Efficient Risk-sensitive Planning via Entropic Risk Measures

Alexandre Marthe
ENS de Lyon
Lyon, France
*

Samuel Bounan
ENS de Lyon
Lyon, France

Aurélien Garivier
ENS de Lyon
Lyon, France

Claire Vernade
University of Tübingen
Tübingen, Germany

Abstract

Risk-sensitive planning aims to identify policies maximizing some tail-focused metrics in Markov Decision Processes (MDPs), such as (Conditional) Values at Risk. In general, such optimization problems can be very costly as it is known that only Entropic Risk Measures (EntRM) can be efficiently optimized through dynamic programming. We show that EntRM can serve as an approximation of other metrics of interest and we propose a dual optimization problem that requires to compute the set of optimal policies for EntRM across all parameter values. We prove that this *optimality front* can be computed effectively thanks to a novel structural analysis of the smoothness properties of entropic risks. Empirical results demonstrate that our approach achieves strong performance in risk-sensitive decision-making scenarios.

1 Introduction

Markov Decision Processes (MDPs) [35] capture sequential decision making in domains as diverse as robotics, finance, healthcare, and operations research [41, 39, 16, 34]. Standard planning problems are about maximizing the *expected* cumulative reward, which can be done efficiently via dynamic programming. However, in many high-stakes applications, such as healthcare, average performance alone is not sufficient. In fact, it may be critical to limit the probability of catastrophic outcomes or to ensure that returns remain above certain thresholds with high confidence. This need has driven research on *risk-sensitive* control, which incorporates measures of uncertainty and tail behavior into the objective [14, 42].

One of the central challenges in risk-sensitive optimization is to identify risk criteria that are both *meaningful* for real-world decision making and *tractable* in an MDP context. Popular approaches often revolve around the quantile-based Value at Risk (VaR) and Conditional Value at Risk (CVaR) [4, 36], which are widely used to control tail risk. Another common objective relies on controlling a Threshold Probability [45], that is the probability that total returns fall below a specified level.

Despite their practical interest and interpretability properties, these common risk metrics cannot be directly and efficiently optimized in MDPs [31, 37]. Our work addresses precisely this gap by connecting the Threshold Probability and the (C)VaR metrics to the moment-generating function of the return of a policy. In fact, in the context of MDPs, these two optimization problems can be successfully approximated by the Entropic (Exponential) Risk Measure (EntRM) [25], the unique

*alexandre.marthe@ens-lyon.fr

18th European Workshop on Reinforcement Learning (EWRL 2025).

functional admitting a dynamic programming decomposition [12, 22, 31]. Thus, EntRM arguably leads to the "best possible" extension of the risk-neutral Bellman recursion to a risk-aware MDP setting. Yet, despite this computational appeal, practical usage of EntRM can be hindered by interpretability concerns, especially around selecting the risk-tolerance parameter β.

We propose a unifying framework for risk-sensitive planning in MDPs through the study of the EntRM. Rather than directly optimizing the latter as a proxy for other target metrics, we prove that optimal policies evolve in a structured manner as the risk parameter changes, which allows us to derive an efficient algorithm to compute all the optimal policies for EntRM, a set that we call *the Optimality Front*. We show how this set can then be used to optimize tail-focused risk measures through the *Generalized Policy Improvement* principle [6] (see Sec. 3). We demonstrate the consequence of this new method on the optimization of the Threshold Probability, the VaR and CVaR via an empirical study on an inventory management problem.

2 Risk Sensitive Planning

2.1 Risk-sensitivity in MDPs

We consider a finite-horizon MDP $\mathcal{M} = \big(\mathcal{X}, \mathcal{A}, p, r, H, p_1\big)$, where $\mathcal{X}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $p(x' \mid x, a)$ is the probability of transitioning to state x' from state x when action a is chosen, $r(x, a)$ is the reward function specifying the immediate reward for taking action a in state x (bounded in $[-1, 1]$), H is the finite horizon, i.e., the number of decision steps, and p_1 is the distribution of initial states $x_1 \sim p_1(\cdot)$. To simplify notation, we assume all rewards and transitions are stationary, i.e. $p_t = p$ and $r_t = r\ \forall t$, and the rewards are deterministic, but our results naturally extend to the non-stationary setting.

A policy $\pi = (\pi_1, \dots, \pi_H)$ is a sequence of decision rules $\pi_t : \mathcal{X} \to \mathcal{A}$ that, at each step $t \in \{1, 2, \dots, H\}$, selects an action based on the current state. For simplicity, we assume policies to be deterministic, but not necessarily stationary (a stationary policy verifies $\forall t, \pi_1 = \pi_t$). The corresponding *cumulative reward* (or *return*) is a random variable, $R^\pi = \sum_{t=1}^{H} r\big(x_t, a_t\big)$, where $x_1 \sim p_1(\cdot)$, $a_t = \pi_t(x_t)$ and $x_{t+1} \sim p(\cdot \mid x_t, a_t)$. We also denote $R^\pi_h(x, a) = \sum_{t=h}^{H} r\big(x_t, a_t\big)$ with $x_h = x, a_h = a$, and $R^\pi_h(x) = R^\pi_h(x, \pi(x))$.

Risk-sensitive control studies generic optimization problems $\max_\pi \rho(R^\pi)$ for a given functional ρ that can capture *tail events* or *uncertainty aversion*. Typical choices of ρ are described below.

2.2 Risk metrics and computational limitations in MDPs

The theory of risk measures is too rich for us to provide more than a few insights below on the most widely used examples. We define the EntRM and common risk metrics, and we discuss the computational challenges that motivate our work.

Entropic Risk Measure (EntRM). Originally introduced by Howard and Matheson [25] for MDPs, the EntRM for a random variable X and parameter $\beta \in \mathbb{R}$ is defined as

$$\mathrm{EntRM}_\beta[X] = \begin{cases} \frac{1}{\beta} \log\big(\mathbb{E}[e^{\beta X}]\big), & \text{if } \beta \neq 0, \\ \mathbb{E}[X], & \text{if } \beta = 0. \end{cases} \tag{1}$$

In the MDP context, maximizing $\mathrm{EntRM}_\beta[R^\pi]$ is also equivalent to maximizing $\mathrm{sign}(\beta)\, \mathbb{E}[\, e^{\beta R^\pi}]$, by composition with a non-decreasing function. We call it the *exponential form* of EntRM, and will use both forms throughout the paper, noting that the core optimization problem remains the same.

The parameter β encodes risk tolerance: large positive β values emphasize a risk-seeking attitude (amplifying high returns), while negative values induce conservativeness. This seemingly intuitive interpretation unfortunately does not hold the test of practice for two main reasons: the exact choice of β for a specific level of risk-sensitivity is unclear in general, and EntRM is not *positively homogeneous*, i.e. $\mathrm{EntRM}_\beta[cX] \neq c\,\mathrm{EntRM}_\beta[X]$ for $c \in \mathbb{R}$. In finance, this is problematic when scaling outcomes by different currencies or units. In the following, the risk parameter is always assumed to be non-positive.

Value at Risk (VaR). VaR represents the quantile of the distribution. At level $\alpha \in (0, 1)$:

$$\mathrm{VaR}_\alpha[R^\pi] \;=\; \inf\Big\{\, x \;\Big|\; \Pr\big(R^\pi \leq x\big) \;\geq\; \alpha \Big\}. \tag{2}$$

Although VaR is widely used, it suffers from a lack of *sub-additivity*, a desirable property for risk measures ensuring that diversification does not increase perceived risk. In the context of MDPs, VaR is particularly challenging to optimize due to its non-linearity and non-monotonicity [19].

Conditional Value at Risk (CVaR). The CVaR [36, 22, 8, 17] at level α is given by

$$\mathrm{CVaR}_\alpha[R^\pi] \;=\; \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_\gamma[R^\pi]d\gamma. \tag{3}$$

CVaR represents the expected return in the α-worst fraction of outcomes, providing a more comprehensive view of the risk than VaR. What makes it particularly interesting is that it is a *coherent risk measure*, meaning that it satisfies both sub-additivity and positive homogeneity, compared to the risk measure mentioned previously.

In this article, we chose the convention of VaR and CVaR representing the low tail of the distributions. In the literature, it is often defined using the high tail. This has no impact on the theory as a change of variable $X \leftarrow -X$ will transpose the results. More details can be found in Section B.

Threshold Probability. Certainly the most intuitive measure of risk is the probability of falling below a user-specified threshold level T. Solving $\min_\pi \Pr\big(R^\pi \;\leq\; T\big)$, means seeking a policy whose probability of yielding a return below T is minimal. The VaR and Threshold optimization are **dual** problems. A farmer who worries about the possibility of a very poor harvest in the coming year will either ask, "What is the chance that my yield will be below 2 tons?" (Threshold Probability viewpoint), or "With 90% confidence, what is the size of the minimal yield I can expect?" (VaR perspective). Lowering the probability of dropping below a certain threshold (Threshold Probability) is directly tied to choosing a quantile-based cutoff for outcomes, as $\mathrm{VaR}_\alpha[X] = T$ just means that $P(X \leq T) = \alpha$ (if the distribution of X is continuous).

Computational limitations. The EntRM verifies a recursive Dynamic Programming equation akin to the Bellman Equation [27, 22, 37, 31]. Thus, for this family of metrics, the optimization problem is tractable (See Section A for details). Yet, optimization is much more complex for the other risk measures. On the other hand, the optimal policies for popular risk measures such as (C)VaR can be non-Markovian[2] [28, 24], and standard optimization strategies rely on augmenting the state space with a continuous variable that contains the cumulated reward so far, which makes the problem intractable except for very simple MDPs. The optimization of VaR and especially CVaR have been studied extensively [17, 18, 19, 1, 8], but remains a challenging task [24].

Approximation schemes have been proposed, such as *Dynamic Risk Measures* [7] (also called *Nested* or *Recursive Risk Measures*), where the goal is to recursively optimize a pseudo Bellman objective $V_h(x) = \max_a \rho[r(x, a) + V_{h+1}(X')]$, where X' denotes the (deterministic) next state given (x, a). Despite its appealing computational benefits, such criterion does not optimize for ρ in general. In fact, the approximation is usually quite poor (see Sec.5) and the actual optimized objective is not law invariant, which makes it harder to interpret. Optimizing the Threshold Probability raises the same challenges as VaR and CVaR [45, 46, 26] but seems to have been less studied and the literature lacks approximation schemes. In short, while the EntRM is the only risk measure that can be efficiently optimized in MDPs, it is not what people use in practice due to a lack of interpretability. We propose a novel framework to answer the question: *How can we leverage the computational properties of EntRM to optimize those more preferred measures of risk?*

3 A unified framework for risk-sensitive optimization

The Threshold Probability and VaR/CVaR objectives are all related to the tail probabilities of the return distribution. These tail probabilities can be approximated with the help of exponential moments

[2] It means that the policy in state S_t depends on the full history up to this state and the accumulated reward so far.

of the distribution by Chernoff's bound:

$$\Pr\big(X \leq T\big) \quad \leq \quad \inf_{\beta \leq 0} \, \exp\big(-\beta\, T\big)\, \mathbb{E}\big[e^{\beta\, X}\big]. \tag{4}$$

Exponential moments are the core of Entropic Risk Measure and we explain in this section how Inequality (4) leads to proxies for the risk metrics introduced above.

From (C)VaR to EVaR. Solving for the rhs of (4) to equal α, Ahmadi-Javid [2] introduced the *Entropic Value at Risk (EVaR)* as a proxy for the VaR defined by

$$\mathrm{EVaR}_\alpha[X] = \sup_{\beta<0} \mathrm{EntRM}_\beta\big[X\big] - \frac{1}{\beta}\log(\alpha).$$

EVaR has been shown to be a tight approximation of the VaR, and an even tighter one for CVaR, due to $\mathrm{VaR}_\alpha[X] \geq \mathrm{CVaR}_\alpha[X] \geq \mathrm{EVaR}_\alpha[X]$ [2]. It is a *coherent* risk measure, and its use for approximating VaR for bandit algorithms was already noted by Maillard [30] and has received growing interest recently in MDPs [33, 23, 40]. The related proxy for VaR and CVaR is $\max_\pi \mathrm{VaR}_\alpha\left[R^\pi\right] \geq \sup_{\beta<0} \max_\pi \mathrm{EntRM}_\beta\big[R^\pi\big] - \frac{\log(\alpha)}{\beta}$, where the $\sup$ and $\max$ can be swapped as the policy space is finite.

A proxy for the Threshold Probability. Using a similar idea, we derive a proxy for the Threshold Probability: $\min_\pi \ \Pr\big(R^\pi \ \leq \ T\big) \leq \min_{\beta<0} \min_\pi e^{-\beta\, T}\, \mathbb{E}\big[e^{\beta R^\pi}\big]$. More details can be found in Section C.1. To the best of our knowledge, this metric has not been studied so far, and this simple approximation scheme seems novel.

Quality of the approximations. The quality of deviations bounds used to derive the proxies above depend on the tail of the distributions and are known to be more accurate for distributions with light tails [44]. In MDPs, the return of a policy is more concentrated around its mean when there is a rich and bounded reward signals along the trajectory, leading to thinner tails and tighter EVaR approximations.

On the other hand, existing methods to optimize VaR, CVaR and Threshold Probability rely on dynamic programming on extended MDPs, where the state space is augmented with a continuous variable that correspond to the achievable values of the return [17, 45]. Thus, even for small MDPs, rich reward signals may quickly increase the dimension of the state space to an extent that renders the optimization intractable. Conveniently, this is when our approximation method is relevant.

Relaxed optimization problems. Having derived proxy targets, a natural idea is to attempt to optimize them *instead of the true objective*. We comment on these proxy optimization problems and propose below a better and more general method that builds on these intermediate solutions.

For a given parameter β, we denote π^*_β the optimal policy for the EntRM_β [3]:

$$\pi^*_\beta := \operatorname*{argmax}_{\pi\in\Pi} \mathrm{EntRM}_\beta[R^\pi]\,. \tag{5}$$

For all objectives, as suggested by the proxy derivations above, risk-sensitive optimization problems can be reduced to finding the right parameter β:

$$\beta^*_{\text{(C)VAR}} = \arg\sup_{\beta<0} \mathrm{EntRM}_\beta\big[R^{\pi^*_\beta}\big] - \frac{1}{\beta}\log(\alpha), \quad \beta^*_{\text{TP}} = \arg\inf_{\beta<0} e^{-\beta\, T}\, \mathbb{E}\big[e^{\beta R^{\pi^*_\beta}}\big], \tag{6}$$

with the corresponding $\pi^*_{\beta^*}$ as the optimal policy for the optimized metric. Both proxy problems are effectively optimizing the EntRM, inducing important properties of the resulting optimal policies.

