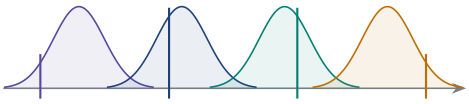
5

Utilities, Preferences and Distributional Planning



Planning for the expected return rests on two properties of the expectation, linearity and monotonicity, which make backward induction exact and deterministic Markov policies sufficient. This chapter asks which other criteria enjoy the same privilege. The most general dynamic programming scheme—propagate return distributions, choose greedily for the criterion at each state—is shown to be exact only for criteria that are compatible with mixtures and with shifts. Compatibility with mixtures is the independence axiom of von Neumann and Morgenstern, whose expected-utility theorem is proved in the setting of return distributions; compatibility with shifts then forces the utility to be exponential, by the closure theorem of the previous chapter. The exponential-utility planning algorithm is developed and illustrated on the windy hike, where the two hikers' attitudes turn out to be two ranges of β. The chapter ends with what can be done for the other criteria: counterexamples showing that memory may be necessary, and the state augmentation technique that restores dynamic programming for any expected utility, at a cost.

5.1 Preferences over return distributions

An agent who chooses between policies compares their return distributions: a *criterion* is a real-valued statistic in the sense of Section 4.2, that is, a functional $\psi : \mathcal{M}_1(\mathbb{R}) \to \mathbb{R}$ defined on laws with bounded support, and the agent prefers π to π' from state s when $\psi(\nu^\pi(s)) > \psi(\nu^{\pi'}(s))$. The mean, the β-means, the quantiles and the mean–variance criterion $\int x \, \mathrm{d}\nu - \theta \mathbb{V}\mathrm{ar}[\nu]$ are examples. Example 1.19 is the question in its simplest form: the sure unit δ_1 and the fair coin $\frac{1}{2}(\delta_0 + \delta_2)$ have the same mean, a β-mean with $\beta < 0$ prefers the first, one with $\beta > 0$ the second, the lower median $F_\nu^{-1}(\frac{1}{2})$ (lower quantile, Appendix D) prefers the first and the threshold probability $\nu([2, \infty))$ the second—each criterion encodes a different attitude toward the same two laws. Three properties of a criterion will matter.

Definition 5.1 (Monotone, continuous and independent criteria). A criterion ψ is

(i) *monotone* if $\nu \leq_{\mathrm{st}} \nu'$ implies $\psi(\nu) \leq \psi(\nu')$;

(ii) *continuous* if it is continuous for $\mathcal{W}_1$ on the laws supported in $[a, b]$, for every compact interval $[a, b]$;

(iii) *independent* (or: satisfies the *independence property*) if, for all finitely supported laws ν_1, ν_2, ν_3 and every $\lambda \in [0, 1]$,

$$\psi(\nu_1) \geq \psi(\nu_2) \quad \Longrightarrow \quad \psi\big(\lambda\nu_1 + (1-\lambda)\nu_3\big) \geq \psi\big(\lambda\nu_2 + (1-\lambda)\nu_3\big).$$

Monotonicity says that a sure improvement of the return is never disliked: by Proposition D.15, $\nu \leq_{\mathrm{st}} \nu'$ means that ν' can be obtained from ν by moving mass to the right. Continuity says that a small perturbation of the return distribution, in the sense of the Wasserstein distance, changes the criterion little; the mean, the β-means and the conditional values at risk (Theorem 6.11) are continuous, the quantiles are not (Exercise 5.3). Independence is the key axiom of expected-utility theory: if ν_1 is at least as good as ν_2, then replacing ν_2 by ν_1 inside a mixture cannot hurt, whatever the other component of the mixture. Applied twice (to $\psi(\nu_1) \geq \psi(\nu_2)$ and to $\psi(\nu_2) \geq \psi(\nu_1)$), it gives the *indifference form*: $\psi(\nu_1) = \psi(\nu_2)$ implies $\psi(\lambda\nu_1 + (1-\lambda)\nu_3) = \psi(\lambda\nu_2 + (1-\lambda)\nu_3)$. Utilities $\psi_f(\nu) = \int f \,\mathrm{d}\nu$ are independent because they are affine in ν; the converse is the content of the expected-utility theorem, which we prove in the form needed here. We say that two criteria ψ, ψ' are *ordinally equivalent* if $\psi(\nu) \geq \psi(\nu') \iff \psi'(\nu) \geq \psi'(\nu')$ for all ν, ν': they induce the same preferences.

Theorem 5.2 (Expected-utility theorem of von Neumann and Morgenstern). Let ψ be a monotone, continuous and independent criterion, not constant on Dirac masses. Then there exists a continuous nondecreasing nonconstant $f : \mathbb{R} \to \mathbb{R}$ such that ψ is ordinally equivalent to the utility ψ_f:

$$\psi(\nu) \geq \psi(\nu') \iff \int f \,\mathrm{d}\nu \geq \int f \,\mathrm{d}\nu' \qquad \text{for all laws } \nu, \nu' \text{ with bounded support.}$$

The function f is unique up to positive affine transformations $f \mapsto \alpha f + \gamma$, $\alpha > 0$.

Proof. Fix a compact interval $[a, b]$ such that $\psi(\delta_a) < \psi(\delta_b)$; such intervals exist since $x \mapsto \psi(\delta_x)$ is nondecreasing (monotonicity) and not constant. All laws in this proof are supported in $[a, b]$ until the last step. For $\lambda \in [0, 1]$, let $\nu_\lambda := (1-\lambda)\delta_a + \lambda\delta_b$ and $G(\lambda) := \psi(\nu_\lambda)$.

Step 1: G is a continuous strictly increasing bijection from $[0, 1]$ onto $[\psi(\delta_a), \psi(\delta_b)]$. Continuity follows from that of ψ, since $\mathcal{W}_1(\nu_\lambda, \nu_{\lambda'}) = |\lambda - \lambda'|(b-a)$, and monotonicity from $\nu_\lambda \leq_{\mathrm{st}} \nu_{\lambda'}$ for $\lambda \leq \lambda'$. Suppose $G(\lambda) = G(\lambda')$ for some $\lambda < \lambda'$, and write $\nu_{\lambda'} = (1-\kappa)\nu_\lambda + \kappa\delta_b$ with $\kappa = (\lambda' - \lambda)/(1-\lambda) \in (0, 1]$. The indifference form of independence allows ν_λ to be replaced by $\nu_{\lambda'}$ in this mixture: $G(\lambda') = \psi\big((1-\kappa)\nu_{\lambda'} + \kappa\delta_b\big) = G(\lambda'')$ with $1 - \lambda'' = (1-\kappa)(1-\lambda')$. Iterating, G takes the value $G(\lambda)$ at the points λ_n defined by $\lambda_0 = \lambda$ and $1 - \lambda_{n+1} = (1-\kappa)(1-\lambda_n)$, which increase to 1; being nondecreasing, G is constant on $[\lambda, 1)$, hence on $[\lambda, 1]$ by continuity. Mixing with δ_a instead of δ_b shows in the same way that G is constant on $[0, \lambda']$. Hence G is constant on $[0, 1]$, contradicting $\psi(\delta_a) < \psi(\delta_b)$.

Step 2: the calibrated criterion. Every ν supported in $[a, b]$ satisfies $\delta_a \leq_{\mathrm{st}} \nu \leq_{\mathrm{st}} \delta_b$, hence $\psi(\delta_a) \leq \psi(\nu) \leq \psi(\delta_b)$, so that $\Phi(\nu) := G^{-1}(\psi(\nu)) \in [0, 1]$ is well defined; it is the unique λ such that $\psi(\nu) = \psi(\nu_\lambda)$, and Φ is ordinally equivalent to ψ since G^{-1} is strictly increasing.

Step 3: Φ preserves mixtures of finitely supported laws. Let ν, ν' be finitely supported and $\kappa \in [0, 1]$. Since $\psi(\nu) = \psi(\nu_{\Phi(\nu)})$ and $\psi(\nu') = \psi(\nu_{\Phi(\nu')})$, two applications of the indifference form of independence give

$$\psi\big(\kappa\nu + (1-\kappa)\nu'\big) = \psi\big(\kappa\nu_{\Phi(\nu)} + (1-\kappa)\nu'\big) = \psi\big(\kappa\nu_{\Phi(\nu)} + (1-\kappa)\nu_{\Phi(\nu')}\big) = \psi\big(\nu_{\kappa\Phi(\nu)+(1-\kappa)\Phi(\nu')}\big),$$

the last equality being an identity between measures. By uniqueness in Step 2, $\Phi(\kappa\nu + (1-\kappa)\nu') = \kappa\Phi(\nu) + (1-\kappa)\Phi(\nu')$.

Step 4: the utility function. Set $f(x) := \Phi(\delta_x)$ for $x \in [a, b]$. By Step 3 and induction on the number of atoms, $\Phi(\sum_i p_i\delta_{x_i}) = \sum_i p_i f(x_i)$: $\Phi(\nu) = \int f \,\mathrm{d}\nu$ for every finitely supported ν. The function f is nondecreasing (monotonicity of ψ), continuous (as $G^{-1} \circ \psi \circ (x \mapsto \delta_x)$, a composition of continuous maps, since $\mathcal{W}_1(\delta_x, \delta_y) = |x - y|$), and $f(a) = 0$, $f(b) = 1$.

Step 5: extension to laws with bounded support. For ν supported in $[a, b]$, choose finitely supported $\nu_n \to \nu$ in $\mathcal{W}_1$ (for instance the quantile discretizations of Exercise 4.12). Then $\psi(\nu_n) = G(\int f \,\mathrm{d}\nu_n)$

Algorithm 5.1: Greedy distributional planning for a criterion ψ

Input: MDP $\mathcal{M}$, criterion ψ

$\eta_{H+1}(s) \leftarrow \delta_0$ for all $s \in \mathcal{S}$

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

  for $s \in \mathcal{S}$ **do**

    for $a \in \mathcal{A}$ **do**

      $\eta_t(s,a) \leftarrow \sum_{s'} P_t(s' \mid s,a)\,(b_{r_t(s,a,s')})_{\#}\eta_{t+1}(s')$

    $a_t^\star(s) \leftarrow$ any element of $\arg\max_a \psi(\eta_t(s,a))$; $\quad \eta_t(s) \leftarrow \eta_t(s, a_t^\star(s))$

Output: deterministic Markov policy $(a_t^\star)$; distributions (η_t), which are its return distributions

for all n; $\psi(\nu_n) \to \psi(\nu)$ by continuity of ψ, and $\int f \,\mathrm{d}\nu_n \to \int f \,\mathrm{d}\nu$ because f, being continuous on $[a,b]$, is the uniform limit of Lipschitz functions g, for which $|\int g \,\mathrm{d}\nu_n - \int g \,\mathrm{d}\nu| \le L\,\mathcal{W}_1(\nu_n,\nu)$ (Corollary D.5). Hence $\psi(\nu) = G(\int f \,\mathrm{d}\nu)$ for every ν supported in $[a,b]$, and, G being strictly increasing, $\psi(\nu) \ge \psi(\nu') \iff \int f \,\mathrm{d}\nu \ge \int f \,\mathrm{d}\nu'$ on such laws.

Step 6: independence of the interval and uniqueness. If f' is another continuous function such that $\psi_{f'}$ is ordinally equivalent to ψ on laws in $[a,b]$, then $f'(a) < f'(b)$ and the map $\Phi'(\nu) \coloneqq (\int f' \,\mathrm{d}\nu - f'(a))/(f'(b) - f'(a))$ is affine in ν, ordinally equivalent to ψ, and satisfies $\Phi'(\nu_\lambda) = \lambda$; hence $\Phi'(\nu) = \Phi(\nu)$ for all finitely supported ν (both equal the unique λ with $\psi(\nu) = \psi(\nu_\lambda)$), and $f' = f'(a) + (f'(b) - f'(a))f$ on $[a,b]$. This proves uniqueness up to positive affine maps, and shows that the functions obtained on two nested intervals agree on the smaller one after normalization. Fix $x_0 < x_1$ with $\psi(\delta_{x_0}) < \psi(\delta_{x_1})$ and normalize every f by $f(x_0) = 0$, $f(x_1) = 1$: the functions obtained on the intervals $[-n,n] \supset [x_0,x_1]$ are compatible and define f on $\mathbb{R}$. Two laws with bounded support are supported in a common interval, on which the equivalence has been proved. ∎

The theorem is a version of the classical result of von Neumann and Morgenstern (1944), in the axiomatization of Herstein and Milnor (1953), whose independence axiom is precisely the indifference form used above, stated for $\lambda = \frac{1}{2}$ only (the general λ then follows from their continuity axiom, Exercise 5.1(b)); monotonicity and $\mathcal{W}_1$-continuity replace their abstract continuity axiom and make the proof short. For the return distributions of finite-horizon MDPs, which are finitely supported, Steps 1–4 suffice.

5.2 Which criteria can be optimized by dynamic programming?

Given a criterion ψ, the most general dynamic programming scheme one can imagine propagates full return distributions, as in Algorithm 3.1, and chooses at each state the action whose return distribution is best for ψ (Algorithm 5.1). If ψ is the mean, this is Algorithm 3.3, which is exact (Theorem 3.15). For which other criteria is it exact?

Definition 5.3 (Bellman-optimizable criterion, Marthe et al., 2023). A criterion ψ is *Bellman optimizable* if, for every finite-horizon MDP and every choice of the maximizers in Algorithm 5.1, the output satisfies

$$\psi\big(\eta_t(s)\big) = \sup_{\pi \in \Pi_{\mathrm{MR}}} \psi\big(\nu_t^\pi(s)\big) \qquad \text{for all } t \in [H],\ s \in \mathcal{S}.$$

The supremum is taken over Markov policies; for the criteria characterized below it coincides with the supremum over all history-dependent policies Π_{H} (Theorem 5.8). The mean is Bellman optimizable, and so are the β-means, as we shall prove in Section 5.3. Two properties are necessary.

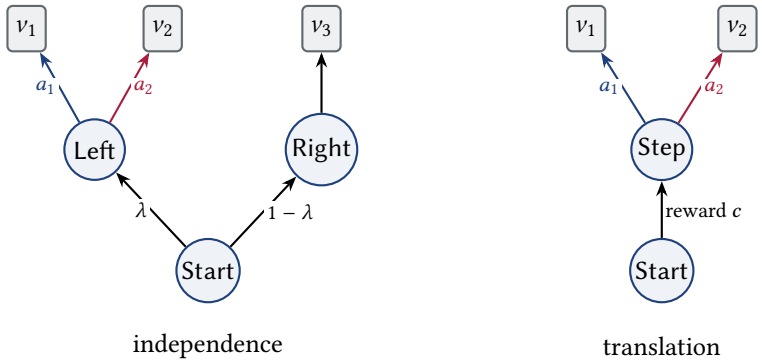


Figure 5.1. The two MDPs of Lemma 5.4, after Marthe et al. (2023). A gray box labeled ν is a *fan*: one action leading to absorbing terminal states with the probabilities and the rewards of the atoms of ν. Unlabeled transitions have reward 0.

