any $\boldsymbol{d} \in \mathcal{D}$, the exponential transition matrix is monotone:*

$$\boldsymbol{x} \geq \boldsymbol{y} \quad \Rightarrow \quad \boldsymbol{B}^{\boldsymbol{d}} \boldsymbol{x} \geq \boldsymbol{B}^{\boldsymbol{d}} \boldsymbol{y}, \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^S.$$

Proof. The result follows immediately from the fact that $\boldsymbol{B}^{\boldsymbol{d}}$ is a non-nonnegative matrix. □

Lemma C.7. *For each $t = 0, \ldots$ and each policy $\pi = (\boldsymbol{d})_\infty \in \Pi_{\mathrm{SR}}$ and each $\boldsymbol{\mu} \in \Delta_S$:*

$$\boldsymbol{\mu}^\top (\boldsymbol{B}^{\boldsymbol{d}})^t \boldsymbol{b}^{\boldsymbol{d}} = 0 \qquad \Leftrightarrow \qquad \boldsymbol{\mu}^\top (\boldsymbol{P}^{\boldsymbol{d}})^t \boldsymbol{p}^{\boldsymbol{d}} = 0. \tag{27}$$

Proof. Algebraic manipulation from the definition in (10) shows that

$$\begin{array}{rcccl} c_{\mathrm{l}} \cdot \boldsymbol{p}^{\boldsymbol{d}} & \leq & \boldsymbol{b}^{\boldsymbol{d}} & \leq & c_{\mathrm{u}} \cdot \boldsymbol{p}^{\boldsymbol{d}}, \\ c_{\mathrm{l}} \cdot \boldsymbol{P}^{\boldsymbol{d}} \boldsymbol{x} & \leq & \boldsymbol{B}^{\boldsymbol{d}} \boldsymbol{x} & \leq & c_{\mathrm{u}} \cdot \boldsymbol{P}^{\boldsymbol{d}} \boldsymbol{x}, \qquad \forall \boldsymbol{x} \in \mathbb{R}^S, \end{array} \tag{28}$$

where

$$c_{\mathrm{l}} := \min_{s, s' \in \bar{\mathcal{S}}, a \in \mathcal{A}} \exp\left(-\beta \cdot r(s, a, s')\right), \qquad c_{\mathrm{u}} := \max_{s, s' \in \bar{\mathcal{S}}, a \in \mathcal{A}} \exp\left(-\beta \cdot r(s, a, s')\right).$$

Note that $\infty > c_{\mathrm{u}} > c_{\mathrm{l}} > 0$.

We now extend the inequalities in (28) to multiple time steps. Suppose that $\boldsymbol{y}_{\mathrm{l}} \leq \boldsymbol{y} \leq \boldsymbol{y}_{\mathrm{u}}$, then, for $t = 0, 1, \ldots$:

$$c_{\mathrm{l}}^t \cdot (\boldsymbol{P}^{\boldsymbol{d}})^t \boldsymbol{y}_{\mathrm{l}} \;\leq\; (\boldsymbol{B}^{\boldsymbol{d}})^t \boldsymbol{y} \;\leq\; c_{\mathrm{u}}^t \cdot (\boldsymbol{P}^{\boldsymbol{d}})^t \boldsymbol{y}_{\mathrm{u}}. \tag{29}$$

For the left inequality in (29), the induction proceeds as follows. The base case $t = 0$ holds immediately. For the inductive step, suppose that the left inequality in (29) property holds for $t = 0, \ldots$ then it also holds for $t + 1$ for each $\boldsymbol{y} \in \mathbb{R}^S$ as

$$(\boldsymbol{B}^{\boldsymbol{d}})^{t+1} \boldsymbol{y} \;=\; \boldsymbol{B}^{\boldsymbol{d}} (\boldsymbol{B}^{\boldsymbol{d}})^t \boldsymbol{y} \;\geq\; c_{\mathrm{l}}^t \boldsymbol{B}^{\boldsymbol{d}} (\boldsymbol{P}^{\boldsymbol{d}})^t \boldsymbol{y}_{\mathrm{l}} \;\geq\; c_{\mathrm{l}}^{t+1} \boldsymbol{P}^{\boldsymbol{d}} (\boldsymbol{P}^{\boldsymbol{d}})^t \boldsymbol{y}_{\mathrm{l}} \;=\; c_{\mathrm{l}}^{t+1} (\boldsymbol{P}^{\boldsymbol{d}})^{t+1} \boldsymbol{y}_{\mathrm{l}}.$$

Above, the first inequality follows from Lemma C.6 and from the inductive assumption, and the second inequality follows from (29) by setting $\boldsymbol{x} = \boldsymbol{P}^{\boldsymbol{d}} \boldsymbol{y}$. The right inequality in (29) follows analogously.

