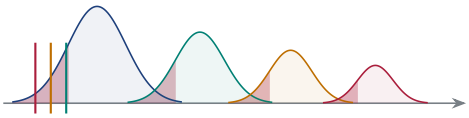
6

Risk Measures: from Chernoff to EVaR



A risk-sensitive agent summarizes a return distribution by a number that penalizes bad outcomes: a quantile, the average of the worst outcomes, a β-mean. Which of these summaries are sensible, and which can be optimized? This chapter introduces the axiomatic theory of coherent risk measures in the gain convention of the book, proves the two standard representations of the conditional value at risk, and builds the entropic value at risk (EVaR) directly from the Chernoff bound of Chapter 2: it is the tightest Chernoff lower bound on a quantile, it is coherent, and it is the worst-case expectation over all models within a fixed relative entropy of the nominal one; its mirror image for a risk-seeking agent, the best case over the same set of models, is the optimism principle of Part IV. Because EVaR is a supremum over β of the β-means that dynamic programming can optimize (Chapter 5), EVaR-optimal policies are obtained from a grid of exponential-utility plans, with an explicit accuracy guarantee. Most points of the grid are redundant: the β-optimal policies form a finite family, the optimality front*, which is computed once and then serves every risk measure by selection. The chapter ends with the tension between law invariance and time consistency, which the entropic family alone resolves.*

6.1 Value at risk and its defects

Throughout, X is a bounded random variable—the return of a finite-horizon MDP, in the applications—with distribution function F_X and lower quantile function $F_X^{-1}(u) = \inf\{x : F_X(x) \geq u\}$, and $\delta \in (0,1)$ is a *risk level*: small δ means a strong concern for bad outcomes. Rewards are gains, so bad outcomes are *small* values of X and risk measures are lower tail statistics, to be maximized. The loss convention of finance (where X is a loss and risk measures are upper tail statistics, to be minimized) is recovered by changing the sign of X. One more convention concerns the sign of β: in the sections on the value at risk and the entropic value at risk (Section 6.1 to Section 6.7), $\beta > 0$ denotes the *rate* of the risk-averse β-mean $U_{-\beta}$, as is customary for the entropic risk measure (and, in Section 6.5, the rate of the risk-seeking β-mean U_β), whereas Chapters 4 and 5, and the sections on the optimality front and on time consistency below (from Section 6.8 on), use the signed β of Definition 2.16, for which risk aversion means $\beta < 0$.

Definition 6.1 (Value at risk). The *value at risk* at level δ is the lower δ-quantile $\mathrm{VaR}_\delta(X) \coloneqq F_X^{-1}(\delta) = \inf\{x \in \mathbb{R} : \mathbb{P}(X \leq x) \geq \delta\}$.

Proposition 6.2 (Guarantee provided by the value at risk). $\mathbb{P}\big(X < \mathrm{VaR}_\delta(X)\big) \leq \delta \leq \mathbb{P}\big(X \leq \mathrm{VaR}_\delta(X)\big)$; in particular $\mathbb{P}(X \geq \mathrm{VaR}_\delta(X)) \geq 1-\delta$, and $\mathrm{VaR}_\delta(X)$ is the largest number with this property in the following sense: $\mathbb{P}(X \geq x) > 1-\delta$ implies $x \leq \mathrm{VaR}_\delta(X)$.

Proof. Let $q = \mathrm{VaR}_\delta(X)$. For $x < q$, $F_X(x) < \delta$ by definition of the infimum, so $\mathbb{P}(X < q) = \lim_{x \uparrow q} F_X(x) \le \delta$; and $F_X(q) \ge \delta$ because F_X is right-continuous, so that the infimum in Definition 6.1 is attained. If $\mathbb{P}(X \ge x) > 1 - \delta$ then $F_X(y) \le \mathbb{P}(X < x) < \delta$ for every $y < x$, so $x \le q$. ∎

The value at risk is thus the level that the return exceeds with probability at least $1 - \delta$: a guarantee, valid with confidence $1 - \delta$. It is monotone ($X \le Y$ a.s. implies $\mathrm{VaR}_\delta(X) \le \mathrm{VaR}_\delta(Y)$), translation invariant ($\mathrm{VaR}_\delta(X + c) = \mathrm{VaR}_\delta(X) + c$), positively homogeneous ($\mathrm{VaR}_\delta(\lambda X) = \lambda \mathrm{VaR}_\delta(X)$ for $\lambda \ge 0$), law invariant (it depends on X only through its law; Definition 6.5 below), and it is the criterion of Corollary 5.19(ii). It has three defects.

Example 6.3 (Defects of the value at risk).

(i) *It ignores what happens beyond the level δ.* The laws $(1 - \varepsilon)\delta_1 + \varepsilon\delta_0$ and $(1 - \varepsilon)\delta_1 + \varepsilon\delta_{-10^6}$ have the same value at risk for every $\delta > \varepsilon$, although the second one is catastrophic with probability ε.

(ii) *It penalizes diversification.* Let X, Y be independent with $\mathbb{P}(X = -1) = 0.04$, $\mathbb{P}(X = 0) = 0.96$, and the same law for Y. Then $\mathrm{VaR}_{0.05}(X) = \mathrm{VaR}_{0.05}(Y) = 0$, but $\mathbb{P}(X+Y \le -1) = 1-0.96^2 \approx 0.078 \ge 0.05$ gives $\mathrm{VaR}_{0.05}(X+Y) = -1 < \mathrm{VaR}_{0.05}(X)+\mathrm{VaR}_{0.05}(Y)$: pooling two independent risks looks worse than either. A sensible risk measure should be *superadditive* in the gain convention (subadditive for losses), rewarding diversification.

(iii) *It is discontinuous.* On the return distribution of the risky policy in the windy hike (Figure 6.2), $\delta \mapsto \mathrm{VaR}_\delta$ is 0 for $\delta \le 0.34$, climbs through all of $1, \dots, 7$ on the short range $(0.34, 0.47]$, and equals its maximum 8 for $\delta > 0.47$; and as a function of the law, it is not continuous for $\mathcal{W}_1$ (Exercise 5.3).

The Chernoff bound of Chapter 2 provides a family of lower bounds on the value at risk, one for each rate β, which do not have these defects. Recall from Corollary 2.19 that for $\beta > 0$ and every u, $\mathbb{P}(X \le u) \le \exp(-\beta(U_{-\beta}(X) - u))$.

Proposition 6.4 (Chernoff lower bounds on the value at risk). For every $\beta > 0$,

$$\mathrm{VaR}_\delta(X) \;\ge\; U_{-\beta}(X) - \frac{\ln(1/\delta)}{\beta}.$$

Proof. Let u_β be the right-hand side; the Chernoff bound gives $\mathbb{P}(X \le u_\beta) \le \delta$. Let $q = \mathrm{VaR}_\delta(X)$. The function $u \mapsto e^{\beta u}\mathbb{E}e^{-\beta X}$ is increasing and equals δ at u_β, while at $u = q$ it satisfies $e^{\beta q}\mathbb{E}e^{-\beta X} \ge \mathbb{E}[e^{\beta(q-X)}\mathbb{1}\{X \le q\}] \ge \mathbb{P}(X \le q) \ge \delta$ (Proposition 6.2). The first two inequalities are equalities only if $X = q$ almost surely, in which case the last one is strict since $\delta < 1$; in every case $e^{\beta q}\mathbb{E}e^{-\beta X} > \delta = e^{\beta u_\beta}\mathbb{E}e^{-\beta X}$, hence $u_\beta < q$: the inequality of the proposition is in fact strict for every X and every $\beta > 0$. ∎

The best of these bounds, $\sup_{\beta>0}\{U_{-\beta}(X) - \ln(1/\delta)/\beta\}$, is the entropic value at risk of Section 6.4. Before studying it, we set up the axiomatic framework in which its qualities can be stated.

6.2 Coherent risk measures

Let L^∞ denote the bounded random variables on a probability space $(\Omega, \mathfrak{F}, \mathbb{P})$. A *risk measure* is a map $\rho : L^\infty \to \mathbb{R}$. The following axioms, due to Artzner et al. (1999), are written in the gain convention.

Definition 6.5 (Coherent and convex risk measures). A risk measure ρ is *coherent* if it satisfies

(M) *monotonicity:* $X \le Y$ almost surely implies $\rho(X) \le \rho(Y)$;

(T) *translation invariance* (the term of Artzner et al. (1999); the property is in fact an *equivariance*, and is also called so): $\rho(X + c) = \rho(X) + c$ for every $c \in \mathbb{R}$;

(S) *superadditivity:* $\rho(X + Y) \ge \rho(X) + \rho(Y)$;

(H) *positive homogeneity:* $\rho(\lambda X) = \lambda\rho(X)$ for every $\lambda \ge 0$.

It is a *convex risk measure* if it satisfies (M), (T) and

(C) *concavity:* $\rho(\lambda X + (1-\lambda)Y) \ge \lambda\rho(X) + (1-\lambda)\rho(Y)$ for $\lambda \in [0,1]$.

It is *law invariant* if $\rho(X)$ depends only on the law of X; we then write $\rho(\nu)$ for ν the law of X.

(T) says that adding a sure amount c to the return improves its risk assessment by exactly c: $\rho(X)$ is measured in the units of the return, and $\rho(X) - \rho(Y)$ is the sure amount that makes Y as acceptable as X. (S) rewards diversification, and (H) says that scaling a position scales its risk. Coherence implies concavity ($\rho(\lambda X + (1-\lambda)Y) \ge \rho(\lambda X) + \rho((1-\lambda)Y) = \lambda\rho(X) + (1-\lambda)\rho(Y)$), so coherent measures are convex risk measures—the name comes from the loss convention, in which the risk of a loss L is $-\rho(-L)$, a convex function of L; the latter class, introduced by Föllmer and Schied (2002) and Frittelli and Rosazza Gianin (2002), drops (H), allowing the risk of a position to grow more than linearly with its size.

Proposition 6.6 (Elementary consequences). A risk measure satisfying (M) and (T)—in particular a convex risk measure—satisfies $\rho(c) = c + \rho(0)$ for $c \in \mathbb{R}$, and $\operatorname{ess\,inf} X + \rho(0) \le \rho(X) \le \operatorname{ess\,sup} X + \rho(0)$; it is 1-Lipschitz for the supremum norm: $|\rho(X) - \rho(Y)| \le \|X - Y\|_\infty$. A coherent risk measure has $\rho(0) = 0$.

Proof. (T) gives $\rho(c) = \rho(0 + c) = \rho(0) + c$. Since $\operatorname{ess\,inf} X \le X \le \operatorname{ess\,sup} X$ a.s., (M) and (T) give the two bounds. From $X \le Y + \|X - Y\|_\infty$, (M) and (T) give $\rho(X) \le \rho(Y) + \|X - Y\|_\infty$, and symmetrically. (H) with $\lambda = 0$ gives $\rho(0) = 0$. ∎

Example 6.7 (Risk measures reviewed).

(i) The expectation is coherent: the risk-neutral agent's measure.

(ii) The essential infimum is coherent: the most conservative measure, which only looks at the worst possible outcome (the criterion of Proposition 4.20).

(iii) The value at risk satisfies (M), (T), (H) but not (S) (Example 6.3); it is not concave either.

(iv) For $\beta < 0$, the β-mean U_β is a convex risk measure: (M) and (T) are Proposition 2.17(iii)–(iv), and (C) holds because $Y \mapsto \ln \mathbb{E}e^Y$ is convex (Hölder's inequality, Lemma C.4), so that $X \mapsto \ln \mathbb{E}e^{\beta X}$ is convex and $U_\beta(X) = \frac{1}{\beta} \ln \mathbb{E}e^{\beta X}$ is concave for $\beta < 0$. It is not positively homogeneous (Proposition 2.17(iii)): $U_\beta(\lambda X) = \lambda U_{\lambda\beta}(X)$, and doubling a position doubles its risk aversion. In this context $U_{-\beta}$, $\beta > 0$, is called the *entropic risk measure* (Föllmer and Schied, 2016, Chap. 4).

(v) The mean–variance criterion violates (M) (Example 5.16).

(vi) The conditional and entropic values at risk are coherent (Theorems 6.11 and 6.15).

On a finite probability space, coherent risk measures have a simple structure: they are the worst-case expectations over a set of alternative probability measures. This is the fundamental representation theorem of Artzner et al. (1999); we prove it in the finite case, where it is an exercise in convex analysis.

Theorem 6.8 (Representation of coherent risk measures). Let Ω be finite and $\mathbb{P}$ have full support, so that $L^\infty = \mathbb{R}^\Omega$. A map $\rho : \mathbb{R}^\Omega \to \mathbb{R}$ is a coherent risk measure if and only if there exists a nonempty closed convex set $\mathcal{Q}$ of probability measures on Ω such that

$$\rho(X) = \min_{Q \in \mathcal{Q}} \mathbb{E}_Q[X] \qquad \text{for all } X \in \mathbb{R}^\Omega.$$

One can take $\mathcal{Q} = \{Q : \mathbb{E}_Q[X] \geq \rho(X) \text{ for all } X\}$.

Proof. If $\rho(X) = \min_{Q\in\mathcal{Q}} \mathbb{E}_Q[X]$, then (M), (T), (H) hold for each $\mathbb{E}_Q$ and pass to the minimum, and (S) holds because $\min_Q(\mathbb{E}_Q X + \mathbb{E}_Q Y) \geq \min_Q \mathbb{E}_Q X + \min_Q \mathbb{E}_Q Y$. Conversely, let ρ be coherent. It is concave and finite on $\mathbb{R}^\Omega$, hence continuous, and admits at every X a supergradient $q_X \in \mathbb{R}^\Omega$: $\rho(Y) \leq \rho(X) + \langle q_X, Y - X\rangle$ for all Y (Lemma C.3, applied to the convex function $-\rho$). Taking $Y = 2X$ and $Y = 0$ and using (H) gives $\rho(X) \leq \langle q_X, X\rangle$ and $-\rho(X) \leq -\langle q_X, X\rangle$, so $\langle q_X, X\rangle = \rho(X)$ and then $\rho(Y) \leq \langle q_X, Y\rangle$ for all Y. Let $\mathcal{Q}_0 = \{q \in \mathbb{R}^\Omega : \langle q, Y\rangle \geq \rho(Y)\ \forall Y\}$, a closed convex set containing every q_X. For $q \in \mathcal{Q}_0$: taking $Y = \mathbb{1}_{\{\omega\}} \geq 0$ and $\rho(Y) \geq \rho(0) = 0$ by (M) gives $q(\omega) \geq 0$; taking $Y = \pm 1$ and $\rho(\pm 1) = \pm 1$ by (T) gives $\sum_\omega q(\omega) = 1$. Thus $\mathcal{Q}_0$ is a set of probability measures, and $\rho(X) \leq \min_{q\in\mathcal{Q}_0}\langle q, X\rangle \leq \langle q_X, X\rangle = \rho(X)$. ∎

On general probability spaces the same representation holds, with $\mathcal{Q}$ a set of probability measures absolutely continuous with respect to $\mathbb{P}$, under a continuity condition (the Fatou property); see Delbaen (2002) and Föllmer and Schied (2016, Chap. 4). Convex risk measures have the analogous representation $\rho(X) = \inf_Q\{\mathbb{E}_Q[X] + \alpha(Q)\}$ with a convex *penalty* α on the alternative measures, nonnegative when ρ is normalized by $\rho(0) = 0$; for the entropic risk measure the penalty is a multiple of the relative entropy (Corollary 6.23).

6.3 The conditional value at risk

The first defect of the value at risk is cured by averaging the quantiles below the level δ.

Definition 6.9 (Conditional value at risk). The *conditional value at risk* (or *expected shortfall*, or *average value at risk*) at level $\delta \in (0, 1)$ is

$$\mathrm{CVaR}_\delta(X) \;\coloneqq\; \frac{1}{\delta}\int_0^\delta \mathrm{VaR}_u(X)\,\mathrm{d}u = \frac{1}{\delta}\int_0^\delta F_X^{-1}(u)\,\mathrm{d}u.$$

By the quantile coupling $X = F_X^{-1}(U)$, U uniform (Theorem D.3), $\mathrm{CVaR}_\delta(X)$ is the mean of X conditionally on $U \leq \delta$, that is, the average of the worst fraction δ of the outcomes. When $\mathbb{P}(X \leq \mathrm{VaR}_\delta(X)) = \delta$—for instance when the law of X has no atoms—it equals $\mathbb{E}[X \mid X \leq \mathrm{VaR}_\delta(X)]$; with atoms, the quantile formula splits the atom at $\mathrm{VaR}_\delta(X)$ in the right proportion (Rockafellar and Uryasev, 2002).