The second involves shifts: we say that ψ satisfies the *translation property* if, for all finitely supported ν_1, ν_2 and every $c \in \mathbb{R}$, $\psi(\nu_1) \geq \psi(\nu_2)$ implies $\psi((b_c)_\# \nu_1) \geq \psi((b_c)_\# \nu_2)$. Applied to the two inequalities $\psi(\nu_1) \geq \psi(\nu_2)$ and $\psi(\nu_2) \geq \psi(\nu_1)$, the translation property implies the translation compatibility of Proposition 4.5, just as independence implies mixture compatibility (its indifference form): the order-preserving properties of this chapter strengthen the equality-preserving ones of Chapter 4, and the converse implications do not hold in general.

Lemma 5.4 (Necessary conditions, Marthe et al., 2023). A Bellman-optimizable criterion satisfies the independence property and the translation property.

Proof. Both proofs exhibit an MDP with horizon $H = 2$ in which Algorithm 5.1 fails (Figure 5.1). Finitely supported laws are realized as return distributions by *fans*: a state from which an action leads, with probabilities p_i, to absorbing terminal states (zero reward) through transitions of rewards x_i, realizes $\sum_i p_i \delta_{x_i}$. At states where "a single action" is available, all actions of $\mathcal{A}$ have the same effect.

Independence. Suppose $\psi(\nu_1) \geq \psi(\nu_2)$ but $\psi(\lambda\nu_1 + (1-\lambda)\nu_3) < \psi(\lambda\nu_2 + (1-\lambda)\nu_3)$ for finitely supported ν_1, ν_2, ν_3 and $\lambda \in (0,1)$. From the initial state Start, the single action leads with probability λ to Left and with probability $1-\lambda$ to Right, with zero reward. At Left, actions a_1 and a_2 are fans realizing ν_1 and ν_2; at Right, the single action realizes ν_3. At time 2, the algorithm compares $\psi(\eta_2(\text{Left}, a_1)) = \psi(\nu_1)$ with $\psi(\nu_2)$ and may select a_1 (it must if the inequality is strict; it may by the choice of the maximizer otherwise), so that $\eta_1(\text{Start}) = \lambda\nu_1 + (1-\lambda)\nu_3$. But the policy choosing a_2 has return distribution $\lambda\nu_2 + (1-\lambda)\nu_3$ from Start, strictly better for ψ: the output is not optimal.

Translation. Suppose $\psi(\nu_1) \geq \psi(\nu_2)$ but $\psi((b_c)_\# \nu_1) < \psi((b_c)_\# \nu_2)$. From Start, the single action leads to Step with reward c; at Step, actions a_1, a_2 realize ν_1, ν_2. The algorithm may select a_1 at time 2 and output $\eta_1(\text{Start}) = (b_c)_\# \nu_1$, whereas a_2 yields $(b_c)_\# \nu_2$, strictly better. ∎

Independence and Theorem 5.2 reduce continuous monotone criteria to utilities; the translation property and the closure theorem of Chapter 4 then leave only exponential ones.

Theorem 5.5 (Bellman-optimizable criteria). Let ψ be a monotone continuous criterion, not constant on Dirac masses. Then ψ is Bellman optimizable if and only if there exist $\beta \in \mathbb{R}$ and a strictly increasing function G such that $\psi = G \circ U_\beta$, that is, if and only if ψ is ordinally equivalent to a β-mean ($\beta = 0$ being the mean).

Proof. Necessity. By Lemma 5.4, ψ is independent; by Theorem 5.2, there is a continuous nondecreasing

nonconstant f with ψ ordinally equivalent to ψ_f on laws with bounded support. Let ν, ν' be finitely supported with $\int f\,\mathrm{d}\nu = \int f\,\mathrm{d}\nu'$. Then $\psi(\nu) = \psi(\nu')$; the translation property, applied in both directions, gives $\psi((b_c)_\# \nu) = \psi((b_c)_\# \nu')$ for every c, hence $\int f\,\mathrm{d}((b_c)_\# \nu) = \int f\,\mathrm{d}((b_c)_\# \nu')$: the utility ψ_f is translation compatible on finitely supported laws. The proof of Lemma 4.7 uses only finitely supported laws, so $f(x+c) = \alpha(c)f(x) + \gamma(c)$ for all x, c, and Steps 1–3 of the proof of Theorem 4.8, which start from exactly this functional equation, yield $f(x) = c_1 x + c_0$ or $f(x) = c_1 e^{\beta x} + c_0$ with $c_1 \neq 0$, $\beta \neq 0$. Since f is nondecreasing, $c_1 > 0$ in the first case and $c_1\beta > 0$ in the second. In the first case ψ_f is ordinally equivalent to the mean U_0. In the second, if $\beta > 0$ then $c_1 > 0$ and $\int f\,\mathrm{d}\nu \geq \int f\,\mathrm{d}\nu'$ iff $\int e^{\beta x}\,\mathrm{d}\nu \geq \int e^{\beta x}\,\mathrm{d}\nu'$ iff $U_\beta(\nu) \geq U_\beta(\nu')$; if $\beta < 0$ then $c_1 < 0$ and the first inequality is equivalent to $\int e^{\beta x}\,\mathrm{d}\nu \leq \int e^{\beta x}\,\mathrm{d}\nu'$, which is again $U_\beta(\nu) \geq U_\beta(\nu')$ because $u \mapsto \frac{1}{\beta}\ln u$ is decreasing. Thus ψ is ordinally equivalent to U_β. Finally, $G(u) \coloneqq \psi(\delta_u)$ is strictly increasing (ordinal equivalence with U_β, and $U_\beta(\delta_u) = u$), and $\psi(\nu) = \psi(\delta_{U_\beta(\nu)}) = G(U_\beta(\nu))$ since ν and $\delta_{U_\beta(\nu)}$ have the same β-mean.

Sufficiency. The maximizers in Algorithm 5.1 are the same for ψ and for U_β, so it suffices to treat $\psi = U_\beta$; this is Theorem 5.8 below. ∎

Remark 5.6 (Scope of the theorem ♦).

(i) Continuity is used only through Theorem 5.2, and for the necessity part on finitely supported laws—the return distributions of finite MDPs—only two consequences of it are needed: the continuity of $\lambda \mapsto \psi((1-\lambda)\delta_a + \lambda\delta_b)$ on two-point laws and that of $x \mapsto \psi(\delta_x)$ (Steps 1–4 of the proof of Theorem 5.2, then Steps 1–3 of the proof of Theorem 4.8). Discontinuous criteria such as quantiles are not covered by the theorem, but they must still satisfy the necessary conditions of Lemma 5.4, and quantiles do not (Example 5.14 below).

(ii) Monotonicity can be dropped at the price of a symmetric conclusion: a nonconstant continuous Bellman-optimizable criterion is ordinally equivalent to U_β or to $-U_\beta$ (an agent who *minimizes* the return). The proof combines the classical expected-utility theorem, which does not assume monotonicity (Herstein and Milnor, 1953), with the same translation argument; the monotone case treated in Theorem 5.2 is the one that matters for criteria.

(iii) Nothing is said about criteria that are not Bellman optimizable beyond the fact that Algorithm 5.1—and any algorithm that decides at each state on the basis of the next-state return distributions—can fail to find the optimal policy. This includes the distributional reinforcement learning algorithms of Chapters 12 and 13 when they are used greedily for a non-optimizable criterion: they may behave well in practice, but no optimality guarantee holds in general. Conversely, the failure of the greedy recursion does not mean that the criterion cannot be optimized at all: $\pi \mapsto \psi(\nu^\pi(s))$ is still a well-defined objective, it can be optimized exactly over an augmented state for expected utilities and quantiles (Section 5.5), and when it is a differentiable function of the parameters of a randomized policy, policy-gradient methods (Chapter 14) can climb it, with the local guarantees of that chapter.

5.3 Exponential-utility planning

We now prove the positive half of Theorem 5.5, and more: for the criterion U_β, backward induction over β-values yields a deterministic Markov policy that is optimal among all history-dependent randomized policies. The proof parallels that of Theorem 3.15, with the tower property of β-means (Proposition 2.17(vii)) in place of the tower property of conditional expectations.

For $\pi \in \Pi_{\mathrm{H}}$ and a history h_t, let $U^\pi_{\beta,t}(h_t)$ be the β-mean of $W_{t:H}$ under the law of the continuation of h_t by π, as in Section 3.3; for a Markov policy it is the β-value $U^\pi_{\beta,t}(s_t)$ of Definition 4.14. (We overload

the letter U as Section 3.3 overloads V: the argument—a history, a state, or a law as in $U_\beta(\nu)$—tells which object is meant.) Conditioning on (A_t, S_{t+1}) as in the proof of Corollary 4.16 gives the recursion

$$U^\pi_{\beta,t}(h_t) = U_\beta\Big(r_t(s_t, A, S') + U^\pi_{\beta,t+1}(h_t, A, S')\Big), \qquad U^\pi_{\beta,H+1} \equiv 0, \tag{5.1}$$

where $A \sim \pi_t(\cdot \mid h_t)$ and, given $A = a$, $S' \sim P_t(\cdot \mid s_t, a)$.

Definition 5.7 (Optimal β-values). For $\beta \neq 0$, define $U^\star_{\beta,H+1} \equiv 0$ and, for $t = H, \dots, 1$,

$$Q^\star_{\beta,t}(s,a) \coloneqq \frac{1}{\beta} \ln \sum_{s'} P_t(s' \mid s, a)\, e^{\beta[r_t(s,a,s') + U^\star_{\beta,t+1}(s')]}, \qquad U^\star_{\beta,t}(s) \coloneqq \max_{a \in \mathcal{A}} Q^\star_{\beta,t}(s,a).$$

A deterministic Markov policy is *β-greedy* if $\pi_t(s) \in \arg\max_a Q^\star_{\beta,t}(s,a)$ for all t, s.

Theorem 5.8 (Exponential-utility planning). Let $\beta \neq 0$.

(i) For every $\pi \in \Pi_{\mathrm{H}}$, $t \in [H+1]$ and $h_t \in \mathcal{H}_t$, $U^\pi_{\beta,t}(h_t) \le U^\star_{\beta,t}(s_t)$.

(ii) Every β-greedy policy $\pi^\star$ satisfies $U^{\pi^\star}_{\beta,t} = U^\star_{\beta,t}$ for all t. Hence $\sup_{\pi \in \Pi_{\mathrm{H}}} U^\pi_{\beta,t}(h_t) = U^\star_{\beta,t}(s_t)$ for every history h_t; in particular $\sup_{\pi \in \Pi_{\mathrm{H}}} U_\beta(\nu^\pi(s)) = \max_{\pi \in \Pi_{\mathrm{MD}}} U_\beta(\nu^\pi(s)) = U^\star_{\beta,1}(s)$, and U_β is Bellman optimizable.

Proof. We use two properties of β-means of finitely supported laws: monotonicity in the variable (Proposition 2.17(iv)), and the *mixture bound*

$$U_\beta\Big(\sum_i p_i \mu_i\Big) \le \max_i U_\beta(\mu_i), \tag{5.2}$$

which follows from $e^{\beta U_\beta(\sum_i p_i \mu_i)} = \sum_i p_i e^{\beta U_\beta(\mu_i)}$: with $y_i = U_\beta(\mu_i)$, the right-hand side is a weighted average of the numbers $e^{\beta y_i}$, hence lies between $\min_i e^{\beta y_i}$ and $\max_i e^{\beta y_i}$; since $y \mapsto e^{\beta y}$ and $u \mapsto \frac{1}{\beta} \ln u$ are both increasing for $\beta > 0$ and both decreasing for $\beta < 0$, in either case $\frac{1}{\beta}\ln$ of the average is at most $\max_i y_i$.

(i) By backward induction on t; the case $t = H+1$ is trivial. Assume the claim at $t+1$. In (5.1), replace $U^\pi_{\beta,t+1}(h_t, A, S')$ by the larger $U^\star_{\beta,t+1}(S')$: by monotonicity, $U^\pi_{\beta,t}(h_t) \le U_\beta(r_t(s_t, A, S') + U^\star_{\beta,t+1}(S'))$. The law of $r_t(s_t, A, S') + U^\star_{\beta,t+1}(S')$ is the mixture, with weights $\pi_t(a \mid h_t)$, of the laws μ_a of $r_t(s_t, a, S') + U^\star_{\beta,t+1}(S')$ under $S' \sim P_t(\cdot \mid s_t, a)$, and $U_\beta(\mu_a) = Q^\star_{\beta,t}(s_t, a)$. By (5.2), $U^\pi_{\beta,t}(h_t) \le \max_a Q^\star_{\beta,t}(s_t, a) = U^\star_{\beta,t}(s_t)$.

(ii) For a β-greedy policy, both inequalities are equalities: by induction $U^{\pi^\star}_{\beta,t+1} = U^\star_{\beta,t+1}$, and the mixture is concentrated on a maximizing action. Since $\pi^\star \in \Pi_{\mathrm{MD}}$, (i) and (ii) give the identity of the suprema. Finally, in Algorithm 5.1 run with $\psi = U_\beta$, induction shows that $\eta_{t+1}(s')$ is the return distribution of a β-greedy continuation, so that $U_\beta(\eta_t(s,a)) = Q^\star_{\beta,t}(s,a)$ by the exponential Bellman equation (Corollary 4.16); the maximizers are the β-greedy actions, and $U_\beta(\eta_t(s)) = U^\star_{\beta,t}(s)$, which is the supremum over Π_{MR}. ∎

Algorithm 5.2 has the cost $O(HS^2A)$ of Algorithm 3.3: risk sensitivity, in the exponential form, is free. Three remarks before the example.

Remark 5.9 (Limits and scaling).

Algorithm 5.2: Backward induction for β-optimal planning

Input: MDP $\mathcal{M}$, $\beta \neq 0$

$U_{H+1}(s) \leftarrow 0$ for all s
for $t = H, \dots, 1$ **do**
  for $s \in \mathcal{S}$ **do**
    for $a \in \mathcal{A}$ **do**
      $Q_t(s,a) \leftarrow \frac{1}{\beta} \ln \sum_{s'} P_t(s' \mid s,a) \exp\left(\beta[r_t(s,a,s') + U_{t+1}(s')]\right)$ ▷ *log-sum-exp*
    $U_t(s) \leftarrow \max_a Q_t(s,a)$; $\pi^\star_t(s) \leftarrow$ any $a \in \arg\max_a Q_t(s,a)$

Output: β-optimal policy $\pi^\star$, optimal β-values (U_t)

(i) As $\beta \to 0$, $Q^\star_{\beta,t} \to Q^\star_t$ (by backward induction, since $U_\beta(\mu) \to \int x\, \mu(\mathrm{d}x)$ for each of the finitely many laws involved), and for $|\beta|$ small enough every β-greedy policy is greedy for the mean, at every stage and state simultaneously. Indeed, there are finitely many triples (t,s,a); whenever a is not greedy for the mean at (t,s), it is beaten by some a' with a gap $Q^\star_t(s,a') - Q^\star_t(s,a) > 0$, and the smallest of these finitely many gaps exceeds $2\max_{t,s,a} |Q^\star_{\beta,t}(s,a) - Q^\star_t(s,a)|$ for $|\beta|$ small enough, so that a is not β-greedy either. The converse fails in general: among actions that tie for the mean, the sign of β decides, and different mean-optimal policies may be selected for $\beta > 0$ and $\beta < 0$. As $\beta \to +\infty$, $Q^\star_{\beta,t}(s,a) \to \max\{r_t(s,a,s') + U^\star_{\infty,t+1}(s') : P_t(s' \mid s,a) > 0\}$, where $U^\star_{\infty,t}$ denotes the limit of $U^\star_{\beta,t}$, the largest attainable return of Exercise 3.1: the agent plans for the best case ("max–max"). As $\beta \to -\infty$ the agent plans for the worst case ("max–min"), the criterion of a hiker who cannot afford the pit at any price. In both limits the optimal policy is again deterministic and Markov.