Proposition 3.1 (Hau et al. [23])**.** *For each proxy problem in* (6)*, there exists $\beta < 0$ such that the optimal policy is also optimal for the EntRM with parameter β. The optimal policies are deterministic and Markovian.*

Compared to the initial problems where optimal policies are usually not Markovian, the transformation to the EntRM allows finding optimal policies that are easier to compute and implement. However, the proxies in (6) now involve an optimization over a *continuous* range of values of $\beta < 0$, and for each value, the return of the optimal policy π^*_β must be computed. While we know efficient algorithms to

[3] Recall that this optimal policy can be efficiently computed by DP.

compute this quantity for one fixed β, optimizing over a continuous range is significantly harder and the only known approaches use discretization schemes [23]. See Section C.1 for more details.

Our main contribution is to show that this continuous search problem can be done more efficiently, and more importantly, that the set of all optimal policies $(\pi^*_\beta)_{\beta<0}$ is nicely structured and can be stored to further optimize the true objectives rather than their proxy.

3.1 Generalized Policy Improvement

The key observation is that the optimal policy $\pi^*_{\beta^*}$ for the proxy target (6) need not be the best one for the true objective. In general, there may be $\beta' \neq \beta^*$ such that $\pi^*_{\beta'}$ achieves better performance on the true objective.

Conveniently, optimizing the proxies already requires sweeping through all β values and computing all the optimal policies for EntRM (5), $\Pi^* = \{\pi^*_\beta \mid \beta \leq 0\}$, that we call the *Optimality Front*. While previous work discard this set during the optimization process [23], we propose to store it, and apply the *Generalized Policy Improvement* principle (GPI) proposed by Barreto et al. [6]. Namely, multiple policies are computed using tractable objectives (here, EntRM_β for various β), and then the one performing best under the original, possibly intractable, risk criterion is selected. Our novel and original approach is to leverage the Optimality Front such that for any risk measure ρ, we compute the optimal policy as

$$\max_{\pi \in \Pi^*} \rho\big(R^\pi\big) \quad \text{with} \quad \Pi^* = \{\pi^*_\beta \mid \beta \leq 0\}\ . \tag{7}$$

In the next section, we give all the elements to justify that the Optimality Front of the EntRM is indeed a good set of policies to perform GPI. Mainly, we show that it can be computed efficiently and that it is *small* in general because there is a bounded number of β values for which the optimal policy changes. Importantly, it can be computed once and reused across multiple downstream risk objectives, such as VaR, CVaR, or threshold-based criteria. Evaluating each candidate policy under the target risk measure can be done efficiently using distributional planning [10], which provides access to the full return distribution of each π^*_β.

4 Structural Insights into Entropic Risk Measures

Optimizing EntRM for an entire range of risk parameters requires understanding the structure of EntRM optimal policies. We now show that exploiting the regularity of the EntRM_β function leads to an efficient algorithm that computes *all the optimal policies* along $\beta \in \mathbb{R}$ much more efficiently than using a grid, and with better guarantees. In passing, understanding how a small perturbation of the risk parameter can influence the optimal policy is also helpful from the point of view of interpretability and robustness [9]. More formal statements[4] and all the proofs are given in the appendix.

4.1 Structure Analysis

Definition 4.1 (Optimality Front)**.** For any MDP, the *optimality front* is defined as $\Gamma = (\pi_k, I_k)_k$, where $(I_k)_k$ is the partition of $\mathbb{R}^-$ and where π_k is the optimal policy of EntRM for all risk tolerance parameters $\beta \in I_k$.

This definition is justified by the following property, formalizing the intuition that a small perturbation of the risk parameter typically does not change the optimal policy.

Proposition 4.2. *The Optimality Front Γ contains a finite set of policies. Each policy in Γ is optimal on a finite union of closed intervals. In other words, the mapping $\beta \mapsto \pi^*_\beta$ is piecewise constant.*

Crucially, we are able to prove lower bounds on the length of these intervals locally around a specific risk parameter β, knowing the Advantage function at this specific point.

Theorem 4.3. *Let $\beta \in \mathbb{R}$ be such that there is a unique deterministic policy π^*_β optimizing EntRM_β. Define the Generalized Advantage function:*

$$A^\pi_{h,\beta}(x,a) \;=\; \mathrm{EntRM}_\beta[R^\pi_h(x)] \;-\; \mathrm{EntRM}_\beta[R^\pi_h(x,a)].$$

[4] To make the paper easier to read, we present slightly informal statements in the main text and refer curious readers to the appendix for full details.

Then, define optimality gaps as the smallest differences over the entire MDP:

$$\Delta = \begin{cases} \frac{|\beta|}{2} \min_{h,x} \min_{a \neq \pi^*_{\beta,h}(x)} \frac{1}{H-h} A^{\pi^*_\beta}_{h,\beta}(x,a) & \text{if } \beta \neq 0 \\ 2 \min_{h,x} \min_{a \neq \pi^*_{\beta,h}(x)} \frac{1}{(H-h)^2} A^{\pi^*_\beta}_{h,\beta}(x,a) & \text{if } \beta = 0. \end{cases}$$

Then, for all $\beta' \in [\beta - \Delta, \beta + \Delta]$, the optimal policy for $\mathrm{EntRM}_{\beta'}$ remains π^*_β.

The first bound is relevant when the risk parameter is not too small, as it scales with β. For large values of β, it balances the small optimality gap (remember that $\mathrm{EntRM}_\beta[R^\pi] \to \operatorname{ess\,inf} R^\pi$ when $\beta \to -\infty$, so the gaps tend to 0). The degeneracy at $\beta = 0$ is circumvented by the second bound.

Knowing this structure of intervals, the only information we need to determine the optimality front is the location of these subinterval boundaries. We call *breakpoints* these risk-parameter values at which the optimal policy changes, and we show that the resulting change is generally only local (see Section D.4 for the proof and more details).

Proposition 4.4. *Consider a random MDP with reward functions and transitions $(r_t(x,a))_{t,x,a}$ and $(p_t(x|a,x'))_{t,x,a,x'}$ generated from, say, independent uniform distributions. With probability 1, for each breakpoint $\beta \in \mathcal{B}$ there is a single state-horizon pair for which the optimal action changes: if π^1 is optimal for $\beta \in [\beta_1, \beta_2]$ and π^2 is optimal for $\beta \in [\beta_2, \beta_3]$, then there exists a unique state x and time step t such that $\pi^1_t(x) \neq \pi^2_t(x)$.*

4.2 Computing the Optimality Front

The first step towards computing the Optimality Front is to *find the breakpoints*. We first show that a direct approach is not feasible but instead we can exploit Theorem 4.3. Then, we present our algorithm, *Distributional Optimality Front Iteration* (DOLFIN), based on efficient (distributional) value iteration [10].

Finding the Breakpoints. Breakpoints mark the transition between two different intervals of optimality. According to Theorem 4.2 there must be at least two optimal policies for those particular parameter values. This yields a system of equations characterizing the breakpoints.

Proposition 4.5. *Assume π^1 and π^2 are such that π^1 is optimal for $\beta \in [\beta_1, \beta_b]$ and π^2 is optimal for $\beta \in [\beta_b, \beta_2]$. Then β_b satisfies:*

$$\forall h, x, \quad \mathrm{EntRM}_{\beta_b}[R^{\pi^1}_h(x)] \;=\; \mathrm{EntRM}_{\beta_b}[R^{\pi^2}_h(x)]$$

.

Note that since the policies only differ in one state-horizon pair (Prop. 4.4), most of these equations are trivial. Nevertheless, one could attempt to use them to directly compute the breakpoints by resolving a system of equations. Unfortunately, we argue in Section E that this is unfeasible due to the lack of regularity of the EntRM functions and to the computational complexity of the problem. In general, we show that the lack of "regularity" in these functions makes it impossible in practice to know in advance how many breakpoints, or optimal policies, might appear between two given points.

However, the advantage of using the EntRM in MDPs is that the optimal policies can be computed recursively by Dynamic Programming and we show that this implies a structure on the breakpoints.

Proposition 4.6. *(Informal) Let $\mathcal{B}^h$ be the set of breakpoints at time h and $\mathcal{B}^h(x)$ the corresponding set of breakpoints when starting at state x. Then, the sets verify the recursive equation*

$$\mathcal{B}^h = \mathcal{B}^{h+1} \cup \left(\bigcup_{x \in \mathcal{X}} \mathcal{B}^h(x) \right)$$

.

A formal statement and the proof can be found in Section D.7. This result shows that the set of breakpoints is simply a union over the per-state breakpoints, and they can be computed via a backward recursion.

We now have all the tools to build an incremental approach that we call *FindBreaks* and a detailed pseudo-code is in Section G but we give the intuition here. Theorem 4.3 tells us that knowing the

Entropic risk at a given point allows us to identify an interval of β values over which the optimal policy does not change. This fact can be utilized to 'jump' over β values. The process is the following. At a given state, assume the distribution of the return for each action is known $(\eta(x,a))_{a\in\mathcal{A}}$. Start with $\beta = 0$ and iterate the following steps:

1. Evaluate the Generalized Advantage function and the optimality gaps (see Theorem 4.3),
2. Use the optimality gaps to get a lower bound $\beta - \Delta$ on the next breakpoint, and 'jump': $\beta \leftarrow \beta - \Delta$

At some point, when getting close to a breakpoint, the increments Δ will get closer to 0. Then use a minimal increment ϵ until the optimal action changes. This can be done in parallel over states thanks to Theorem 4.6.

This may not be the optimal way to compute the breakpoints but it exploits all the structure of the problem: both the regularity of the exponential functions and the recursive properties of the MDP optimization allow to speed up the process. The general question of characterizing optimality for this problem is a challenging open problem we leave for future work. Our final risk-sensitive optimization algorithm below is fully modular and could integrate any other breakpoint-search algorithm.

Distributional Optimality Front Iteration. Combining both insights from Dynamic Programming and the approximation of Optimality Intervals, we derive an algorithm to compute the Optimality Front up to a desired accuracy.

This algorithm keeps in memory the distribution of the return recursively. While not compulsory, it accelerates the computation of several values of the EntRM with same reward distribution. For more details on the Dynamic Programming computation of return distributions, see Bellemare et al. [10].

Algorithm 1 DOLFIN- **D**istributional **O**ptima**l**ity **F**ront **I**teratio**n**

Require: Precision $\varepsilon \in (0,1)$; MDP $\mathcal{M}(\mathcal{X},\mathcal{A},p,R,H)$ parameters.

Select lower bound $\beta_{\min}$ ▷ Computed or handpicked
$\mathcal{I}_H \leftarrow [\beta_{\min}, 0]$ ▷ Starting interval
$\nu_H(x) \leftarrow \delta_0$ ▷ Optimal return distribution at timestep H
for $h \leftarrow H$ **to** 1 **do**
  for $x \in \mathcal{X}$ **do**
    for $I \in \mathcal{I}_h$ **do**
      $\eta_h^I(x,a) \leftarrow \varrho(x,a) * \sum_{x'} p(x' \mid x,a)\, \nu_{h+1}^I(x')$ ▷ Return distributions
      $\{\mathcal{J}, (a_j^*)_{j\in\mathcal{J}}\} \leftarrow$ FINDBREAKS$\big(\epsilon,\ (\eta_h^I(x,a))_a, I\big)$ ▷ Apply Algorithm 2 on $(\eta_h^I(x,a))_a$ as the reward distribution for each action
      for $j \in \mathcal{J}$ **do**
        Add j to $\mathcal{I}_{h-1}$ ▷ Update intervals for next timestep
        $\nu_h^j(x) \leftarrow \eta_h^I\big(x, a_j^*\big)$ ▷ Store optimal return distribution
Output $\Gamma = (\pi_k, I_k)_k\,, (\eta_0^k)^k$ ▷ Optimality Front, distributions

DOLFIN outputs the Optimality Front Γ as well as the return distributions of each optimal policy and it remains to solve (7): $\max_k \rho(\pi_k)$, following the Generalized Policy Improvement principle discussed above. In the experimental section below, we simply call this combined optimization the *Optimality Front* method.

About the computational cost. Calling B the number of breakpoints in the optimality front of the MDP, the number of calls to FindBreaks is bounded by $\Theta(|\mathcal{X}|HB)$ and thus heavily depend on the number of optimal policies. For a low number, only a few calls will be made and only a few Q-value evaluations will have to be computed. Using the empirical observations on the number of breakpoints, the number of calls to FindBreaks can be estimated to be in the order of $(|\mathcal{X}|H)^2$. The total complexity ultimately depends on the complexity of FindBreaks. An empirical study of our implementation can be found in Section G. In practice we observe that FindBreaks runs in $O(|A|f(1/\varepsilon))$ time, with some sublinear f. When the support of the accumulated reward is too large, the implementation of the distributional induction can benefit from the approximation schemes described in [10].

(a) Evaluation of $P(R^\pi \leq T)$ $(\downarrow)$

T/μ^*	0.25	0.33
Optimality Front	$\mathbf{1.26e^{-5}}$	$\mathbf{8.40e^{-5}}$
Proxy Optimization	$2.33e^{-5}$	$1.18e^{-4}$
Risk neutral optimal	$1.11e^{-4}$	$4.24e^{-4}$
Nested Risk Meas.	$1.54e^{-3}$	$8.37e^{-3}$
Optimal value	$6.29e^{-7}$	$8.22e^{-6}$

(b) Evaluation of $(\mathrm{C})\mathrm{VaR}_\alpha[R^\pi]$ $(\uparrow)$

Risk Measure	VaR		CVaR	
Risk parameter α	0.05	0.1	0.05	0.1
Optimality Front	**1.25**	**1.33**	**1.14**	**1.21**
Proxy Optimization	1.22	1.30	1.13	1.20
Risk neutral optimal	1.22	**1.33**	1.11	1.19
Nested Risk Measure	0.88	0.95	0.75	0.84

Table 1: Comparison of our two evaluation metrics under different policies.

5 Numerical Experiments

This section evaluates the effectiveness and efficiency of the Optimality Front approach on simple risk-sensitive planning tasks. Our goal is primarily to demonstrate the performance of our method compared to existing baselines mentioned earlier, namely the EVaR optimization approach of Hau et al. [23] (*Proxy Optimization* thereafter) and the *Nested Risk Measure* [7] implementing an approximate (invalid) value iteration algorithm. All the experiments in this section are of relative small scale and run on a standard laptop in a few seconds.

