8

Optimal Planning in Discounted MDPs



The planning problem of a discounted MDP has a fixed-point formulation: the optimal value function is the unique fixed point of the Bellman optimality operator, a contraction, and any policy greedy with respect to it is optimal—stationary, deterministic, and optimal among all history-dependent policies. Three classical algorithms compute it: value iteration, policy iteration and linear programming; they are compared on the bike store, where policy iteration converges in three steps, and value iteration in a thousand with a norm stopping rule but in ten with a span stopping rule, valid for every monotone Bellman operator; the number of iterations of policy iteration is bounded by a polynomial in the numbers of states and actions. The dual variable of the linear program, the occupation measure, gets a section of its own: it ranges over a polytope on which every expected additive criterion is linear and the law of the return is not a function, and its version started from a fixed state–action pair, the successor measure, evaluates a policy for every reward at once. Adding linear constraints to the linear program gives constrained MDPs, whose optimal policies must randomize. A general bound relates the quality of a value function to that of its greedy policy, and prepares the approximate methods of Chapter 13. Propagating the whole return distribution while selecting actions for the mean is possible too, but the operator that does it is not a contraction, and its iterates converge to the return law of one *optimal policy among possibly several. Risk-sensitive planning comes next: for β-means, the rate change caused by discounting makes the optimal policies non-stationary, but a backward recursion over decreasing rates, followed by a risk-neutral tail, yields policies with an explicit guarantee, and the grid of Chapter 6 extends to the entropic value at risk. The last section lets the episode end at a random time instead of discounting: the law of the return exists without discount, its moments and exponential moments are Bellman closed at a fixed rate, and risk-sensitive optimal policies are stationary again, up to a critical risk level given by a spectral radius.*

8.1 Bellman optimality

Throughout, $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ is a discounted MDP as in Chapter 7. The *optimal value function* is $V^\star(s) \coloneqq \sup_{\pi \in \Pi_{\mathrm{H}}} V^\pi(s)$, where for a history-dependent policy $V^\pi(h_t)$ denotes, as at the end of Section 7.4, the expectation of the discounted reward-to-go $W_{t:\infty} = \sum_{k \geq t} \gamma^{k-t} R_k$ under the continuation of h_t, and $V^\pi(s) = V^\pi(h_1)$ for $h_1 = s$. A policy is *optimal* if $V^\pi = V^\star$.

Definition 8.1 (Bellman optimality operator, greedy policies). The *Bellman optimality operator* $T^* : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ is

$$(T^*V)(s) \coloneqq \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P(s' \mid s, a)\big[r(s, a, s') + \gamma V(s')\big].$$

A stationary deterministic policy π_V is *greedy* with respect to $V \in \mathbb{R}^{\mathcal{S}}$ if $\pi_V(s)$ attains the maximum for every s; then $T^{\pi_V} V = T^* V$, and $T^* V = \max_\pi T^\pi V$ over $\pi \in \Pi_{\mathrm{SD}}$ (componentwise).

Proposition 8.2 (Properties of T^*). T^* is monotone ($V \le V'$ implies $T^*V \le T^*V'$), satisfies $T^*(V+c) = T^*V + \gamma c$ for constants c, and is a γ-contraction for the supremum norm. It has a unique fixed point $\widehat{V}$, and for every V, $\|(T^*)^k V - \widehat{V}\|_\infty \le \gamma^k \|V - \widehat{V}\|_\infty$ and $\|V - \widehat{V}\|_\infty \le \|T^*V - V\|_\infty/(1-\gamma)$.

Proof. Monotonicity and the translation property hold for each T^π and pass to the maximum over π. For the contraction, $|\max_a x_a - \max_a y_a| \le \max_a |x_a - y_a|$ gives $|(T^*V)(s) - (T^*V')(s)| \le \max_a \gamma \sum_{s'} P(s' \mid s,a)|V(s') - V'(s')| \le \gamma \|V - V'\|_\infty$. The rest is Banach's theorem (Theorem D.16) and the argument of (7.3). ∎

Theorem 8.3 (Bellman optimality in discounted MDPs).