(ii) By Proposition 2.17(iii), multiplying all rewards by $\lambda > 0$ turns the β-optimal policy into the $\lambda\beta$-optimal policy of the original problem: β has the dimension of an inverse reward, and a risk attitude is a value of β *relative to the scale of the returns*. In the hike, where the return of every policy lies in $\{0, \dots, 8\}$ (the largest attainable return of Exercise 3.1, the $\beta \to +\infty$ limit of item (i)), $|\beta| = 1$ is a strong attitude; for returns in the thousands it would be extreme.

(iii) The optimal β-values are computed for all states at once, but they are the β-means of the return distributions of one policy, $\pi^\star$; they say nothing about the other statistics of these distributions. When these matter, the distributions of $\pi^\star$ can be computed afterwards by Algorithm 3.1.

Example 5.10 (The hikers as two values of β). Algorithm 5.2 applied to the windy hike of Example 3.6 produces, for each β, an optimal policy; Figure 5.2 shows the highest row visited by the path that this policy follows in the absence of wind, and the table gives the optimal β-value together with the mean and the probability of a zero return of the corresponding return distribution (computed afterwards by Algorithm 3.1):

β	-2	-1	-0.5	0	0.5	1	2
$U^\star_{\beta,1}(\mathrm{S})$	1.79	2.81	3.84	5.09	6.95	7.42	7.69
row of the path	4	4	3	3	2	2	2
mean of the return	3.18	3.72	4.90	5.09	5.02	5.02	5.02
$\mathbb{P}^{\pi^\star}_{\mathrm{S}}(W_H = 0)$	0.02	0.03	0.07	0.15	0.34	0.34	0.34

Three regimes appear, with transitions at $\beta \approx -0.81$ and $\beta \approx 0.017$. (The regimes are those of

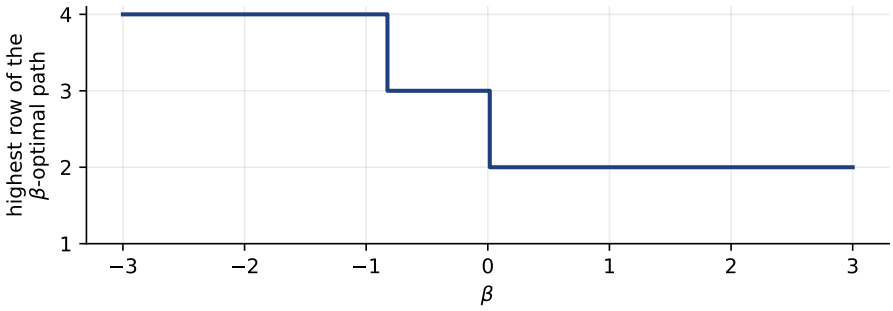


Figure 5.2. Highest row visited by the β-optimal path in the windy hike, as a function of β (Example 5.10): row 4 is the top of the grid, row 2 borders the pit.

the policy returned by Algorithm 5.2 with a fixed tie-breaking rule, described through its path from S; at a given β other β-optimal policies may differ from it at states this path does not visit, or induce the same return law, and only statements about the return law from S, or about the selected policy, are meaningful as β varies. The thresholds should moreover not be confused with the crossing $\beta_\times \approx -0.65$ of Example 4.18, which compared two hand-made policies: the reroute at -0.81 is between optimal policies, which do better than both, and the two numbers have no reason to coincide. A third object, the maximizing rate $\beta^\star(\delta)$ in the definition of the entropic value at risk, appears in Chapter 6.) For $\beta > 0.017$ the optimal policy is exactly the risky policy of Example 3.6: the child is an optimist, whose criterion rewards the upside of arriving early and discounts the pit. For $\beta < -0.8$ the optimal policy climbs to the top row, like the mother's safe policy—but it is not that policy: at $\beta = -1$ it reaches the bar with probability 0.97 against 0.78 for the hand-made safe policy, and its mean return is 3.72 against 2.43, because it adapts to the remaining time: early in the hike it climbs back after a gust, but when time runs short it proceeds east on the row where it stands, whereas the hand-made policy climbs back stubbornly and often arrives too late. Planning, not only prudence, makes the difference. In between, the mean-optimal path ($\beta = 0$) runs along the third row: a compromise that no fixed hiker had considered, with a mean return of 5.09, slightly above the child's 5.02, and a zero return with probability 0.15 instead of 0.34. The child's path is literally a β-optimal policy; the mother's attitude is a range of negative β, for which planning does better than her fixed rule; the mean is a third attitude, with its own path.

Backward induction computes optimal values, and it can carry a property of the value functions from one stage to the previous, hence a property of the optimal policies: this is how Theorem 3.22 was proved. The argument depends on the criterion only through the monotonicity of the one-step evaluation, and applies to every criterion that dynamic programming handles. For a criterion ψ and a bounded function V on $\mathcal{S}$, write

$$Q_t^{\psi,V}(s,a) \;\coloneqq\; \psi\Big(\mathcal{L}\big(r_t(s,a,S') + V(S')\big)\Big), \qquad S' \sim P_t(\cdot \mid s,a), \tag{5.3}$$

and define $V_{H+1}^{\psi} \equiv 0$, $Q_t^{\psi} \coloneqq Q_t^{\psi,V_{t+1}^{\psi}}$ and $V_t^{\psi} \coloneqq \max_a Q_t^{\psi}(\cdot,a)$ for $t = H,\dots,1$. For $\psi = \mathbb{E}$ these are the $Q_t^\star$ and $V_t^\star$ of Definition 3.14; for $\psi = U_\beta$, the $Q_{\beta,t}^\star$ and $U_{\beta,t}^\star$ of Definition 5.7; for a static risk measure ρ satisfying the assumptions of Section 6.10, $\psi = \rho$ gives the optimal values of the nested criterion (6.2). In these three cases a policy greedy for (Q_t^{ψ}) is optimal for the corresponding criterion among all policies of Π_{H} (Theorems 3.15 and 5.8 and Section 6.10).

Proposition 5.11 (Structured optimal policies). Let ψ be one of the three criteria above, $\mathcal{V}$ a set of bounded functions on $\mathcal{S}$ containing 0, and $\mathcal{D}$ a set of deterministic decision rules $d : \mathcal{S} \to \mathcal{A}$. Assume that for every t and every $V \in \mathcal{V}$, (i) the function $s \mapsto \max_a Q_t^{\psi,V}(s,a)$ belongs to $\mathcal{V}$, and (ii) some $d \in \mathcal{D}$ satisfies $d(s) \in \arg\max_a Q_t^{\psi,V}(s,a)$ for every s. Then $V_t^{\psi} \in \mathcal{V}$ for every t, and some optimal policy has all its decision rules in $\mathcal{D}$.

Proof. $V_{H+1}^{\psi} = 0 \in \mathcal{V}$. If $V_{t+1}^{\psi} \in \mathcal{V}$, then $V_t^{\psi} = \max_a Q_t^{\psi,V_{t+1}^{\psi}}(\cdot,a)$ belongs to $\mathcal{V}$ by (i), and (ii) provides $d_t \in \mathcal{D}$ greedy for Q_t^{ψ}. The policy $(d_1,\dots,d_H)$ is greedy for (Q_t^{ψ}), hence optimal. ∎

The work lies in checking (i) and (ii), and there the criterion only enters through its *monotonicity*: $\psi(\mathcal{L}(Y)) \le \psi(\mathcal{L}(Y'))$ whenever $Y \le Y'$ almost surely—for the β-means, Proposition 2.17(iv); for laws, by Proposition D.15, monotonicity for $\le_{\mathrm{st}}$.

Corollary 5.12 (Control limits under risk). In the machine replacement problem (Example 1.8), under the assumptions of Section 3.6, let ψ be the mean, a β-mean ($\beta \neq 0$), or a nested risk measure as above. For every t, V_t^{ψ} is nonincreasing in the wear, the set of wear levels at which replacing is greedy for Q_t^{ψ} is an upper set of $\mathbb{R}_+$, and the policy that replaces at time t if and only if S_t lies in this set is optimal for ψ among all policies.

Proof. The continuous state space is handled as in Section 3.6: the values lie in $[-H(C+c(0)),0]$, monotone functions are measurable, and there are two actions, so that the proofs of the optimality theorems apply verbatim. Apply Proposition 5.11 with $\mathcal{V}$ the bounded nonincreasing functions on $\mathbb{R}_+$ and $\mathcal{D}$ the rules "replace if and only if $x \in B$", B an upper set. Let $V \in \mathcal{V}$ and let X be the yearly increment. For $x \le x'$, $-c(x) + V(x+X) \ge -c(x') + V(x'+X)$ pointwise, since c is nondecreasing and V nonincreasing; by monotonicity of ψ, $x \mapsto Q_t^{\psi,V}(x,\mathrm{keep})$ is nonincreasing, whereas $Q_t^{\psi,V}(x,\mathrm{replace}) = \psi(\mathcal{L}(-C - c(0) + V(X)))$ does not depend on x. Hence the maximum of the two is bounded and nonincreasing, which is (i), and $B = \{x : Q_t^{\psi,V}(x,\mathrm{keep}) \le Q_t^{\psi,V}(x,\mathrm{replace})\}$ is an upper set, which is (ii). ∎

For $\psi = \mathbb{E}$ this is Theorem 3.22, with the same proof: linearity was never used. Structure is a matter of monotonicity, and the translation property, which decides whether a criterion can be optimized at all (Theorem 5.5), plays no role in it. The same argument on the augmented MDP of Section 5.5 gives control limits that depend on the cost already paid (Exercise 5.11).

Example 5.13 (The risk-averse owner replaces later). With the data of Section 3.6 (exponential increments of mean 1, $c(x) = x$, $C = 4$), backward induction on a fine grid gives the following β-optimal thresholds when 2, 3 and 25 years remain (with one year left, $x_H^{\star} = C = 4$ for every β, since both continuations vanish):

β	-1	-0.5	-0.2	0	0.2	0.5
$x_{H-1}^{\star}$	2.31	2.15	2.09	2.06	2.04	2.02
$x_{H-2}^{\star}$	2.46	2.12	1.90	1.78	1.67	1.55
$x_{H-24}^{\star}$	2.46	2.23	2.09	2.00	1.91	1.79

(script `replacement_risk.py`; the grid step 0.01 and the step 0.005 agree to 10^{-4}). The more risk-averse the owner, the *later* she replaces, and the direction is the same for $C = 2$ or 8 and for $c(x) = x^2/2$. Replacing a worn machine may look like the prudent choice, but the price of a new machine and the maintenance of the old one are both paid for sure; what differs is the

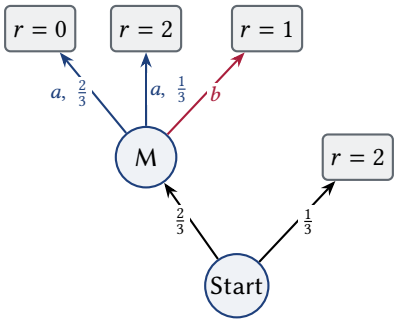


Figure 5.3. The MDP of Example 5.14: greedy planning for the median selects b at M (median 1 against 0), but a is optimal from Start (median 2 against 1).

continuation. Close to the long-run threshold, keeping means replacing next year for sure, since the wear can only increase: at $x^\star = 2.00$ in the risk-neutral problem, the continuation value $V^\star_{t+1}(x^\star + X)$ is a constant, whereas the continuation after a replacement, $V^\star_{t+1}(X)$—the future of a new machine, whose wear is unknown—has variance 1.73, and its β-mean at $\beta = -0.5$ lies 0.43 below its mean. A risk-averse owner discounts the uncertain future of the new machine, and keeps the old one longer; a risk-seeking owner does the opposite. Risk aversion is not caution in every coordinate: the bike store owners of Example 8.17 keep smaller stocks, the machine owner keeps older machines.

5.4 When dynamic programming fails

The negative half of Theorem 5.5 is best understood on examples. They show, in increasing order of gravity, that greedy planning may return a suboptimal policy, that the optimal policy may require memory, and that a criterion may be incompatible with sequential decision making altogether.

Example 5.14 (The median). Let $\psi(\nu) = F_\nu^{-1}(\frac{1}{2})$ be the lower median, and consider the MDP of Figure 5.3 with $H = 2$: from Start, the single action leads to M with probability $\frac{2}{3}$ and, with probability $\frac{1}{3}$, to an absorbing terminal state (zero reward afterwards) through a transition of reward 2. At M, action a yields reward 0 or 2 with probabilities $\frac{2}{3}, \frac{1}{3}$, and action b yields reward 1. Algorithm 5.1 compares, at M, $\psi(\frac{2}{3}\delta_0 + \frac{1}{3}\delta_2) = 0$ with $\psi(\delta_1) = 1$ and selects b; its output at Start is $\eta_1 = \frac{2}{3}\delta_1 + \frac{1}{3}\delta_2$, of median 1. The policy choosing a has return distribution $\frac{4}{9}\delta_0 + \frac{5}{9}\delta_2$, of median 2: greedy planning is suboptimal. The failure is one of independence: $\psi(\delta_1) \geq \psi(\frac{2}{3}\delta_0 + \frac{1}{3}\delta_2)$, yet mixing both with δ_2 at weight $\frac{1}{3}$ reverses the order. The median is not continuous, so Theorem 5.2 does not apply to it; Lemma 5.4 does.

Example 5.15 (The third moment: memory is needed). In the MDP of Figure 5.4, with $H = 3$, the agent first receives a reward $R_1 \in \{0, 1\}$ with probability $\frac{1}{2}$ each, moves on (with zero reward) to D, and chooses there between a (reward -1 or 1 with probabilities $\frac{2}{3}, \frac{1}{3}$) and b (reward 0). For the criterion $\psi(\nu) = \int x^3 \, \mathrm{d}\nu$, greedy planning selects b at D (0 against $-\frac{1}{3}$), the Markov policy a is better from Start ($\frac{7}{6}$ against $\frac{1}{2}$), and the history-dependent policy "a if $R_1 = 1$, b if $R_1 = 0$" is better still ($\frac{4}{3}$); see Exercise 5.2. Here the optimal policy is not Markov: the best action at D depends on the reward already collected, because the third moment of a sum is not determined by the third moments of its parts. The third moment is a utility, hence independent; it is the translation

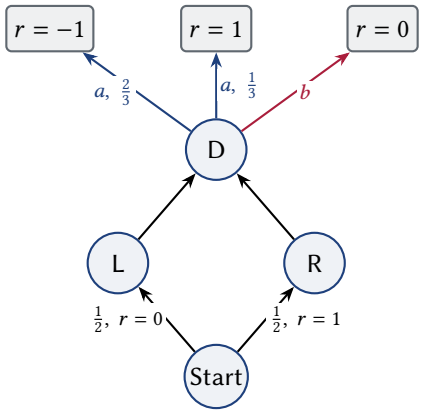


Figure 5.4. The MDP of Example 5.15: for the third moment of the return, the optimal action at D depends on the reward collected on the way.

property that fails. Threshold probabilities and quantiles may need memory too: in this MDP the policy "b if $R_1 = 1$, a if $R_1 = 0$" achieves $\mathbb{P}(W_H \geq 1) = \frac{2}{3}$, hence a lower median equal to 1, whereas every Markov policy, randomized or not, has $\mathbb{P}(W_H \geq 1) = \frac{1}{2} - \frac{q}{6} \leq \frac{1}{2}$ (q the probability of a at D), hence a lower median at most 0.