Environment. We test our method on two settings, the Inventory Management MDP [11, 38], which is a standard model for an important logistics problem, and the Cliff environment, that we mainly use to visualize the optimality front test the empirical efficiency of *Find Breaks* compared to a naive grid-based discretization.

In the Inventory Management problem, the goal is to maximize the profit of a store selling one extensive good. The store has a strict maximal capacity of $M = 10$. At each time step, the state of the store is its number of available goods, $x_t \in [M]$, and it can buy (action) a quantity $a_t \in [M]$ of new goods. The reward obtained is the profit minus the costs: $r_t = [f(D_t, x_t, a_t) - C_m(x_t) - C_c(a_t)]/4M$, where D_t is the random demand modeled by a binomial $D_t \sim B(0.5, M)$, $f(D_t, x_t, a_t) = 4\min(D_t, x_t + a_t)$ is the sales profit, $C_m(x_t) = 1x_t$ is the maintenance cost, and $C_c(a_t) = 3 + 2a_t$ is the order cost. We considered a horizon $H = 10$ with $x_1 = 0$. Optimal policies in the Inventory Management MDP can be parametrized by two thresholds (s_t, S_t) [38]: at time t, if the stock x_t is less than s_t, then agent should buy goods so that they have a stock of exactly S_t, i.e. $a_t^* = S_t - x_t$.

Optimality Front For our method, *Optimality Front*, we executed DOLFIN only once, with accuracy $\varepsilon = 10^{-2}$ on the chosen environment. It outputs a set of return distribution for all EntRM-optimal policies. We then computed all the metrics following Equation (7) by applying the functional of choice on all returned optimal distributions and selecting the optimal value. For the Inventory Management problem described above, with our choice of rewards and costs, the Optimality Front contained 18 different optimal policies.

5.1 Threshold Probability

We first optimize policies to minimize the probability that the return falls below a threshold $T < \mu^*$ where μ^* is the optimal mean return. For this problem, we were able to compute the non-Markovian optimal policies via Dynamic Programming on the augmented state space. On Table 1 (a), we report the estimated value of the probability of falling below the imposed threshold. We observe that using the Optimality Front method outperforms the risk-neutral optimal policy by up to a factor 10. On the other hand, *Nested Risk Measure* fails to find a good policy and performs worse than the risk-neutral one. While *Proxy Optimization* achieves reasonable results, this experiment shows that Generalized Policy Optimization on the Optimality Front performs better.

The true optimal value here is significantly better than what any Markovian policy can achieve, especially for low thresholds. This is due to the density and high randomness of the reward that strongly affect the distance to the threshold at each step. In goal-oriented MDPs, with scarce reward, this gap mostly vanishes (ex. see Cliff in Section H).

5.2 Value at Risk family

In this second experiment, we maximize lower quantiles of the return, aiming for policies with thin tails again, but through a different optimization objective. Hau et al. [23] already showed that optimizing over the EVaR performed better than previous methods to optimize VaR or CVaR, including using augmented MDPs, so we do not compute the optimal augmented-state DP policies for this second experiment and directly compare to *Proxy optimization*. We observe in Table 1b that for both VaR and CVaR, DOLFIN achieves slightly better or comparable performance to the state-of-the-art *Proxy Optimization*.

5.3 Optimality Front and sample efficiency

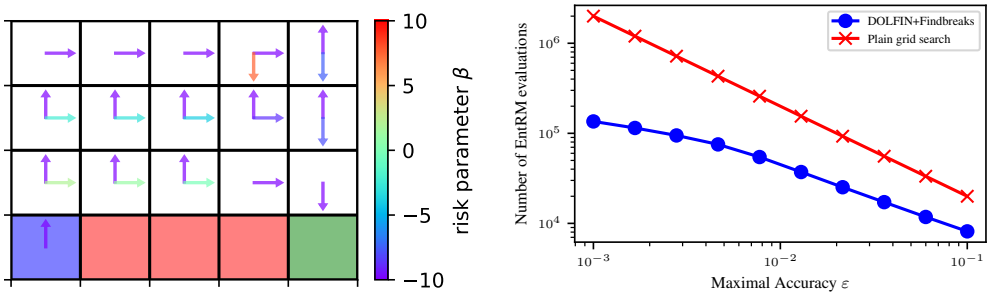

Figure 1: (Left) Illustration of all cliffs β-optimal policies. Arrows are of the color of the smallest β value corresponding to the optimal action $[\pi^*_\beta]$. (Right) Number of evaluations of EntRM required to obtain an accuracy ε on the breakpoints using our method (blue) vs. a naive grid (red). The efficiency ratio goes up to 15.

We illustrate the Optimality Front on the classic Cliff environment [41] (Figure 1 Left). As the risk parameter β grows on a continuous scale from $[-10, 10]$, the optimal policy shifts on a finite number of breakpoints reflected in the color of the corresponding arrows. For each β evaluated, we compute the optimal policy π^*_β (5) by Dynamic Programming. The Optimality Front approach relies on *FindBreaks* to avoid computing the optimal policy on a naive grid of β values. It is hard to theoretically bound the number of evaluations needed to find all the breakpoints, as it strongly depends on the structure of the MDP. We illustrate the computational gains on this simple example on Fig. 1 (Right), showing that *FindBreaks* can save up to an order of magnitude of evaluations.

6 Limitations and Discussions

Planning Setting. This paper only considers the *planning* setting, when the environment dynamics are fully known, and only small-scale problems are considered. The extension to more general settings is natural yet challenging because of the new problems raised. Optimizing the EntRM in the reinforcement learning setting has been studied before [29], but only for a fixed risk parameter, not a full range as studied here, so it is not clear how this can be done efficiently. Considering larger-scale environment would require using function approximations. In this case, computing the exact Optimality Front may not be tractable anymore, but the principle of Generalized Policy Improvement still holds. One idea could be to compute only a few policies associated to hand-picked risk parameters, and apply the same method. Those challenges are left for future work.

Approximation tightness. The approximations used in Section 3 are fully dependent on the return distribution and hard to control precisely in general. We are not aware of distribution-dependent measure of the tightness of EVaR, and it is unclear to us whether it is possible to prove any relevant error bounds for our *Optimality Front*.

Complexity of the algorithm. The theoretical complexity of DOLFIN is a function of the number of breakpoints. Those breakpoints arise from the MDP dynamics and their number cannot be simply expressed as a function of the main parameters of the MDP ($\mathcal{X}$, $\mathcal{A}$ or H). The only bound we can obtain is highly pessimistic and does not allow us to provide meaningful and problem-dependent complexity results. Empirical results show a clear gain on a toy problem and these gains only get bigger with the size of the MDP (see Section H).

7 Conclusion

We propose a unified framework for optimizing risk-sensitive objectives in Markov Decision Processes. Leveraging the computational advantages of the Entropic Risk Measure (EntRM), we provide an efficient algorithm for computing the optimality front, a family of policies which are optimal on a range of risk tolerance values. This also allows us to approximate key metrics such as Threshold Probabilities, Values at Risk and Conditional Values at Risk. Our algorithm demonstrates significant practical benefits in both efficiency and policy quality. This approach not only enhances risk-sensitive planning but also provides a versatile tool for tackling a variety of decision-making problems under uncertainty. It remains to be extended beyond planning in the context of learning unknown transition probabilities and reward distributions.

Acknowledgments

C. Vernade is funded by the Deutsche Forschungsgemeinschaft (DFG) under both the project 468806714 of the Emmy Noether Program and under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. CV also thanks the international Max Planck Research School for Intelligent Systems (IMPRS-IS). Aurélien Garivier and Alexandre Marthe thank the Chaire SeqALO (ANR-20-CHIA-0020-01) and PEPR IA project FOUNDRY (ANR-23-PEIA-0003).

References

[1] M. Achab and G. Neu. Robustness and risk management via distributional dynamic programming. *arXiv preprint arXiv:2112.15430*, 2021.

[2] A. Ahmadi-Javid. Entropic Value-at-Risk: A New Coherent Risk Measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, Dec. 2012. ISSN 1573-2878. doi: 10.1007/s10957-011-9968-2.

[3] A. Akritas, A. Strzebonski, and P. Vigklas. Improving the performance of the continued fractions method using new bounds of positive roots. *Nonlinear Analysis: Modelling and Control*, 13(3): 265–279, 2008.

[4] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

[5] M. J. Atallah. Some dynamic computational geometry problems. *Computers & Mathematics with Applications*, 11(12):1171–1181, 1985.

[6] A. Barreto, S. Hou, D. Borsa, D. Silver, and D. Precup. Fast reinforcement learning with generalized policy updates. *Proceedings of the National Academy of Sciences*, 117(48):30079–30087, 2020.

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

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

[9] N. Bäuerle, M. Pitera, and Ł. Stettner. Blackwell optimality and policy stability for long-run risk-sensitive stochastic control. *SIAM Journal on Control and Optimization*, 62(6):3172–3194, 2024.

[10] M. G. Bellemare, W. Dabney, and M. Rowland. *Distributional reinforcement learning*. MIT Press, 2023.

[11] R. Bellman, I. Glicksberg, and O. Gross. On the optimal inventory equation. *Management Science*, 2(1):83–104, 1955.

[12] A. Ben-Tal and M. Teboulle. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.

[13] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK ; New York, 2004. ISBN 978-0-521-83378-3.

[14] N. Bäuerle and U. Rieder. More Risk-Sensitive Markov Decision Processes. *Mathematics of Operations Research*, 39(1):105–120, Feb. 2014. ISSN 0364-765X. doi: 10.1287/moor.2013.0601. Publisher: INFORMS.

[15] A. L. B. Cauchy. *Exercices de mathématiques*, volume 3. De Bure Frères, 1828.

[16] A. Charpentier, R. Elie, and C. Remlinger. Reinforcement learning in economics and finance. *Computational Economics*, pages 1–38, 2021.

[17] Y. Chow and M. Ghavamzadeh. Algorithms for CVaR Optimization in MDPs. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[18] Y. Chow, A. Tamar, S. Mannor, and M. Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. *Advances in neural information processing systems*, 28, 2015.

[19] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(167):1–51, 2018.

[20] G. E. Collins. Polynomial minimum root separation. *Journal of Symbolic Computation*, 32(5): 467–473, 2001.

[21] R. Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[22] H. Föllmer and A. Schied. *Stochastic finance: an introduction in discrete time*. Walter de Gruyter, 2011.

[23] 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, 2023.

[24] J. L. Hau, E. Delage, M. Ghavamzadeh, and M. Petrik. On dynamic programming decompositions of static risk measures in markov decision processes. *Advances in Neural Information Processing Systems*, 36, 2024.

[25] R. A. Howard and J. E. Matheson. Risk-sensitive markov decision processes. *Management science*, 18(7):356–369, 1972.

[26] A. Kira, T. Ueno, and T. Fujita. Threshold probability of non-terminal type in finite horizon markov decision processes. *Journal of Mathematical Analysis and Applications*, 386(1):461–472, 2012.

[27] M. Kupper and W. Schachermayer. Representation results for law invariant time consistent functions. *Mathematics and Financial Economics*, 2(3):189–210, Sept. 2009. ISSN 1862-9660. doi: 10.1007/s11579-009-0019-9.

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

[29] H. Liang and Z.-Q. Luo. Bridging distributional and risk-sensitive reinforcement learning with provable regret bounds. *Journal of Machine Learning Research*, 25(221):1–56, 2024.

[30] O.-A. Maillard. Robust risk-averse stochastic multi-armed bandits. In *Algorithmic Learning Theory: 24th International Conference, ALT 2013, Singapore, October 6-9, 2013. Proceedings 24*, pages 218–233. Springer, 2013.

[31] A. Marthe, A. Garivier, and C. Vernade. Beyond average return in Markov Decision Processes. *Advances in Neural Information Processing Systems*, 36, 2024.

[32] P. Massart. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

[33] X. Ni and L. Lai. Evar optimization for risk-sensitive reinforcement learning. *IEEE Transactions on Information Theory*, 2022.

[34] A. S. Polydoros and L. Nalpantidis. Survey of model-based reinforcement learning: Applications on robotics. *Journal of Intelligent & Robotic Systems*, 86(2):153–173, 2017.

[35] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[36] R. T. Rockafellar, S. Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2: 21–42, 2000.

[37] M. Rowland, R. Dadashi, S. Kumar, R. Munos, M. G. Bellemare, and W. Dabney. Statistics and samples in distributional reinforcement learning. In *International Conference on Machine Learning*, pages 5528–5536. PMLR, 2019.

[38] H. Scarf, K. Arrow, S. Karlin, and P. Suppes. The optimality of (s, s) policies in the dynamic inventory problem. *Optimal pricing, inflation, and the cost of price adjustment*, pages 49–56, 1960.

[39] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550 (7676):354–359, 2017.

[40] X. Su, M. Petrik, and J. Grand-Clément. Evar optimization in mdps with total reward criterion. In *Seventeenth European Workshop on Reinforcement Learning*, 2024.

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

[42] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor. Policy gradient for coherent risk measures. *Advances in neural information processing systems*, 28, 2015.

[43] T. Tossavainen. The lost cousin of the fundamental theorem of algebra. *Mathematics Magazine*, 80(4):290–294, 2007.

[44] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[45] D. J. White. Minimizing a threshold probability in discounted markov decision processes. *Journal of mathematical analysis and applications*, 173(2):634–646, 1993.

[46] C. Wu and Y. Lin. Minimizing risk models in markov decision processes with policies depending on target values. *Journal of mathematical analysis and applications*, 231(1):47–67, 1999.

A EntRM supplements

In this section, we give more insights about the Entropic Risk Measure. We first recall the definition of the EntRM.

Let X be a random variable. The Entropic Risk $\mathrm{EntRM}_\beta[X]$ of parameter $\beta \in \mathbb{R}$ is defined as

$$\mathrm{EntRM}_\beta[X] = \begin{cases} \frac{1}{\beta}\log\big(\mathbb{E}[e^{\beta X}]\big), & \text{if } \beta \neq 0, \\ \mathbb{E}[X], & \text{if } \beta = 0. \end{cases}$$

Risk parameter interpretation. The parameter β of EntRM measures risk tolerance. Large (positive) values of β encourage a riskier behavior, as only the largest values taken by X contribute significantly to the expectation. Conversely, negative values of β promote a conservative behavior that minimizes potential losses, irrespective of the maximum reward. This behavior is illustrated by the values of the EntRM when the risk parameter approaches $\pm\infty$.