Proposition 6.10 (Rockafellar–Uryasev representation). For every $\delta \in (0, 1)$,

$$\mathrm{CVaR}_\delta(X) \;=\; \max_{\eta\in\mathbb{R}} \left\{\eta - \frac{1}{\delta}\mathbb{E}\big[(\eta - X)_+\big]\right\},$$

and the maximum is attained at $\eta = \mathrm{VaR}_\delta(X)$.

Proof. Write $X = F^{-1}(U)$ with U uniform, so that $\mathbb{E}[(\eta - X)_+] = \int_0^1 (\eta - F^{-1}(u))_+ \,\mathrm{d}u$. For every η, dropping the part of the integral over $(\delta, 1)$ and then the positive part,

$$\eta - \frac{1}{\delta}\int_0^1 \big(\eta - F^{-1}(u)\big)_+ \,\mathrm{d}u \le \eta - \frac{1}{\delta}\int_0^\delta \big(\eta - F^{-1}(u)\big)_+ \,\mathrm{d}u \le \eta - \frac{1}{\delta}\int_0^\delta \big(\eta - F^{-1}(u)\big) \,\mathrm{d}u = \frac{1}{\delta}\int_0^\delta F^{-1}(u) \,\mathrm{d}u.$$

For $\eta = q \coloneqq F^{-1}(\delta)$ both inequalities are equalities: $F^{-1}(u) \ge q$ for $u > \delta$, so the dropped integrand vanishes, and $F^{-1}(u) \le q$ for $u \le \delta$, so the positive part is inactive. ∎

Theorem 6.11 (Dual representation and coherence of the conditional value at risk). For every $\delta \in (0, 1)$,

$$\mathrm{CVaR}_\delta(X) \;=\; \min\Big\{\mathbb{E}[XZ] :\; Z \ge 0,\; Z \le \tfrac{1}{\delta} \text{ a.s.},\; \mathbb{E}Z = 1\Big\} = \min\Big\{\mathbb{E}_Q[X] :\; Q \ll \mathbb{P},\; \frac{\mathrm{d}Q}{\mathrm{d}\mathbb{P}} \le \frac{1}{\delta}\Big\}.$$

Consequently CVaR_δ is a coherent, law-invariant risk measure. It is $\frac{1}{\delta}$-Lipschitz for $\mathcal{W}_1$, nondecreasing in δ, and $\mathrm{CVaR}_\delta(X) \le \mathbb{E}[X]$, with $\mathrm{CVaR}_\delta(X) \to \operatorname{ess\,inf} X$ as $\delta \to 0$ and $\mathrm{CVaR}_\delta(X) \to \mathbb{E}[X]$ as $\delta \to 1$.

Proof. Let Z be admissible and $\eta \in \mathbb{R}$. Since $X \ge \eta - (\eta - X)_+$ pointwise, $Z \ge 0$ and $Z \le 1/\delta$,

$$\mathbb{E}[XZ] \ge \eta\,\mathbb{E}Z - \mathbb{E}\big[(\eta - X)_+ Z\big] \ge \eta - \frac{1}{\delta}\mathbb{E}\big[(\eta - X)_+\big],$$

and taking $\eta = \mathrm{VaR}_\delta(X)$ gives $\mathbb{E}[XZ] \ge \mathrm{CVaR}_\delta(X)$ by Proposition 6.10. Conversely, let $q = \mathrm{VaR}_\delta(X)$ and define $Z^\star = \frac{1}{\delta}\mathbb{1}\{X < q\} + c\,\mathbb{1}\{X = q\}$ with c chosen so that $\mathbb{E}Z^\star = 1$: $c = (1 - \mathbb{P}(X < q)/\delta)/\mathbb{P}(X = q)$ if $\mathbb{P}(X = q) > 0$, and $c = 0$ otherwise (then $\mathbb{P}(X < q) = \delta$). Since $\mathbb{P}(X < q) \le \delta \le \mathbb{P}(X \le q)$ (Proposition 6.2), $c \in [0, 1/\delta]$ and $Z^\star$ is admissible. Then $\mathbb{E}[XZ^\star] = \frac{1}{\delta}\big(\mathbb{E}[X\mathbb{1}\{X < q\}] + (\delta - \mathbb{P}(X < q))\,q\big)$, which is the mean of the lower δ-tail in the quantile representation, that is, $\mathrm{CVaR}_\delta(X)$. The second form is the first with $Z = \mathrm{d}Q/\,\mathrm{d}\mathbb{P}$.

Coherence follows from the representation as a minimum of expectations over a convex set of probability measures (first part of the proof of Theorem 6.8); law invariance is clear from the definition. The Lipschitz property is Corollary D.5 for the distorted mean with weight $\frac{1}{\delta}\mathbb{1}_{[0,\delta]}$. Monotonicity in δ: the average of $u \mapsto F^{-1}(u)$, a nondecreasing function, over $[0, \delta]$ is nondecreasing in δ; its limits are $F^{-1}(0^+) = \operatorname{ess\,inf} X$ and $\int_0^1 F^{-1} = \mathbb{E}X$, and the bound $\mathrm{CVaR}_\delta(X) \le \mathbb{E}X$ is the case $\delta = 1$ of the monotonicity. ∎

The conditional value at risk is thus a coherent repair of the value at risk, and the risk measure adopted by banking regulation (Basel III) in place of the value at risk. From the point of view of this book it has one drawback: it is neither Bellman closed nor Bellman optimizable (Exercise 4.3 and Exercise 5.4), so that greedy distributional planning may fail for it. Its optimization in an MDP is nevertheless possible, at the price of the augmented state of Section 5.5: the Rockafellar–Uryasev representation reduces it to a one-dimensional family of expected-utility problems.

Corollary 6.12 (CVaR planning by state augmentation). In a finite-horizon MDP, for every state s and $\delta \in (0, 1)$,

$$\max_{\pi \in \Pi_{\mathrm{H}}} \mathrm{CVaR}_\delta\left(v^\pi(s)\right) \;=\; \max_{u \in \mathcal{W}_{H+1}} \Big\{u - \frac{1}{\delta}\min_{\pi \in \Pi_{\mathrm{H}}} \mathbb{E}_s^\pi\big[(u - W_H)_+\big]\Big\},$$

where $\mathcal{W}_{H+1}$ is the finite set of values of the return (Definition 5.17) and each inner minimization is an instance of Theorem 5.18 with $f(w) = -(u-w)_+$; an optimal policy is Markov in the augmented state $(s, W_{1:t-1})$.

Proof. By Proposition 6.10, for each fixed π the maximum over η is attained at a δ-quantile of $v^\pi(s)$, hence at a point of $\mathcal{W}_{H+1}$, so that $\max_\pi \mathrm{CVaR}_\delta(v^\pi(s)) = \max_\pi \max_{u\in\mathcal{W}_{H+1}}\{u - \delta^{-1}\mathbb{E}_s^\pi[(u-W_H)_+]\}$; exchanging the two maxima gives the formula, and Theorem 5.18 the form of the policy attaining the inner minimum at the optimal u. ∎

6.4 The entropic value at risk

The Chernoff bound of Proposition 6.4 controls the probability of the lower tail $\{X \le u\}$, and it carries information only when this probability is small: a bound $\mathbb{P}(X \le u) \le \delta$ with $\delta \ge \frac{1}{2}$ says nothing that the median does not say better. The construction of this section is therefore meaningful for small levels δ—a few percent, when δ stands for the tolerated probability of a disaster—and makes essentially no sense for $\delta \ge \frac{1}{2}$: the quantity defined below remains a lower bound on the δ-quantile, but one that never exceeds the mean, whereas the quantile itself rises to $\operatorname{ess\,sup} X$ as $\delta \to 1$. This is why the limit $\delta \to 1$ in Theorem 6.15(iii) is the expectation and not the essential supremum: as δ grows, the measure forgets the tail and tends to the risk-neutral assessment. An agent interested in the upper half of the distribution—in what can be hoped for rather than in what is to be feared—needs the mirror construction of Section 6.5.

Definition 6.13 (Entropic value at risk). The *entropic value at risk* at level $\delta \in (0,1)$ is

$$\mathrm{EVaR}_\delta(X) \coloneqq \sup_{\beta>0}\left\{U_{-\beta}(X) - \frac{\ln(1/\delta)}{\beta}\right\} = \sup_{\beta>0}\left\{-\frac{1}{\beta}\ln\mathbb{E}\left[e^{-\beta X}\right] - \frac{\ln(1/\delta)}{\beta}\right\}.$$

By Proposition 6.4, $\mathrm{EVaR}_\delta(X)$ is the tightest lower bound on $\mathrm{VaR}_\delta(X)$ that the Chernoff method can produce; this is how Ahmadi-Javid (2012) introduced it. Each term of the supremum is a translation of the entropic risk measure $U_{-\beta}$, and the supremum will turn out to restore the positive homogeneity that the β-means lack, without losing their other properties. On the two actions of Example 1.19, $\mathrm{EVaR}_\delta(\delta_1) = 1$ for every δ, while for the fair coin $\frac{1}{2}(\delta_0+\delta_2)$ the supremum equals $\sup_{\beta>0}\{-\frac{1}{\beta}\ln(1+e^{-2\beta}) + \frac{\ln(2\delta)}{\beta}\}$: it is 0 at $\delta = \frac{1}{2}$ (the mass $\frac{1}{2}$ on the outcome 0 is then the "event of probability δ" that the measure guards against), and it increases to the mean 1 as $\delta \to 1$. The key is a convexity lemma.

Lemma 6.14 (Perspective of a convex function). Let $g : E \to \mathbb{R}$ be a convex function on a real vector space E. Then $(x,t) \mapsto t\,g(x/t)$ is convex on $E \times (0,\infty)$.

Proof. For $\lambda \in (0,1)$, (x_1,t_1), (x_2,t_2) and $(x,t) = \lambda(x_1,t_1) + (1-\lambda)(x_2,t_2)$, write $\frac{x}{t} = \frac{\lambda t_1}{t}\frac{x_1}{t_1} + \frac{(1-\lambda)t_2}{t}\frac{x_2}{t_2}$, a convex combination. Convexity of g gives $g(x/t) \le \frac{\lambda t_1}{t}g(x_1/t_1) + \frac{(1-\lambda)t_2}{t}g(x_2/t_2)$; multiply by t. ∎

Theorem 6.15 (Properties of the entropic value at risk). For every $\delta \in (0,1)$:

(i) EVaR_δ is a coherent, law-invariant risk measure;

(ii) $\operatorname{ess\,inf} X \le \mathrm{EVaR}_\delta(X) \le \mathrm{CVaR}_\delta(X) \le \mathrm{VaR}_\delta(X)$, and $\mathrm{EVaR}_\delta(X) \le \mathbb{E}[X]$;

(iii) $\delta \mapsto \mathrm{EVaR}_\delta(X)$ is nondecreasing and continuous on $(0,1)$, with $\mathrm{EVaR}_\delta(X) \to \operatorname{ess\,inf} X$ as

$\delta \to 0$ and $\mathrm{EVaR}_\delta(X) \to \mathbb{E}[X]$ as $\delta \to 1$.

Proof. (i) Monotonicity, translation invariance and law invariance hold for each $U_{-\beta}$ (Proposition 2.17) and pass to the supremum. Positive homogeneity: for $c > 0$, $U_{-\beta}(cX) = c\,U_{-c\beta}(X)$ by Proposition 2.17(iii), so $\mathrm{EVaR}_\delta(cX) = \sup_\beta c\{U_{-c\beta}(X) - \ln(1/\delta)/(c\beta)\} = c\,\mathrm{EVaR}_\delta(X)$ (substitute $\beta' = c\beta$); the case $c = 0$ is $\mathrm{EVaR}_\delta(0) = 0$. Concavity: with $t = 1/\beta$ and $g(Y) \coloneqq \ln \mathbb{E}e^{-Y}$, a convex function of the bounded random variable Y (Hölder's inequality),

$$\mathrm{EVaR}_\delta(X) = \sup_{t>0} \big\{ -t\,g(X/t) - t\ln(1/\delta)\big\} = -\inf_{t>0} \big\{t\,g(X/t) + t\ln(1/\delta)\big\}.$$

By Lemma 6.14 with $E = L^\infty$, $(X,t) \mapsto t\,g(X/t) + t\ln(1/\delta)$ is jointly convex, and the infimum over t of a jointly convex function is convex in X (for $X = \lambda X_1 + (1-\lambda)X_2$, compare with $t = \lambda t_1 + (1-\lambda)t_2$ for near-optimal t_1, t_2); hence EVaR_δ is concave. Concavity and positive homogeneity give superadditivity: $\rho(X+Y) = 2\rho(\frac{X+Y}{2}) \geq \rho(X) + \rho(Y)$.

(ii) The lower bound: $U_{-\beta}(X) \geq \operatorname{ess\,inf} X$ and $\ln(1/\delta)/\beta \to 0$, so the supremum is at least $\lim_{\beta\to\infty} U_{-\beta}(X) = \operatorname{ess\,inf} X$ (Proposition 2.17(ii)). The comparison with the conditional value at risk: for every $\beta > 0$ and $\eta \in \mathbb{R}$, the inequality $s \leq e^{s-1}$ ($s \in \mathbb{R}$) applied to $s = \beta(\eta - X)$ gives $(\eta - X)_+ \leq \frac{1}{\beta} e^{\beta(\eta - X)-1}$, hence, by Proposition 6.10,

$$\mathrm{CVaR}_\delta(X) \geq \eta - \frac{1}{\delta}\mathbb{E}(\eta - X)_+ \geq \eta - \frac{e^{\beta\eta - 1}}{\beta\delta}\mathbb{E}e^{-\beta X}.$$

The right-hand side is maximized at $\beta\eta = 1 - \ln(1/\delta) - \ln \mathbb{E}e^{-\beta X}$, where it equals $\eta - \frac{1}{\beta} = U_{-\beta}(X) - \frac{\ln(1/\delta)}{\beta}$. Taking the supremum over β gives $\mathrm{CVaR}_\delta(X) \geq \mathrm{EVaR}_\delta(X)$. The inequality $\mathrm{CVaR}_\delta(X) \leq \mathrm{VaR}_\delta(X)$ holds because $\mathrm{CVaR}_\delta(X) = \frac{1}{\delta}\int_0^\delta F_X^{-1}(u)\,\mathrm{d}u \leq F_X^{-1}(\delta)$, the average of a nondecreasing function over $[0,\delta]$ being at most its value at δ; and $\mathrm{EVaR}_\delta(X) \leq \mathbb{E}X$ because $U_{-\beta}(X) \leq \mathbb{E}X$ for every $\beta > 0$.

(iii) $\mathrm{EVaR}_\delta(X)$ is the supremum over β of affine functions of $y = \ln(1/\delta)$, hence a convex function of y, finite on $(0,\infty)$, hence continuous; it is nonincreasing in y, that is, nondecreasing in δ. As $\delta \to 0$, $y \to \infty$: for fixed β the term tends to $-\infty$, but $\mathrm{EVaR}_\delta(X) \geq \operatorname{ess\,inf} X$ always, and $\mathrm{EVaR}_\delta(X) \leq \mathrm{VaR}_\delta(X) \to \operatorname{ess\,inf} X$ by (ii). As $\delta \to 1$, $\mathrm{EVaR}_\delta(X) \leq \mathbb{E}X$, and the choice $\beta = \sqrt{y}$ gives $\mathrm{EVaR}_\delta(X) \geq U_{-\sqrt{y}}(X) - \sqrt{y} \to \mathbb{E}X$. ∎

In the gain convention the more conservative measure is the smaller one, so the chain of (ii) runs from the most to the least conservative; in the loss convention of finance every inequality is reversed, and the chain reads VaR ≤ CVaR ≤ EVaR ≤ ess sup.