Example 5.16 (Mean–variance criteria). The criterion $\psi_\theta(\nu) = \int x \, \mathrm{d}\nu - \theta \mathbb{V}\mathrm{ar}[\nu]$, $\theta > 0$, is the historical formalization of risk aversion (Markowitz, 1952). It is not monotone: $\frac{1}{2}(\delta_0 + \delta_M) \geq_{\mathrm{st}} \delta_0$ for $M > 0$, yet $\psi_\theta(\frac{1}{2}(\delta_0 + \delta_M)) = \frac{M}{2} - \frac{\theta M^2}{4} < 0 = \psi_\theta(\delta_0)$ for $M > 2/\theta$. A sure gain can be refused because it comes with variance. It is not independent either, and the optimal policy may depend on the accumulated reward in a striking way: an agent may deliberately choose a *worse* continuation after a lucky start, to reduce the variance of the total (Exercise 5.5). Such a plan is *time inconsistent* (Section 6.10): re-evaluated from the intermediate state, the continuation is not what the agent would choose then. Mannor and Tsitsiklis (2013) show that variance-constrained planning is NP-hard in general.

The three examples display three ways of failing. In the median example, the optimal policy is deterministic and Markov, but greedy planning does not find it, because the criterion compares next-state distributions in a way that the mixture at the current state does not respect. In the third-moment example, no Markov policy is optimal: the accumulated reward is a relevant piece of information, which a Markov policy ignores. For the mean–variance criterion, the notion of an optimal continuation is itself problematic. The next section shows that the first two kinds of failure can be repaired by enlarging the state: directly for expected utilities such as the third moment, and for quantiles through threshold probabilities.

5.5 State augmentation

Utilities $\psi_f(\nu) = \int f \, \mathrm{d}\nu$—*expected utilities*, in the language of decision theory—are independent but, unless f is affine or exponential, not translation compatible: the value of a continuation depends on the reward already collected. Recording this reward in the state removes the obstacle.

Definition 5.17 (Augmented MDP). Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, (P_t), (r_t), H)$ and let $\mathcal{W}_t$ be the (finite) set of values that the accumulated reward $W_{1:t-1} = \sum_{k<t} r_k(S_k, A_k, S_{k+1})$ can take. The *augmented MDP* $\widetilde{\mathcal{M}}_f$ has states $(s, w) \in \mathcal{S} \times \mathcal{W}_t$ at time t, the same actions, kernels $\widetilde{P}_t\big((s', w + r_t(s, a, s')) \mid (s, w), a\big) = P_t(s' \mid s, a)$, zero rewards at times $1, \dots, H-1$, and the reward $\widetilde{r}_H\big((s, w), a, (s', w')\big) = f(w')$ on the last transition, so that its return is $f(W_H)$.

Histories of $\widetilde{\mathcal{M}}_f$ are histories of $\mathcal{M}$ with a redundant coordinate (the accumulated reward is a function of the history), so that the policy sets Π_{H} of the two MDPs coincide. In the backward induction below, the last-step reward is accounted for through the terminal values $\widetilde{V}_{H+1}(s, w) = f(w)$, a bookkeeping equivalent to the convention $V_{H+1} \equiv 0$ with the reward $f(w')$ on the last transition.

Theorem 5.18 (Planning for an expected utility). For every $f : \mathbb{R} \to \mathbb{R}$ (continuity is not needed here), $\sup_{\pi \in \Pi_{\mathrm{H}}} \mathbb{E}_s^{\pi}[f(W_H)]$ is attained by a policy of the form $\pi_t(h_t) = g_t(s_t, W_{1:t-1})$, computed by backward induction on $\widetilde{\mathcal{M}}_f$:

$$\widetilde{V}_{H+1}(s, w) = f(w), \qquad \widetilde{V}_t(s, w) = \max_a \widetilde{Q}_t(s, w, a),$$
$$\widetilde{Q}_t(s, w, a) = \sum_{s'} P_t(s' \mid s, a)\, \widetilde{V}_{t+1}\big(s', w + r_t(s, a, s')\big),$$

with $g_t(s, w) \in \arg\max_a \widetilde{Q}_t(s, w, a)$. The cost is $O(HS^2A \max_t |\mathcal{W}_t|)$; with integer rewards in $\{0, \dots, R_{\max}\}$, $|\mathcal{W}_t| \le HR_{\max} + 1$ and the cost is $O(H^2S^2AR_{\max})$.

Proof. Apply Theorem 3.15 to $\widetilde{\mathcal{M}}_f$, whose expected return under any policy is $\mathbb{E}[f(W_H)]$ under the same policy in $\mathcal{M}$; a deterministic Markov policy of $\widetilde{\mathcal{M}}_f$ is a policy of $\mathcal{M}$ depending on $(t, s_t, W_{1:t-1})$. ∎

Corollary 5.19 (Threshold probabilities and quantiles).

(i) For every $u \in \mathbb{R}$, $\max_{\pi \in \Pi_{\mathrm{H}}} \mathbb{P}_s^{\pi}(W_H \ge u)$ is computed by Theorem 5.18 with $f = \mathbb{1}_{[u,\infty)}$: the probability of reaching the set $\{w \ge u\}$ at the last stage of the augmented MDP, in the sense of Proposition 3.12.

(ii) For $\delta \in (0, 1)$, $\max_{\pi \in \Pi_{\mathrm{H}}} \mathrm{VaR}_\delta(\nu^\pi(s))$, with $\mathrm{VaR}_\delta(\nu) = F_\nu^{-1}(\delta)$ the lower δ-quantile (the *value at risk* of Definition 6.1), equals the largest $u \in \mathcal{W}_{H+1}$ such that $\max_\pi \mathbb{P}_s^{\pi}(W_H \ge u) > 1 - \delta$; it is found by at most $|\mathcal{W}_{H+1}|$ applications of (i).

The conditional value at risk, the mean of the worst fraction δ of the outcomes, is treated in the same way in Chapter 6 (Corollary 6.12), once its variational representation is available.

Proof. (i) is immediate. (ii) $F_\nu^{-1}(\delta) \ge u$ iff $F_\nu(x) < \delta$ for all $x < u$. Since $\nu = \nu^\pi(s)$ is supported in the finite set $\mathcal{W}_{H+1}$ and $u \in \mathcal{W}_{H+1}$, the supremum $\sup_{x<u} F_\nu(x) = F_\nu(u^-)$ is attained, so the condition is equivalent to $F_\nu(u^-) < \delta$, that is, to $\nu([u, \infty)) > 1 - \delta$. Hence $\mathrm{VaR}_\delta(\nu^\pi(s)) \ge u$ for some π iff $\max_\pi \mathbb{P}_s^{\pi}(W_H \ge u) > 1 - \delta$, and the quantile of a law supported in $\mathcal{W}_{H+1}$ belongs to $\mathcal{W}_{H+1}$. ∎

State augmentation is the planning counterpart of exact distributional evaluation: it costs exactly as much as Algorithm 3.1 on a lattice, and explodes in the same way with general rewards (Remark 3.9), since $|\mathcal{W}_t|$ is at most $(SA)^{t-1}$ and can be of that order. It is due to Bäuerle and Ott (2011) for the conditional value at risk and to Bäuerle and Rieder (2014) for general utilities; see also White (1988) for the earlier literature on probabilistic criteria.

Example 5.20 (Reaching the bar with high probability). In the windy hike, Corollary 5.19(i) gives the largest probability of spending at least u steps at the bar:

u	1	2	3	4	5	6	7	8
$\max_\pi \mathbb{P}^\pi_S(W_H \geq u)$	0.98	0.98	0.95	0.94	0.85	0.82	0.59	0.53
row of the optimal path	4	4	4	4	3	3	2	2
safe policy	0.78	0.75	0.47	0.43	0	0	0	0
risky policy	0.66	0.66	0.66	0.65	0.65	0.64	0.59	0.53

A hiker who only wants to reach the bar can do so with probability 0.98, along the top row; one who insists on five steps there still has probability 0.85, along the middle row—a policy that neither hiker had considered. For $u \geq 7$ the risky path is the only way, and the optimum is attained by it. By (ii), the optimal median is 8 (since $0.53 > \frac{1}{2}$, while no return exceeds 8): the child's path is median-optimal, an illustration of how much a single quantile can hide (its return is 0 with probability 0.34). Here the accumulated reward is positive only at the bar, which is absorbing, so the augmented policies are in fact Markov in (t, s); they nevertheless depend on u, and no single policy is optimal for all thresholds.

Example 5.21 (Pulling the goalie). A hockey team trailing by d goals near the end of a game may replace its goalkeeper by an extra skater: it then scores more often, and concedes far more often into the empty net. Following Puterman and Chan (2025, Sec. 3.9.1), after Morrison (1976) and Washburn (1991), the coach decides every $h = 5$ seconds, and the rates are those of the NHL seasons 2013–2020 compiled by Hall (2020): at full strength each team scores 2.25 goals per 60 minutes; with the goalie pulled, the trailing team scores 6.39 and concedes 19.16. In an interval of length h each team thus scores with probability $2.25\,h/3600$ in the first case, and the trailing team scores and concedes with probabilities $6.39\,h/3600$ and $19.16\,h/3600$ in the second (at most one goal per interval). The reward of a step is +1, −1 or 0 according as the team scores, concedes, or neither: the MDP has a single state and two actions, and the return W_H over the remaining time is the change of the goal difference. Each minute with the goalie pulled lowers $\mathbb{E}[W_H]$ by $(19.16 - 6.39)/60 \approx 0.21$ goal, so that the mean-optimal policy never pulls. But the team wants not to be behind at the end of regulation time, an event of probability $\mathbb{P}(W_H \geq d)$: this is Corollary 5.19(i) with $u = d$, and the augmented state $(t, W_{1:t-1})$ is the scoreboard—the time left and the current deficit. Backward induction (script `goalie.py`; 240 intervals for the last 20 minutes, deficits not capped, a tie not absorbing since the team may concede again) gives an optimal policy that pulls the goalie with a deficit of 1, 2 and 3 goals from 4 min, 8 min 20 s and 12 min 45 s before the end, never when the team is not behind; coaches pull on average at 1.4, 2.3 and 4 minutes (Puterman and Chan, 2025). With 5 minutes left,

	one goal behind		two goals behind	
policy	$\mathbb{P}$(not behind)	$\mathbb{E}[W_H]$	$\mathbb{P}$(not behind)	$\mathbb{E}[W_H]$
never pull	0.145	0	0.013	0
coaches' average times	0.193	−0.26	0.027	−0.43
optimal	0.220	−0.71	0.041	−1.00

The optimal policy gives away 0.7 goal on average and multiplies the probability of not losing in regulation time by 1.5: Example 1.19, with data, where the mean and a threshold probability prescribe opposite actions. For each deficit d the pull region is an interval of times, from the time given above until the last $d - 1$ intervals, during which the deficit can no longer be recovered and both actions are worth 0. The decision depends on the time and on the deficit together: the policy

is Markov in the augmented state, not in the time alone. (The model of Puterman and Chan, 2025 makes the tie absorbing, a different criterion—reaching a tie at some time—which pulls earlier, from 4 min 25 s; Exercise 5.12.)

5.6 Which policies suffice?

For the mean, deterministic Markov policies are optimal among all policies (Theorem 3.15); for other criteria, the examples of this chapter required memory. Randomization, the other ingredient of the general policies of Definition 1.13, is needed for other reasons: constraints, regularization, adversaries. A classical result shows that randomization never requires memory as long as the criterion only involves expectations.

Proposition 5.22 (Derman–Strauch). Let $\pi \in \Pi_{\mathrm{H}}$ and $\mu \in \mathcal{M}_1(\mathcal{S})$. Define $\pi' \in \Pi_{\mathrm{MR}}$ by $\pi'_t(a \mid s) \coloneqq \mathbb{P}^{\pi}_{\mu}(A_t = a \mid S_t = s)$ when $\mathbb{P}^{\pi}_{\mu}(S_t = s) > 0$, and arbitrarily otherwise. Then for every $t \in [H]$ the triples (S_t, A_t, S_{t+1}) have the same law under $\mathbb{P}^{\pi}_{\mu}$ and $\mathbb{P}^{\pi'}_{\mu}$. Consequently $\mathbb{E}^{\pi'}_{\mu}[\sum_t g_t(S_t, A_t, S_{t+1})] = \mathbb{E}^{\pi}_{\mu}[\sum_t g_t(S_t, A_t, S_{t+1})]$ for all functions $g_1, \dots, g_H$: the expected return, and the expectation of any other additive cost, are the same.

Proof. Let p_t and p'_t be the laws of S_t under the two policies; $p_1 = p'_1 = \mu$. If $p_t = p'_t$, then $\mathbb{P}^{\pi'}_{\mu}(S_t = s, A_t = a) = p_t(s)\pi'_t(a \mid s) = \mathbb{P}^{\pi}_{\mu}(S_t = s, A_t = a)$ for all (s, a) (both sides vanish when $p_t(s) = 0$), and by Theorem 1.14(iii), summed over the histories ending in s, both laws of (S_t, A_t, S_{t+1}) are $(s, a, s') \mapsto \mathbb{P}^{\pi}_{\mu}(S_t = s, A_t = a)P_t(s' \mid s, a)$. Summing over (s, a) gives $p_{t+1} = p'_{t+1}$, and the induction proceeds. The last claim follows by linearity of the expectation. ∎

The result is due to Derman and Strauch (1966); see also Puterman (1994, Thm. 5.5.1). It says that the occupation measures of Exercise 3.2 are the same for Π_{H} and Π_{MR}: every criterion that is a function of finitely many expected additive rewards—the expected return, expected constraints $\mathbb{E}[\sum_t c_t(S_t, A_t)] \le \kappa$—has the same supremum over Π_{H} and over Π_{MR}. (For the regularized criteria of Chapter 16, π' does at least as well as π: by convexity of the divergence, $\mathrm{KL}\big(\pi'_t(\cdot \mid s) \,\big\|\, \pi^{\mathrm{ref}}_t(\cdot \mid s)\big) \le \mathbb{E}^{\pi}_{\mu}\big[\,\mathrm{KL}\big(\pi_t(\cdot \mid H_t) \,\big\|\, \pi^{\mathrm{ref}}_t(\cdot \mid s)\big) \;\big|\; S_t = s\big]$, since $\pi'_t(\cdot \mid s)$ is the conditional mean of $\pi_t(\cdot \mid H_t)$.) It says nothing about the law of the return, which depends on the joint law of the trajectory and not only on its one-step marginals. In Example 5.15, the history-dependent policy π "a if $R_1 = 1$, b if $R_1 = 0$" has $\mathbb{P}^{\pi}_{\mathrm{Start}}(A_3 = a \mid S_3 = \mathrm{D}) = \frac{1}{2}$, so that π' plays a with probability $\frac{1}{2}$ at D: both have mean return $\frac{1}{3}$, but the third moment of the return is $\frac{4}{3}$ under π and $\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{7}{6} = \frac{5}{6}$ under π', and no Markov policy, randomized or not, exceeds $\frac{7}{6}$.