If the support of X is $(R_{\min}, R_{\max})$, then

$$\lim_{\beta\to-\infty} \mathrm{EntRM}_\beta[X] = R_{\min} \text{ and } \lim_{\beta\to+\infty} \mathrm{EntRM}_\beta[X] = R_{\max}.$$

When β approaches zero, the EntRM converges to the expected value, leading to risk-neutral behavior:

$$\mathrm{EntRM}_\beta(X) \underset{\beta\to 0}{=} \mathbb{E}[X] + \frac{\beta}{2}\mathbb{V}[X] + o(\beta).$$

The risk neutral behavior obtained for $\beta = 0$ is just the expectation, and the variance appears as the sensitivity of EntRM. For a Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, it holds exactly that $\mathrm{EntRM}_\beta(X) = \mu + \frac{\beta}{2}\sigma^2$.

Dynamic programming. Q-value function in the case of Entropic Risk Measure is defined as follows:

$$Q^\pi_{h,\beta}(x,a) = \frac{1}{\beta}\log\left(\mathbb{E}\left[\exp\left(\beta R^\pi_h(x,a)\right)\right]\right). \tag{8}$$

As shown in Howard and Matheson [25], this Q-value also satisfies a Bellman equation and thus it can be optimized efficiently by value iteration, and can result in a deterministic optimal policy.

The Q-value function for the entropic risk of parameter β satisfies the following optimal Bellman equations:

$$Q^\pi_{h,\beta}(x,a) = \mathrm{EntRM}_\beta[r_h(x,a)] + \mathrm{EntRM}_\beta\left[Q^\pi_{h+1,\beta}(X', \pi(X'))\right] \tag{9}$$

$$Q^*_{h,\beta}(x,a) = \mathrm{EntRM}_\beta[r_h(x,a)] + \mathrm{EntRM}_\beta\left[\max_{a'} Q^*_{h+1,\beta}(X', a')\right] \tag{10}$$

Where $X' \sim p(\cdot|x,a)$.

Principle of Optimality. Another implication of the Bellman equation is that the Entropic Risk Measure verifies the principle of Optimality, meaning there exists an optimal policy that is optimal for any subproblem:

$$\exists \pi^*,\ \forall x, h, \quad \pi^*_{h:H} \in \operatorname*{argmax}_\pi \mathrm{EntRM}_\beta[R^\pi(x)] \tag{11}$$

with $\pi_{h:H} = (\pi_h, \dots, \pi_H)$. This is in general not true for other risk measures, where a policy may only be optimal for a specific timestep and state, and can be suboptimal at intermediate timestep.

Exponential form While the EntRM is often defined as in Equation (1), the $\frac{1}{\beta}\ln$ serves solely as a renormalization factor. It does not influence the preference ordering between returns, and thus has no impact on the optimal policy in an MDP. For this reason, we also consider the exponential form $E_\beta[X] = \mathrm{sign}(\beta)\mathbb{E}_\pi[\exp(\beta X)]$ (also called *exponential utility*), which can appear in some applications, like for the Threshold Probability objective. If a policy π is optimal for EntRM_β, it is also optimal for E_β, and vice versa. In general:

$$\mathrm{EntRM}_\beta(X_1) > \mathrm{EntRM}_\beta(X_2) \iff \mathrm{sign}(\beta)\cdot\mathbb{E}[\exp(\beta X_1)] > \mathrm{sign}(\beta)\cdot\mathbb{E}[\exp(\beta X_2)]$$

and thus,

$$\operatorname*{argmax}_{\pi} \mathrm{EntRM}_{\beta}[R^{\pi}] = \operatorname*{argmax}_{\pi} \mathrm{sign}(\beta)\mathbb{E}_{\pi}[\exp(\beta R^{\pi})].$$

This equivalence is used at several occasions below.

B More on the Value at Risk family of risk measures

In this section, we provide additional details on the *Value at Risk* (VaR), *Conditional Value at Risk* (CVaR), and *Entropic Value at Risk* (EVaR) measures discussed in the main text. Two main points are covered: (1) a clarification of conventions and notation, and (2) the Derivation of EVaR.

B.1 Conventions and Notation

In the finance literature, VaR is generally introduced as the $\big(1-\alpha\big)$-quantile of a *loss* distribution, referring to a worst-case tail. It can also be defined via an α-quantile for *gains* (or returns), leading to slightly different formulas or sign conventions. For instance, one would usually write

$$\Pr\big(X \geq \mathrm{VaR}_{1-\alpha}[X]\big) = \alpha. \quad \text{while others use} \quad \Pr\big(X \leq \mathrm{VaR}_{\alpha}[X]\big) = \alpha,$$

Both definitions capture the idea of a critical quantile but from opposite sides of the distribution (gains vs. losses).

Throughout this work, we consistently adopt the low-tail perspective, defining

$$\mathrm{VaR}_{\alpha}[X] \;=\; \inf\big\{\; x \;\big|\; \Pr(X \leq x) \;\geq\; \alpha\big\},$$

so that VaR at level α is simply the α-quantile of X from below. We focus on this definition to remain consistent with our usage of "negative outcomes" or "low returns" as the main source of risk. Should the convention change, one can substitute $X \leftarrow -X$ to convert between definitions without altering the mathematical properties.

To follow the change of convention, we also define the *Conditional Value at Risk* (CVaR) as

$$\mathrm{CVaR}_{\alpha}[X] \;=\; \frac{1}{\alpha}\int_0^{\alpha} \mathrm{VaR}_{\gamma}[X]\, d\gamma. \quad \text{instead of} \quad \mathrm{CVaR}_{1-\alpha}[X] \;=\; \frac{1}{\alpha}\int_0^{\alpha} \mathrm{VaR}_{1-\gamma}[X]\, d\gamma.$$

and the *Entropic Value at Risk* (EVaR) as

$$\mathrm{EVaR}_{\alpha}[X] \;=\; \sup_{\beta<0}\left\{\frac{1}{\beta}\ln \mathbb{E}[e^{\beta X}] - \frac{1}{\beta}\ln(\alpha)\right\} \quad \text{instead of} \quad \inf_{\beta>0}\left\{\frac{1}{\beta}\ln \mathbb{E}[e^{\beta X}] - \frac{1}{\beta}\ln(1-\alpha)\right\}.$$

B.2 Entropic Value at Risk (EVaR)

Entropic Value at Risk (EVaR) was proposed as a tighter exponential-based bound on VaR and CVaR (2). Its derivation stems from the fact that exponential moment bounding can yield a close approximation to quantile-based measures.

Proof. By Chernoff's inequality, for any $\beta < 0$ we have:

$$\Pr\big(X \leq \ell\big) \;\leq\; \mathbb{E}\big[e^{\beta X}\big]\exp\big(-\beta\,\ell\big).$$

Solving the equation $\mathbb{E}\big[e^{\beta X}\big]\exp\big(-\beta\,\ell\big) \;=\; \alpha$ for ℓ, we obtain

$$\ell \;=\; a_X(\beta,\alpha) \;=\; \frac{1}{\beta}\,\ln\big(\mathbb{E}[e^{\beta X}]\big) \;-\; \frac{1}{\beta}\ln(\alpha).$$

Hence, for any $\beta < 0$ and $\alpha \in (0,1]$, the inequality

$$\Pr\big(X \leq a_X(\beta,\alpha)\big) \;\leq\; \alpha$$

implies

$$a_X(\beta,\alpha) \;\leq\; \mathrm{VaR}_{\alpha}(X).$$

Hence,

$$\mathrm{EVaR}_{\alpha}(X) \;=\; \sup_{\beta<0} a_X\big(\beta,\alpha\big) \;\leq\; \mathrm{VaR}_{\alpha}(X).$$

□

C Details on Section 3

In this section we provide additional details on the proxy method. We explain : (1) how to derive the problemes, (2) how to optmize them using a grid-based method, and (3) how to optimize them using the optimality front.

C.1 Derivation of Proxy Problems

We first detail how the proxy problems are derived from the original problem.

Probability Threshold Problem. The inital problem is

$$\min_{\pi} \ \Pr\big(R^{\pi} \ \leq \ T\big),$$

Using the Chernoff bound, and a few manipulations.

$$\begin{aligned}
\min_{\pi} \ \Pr\big(R^{\pi} \leq T\big) &\leq \min_{\pi} \ \min_{\beta<0} \mathbb{E}\big[e^{\beta(R^{\pi}-T)}\big] \\
&= \min_{\beta<0} \ \min_{\pi} \mathbb{E}\big[e^{\beta(R^{\pi}-T)}\big] \\
&= \min_{\beta<0} \mathbb{E}\big[e^{\beta\left(R^{\pi^*_\beta}-T\right)}\big].
\end{aligned}$$

Where the inversion between the minimum on the policy and the minimum on β is justified by the fact that there is only a finite number of policies.

Value at Risk Problem. The initial problem is

$$\max_{\pi} \ (\mathrm{C})\mathrm{VaR}_{\alpha}\big[R^{\pi}\big].$$

The similar manipulations lead to

$$\begin{aligned}
\max_{\pi} \ (\mathrm{C})\mathrm{VaR}_{\alpha}\big[R^{\pi}\big] &\geq \max_{\pi} \ \mathrm{EVaR}_{\alpha}\big[R^{\pi}\big] \\
&= \max_{\pi} \sup_{\beta<0} \mathrm{EntRM}_{\beta}\big[R^{\pi}\big] + \frac{1}{\beta}\ln(\alpha) \\
&= \sup_{\beta<0} \max_{\pi} \mathrm{EntRM}_{\beta}\big[R^{\pi}\big] + \frac{1}{\beta}\ln(\alpha) \\
&= \sup_{\beta<0} \mathrm{EntRM}_{\beta}\big[R^{\pi^*_\beta}\big] + \frac{1}{\beta}\ln(\alpha)
\end{aligned}$$

C.2 Grid-Based Optimization of the Risk Parameter

A natural first idea to optimize the EVaR in MDPs is to discretize β on a grid. By considering well-chosen values of β, Hau et al. [23] show that they can get an ε-approximation of the optimal policy for the EVaR problem with a complexity of $O\big(|\mathcal{S}|^2\,|\mathcal{A}|\,H\frac{\log(1/\varepsilon)}{\varepsilon^2}\big)$.

We prove a similar result for the Chernoff approximation of the Threshold Probability problem.

Proposition C.1. *Let R be the return of the MDP such that there exists $a<0$ and $p>0$ with the property that for any policy π, $\Pr(R \leq a) \geq p$. Then solving the proxy problem with accuracy $2\log(1+\varepsilon)/H$ on β and $\beta_{\min} = \ln(p)/a$, finds a policy π that satisfies*

$$\mathbb{P}\big(R^{\pi} \leq 0\big) \ \leq \ \tilde{B} \quad \textit{and} \quad B \ \leq \ \tilde{B} \ \leq \ (1+\varepsilon)\,B,$$

where B is the true optimal value.

Hence, one needs to compute $\frac{H\,\beta_{\min}}{2\,\log(1+\varepsilon)}$ policies to obtain a value within a factor $(1+\varepsilon)$ of the optimum. Given that computing an optimal policy for EntRM involves a complexity of $O\big(|\mathcal{S}|^2\,|\mathcal{A}|\,H\big)$, the overall complexity to achieve an approximation ratio of $(1+\varepsilon)$ is $O\Big(\frac{H^2\,|\mathcal{S}|^2\,|\mathcal{A}|\,\beta_{\min}}{\log(1+\varepsilon)}\Big)$.

Note that this result translates to any value of threshold other than 0, by just translating the rewards of the MDP so that the threshold becomes equivalent to 0.

As expected, those bounds show that the complexity explodes as the grid is refined to obtain more accurate approximations. However, this method does not use the knowledge about the structure of the Optimality Front, and computes the same optimal policies over and over again.

We show in Section 5 that DOLFIN computes the Optimality Front orders of magnitude more efficiently than computing on a grid. Conveniently, we can use this Optimality Front to then compute those relaxed optimization problems in a more efficient way.

Consider the EntRM optimal policies $\pi_1, \dots, \pi_K$ and the corresponding optimality intervals $I_1, \dots, I_K$. We have

$$\begin{aligned}\min_{\pi} \ \Pr\big(R^{\pi} \leq T\big) &\leq \min_{\beta<0} \mathbb{E}\big[e^{\beta(R^{\pi^*_\beta}-T)}\big] \\ &= \min_{k} \ \min_{\beta\in I_k} \mathbb{E}\big[e^{\beta(R^{\pi_k}-T)}\big].\end{aligned}$$

and

$$\begin{aligned}\max_{\pi} \ (\mathrm{C})\mathrm{VaR}_{\alpha}\big[R^{\pi}\big] &\geq \sup_{\beta<0} \mathrm{EntRM}_{\beta}\big[R^{\pi^*_\beta}\big] + \frac{1}{\beta}\ln(\alpha) \\ &= \max_{k} \sup_{\beta\in I_k} \mathrm{EntRM}_{\beta}\big[R^{\pi_k}\big] + \frac{1}{\beta}\ln(\alpha).\end{aligned}$$

These problems are reduced to more simple optimization problems on small intervals that can be solved using gradient methods. Indeed, the first one, for the Threshold Probability, is concave [13], and the second one is quasiconcave[5] [23].

C.3 Proof of Theorem C.1

We write $M_R(\beta) = \mathbb{E}[\exp(\beta R)]$ the moment generating function of the random variable R.

Lemma C.2. *Assume that R is bounded between $-H$ and H, then for a fixed stepsize $\epsilon > 0$, for $k \in \mathbb{N}$ and $\beta \in [\epsilon k, \epsilon(k+1)]$, noting $b_k = M_R(-\epsilon k)$, we have*

$$M_R(-\beta) \geq \exp(-\epsilon H) b_k \quad \textit{and} \quad M_R(-\beta) \geq \exp(\epsilon H) b_{k+1},$$

and,

$$\frac{M_R(-\beta)}{\min\{b_k, b_{k+1}\}} \geq \exp\left(\epsilon\frac{-H}{2}\right)\left(\frac{b_k}{b_{k+1}}\right)^{\frac{1}{2H}} \geq \exp\left(\epsilon\frac{-H}{2}\right),$$

otherwise the ratio is greater or equal to 1.