Example 6.16 (Gaussian returns, Ahmadi-Javid, 2012, Sect. 3). For $X \sim \mathcal{N}(m, \sigma^2)$, with Φ and φ the standard Gaussian distribution and density functions and $z_\delta = \Phi^{-1}(\delta) < 0$ for $\delta < \frac{1}{2}$,

$$\mathrm{VaR}_\delta(X) = m + \sigma z_\delta, \qquad \mathrm{CVaR}_\delta(X) = m - \sigma\frac{\varphi(z_\delta)}{\delta}, \qquad \mathrm{EVaR}_\delta(X) = m - \sigma\sqrt{2\ln(1/\delta)},$$

the last one because $U_{-\beta}(X) = m - \beta\sigma^2/2$ (Proposition 2.17(vi)) and $\sup_\beta\{-\beta\sigma^2/2 - \ln(1/\delta)/\beta\}$ is attained at $\beta = \sqrt{2\ln(1/\delta)}/\sigma$. At $\delta = 0.05$ the three measures are $m - 1.645\sigma$, $m - 2.063\sigma$ and $m - 2.448\sigma$: the entropic value at risk is the most conservative, and Figure 6.1 shows that the gap persists at all levels. For the Gaussian law, EVaR_δ is simply the mean minus a multiple of the standard deviation—a mean–variance criterion which, unlike Markowitz's, is monotone.

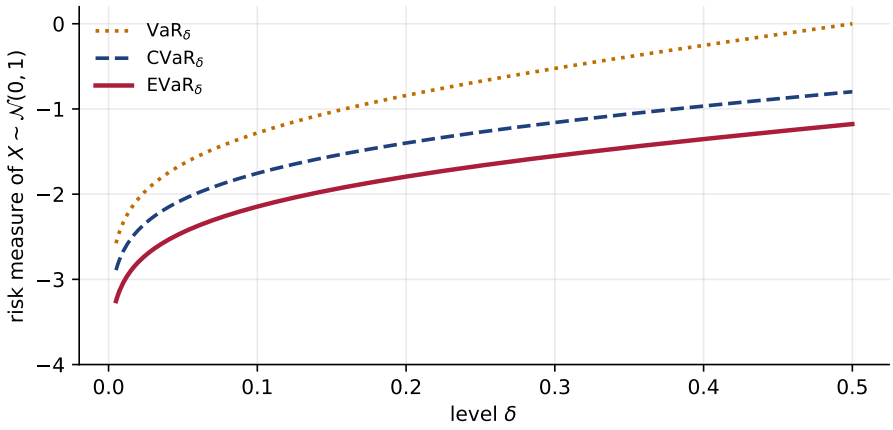


Figure 6.1. Value at risk, conditional value at risk and entropic value at risk of a standard Gaussian return as functions of the level δ (Example 6.16); the comparison is Figure 1 of Ahmadi-Javid (2012).

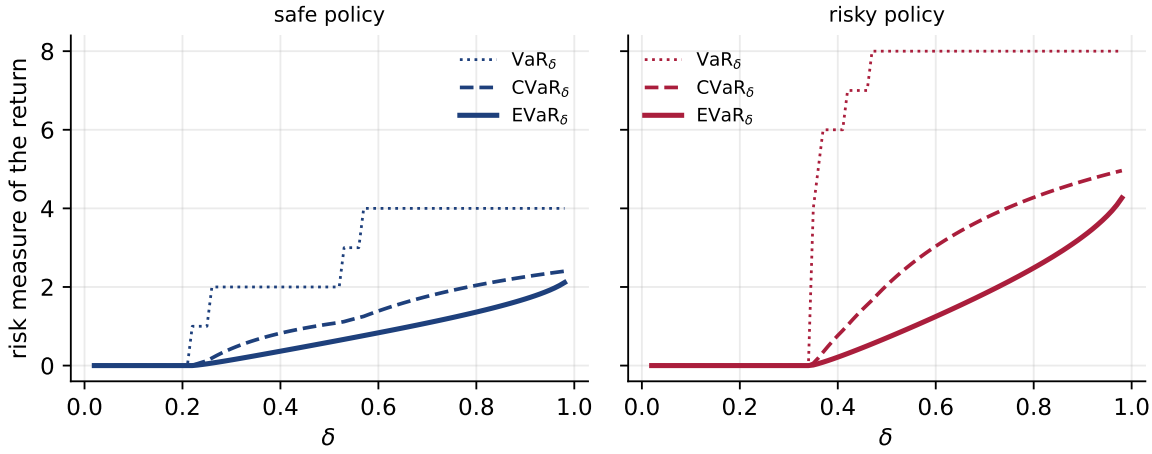


Figure 6.2. Value at risk, conditional value at risk and entropic value at risk of the return of the safe (left) and risky (right) policies in the windy hike, as functions of δ (Example 6.17).

Example 6.17 (The two hikers). Figure 6.2 shows the three risk measures of the return distributions of the safe and risky policies of Example 3.6, as functions of δ. All three vanish for δ below the probability of a zero return (0.22 for the safe policy, 0.34 for the risky one): none of the three measures at these levels can distinguish a hiker who fails from one who succeeds, since failure is not rare enough. Above these thresholds the value at risk climbs by jumps through the atoms of the return, whereas the two coherent measures increase continuously. At $\delta = 0.25$, 0.5 and 0.75 the entropic values at risk are 0.05, 0.60, 1.22 for the safe policy and 0, 0.71, 2.15 for the risky one: the safe policy is preferred only for $\delta \in (0.22, 0.47)$—the levels at which its failures are rare enough to be tolerated while the risky hiker's are not. Compare with the β-means, which prefer the safe policy for $\beta < \beta_\times \approx -0.65$, that is, at rates $-\beta > 0.65$ (Example 4.18): the two parametrizations of risk aversion, by a level δ or by a rate $\beta > 0$ (the β-mean being $U_{-\beta}$), are related by the supremum in Definition 6.13, whose maximizer $\beta^\star(\delta)$ decreases as δ increases (Example 6.32).

6.5 The risk-seeking counterpart

A risk-seeking agent—the child of the windy hike, a gambler, and, as Chapters 10 and 15 will show, every learner who explores—summarizes a return distribution by what it can hope for rather than by what it must fear: a high quantile, the average of the best outcomes, a β-mean U_β with $\beta > 0$. This section is the mirror image of the previous one for such an agent. Every statement follows from its counterpart applied to $-X$, whose lower tail is the upper tail of X, and the proofs say so; the statements are nevertheless written in their own terms, with $\beta > 0$ and U_β the β-mean of Definition 2.16 itself, because the object they define is the one that returns as the optimism principle of Part IV, and because the mirror is not perfect: planning for the risk-seeking agent has a structure of its own.

The first inequality of Corollary 2.19, $\mathbb{P}(X \geq u) \leq \exp(-\beta(u - U_\beta(X)))$ for $\beta > 0$, controls the upper tail. The relevant quantile is $\mathrm{VaR}_{1-\delta}(X) = F_X^{-1}(1-\delta)$: by Proposition 6.2 at level $1-\delta$, $\mathbb{P}(X \geq \mathrm{VaR}_{1-\delta}(X)) \geq \delta$, so that $\mathrm{VaR}_{1-\delta}(X)$ is a level that the return reaches with probability at least δ—a hope valid with probability δ, and the largest such level in the sense of that proposition.

Proposition 6.18 (Chernoff upper bounds on the upper quantile). For every $\beta > 0$,

$$\mathrm{VaR}_{1-\delta}(X) \;\leq\; U_\beta(X) + \frac{\ln(1/\delta)}{\beta},$$

and the inequality is strict for every X.

Proof. Let u_β be the right-hand side and $q = \mathrm{VaR}_{1-\delta}(X)$. The function $u \mapsto e^{-\beta u}\mathbb{E}e^{\beta X}$ is decreasing and equals δ at u_β, while at $u = q$

$$e^{-\beta q}\,\mathbb{E}e^{\beta X} \;\geq\; \mathbb{E}\big[e^{\beta(X-q)}\mathbb{1}\{X \geq q\}\big] \;\geq\; \mathbb{P}(X \geq q) \;\geq\; \delta,$$

where the first two inequalities are equalities only if $X = q$ almost surely, in which case the last one is strict since $\delta < 1$. Hence $e^{-\beta q}\mathbb{E}e^{\beta X} > \delta$ and $q < u_\beta$. ∎

Definition 6.19 (Optimistic entropic value at risk). The *optimistic entropic value at risk* at level $\delta \in (0,1)$ is

$$\overline{\mathrm{EVaR}}_\delta(X) \;\coloneqq\; \inf_{\beta>0}\left\{U_\beta(X) + \frac{\ln(1/\delta)}{\beta}\right\} = \inf_{\beta>0}\left\{\frac{1}{\beta}\ln\mathbb{E}\big[e^{\beta X}\big] + \frac{\ln(1/\delta)}{\beta}\right\}.$$

It is the tightest Chernoff upper bound on the level that the return reaches with probability δ: the best outcome one may hope for with probability δ, as certified by the exponential moments. Since $U_{-\beta}(-X) = -U_\beta(X)$ by the definition of the β-mean, $\overline{\mathrm{EVaR}}_\delta(X) = -\,\mathrm{EVaR}_\delta(-X)$: this is the reduction, and the only place where a sign is changed. As in Section 6.4, the construction is meaningful for small δ only: for $\delta \geq \frac{1}{2}$ it is an upper bound on the $(1-\delta)$-quantile which never falls below the mean.

Theorem 6.20 (Properties of the optimistic entropic value at risk). For every $\delta \in (0,1)$:

(i) $\overline{\mathrm{EVaR}}_\delta$ is monotone, translation invariant, positively homogeneous and law invariant, and it is *convex* and *subadditive*: $\overline{\mathrm{EVaR}}_\delta(X+Y) \leq \overline{\mathrm{EVaR}}_\delta(X) + \overline{\mathrm{EVaR}}_\delta(Y)$. In the loss convention of finance it is a coherent risk measure.

(ii) $\mathbb{E}[X] \leq \overline{\mathrm{EVaR}}_\delta(X)$ and $\mathrm{VaR}_{1-\delta}(X) \leq \overline{\mathrm{CVaR}}_\delta(X) \leq \overline{\mathrm{EVaR}}_\delta(X) \leq \operatorname{ess\,sup} X$, where $\overline{\mathrm{CVaR}}_\delta(X) \coloneqq \frac{1}{\delta}\int_{1-\delta}^1 F_X^{-1}(u)\,du$ is the average of the best fraction δ of the outcomes.

(iii) $\delta \mapsto \overline{\mathrm{EVaR}}_\delta(X)$ is nonincreasing and continuous on $(0,1)$, with $\overline{\mathrm{EVaR}}_\delta(X) \to \operatorname{ess\,sup} X$ as $\delta \to 0$ and $\overline{\mathrm{EVaR}}_\delta(X) \to \mathbb{E}[X]$ as $\delta \to 1$; and $\overline{\mathrm{EVaR}}_\delta(X) = \operatorname{ess\,sup} X$ as soon as $\delta \le \mathbb{P}(X = \operatorname{ess\,sup} X)$.

Proof. All claims are those of Theorem 6.15 for $-X$, read through $\overline{\mathrm{EVaR}}_\delta(X) = -\mathrm{EVaR}_\delta(-X)$. The change of sign preserves monotonicity, translation invariance, positive homogeneity and law invariance, turns concavity and superadditivity into convexity and subadditivity, and exchanges essential infimum and supremum as well as the two limits in (iii). For the chain of (ii): the lower quantile function of $-X$ is $u \mapsto -F_X^+(1-u)$, where $F_X^+(v) \coloneqq \inf\{x : F_X(x) > v\}$ is the upper quantile function of X, which differs from F_X^{-1} only at the countably many levels of the flat parts of F_X; hence $\mathrm{CVaR}_\delta(-X) = -\overline{\mathrm{CVaR}}_\delta(X)$ and $\mathrm{VaR}_\delta(-X) = -F_X^+(1-\delta) \le -\mathrm{VaR}_{1-\delta}(X)$. The first inequality of (ii) holds because $U_\beta(X) \ge \mathbb{E}[X]$ for $\beta > 0$ (Proposition 2.17(i)), and the last claim of (iii) is Exercise 6.4(c) for $-X$. ∎

Example 6.21 (Gaussian returns and the two hikers, seen from above). For $X \sim \mathcal{N}(m, \sigma^2)$ the symmetry of the Gaussian law turns the formulas of Example 6.16 into

$$\mathrm{VaR}_{1-\delta}(X) = m + \sigma|z_\delta|, \qquad \overline{\mathrm{CVaR}}_\delta(X) = m + \sigma\frac{\varphi(z_\delta)}{\delta}, \qquad \overline{\mathrm{EVaR}}_\delta(X) = m + \sigma\sqrt{2\ln(1/\delta)},$$

the infimum of $m + \beta\sigma^2/2 + \ln(1/\delta)/\beta$ being attained at $\beta = \sqrt{2\ln(1/\delta)}/\sigma$; at $\delta = 0.05$ the three values are $m + 1.645\sigma$, $m + 2.063\sigma$ and $m + 2.448\sigma$, and the optimistic entropic value at risk is the most optimistic of the three. In the windy hike (Figure 6.3) the best outcome has a sizeable probability—four steps at the bar with probability 0.43 for the safe policy, eight with probability 0.53 for the risky one—so that by Theorem 6.20(iii) $\overline{\mathrm{EVaR}}_\delta$ equals 4 for $\delta \le 0.43$ and 8 for $\delta \le 0.53$: at these levels, hope sees nothing but the best case. Beyond, at $\delta = 0.6$, 0.75 and 0.9, the optimistic entropic values at risk are 3.80, 3.52, 3.12 for the safe policy and 7.93, 7.46, 6.61 for the risky one (the upper conditional values at risk are 3.51, 3.20, 2.70 and 7.86, 6.70, 5.58; the upper quantiles 2, 1, 0 and 6, 0, 0). The child prefers the risky path at every level, by a margin of at least 2.9: unlike the β-means, which cross at $\beta_\times \approx -0.65$ (Example 4.18), the two upper tails never cross. The rate at which the infimum is attained decreases with δ (1.05, 0.63, 0.33 for the safe policy at the three levels), as it does in Example 6.17.

Duality. Theorem 6.24 below identifies $\mathrm{EVaR}_\delta(X)$ with the worst expectation over the models Q with $\mathrm{KL}(Q \,\|\, \mathbb{P}) \le \ln(1/\delta)$. Applied to $-X$ it gives (Corollary 6.25)

$$\overline{\mathrm{EVaR}}_\delta(X) \;=\; \max\left\{\mathbb{E}_Q[X] \,:\, \mathrm{KL}(Q \,\|\, \mathbb{P}) \le \ln\tfrac{1}{\delta}\right\},$$

the *best* case over the same relative-entropy ball, attained at a tilted measure $Q \propto e^{\beta X}\mathbb{P}$ on its boundary. The two measures are the two halves of one picture: as the radius $y = \ln(1/\delta)$ grows from 0 to ∞, the worst case decreases from $\mathbb{E}[X]$ to $\operatorname{ess\,inf} X$ along tilts of negative rate, in the signed convention of Definition 2.16, and the best case increases from $\mathbb{E}[X]$ to $\operatorname{ess\,sup} X$ along tilts of positive rate (Remark 6.26). The parameter running from the essential infimum to the essential supremum is therefore the signed rate, or the signed radius, and not a level $\delta \in (0,1)$: EVaR_δ and $\overline{\mathrm{EVaR}}_\delta$ do not meet at $\delta = \frac{1}{2}$—for a Gaussian they are $m \mp \sigma\sqrt{2\ln 2}$—and both tend to the mean as the ball shrinks. The signed rate is also the axis on which the optimality front of Section 6.8 lives; a front computed for both signs of β serves both agents. The best case over a relative-entropy ball is the form under which the optimistic entropic value at risk returns in Part IV: the KL-UCB index of Chapter 10, (10.2), is the

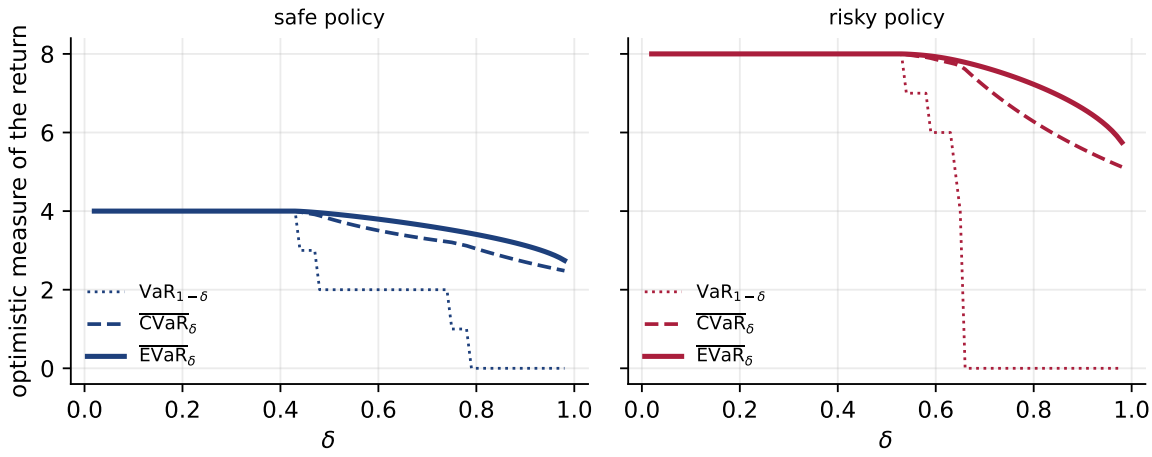


Figure 6.3. The mirror image of Figure 6.2: upper quantile $\mathrm{VaR}_{1-\delta}$, upper conditional value at risk $\overline{\mathrm{CVaR}}_\delta$ and optimistic entropic value at risk $\overline{\mathrm{EVaR}}_\delta$ of the return of the safe (left) and risky (right) policies in the windy hike, as functions of δ (Example 6.21). The curves start at the best reachable return and decrease to the mean; the optimistic entropic value at risk is the largest of the three, and equals the best return up to $\delta = 0.43$ (safe) and 0.53 (risky), the probabilities of that return.

largest mean compatible with the observations within such a ball, and Proposition 10.13 of Chapter 15 is its β-mean form.