Conversely, randomization can be strictly necessary. With a single decision between an action of reward 1 and cost 1 and an action of reward 0 and cost 0, under the constraint that the expected cost be at most $\frac{1}{2}$, every deterministic feasible policy earns 0, whereas playing the first action with probability $\frac{1}{2}$ earns $\frac{1}{2}$; constrained MDPs, where this phenomenon is generic, are treated by linear programming in Section 8.5. The unique optimal policy of a regularized MDP is randomized (Theorem 16.1), and so are the equilibrium strategies of a zero-sum game without a pure saddle point (Example 17.6). Table 5.1 collects the answers met in the book.

5.7 Hidden states: the belief as a state ♦

State augmentation answered the question "what must the state carry?" for criteria other than the mean. The question returns when the agent sees only noisy signals of the state. The answer is then

Table 5.1. Smallest class of policies containing an optimal policy, for the criteria of the book (finite spaces; (s, w) is the state augmented by the accumulated reward $W_{1:t-1}$; "stationary" refers to infinite-horizon problems).

Criterion	Sufficient class	Where
Expected return, finite horizon	deterministic Markov	Theorem 3.15
Discounted return; average reward (unichain)	stationary deterministic	Theorems 8.3 and 9.8
β-mean; entropic value at risk	deterministic Markov; under discounting non-stationary, and ε-optimal for an infinite horizon	Theorems 5.8, 6.30 and 8.15
Nested risk measure	deterministic Markov	Section 6.10
Robust MDP, rectangular sets	stationary deterministic	Theorem 17.12
Expected utility, threshold probability, quantile, CVaR	deterministic, Markov in (s, w); Markov in s is not enough	Theorem 5.18 and Corollaries 5.19 and 6.12; Example 5.15
Mean–variance	history-dependent; optimal plans time inconsistent	Example 5.16 and Exercise 5.5
Hidden states, finite horizon: expected return; β-mean with observed rewards	deterministic, Markov in the belief b_t	Theorem 5.26; Theorem 5.34(iii)
Hidden states, β-mean with hidden rewards	deterministic, Markov in the tilted belief; b_t is not enough	Theorem 5.34; Exercise 5.15
Expected constraints	Markov randomized (stationary randomized under discounting); randomization may be needed	Proposition 5.22; Section 8.5
KL- or entropy-regularized return	Markov randomized; the optimum is randomized	Theorem 16.1
Zero-sum game, worst case over the opponent	randomized (mixed) strategies	Theorem 17.1 and Example 17.6

the conditional law of the hidden state given the observations, the *belief*: an augmentation of the information rather than of the criterion. For the mean, the belief restores backward induction, with a convex piecewise-linear value; for other criteria, the law of the return from a belief is a mixture, and this decides which properties survive.

The model. To lighten notation the kernels and rewards below do not depend on t; everything holds verbatim with time-dependent ones.

Definition 5.23 (Partially observable MDP). A *partially observable MDP* (POMDP) consists of an MDP $(\mathcal{S}, \mathcal{A}, P, r, H)$ as in Definition 1.10, a finite *observation space* $\mathcal{Y}$, an *observation kernel* $\kappa(\cdot \mid s, a, s') \in \mathcal{M}_1(\mathcal{Y})$, and an initial law $b_1 \in \mathcal{M}_1(\mathcal{S})$. The state $S_1 \sim b_1$ is not observed. At each time $t \in [H]$ the agent chooses A_t knowing only the *information* $I_t = (A_1, Y_2, A_2, \dots, A_{t-1}, Y_t)$ (I_1 is empty); then (S_{t+1}, Y_{t+1}) is drawn with probability $P(s' \mid S_t, A_t)\kappa(y \mid S_t, A_t, s')$, the reward $r(S_t, A_t, S_{t+1})$ counts in the return $W_H = \sum_{t \le H} r(S_t, A_t, S_{t+1})$, and Y_{t+1} is revealed. An *observation-based policy* is a sequence of maps $\pi_t : \mathcal{I}_t \to \mathcal{M}_1(\mathcal{A})$, $\mathcal{I}_t := (\mathcal{A} \times \mathcal{Y})^{t-1}$; their set is denoted $\Pi_{\mathcal{Y}}$. We write $M_{a,y} \in \mathbb{R}^{\mathcal{S} \times \mathcal{S}}$ for the matrix $M_{a,y}(s, s') = P(s' \mid s, a)\kappa(y \mid s, a, s')$, so that $\sum_y M_{a,y} = P_a$, and $r_a \in \mathbb{R}^{\mathcal{S}}$ for the vector $r_a(s) = r(s, a)$.

Convention on rewards. Rewards are *hidden*: they count, but are not revealed. They are *observed* if there is a function ρ such that $r(s, a, s') = \rho(a, y)$ whenever $M_{a,y}(s, s') > 0$, for instance when the observation contains the reward. For the mean the distinction is immaterial; for other criteria it is not (Theorem 5.34). An MDP is the case $\mathcal{Y} = \mathcal{S}$, $\kappa(y \mid s, a, s') = \mathbb{1}\{y = s'\}$. As in Theorem 1.14, the law $\mathbb{P}^{\pi}_{b_1}$

of $(S_1, A_1, Y_2, S_2, \dots, A_H, Y_{H+1}, S_{H+1})$ is given by the product formula

$$b_1(s_1)\prod_{t=1}^{H}\pi_t(a_t \mid i_t)\,M_{a_t,y_{t+1}}(s_t,s_{t+1}), \qquad i_t=(a_1,y_2,\dots,a_{t-1},y_t). \tag{5.4}$$

Example 5.24 (Replacement with hidden wear). Simplify the machine of Example 1.8 to two hidden wear levels, good (G) and worn (B). Each year a good machine becomes worn with probability $\theta = 0.15$, a worn one stays worn; *replace* (cost $C = 1.5$) installs a good machine at the beginning of the year, which then wears out like any other. A worn machine costs 1 per year, paid later as warranty claims: the cost is hidden. At the end of the year the owner sees the quality of the output, a defect occurring with probability 0.2 for a good machine and 0.7 for a worn one; the action *inspect* costs 0.25 more than *keep* and reveals the state at the end of the year. The belief is $p = b(\mathrm{B}) \in [0,1]$.

The Bayes filter. For $b \in \mathcal{M}_1(\mathcal{S})$, $a \in \mathcal{A}$ and $y \in \mathcal{Y}$, let

$$p(y \mid b,a) \coloneqq \sum_{s,s'} b(s)M_{a,y}(s,s'), \qquad \Phi(b,a,y) \coloneqq \frac{M_{a,y}^{\top}b}{p(y \mid b,a)} \quad \text{if } p(y \mid b,a) > 0,$$

the probability of observing y after action a from belief b, and the updated belief.

Proposition 5.25 (Bayes filter). Let $\pi \in \Pi_{\mathcal{Y}}$ and $i_t \in \mathcal{I}_t$ with $\mathbb{P}^{\pi}_{b_1}(I_t = i_t) > 0$. The conditional law $b_t(s) \coloneqq \mathbb{P}^{\pi}_{b_1}(S_t = s \mid I_t = i_t)$ does not depend on π, and $b_{t+1} = \Phi(b_t, a_t, y_{t+1})$ along $i_{t+1} = (i_t, a_t, y_{t+1})$. Moreover, for every a with $\pi_t(a \mid i_t) > 0$,

$$\mathbb{P}^{\pi}_{b_1}(Y_{t+1} = y \mid I_t = i_t, A_t = a) = p(y \mid b_t, a), \qquad \mathbb{E}^{\pi}_{b_1}\big[r(S_t,A_t,S_{t+1}) \,\big|\, I_t = i_t, A_t = a\big] = \langle r_a, b_t\rangle.$$

Proof. Let $\tilde b_1 = b_1$, $\tilde b_{k+1} = M^{\top}_{a_k,y_{k+1}}\tilde b_k$ along i_t, and $\Pi^{\pi}(i_t) = \prod_{k<t}\pi_k(a_k \mid i_k)$. Summing (5.4) over the states before t and over the future gives, by induction on t, $\mathbb{P}^{\pi}_{b_1}(I_t = i_t, A_t = a, S_t = s, S_{t+1} = s', Y_{t+1} = y) = \Pi^{\pi}(i_t)\pi_t(a \mid i_t)\tilde b_t(s)M_{a,y}(s,s')$. The policy factors do not depend on the hidden states and cancel in Bayes' formula, as in the Beta computation of Section 10.4: $b_t = \tilde b_t / \langle \mathbf{1}, \tilde b_t\rangle$; summing over s and normalizing gives the filter, summing over (s,s') the law of Y_{t+1}, and summing over y gives $\mathbb{P}^{\pi}_{b_1}(S_t = s, S_{t+1} = s' \mid I_t = i_t, A_t = a) = b_t(s)P(s' \mid s,a)$, hence the reward. ∎

The belief as a state. Let $\mathcal{B}_1 = \{b_1\}$ and $\mathcal{B}_{t+1} = \{\Phi(b,a,y) : b \in \mathcal{B}_t,\ a \in \mathcal{A},\ y \in \mathcal{Y},\ p(y \mid b,a) > 0\}$: the beliefs reachable at time t form a finite set, with at most $(A|\mathcal{Y}|)^{t-1}$ elements.

Theorem 5.26 (The belief MDP). Define on $\mathcal{M}_1(\mathcal{S})$ the functions $V_{H+1} \equiv 0$ and, for $t = H, \dots, 1$,

$$Q_t(b,a) \coloneqq \langle r_a, b\rangle + \sum_{y:\,p(y|b,a)>0} p(y \mid b,a)\,V_{t+1}\big(\Phi(b,a,y)\big), \qquad V_t(b) \coloneqq \max_{a\in\mathcal{A}} Q_t(b,a). \tag{5.5}$$

Then $\sup_{\pi\in\Pi_{\mathcal{Y}}}\mathbb{E}^{\pi}_{b_1}[W_H] = V_1(b_1)$, and the supremum is attained by the deterministic policy $A_t = g_t(b_t)$, for any choice of $g_t(b) \in \arg\max_a Q_t(b,a)$: the belief is a sufficient statistic of the information, and only the values of V_t on the finite set $\mathcal{B}_t$ are used.

Proof. We reduce the problem to a finite MDP $\bar{\mathcal{M}}$, the tree of informations. Its states at time t are the $i_t \in \mathcal{I}_t$ (the disjoint union of $\mathcal{I}_1, \dots, \mathcal{I}_{H+1}$ is a finite state space); from i_t, action a leads to (i_t, a, y) with probability $p(y \mid b_t(i_t), a)$ and earns $\langle r_a, b_t(i_t)\rangle$, where $b_t(i_t)$ is computed from b_1 by the

filter along i_t (it is defined on the informations of positive probability, the only ones that $\bar{\mathcal{M}}$ visits; elsewhere let it be arbitrary). A policy of $\Pi_{\mathcal{Y}}$ is a Markov policy of $\bar{\mathcal{M}}$, whose state is the whole information. By Proposition 5.25, the law of $(I_1, A_1, \dots, I_H, A_H, I_{H+1})$ under $\mathbb{P}^{\pi}_{b_1}$ satisfies the conditions of Theorem 1.14 for $\bar{\mathcal{M}}$, hence is the law of the trajectory of $\bar{\mathcal{M}}$ under π; and by the tower property, $\mathbb{E}^{\pi}_{b_1}[W_H] = \sum_t \mathbb{E}^{\pi}_{b_1}[\langle r_{A_t}, b_t(I_t)\rangle]$ is the expected return of π in $\bar{\mathcal{M}}$. By Theorem 3.15 applied to $\bar{\mathcal{M}}$, the supremum is the optimal value $\bar{V}^{\star}_1(i_1)$, attained by every greedy deterministic Markov policy. Finally $\bar{V}^{\star}_t(i_t) = V_t(b_t(i_t))$, by backward induction: if it holds at $t+1$, then, since $b_{t+1}(i_t, a, y) = \Phi(b_t(i_t), a, y)$, the optimal state–action value of $\bar{\mathcal{M}}$ at (i_t, a) is $Q_t(b_t(i_t), a)$, and the actions $g_t(b_t(i_t))$ are greedy in $\bar{\mathcal{M}}$. ∎

The belief MDP of the literature has the whole simplex as state space (Åström, 1965), which requires measurable policies on a continuum; the proof uses only the finite part reachable from b_1. In the example, $|\mathcal{B}_{12}| = 4096$ from $b_1 = \delta_{\mathrm{G}}$ (the bound is $6^{11} \approx 3.6 \cdot 10^8$; `hidden_replacement.py`).

Piecewise-linear convex values. The recursion (5.5) has a finite representation on the whole simplex.

Theorem 5.27 (Smallwood–Sondik). Let $\Gamma_{H+1} = \{0\} \subset \mathbb{R}^{\mathcal{S}}$ and, for $t = H, \dots, 1$,

$$\Gamma_t \coloneqq \Big\{ r_a + \sum_{y \in \mathcal{Y}} M_{a,y} \alpha_y \,:\, a \in \mathcal{A},\ (\alpha_y)_{y \in \mathcal{Y}} \in \Gamma_{t+1}^{\mathcal{Y}} \Big\}.$$

Then, for every t and every $b \in \mathcal{M}_1(\mathcal{S})$,

$$V_t(b) = \max_{\alpha \in \Gamma_t} \langle \alpha, b \rangle, \qquad Q_t(b,a) = \langle r_a, b \rangle + \sum_{y \in \mathcal{Y}} \max_{\alpha \in \Gamma_{t+1}} \langle M_{a,y}\alpha, b \rangle.$$

Hence V_t and each $Q_t(\cdot, a)$ are convex and piecewise linear, and $|\Gamma_t| \le A\,|\Gamma_{t+1}|^{|\mathcal{Y}|}$. If $\alpha = r_a + \sum_y M_{a,y}\alpha_y$ attains the maximum at b, then a is optimal at (t, b): the simplex is covered by finitely many polyhedra on each of which V_t is linear and one action is optimal.