Proof. First we have

$$\begin{aligned}M_R(-\beta) &= \mathbb{E}[\exp(-(\beta-\epsilon k)R)\exp(-\epsilon k R)] \\ &\geq \exp(-(\beta-\epsilon k)H)\mathbb{E}[\exp(-\epsilon k R)] \geq \exp(-\epsilon H) b_k, \\ M_R(-\beta) &= \mathbb{E}[\exp((-\beta+\epsilon(k+1))R)\exp(-\epsilon(k+1)R)] \\ &\geq \exp(-(\epsilon(k+1)-\beta)H)\mathbb{E}[\exp(-\epsilon(k+1)R)] \geq \exp(-\epsilon H) b_{k+1}\ .\end{aligned}$$

To obtain the following inequality it is enough to consider the intersection of the functions $t \mapsto \exp(-tH)b_k$ and $t \mapsto \exp((-\epsilon - t)H)b_{k+1}$ on $t \in [0, \epsilon]$: as $M_R(-\beta)$ follows both left decreasing and right increasing constraint, the minimal value possible is on the intersection. □

We can now give a proof of Proposition C.1.

[5] It becomes a concave problem after using the change of variable $\beta \leftarrow \frac{1}{\beta}$ and thus can still be optimized efficiently

Proof. Let π_c and β_c be the policy and the β optimizing the Chernoff bound B. First, we can consider that $\beta_c < \ln(1/p)/a\}$. Indeed if not we can choose $\beta_c = 0$ instead:

$$\begin{aligned} M_R(\beta_c) &= \mathbb{E}[\exp(-\beta R)\mathbb{1}\{R \leq a\}] + \mathbb{E}[\exp(-\beta R)\mathbb{1}\{R > a\}] \\ &\geq \exp(-\beta_c a)\mathbb{P}[R \leq a] + \exp(-\beta M)\mathbb{P}[R \geq a] \\ &\geq \exp(-\beta_c a)p \geq 1 = M_R(0) \quad \text{using the hypothesis on } \beta_c \ . \end{aligned}$$

Let then $k \in \mathbb{N}$ be such that $\epsilon k \leq \beta_c < \epsilon(k+1)$, and suppose $M_{R^{\pi_c}}(-\epsilon k) \leq M_{R^{\pi_c}}(-\epsilon(k+1))$ without loss of generality. We have

$$\tilde{B} \leq \min_{\pi} M_{R^{\pi}}(-\epsilon k) \leq M_{R^{\pi_c}}(-\epsilon k) \leq \exp\left(\epsilon\frac{H}{2}\right) M_{R^{\pi_c}}(\beta_c) \leq \exp\left(\epsilon\frac{H}{2}\right) B,$$

using Theorem C.2.

For $\epsilon \leftarrow 2\log(1\varepsilon)/H$, we have that $\tilde{B} \leq (1+\varepsilon)B$. □

D Complements on Section 4

D.1 Optimality Front definition

Theorem 4.1 has a some shortcomings and the optimality front is not well defined this way. A first issue is that, for a specific β, the optimal policy may not be unique. There might be different possible choices, that could lead to also different optimality intervals. For the optimality front to be well defined, we would need unicity. A second issue comes from the fact that a policy might be optimal for a unique value of β, and not an interval of length > 0. To match with the definition of *breakpoints*, we would like to ensure that the optimality intervals are always of length > 0, and that two optimality intervals may only overlaps on their endpoints. For both those reasons, we introduce the following notions.

We say that two policies π_1 and π_2 are equivalent, noted $\pi_1 \sim \pi_2$ if the associated return distributions are equal : $\pi_1 \sim \pi_2 \iff R^{\pi_1} \overset{d}{=} R^{\pi_2}$. It is an *equivalence relation* (in the sense of mathematical binary relations) so it is possible to consider the quotient set $\bar{\Pi} = \Pi/\sim$ of equivalence classes. For $\bar{\pi} \in \bar{\Pi}$, the return distribution $R^{\bar{\pi}}$ is well defined. Considering the equivalence classes allows to avoid the issue of having several optimal policies on any given intervals of the risk parameter, instead, all such policies are in the same equivalence class $\bar{\pi}$.

For $\bar{\pi} \in \bar{\Pi}$, we consider $O_{\bar{\pi}} = \{\beta \in \mathbb{R} \text{ s.t. } \mathrm{EntRM}_\beta[R^{\bar{\pi}}] = \sup_\pi \mathrm{EntRM}_\beta[R^\pi]\}$ the set of values for which $\bar{\pi}$ is optimal, and $I_{\bar{\pi}} = \overline{\mathrm{int}\, O_{\bar{\pi}}}$ that set with its isolated points removed.

Definition D.1 (Optimality front)**.** The optimality front Γ is defined as $\Gamma = (\bar{\pi}_k, I_k)_{k\in[1,K]}$ where $I_k = I_{\bar{\pi}_k}$ and $(\bar{\pi}_k)$ is the set of class of policies such that $I_k \neq \emptyset$.

With such definition, the optimality front is uniquely defined, up to permutations. Computing it consists in computing *one* representative policy of each optimal class $\bar{\pi}_k$, and the corresponding optimality interval I_k.

D.2 Complements on Theorem 4.2 and definition of Breakpoints.

We first provide a more formal statement of the result, and then give a proof of it.

Proposition D.2. *The Optimality Front is unique, and finite. Also, $\forall k$, I_k is a finite union of closed intervals in $\mathbb{R}\bigcup\{-\infty\}$ with no isolated points, and $\bigcup_k I_k = \mathbb{R}^- \bigcup\{-\infty\}$. Finally, for $k_1 \neq k_2$, $I_{k_1} \cap I_{k_2} = \partial I_{k_1} \cap \partial I_{k_2}$.*

Proof of Theorem 4.2. Let us first show that the optimality front is well defined.

Let $\bar{\pi} \in \bar{\Pi}$. $\forall \pi_1, \pi_2 \in \bar{\pi}$, the distributions of the return are equal by definition: $R^{\pi_1} \overset{d}{=} R^{\pi_2}$. Hence, $R^{\bar{\pi}}$ is well defined. Similarly

$$O_{\pi_1} = \left\{\beta \in \mathbb{R} \text{ s.t. } \mathrm{EntRM}_\beta[R^{\pi_1}] = \sup_\pi \mathrm{EntRM}_\beta[R^\pi]\right\}$$
$$= \left\{\beta \in \mathbb{R} \text{ s.t. } \mathrm{EntRM}_\beta[R^{\pi_2}] = \sup_\pi \mathrm{EntRM}_\beta[R^\pi]\right\} = O_{\pi_2}$$

and so $O_{\bar{\pi}}$ (and consequently, $I_{\bar{\pi}}$) is well defined. Finally, the Optimality Front $\left\{(\bar{\pi}, I_{\bar{\pi}}) \mid \bar{\pi} \in \bar{\Pi}, I_{\bar{\pi}} \neq \emptyset\right\}$ is uniquely defined.

We now proceed to show the intervals part. In order to do so, we introduce the following result about analytic functions.

Lemma D.3. *Let f_1 and f_2 be two analytic functions on a finite interval $I \subset \mathbb{R}$. If f_1 and f_2 are equal on I, then they are equal everywhere. If they are not equal on I, then they can only intersect at a finite number of points in I.*

We also recall that $\bar{\Pi}$ is finite since the number of markov deterministic policies is finite in our finite horizon setting, and that optimizing on $\mathrm{EntRM}_\beta[R^\pi]$ is equivalent to optimizing on $-\mathbb{E}\left[\exp(\beta R^\pi)\right]$

Consider $f_{\bar{\pi}}(\beta) = -\mathbb{E}\left[\exp(\beta R^{\bar{\pi}})\right]$. Those functions are analytic as a convex combination of exponential functions, which are analytic.

Assume that there exist $\bar{\pi}_1 \neq \bar{\pi}_2$ such that $f_{\bar{\pi}_1}(\beta) = f_{\bar{\pi}_2}(\beta)$ for all β. The moment generating function defines uniquely the distribution of a random variable, so it implies that $R^{\bar{\pi}_1} \overset{d}{=} R^{\bar{\pi}_2}$. This contradicts the fact that $\bar{\pi}_1$ and $\bar{\pi}_2$ are different policies. Hence, the functions $f_{\bar{\pi}}$ are distinct pairwise.

Using Theorem D.8, we can restrict ourselves to only considering the risk parameter in the interval $[\beta_{\min}, 0]$. The functions $f_{\bar{\pi}}$ are analytic on this interval, and thus they can only intersect at a finite number of points. This implies that the set $\{\beta \in \mathbb{R}^- \mid \exists \bar{\pi}_1 \neq \bar{\pi}_2 \in \bar{\Pi}, f_{\bar{\pi}_1}(\beta) = f_{\bar{\pi}_2}(\beta)\}$ is finite. We write $\beta_0 < \beta_1 < \ldots < \beta_K$ the ordered set of those points. By definition, on each interval $]\beta_k, \beta_{k+1}[$, there is a unique $\bar{\pi}_k$ such that $f_{\bar{\pi}_k}(\beta) = \sup_{\bar{\pi}} f_{\bar{\pi}}(\beta)$, and thus that $\mathrm{EntRM}_\beta[R^{\bar{\pi}_k}] = \sup_{\bar{\pi}} \mathrm{EntRM}_\beta[R^{\bar{\pi}}]$. Also, by continuity, both π_k and π_{k+1} are optimal for β_{k+1}. Hence, consider $\bar{\pi} \in \bar{\Pi}$ and $k_1, \ldots, k_m$ the set of indices such that $\forall i, \bar{\pi}_{k_i} = \bar{\pi}$. Then, we have $I_{\bar{\pi}} \supset \bigcup_{i=1}^m [\beta_k, \beta_{k+1}]$. Furthermore, the only other values of β for which $\bar{\pi}$ can be optimal, are isolated points among $\beta_0, \ldots, \beta_K$. By definition, the intervals $I_{\bar{\pi}}$ do not include isolated points, so we have the equality $I_{\bar{\pi}} = \bigcup_{i=1}^m [\beta_k, \beta_{k+1}]$. Hence, the intervals $I_{\bar{\pi}}$ are finite unions of closed intervals in $\mathbb{R}^-$.

By construction, the intervals $I_{\bar{\pi}}$ are disjoint except for their endpoints, and the union of all the intervals is $\bigcup_{\bar{\pi}} I_{\bar{\pi}} = \mathbb{R}^-$, which concludes the proof of the proposition. □

Using this result, we can now formally define the breakpoints of the optimality front.

Definition D.4. The breakpoints of the optimality front are defined as the finite set of points $\bigcup_{\bar{\pi} \in \bar{\Pi}} \partial I_{\bar{\pi}}$. The breakpoints are the points where the optimality intervals $I_{\bar{\pi}_k}$ meet. We denote by $\mathcal{B} = \{\beta_1, \ldots, \beta_K\}$ the set of breakpoints.

D.3 Proof of Theorem 4.3.

We first start by deriving bounds on the growth of the EntRM when the risk parameter is changed slightly.

Proposition D.5. *Let X be a random variable, $\beta \in \mathbb{R}^-$, and $0 < \varepsilon < |\beta|$. We assume X is bounded in $[r_{\min}, r_{\max}]$. Then:*

$$\mathrm{EntRM}_\beta[X] \leq \mathrm{EntRM}_{\beta+\varepsilon}[X] \leq \frac{\beta}{\beta+\varepsilon}\mathrm{EntRM}_\beta[X] + \frac{\varepsilon}{\beta+\varepsilon} r_{\min}, \quad (12)$$

$$\frac{\beta}{\beta-\varepsilon}\mathrm{EntRM}_\beta[X] - \frac{\varepsilon}{\beta-\varepsilon} r_{\min} \leq \mathrm{EntRM}_{\beta-\varepsilon}[X] \leq \mathrm{EntRM}_\beta[X] \quad (13)$$

Proof. In the following, $0 < \varepsilon < |\beta|$. We write $X = \sum \mu_i \delta_{x_i}$, with $r_{\min} = \min_i x_i$.

$$\begin{aligned}
\mathrm{EntRM}_{\beta+\varepsilon}[X] &= \frac{1}{\beta+\varepsilon}\ln\left(\sum \mu_i e^{(\beta+\varepsilon)x_i}\right) \\
&= \frac{1}{\beta+\varepsilon}\ln\left(\sum \mu_i e^{\beta x_i} e^{\varepsilon x_i}\right) \\
&\leq \frac{1}{\beta+\varepsilon}\ln\left(e^{\varepsilon r_{\min}}\sum \mu_i e^{\beta x_i}\right) \quad (\text{since } \frac{1}{\beta+\varepsilon} < 0) \\
&= \frac{1}{\beta+\varepsilon}\ln\left(\sum \mu_i e^{\beta x_i}\right) + \frac{\varepsilon}{\beta+\varepsilon} r_{\min} \\
&= \frac{\beta}{\beta+\varepsilon}\mathrm{EntRM}_{\beta}[X] + \frac{\varepsilon}{\beta+\varepsilon} r_{\min},
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{EntRM}_{\beta-\varepsilon}[X] &= \frac{1}{\beta-\varepsilon}\ln\left(\sum \mu_i e^{(\beta-\varepsilon)x_i}\right) \\
&= \frac{1}{\beta-\varepsilon}\ln\left(\sum \mu_i e^{\beta x_i} e^{-\varepsilon x_i}\right) \\
&\geq \frac{1}{\beta-\varepsilon}\ln\left(e^{-\varepsilon r_{\min}}\sum \mu_i e^{\beta x_i}\right) \quad (\text{since } \frac{1}{\beta-\varepsilon} < 0) \\
&= \frac{1}{\beta-\varepsilon}\ln\left(\sum \mu_i e^{\beta x_i}\right) - \frac{\varepsilon}{\beta-\varepsilon} r_{\min} \\
&= \frac{\beta}{\beta-\varepsilon}\mathrm{EntRM}_{\beta}[X] - \frac{\varepsilon}{\beta-\varepsilon} r_{\min}.
\end{aligned}$$

The other sides of the inequalities are obtained using the monotony of the function $\beta \mapsto \mathrm{EntRM}_\beta[X]$ [2]. We assumed X to be a discrete random variable, but the proof can be extended to continuous random variables by using the Lebesgue integral instead of the sum. □

The next theorem is a special case of Theorem D.6 for the case of a single state. While not necessary to prove the main result, the result is used in the algorithm *FindBreaks*, see Section G.