Exploiting the fact that $\boldsymbol{\mu} \geq \mathbf{0}$ and substituting $\boldsymbol{y} = \boldsymbol{b}^{\boldsymbol{d}}$, $\boldsymbol{y}_{\mathrm{l}} = c_{\mathrm{l}} \cdot \boldsymbol{p}^{\boldsymbol{d}}$, $\boldsymbol{y}_{\mathrm{u}} = c_{\mathrm{u}} \cdot \boldsymbol{p}^{\boldsymbol{d}}$ into (29) and using the bounds in (28), we get that

$$0 \;\leq\; c_{\mathrm{l}}^{t+1} \cdot \boldsymbol{\mu}^\top (\boldsymbol{P}^{\boldsymbol{d}})^t \boldsymbol{p}^{\boldsymbol{d}} \;\leq\; \boldsymbol{\mu}^\top (\boldsymbol{B}^{\boldsymbol{d}})^t \boldsymbol{b}^{\boldsymbol{d}} \;\leq\; c_{\mathrm{u}}^{t+1} \cdot \boldsymbol{\mu}^\top (\boldsymbol{P}^{\boldsymbol{d}})^t \boldsymbol{p}^{\boldsymbol{d}},$$

where the terms are non-negative because all constants, matrices, and vectors are non-negative. Therefore,

$$\boldsymbol{\mu}^\top (\boldsymbol{B}^{\boldsymbol{d}})^t \boldsymbol{b}^{\boldsymbol{d}} = 0 \quad \Leftrightarrow \quad \boldsymbol{\mu}^\top (\boldsymbol{P}^{\boldsymbol{d}})^t \boldsymbol{p}^{\boldsymbol{d}} = 0.$$

□

Lemma C.8. *Assume a transient MDP and a $\pi = (\boldsymbol{d})_\infty \in \Pi_{\mathrm{SR}}$. Then there exists $\boldsymbol{f} \in \mathbb{R}^S$ such that $\boldsymbol{f}^\top \boldsymbol{B}^{\boldsymbol{d}} = \rho(\boldsymbol{B}^{\boldsymbol{d}}) \cdot \boldsymbol{f}^\top$ and $\boldsymbol{f} \geq \mathbf{0}, \boldsymbol{f} \neq \mathbf{0}$ and*

$$\boldsymbol{f}^\top \boldsymbol{b}^{\boldsymbol{d}} > 0.$$

Proof. Because $\boldsymbol{B}^{\boldsymbol{d}}$ is non-negative, the required vector $\boldsymbol{f}$ exists from the Perron-Frobenius theorem, e.g. [Horn and Johnson, 2013, Theorem 8.3.1]. Therefore,

$$\boldsymbol{f}^\top \boldsymbol{b}^{\boldsymbol{d}} \geq \mathbf{0},$$

since $\boldsymbol{b}^{\boldsymbol{d}} \geq \mathbf{0}$.

It remains to show that $\boldsymbol{f}^\top \boldsymbol{b}^{\boldsymbol{d}} \neq 0$, which we do by deriving a contradiction. Without loss of generality, assume that $\mathbf{1}^\top \boldsymbol{f} = 1$ and suppose that $\boldsymbol{f}^\top \boldsymbol{b}^{\boldsymbol{d}} = 0$. Then:

$$\begin{aligned}
\boldsymbol{f}^\top \boldsymbol{b}^{\boldsymbol{d}} &= 0 && \\
\boldsymbol{f}^\top (\boldsymbol{B}^{\boldsymbol{d}})^t \boldsymbol{b}^{\boldsymbol{d}} &= 0, \forall t = 0, 1, \ldots && \Downarrow \text{ from } \boldsymbol{f}^\top \boldsymbol{B}^{\boldsymbol{d}} = \rho(\boldsymbol{B}^{\boldsymbol{d}}) \cdot \boldsymbol{f}^\top \\
\boldsymbol{f}^\top (\boldsymbol{P}^{\boldsymbol{d}})^t \boldsymbol{p}^{\boldsymbol{d}} &= 0, \forall t = 0, 1, \ldots && \Downarrow \text{ from Lemma C.7} \\
\sum_{k=0}^{t} \boldsymbol{f}^\top (\boldsymbol{P}^{\boldsymbol{d}})^k \boldsymbol{p}^{\boldsymbol{d}} &= 0, \forall t = 0, 1, \ldots && \Downarrow \text{ by summing elements} \\
\boldsymbol{f}^\top (\boldsymbol{I} - (\boldsymbol{P}^{\boldsymbol{d}})^{t+1}) \mathbf{1} &= 0, \forall t = 0, 1, \ldots && \Downarrow \text{ from Lemma C.5} \\
\lim_{t \to \infty} \boldsymbol{f}^\top (\boldsymbol{I} - (\boldsymbol{P}^{\boldsymbol{d}})^{t+1}) \mathbf{1} &= 0, && \Downarrow \text{ limit} \\
\boldsymbol{f}^\top \mathbf{1} &= 0, && \Downarrow \text{ from Lemma 2.2}
\end{aligned}$$

which is a contradiction with $\mathbf{1}^\top \boldsymbol{f} \neq 0$. The last step in the derivation follows from $\rho(\boldsymbol{P}^{\boldsymbol{d}}) < 1$ and therefore $\lim_{t \to \infty} (\boldsymbol{P}^{\boldsymbol{d}})^{t+1} = \mathbf{0}$ [Kallenberg, 2021, Theorem 4.5]. □