Proof. By backward induction; the case $t = H+1$ is trivial. If $V_{t+1} = \max_{\Gamma_{t+1}} \langle \alpha, \cdot \rangle$, then

$$p(y \mid b, a)\, V_{t+1}\big(\Phi(b,a,y)\big) = \max_{\alpha \in \Gamma_{t+1}} \langle \alpha, M_{a,y}^{\top} b \rangle = \max_{\alpha \in \Gamma_{t+1}} \langle M_{a,y}\alpha, b \rangle :$$

the normalization of the filter cancels the observation probability. When $p(y \mid b, a) = 0$, the nonnegative vector $M_{a,y}^{\top} b$ has zero sum, so both sides vanish and the sum in (5.5) may run over all y. A sum of maxima over independent choices is the maximum of the sum, so $Q_t(b,a) = \max_{(\alpha_y)} \langle r_a + \sum_y M_{a,y}\alpha_y, b \rangle$, and $V_t = \max_a Q_t(\cdot, a)$. If $\langle \alpha, b \rangle = V_t(b)$ for α built from a, then $V_t(b) \ge Q_t(b,a) \ge \langle \alpha, b \rangle = V_t(b)$. The set where α is maximal, $\{b : \langle \alpha - \alpha', b \rangle \ge 0\ \forall \alpha' \in \Gamma_t\}$, is a polyhedron. ∎

The theorem is due to Smallwood and Sondik (1973). Each $\alpha \in \Gamma_t$ is the value of a *conditional plan*: an action a followed, for each observation y, by a plan for the remaining stages (a decision tree on observations, executed without looking at the state): conditioning on (Y_{t+1}, S_{t+1}), its expected return from $S_t = s$ is $r_a(s) + \sum_{y,s'} M_{a,y}(s,s')\alpha_y(s')$, with α_y the values of the subplans. From b a plan yields $\langle \alpha, b \rangle$, linear because the plan does not see the state, and the best observation-based behavior is the best plan. There are $A^{(|\mathcal{Y}|^n - 1)/(|\mathcal{Y}| - 1)}$ plans with n stages: the backward representation is doubly exponential in the remaining horizon, the forward one ($\mathcal{B}_t$) exponential in the elapsed time. Exact algorithms prune the vectors that are nowhere maximal, with one linear program per vector (Kaelbling

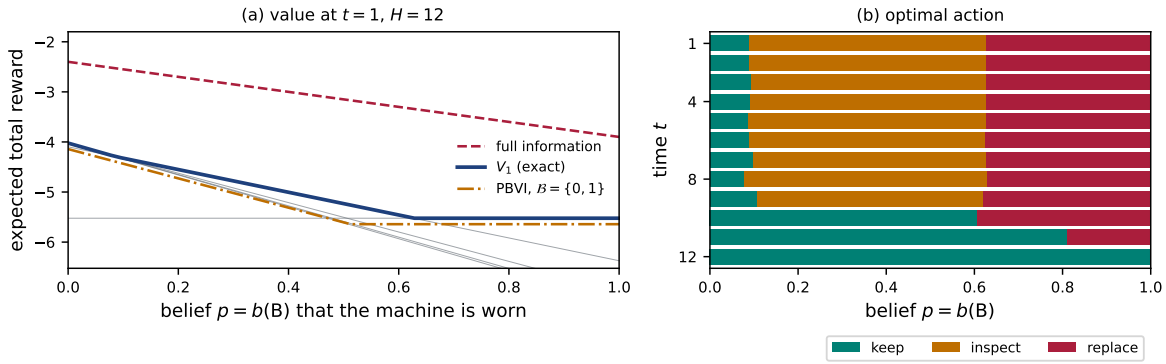


Figure 5.5. Replacement with hidden wear (Example 5.24, $H = 12$). (a) The optimal value V_1 as a function of the belief $p = b(\mathrm{B})$, upper envelope of the five vectors of Γ_1 (thin lines); the full-information bound of Proposition 5.29 (dashed) and the point-based lower bound of Algorithm 5.3 with $\mathcal{B} = \{0, 1\}$ (dash-dotted). (b) Optimal action at each time and belief: the replace region is an interval, as Proposition 5.28 predicts, here of the form $[p_t^\star, 1]$; inspection pays only when enough decisions remain. Script `hidden_replacement.py`.

et al., 1998). Neither growth can be avoided in general: the finite-horizon planning problem is PSPACE-complete (Papadimitriou and Tsitsiklis, 1987), and in infinite horizon the existence of a policy reaching a given value is undecidable (Madani et al., 2003).

In Example 5.24, after the removal of the vectors that are nowhere maximal, $|\Gamma_t| \le 5$ for every t, while the bound of the theorem reaches $3 \cdot 5^2 = 75$ (Figure 5.5). With twelve years to go, the owner keeps the machine when $p < 0.09$, inspects when $0.09 < p < 0.63$ and replaces beyond. Inspection, which buys information, is never optimal in the last three years: there is too little time left to use what it reveals.

Resetting actions. Replacing a machine erases what is known about it, and this has a geometric consequence.

Proposition 5.28 (A convex region for a resetting action). Suppose that the law of (S_{t+1}, Y_{t+1}) under an action a_0 does not depend on the current state, $M_{a_0,y}(s, s') = m_y(s')$ for all s, and that $r_{a_0}(s) = c_0$ is constant. Then $Q_t(\cdot, a_0)$ is constant on $\mathcal{M}_1(\mathcal{S})$, and the set of beliefs at which a_0 is optimal at time t is convex. With two hidden states, it is an interval of values of p.

Proof. The vector $M_{a_0,y}\alpha$ has all its entries equal to $\sum_{s'} m_y(s')\alpha(s')$, so $\langle M_{a_0,y}\alpha, b\rangle$ does not depend on b, and $Q_t(\cdot, a_0)$ is a constant q_t by Theorem 5.27. The set where a_0 is optimal is $\{b : \max_{a \ne a_0} Q_t(b, a) \le q_t\}$, a sublevel set of a convex function by the same theorem. ∎

This is the belief counterpart of Theorem 3.22, and weaker: there the replace set was an upper set of wear levels; here it is convex, an interval that need not contain $p = 1$. That it is of the form $[p_t^\star, 1]$, a control limit in the belief, requires monotonicity of the value for a likelihood-ratio order on beliefs, under conditions on the kernels and observations (Lovejoy, 1987; Krishnamurthy, 2016). In the example it holds, with $p_t^\star \in [0.62, 0.63]$ for $t \le 9$, $p_{10}^\star = 0.61$, $p_{11}^\star = 0.81$, and no replacement in the last year.

Bounds. Every finite set $\mathcal{P}$ of plans gives a lower bound $\max_{\alpha \in \mathcal{P}} \langle \alpha, b\rangle \le V_t(b)$, attained by executing the best plan of $\mathcal{P}$. Upper bounds come from the observed problem.

Algorithm 5.3: Point-based value iteration (finite horizon)

Input: POMDP, finite belief sets $\mathcal{B}_1, \dots, \mathcal{B}_H \subset \mathcal{M}_1(\mathcal{S})$

$\hat{\Gamma}_{H+1} \leftarrow \{0\}$
for $t = H, \dots, 1$ **do**
  $\hat{\Gamma}_t \leftarrow \emptyset$
  for $b \in \mathcal{B}_t$ **do**
    for $a \in \mathcal{A}$ **do**
      for $y \in \mathcal{Y}$ **do** $\alpha_{a,y} \leftarrow$ any $\alpha \in \arg\max_{\alpha \in \hat{\Gamma}_{t+1}} \langle M_{a,y}\alpha, b\rangle$
      $\alpha_{b,a} \leftarrow r_a + \sum_y M_{a,y}\alpha_{a,y}$
    add $\alpha_{b,a}$ for some $a \in \arg\max_a \langle \alpha_{b,a}, b\rangle$ to $\hat{\Gamma}_t$, labeled by a

Output: $\hat{V}_t(b) = \max_{\alpha \in \hat{\Gamma}_t} \langle \alpha, b\rangle$; policy: at b_t, play the label of a maximizing α

Proposition 5.29 (Value of full information). (i) For every $\pi \in \Pi_{\mathcal{Y}}$ and b_1, there is $\pi' \in \Pi_{\mathrm{H}}$ for the MDP $(\mathcal{S}, \mathcal{A}, P, r, H)$ such that the state–action trajectory $(S_1, A_1, \dots, S_{H+1})$, hence the return, has the same law under $\mathbb{P}^{\pi}_{b_1}$ as under $\mathbb{P}^{\pi'}_{b_1}$. (ii) Consequently $V_t(b) \le \sum_s b(s) V_t^\star(s)$ for every t and b, where $V_t^\star$ are the optimal values of the MDP, in which the state is observed.

Proof. (i) Let $H_t = (S_1, A_1, \dots, S_t)$ and $\pi'_t(a \mid h_t) \coloneqq \mathbb{P}^{\pi}_{b_1}(A_t = a \mid H_t = h_t)$ when $\mathbb{P}^{\pi}_{b_1}(H_t = h_t) > 0$, arbitrary otherwise. Summing (5.4) over y_{t+1} shows that $\mathbb{P}^{\pi}_{b_1}(S_{t+1} = s' \mid H_t, I_t, A_t) = P(s' \mid S_t, A_t)$, hence, by the tower property, $\mathbb{P}^{\pi}_{b_1}(S_{t+1} = s' \mid H_t, A_t) = P(s' \mid S_t, A_t)$. The law of the state–action trajectory thus satisfies (i)–(iii) of Theorem 1.14 for π', and is $\mathbb{P}^{\pi'}_{b_1}$ by uniqueness. (ii) $V_t(b)$ is the optimal value of the POMDP started at time t from b (Theorem 5.26); by (i) and Theorem 3.15, each of its policies earns at most $\sum_s b(s) V_t^\star(s)$. ∎

Point-based value iteration. The lower bound by plans becomes an algorithm if the plans are chosen by backups at a finite set of beliefs (Pineau et al., 2003): at each b of a set $\mathcal{B}_t$, keep only the vector of Γ_t that the exact backup would select at b, built from the current vectors.

A sweep costs $O(|\mathcal{B}_t|\, A\, |\mathcal{Y}|\, |\hat{\Gamma}_{t+1}|\, S^2)$ operations, and $|\hat{\Gamma}_t| \le |\mathcal{B}_t|$: the doubly exponential cross-sum is gone. What remains is a guarantee.

Proposition 5.30 (Point-based value iteration). Let $\hat{V}_t$ be computed by Algorithm 5.3, and let $\Delta = \max r - \min r$ be the range of the rewards.

(i) $\hat{V}_t \le V_t$ on $\mathcal{M}_1(\mathcal{S})$, and the output policy, run from time t at belief b, has an expected return at least $\hat{V}_t(b)$.

(ii) If $\mathcal{B}_t \supseteq \{\text{beliefs reachable at time } t \text{ from } b_1\}$ for every t, then $\hat{V}_1(b_1) = V_1(b_1)$.

(iii) With $\delta_t \coloneqq \max_{b \in \mathcal{M}_1(\mathcal{S})} \min_{b' \in \mathcal{B}_t} \|b - b'\|_1$, $\sup_b \left(V_t(b) - \hat{V}_t(b)\right) \le \Delta \sum_{k=t}^{H} (H - k + 1)\, \delta_k$; for constant $\delta_k = \delta$ the error at $t = 1$ is at most $\Delta\, \delta\, H(H+1)/2$.

Proof. (i) Each vector of $\hat{\Gamma}_t$ is of the form $r_a + \sum_y M_{a,y}\alpha_y$ with $\alpha_y \in \hat{\Gamma}_{t+1}$, hence, by induction, the value of a plan; so $\hat{\Gamma}_t$ is contained in the set Γ_t of all plan values and $\hat{V}_t \le V_t$. Let $G_t(b)$ be the expected return of the output policy from (t, b), and $\alpha = r_a + \sum_y M_{a,y}\alpha_y$ the vector it uses at b. Assuming $G_{t+1} \ge \hat{V}_{t+1}$,

$$G_t(b) = \langle r_a, b\rangle + \sum_y p(y \mid b, a) G_{t+1}(\Phi(b, a, y)) \ge \langle r_a, b\rangle + \sum_y \langle M_{a,y}\alpha_y, b\rangle = \hat{V}_t(b),$$

since $p(y \mid b, a)\hat{V}_{t+1}(\Phi(b, a, y)) \ge \langle \alpha_y, M_{a,y}^{\top} b\rangle$. (ii) For $b \in \mathcal{B}_t$ and any a, by the choice of the $\alpha_{a,y}$ and

the cancellation in the proof of Theorem 5.27,

$$\hat{V}_t(b) \geq \langle \alpha_{b,a}, b\rangle = \langle r_a, b\rangle + \sum_y p(y \mid b, a)\, \hat{V}_{t+1}\big(\Phi(b,a,y)\big). \tag{5.6}$$

If $\hat{V}_{t+1} = V_{t+1}$ on the successors of b, the right-hand side maximized over a is $V_t(b)$; with (i), $\hat{V}_t(b) = V_t(b)$, and backward induction along the reachable beliefs gives the claim. (iii) Let $\varepsilon_t = \sup_b (V_t - \hat{V}_t)(b)$, $\varepsilon_{H+1} = 0$. For $b' \in \mathcal{B}_t$, (5.6) with an action optimal at b' gives $V_t(b') - \hat{V}_t(b') \leq \varepsilon_{t+1}$. For any b, take $b' \in \mathcal{B}_t$ with $\|b - b'\|_1 \leq \delta_t$, $\alpha^\star \in \Gamma_t$ maximal at b and $\hat{\alpha} \in \hat{\Gamma}_t$ maximal at b':

$$V_t(b) - \hat{V}_t(b) \leq \langle \alpha^\star - \hat{\alpha}, b\rangle = \langle \alpha^\star, b - b'\rangle + \big(\langle \alpha^\star, b'\rangle - \hat{V}_t(b')\big) + \langle \hat{\alpha}, b' - b\rangle.$$

The middle term is at most $V_t(b') - \hat{V}_t(b') \leq \varepsilon_{t+1}$. Since $b - b'$ has zero sum, $\langle \alpha, b - b'\rangle \leq \frac{1}{2}(\max \alpha - \min \alpha)\|b - b'\|_1$, and the entries of a plan value with $H - t + 1$ stages lie in an interval of length $(H - t + 1)\Delta$. Hence $\varepsilon_t \leq \varepsilon_{t+1} + (H - t + 1)\Delta\delta_t$. ∎

By (ii), the algorithm interpolates between the exact computation of Theorem 5.26 (all reachable beliefs) and a cheap lower bound (a few points). In the example (Figure 5.5), $\mathcal{B} = \{0, 1\}$ gives a lower bound within 0.38 of V_1, $\{0, \frac{1}{2}, 1\}$ within 0.021, while the worst-case bound (iii) for the latter is 58.5; the full-information bound, which credits the owner with knowing the state at all later times, is off by up to 2.18. In discounted problems the same argument gives $\Delta\delta/(1-\gamma)^2$ (Pineau et al., 2003); modern solvers differ in how they choose the belief points, for instance by simulating the current policy (Spaan and Vlassis, 2005).

> **Remark 5.31 (Discounting and bandits).**
>
> (i) *Discounted POMDPs.* With a factor γ before the sum, (5.5) is a γ-contraction for the supremum norm (Chapter 8); its fixed point is convex, as a limit of convex functions, but in general not piecewise linear (Sondik, 1978).
>
> (ii) *Bayesian bandits* are POMDPs with a static hidden parameter ($P_a = \mathrm{Id}$) and observed rewards, the belief being the posterior; conjugacy reduces it to the counts of Exercise 1.6, and the Gittins index (Theorem 10.12) solves this belief MDP in the discounted case with independent arms.