Theorem D.6 (Interval of Action Optimality)**.** *Let $r_{\min}$ (resp. $r_{\max}$) be the minimum (resp. maximum) achievable reward, and $\Delta R = r_{\max} - r_{\min}$. Suppose we have actions $(a_{(i)})_i$ ordered so that*

$$U^1_\beta = \mathrm{EntRM}_\beta\big[R(a_{(1)})\big] \;>\; U^2_\beta = \mathrm{EntRM}_\beta\big[R(a_{(2)})\big] \;\geq\; \ldots \;\geq\; U^n_\beta = \mathrm{EntRM}_\beta\big[R(a_{(n)})\big].$$

In particular, $a_{(1)}$ is the unique optimal action and $a_{(2)}$ is the second-best. Define $\Delta U = U^1_\beta - U^2_\beta$. Then:

- *If $\beta \neq 0$, for all $\beta' \in \left[\beta\left(1 - \frac{\Delta U}{U_2 - r_{\min}}\right), \beta\left(1 + \frac{\Delta U}{U_1 - r_{\min}}\right)\right]$ and for all $i \geq 2$, we have*

$$\mathrm{EntRM}_{\beta'}\big[R(a_{(1)})\big] \;>\; \mathrm{EntRM}_{\beta'}\big[R(a_{(i)})\big].$$

- *If $\beta = 0$, then for all $\beta' \in \left[-\frac{8\,\Delta U}{\Delta R^2},\ \frac{8\,\Delta U}{\Delta R^2}\right]$ and all $i \geq 2$, we have*

$$\mathrm{EntRM}_{\beta'}\big[R(a_{(1)})\big] \;>\; \mathrm{EntRM}_{\beta'}\big[R(a_{(i)})\big].$$

In particular, the action $a_{(1)}$ remains strictly optimal for all β' in the specified range.

Proof of Theorem D.6. Assume $\beta \neq 0$. Let

By hypothesis, $U^1_\beta > U^2_\beta$. We aim to show that if β' remains within the specified range around β, action $a_{(1)}$ remains strictly optimal for $\mathrm{EntRM}_{\beta'}$.

Case $\beta' > \beta, \beta \neq 0$

Write $\beta' = \beta + \varepsilon$ with $\varepsilon > 0$. Assume

$$\varepsilon < \beta \, \frac{\Delta U}{r_{\min} - U_\beta^1}.$$

From the previously established bounds (see Theorem D.5), we know

$$U_\beta^1 \;\leq\; \mathrm{EntRM}_{\beta+\varepsilon}\big[R(a_{(1)})\big] \;=\; U_{\beta'}^1,$$

and

$$\mathrm{EntRM}_{\beta+\varepsilon}\big[R(a_{(2)})\big] \;=\; U_{\beta'}^2 \;\leq\; \frac{\beta}{\beta+\varepsilon}\, U_\beta^2 \;+\; \frac{\varepsilon}{\beta+\varepsilon}\, r_{\min}.$$

Thus, showing

$$U_\beta^1 \;>\; \frac{\beta}{\beta+\varepsilon}\, U_\beta^2 \;+\; \frac{\varepsilon}{\beta+\varepsilon}\, r_{\min}$$

implies $U_{\beta'}^1 > U_{\beta'}^2$. Rewriting, we get

$$\beta\big(U_\beta^1 - U_\beta^2\big) \;<\; \varepsilon\,\big(r_{\min} - U_\beta^1\big),$$

i.e.

$$\beta \, \frac{\Delta U}{r_{\min} - U_\beta^1} \;>\; \varepsilon.$$

The changes of sign in the inequality are due to the fact that $\beta + \varepsilon < 0$ and $U_\beta^1 > r_{\min}$.

This is exactly our assumption on ε. Hence $a_{(1)}$ remains strictly better than $a_{(2)}$ at $\beta' = \beta + \varepsilon$, and by extension, better than all other actions.

Case $\beta' < \beta, \beta \neq 0$**.**

The similar argument holds using the two inequalities

$$U_\beta^2 \;\geq\; \mathrm{EntRM}_{\beta-\varepsilon}\big[R(a_{(2)})\big] \;=\; U_{\beta'}^2,$$

and

$$\mathrm{EntRM}_{\beta-\varepsilon}\big[R(a_{(1)})\big] \;=\; U_{\beta'}^1 \;\geq\; \frac{\beta}{\beta-\varepsilon}\, U_\beta^1 \;-\; \frac{\varepsilon}{\beta-\varepsilon}\, r_{\min}.$$

Case $\beta = 0$**.**

This second part of Theorem D.6, when $\beta = 0$, uses *Hoeffding's lemma* (see e.g. Massart [32]):

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}[\exp(\lambda X)] \leq \exp\left(\lambda \mathbb{E}[X] + \frac{\lambda^2 \Delta R^2}{8}\right). \tag{14}$$

Hoeffding's lemma gives

$$\mathrm{EntRM}_\beta[X] \geq \mathbb{E}[X] - \beta \frac{\Delta R^2}{8}. \tag{15}$$

We can then proceed similarly as the previous proof. Consider $0 > \beta' > \frac{8\Delta U}{\Delta R^2}$. Using the previous equation and the fact that $U_\beta^1 \geq E[R(a_1)]$, we have

$$U_{\beta'}^1 - U_{\beta'}^2 \geq \mathbb{E}[R(a_1)] - \mathbb{E}[R(a_2)] - \beta' \frac{\Delta R^2}{8} \geq \Delta U - \frac{8\Delta U}{\Delta R^2} \frac{\Delta R^2}{8} = 0\,,$$

hence $U_\beta^1 > U_\beta^2$.

This concludes the proof of Theorem D.6. □

Finally, we can prove Theorem 4.3. We use the same principle as in the proof of Theorem D.6 and the principle of optimality (i.e. the optimal policy is always greedy with respect to itself) to show that the optimal policy remains the same. The inequalities from Theorem D.5 are adapted by replacing $r_{\max}$ by $H - h$ and $r_{\min}$ by $h - H$ (for timestep h, the return is in $[h - H, H - h]$ as the reward is bounded by $[-1, 1]$ at each step). The value $r_{\min} - U_\beta^1$ is replaced by its upper bound $2(H - h)$ to obtain a symetric bound and simplify the formula.

Proof. (of Theorem 4.3) Let us address the case where $\beta < 0$ and $\varepsilon > 0$. By the principle of optimality, we have:

$$\forall h, x, \quad \pi^*_{h,\beta}(x) = \arg\max_a \mathrm{EntRM}_\beta[R_h^{\pi^*}(x,a)] \ .$$

Let $\varepsilon < \Delta$ (with $\varepsilon < |\beta|$). We then obtain the inequalities:

$$\begin{aligned} \forall h, x, a, \quad & \mathrm{EntRM}_\beta[R_h^{\pi^*}(x,a)] \ \leq \ \mathrm{EntRM}_{\beta+\varepsilon}[R_h^{\pi^*}(x,a)] \\ \text{and} \quad & \mathrm{EntRM}_{\beta+\varepsilon}[R_h^{\pi^*}(x,a)] \ \leq \ \frac{\beta}{\beta+\varepsilon}\mathrm{EntRM}_\beta[R_h^{\pi^*}(x,a)] + \frac{\varepsilon}{\beta+\varepsilon}(H-h) \ . \end{aligned}$$

By the definition of ε, and following the same method as for Theorem D.6 we get $\forall h, x, \forall a \neq \pi^*_h(x)$,

$$\frac{\beta}{\beta+\varepsilon}\mathrm{EntRM}_\beta[R_h^{\pi^*}(x,a)] + \frac{\varepsilon}{\beta+\varepsilon}(H-H) \ \leq \ \mathrm{EntRM}_\beta[R_h^{\pi^*}(x,\pi^*_h(x))].$$

Therefore,

$$\begin{aligned} \mathrm{EntRM}_{\beta'}[R_h^{\pi^*}(x,a)] &\leq \frac{\beta}{\beta+\varepsilon}\mathrm{EntRM}_\beta[R_h^{\pi^*}(x,a)] + \frac{\varepsilon}{\beta+\varepsilon}(H-h) \\ &\leq \mathrm{EntRM}_\beta[R_h^{\pi^*}(x,\pi^*_h(x))] \\ &\leq \mathrm{EntRM}_{\beta'}[R_h^{\pi^*}(x,\pi^*_h(x))] \ . \end{aligned}$$

Thus, π^* remains greedy with respect to itself and remains the optimal policy.

The cases $\beta = 0$ and $\varepsilon < 0$ follow similarly, using the associated inequalities. □

D.4 About Theorem 4.4.

We first start with a lemma.

Lemma D.7. *Let Y be a random variable with continuous law. Let $X_1, \dots, X_n$ be random variables with Y independant from $(X_1, \dots, X_n)$. Then,*

$$\Pr\big(Y = f(X_1, \dots, X_n) \mid X_1, \dots, X_n\big) = 0$$

Proof. This is a special case of Exercise 2.1.5 in Durrett [21]. It comes down to writing the definition of conditional probabilities and computing the integrals with Fubini's theorem. □

Proof. (of Theorem 4.4)

Let π be a policy. Assume the probability transitions are fixed and only the rewards are random. We consider, for $x, h, a, a' \in \mathcal{X} \times [H] \times \mathcal{A} \times \mathcal{A}$ the set $\mathcal{B}^{x,h}_{a,a'}$ of parameters β such that $\mathrm{EntRM}_\beta[R_h^\pi(x,a)] = \mathrm{EntRM}_\beta[R_h^\pi(x,a')]$. We aim to show that the sets $(\mathcal{B}^{x,h}_{a,a'})_{x,h,a,a'}$ have no element in common pairwise, with probability one.

Consider orders on both $\mathcal{X}$ and $\mathcal{A}$, and consider the associated lexigocraphic order of $[H,1] \times \mathcal{X} \times \mathcal{A}$. We proceed by induction on this order.

Case $t = H$.

Let $x_1, x_2 \in \mathcal{X}$, $a_1, a_1', a_2, a_2' \in \mathcal{A}$, with $(x_1, a_1, a_1') \neq (x_2, a_2, a_2')$ (the triplets are different, but some elements of the triplets can be equal). Consider known $\mathcal{B}^{x_1,H}_{a_1,a_1'}$.

$$\begin{aligned} \forall \beta \in \mathbb{R}, \quad \Pr\Big(\mathrm{EntRM}_\beta[R_H^\pi(x_2,a_2)] = \mathrm{EntRM}_\beta[R_H^\pi(x_2,a_2')] \ \Big| \ r_H(x_1,a_1), r_H(x_1,a_1')\Big) \\ = \Pr\Big(r_H(x_2,a_2) = r_H(x_2,a_2') \ \Big| \ r_H(x_1,a_1), r_H(x_1,a_1')\Big) \\ = 0. \end{aligned}$$

Because $r_H(x_2,a_2)$ and $r_H(x_2,a_2')$ are continuous random variables and at least one of the two is not conditioned on, the probability that they are equal is zero according to Theorem D.7.

It is in particular true for all $\beta \in \mathcal{B}^{x_1,H}_{a_1,a_1'}$. Using the union bound, we get that $\mathcal{B}^{x_1,H}_{a_1,a_1'}$ and $\mathcal{B}^{x_2,H}_{a_2,a_2'}$ have no element in common with probability one.

By considering elements in order and conditioning on the previously *observed* rewards, the induction is verified for $t = H$.

Case $t < H$.

Let $u = h_0, x_0, a_0$. Consider all breakpoints observed before

$$\mathcal{B} = \bigcup_{\substack{(h,x,a)<u \\ (h,x,a')\leq u}} \mathcal{B}^{x,h}_{a,a'}.$$

By induction, all of them are disjoint pairwise with probability one.

$$\begin{aligned}
&\forall \beta \in \mathcal{B}, \quad \Pr\left(\mathrm{EntRM}_\beta[R^\pi_{h_0}(x_0,a_0)] = \mathrm{EntRM}_\beta[R^\pi_{h_0}(x_0,a_0')] \,\middle|\, \big(r_h(x,a)\big)_{(h,x,a)\leq u}\right) \\
&\quad = \Pr\left(r_{h_0}(x_0,a_0) + \mathrm{EntRM}_\beta[R^\pi_{h_0+1}(X_0)] = r_{h_0}(x_0,a_0') + \mathrm{EntRM}_\beta[R^\pi_{h_0+1}(X_0')] \,\middle|\, \big(r_h(x,a)\big)_{(h,x,a)\leq u}\right) \\
&\quad = \Pr\left(r_{h_0}(x_0,a_0) = f\left(\big(r_h(x,a)\big)_{(h,x,a)\leq u}\right) \,\middle|\, \big(r_h(x,a)\big)_{(h,x,a)\leq u}\right) \\
&\quad = 0.
\end{aligned}$$

Where $f\left(\big(r_h(x,a)\big)_{(h,x,a)\leq u}\right) = r_{h_0}(x_0,a_0') + \mathrm{EntRM}_\beta[R^\pi_{h_0+1}(X_0')] - \mathrm{EntRM}_\beta[R^\pi_{h_0+1}(X_0)]$ and using Theorem D.7. The induction is then proved.

As there is a finite number of policy, this result remains when considering the set of all policies.

To conclude, notice that a breakpoint β_0 is a value of risk parameter for which two different action a_1, a_2 have the same expected return for some state x and timestep h. Hence, $\beta_0 \in \mathcal{B}^{x,h}_{a_1,a_2}$. With probability one, those set are pairwise disjoint, meaning a single action changes in β_0.

As this proof works for any value of the probability transitions, the result also remains for p random. □

The result is what we observe for all the MDPs used in practice. Yet it is simple to construct an MDP where several actions change at the same time, it is enough to have two states x_1, x_2 have the same transitions function and rewards : $\forall a, x', p(x'|x_1,a) = p(x'|x_2,a)$ and $r(x_1,a) = r(x_2,a)$.

D.5 Proof of Theorem 4.5.

Proof. (of Theorem 4.5) It all relies on the continuity of the function $\beta \to \mathrm{EntRM}_\beta[R^\pi]$.

$$\begin{aligned}
\forall h, x, \quad &\mathrm{EntRM}_{\beta_1}[R^{\pi^1}_h(x)] \geq \mathrm{EntRM}_{\beta_1}[R^{\pi^2}_h(x)] \\
&\mathrm{EntRM}_{\beta_3}[R^{\pi^1}_h(x)] \leq \mathrm{EntRM}_{\beta_3}[R^{\pi^2}_h(x)]
\end{aligned}$$

By continuity, there exists β_2 such that there is an equality, and, since we assume that there are no other optimal policy in $[\beta_1, \beta_3]$, the equality is verified for all x, h. □

D.6 On the lowest breakpoint.