Planning. The mirror breaks here, in an instructive way. Write $y = \ln(1/\delta)$, let $U^\star_{\beta,1}(s)$ be the optimal β-value at a rate $\beta > 0$ (Definition 5.7, computed by Algorithm 5.2), and consider

$$\bar{h}_\delta(\beta) \coloneqq U^\star_{\beta,1}(s) + \frac{y}{\beta}, \qquad U^\star_{\beta,1}(s) = \max_{\pi \in \Pi_{\mathrm{MD}}} U_\beta\big(\nu^\pi(s)\big), \qquad \beta > 0,$$

the counterpart of the function h_δ of Proposition 6.29. For every policy π and rate β, $\overline{\mathrm{EVaR}}_\delta(\nu^\pi(s)) \le U_\beta(\nu^\pi(s)) + y/\beta \le \bar{h}_\delta(\beta)$, so that $\inf_\beta \bar{h}_\delta$ bounds the risk-seeking optimum from above; but the argument of Proposition 6.29 exchanged two suprema, and a supremum over policies does not commute with an infimum over rates. The gap is real, and it is filled by randomization.

Proposition 6.22 (Risk-seeking planning). For every $s \in \mathcal{S}$ and $\delta \in (0,1)$,

$$\sup_{\pi \in \Pi_{\mathrm{H}}} \overline{\mathrm{EVaR}}_\delta\big(\nu^\pi(s)\big) = \inf_{\beta > 0} \bar{h}_\delta(\beta),$$

and the supremum is attained. If the infimum is attained at $\bar{\beta}$, the supremum is attained by a policy that draws, before the first step, one of the deterministic policies that are $\bar{\beta}$-optimal, and then follows it. The maximum over deterministic policies can be strictly smaller (Exercise 6.5(d)).

Proof. Let C be the set of return laws from s of the policies of Π_{H}. A policy of Π_{H} draws its actions from independent uniform variables, as a deterministic function of the history and of these variables; conditionally on their values it is a deterministic history-dependent policy, so that its return law is a mixture of the finitely many return laws of the deterministic policies of Π_{H}, and conversely every such mixture is the return law of a policy of Π_{H}—draw one of them at random, then follow it. Thus C is the convex hull of a finite set of laws on a finite set of atoms: a compact convex polytope. For $\nu \in C$ and $t > 0$ let $f(\nu, t) \coloneqq t \ln \int e^{x/t} \nu(\mathrm{d}x) + yt$, so that $f(\nu, 1/\beta) = U_\beta(\nu) + y/\beta$

and $\overline{\mathrm{EVaR}}_\delta(\nu) = \inf_{t>0} f(\nu,t)$. The function f is continuous, concave in ν for fixed t (the integral is linear in ν and the logarithm is concave) and convex in t for fixed ν (Lemma 6.14). The minimax theorem (Theorem C.7, applied to $-f$ with the compact convex set C on the side of the minimum) gives $\max_{\nu\in C}\inf_{t>0} f(\nu,t) = \inf_{t>0}\max_{\nu\in C} f(\nu,t)$. For fixed t, $f(\cdot,t)$ is an increasing function of a linear form, so that its maximum over the polytope C is attained at a vertex, the return law of a deterministic policy, and $\max_{\nu\in C} f(\nu,t) = \bar{h}_\delta(1/t)$ by Theorem 5.8. This proves the identity; the maximum on the left is attained because C is compact and $\inf_t f(\cdot,t)$ is upper semicontinuous. If the infimum is attained at $\bar\beta$, a maximizer $\bar\nu$ satisfies $f(\bar\nu,1/\bar\beta) \geq \inf_t f(\bar\nu,t) = \bar{h}_\delta(\bar\beta) = \max_{\nu\in C} f(\nu,1/\bar\beta)$, so that $\bar\nu$ maximizes the linear form $\nu \mapsto \int e^{\bar\beta x}\nu(\mathrm{d}x)$ over C: it is a mixture of the vertices attaining this maximum, which are the return laws of the $\bar\beta$-optimal deterministic policies. ∎

The risk-averse planner of Section 6.7 never needs to randomize, and a deterministic Markov policy is optimal (Theorem 6.30); the risk-seeking planner may need a coin, and the coin is thrown exactly where two β-optimal policies tie. In the one-state example of Exercise 6.5(d)—a sure reward 1, or 2 with probability 0.3 and 0 otherwise—the two deterministic policies have optimistic entropic values at risk 1 and 1.04 at $\delta = 0.9$, the surrogate $\inf_\beta \bar{h}_\delta$ equals 1.12, and it is reached by playing the sure action with probability 0.69: the two actions tie at the rate $\bar\beta \approx 0.85$ where the β-mean of the lottery equals 1, and the coin is tuned so that the mixture's own optimal rate is $\bar\beta$. The reason lies in the behaviour of β-means on mixtures: since $e^{\beta U_\beta(\sum_i p_i\mu_i)} = \sum_i p_i e^{\beta U_\beta(\mu_i)}$, the concavity of the logarithm gives $U_\beta(\sum_i p_i\mu_i) \geq \sum_i p_i U_\beta(\mu_i)$ for $\beta > 0$, with the reverse inequality for $\beta < 0$: mixing can only raise a criterion built from β-means at positive rates, and can only lower one built at negative rates—which is why Theorem 6.30 needs no coin. Computationally the mirror holds: the grid of Algorithm 6.1, run at the rates $+\beta_k$, approaches $\inf_\beta \bar{h}_\delta$ within ε with the same number of points, by the proof of Theorem 6.31 with every inequality reversed (Exercise 6.5(e)); and the front of Section 6.8, computed for both signs of β, supplies the vertices of every mixture one may need.

6.6 Duality: worst cases in a relative-entropy ball

Theorem 6.8 guarantees that the entropic value at risk is a worst-case expectation over some set of alternative measures. The Gibbs variational principle (Theorem C.11) identifies this set. We work on a finite probability space $(\Omega,\mathbb{P})$ with $\mathbb{P}$ of full support, which is the case of the return of a finite MDP viewed as a function of the trajectory (after discarding the trajectories of probability zero under the policy); the results extend to general spaces (Ahmadi-Javid, 2012).

Corollary 6.23 (Dual form of the entropic risk measure). For $\beta > 0$, $U_{-\beta}(X) = \min_{Q\ll\mathbb{P}} \left\{\mathbb{E}_Q[X] + \frac{1}{\beta}\mathrm{KL}(Q\,\|\,\mathbb{P})\right\}$, the minimum over probability measures on Ω being attained at the tilted measure $Q_\beta \propto e^{-\beta X}\mathbb{P}$.

Proof. Theorem C.11 (with $\mathbb{P}$ as the reference measure P) applied to $-\beta X$: $\ln \mathbb{E}e^{-\beta X} = \max_Q\{-\beta\mathbb{E}_Q X - \mathrm{KL}(Q\,\|\,\mathbb{P})\}$; divide by $-\beta$. ∎

Theorem 6.24 (Relative-entropy ball representation of EVaR). For every $\delta \in (0,1)$,

$$\mathrm{EVaR}_\delta(X) \;=\; \min\left\{\mathbb{E}_Q[X] \;:\; \mathrm{KL}(Q\,\|\,\mathbb{P}) \leq \ln\tfrac{1}{\delta}\right\}.$$

Proof. By Corollary 6.23, with $t = 1/\beta$,

$$\mathrm{EVaR}_\delta(X) = \sup_{t>0} \min_Q \left\{ \mathbb{E}_Q X + t\big(\mathrm{KL}(Q \,\|\, \mathbb{P}) - \ln \tfrac{1}{\delta}\big) \right\}.$$

The function $L(Q,t) = \mathbb{E}_Q X + t(\mathrm{KL}(Q \,\|\, \mathbb{P}) - \ln \frac{1}{\delta})$ is continuous, convex in Q (relative entropy is convex, and continuous on the simplex since $\mathbb{P}$ has full support) on the compact convex set of probability measures on Ω, and affine in t on the convex set $(0,\infty)$. The minimax theorem (Theorem C.7) allows the exchange of $\sup_t$ and $\min_Q$. For fixed Q, $\sup_{t>0} L(Q,t)$ equals $\mathbb{E}_Q X$ if $\mathrm{KL}(Q \,\|\, \mathbb{P}) \le \ln \frac{1}{\delta}$ and $+\infty$ otherwise. Hence $\mathrm{EVaR}_\delta(X) = \min_Q \sup_t L = \min\{\mathbb{E}_Q X : \mathrm{KL}(Q \,\|\, \mathbb{P}) \le \ln \frac{1}{\delta}\}$. ∎

Corollary 6.25 (Best case over the ball). For every $\delta \in (0,1)$, $\overline{\mathrm{EVaR}}_\delta(X) = \max\left\{\mathbb{E}_Q[X] : \mathrm{KL}(Q \,\|\, \mathbb{P}) \le \ln \frac{1}{\delta}\right\}$.

Proof. Apply Theorem 6.24 to $-X$ and use $\overline{\mathrm{EVaR}}_\delta(X) = -\mathrm{EVaR}_\delta(-X)$; the ball does not depend on X. ∎

The theorem gives the entropic value at risk its second meaning: it is the worst expected return over all models that cannot make certain an event which the nominal model deems to have probability less than δ—if $\mathrm{KL}(Q \,\|\, \mathbb{P}) \le \ln(1/\delta)$ and $Q(A) = 1$, then $\mathbb{P}(A) \ge \delta$, by the data-processing inequality (Lemma D.11(ii)) applied to the partition $\{A, A^c\}$: $\mathrm{KL}(Q \,\|\, \mathbb{P}) \ge \mathrm{kl}(Q(A), \mathbb{P}(A))$ for every event A, and $\mathrm{kl}(1,p) = \ln(1/p)$. The parameter δ thus quantifies *model uncertainty* as much as risk aversion, and EVaR_δ is a distributionally robust criterion. The alternative models are laws of the whole trajectory, not necessarily Markov: this is not a robust MDP with a divergence ball around each transition kernel (Section 17.4); the two constructions meet only one step at a time, where the worst case over a Kullback–Leibler ball around the next-state law is again an entropic value at risk (Proposition 17.13). The two extreme cases are consistent with Theorem 6.15: as $\delta \to 1$ the ball shrinks to $\{\mathbb{P}\}$ and the criterion is the expectation; as $\delta \to 0$ the ball fills the whole simplex and the criterion is the essential infimum. The same duality read in the other direction, Corollary 6.25, is the source of the optimistic exploration bonuses of Chapter 15.

Remark 6.26 (The optimal rate sits on the boundary of the ball). Let $\Lambda(\beta) \coloneqq \Lambda_{-X}(\beta) = \ln \mathbb{E} e^{-\beta X}$ be the cumulant generating function of $-X$, in the notation of Proposition 2.15, so that $U_{-\beta}(X) = -\Lambda(\beta)/\beta$, and let $Q_\beta \propto e^{-\beta X}\mathbb{P}$ be the tilted measure of Corollary 6.23. Then $\Lambda'(\beta) = -\mathbb{E}_{Q_\beta}[X]$ and $\mathrm{KL}\big(Q_\beta \,\big\|\, \mathbb{P}\big) = \mathbb{E}_{Q_\beta}[-\beta X - \Lambda(\beta)] = \beta\Lambda'(\beta) - \Lambda(\beta)$, whence, with $y = \ln(1/\delta)$,

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\left[U_{-\beta}(X) - \frac{y}{\beta}\right] = \frac{-\beta\Lambda'(\beta) + \Lambda(\beta) + y}{\beta^2} = \frac{y - \mathrm{KL}\big(Q_\beta \,\big\|\, \mathbb{P}\big)}{\beta^2}.$$

Since $\frac{\mathrm{d}}{\mathrm{d}\beta}\mathrm{KL}\big(Q_\beta \,\big\|\, \mathbb{P}\big) = \beta\Lambda''(\beta) = \beta\mathbb{V}\mathrm{ar}_{Q_\beta}[X] \ge 0$, the relative entropy of the tilted measure increases with the rate, from 0 at $\beta = 0$ to $\ln(1/\mathbb{P}(X = \operatorname{ess\,inf} X))$ as $\beta \to \infty$. The supremum defining $\mathrm{EVaR}_\delta(X)$ is thus attained, at the unique $\beta^\star$ with $\mathrm{KL}\big(Q_{\beta^\star} \,\big\|\, \mathbb{P}\big) = \ln(1/\delta)$, exactly when $\delta > \mathbb{P}(X = \operatorname{ess\,inf} X)$; the minimizer of Theorem 6.24 is then $Q_{\beta^\star}$, on the boundary of the ball, and $\mathbb{E}_{Q_{\beta^\star}}[X] = U_{-\beta^\star}(X) - y/\beta^\star = \mathrm{EVaR}_\delta(X)$ by the Gibbs identity. Otherwise $\mathrm{EVaR}_\delta(X) = \operatorname{ess\,inf} X$ (Exercise 6.4(c)). The rate $\beta^\star(\delta)$ decreases with δ, as observed on the hike (Examples 6.17 and 6.32). Applied to $-X$, the same computation shows that the infimum defining $\overline{\mathrm{EVaR}}_\delta(X)$ is attained at the unique $\beta > 0$ with $\mathrm{KL}(Q \,\|\, \mathbb{P}) = \ln(1/\delta)$ for $Q \propto e^{\beta X}\mathbb{P}$, provided $\delta > \mathbb{P}(X = \operatorname{ess\,sup} X)$, and that this rate decreases with δ too (Example 6.21).

Remark 6.27 (Other divergence balls, optimized certainty equivalents ♦). Replacing the relative entropy by another f-divergence $\int f(\mathrm{d}Q/\mathrm{d}\mathbb{P})\,\mathrm{d}\mathbb{P}$ (f convex with $f(1)=0$; the relative entropy is the case $f(x)=x\ln x$) defines the *g-entropic* risk measures of Ahmadi-Javid (2012) (in his terminology); the conditional value at risk is one of them, with the degenerate generator $f(x)=0$ for $x\le 1/\delta$ and $+\infty$ beyond, whose ball of radius 0 is the set $\{\mathrm{d}Q/\mathrm{d}\mathbb{P}\le 1/\delta\}$ of Theorem 6.11. A different unifying viewpoint is the *optimized certainty equivalent* of Ben-Tal and Teboulle (2007), $\mathrm{OCE}_u(X)=\sup_\eta\{\eta+\mathbb{E}[u(X-\eta)]\}$ for a concave nondecreasing u with $u(0)=0$ and $u(x)\le x$ (which makes the supremum finite and $\mathrm{OCE}_u(c)=c$): the conditional value at risk is OCE_u for $u(x)=-\frac{1}{\delta}(-x)_+$ (Proposition 6.10), and the entropic risk measure $U_{-\beta}$ is OCE_u for $u(x)=(1-e^{-\beta x})/\beta$: the optimal η solves $\mathbb{E}e^{-\beta(X-\eta)}=1$, that is, $\eta=U_{-\beta}(X)$, and the value is η. The entropic value at risk is not an optimized certainty equivalent, but a supremum of translated ones.