The law of the return from a belief. For $\pi \in \Pi_y$, let $\nu^\pi(b)$ be the law of W_H under $\mathbb{P}_b^\pi$ and $\nu_s^\pi = \nu^\pi(\delta_s)$. The product formula (5.4) is linear in the initial law, and the policy, a function of the information only, does not depend on it; hence

$$\nu^\pi(b) = \sum_{s \in \mathcal{S}} b(s)\, \nu_s^\pi. \tag{5.7}$$

For a criterion ψ (Section 5.1), let $V_\psi(b) \coloneqq \sup_{\pi \in \Pi_y} \psi(\nu^\pi(b))$. Call ψ *mixture-convex* if $\psi(\lambda\nu + (1-\lambda)\nu') \leq \lambda\psi(\nu) + (1-\lambda)\psi(\nu')$ for all finitely supported ν, ν' and $\lambda \in [0,1]$.

> **Proposition 5.32 (Convex belief values).**
>
> (i) If ψ is mixture-convex, V_ψ is convex on $\mathcal{M}_1(\mathcal{S})$.
>
> (ii) For every ψ, $V_\psi(b) \leq V_\psi^{\mathrm{full}}(b) \coloneqq \sup_{\pi' \in \Pi_{\mathrm{H}}} \psi\big(\sum_s b(s)\nu^{\pi'}(s)\big)$, the value of the MDP whose initial state is drawn from b and then observed. If ψ is mixture-convex, $V_\psi^{\mathrm{full}}(b) \leq \sum_s b(s) \sup_{\pi' \in \Pi_{\mathrm{H}}} \psi(\nu^{\pi'}(s))$.

Proof. (i) For fixed π, $b \mapsto v^\pi(b)$ is affine by (5.7), so $b \mapsto \psi(v^\pi(b))$ is convex, and a supremum of convex functions is convex. (ii) The first inequality is Proposition 5.29(i); the second follows from $\psi(\sum_s b(s)v_s) \le \sum_s b(s)\psi(v_s)$, obtained from mixture convexity by induction on the number of components. ∎

Proposition 5.33 (Mixture-convex criteria). The following criteria are mixture-convex: (i) the expected utilities $\psi_f(\nu) = \int f \, \mathrm{d}\nu$, which are mixture-affine; (ii) the β-means U_β with $\beta < 0$; (iii) the conditional value at risk CVaR_δ of Definition 6.9, $\delta \in (0,1)$; (iv) the entropic value at risk EVaR_δ of Definition 6.13. For $\beta > 0$, U_β is mixture-concave, and V_{U_β} need not be convex.

Proof. (i) is clear. (ii) With $Z(\nu) = \int e^{\beta x} \nu(\mathrm{d}x) > 0$, affine in ν, concavity of the logarithm gives $\ln Z(\lambda\nu + (1-\lambda)\nu') \ge \lambda \ln Z(\nu) + (1-\lambda)\ln Z(\nu')$; multiplying by $\frac{1}{\beta} < 0$ reverses the inequality. For $\beta > 0$ the same computation gives mixture concavity. (iii) By Proposition 6.10, $\mathrm{CVaR}_\delta(\nu) = \max_\eta\{\eta - \frac{1}{\delta}\int (\eta - x)_+ \nu(\mathrm{d}x)\}$ is a maximum of functionals affine in ν, hence mixture-convex. (iv) $\mathrm{EVaR}_\delta(\nu) = \sup_{\beta>0}\{U_{-\beta}(\nu) - \ln(1/\delta)/\beta\}$ is a supremum of mixture-convex functionals by (ii). Finally, take $H = 1$, one action, two hidden states with rewards 0 and 1: then $\nu(b) = b(s_1)\delta_0 + b(s_2)\delta_1$ and $V_{U_\beta}(b) = \frac{1}{\beta}\ln\big(b(s_1) + b(s_2)e^\beta\big)$, strictly concave in b for $\beta > 0$. ∎

For the mean, Proposition 5.32(i) is the convexity of Theorem 5.27 seen without the recursion, V being a supremum of the linear functions $b \mapsto \sum_s b(s)\int x\, \nu_s^\pi(\mathrm{d}x)$: the independence axiom of Theorem 5.2 appears as linearity in the belief. For the risk-averse criteria of the book—$\beta < 0$, CVaR, EVaR—the belief value keeps this convex geometry; for $\beta > 0$ it does not. Convexity is not the statement that information has a nonnegative value, which holds for every criterion by Proposition 5.32(ii). In the example of the proof, the average $b(s_2)$ of the full-information values is below $V_{U_\beta}(b)$: for criteria that are not mixture-convex the last bound of (ii) fails, while V_ψ^{full} remains an upper bound. For β-means it is explicit, $V_{U_\beta}^{\mathrm{full}}(b) = \frac{1}{\beta}\ln\sum_s b(s)e^{\beta U^\star_{\beta,1}(s)}$, since $e^{\beta U_\beta(\sum_s b(s)\nu_s)} = \sum_s b(s)e^{\beta U_\beta(\nu_s)}$ and a β-greedy policy maximizes all the $U_\beta(\nu^{\pi'}(s))$ at once (Theorem 5.8).

Exponential utility and the tilted belief. For the β-means, the belief of Proposition 5.25 is not the right state when the rewards are hidden. Let $M^\beta_{a,y}(s,s') \coloneqq M_{a,y}(s,s')e^{\beta r(s,a,s')}$, and define the *tilted belief* along an information i_t by $\sigma_1 = b_1$ and $\sigma_{t+1} = (M^\beta_{a_t,y_{t+1}})^\top\sigma_t$: the computation of Proposition 5.25 with the factor $e^{\beta r}$ gives $\mathbb{E}^\pi_{b_1}[e^{\beta W_{1:t-1}}\mathbb{1}\{S_t = s, I_t = i_t\}] = \Pi^\pi(i_t)\,\sigma_t(s)$, an unnormalized belief reweighted by the exponential of the hidden reward accumulated on each path. Recall that M_β denotes the maximum if $\beta > 0$ and the minimum if $\beta < 0$.

Theorem 5.34 (Tilted belief). Let $\beta \neq 0$. Define on $\mathbb{R}^S_+$ the functions $Z_{H+1}(\sigma) = \langle \mathbf{1}, \sigma\rangle$ and $Z_t(\sigma) = \mathsf{M}_{\beta_a}\sum_y Z_{t+1}\big((M^\beta_{a,y})^\top\sigma\big)$.

(i) For every $\pi \in \Pi_{\mathcal{Y}}$, $U_\beta(v^\pi(b_1)) \le \frac{1}{\beta}\ln Z_1(b_1)$, with equality for the policy that plays at time t an action attaining M_β in $Z_t(\sigma_t)$.

(ii) $Z_t(\sigma) = \mathsf{M}_{\beta_{\alpha\in\Gamma_t^\beta}}\langle\alpha,\sigma\rangle$, where $\Gamma^\beta_{H+1} = \{\mathbf{1}\}$ and $\Gamma^\beta_t = \{\sum_y M^\beta_{a,y}\alpha_y : a \in \mathcal{A},\ \alpha_y \in \Gamma^\beta_{t+1}\}$. In particular Z_t is piecewise linear and positively homogeneous, so the optimal action depends on σ_t only through the normalized tilted belief $\sigma_t/\langle\mathbf{1},\sigma_t\rangle$.

(iii) If the rewards are observed, σ_t is proportional to the belief b_t, and the policy $A_t = g_t(b_t)$ with $g_t(b)$ attaining M_β in $Z_t(b)$ is β-optimal.

Proof. (i) Multiplying (5.4) by $e^{\beta W_H} = \prod_t e^{\beta r(s_t,a_t,s_{t+1})}$ and summing over the hidden states as in the

proof of Proposition 5.25, $\mathbb{E}^{\pi}_{b_1}[e^{\beta W_H}] = \sum_{i_{H+1}} \Pi^{\pi}(i_{H+1})\langle \mathbf{1}, \sigma_{H+1}(i_{H+1})\rangle$. Let $G^{\pi}_t(i_t)$ be the same sum restricted to the continuations of i_t, with the factors π_k, $k < t$, removed; then $G^{\pi}_{H+1}(i_{H+1}) = \langle \mathbf{1}, \sigma_{H+1}\rangle$, $G^{\pi}_t(i_t) = \sum_a \pi_t(a \mid i_t) \sum_y G^{\pi}_{t+1}(i_t, a, y)$, and $G^{\pi}_1 = \mathbb{E}^{\pi}_{b_1}[e^{\beta W_H}]$. By backward induction, $G^{\pi}_t(i_t) \le Z_t(\sigma_t(i_t))$ if $\beta > 0$ and $\ge$ if $\beta < 0$: replace G^{π}_{t+1} by $Z_{t+1}(\sigma_{t+1})$, with $\sigma_{t+1}(i_t, a, y) = (M^{\beta}_{a,y})^{\top}\sigma_t(i_t)$, and the average over a by M_{β_a}; both steps are equalities for the stated policy. Since $\frac{1}{\beta}\ln$ is increasing for $\beta > 0$ and decreasing for $\beta < 0$, the claim follows. (ii) By induction, $Z_{t+1}((M^{\beta}_{a,y})^{\top}\sigma) = \mathsf{M}_{\beta_\alpha}\langle M^{\beta}_{a,y}\alpha, \sigma\rangle$, and the sum over y of M_{β} over independent choices is the M_{β} of the sums. A function of the form $\mathsf{M}_{\beta_\alpha}\langle \alpha, \sigma\rangle$ is positively homogeneous, and so are the quantities compared by M_{β_a}. (iii) If $r(s, a, s') = \rho(a, y)$ whenever $M_{a,y}(s, s') > 0$, then $M^{\beta}_{a,y} = e^{\beta\rho(a,y)}M_{a,y}$, so $\sigma_t = e^{\beta W_{1:t-1}}\tilde{b}_t$ is proportional to b_t, and (ii) applies. ∎

With hidden rewards, b_t is not sufficient: two informations with the same belief can call for different actions, because the hidden reward already collected weighs the states differently (Exercise 5.15); for $\beta < 0$ the tilt overweights the states reached along paths where things went badly, and the agent hedges. As $\beta \to 0$ the ordinary belief is recovered. For $\beta < 0$, $\frac{1}{\beta}\ln Z_1(b) = \max_\alpha \frac{1}{\beta}\ln\langle \alpha, b\rangle$ is convex, as Proposition 5.33(ii) predicts. The information state σ_t goes back to James et al. (1994), who obtain it by a change of measure; its piecewise linearity is shown by Fernández-Gaucherand and Marcus (1997).

For other expected utilities, augmentation and filtering compose: apply Theorem 5.26 to the POMDP whose hidden state is the pair $(S_t, W_{1:t-1})$ of Definition 5.17, with the same observations and the terminal reward $f(W_H)$. The information state is the conditional law of $(S_t, W_{1:t-1})$; it reduces to $(b_t, W_{1:t-1})$ if the rewards are observed, and for $f(w) = e^{\beta w}$ to its image $\sigma_t(s) \propto \sum_w e^{\beta w}\,\mathbb{P}(S_t = s, W_{1:t-1} = w \mid I_t)$, the tilted belief. Quantiles and the conditional value at risk follow as in Corollaries 5.19 and 6.12. In general spaces, this embedding is due to Bäuerle and Rieder (2017), who recover the tilted information state without change of measure, and the ordinary filter when the rewards do not depend on the hidden state.

Distributional viewpoint

The greedy distributional planning scheme (Algorithm 5.1) is the most general form of dynamic programming: any criterion that dynamic programming can optimize must be compatible with the two operations of the distributional Bellman operator, mixtures and shifts (Lemma 5.4). Compatibility with mixtures, in its order-preserving form, is the independence axiom, and the expected-utility theorem (Theorem 5.2) turns it into an expected utility; compatibility with shifts, in the same form, is the translation property, which implies the translation compatibility of Chapter 4 and leaves only exponential utilities and the mean (Theorem 5.5). For these, the exponential Bellman equation optimizes as easily as the classical one, and deterministic Markov policies are optimal among all policies (Theorem 5.8). Every other criterion requires either more state—the accumulated reward, which restores dynamic programming for all expected utilities (Theorem 5.18)—or an approximation with no optimality guarantee.

Three properties of a statistic must not be confused: that its exact evaluation is a closed recursion (Chapter 4), that greedy optimization is valid for it (this chapter), and that a single transition gives an unbiased sample of its Bellman target, which is what temporal-difference learning needs (Chapter 12). Table 5.2 sets them side by side; a fourth, the contraction of the discounted iteration, is added in Chapter 8.

Table 5.2. Three properties of the statistics of the return met so far (finite horizon; V, $M^{(2)}$, Z_β, $\overline{x}$ are the next-state tables of the mean, the second moment, the exponential moment and the largest attainable return).

Statistic	Exact evaluation	Greedy optimization	One-transition sample target
Mean	closed, scalar (Corollary 3.11)	valid (Theorem 3.15)	unbiased: $r + V(s')$ (TD(0), Section 12.3)
Variance	closed jointly with the mean, not alone (Corollary 4.2 and Example 4.6)	mean–variance criteria: not monotone, not independent, time inconsistent (Example 5.16)	unbiased for the pair (mean, second moment): $r + V(s')$ and $r^2 + 2rV(s') + M^{(2)}(s')$ (moment TD, Theorem 12.25); the variance is $M^{(2)} - V^2$
Exponential moment, β-mean	Z_β closed alone (Theorem 4.15); with discounting the rate changes (Remark 4.13)	valid; deterministic Markov policies optimal (Theorem 5.8)	unbiased for Z_β: $e^{\beta r} Z_\beta(s')$; not for $U_\beta = \frac{1}{\beta} \ln Z_\beta$ (Section 12.6)
Quantile, CVaR	not closed: the whole law is needed (Example 4.6 and Exercise 4.3)	greedy may fail (Example 5.14 and Exercise 5.4); exact on the state augmented by the accumulated reward (Corollary 5.19; Bäuerle and Ott, 2011 for CVaR)	none on a scalar table; quantile TD is a pinball-loss step on a distributional table, not an unbiased Bellman target (Section 12.7)
Support extrema	closed (Proposition 4.20)	valid: $(\max, +)$ and $(\min, +)$ recursions (Exercise 3.1, Remark 5.9(i))	none unbiased: $\overline{x}(s)$ is the maximum of $r + \overline{x}(s')$ over the support, not an expectation; no finite sample certifies an unseen atom

Hands-on

The agent to be submitted for session H03 (*windy hike around the pit*) exposes a method `setObjective` that selects either a β-mean or a quantile of the return to optimize, and is evaluated on several grids and wind strengths. For β-means, Algorithm 5.2 is the answer, in $O(HS^2A)$ operations. For quantiles, Corollary 5.19(ii) reduces the problem to threshold probabilities on the augmented state space; since the accumulated reward is only positive at the bar, the augmentation is cheap here. The evaluation script also rewards computational efficiency: the distributional representation as pairs of vectors (atoms, probabilities) of the notebook is exact but slower than the value recursions of this chapter.

Exercises

Exercise 5.1 (Around the expected-utility theorem). [verify] (a) Show that in Theorem 5.2 the assumption "not constant on Dirac masses" cannot be dropped, and that a monotone continuous independent criterion that is constant on Dirac masses is constant. (b) Where does the proof use that λ ranges over all of $[0, 1]$ in the independence property, and not only over $\lambda = \frac{1}{2}$? (c) Show that if ψ is Bellman optimizable, then so is $G \circ \psi$ for every strictly increasing G. (d) Show that U_β and $U_{\beta'}$ are not ordinally equivalent when $\beta \neq \beta'$ (*hint:* compare a Dirac mass with a two-point law).