Proof of Lemma 3.6. From Lemma C.8, there exists an $\boldsymbol{f} \in \mathbb{R}_+^S$ that $\boldsymbol{f}^\top \boldsymbol{B}^{\boldsymbol{d}} = \rho(\boldsymbol{B}^{\boldsymbol{d}}) \cdot \boldsymbol{f}^\top$ and $\boldsymbol{f} \geq \mathbf{0}, \boldsymbol{f} \neq \mathbf{0}$. Then from (12):

$$-\infty < \boldsymbol{\mu}^\top \boldsymbol{w}^t(\pi) = -\boldsymbol{\mu}^\top (\boldsymbol{B}^{\boldsymbol{d}})^t \boldsymbol{z} - \boldsymbol{\mu}^\top \sum_{k=0}^{t-1} (\boldsymbol{B}^{\boldsymbol{d}})^k \boldsymbol{b} \leq -\sum_{k=0}^{t-1} \rho(\boldsymbol{B}^{\boldsymbol{d}})^k \boldsymbol{\mu}^\top \boldsymbol{b}.$$

The second inequality follows because $\boldsymbol{z} \geq \mathbf{0}$ and $\boldsymbol{B}^{\boldsymbol{d}}$ is non-negative. Since $\boldsymbol{\mu}^\top \boldsymbol{b} > 0$ from Lemma C.8, we can cancel it from the inequality getting that

$$\sum_{k=0}^{t-1} \rho(\boldsymbol{B}^{\boldsymbol{d}})^k < \infty.$$

Then $\rho(\boldsymbol{B}^{\boldsymbol{d}}) < 1$ because $\rho(\boldsymbol{B}^{\boldsymbol{d}}) \geq 0$ the geometric series is bounded. □

D Proofs for Section 4

D.1 Proof of Theorem 4.1

Proof of Theorem 4.1. To streamline the notation, we use

$$\psi(\beta,\pi) := \lim_{T\to\infty} \mathrm{ERM}_{\beta}^{\pi,\boldsymbol{\mu}}\left[\sum_{t=0}^{T} r(\tilde{s}_t,\tilde{a}_t,\tilde{s}_{t+1})\right].$$

First, to prove claim 1., suppose that $\sup_{\beta>0} h(\beta)$ is attained in some $\beta^\star > 0$:

$$\sup_{\beta>0} h(\beta) = h(\beta^\star),$$

and $\pi^\star \in \operatorname{argmax}_{\pi\in} \psi(\beta^\star,\pi)$. Then, the objective in (14) can be expressed as

$$\begin{aligned}
\sup_{\pi\in\Pi_{\mathrm{HR}}} \mathrm{EVaR}_{\alpha}^{\pi,\boldsymbol{\mu}}\left[\sum_{t=0}^{\infty} r(\tilde{s}_t,\tilde{a}_t,\tilde{s}_{t+1})\right] &= \sup_{\pi\in\Pi_{\mathrm{HR}}}\sup_{\beta>0} \psi(\beta,\pi) + \beta^{-1}\cdot\log\alpha \\
&\overset{\text{(a)}}{=} \sup_{\beta>0}\sup_{\pi\in\Pi_{\mathrm{HR}}} \psi(\beta,\pi) + \beta^{-1}\cdot\log\alpha \\
&\overset{\text{(b)}}{=} \sup_{\beta>0}\max_{\pi\in\Pi_{\mathrm{SD}}} \psi(\beta,\pi) + \beta^{-1}\cdot\log\alpha \\
&\overset{\text{(c)}}{=} \max_{\pi\in\Pi_{\mathrm{SD}}} \psi(\beta^\star,\pi) + (\beta^\star)^{-1}\cdot\log\alpha \\
&\overset{\text{(d)}}{=} \psi(\beta^\star,\pi^\star) + (\beta^\star)^{-1}\cdot\log\alpha \\
&\leq \sup_{\beta>0} \psi(\beta,\pi^\star) + (\beta)^{-1}\cdot\log\alpha \\
&= \mathrm{EVaR}_{\alpha}^{\pi^\star,\boldsymbol{\mu}}\left[\sum_{t=0}^{\infty} r(\tilde{s}_t,\tilde{a}_t,\tilde{s}_{t+1})\right].
\end{aligned} \tag{30}$$