Remark 6.28 (The entropic value at risk determines the law ♦). The map $\delta\mapsto\mathrm{EVaR}_\delta(X)$ is, up to the change of variables $y=\ln(1/\delta)$ and $t=1/\beta$, the Legendre transform of the convex function $t\mapsto t\ln\mathbb{E}e^{-X/t}$; inverting the transform recovers the cumulant generating function of X on $(-\infty,0)$, hence the law of X (Exercise 6.6). Knowing the entropic value at risk at every level is thus knowing the distribution, exactly as for the value at risk (the quantile function) and the conditional value at risk.

6.7 Planning with the entropic value at risk

We now return to finite-horizon MDPs and to the planning problem for the criterion $\psi=\mathrm{EVaR}_\delta$: find $\pi\in\Pi_{\mathrm{H}}$ maximizing $\mathrm{EVaR}_\delta(v^\pi(s))$. Being monotone and continuous but not ordinally equivalent to a β-mean, EVaR_δ is not Bellman optimizable (Theorem 5.5 and Exercise 6.12; Remark 5.6(i) states the continuity that the necessity part actually requires), and greedy distributional planning may fail. But its definition as a supremum of β-means, each of which *is* optimizable, gives a direct route.

Proposition 6.29 (Reduction to exponential-utility planning, Hau et al., 2023). For every $s\in\mathcal{S}$ and $\delta\in(0,1)$,

$$\sup_{\pi\in\Pi_{\mathrm{H}}}\mathrm{EVaR}_\delta\left(v^\pi(s)\right)=\sup_{\beta>0}h_\delta(\beta),\qquad h_\delta(\beta)\coloneqq U^\star_{-\beta,1}(s)-\frac{\ln(1/\delta)}{\beta},$$

where $U^\star_{-\beta,1}(s)=\max_\pi U_{-\beta}(v^\pi(s))$ is the optimal β-value of Definition 5.7, computed by Algorithm 5.2.

Proof. Exchange the two suprema: $\sup_\pi\sup_\beta\{U_{-\beta}(v^\pi(s))-\ln(1/\delta)/\beta\}=\sup_\beta\{\sup_\pi U_{-\beta}(v^\pi(s))-\ln(1/\delta)/\beta\}$, and use Theorem 5.8 for the inner supremum. ∎

Theorem 6.30 (EVaR-optimal policies, Hau et al., 2023). The function h_δ is continuous on $(0,\infty)$, tends to $-\infty$ as $\beta\to 0$, and tends to the max–min value $m_\infty(s)\coloneqq\max_{\pi\in\Pi_{\mathrm{MD}}}\operatorname{ess\,inf}v^\pi(s)$ as $\beta\to\infty$ (the worst-case value of Remark 5.9(i), $\lim_{\beta\to-\infty}U^\star_{\beta,1}(s)$ in the signed notation of Chapter 5). Consequently there exists a deterministic Markov policy that is EVaR_δ-optimal among all history-dependent policies: the $(-\beta^\star)$-greedy policy if h_δ attains its supremum at $\beta^\star$, and a max–min optimal policy otherwise.

Proof. $U^\star_{-\beta,1}(s)$ is the maximum over the finitely many deterministic Markov policies of $U_{-\beta}(v^\pi(s))$

Algorithm 6.1: EVaR-optimal planning by a grid of exponential-utility plans, after Hau et al. (2023)

Input: MDP $\mathcal{M}$, state s, level $\delta \in (0,1)$, accuracy $\varepsilon > 0$, range Δ

$\beta_1 \leftarrow 8\varepsilon/\Delta^2$; $\quad k \leftarrow 1$

while $\beta_k < \ln(1/\delta)/\varepsilon$ **do**

  $\beta_{k+1} \leftarrow \dfrac{\beta_k \ln(1/\delta)}{\ln(1/\delta) - \beta_k \varepsilon}$; $\quad k \leftarrow k+1$ ▷ $1/\beta_{k+1} = 1/\beta_k - \varepsilon/\ln(1/\delta)$

$K \leftarrow k$

for $k = 1, \dots, K$ **do**

  run Algorithm 5.2 with $\beta = -\beta_k$: optimal value $U^\star_{-\beta_k,1}(s)$, policy π_k

  $h_k \leftarrow U^\star_{-\beta_k,1}(s) - \ln(1/\delta)/\beta_k$

Output: $\pi_{k^\star}$ with $k^\star \in \arg\max_k h_k$

(Theorem 5.8), each of which is continuous in β and bounded by $\max_\pi \operatorname{ess\,sup} v^\pi(s)$, and tends to $\operatorname{ess\,inf} v^\pi(s)$ as $\beta \to \infty$ (Proposition 2.17); this gives the continuity and the two limits. If the supremum of h_δ is attained at $\beta^\star$, the $(-\beta^\star)$-greedy policy $\pi^\star$ satisfies $\mathrm{EVaR}_\delta(v^{\pi^\star}(s)) \geq U_{-\beta^\star}(v^{\pi^\star}(s)) - \ln(1/\delta)/\beta^\star = h_\delta(\beta^\star)$, which equals the supremum of Proposition 6.29. Otherwise $\sup h_\delta = m_\infty(s)$, and a policy π_∞ attaining the max–min value satisfies $\mathrm{EVaR}_\delta(v^{\pi_\infty}(s)) \geq \operatorname{ess\,inf} v^{\pi_\infty}(s) = m_\infty(s)$ by Theorem 6.15(ii). ∎

The supremum over β is a one-dimensional problem, but h_δ is not concave in general (it is the maximum over policies of functions which are concave in $1/\beta$, not in β; Hau et al. 2023 give an example where the maximum of two such functions is not even quasi-concave). A grid search is therefore the safe option, and a suitably chosen grid comes with a guarantee. Let Δ be the length of an interval containing all returns of all policies from s ($\Delta \leq HR_{\max}$ for rewards in $[0, R_{\max}]$; the extreme values of Proposition 4.20, optimized over policies, give a sharper interval).

Theorem 6.31 (Accuracy of the grid, Hau et al., 2023). The policy returned by Algorithm 6.1 satisfies

$$\mathrm{EVaR}_\delta\left(v^{\pi_{k^\star}}(s)\right) \geq \sup_{\pi \in \Pi_{\mathrm{H}}} \mathrm{EVaR}_\delta\left(v^{\pi}(s)\right) - \varepsilon,$$

and the number of grid points is $K \leq 1 + \Delta^2 \ln(1/\delta)/(8\varepsilon^2)$. The total cost is $O(KHS^2A)$.

Proof. By Proposition 6.29 it suffices to show $\sup_{\beta>0} h_\delta(\beta) \leq \max_k h_k + \varepsilon$; then $\mathrm{EVaR}_\delta(v^{\pi_{k^\star}}(s)) \geq h_{k^\star}$ concludes as in the proof of Theorem 6.30. Write $U^\star_{-\beta}$ for $U^\star_{-\beta,1}(s)$ and $y = \ln(1/\delta)$; recall that $\beta \mapsto U^\star_{-\beta}$ is nonincreasing (Proposition 2.17(i)). Three cases.

Small rates, $\beta < \beta_1$. For any policy, Hoeffding's lemma (Lemma 2.10) applied to the return $W \in [a, a+\Delta]$ gives $\ln \mathbb{E} e^{-\beta W} \leq -\beta \mathbb{E}W + \beta^2\Delta^2/8$, that is, $U_{-\beta}(W) \geq \mathbb{E}W - \beta\Delta^2/8$; combined with $U_{-\beta}(W) \leq \mathbb{E}W$ this yields $U^\star_{-\beta} \leq \max_\pi \mathbb{E}[W^\pi] \leq U^\star_{-\beta_1} + \beta_1\Delta^2/8 = U^\star_{-\beta_1} + \varepsilon$. Since $-y/\beta \leq -y/\beta_1$, $h_\delta(\beta) \leq h_1 + \varepsilon$.

Intermediate rates, $\beta_k \leq \beta \leq \beta_{k+1}$. By monotonicity, $U^\star_{-\beta} \leq U^\star_{-\beta_k}$ and $-y/\beta \leq -y/\beta_{k+1}$, so $h_\delta(\beta) \leq U^\star_{-\beta_k} - y/\beta_{k+1} = h_k + y\,(1/\beta_k - 1/\beta_{k+1}) = h_k + \varepsilon$ by the choice of the grid.

Large rates, $\beta > \beta_K$. $h_\delta(\beta) \leq U^\star_{-\beta} \leq U^\star_{-\beta_K} = h_K + y/\beta_K \leq h_K + \varepsilon$, since $\beta_K \geq y/\varepsilon$.

Finally, $1/\beta_k$ decreases by ε/y at each step from $1/\beta_1 = \Delta^2/(8\varepsilon)$ until it falls to ε/y or below, which takes at most $\Delta^2 y/(8\varepsilon^2)$ steps. ∎

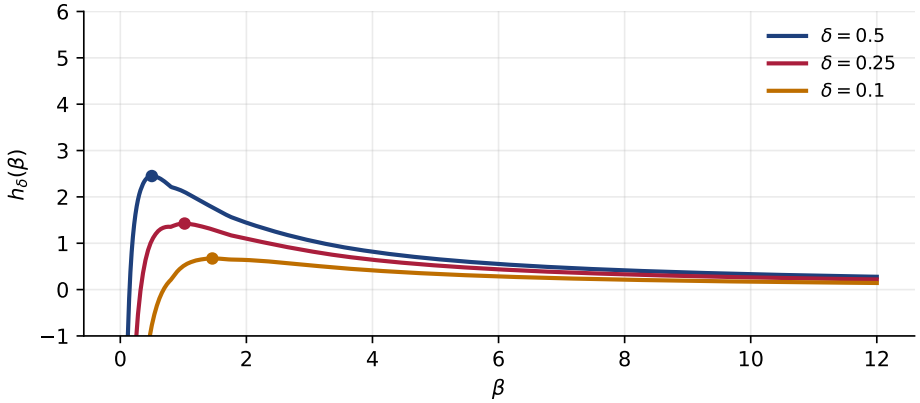


Figure 6.4. The function $h_\delta(\beta) = U^\star_{-\beta,1}(\mathrm{S}) - \ln(1/\delta)/\beta$ of Proposition 6.29 in the windy hike, for three levels δ; the dots mark the maximizers $\beta^\star(\delta)$ (Example 6.32).

Example 6.32 (EVaR-optimal hikers). Figure 6.4 shows h_δ for the windy hike at three levels, computed on a fine grid of rates by Algorithm 5.2: the curves rise steeply from $-\infty$, peak, and decay slowly towards the max–min value 0 (no path avoids the pit and the deadline with certainty). The optimal levels, rates and paths are

δ	0.9	0.75	0.5	0.25	0.1
$\max_\pi \mathrm{EVaR}_\delta$	4.08	3.41	2.45	1.43	0.67
$\beta^\star(\delta)$	0.20	0.32	0.50	1.02	1.46
row of the path	3	3	3	4	4
EVaR_δ of the safe policy	1.70	1.22	0.60	0.05	0
EVaR_δ of the risky policy	3.28	2.15	0.71	0	0

The EVaR-optimal hiker never takes the risky path: for $\delta \geq 0.5$ she follows the middle row, like the mean-optimal hiker, and for $\delta \leq 0.25$ the top row, like the prudent one—but with the time-adaptive behavior of Example 5.10, which is why she does far better than the hand-made safe policy at every level. The rate $\beta^\star(\delta)$ selected by the supremum decreases with δ: a smaller tolerance for bad outcomes translates into a larger risk aversion in the exponential-utility plan. For comparison, the conditional value at risk at $\delta = 0.5$ can be pushed to 4.17 by the augmented-state method of Corollary 6.12 (against 1.06 and 2.04 for the safe and risky policies), a much larger number than the 2.45 of the entropic value at risk, as Theorem 6.15(ii) requires. Both optimal policies depend on the level δ; the difference lies in the state they need. The CVaR-optimal policy is obtained from a family of augmented problems indexed by u and is in general Markov only in the augmented state $(s, W_{1:t-1})$, whereas the EVaR-optimal one is Markov in s (Theorem 6.30). In this example the CVaR-optimal policy happens to be Markov as well, because the accumulated reward carries no information: rewards are collected only in the absorbing exit.

6.8 The optimality front

From this section on we return to the signed β of Chapter 5: risk aversion is $\beta < 0$, and $\pi^\star_\beta$ denotes a β-greedy policy of Definition 5.7, computed from a fixed initial state s_1. Algorithm 6.1 solves one exponential-utility plan per grid point, and Theorem 6.31 asks for many points: K grows like Δ^2/ε^2.

Most of these plans are the same policy. The level δ is one way of asking for a property of the return distribution—a small probability of disaster, a good worst case—and translating it into a rate β; every such translation goes through the map $\beta \mapsto \pi^\star_\beta$ of Theorem 5.8, which this section studies as a whole. It is a small object.

Definition 6.33 (Optimality front). Call two policies equivalent when their return distributions from s_1 coincide, and write $\bar\pi$ for the class of π. For a class $\bar\pi$, let $O_{\bar\pi} = \{\beta \in \mathbb{R} : U_\beta(\nu^{\bar\pi}) = \sup_\pi U_\beta(\nu^\pi)\}$ be the set of rates for which it is optimal, and $I_{\bar\pi}$ the closure of the interior of $O_{\bar\pi}$. The *optimality front* is the family $\Gamma = (\bar\pi_k, I_k)_{k \le N_\Gamma}$ of the classes with $I_{\bar\pi} \neq \emptyset$, and the *breakpoints* are the endpoints $\bigcup_k \partial I_k$.

Taking the closure of the interior discards the classes that are optimal at isolated rates only, which are artefacts of ties; what remains is a genuine partition of the axis of rates. The front is defined through the law of the return from the initial state s_1, although the policies of Theorem 5.8 are optimal at every state simultaneously; and statements that $\beta \mapsto \pi^\star_\beta$ is piecewise constant refer to classes, or to a policy selected by a fixed deterministic tie-breaking rule in the backward induction: otherwise, policies inducing the same law could be exchanged arbitrarily as β varies.

Proposition 6.34 (Structure of the front, Marthe et al., 2025). In a finite-horizon MDP with finitely many states and actions (so that the return takes finitely many values, Remark 1.18), the optimality front is unique and finite. Each I_k is a finite union of closed intervals without isolated points, $\bigcup_k I_k = \mathbb{R}$, and two distinct I_{k_1}, I_{k_2} meet only along their boundaries. The map $\beta \mapsto \bar\pi^\star_\beta$, the class (return law from s_1) of the optimal policies, unique outside a finite set of rates, is therefore piecewise constant, with finitely many breakpoints.

Proof. By Theorem 5.8 one deterministic Markov policy is optimal at each β, and there are finitely many of those; a class with $I_{\bar\pi} \neq \emptyset$ is optimal on an interval, hence is the class of some $\pi^\star_\beta \in \Pi_{\mathrm{MD}}$, so there are finitely many classes in the front. Each return W_H is supported on the finite set $\mathcal{W}_{H+1}$ of Definition 5.17 (finite by Remark 1.18), so $M_{\bar\pi}(\beta) = \mathbb{E}[e^{\beta W_H}]$ is a finite sum $\sum_w \nu^{\bar\pi}(w) e^{\beta w}$ of exponentials: an entire function of β. Two distinct classes have distinct laws, hence distinct moment generating functions, so $M_{\bar\pi_1} - M_{\bar\pi_2}$ is a nonzero entire function and its zeros are isolated. Each such difference is a nonzero combination of at most $|\mathcal{W}_{H+1}|$ distinct exponentials, hence has at most $|\mathcal{W}_{H+1}| - 1$ real zeros on the whole line (Exercise 6.13(a)); there are finitely many pairs, so the finitely many crossing points cut the line into finitely many open pieces on each of which the ordering of the $M_{\bar\pi}$ does not change—and with it the ordering of the U_β, which are monotone functions of them, increasing for $\beta > 0$ and decreasing for $\beta < 0$, so that a constant ordering of the $M_{\bar\pi}$ gives a constant ordering of the U_β on each piece not containing 0. Hence one class is optimal throughout each piece. Passing to the closure of the interior removes the isolated optimality points. Note that every difference vanishes at $\beta = 0$, where all the β-means reduce to the mean, so 0 is always one of the cutting points. ∎

How wide are the pieces? The answer is a perturbation bound, and it is what makes the front cheap to compute: one backward induction at β certifies an interval around β on which nothing changes. The idea is simple: β-means move monotonically with β, and at a controlled speed; the certified interval is the set of rates on which the leading action's lead cannot be overturned.