Theorem D.8. *Consider $\pi_{\inf} = \lim_{\beta\to-\infty} \pi^*_\beta$. Consider the return R taking value in $x_1 \leq \ldots \leq x_n$. We write $a_i(\pi) = P(R^{\pi_{\inf}} = x_i) - \Pr(R^\pi = x_i)$. Then, the lowest breakpoint $\beta_{\inf}$ verifies*

$$\beta_{\inf} \geq \frac{-\log\left(1 + \max_{\pi\neq\pi_{\inf}} \max_{i>0} \frac{|a_i(\pi)|}{|a_0(\pi)|}\right)}{\min_{i\neq j} |x_i - x_j|}.$$

Proof. (of Theorem D.8)

We write $\Delta R = R_{\max} - R_{\min}$. We assume here for simplicity that all the rewards are evenly spaced on a grid. This mean that the values of the return are of the form $R_i = R_{\min} + i\frac{\Delta R}{n}$ for $i \in \{0, \dots, n\}$.

With this natural assumption, the problem of finding β such that,

$$\mathrm{EntRM}_\beta[R^\pi] = \mathrm{EntRM}_\beta[R^{\pi'}]$$

can be transformed into a problem of finding the roots of a polynomial.

$$\begin{aligned}
\mathrm{EntRM}_\beta\big[R^\pi\big] \;=\; \mathrm{EntRM}_\beta\big[R^{\pi'}\big] \;&\Longrightarrow\; \mathbb{E}\big[\exp(\beta R^\pi)\big] \;=\; \mathbb{E}\big[\exp(\beta R^{\pi'})\big] \\
&\iff \sum_{i=0}^{n} \mu_i \,\exp\big(\beta R_i\big) \;=\; \sum_{i=0}^{n} \mu'_i \,\exp\big(\beta R'_i\big) \\
&\iff \sum_{i=0}^{n} a_i \,\exp\big(\beta R_i\big) \;=\; 0 \\
&\iff \sum_{i=0}^{n} a_i \,\exp\Big(\beta\, i\,\frac{\Delta R}{n}\Big) \;=\; 0 \\
&\iff \sum_{i=0}^{n} a_i \, X^{-i} \;=\; 0\,. \\
&\iff \sum_{i=0}^{n} a_{n-i} \, X^{i} \;=\; 0\,.
\end{aligned}$$

Where $X = \exp\Big(-\beta\,\frac{\Delta R}{n}\Big)$. The first implication is not an equivalence because of the case $\beta = 0$, where the exponential form (rhs) is always equal to 1 and thus the equality is always verified. The last implication is verified by multiplying by X^n because of that same fact that 0 is already a root of the equation.

Hence, if X is a solution, β verifies $\beta = -\frac{\log(X)}{\frac{\Delta R}{n}}$.

Cauchy's bound on the size of the largest polynomial root [15] claims that the largest root verifies

$$1 + \max_{i>0} \frac{|a_i|}{|a_0|}.$$

The lowest breakpoints corresponds to the largest breakpoint solution between the $\pi_{\inf}$ and any other policy. Hence, the lowest breakpoint verifies

$$\beta_{\inf} \geq \frac{-\log\Big(1 + \max_{\pi \neq \pi_{\inf}} \max_i \frac{|a_i(\pi)|}{|a_0(\pi)|}\Big)}{\frac{\Delta R}{n}}\,.$$

□

This proof uses Cauchy's bound on the size of the largest polynomial root as it is simple to write and an efficient bound. Tighter bound have been developed in the litterature that could also be used here. See Akritas et al. [3] for example.

Using this polynomial formulation, it is also possible to derive a theoretical bound on the smallest distance between breakpoints. Mignotte's separation bound [20] gives a lower bound on the distance between two roots of the polynomial, as a function of the coefficient of such polynomial. This bound can then be transfered to a bound on the distance between breakpoints. This kind of bound could be useful to choose the necessary precision for the computation of the breakpoints, but are intractable to compute and the obtained values are too small to be relevant in practice.

D.7 About Theorem 4.6.

Proposition D.9 (Formal)**.** *Let $\mathcal{B}^h$ and Γ^h respectively be the set of breakpoints and the optimality front when the MDP starts at timestep h. Let $\mathcal{B}\left((X_i)_i, I\right)$ the set of breakpoints for the MDP with a single state and reward distributions $(X_i)_i$ (not assumed deterministic).*

$$\mathcal{B}^h = \mathcal{B}^{h+1} \cup \left(\bigcup_{x \in \mathcal{X}, (\pi_k^h, I_k^h) \in \Gamma^h} \mathcal{B}\left([R_h^{\pi_k^h}(x,a)]_a, I_k^h \right) \right)$$

Proof. Consider β_0 a breakpoint. There exists h such that $\beta_0 \in \mathcal{B}^h \setminus \mathcal{B}^{h+1}$. This h corresponds to the last timestep where the optimal policy changes when the risk parameter crosses β_0.

In particular, let $(\pi^I, I) \in \Gamma^{h+1}$ such that $\beta_0 \in I$. By definition,

$$\forall t \geq h+1, \forall \beta \in I, \quad [\pi^*_\beta]_t(x) = [\pi^I]_t(x).$$

Also, since β_0 is a breakpoint, there exists $x, a_1 \neq a_2, \pi_1 \neq \pi_2$ such that

$$\mathrm{EntRM}_{\beta_0}[R_h^{\pi^1}(x,a_1)] = \mathrm{EntRM}_{\beta_0}[R_h^{\pi^2}(x,a_2)],$$

By definition of π^I, it is also equal to

$$\mathrm{EntRM}_{\beta_0}[R_h^{\pi^I}(x,a_1)] = \mathrm{EntRM}_{\beta_0}[R_h^{\pi^I}(x,a_2)],$$

Hence, $\beta_0 \in \mathcal{B}\left([R_h^{\pi^I}(x,a)]_a, I \right)$.

□

E Solving the exact breakpoints

Several issues rises when trying to computing the Optimality front with this method. The first one is that there is no easy way to compute the function $\beta \mapsto Q^\pi_{h,\beta}(x, \pi(x))$ without having access to the exact distribution of return for this policy, or to perform a Policy Evaluation step for each value of β, which quickly becomes computationally inefficient.

A second issue is that this system of equation is only valid if there is a single breakpoint (i.e., no other optimal policy) between those two policies. Assume that we know π_1 is optimal for β_1 and π_2 is optimal for π_2, solving the equations gives value of the risk parameter for which one policy becomes better than the other. However, there could also be a third (or more) policy π_3 which is optimal for values of β between β_1 and β_2. Computing the optimality front would require computing the breakpoints between π_1 and π_3, and between π_3 and π_2. As there are no simple conditions to verify the existence of another optimal policy between two known one, solving the exact breakpoint for a policy π_1 optimal for β_1 would require to solve the system of equation for all possible policy π_2, and retrieving the lowest breakpoint obtained among them all. Selecting this lowest breakpoint β_2 ensures that π_1 is only optimal for the a risk parameter up to β_2. Because of the exponential number of possible policies, such method is intractable in practice, and some approximation will be required.

Lastly, the equations themselves are non-trivial. Indeed, knowing the distribution of rewards, such an equation takes the form $\sum_{i=1}^k a_i \, e^{\alpha_i \beta} = 0$, where the coefficient a_i may be non-positive. The function in question can be expressed as the difference of two convex functions, but it lacks many of the regularity properties needed for simple optimization methods. To the best of our knowledge, the best algorithms for solving such problems reduce it to finding roots of polynomial and using efficient solvers. Yet, the non-linear transformation required (see proof of Theorem D.8) makes the approximation error become significant when transposed back to our initial problem.

F Analysis of the number of breakpoints

F.1 Theoretical bound on the number of breakpoints

Proposition F.1. *Let n the number of possible values of R^π. The number of breakpoints B verifies*

$$B \leq n \cdot |\mathcal{A}|^{2|X|H}$$

Proof. (of Theorem F.1)

We first consider this lemma, on the number of roots of an exponential sum.

Lemma F.2 ([43]). *Let $f_n(x) = \sum_0^n b_i k_i^x$, $b_i \in \mathbb{R}, k_i > 0$. Then f_n has at most $n-1$ roots.*

This lemma allows for bounding the number of values of the risk parameter β for which two different policies have the same entropic risk.

Proposition F.3. *Consider two distributions $\mu_1 \neq \mu_2$ with support on $X = \{x_1, \dots, x_n\} \subset \mathbb{R}$ of size n. Consider $X_1 \sim \mu_1, X_2 \sim \mu_2$ random variables following those distributions. Consider $\mathcal{B} = \{\beta \in \mathbb{R} \mid \mathrm{EntRM}_\beta(X_1) = \mathrm{EntRM}_\beta(X_2)\}$.*

Then: $|\mathcal{B}| \leq n-1$

The proposition is straightforward by considering the exponential form of the EntRM, which is a sum of exponentials.

Finally we conclude by considering that there are $|\mathcal{A}|^{|X|H}$ deterministic markovian policies for a specific MDP, and thus $|\mathcal{A}|^{2|X|H}$ pairs of policies. The breakpoints are values for which the entropic risk of two policies is equal, and thus are included in the union of the "breakpoints" of all pairs of policies. By Theorem F.3, the number of breakpoints is at most $(n-1)|\mathcal{A}|^{2|X|H}$. □

This proposition hat no reason to be tight in general. In the conclusion we consider the number of point of equality between any two policies, but those points cannot all be breakpoints. Only the points where one of the two policy is optimal count as breakpoints. This problem is reduced to finding the number of components of the function $h(\beta) = \max\{f_\pi(\beta)\}_{\pi \in \Pi}$, where $f_\pi(\beta) = \mathrm{EntRM}_\beta[R^\pi]$. This combinatorial problem has been studied before (see Lemma 2.4 in Atallah [5]), but to the best of our knowledge, no better explicit formula has been found.

F.2 Evolution with the number of actions and atoms in the distributions

We conducted experiments in the single-state setting (the *FindBreaks* setting) to determine whether the theoretical bound on the number of breakpoints is reached in practice. We generated numerous random reward distributions and used our algorithm to find the number of breakpoints. We performed two experiments: one evaluating how the number of breakpoints evolves with respect to the number of actions, and the other with respect to the number of atoms in the reward distributions.

All the distributions considered have values in $[0, 1]$ with atoms evenly spaced over this interval. To generate these distributions, we treat each as an element of $[0, 1]^n$, where the sum of the elements equals 1, representing a point on the n-dimensional simplex. The distributions are thus generated by uniformly sampling a point on this simplex.

For the first experiment, we fixed the number of atoms to 10 and varied the number of actions from 5 to 50. For the second experiment, we fixed the number of actions to 10 and varied the number of atoms from 5 to 50 as well. For each action, the reward distribution was generated randomly as previously described. In the solving algorithm, we searched for breakpoints for β values in the range $[-15, 15]$ with a precision of 0.01. For each plot, we generated 100 independent problems and displayed a histogram of the number of breakpoints found across these 100 problems.

The results are shown in Figure 2. We observe that the number of breakpoints increases according to the studied parameters, with an average number of crossings of 1.25, 1.9, 2.19, and 2.36 for the first experiment, and 1.42, 1.66, 1.93, and 2.05 for the second. However, this increase is far from the theoretical bound established in Theorem F.1. Indeed, the number of breakpoints is much lower than the theoretical bound, showing sublinear growth. This experiment confirms the efficiency of our algorithm, whose complexity is strongly tied to the number of breakpoints.

G FindBreaks Algorithm

Here *FindBreaks* is designed to be able to handle values of the risk parameter both positive and negative. In practice, using DOLFIN, all the values will be non-positive. The values of the increments Δ come from Theorem D.6.

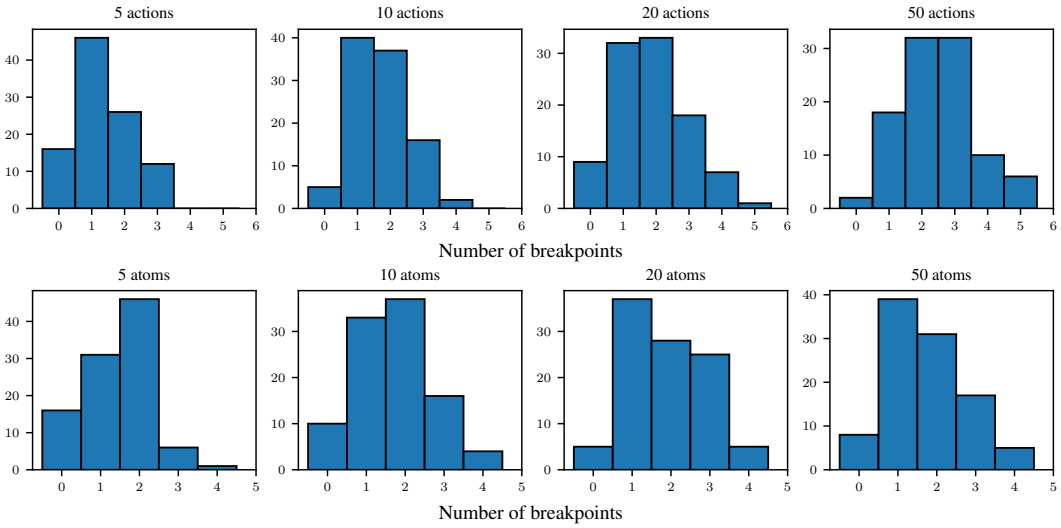

Figure 2: Illustration of the evolution of the number of breakpoints when the number of actions (Up) and atoms in the distributions (Down) is increasing. While the number of actions and atoms is multiplied by 10, the average number of breakpoints increases by less than a factor 2.

Algorithm 2 FindBreaks: Computing all optimal actions

Require: Precision $\varepsilon \in (0,1)$, Random rewards $(R(a_i))_i$, interval I.