Above, the equality in (a) follows by exchanging the suprema [Rockafellar and Wets, 2009, proposition 1.35], (b) follows from Corollary 3.3, (c) follows from the supremum of h being attained, and (d) from the optimality of $\pi^\star$. Then, since $\Pi_{\mathrm{SD}} \subseteq \Pi_{\mathrm{HR}}$, we get that

$$\sup_{\pi\in\Pi_{\mathrm{HR}}} \mathrm{EVaR}_{\alpha}^{\pi,\boldsymbol{\mu}}\left[\sum_{t=0}^{\infty} r(\tilde{s}_t,\tilde{a}_t,\tilde{s}_{t+1})\right] = \mathrm{EVaR}_{\alpha}^{\pi^\star,\boldsymbol{\mu}}\left[\sum_{t=0}^{\infty} r(\tilde{s}_t,\tilde{a}_t,\tilde{s}_{t+1})\right],$$

which proves the first part of the statement of the theorem.

To prove the second part of the statement, suppose that $\sup_{\beta>0} h(\beta)$ is unattained, then because Π_{SD} is finite, there must exist $\pi^\star \in \Pi_{\mathrm{SD}}$ such that

$$\sup_{\beta>0} h(\beta) = \sup_{\beta>0} \psi(\beta,\pi^\star) + \beta^{-1}\cdot\log\alpha, \tag{31}$$

and the supremum on the right-hand side is not attained. This is true because if the supremum were attained for all policies $\pi \in \Pi_{\mathrm{SD}}$, then the supremum of the maxima would also be attained. From the properties of EVaR, we have that when the supremum is unattained, then [Ahmadi-Javid and Pichler, 2017, proposition 2.11]

$$\mathrm{EVaR}_{\alpha}^{\pi^\star,\boldsymbol{\mu}}\left[\sum_{t=0}^{\infty} r(\tilde{s}_t,\tilde{a}_t,\tilde{s}_{t+1})\right] = \operatorname{ess\,inf}^{\pi^\star,\boldsymbol{\mu}}\left[\sum_{t=0}^{\infty} r(\tilde{s}_t,\tilde{a}_t,\tilde{s}_{t+1})\right].$$

Then, following an analogous reasoning to (30), we get that

$$\begin{aligned}
\sup_{\pi\in\Pi_{\mathrm{HR}}} \mathrm{EVaR}_{\alpha}^{\pi,\boldsymbol{\mu}}\left[\sum_{t=0}^{\infty} r(\tilde{s}_t,\tilde{a}_t,\tilde{s}_{t+1})\right] &= \sup_{\pi\in\Pi_{\mathrm{HR}}}\sup_{\beta>0} \psi(\beta,\pi)+\beta^{-1}\cdot\log\alpha \\
&= \sup_{\beta>0}\sup_{\pi\in\Pi_{\mathrm{HR}}} \psi(\beta,\pi)+\beta^{-1}\cdot\log\alpha \\
&= \sup_{\beta>0}\max_{\pi\in\Pi_{\mathrm{SD}}} \psi(\beta,\pi)+\beta^{-1}\cdot\log\alpha \\
&= \sup_{\beta>0} \psi(\beta,\pi^\star)+\beta^{-1}\cdot\log\alpha \\
&= \operatorname{ess\,inf}^{\pi^\star,\boldsymbol{\mu}}\left[\sum_{t=0}^{\infty} r(\tilde{s}_t,\tilde{a}_t,\tilde{s}_{t+1})\right],
\end{aligned}$$

which proves the optimality of $\pi^\star$. Finally, note from (31) we get that

$$\begin{aligned}
\operatorname{ess\,inf}^{\pi^\star,\boldsymbol{\mu}}\left[\sum_{t=0}^{\infty} r(\tilde{s}_t,\tilde{a}_t,\tilde{s}_{t+1})\right] &= \sup_{\beta>0} \psi(\beta,\pi^\star)+\beta^{-1}\cdot\log\alpha \\
&= \sup_{\beta>0}\max_{\pi\in\Pi_{\mathrm{SD}}} \psi(\beta,\pi)+\beta^{-1}\cdot\log\alpha \\
&= \sup_{\pi\in\Pi_{\mathrm{SD}}}\sup_{\beta>0} \psi(\beta,\pi)+\beta^{-1}\cdot\log\alpha \\
&\geq \operatorname{ess\,inf}^{\pi,\boldsymbol{\mu}}\left[\sum_{t=0}^{\infty} r(\tilde{s}_t,\tilde{a}_t,\tilde{s}_{t+1})\right], \quad \forall\pi\in\Pi_{\mathrm{SD}},
\end{aligned}$$

which shows that $\pi^\star$ maximizes the essential infimum, as desired. □