Theorem 6.35 (Interval of optimality, Marthe et al., 2025). Fix a state and a stage, let $\mu_1, \dots, \mu_n$ be the return laws of the actions available there, and let $U^1_\beta > U^2_\beta \ge \cdots \ge U^n_\beta$ be their β-means, ordered, with U^1_β attained by a unique action; let $\Delta U = U^1_\beta - U^2_\beta$ be the gap. Let the

returns lie in $[r_{\min}, r_{\max}]$ and $\Delta = r_{\max} - r_{\min}$ be their range. Then the same action maximizes $U_{\beta'}$ for every β' in

$$\begin{cases} \left[\beta\left(1 + \dfrac{\Delta U}{U^2_\beta - r_{\min}}\right),\ \beta\left(1 - \dfrac{\Delta U}{U^1_\beta - r_{\min}}\right)\right] & \text{if } \beta < 0, \\ \left[-\dfrac{8\Delta U}{\Delta^2},\ \dfrac{8\Delta U}{\Delta^2}\right] & \text{if } \beta = 0, \\ \left[\beta\left(1 - \dfrac{\Delta U}{r_{\max} - U^2_\beta}\right),\ \beta\left(1 + \dfrac{\Delta U}{r_{\max} - U^1_\beta}\right)\right] & \text{if } \beta > 0. \end{cases}$$

In an MDP, apply this at every stage t and state s, after merging with the action of $\pi^\star_\beta$ the actions whose return law $\nu_t^{\pi^\star_\beta}(s,a)$ coincides with its own (they remain tied at every rate), with the generalized advantage $A^{\pi^\star_\beta}_{t,\beta}(s,a) = U_\beta(\nu_t^{\pi^\star_\beta}(s)) - U_\beta(\nu_t^{\pi^\star_\beta}(s,a))$ (a nonnegative gap: the sign is opposite to that of the usual advantage $Q - V$) in the role of ΔU and with the range of the returns-to-go from stage t in the role of $[r_{\min}, r_{\max}]$; the intersection of the resulting intervals is an interval containing β on which $\pi^\star_\beta$ remains optimal.

Proof idea. The key is a two-sided comparison of β-means at neighbouring rates. For $\beta < 0$ and $0 < \epsilon < |\beta|$,

$$U_\beta(X) \ \le\ U_{\beta+\epsilon}(X) \ \le\ \frac{\beta}{\beta+\epsilon} U_\beta(X) + \frac{\epsilon}{\beta+\epsilon} r_{\min},$$

obtained by factoring $e^{\epsilon x} \ge e^{\epsilon r_{\min}}$ out of the exponential sum and using that $1/(\beta+\epsilon) < 0$ reverses the inequality; the left-hand side is the monotonicity of Proposition 2.17. Rearranged, it reads $U_{\beta'} - r_{\min} \le (\beta/\beta')(U_\beta - r_{\min})$ for $\beta < \beta' < 0$: every U^i increases with the rate, by at most that factor. The two ends of the interval are therefore not symmetric, and this is the whole point: on each side only one of the two leading β-means has to be controlled, because both move the same way. Raising β toward 0, the leader can only gain, so it is enough that the runner-up stay below it, which gives the right endpoint $\beta(1 - \Delta U/(U^1 - r_{\min}))$; lowering β, the runner-up can only lose, so it is enough that the leader not fall to it, which gives the left endpoint $\beta(1 + \Delta U/(U^2 - r_{\min}))$. For $\beta > 0$ the same computation with $e^{\epsilon x} \le e^{\epsilon r_{\max}}$ replaces $U^i - r_{\min}$ by $r_{\max} - U^i$ and exchanges the two sides. At $\beta = 0$ the bounds degenerate and are replaced by Hoeffding's lemma (Lemma 2.10), which gives $|U_\beta(X) - \mathbb{E}[X]| \le |\beta|\Delta^2/8$; all β-means move in the same direction (upward for $\beta > 0$, downward for $\beta < 0$), so here again only one of the two leading ones has to be controlled—the runner-up for $\beta > 0$, the leader for $\beta < 0$—and the lead ΔU is preserved as long as $|\beta|\Delta^2/8 < \Delta U$: a window of half-width $8\Delta U/\Delta^2$ (a two-sided control of both would only give $4\Delta U/\Delta^2$). In the MDP, the optimal policy is greedy with respect to its own β-values; if at every stage and state the generalized advantage of every other action exceeds the possible perturbation, the policy is still greedy for itself at β', hence still optimal. See Marthe (2026, Thm. 4.2 and 4.4), whose statement of the single-state bound has the two endpoints exchanged. ■

The two ends of the front are the two extreme attitudes, and they have a combinatorial description that refines "optimize the best case" and "optimize the worst case".

Proposition 6.36 (The ends of the front, Marthe et al., 2025 ♦). Order the possible returns $x_1 < \cdots < x_n$ and associate with a policy the vector $L(\pi) = (\mathbb{P}(W_H = x_1), \ldots, \mathbb{P}(W_H = x_n))$. Then $\pi^\star_\beta$ converges, as $\beta \to -\infty$, to the policy minimizing $L(\pi)$ in the lexicographic order—it minimizes the probability of the worst return, then among those the probability of the second worst, and so on—and, as $\beta \to +\infty$, to the policy maximizing L read from the right. Moreover the extreme breakpoints are finite: if the returns are evenly spaced with gap Δx, the smallest breakpoint

Algorithm 6.2: FindBreaks: the front of a one-stage problem (Marthe et al., 2025)

Input: return laws $(\nu_a)_{a\in\mathcal{A}}$ of the actions, pairwise distinct (actions with the same law are merged first: they are tied at every rate), interval $[\beta_{\min}, \beta_{\max}]$, precision ϵ_β on the rate axis (not the value accuracy ε of Algorithm 6.1)

$\beta \leftarrow \beta_{\min}$; $\beta_\ell \leftarrow \beta_{\min}$; $a_{\text{prev}} \leftarrow \varnothing$

while $\beta \le \beta_{\max}$ **do**

  compute $U_\beta(\nu_a)$ for all a; $a_\beta \leftarrow \arg\max_a U_\beta(\nu_a)$ (ties, which occur only at isolated rates, broken by a fixed order of the actions)

  if $a_\beta \neq a_{\text{prev}}$ *and* $a_{\text{prev}} \neq \varnothing$ **then** record the interval $[\beta_\ell, \beta]$ for a_{prev}; $\beta_\ell \leftarrow \beta$

  $a_{\text{prev}} \leftarrow a_\beta$; $\Delta U \leftarrow U_\beta(\nu_{a_\beta}) - \max_{a\neq a_\beta} U_\beta(\nu_a)$

  $\beta \leftarrow \beta + \max\{w(\beta, \Delta U), \epsilon_\beta\}$ ▷ *w = certified half-width: distance from β to the* right *endpoint of the interval of Theorem 6.35*

record the interval $[\beta_\ell, \beta_{\max}]$ for a_{prev}

Output: the optimal actions and their intervals, breakpoints located within ϵ_β (intervals of optimality shorter than ϵ_β may be missed)

satisfies $\beta_{\inf} \ge -\ln\big(1 + \max_\pi \max_{i>i_0(\pi)} |a_i(\pi)|/|a_{i_0(\pi)}(\pi)|\big)/\Delta x$, where $a_i(\pi) = \mathbb{P}(W_H = x_i \mid \pi_{\inf}) - \mathbb{P}(W_H = x_i \mid \pi)$, the maximum runs over the policies whose return law differs from that of $\pi_{\inf}$, and $i_0(\pi) = \min\{i : a_i(\pi) \neq 0\}$ is the index of the first nonzero coefficient; symmetrically for the largest.

The bound is proved by turning $U_\beta(\nu) = U_\beta(\nu')$ into a polynomial equation through the substitution $z = e^{-\beta\Delta x}$ and applying Cauchy's bound on the roots (every root of $\sum_{i=1}^n a_i z^{n-i}$ satisfies $|z| \le 1 + \max_{i>i_0} |a_i|/|a_{i_0}|$, where a_{i_0} is the first nonzero coefficient: it vanishes, together with the nominal leading coefficient a_1, whenever the two policies give the worst return the same probability, and the polynomial then has lower degree). It is not computable without knowing every policy's return law, but it says that the front does not extend indefinitely: beyond a finite rate, increasing the risk aversion changes nothing.

How many pieces are there? The worst case is hopeless—at most $(n-1)|\mathcal{A}|^{2|\mathcal{S}|H}$ by counting pairs of policies and using that two distinct laws on n points have at most $n-1$ crossings (Exercise 6.13)—but the worst case is not what happens. Under a generic model (rewards and transition probabilities drawn independently from continuous laws), each breakpoint changes the optimal action at a single state and stage with probability one, because the finitely many equations that could make two actions tie at the same rate almost surely have no common solution (Marthe, 2026, Prop. 4.6). The thesis infers from it, as a heuristic supported by its experiments and not as part of the proposition, that the number of breakpoints should grow linearly with the number of states and the horizon. The proposition excludes simultaneous changes, not repeated ones: it does not bound the number of breakpoints by the number of state–stage pairs, since one pair can change its action several times and an action can come back. Already with one state, one stage and two actions with return laws $\frac{1}{40}(4, 21, 4, 11)$ and $\frac{1}{4}(1, 1, 1, 1)$ on $\{0, 1, 2, 3\}$, the difference of the moment generating functions is $\frac{1}{40}(e^\beta - 1)(e^\beta - 2)(e^\beta - 3)$, and the first action leads on $(0, \ln 2)$, loses the lead on $(\ln 2, \ln 3)$ and recovers it beyond $\ln 3$—a configuration that small perturbations do not destroy. A bound linear in the number of state–stage pairs would need an additional single-crossing property; what is true is that the fronts observed in practice are small, far below the worst case.

Computing the front is then two nested ideas. At one state, Algorithm 6.2 walks along the axis of rates, jumping by the certified width of Theorem 6.35 and falling back on a minimal step ϵ_β only when a breakpoint is near. The guarantee is resolution dependent. A certified jump never crosses a breakpoint; a fallback step of length ϵ_β, taken when the certified width is shorter, is not certified, and it can cross a

Algorithm 6.3: DOLFIN (distributional optimality-front iteration): the front of an MDP (Marthe et al., 2025)

Input: model $(P_t, r_t)_{t \le H}$, interval $[\beta_{\min}, \beta_{\max}]$, precision ϵ_β

$\mathcal{I}_H \leftarrow \{[\beta_{\min}, \beta_{\max}]\}$; $v_{H+1}(s) \leftarrow \delta_0$ for all s

for $t = H, \dots, 1$ **do**

  foreach *interval* $I \in \mathcal{I}_t$ **do**

    foreach *state* s **do**

      $\eta_t^I(s,a) \leftarrow \sum_{s'} P_t(s' \mid s,a)\,(b_{r_t(s,a,s')})_{\#} v_{t+1}^I(s')$ ▷ *distributional backup*

      $(J_s, (a_j^\star(s))_{j \in J_s}) \leftarrow \textsc{FindBreaks}\big(\epsilon_\beta, (\eta_t^I(s,a))_a, I\big)$

    add to $\mathcal{I}_{t-1}$ the cells j of the common refinement of the partitions $(J_s)_s$ of I, and store $\pi_t^j(s) \leftarrow a_{j_s}^\star(s)$ and $v_t^j(s) \leftarrow \eta_t^I(s, a_{j_s}^\star(s))$ for every s, where $j_s \in J_s$ is the subinterval containing j

Output: the front $\Gamma = (\pi_k, I_k)_k$ and the return law v^{π_k} of each of its policies

whole interval on which another action is optimal if that interval is shorter than ϵ_β. What the sweep does guarantee is that every maximal interval of optimality longer than ϵ_β is visited, and each of its endpoints located within ϵ_β; narrower ones can be missed, exactly as a grid of step ϵ_β misses them (the twenty-eighth policy of Example 6.37, optimal on a window of width $6 \cdot 10^{-4}$, is missed by the grid of step 10^{-3}). In an MDP, the return law of an action at stage t depends on the policy used afterwards, which itself changes with β—but only at the breakpoints of the subproblem starting at $t+1$. So the breakpoints propagate backward: those of the problem starting at t are those of the problem starting at $t+1$, together with the breakpoints of the one-stage problems obtained by freezing the continuation on each interval (Marthe, 2026, Prop. 4.10). Different states have different breakpoints, so the intervals passed to the next stage are the cells of the *common refinement* of the partitions found at all states: on each cell the whole decision rule of stage t is constant, the continuation laws of all successor states are fixed, and the cells are nested from one stage to the previous. A policy of the front is then assembled on each cell of $\mathcal{I}_0$ from the decision rules stored, at every stage, on the cells containing it (adjacent cells with the same return law from s_1 being merged). Algorithm 6.3 is this recursion, run distributionally: it carries the return laws rather than their β-means, which costs memory but means that the β-mean at any rate is a sum away, and that each policy of the front is delivered together with the law of its return—the object that the next section needs.

Example 6.37 (The front of the windy hike). On the windy hike, a grid of step 10^{-3} over $\beta \in [-5, 5]$ resolves 27 policies of the front; at step 10^{-4} a twenty-eighth appears, optimal on a window of width $6 \cdot 10^{-4}$. They are genuinely distinct—their return laws differ—but most of the differences are invisible from the outside: the mean return increases steadily from 2.93 at $\beta = -5$ to 5.09 near $\beta = 0$, and the probability of a zero return, which here is essentially the probability of the pit, from 0.018 to 0.153, while the *route* changes only twice, at $\beta \approx -0.81$ and $\beta \approx 0.017$ (Figure 5.2). This is Proposition 6.34's picture and the genericity remark above it: all but one of the 26 breakpoints move the action at a single cell and stage, usually deep in the grid where it barely matters, and only two of them reroute the walk. Figure 6.5 (left) shows the β-means of the policies and the intervals on which each is optimal.

The certified stepping of Algorithm 6.2 is what makes this cheap, and the accounting is instructive. At precision $\epsilon_\beta = 10^{-3}$ a plain grid needs 10 001 backward inductions and the certified sweep 3 288, a factor of 3; at $\epsilon_\beta = 10^{-4}$ the grid needs 100 001 and the sweep 5 524, a factor of 18. The jumps do not shrink with ϵ_β—only the final approach to each breakpoint does—so the saving grows as the precision does. It is nevertheless modest here, and for the reason Theorem 6.35

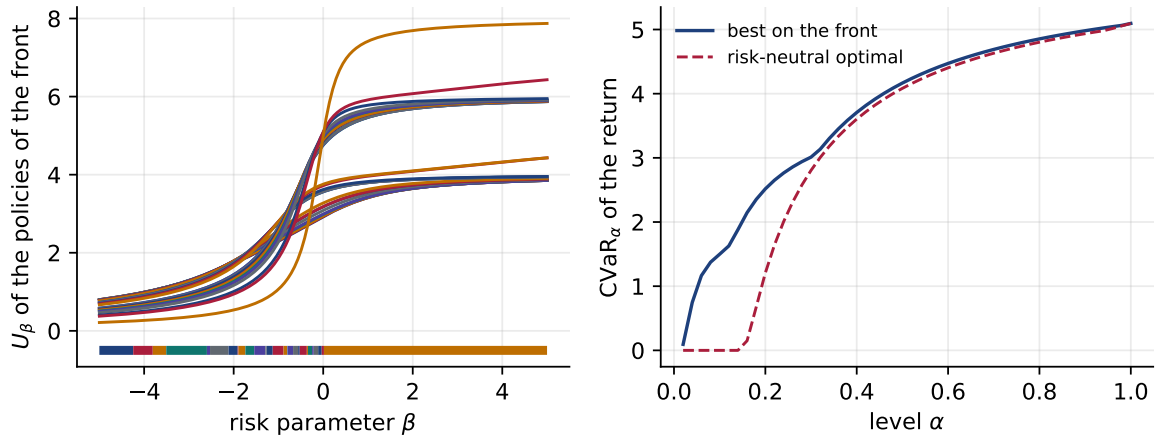


Figure 6.5. Left: the optimality front of the windy hike (Example 6.37); each curve is $\beta \mapsto U_\beta(\nu^\pi)$ for one of the 27 policies of the front, and the coloured segments below the axis are the intervals on which each is optimal. Right: the best conditional value at risk attainable on the front, against the level (denoted α on the figure's axis, δ in the text), compared with the risk-neutral optimal policy (Example 6.39).

makes plain: in an MDP the certificate is the *minimum* over all stages and states, so one cell where two actions are nearly equivalent throttles the whole sweep. Removing that bottleneck is exactly what Algorithm 6.3 does by treating each state separately.