Exercise 5.2 (The third moment). [verify] In Example 5.15, compute $\mathbb{E}[W_H^3]$ for the two Markov policies and for the two history-dependent deterministic policies, and check the values $\frac{7}{6}, \frac{1}{2}, \frac{4}{3}$

and $\frac{1}{3}$. Show that the optimal policy is unique among deterministic history-dependent policies and that randomization does not help. Write the augmented MDP of Definition 5.17 for $f(x) = x^3$ and verify that its backward induction returns the optimal policy.

Exercise 5.3 (The median). [limits] (a) Verify the return distributions and medians of Example 5.14. (b) Show that the lower median is not $\mathcal{W}_1$-continuous: find $\nu_n \to \nu$ in $\mathcal{W}_1$ with $F^{-1}_{\nu_n}(\frac{1}{2}) \not\to F^{-1}_{\nu}(\frac{1}{2})$. (c) Show that no quantile $\nu \mapsto F^{-1}_{\nu}(\tau)$, $\tau \in (0,1)$, satisfies the independence property. (d) Is the median monotone? Does it satisfy the translation property?

Exercise 5.4 (The conditional value at risk is not Bellman optimizable). [limits] Let

$$\mathrm{CVaR}_{1/2}(\nu) = 2\int_0^{1/2} F^{-1}_{\nu}(u)\,\mathrm{d}u$$

be the mean of the lower half of ν. Let $\nu_1 = \frac{1}{2}(\delta_{-1}+\delta_1)$ and $\nu_2 = \delta_{-1}$. (a) Check that $\mathrm{CVaR}_{1/2}(\nu_1) = \mathrm{CVaR}_{1/2}(\nu_2) = -1$, and compute the two mixtures $\frac{1}{2}\nu_i + \frac{1}{2}\delta_5$ and their conditional values at risk. (b) Deduce that $\mathrm{CVaR}_{1/2}$ violates the indifference form of the independence property, and build an MDP with $H = 2$ in which Algorithm 5.1 may return a suboptimal policy. (c) Replace ν_2 by $\delta_{-1+\varepsilon}$ to obtain a violation with a strict inequality $\psi(\nu_2) > \psi(\nu_1)$. (d) Is $\mathrm{CVaR}_{1/2}$ monotone? Continuous? Deduce from Theorem 5.5 that it is not Bellman optimizable, by checking that it is not ordinally equivalent to any U_β.

Exercise 5.5 (A variance-penalized plan that refuses money). [limits] Consider $H = 2$: at time 1 the agent receives, without choice, $R_1 = 0$ or $R_1 = 10$ with probability $\frac{1}{2}$ each and moves to state D; at time 2 she chooses a (reward 1) or c (reward -4). For the mean–variance criterion $\psi_1(\nu) = \int x\,\mathrm{d}\nu - \mathbb{V}\mathrm{ar}[\nu]$ ($\theta = 1$ in Example 5.16), compute the criterion for the two Markov policies and for the policy "c if $R_1 = 10$, a if $R_1 = 0$". Show that the latter is optimal among all policies (including randomized ones), and that after $R_1 = 10$ it prescribes the action that is worse for every monotone criterion of the continuation. Comment on the meaning of "optimal" for such a criterion.

Exercise 5.6 (Details of exponential-utility planning). [verify] (a) Show that the mixture bound (5.2) is an equality if and only if all μ_i with $p_i > 0$ have the same β-mean, and deduce that a randomized Markov policy is β-optimal from every (t,s) if and only if it charges only β-greedy actions. (b) Prove the limits of Remark 5.9(i), and identify the $\beta \to -\infty$ recursion with the $(\max, \min, +)$ recursion for the largest return that can be guaranteed with probability one. (c) Show that the optimal β-value $U^\star_{\beta,1}(s)$ is nondecreasing in β and continuous, and that $\beta \mapsto \beta U^\star_{\beta,1}(s)$ is convex on $\beta > 0$, where it equals $\max_\pi \Lambda_{\nu^\pi(s)}(\beta)$. Show that on $\beta < 0$ it equals $\min_\pi \Lambda_{\nu^\pi(s)}(\beta)$ and need not be convex: exhibit a one-state, two-action example.

Exercise 5.7 (Quantile planning by bisection). [verify] Show that $u \mapsto \max_\pi \mathbb{P}^\pi_s(W_H \geq u)$ is nonincreasing, and deduce from Corollary 5.19(ii) an algorithm computing $\max_\pi \mathrm{VaR}_\delta(\nu^\pi(s))$ with $O(\log|\mathcal{W}_{H+1}|)$ augmented backward inductions. In the windy hike, what are the optimal quantiles of orders $\delta = 0.1, 0.25, 0.5$? (Use the table of Example 5.20.)

Exercise 5.8 (Augmentation versus distributional evaluation). [verify] For a *fixed* Markov policy π and $f = \mathbb{1}_{[u,\infty)}$, show that the backward induction of Theorem 5.18 on $\widetilde{\mathcal{M}}_f$, with the maximum over actions replaced by the average under $\pi_t(\cdot \mid s)$, computes $\mathbb{P}(w + W_{t:H} \geq u \mid S_t = s)$ for all (t, s, w), and that running it for all u amounts to computing the distribution functions of all $v_t^\pi(s)$, exactly as Algorithm 3.1 does. Compare the costs. Then explain why, for planning, the augmented recursion cannot be replaced by Algorithm 5.1 with $\psi(\nu) = \nu([u, \infty))$.

Exercise 5.9 (Reward scaling and the choice of β). [verify] Show that if all rewards of an MDP are multiplied by $\lambda > 0$, the β-greedy policies of the new MDP are the $\lambda\beta$-greedy policies of the original one. A practitioner normalizes the returns of a problem to $[0, 1]$ before applying Algorithm 5.2 with $\beta = -1$; describe the corresponding risk attitude in the original units, and compare with the same choice of β made without normalization when returns range in $[0, 1000]$.

Exercise 5.10 (Numerical study). [implement] Implement Algorithm 5.2 and the augmented recursion of Theorem 5.18 for the windy hike and reproduce the tables of Examples 5.10 and 5.20 (both recursions are exact, so the reference is exact up to rounding). Describe the β-optimal policy at $\beta = -1$: what does it do after being blown one row down? Which policy maximizes $\mathbb{P}(W \geq 3)$, and which row does it use?

Exercise 5.11 (Control limits for expected utilities). [verify] In the machine replacement problem, the owner maximizes $\mathbb{E}[f(W_H)]$, W_H being minus the total cost and f a bounded nondecreasing function, by the augmentation of Theorem 5.18 (take measurability on the state space $\mathbb{R}_+ \times \mathbb{R}$ for granted). Show by backward induction that the augmented values $\widetilde{V}_t(x, w)$ are nonincreasing in the wear x and nondecreasing in the accumulated reward w, and deduce that for every t and w the set of wear levels at which replacing is optimal is an upper set: some optimal policy is a control limit $x_t^\star(w)$ that depends on the cost already paid. For $f = \mathbb{1}_{[u,\infty)}$ and a single year left, compute $x_H^\star(w)$ and check that it depends on w.

Exercise 5.12 (Pulling the goalie). [implement] (a) Implement the backward induction of Example 5.21 on the pairs (intervals to go, deficit) and reproduce the pull times and the table (exact recursion; a fraction of a second). (b) Show that under every policy $\mathbb{E}[W_H] = -(p_B - p_A)\,\mathbb{E}[N]$, where N is the number of intervals played with the goalie pulled and p_A, p_B the per-interval scoring probabilities when pulling, and deduce the mean-optimal policy. (c) Show that with $k < d$ intervals to go and a deficit d both actions have value 0. A program that records the pull region as the set where pulling is *strictly* better, and tests whether it is $\{1, \dots, k_d\}$, reports a region that is not an interval for $d \geq 2$: explain. (d) Solve the variant in which a tie is absorbing (value 1), check that the team pulls earlier with a one-goal deficit, and explain why by comparing the values of a tie in the two models. (e) How does the one-goal pull time depend on the empty-net rate 19.16? Try 12, 15, 25, 30 goals per 60 minutes.

Exercise 5.13 (The tiger problem). [verify] A tiger is behind the left or the right door, with probability $\frac{1}{2}$ each; the belief is $q = b(\text{left})$. The agent may *listen* (reward -1; she hears the tiger on the correct side with probability 0.85, and the tiger does not move) or open a door (reward $+10$ if the tiger is behind the other door, -100 otherwise; then the problem is reset: the tiger is placed

again uniformly at random, and the next observation is uninformative) (Cassandra et al., 1994). (a) Write $M_{a,y}$ and show that after d more "left" than "right" answers since the last reset, $q = 0.85^d/(0.85^d+0.15^d)$. (b) With one decision left, show that $V_H(q) = \max(-1,\ 110q-110,\ 110q-100)$ and that listening is optimal exactly for $q \in [0.1, 0.9]$. (c) With two decisions left, compute Γ_{H-1} by Theorem 5.27, remove the vectors that are nowhere maximal (five remain), and show that $V_{H-1}(\frac{1}{2}) = -2$. (d) With three decisions left, show that the optimal behavior from $q = \frac{1}{2}$ is to listen twice and to open a door at the last step if and only if the two answers agree, and that $V_{H-2}(\frac{1}{2}) = 2.72$. Why does a single answer never justify opening? (e) Check the number of vectors that are nowhere dominated for four, five and six decisions left (5, 9, 13), and compare with the bound $|\Gamma_t| \le A|\Gamma_{t+1}|^{|\mathcal{Y}|}$.

Exercise 5.14 (Information has value: Blackwell's garbling). [verify] Let κ and κ' be two observation kernels for the same MDP, with $\kappa'(y' \mid s,a,s') = \sum_y \kappa(y \mid s,a,s')K(y' \mid y)$ for a stochastic matrix K on $\mathcal{Y}$: the observation y' is a *garbling* of y, obtained by passing it through a noisy channel. Let V_t and V'_t be the corresponding belief values of Theorem 5.26. (a) Show that $M'_{a,y'} = \sum_y K(y' \mid y)M_{a,y}$, and that $\Phi'(b,a,y')$ is a convex combination of the beliefs $\Phi(b,a,y)$, with weights proportional to $K(y' \mid y)p(y \mid b,a)$. (b) Using the convexity of V'_{t+1} (Theorem 5.27), prove by backward induction that $V'_t \le V_t$ on $\mathcal{M}_1(\mathcal{S})$. (c) Deduce that a better sensor never hurts the mean, and that the inequality holds with κ' uninformative (K with identical rows): observing is worth at least as much as not observing. (d) Show that $V'_\psi \le V_\psi$ for every criterion ψ, by letting the agent with the better sensor garble its own observations with an independent randomization. For the β-means, recover this from Theorem 5.34(ii): Z_t is positively homogeneous, and convex for $\beta > 0$, concave for $\beta < 0$, hence subadditive or superadditive. Note that convexity of the belief value is used in (b) but not needed in (d). Blackwell's theorem (Blackwell, 1953) states a converse: if one experiment is worth at least as much as another in every decision problem, it is a garbling of the other.

Exercise 5.15 (Hidden rewards: the belief is not enough). [limits] A static parameter x or y is drawn with probability $\frac{1}{2}$ each. At time 1 a fair coin $c \in \{u,d\}$, independent of the parameter, is tossed and observed. At time 2 a hidden reward $L > 0$ is collected if $(\text{parameter}, c) \in \{(x,u),(y,d)\}$, and 0 otherwise. At time 3 the agent bets on the parameter and receives 1 if right. (a) Write this as a POMDP with $H = 3$ (hidden states x, y at times 1 and 3, pairs $(\text{parameter}, c)$ at time 2; the only choice is at time 3), and check that the belief at time 3 is $(\frac{1}{2}, \frac{1}{2})$ whatever c. (b) Compute the tilted belief σ_3 after $c = u$ and after $c = d$. Show that for $\beta < 0$ the β-optimal bet after $c = u$ is y, the parameter under which no reward was collected, and x after $c = d$; and that for $\beta > 0$ the bets are reversed. (*Hint:* the difference of the two exponential utilities is $\frac{1}{2}(1-e^{\beta L})(1-e^{\beta})$ up to the sign.) (c) Deduce that no policy depending only on (t, b_t) is β-optimal. For $L = 1$ and $\beta = -1$, compare the optimal β-mean (1: the hedging bet makes the return sure) with the best one among policies depending on (t, b_t) (≈ 0.76). (d) What changes if the reward of time 2 is observed? If $\beta \to 0$?

Bibliographical notes

Expected-utility theory begins with von Neumann and Morgenstern (1944); the axiomatization through the indifference form of independence (for mixtures with weight $\frac{1}{2}$) and a continuity axiom is due to Herstein and Milnor (1953), and Fishburn (1970) and Kreps (1988) are standard expositions. Grandmont (1972) studies the continuity of the utility function on spaces of probability measures with the weak

topology; our Theorem 5.2 uses the Wasserstein topology and monotonicity, which shorten the argument considerably. The connection between translative quasi-arithmetic means and exponential functions, which underlies Theorem 5.5, is classical (Hardy et al., 1952, Chap. III).

The notion of a Bellman-optimizable criterion, the two necessary conditions and the characterization theorem are due to Marthe et al. (2023), who prove Theorem 5.5 by a differential argument; the route through the closure theorem of Chapter 4 is new in this form. Exponential-utility planning goes back to Howard and Matheson (1972); Jaquette (1973) proposed moment-based criteria and Denardo and Rothblum (1979) connected exponential utilities to optimal stopping and linear programming. The median and third-moment counterexamples were examination problems of the course.

Mean–variance criteria were introduced by Markowitz (1952) in portfolio theory; White (1988) reviews their use in MDPs, Filar et al. (1989) study variance-penalized MDPs and Mannor and Tsitsiklis (2013) establish the hardness results. State augmentation for probabilistic criteria has a long history reviewed by White (1988); its systematic use for the conditional value at risk is due to Bäuerle and Ott (2011) and for general utilities to Bäuerle and Rieder (2014). The time-inconsistency of non-exponential criteria and the nested (recursive) risk measures designed to restore consistency are discussed in Chapter 6; in discounted MDPs, nested risk measures are developed by Ruszczyński (2010).

Partially observable MDPs go back to Åström (1965). The piecewise-linear convex structure of the finite-horizon value is due to Smallwood and Sondik (1973), the discounted case to Sondik (1978); Kaelbling et al. (1998) brought the model to artificial intelligence, with the tiger problem of Cassandra et al. (1994), and Krishnamurthy (2016) is a modern textbook, which includes the structural results initiated by Lovejoy (1987). Point-based value iteration is due to Pineau et al. (2003), with randomized variants such as that of Spaan and Vlassis (2005); the complexity results are those of Papadimitriou and Tsitsiklis (1987) and Madani et al. (2003). The risk-sensitive information state is due to James et al. (1994), structural results for it to Fernández-Gaucherand and Marcus (1997), and the embedding for general utilities to Bäuerle and Rieder (2017). The comparison of experiments by garbling is due to Blackwell (1953). The mixture viewpoint of Propositions 5.32 and 5.33, and the counterexample of Exercise 5.15, are presented here in this form.