Compute $a^* = \arg\max_a \mathrm{EntRM}_0(R(a))$ ▷ Initial optimal action

Compute $A = \min_{a_2 \neq a^*} \mathrm{EntRM}_0(R(a^*)) - \mathrm{EntRM}_0(R(a_2))$, $\Delta R = r_{\max} - r_{\min}$ ▷ Advantage function and range of rewards

Initialize $\beta = \frac{8\Delta\mu}{\Delta R^2}$, $\beta_{\text{old}} = 0$ ▷ Initial parameter values

Initialize $b_\ell = 0$, $a_{\beta_{\text{old}}} = a^*$

while $\beta_{\min} < \beta < \beta_{\max}$ **do**

  Compute the β risks $\mathrm{EntRM}_\beta(R(a))$ for each action a ▷ Evaluate risks for all actions

  $a_\beta = \arg\max_a \mathrm{EntRM}_\beta(R(a))$ ▷ Select the action with the highest utility

  if $a_\beta \neq a_{\beta_{\text{old}}}$ **then**

    Add $[b_\ell, \beta]$ to the interval set $\mathcal{I}$ and a_β to the optimality front Π^* ▷ Update intervals and front

    $b_\ell \leftarrow \beta$ ▷ Update the lower bound of the next interval

  $a_{\beta_{\text{old}}} \leftarrow a_\beta$ ▷ Update the last optimal action

  $\Delta U = \min_{a \neq a_\beta} \big(\mathrm{EntRM}_\beta(R(a_\beta)) - \mathrm{EntRM}_\beta(R(a))\big)$ ▷ Smallest optimality gap for non-optimal actions

  if $\beta < 0$ **then**

    $\beta \leftarrow \beta - \max\{\beta \frac{\Delta U}{\Delta R}, \varepsilon\}$ ▷ Decrease β for negative values

  if $\beta > 0$ **then**

    $\beta \leftarrow \beta + \max\{\beta \frac{\Delta U}{\Delta R}, \varepsilon\}$ ▷ Increase β for positive values

Add $[b_\ell, \beta_{\max}]$ to $\mathcal{I}$ ▷ Add the last interval to the set

Output Optimality Front Γ^*

Empirical Evaluation and Performance Analysis A simple simulation illustrates the behavior or Algorithm 2. We consider two actions, a_1 and a_2, with reward distributions $\varrho(a_1) = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ and $\varrho(a_2) = \frac{99}{100}\delta_0 + \frac{1}{100}\delta_2$. Action a_1 is better in expectation (lower risk) but action a_2 can achieve a higher reward with small probability and should only outperform action a_1 for large risk parameters. Figure 3 illustrates this: the functions $\mathrm{EntRM}(R(a))$ are plotted for[6] $\beta > 0$. As soon as the risk parameter β is large enough (specifically, $\beta > 3.9$), action a_2 becomes optimal.

[6]This experiment is design to test the transition to a risky action, so it is only relevant to observe the optimality front for $\beta > 0$.

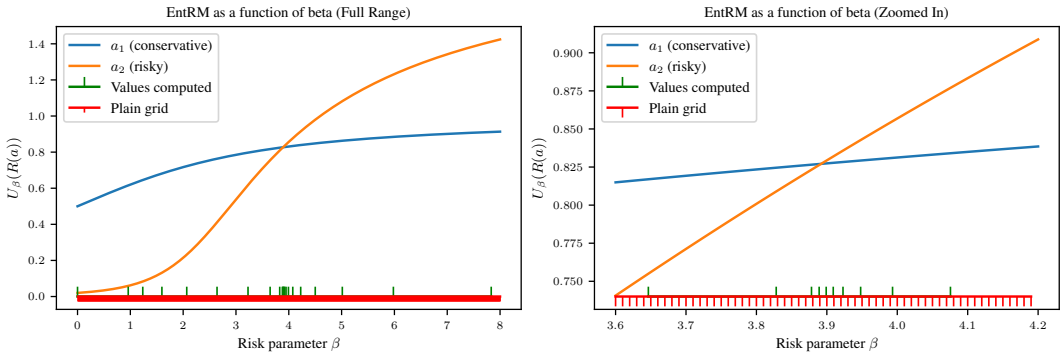

Figure 3: Utility functions on a single-state problem: a conservative (blue) and risky (red) actions, with respective ranges of optimality intersecting at $\beta_{\text{bp}} = 3.9 \pm 10^{-2}$. In red, we show the 800 values of β tested with a naive grid; in green the 22 values tested by Algorithm 2 to identify the breakpoint. Right-hand figure zooms around β_{bp}.

Algorithm 2 was executed on this example for $\beta \in [0, 8]$, with a precision of $\varepsilon = 10^{-2}$, but our theoretical upper bound[7] is $\beta_{\max} = \ln(100) = 4.6$. The green markers correspond to the values of β where the algorithm computed the EntRM, while the red markers represent the values that would be computed using a plain grid search with precision ε. As expected, we observe that the intervals shrink near the breakpoint, but grow significantly larger as we move away from these regions. Figure 3 (Right) zooms in on the interval $\beta \in [3.6, 4.2]$ to better visualize the concentration of intervals. Around $\beta = 3.9$, we observe that a few intervals are indeed capped by the maximal precision.

In this simple example, the naive grid uses 800 evaluations while Algorithm 2 only requires 22, with an *efficiency ratio* of $800/22 = 36$. To better quantify this gain, we run another experiment on random problems with 8 actions and reward function supported on 20 atoms in $[0, 1]$ (hence with possibly much more that 1 breakpoint). On Figure 4, for each level of precision $\varepsilon \in [10^{-3}, 10^{-1}]$, we compare the number of evaluations required by Algorithm 2 with the naive $1/\varepsilon$ obtained with a plain grid search. We report the gain on average over 20 random problem and notice efficiency ratios up to 35 for high-precision ε values.

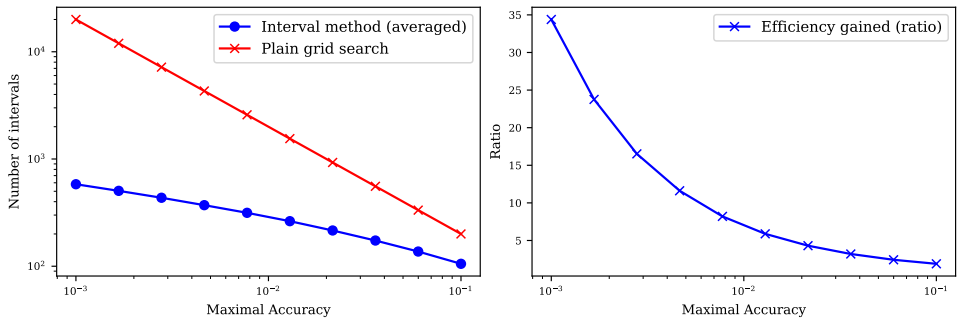

Figure 4: Performance gained using Algorithm 2. (Left) Number of evaluations, (Right) efficiency ratio (red over blue).

The performance of this algorithm is also closely tied to the number of breakpoints. The more breakpoints there are, the smaller the intervals will tend to be, which in turn reduces the algorithm's efficiency compared to a plain grid search. Therefore, the number of breakpoints is a critical factor. Section F show that the number of breakpoints grows much slower than the theoretical bound. This allows our algorithm to be more efficient in practice than predicted by theory.

[7]We show a larger upper bound for visualisation purposes. Our algorithm only evaluates 3 values past this limit so it does not hurt performance significantly.

H More Numerical Experiments

H.1 Inventory Management

We first consider some more values for the Inventory Management environments.

Table 2: Evaluation of $P(R^\pi \leq T)$ for Inventory Management.

T/μ^*	0.25	0.33	0.5	0.66	0.75
Optimality Front	$\mathbf{1.26e^{-5}}$	$\mathbf{8.40e^{-5}}$	$\mathbf{3.26e^{-3}}$	$\mathbf{3.91e^{-2}}$	$\mathbf{8.78e^{-2}}$
Proxy Optimization	$2.33e^{-5}$	$1.18e^{-4}$	$3.28e^{-3}$	$\mathbf{3.91e^{-2}}$	$\mathbf{8.78e^{-2}}$
Risk neutral optimal	$1.11e^{-4}$	$4.24e^{-4}$	$5.73e^{-3}$	$4.62e^{-2}$	$9.77e^{-2}$
Nested Prob. Thresh.	$1.54e^{-3}$	$8.37e^{-3}$	1	1	1
Optimal value	$6.71e^{-8}$	$6.29e^{-7}$	$4.48e^{-5}$	$7.85e^{-4}$	$3.59e^{-3}$

Table 3: Evaluation of $(\mathrm{C})\mathrm{VaR}_\alpha[R^\pi]$ for Inventory Management.

Risk Measure	VaR				CVaR			
Risk parameter α	0.05	0.1	0.2	0.5	0.05	0.1	0.2	0.5
Optimality Front	**1.25**	**1.33**	**1.45**	**1.65**	**1.14**	**1.21**	**1.30**	**1.45**
Proxy Optimization	1.22	1.30	**1.45**	**1.65**	1.13	1.20	1.30	**1.45**
Risk neutral optimal	1.22	**1.33**	1.43	**1.65**	1.11	1.19	1.28	1.44
Nested Risk Measure	0.88	0.95	1.05	1.28	0.75	0.84	0.95	1.12

The values in Table 2 and Table 3 reveal that the improvements of our *Optimality Front* method over the *Proxy Optimization* becomes less significant as the level of risks decreases.

We also plot in Figure 5 the efficiency ratio of using the Findbreaks, similar to Figure 4. This figure highlights the polynomial gain in performances of using DOLFIN and Findbreaks when one wants to compute all the optimal policies up to a certain accuracy.

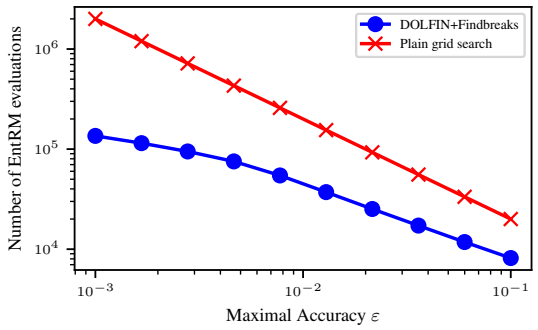

Figure 5: Performance gained using Algorithm 1 on Inventory Management.

H.2 Cliff Grid World

Here we consider the Cliff grid world [41] illustrated in Figure 1. The agent starts in the blue state. At each step, she has a small probability (0.1 here) of moving to another random direction. Due to these random transitions, it is risky to walk too close to the cliff (bottom, in red, negative reward $-\frac{1}{2}$), and conservative policies will prefer to walk further away to reach the goal (in green). The horizon is fixed at $H = 15$, so in principle, the agent has enough time to reach the end using the safe path. The reward when the goal is reached at step h is $1 - \frac{h}{2H}$, which encourages the agent to reach it as fast as possible.

The agent thus faces a dilemma: she could either walk along the cliff, risking to fall but reaching the goal faster and consequently receiving a higher reward; or she could go all the way around, taking

more time but with a lower probability to fall down. This can be observed in Figure 1, where the optimal policies for different values of β are shown. For high values of beta (e.g., $\beta = 10$), the agent takes the risky path, while for low values of beta (e.g., $\beta = -5$), the agent takes the safe path. One can even observe that for extremely negative values of β values (e.g. $\beta = -10$) the agent prefers to stay away from the cliff, not even trying to reach the goal (see purple arrow in top right corner pointing up).

The Cliff environment is very different from Inventory Management in its nature. The agent only receives rewards when a certain state is reached, making the reward scarce and the return distribution simpler. This implies that the optimal policy for different measures of risks, such as Probability Threshold, VaR and CVaR are markovian for this MDP.

For the Threshold Probability problem, we only consider 2 values of the threshold, -0.5 corresponding to falling into the cliff, and 0 corresponding to not reaching the goal. For the first threshold, the objective is to find the policy that is least likely to fall, while for the second it is to find the policy with the most chances of reaching the goal.

Table 1a confirms that our *Optimality Front* method performs better than other methods for the Threshold probability. An important remark is that, here, the real optimal value is reached by the optimality front. Similar performances are observed for the CVaR in Table 5. For the VaR, the gain in performance is limited, which is explained by the scarcity of rewards in the environment (the small changes in the return distribution does not change the value of the VaR).

Compared to the Inventory Management, the gain performance for computing the optimality front is much better, with a ratio up to a factor 100, as seen in Figure 6.

Table 4: Evaluation of $P(R^\pi \leq T)$ for Cliff.

T	-0.5	0
Optimality Front	$\mathbf{3.72e^{-2}}$	$\mathbf{4.65e^{-2}}$
Proxy Optimization	$3.85e^{-2}$	$4.67e^{-2}$
Risk neutral optimal	$4.50e^{-2}$	$4.84e^{-2}$
Nested Prob. Thresh.	1	1
Optimal value	$\mathbf{3.72e^{-2}}$	$\mathbf{4.65e^{-2}}$

Table 5: Evaluation of $(\mathrm{C})\mathrm{VaR}_\alpha[R^\pi]$ for Cliff.

Risk Measure	VaR				CVaR			
Risk parameter α	0.05	0.1	0.2	0.5	0.05	0.1	0.2	0.5
Optimality Front	**0.53**	**0.63**	**0.70**	**0.76**	**−0.37**	**0.11**	**0.38**	**0.58**
Proxy Optimization	0.00	0.6	**0.70**	**0.76**	−0.39	−0.08	0.38	0.58
Risk neutral optimal	**0.53**	**0.63**	**0.70**	**0.76**	−0.43	0.08	0.37	0.58
Nested Risk Measure	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5

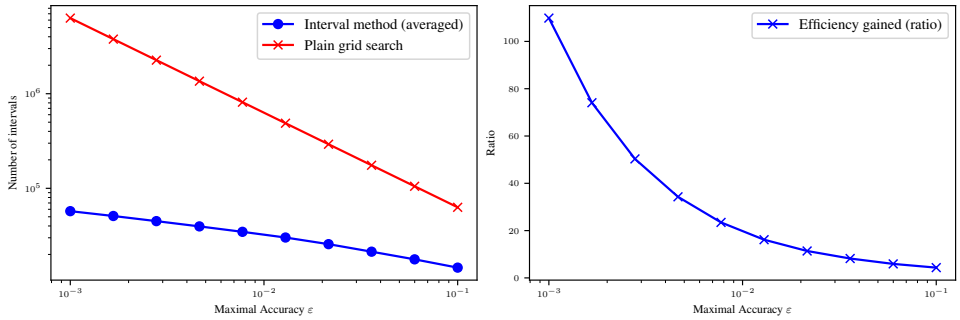

Figure 6: Performance gained using Algorithm 1 on Cliff. (Left) Number of evaluations, (Right) efficiency ratio (red over blue).