6.9 One front, every risk measure

The front is the information that Algorithm 6.1 gathers, without the redundancy: by Proposition 6.34 the map $\beta \mapsto \pi^\star_\beta$ is piecewise constant, so a fine grid recomputes each piece many times, whereas Algorithm 6.3 produces each policy once, together with the law of its return. That changes what one can do with it.

Proposition 6.38 (Planning on the front). Let $\Gamma = (\pi_k, I_k)_{k \le N_\Gamma}$ be the optimality front and ρ any functional (statistic) of the return distribution. Then

$$\max_{k \le N_\Gamma} \rho\big(\nu^{\pi_k}(s_1)\big) \;\ge\; \rho\big(\nu^{\pi^\star_\beta}(s_1)\big) \quad \text{for every } \beta \notin B,$$

where the finite set B consists of the breakpoints and of the isolated rates at which a class outside the front is also optimal (Definition 6.33), so that the left-hand side dominates what any method selecting a single rate—Algorithm 6.1 included—can return. Computing the left-hand side costs N_Γ evaluations of ρ on laws that Algorithm 6.3 has already produced.

This is the *generalized policy improvement* of Barreto et al. (2020) applied to the family of rates: solve the problems one can solve—the β-means, by Theorem 5.5 the only ones—then pick, among their solutions, the one that is best for the criterion one actually cares about. Three remarks fix its status.

First, the front is computed *once* and reused for every ρ. The expensive object is Γ; evaluating CVaR_δ, VaR_δ, a threshold probability or a distorted mean (Corollary D.5) on N_Γ known laws is arithmetic. Second, the inequality of Proposition 6.38 is all that is guaranteed: the front need not contain a ρ-optimal policy, and nothing here bounds the gap to $\sup_\pi \rho(\nu^\pi)$. What Theorem 6.31 bounds is the

error in the *surrogate* problem, and what Theorem 6.15(ii) bounds is the gap between the surrogate and the target; the front can only improve on the former, never repair the latter. Third, the gap is genuinely there, for two reasons. The front contains only policies that are β-optimal for some β, and a ρ-optimal policy need not be among them even when it is Markov (Example 6.39). And for criteria such as the value at risk or a threshold probability, the optimal policy may need the accumulated reward (Example 5.15); then no Markov policy is optimal, and Chapter 5's augmented state space remains the only exact route.

Threshold probabilities. The same Chernoff bound that produced the entropic value at risk produces a surrogate for the probability of disaster. For any $\beta < 0$ and any policy, $\mathbb{P}(W_H \le T) \le e^{-\beta T}\mathbb{E}[e^{\beta W_H}]$, so that

$$\min_{\pi} \mathbb{P}(W_H \le T) \;\le\; \inf_{\beta<0} \min_{\pi}\; e^{-\beta T}\,\mathbb{E}\big[e^{\beta W_H^{\pi}}\big], \tag{6.1}$$

and the inner minimization is again an exponential-utility plan. As for the entropic value at risk, a grid on β suffices: if every policy satisfies $\mathbb{P}(W_H \le T + a) \ge p$ for some $a < 0$ and $p > 0$—so that no rate beyond $\ln(1/p)/a$ can help—then a grid of step ε on $[\ln(1/p)/a, 0]$ returns a policy whose surrogate value is within a factor $\exp\big(\varepsilon(T - r_{\min})(r_{\max} - T)/\Delta\big) \le e^{\varepsilon\Delta/4}$ of the optimal surrogate, $\Delta = r_{\max} - r_{\min}$ being the range of the return (Marthe, 2026, Prop. 4.2). Here again the front does better for free, and the surrogate can be optimized exactly on each of its intervals: on I_k the map $\beta \mapsto e^{-\beta T}\mathbb{E}[e^{\beta W_H^{\pi_k}}]$ is convex, and the corresponding map for the entropic value at risk is quasiconcave, so one-dimensional optimization on finitely many intervals replaces a grid on the whole line.

Example 6.39 (Risk measures on the front of the hike). The front of the windy hike has 27 policies (Example 6.37), and every one of them comes with its exact return law. Optimizing each criterion over them—maximizing, or minimizing for the two probabilities of bad events—gives the table below, against the risk-neutral optimal policy—the one an agent ignoring risk would play.

	$\mathrm{CVaR}_{0.05}$	$\mathrm{CVaR}_{0.1}$	$\mathrm{CVaR}_{0.25}$	$\mathrm{VaR}_{0.1}$	$\mathbb{P}(W_H = 0)$	$\mathbb{P}(W_H \le 2)$
best on the front	1.00	1.50	2.81	4	0.018	0.073
risk-neutral optimal	0	0	2.17	0	0.153	0.154

At the levels where risk is the question the difference is not marginal: the risk-neutral policy falls into the pit with probability 0.153 and the best policy of the front with probability 0.018, eight times less, and the conditional value at risk at level 0.1 goes from 0—the risk-neutral hiker's worst tenth is entirely made of falls—to 1.50. At level $\frac{1}{2}$ the front reaches 4.17 against 4.08 for the risk-neutral policy—and 4.17 is the optimum over *all* policies, computed on the augmented state space in Example 6.32: when half the outcomes are allowed to count, the front already contains a policy that nothing can beat. Figure 6.5 (right) plots the whole curve. Note what the table does not say: nothing here claims these are the best achievable values, only the best on the front. Some of them are optimal nonetheless: the value 4 of $\mathrm{VaR}_{0.1}$ cannot be improved, since by the table of Example 5.20 no policy has $\mathbb{P}(W_H \ge 5) > 0.9$ (Corollary 5.19(ii)). Others are not: the smallest attainable $\mathbb{P}(W_H \le 2)$ is $1 - 0.95 = 0.05$ (Corollary 5.19(i) with $u = 3$), reached by a policy that is Markov in (t, s), since the accumulated reward carries no information here, but by none of the policies of the front.

6.10 Time consistency and nested risk measures

The criteria studied so far evaluate the *law of the total return*: they are law invariant, and static. A sequential decision maker may also want her evaluations at different times to be consistent with each other. Suppose that, at each time t, the reward-to-go $W_{t:H}$ is evaluated conditionally on the past by a risk measure ρ_t (for a static law-invariant ρ, take $\rho_t(W) = \rho$ of the conditional law of W given $\mathfrak{F}_t = \sigma(S_1, A_1, \dots, S_t)$). The family (ρ_t) is *time consistent* if for all t and all bounded rewards-to-go W, W' (measurable with respect to $\mathfrak{F}_{H+1}$),

$$\rho_{t+1}(W) \le \rho_{t+1}(W') \text{ a.s.} \quad \Longrightarrow \quad \rho_t(W) \le \rho_t(W') :$$

a continuation judged better in every state tomorrow is judged better today. For conditional risk measures that are monotone, normalized ($\rho_t(0) = 0$), local (the evaluation on an event of $\mathfrak{F}_t$ does not depend on the reward-to-go outside it) and translation invariant—properties that law invariance does not supply by itself—time consistency is equivalent to the tower property $\rho_t = \rho_t \circ \rho_{t+1}$ (Exercise 6.9), which the expectation and the β-means possess (Proposition 2.17(vii)) and which the values at risk lack: the plan of Exercise 5.5, or a CVaR-optimal plan, can prescribe tomorrow an action that tomorrow's evaluation rejects. A theorem of Kupper and Schachermayer (2009, Thm. 1.10) shows that this is no accident. It is an external result in its own setting, which we do not prove. The probability space is the standard one, $[0,1]^{\mathbb{N}}$ with the product of Lebesgue measures, filtered by the coordinates: a fresh uniform variable arrives at every step, over an unbounded horizon. A dynamic risk measure is a family of maps ρ_t from all bounded variables to the bounded $\mathfrak{F}_t$-measurable ones, normalized, translation invariant and monotone; law invariance is required of ρ_0 only, time consistency in the form $\rho_0(W) = \rho_0(\rho_t(W))$, and *relevance* means that ρ_0 strictly prefers 0 to the loss $-\varepsilon \mathbb{1}_A$, for every $\varepsilon > 0$ and every event A of positive probability (stated in our gain convention: the paper's ρ_t is our $-\rho_t$). Neither convexity nor continuity is assumed. In that setting, the only such families are the conditional entropic measures $\rho_t = U_\beta(\cdot \mid \mathfrak{F}_t)$, $\beta \in [-\infty, \infty)$, with the expectation at $\beta = 0$ and the essential infimum at $\beta = -\infty$ (the essential supremum is not relevant). They are concave—convex risk measures in the paper's sign convention—exactly for $\beta \le 0$, so these are the only law-invariant, time-consistent, relevant *convex* risk measures. Without translation invariance, the companion theorem (Kupper and Schachermayer, 2009, Thm. 1.4) gives, under strict monotonicity and a continuity condition, exactly the certainty equivalents $u^{-1}(\mathbb{E}[u(W)])$ of expected utility. It is not a statement about a given finite MDP, whose tree of histories may carry too little randomness for the classification to bite: with a single decision, for instance, time consistency is essentially vacuous and every static criterion satisfies it.

There is a standard way to manufacture time-consistent measures from any static ρ: apply it recursively, one step at a time. The *nested* (or *recursive*) risk measure of the return is defined backward by

$$\rho^{\text{nest}}_{H+1} \coloneqq 0, \qquad \rho^{\text{nest}}_t(h_t) \coloneqq \rho\Big(r_t(s_t, A_t, S_{t+1}) + \rho^{\text{nest}}_{t+1}(h_t, A_t, S_{t+1}) \,\Big|\, h_t\Big), \tag{6.2}$$

where $\rho(\cdot \mid h_t)$ is applied to the conditional law given the history. Nested measures are time consistent by construction, satisfy a Bellman equation, and can be optimized by backward induction with deterministic Markov optimal policies when ρ is monotone, translation invariant and satisfies the mixture bound $\rho(\sum_i p_i \mu_i) \le \max_i \rho(\mu_i)$ of (5.2)—which holds for CVaR_δ and for every optimized certainty equivalent, and is what excludes randomized decision rules (Ruszczyński, 2010): the proof of Theorem 5.8 then goes through with ρ in place of U_β. The price is law invariance: optimizing the nested measure built from CVaR_δ solves a *different* problem, not the static CVaR_δ problem by other means, and its optimal policy need not be optimal for the static criterion (Example 6.40).

Example 6.40 (Nested conditional value at risk depends on the information structure). Let $\rho = \mathrm{CVaR}_{1/2}$ and $H = 2$, and consider two ways of revealing a return W uniformly distributed on $\{0, 1, 2, 3\}$ with zero intermediate rewards. In the first tree, nothing is learned at time 1 and W is revealed at time 2: $\rho_1^{\mathrm{nest}} = \mathrm{CVaR}_{1/2}(W) = \frac{1}{2}$. In the second, time 1 reveals whether $W \in \{0, 3\}$ or $W \in \{1, 2\}$ (each with probability $\frac{1}{2}$), and time 2 reveals W: the inner evaluations are $\mathrm{CVaR}_{1/2}(\frac{1}{2}(\delta_0+\delta_3)) = 0$ and $\mathrm{CVaR}_{1/2}(\frac{1}{2}(\delta_1+\delta_2)) = 1$, and the outer one is $\mathrm{CVaR}_{1/2}(\frac{1}{2}(\delta_0+\delta_1)) = 0$. The same law of the total return receives the values $\frac{1}{2}$ and 0: the nested measure evaluates a *filtration*, not a distribution, and it compounds risk aversion at every stage. Iancu et al. (2015) analyze how far a nested measure can be from the static one it is built from and give tight bounds on the gap; for the entropic family, nesting changes nothing, by the tower property.

The agent thus faces a choice. Law-invariant criteria such as CVaR_δ assess what matters—the distribution of what she will get—but their optimal plans may not be Markov (Chapter 5), and they are in general time inconsistent, EVaR_δ included; nested criteria are Markov and time consistent but depend on how information arrives. The entropic risk measures sit at the intersection, and in the setting of Kupper and Schachermayer (2009) they alone among convex risk measures; and the entropic value at risk, the upper envelope of their Chernoff translates, inherits from them a planning algorithm with deterministic Markov optimal policies (Theorem 6.30), although it is not itself time consistent (Exercise 6.9).

Distributional viewpoint

Exponential moments give both a computable and a coherent risk measure. The entropic risk measure $U_{-\beta}$ is a convex risk measure with penalty $\mathrm{KL}(Q \,\|\, \mathbb{P})/\beta$ (Corollary 6.23), the one that dynamic programming can optimize (Chapter 5); together with the expectation these measures form, for $\beta \leq 0$ and on a rich enough probability space, the whole family of law-invariant time-consistent convex risk measures. Its lack of positive homogeneity means that β is a rate, tied to the scale of the returns. The entropic value at risk $\mathrm{EVaR}_\delta = \sup_\beta \{U_{-\beta} - \ln(1/\delta)/\beta\}$ is the tightest Chernoff bound on the δ-quantile, a coherent risk measure below the conditional value at risk, and the worst-case expectation over the relative-entropy ball of radius $\ln(1/\delta)$ (Theorem 6.24); its optimization reduces to a grid of exponential-utility plans with an explicit accuracy (Theorem 6.31), and among its optimal policies there is always a deterministic Markov one (Theorem 6.30). Its mirror image for a risk-seeking agent, the optimistic entropic value at risk $\overline{\mathrm{EVaR}}_\delta = \inf_{\beta>0}\{U_\beta + \ln(1/\delta)/\beta\}$, is the tightest Chernoff bound on the level reached with probability δ and the best case over the same ball (Section 6.5); unlike its counterpart it may call for a randomized policy, a coin thrown between two β-optimal plans at the rate where they tie (Proposition 6.22). The β-optimal policies form a finite family, the optimality front (Proposition 6.34), computed once by certified sweeps over the rates (Theorem 6.35, Algorithm 6.3) and then available for any risk measure by selection (Proposition 6.38). The same relative-entropy ball, explored in the optimistic direction, will produce the exploration bonuses of Chapter 15.

Hands-on

The return distributions produced by the distributional recursion of session H03 (*windy hike around the pit*), represented as pairs of vectors (atoms, probabilities), are all that is needed to compute the three risk measures of this chapter for any policy: the quantile function for VaR_δ and CVaR_δ, a one-dimensional maximization over β for EVaR_δ (compute $U_{-\beta}$ in the log domain). Verify the chain of inequalities of Theorem 6.15(ii) and reproduce Figure 6.2. The agent of H03

optimizes β-means through `setObjective`; wrapping it in the grid of Algorithm 6.1 turns it into an EVaR_δ-optimal agent, and Figure 6.4 can be reproduced from the values it returns.

Exercises

Exercise 6.1 (Properties of the value at risk). [verify] (a) Prove the monotonicity, translation equivariance, positive homogeneity and law invariance of VaR_δ. (b) Verify the numbers of Example 6.3(ii), and show that for X, Y i.i.d. with $\mathbb{P}(X=-1)=p$, $\mathbb{P}(X=0)=1-p$, the superadditivity inequality $\mathrm{VaR}_\delta(X+Y) \geq 2\,\mathrm{VaR}_\delta(X)$ fails exactly for $\delta \in (p, 2p-p^2]$. (c) Show that $\delta \mapsto \mathrm{VaR}_\delta(X)$ is nondecreasing and left-continuous. (d) Show that the inequality of Proposition 6.4 is strict for every X and every $\beta>0$, and that nevertheless $\mathrm{EVaR}_\delta(X)=\mathrm{VaR}_\delta(X)$ whenever $\delta \leq \mathbb{P}(X=\operatorname{ess\,inf} X)$.

Exercise 6.2 (Risk measures from sets of models). [verify] Let $\mathcal{Q}$ be a nonempty set of probability measures on a finite Ω. (a) Show that $\rho(X)=\inf_{Q\in\mathcal{Q}} \mathbb{E}_Q X$ is coherent, and that replacing $\mathcal{Q}$ by its closed convex hull does not change ρ. (b) Identify ρ when $\mathcal{Q}=\{\mathbb{P}\}$, when $\mathcal{Q}$ is the set of all probability measures, and when $\mathcal{Q}=\{Q : \mathrm{d}Q/\mathrm{d}\mathbb{P} \leq 1/\delta\}$. (c) Show that the set $\mathcal{Q}$ of Theorem 6.8 is the largest set representing ρ, and compute it for $\rho=\operatorname{ess\,inf}$ and for $\rho=\mathrm{CVaR}_\delta$.

Exercise 6.3 (The entropic risk measure as a convex risk measure). [verify] (a) Prove directly from Hölder's inequality that $Y \mapsto \ln \mathbb{E}e^{Y}$ is convex on bounded random variables. (b) Deduce that $U_{-\beta}$, $\beta>0$, is a convex risk measure, and show by an example that it is not positively homogeneous. (c) Using Corollary 6.23, show that the penalty function $\alpha(Q)=\mathrm{KL}(Q\,\|\,\mathbb{P})/\beta$ is the smallest one representing $U_{-\beta}$ in the form $\inf_Q\{\mathbb{E}_Q X+\alpha(Q)\}$.

Exercise 6.4 (Gaussian and two-point laws, after Ahmadi-Javid, 2012, Sect. 3). [verify] (a) Prove the three formulas of Example 6.16 and check the numerical values at $\delta=0.05$. (b) For $X \sim (1-p)\delta_0 + p\delta_1$, compute VaR_δ, CVaR_δ and show that $\mathrm{EVaR}_\delta(X)=0$ for $\delta \leq 1-p$ and $\mathrm{EVaR}_\delta(X)>0$ for $\delta>1-p$; compare with CVaR_δ in the latter case. (c) Show that for every law, $\mathrm{EVaR}_\delta(X)=\operatorname{ess\,inf} X$ as soon as $\delta \leq \mathbb{P}(X=\operatorname{ess\,inf} X)$, and interpret in the windy hike.

Exercise 6.5 (The optimistic entropic value at risk). [verify] (a) Prove Theorem 6.20 from Theorem 6.15, making the quantile functions of $-X$ explicit, and prove Proposition 6.18 directly from Markov's inequality. (b) Prove the three Gaussian formulas of Example 6.21. (c) Show that $\overline{\mathrm{EVaR}}_\delta(X)=\operatorname{ess\,sup} X$ if and only if $\delta \leq \mathbb{P}(X=\operatorname{ess\,sup} X)$, and check the levels 0.43 and 0.53 of Example 6.21 on the return distributions of Figure 3.2. (d) One state, two actions: a sure reward 1, or 2 with probability 0.3 and 0 otherwise. Show that the rate $\bar\beta$ at which the two actions have the same β-mean does not depend on δ and solves $\ln(0.7+0.3e^{2\bar\beta})=\bar\beta$ ($\bar\beta \approx 0.847$); compute, for $\delta=0.8$ and 0.9, the optimistic entropic values at risk of the two actions, the surrogate $\inf_\beta \bar h_\delta = 1+\ln(1/\delta)/\bar\beta$ of Proposition 6.22, and the probability with which the sure action must be played for the mixture to reach it. For which levels δ is the lottery, played deterministically, optimal? (e) Mirror the proof of Theorem 6.31: show that the grid of Algorithm 6.1, used at the rates $+\beta_k$, satisfies $\min_k \bar h_\delta(\beta_k) \leq \inf_{\beta>0} \bar h_\delta(\beta)+\varepsilon$.

Exercise 6.6 (The entropic value at risk determines the law). [verify] Let X be bounded and $\psi(t) \coloneqq t \ln \mathbb{E} e^{-X/t}$ for $t > 0$. (a) Show that ψ is convex (Lemma 6.14) and that $\mathrm{EVaR}_\delta(X) = \psi^*(\ln \delta)$ with $\psi^*(y) = \sup_{t>0}\{ty - \psi(t)\}$ the Legendre transform (Definition C.5). (b) Show that $\psi^*(y) = +\infty$ for $y > 0$, and deduce from the biconjugation theorem that $\ln \mathbb{E} e^{-\beta X} = \sup_{\delta \in (0,1]}\{\ln \delta - \beta\, \mathrm{EVaR}_\delta(X)\}$ for every $\beta > 0$, with the convention $\mathrm{EVaR}_1(X) = \mathbb{E}[X]$ (the value of the defining formula at $\delta = 1$). (c) Conclude that the map $\delta \mapsto \mathrm{EVaR}_\delta(X)$ determines the law of X.

Exercise 6.7 (Sums of independent returns). [verify] Let $X_1, \dots, X_n$ be independent. (a) Show that $U_{-\beta}(\sum_i X_i) = \sum_i U_{-\beta}(X_i)$ and that $\mathrm{EVaR}_\delta(\sum_i X_i) \geq \sum_i \mathrm{EVaR}_\delta(X_i)$. (b) Explain why $\mathrm{EVaR}_\delta(\sum_i X_i)$ can be computed from the cumulant generating functions of the X_i by a one-dimensional maximization, whereas $\mathrm{CVaR}_\delta(\sum_i X_i)$ requires the law of the sum (a convolution). (c) In the windy hike, the return is not a sum of independent variables; which structure replaces independence in the computation of EVaR_δ-optimal policies?

Exercise 6.8 (The grid). [verify] (a) Verify the bound on K in Theorem 6.31 and evaluate it for $\Delta = 14$, $\delta = 0.5$, $\varepsilon = 0.1$. (b) Show that for a fixed policy π, $t \mapsto U_{-1/t}(\nu^\pi(s)) - t \ln(1/\delta)$ is concave on $(0, \infty)$, and explain why h_δ need not be concave in β nor in $1/\beta$. (c) Propose an adaptive refinement of the grid that exploits the piecewise structure of $\beta \mapsto U^\star_{-\beta,1}(s)$ (a maximum of finitely many smooth functions), and discuss its guarantee.

Exercise 6.9 (Nested risk measures). [limits] (a) Verify the computations of Example 6.40. (b) Show that a family (ρ_t) of monotone, translation-invariant conditional risk measures with the tower property $\rho_t = \rho_t \circ \rho_{t+1}$ is time consistent, that the converse holds if the ρ_t are also normalized and local, and that for a law-invariant static ρ, time consistency of the induced family implies $\rho(W) = \rho(\rho(W \mid \mathfrak{G}))$ for every finite σ-algebra $\mathfrak{G}$. (c) Show that the nested version of U_β coincides with U_β of the total return. (d) Compute the nested and the static $\mathrm{EVaR}_{1/2}$ on the two trees of Example 6.40, and conclude that EVaR_δ is not time consistent.

Exercise 6.10 (CVaR planning in the hike). [implement] Using Corollary 6.12, compute

$$\max_\pi \mathrm{CVaR}_{1/2}\left(\nu^\pi(\mathrm{S})\right)$$

in the windy hike (the answer is 4.17, attained with $u = 6$ in the Rockafellar–Uryasev formula), and describe the optimal policy: how does it depend on the accumulated reward? Compare with the EVaR-optimal policy of Example 6.32 and with the two hand-made policies. The reference is exact: the accumulated reward of the hike takes finitely many values, so the augmented backward induction is a finite computation.

Exercise 6.11 (Continuity in the level and in the law). [limits] (a) Show that $y \mapsto \mathrm{EVaR}_{e^{-y}}(X)$ is convex and nonincreasing on $(0, \infty)$. (b) Show that $\delta \mapsto \mathrm{CVaR}_\delta(X)$ is continuous on $(0, 1)$ and that $\delta \mapsto \mathrm{VaR}_\delta(X)$ is not, in general. (c) Let $\nu_\varepsilon = (1 - \varepsilon)\delta_1 + \varepsilon\delta_0$, so that $\mathcal{W}_1(\nu_\varepsilon, \delta_1) = \varepsilon$. Using Theorem 6.24 with $Q = (1 - q)\delta_1 + q\delta_0$ and $q = \ln(1/\delta)/\ln(1/\varepsilon)$, show that $\mathrm{EVaR}_\delta(\nu_\varepsilon) \leq 1 - \ln(1/\delta)/\ln(1/\varepsilon)$ for ε small, whereas $\mathrm{EVaR}_\delta(\delta_1) = 1$. Conclude that EVaR_δ is not Lipschitz for $\mathcal{W}_1$, unlike CVaR_δ (Theorem 6.11), and interpret: a catastrophe of probability ε lowers the

entropic value at risk by an amount of order $1/\ln(1/\varepsilon)$, not ε.

Exercise 6.12 (EVaR is not Bellman optimizable). [limits] (a) Show that on the laws supported in a fixed finite set, EVaR_δ is a continuous function of the weights (*hint:* use Theorem 6.24 and the continuity of relative entropy on the simplex when the reference weights are positive, treating vanishing weights separately). (b) Show that a positively homogeneous law-invariant criterion ψ with $\psi(\delta_c) = c$ that is ordinally equivalent to some U_β must equal U_β, and deduce that EVaR_δ is not ordinally equivalent to any β-mean. (c) Deduce from Theorem 5.5 that EVaR_δ is not Bellman optimizable, and exhibit directly a violation of the independence property for $\mathrm{EVaR}_{1/2}$ (*hint:* adapt Exercise 5.4, using Exercise 6.4(c) to compute entropic values at risk of two-point laws).

Exercise 6.13 (Counting the breakpoints). [limits] (a) Let $\nu \neq \nu'$ be two laws supported on the same set of n points. Show that $\{\beta : U_\beta(\nu) = U_\beta(\nu')\}$ has at most $n-1$ elements (*hint:* the equation is $\sum_i (\nu_i - \nu'_i) e^{\beta x_i} = 0$; a nonzero combination of n distinct exponentials has at most $n-1$ zeros). (b) Deduce the bound $(n-1)|\mathcal{A}|^{2|\mathcal{S}|H}$ on the number of breakpoints of the optimality front. (c) Explain why this bound is pessimistic, and what the genericity argument after Proposition 6.36 does and does not imply: verify the two-action example given there, and explain why it shows that genericity alone gives no bound of order $|\mathcal{S}|H$. (d) On the windy hike, compare the 26 breakpoints found in $[-5, 5]$ with $|\mathcal{S}|H$ and with the worst-case bound; which of these comparisons is a guarantee, and which an observation?

Exercise 6.14 (The interval of optimality). [verify] (a) Prove the two-sided comparison used in Theorem 6.35: for $\beta < 0$, $0 < \epsilon < |\beta|$ and $X \in [r_{\min}, r_{\max}]$, $U_\beta(X) \leq U_{\beta+\epsilon}(X) \leq \frac{\beta}{\beta+\epsilon} U_\beta(X) + \frac{\epsilon}{\beta+\epsilon} r_{\min}$. (b) Deduce the single-state statement of Theorem 6.35. (c) Show that the certified half-width shrinks to 0 as a breakpoint is approached, so that Algorithm 6.2 must fall back on its minimal step; explain why it can nevertheless not skip a maximal interval of optimality longer than ϵ_β, and why it can skip a shorter one. (d) Implement Algorithm 6.2 on two actions with return laws $\frac{1}{2}(\delta_0 + \delta_1)$ and $\frac{99}{100}\delta_0 + \frac{1}{100}\delta_2$: find the breakpoint, and compare the number of evaluations with a grid of the same precision.

Exercise 6.15 (The ends of the front). [verify] (a) Prove that $U_\beta(\nu) \to \max \operatorname{supp} \nu$ as $\beta \to +\infty$ and $\to \min \operatorname{supp} \nu$ as $\beta \to -\infty$ (Remark 4.17), and deduce that $\pi^\star_\beta$ optimizes the best case and the worst case in the two limits. (b) Show the refinement of Proposition 6.36: among the policies maximizing the worst return, the limit as $\beta \to -\infty$ selects the one minimizing the probability of that return, then of the next, and so on (*hint:* expand $\mathbb{E}[e^{\beta W}]$ in powers of $e^{\beta \Delta x}$). (c) Give a two-action example where the lexicographic tie-breaking changes the answer.

Exercise 6.16 (Numerical study: the front of the hike). [implement] Run `hike_front.py` to reproduce Figure 6.5. (a) Vary the precision ϵ_β from 10^{-1} to 10^{-4} and plot the number of backward inductions used by the grid and by Algorithm 6.2: which one depends on ϵ_β, and how? (b) Vary the wind p and report the number of policies in the front: does a noisier environment produce more or fewer? (c) For each policy of the front, plot the probability of falling into the pit against its mean return: the front traces a frontier in this plane. Which of its policies would you choose, and on what grounds? Each backward induction is exact, but the front itself is resolution dependent:

take the run at $\epsilon_\beta = 10^{-4}$ as the reference, keeping in mind that intervals of optimality shorter than ϵ_β may be missed by both methods.

Bibliographical notes

Coherent risk measures were axiomatized by Artzner et al. (1999), whose representation theorem we proved in the finite case following the convex-analytic route of Rockafellar (1970); Delbaen (2002) extends it to general probability spaces, and convex risk measures were introduced by Föllmer and Schied (2002) and Frittelli and Rosazza Gianin (2002). The monograph of Föllmer and Schied (2016) is the standard reference, including the entropic risk measure and the law-invariant case studied by Kusuoka (2001).

The conditional value at risk and its variational representation are due to Rockafellar and Uryasev (2000) and Rockafellar and Uryasev (2002); its coherence was established by Acerbi and Tasche (2002). The entropic value at risk was introduced by Ahmadi-Javid (2012), together with its coherence (via the perspective argument of Lemma 6.14), its dual representation over a relative-entropy ball, the conjugacy of Exercise 6.6, the comparison with the conditional value at risk and the g-entropic family; optimized certainty equivalents are due to Ben-Tal and Teboulle (2007). The optimistic entropic value at risk of Section 6.5 is the entropic value at risk of the loss convention applied to a gain; the minimax argument of Proposition 6.22 is that of Theorem 6.24, used on the side of the policies. Our development in the gain convention, from the Chernoff bound of Chapter 2, follows the thread of this book rather than the historical order.

EVaR-optimal planning in MDPs by a grid of exponential-utility plans is due to Hau et al. (2023), who treat the discounted setting (Chapter 8) and establish the accuracy bound of Theorem 6.31 and the non-concavity of h_δ; the finite-horizon version given here needs no truncation. Time consistency of dynamic risk measures and risk-averse dynamic programming with nested measures are developed by Ruszczyński (2010); the characterization of the entropic family as the only law-invariant, time-consistent and relevant dynamic risk measures is due to Kupper and Schachermayer (2009), and the shortcomings of nested measures are analyzed by Iancu et al. (2015). Exponential utilities in MDPs go back to Howard and Matheson (1972); the conditional value at risk in MDPs, with state augmentation, to Bäuerle and Ott (2011).

The optimality front of Section 6.8, its structure, the interval of optimality of Theorem 6.35, the lexicographic description of its ends, the algorithms FindBreaks and Dolfin, planning on the front (Section 6.9) and the surrogate (6.1) for threshold probabilities are due to Marthe et al. (2025); Marthe (2026, Chap. 4) gives the proofs in full, the bound on the number of breakpoints, the genericity argument showing that a breakpoint generically changes one action at one state and one stage (with the heuristic, not a theorem, that the number of breakpoints then grows like the number of state–stage pairs), and experiments on an inventory problem, a cliff-walking grid and a synthetic MDP, the fronts of the first and the last containing 18 and 16 policies. The grid over the risk parameter that the front replaces is the method of Hau et al. (2023), and the selection step is the generalized policy improvement of Barreto et al. (2020).