(i) The optimal value function $V^\star$ is the unique fixed point of T^*: $V^\star = T^*V^\star$.

(ii) For every $\pi \in \Pi_{\mathrm{H}}$ and every history h_t, $V^\pi(h_t) \le V^\star(s_t)$.

(iii) Every policy $\pi^\star \in \Pi_{\mathrm{SD}}$ greedy with respect to $V^\star$ is optimal. Hence $\sup_{\pi \in \Pi_{\mathrm{H}}} V^\pi = \max_{\pi \in \Pi_{\mathrm{SD}}} V^\pi = V^\star$: an optimal policy exists, and can be chosen stationary and deterministic.

Proof. Let $\widehat{V}$ be the fixed point of T^* (Proposition 8.2). (ii) with $\widehat{V}$ in place of $V^\star$: for $\pi \in \Pi_{\mathrm{H}}$ and $n \ge 0$, let $V^{\pi,n}(h_t)$ be the expected discounted reward over the next n steps, that is $\sum_{k=t}^{t+n-1} \gamma^{k-t} R_k$, under the continuation of h_t. By induction on n, $V^{\pi,n}(h_t) \le ((T^*)^n 0)(s_t)$: the case $n = 0$ is trivial, and

$$\begin{aligned} V^{\pi,n}(h_t) &= \sum_a \pi_t(a \mid h_t) \sum_{s'} P(s' \mid s_t, a)\big[r(s_t,a,s') + \gamma V^{\pi,n-1}(h_t,a,s')\big] \\ &\le \sum_a \pi_t(a \mid h_t) \sum_{s'} P(s' \mid s_t, a)\big[r + \gamma((T^*)^{n-1}0)(s')\big] \;\le\; ((T^*)^n 0)(s_t). \end{aligned}$$

As $n \to \infty$, $V^{\pi,n}(h_t) \to V^\pi(h_t)$ (the tail is at most $\gamma^n R_{\max}/(1-\gamma)$) and $(T^*)^n 0 \to \widehat{V}$; hence $V^\pi(h_t) \le \widehat{V}(s_t)$, and $V^\star \le \widehat{V}$. (iii) with $\widehat{V}$: if $\pi^\star$ is greedy with respect to $\widehat{V}$, then $T^{\pi^\star}\widehat{V} = T^*\widehat{V} = \widehat{V}$, so $\widehat{V}$ is the fixed point of $T^{\pi^\star}$, which is $V^{\pi^\star}$ by Theorem 7.11. Thus $V^{\pi^\star} = \widehat{V} \ge V^\star \ge V^{\pi^\star}$: all are equal, which proves (i), (ii) and (iii). ∎

The theorem, due to Blackwell (1962) in this generality, has the same content as its finite-horizon counterpart (Theorem 3.15): neither memory nor randomization helps, and—new here—neither does time dependence. Its proof again uses only the linearity and monotonicity of the expectation. Section 8.4 looks at what happens when the whole return distribution is propagated while the selection is still made for the mean, and Section 8.5 explores what remains for other criteria. The Q-function version reads $Q^\star(s,a) = \sum_{s'} P(s' \mid s,a)[r(s,a,s') + \gamma \max_{a'} Q^\star(s',a')]$, and greedy policies are $\pi^\star(s) \in \arg\max_a Q^\star(s,a)$.

Optimal policies are found by improving policies step by step, using the following observation.

Proposition 8.4 (Policy improvement). Let $\pi \in \Pi_{\mathrm{SR}}$ and let $\pi' \in \Pi_{\mathrm{SD}}$ be greedy with respect to V^π, so that $T^{\pi'}V^\pi = T^*V^\pi \ge T^\pi V^\pi = V^\pi$. Then $V^{\pi'} \ge V^\pi$, and $V^{\pi'} = V^\pi$ only if $V^\pi = V^\star$.

Proof. By monotonicity of $T^{\pi'}$, $V^\pi \le T^{\pi'}V^\pi \le (T^{\pi'})^2 V^\pi \le \cdots \le \lim_k (T^{\pi'})^k V^\pi = V^{\pi'}$. If $V^{\pi'} = V^\pi$, all these inequalities are equalities, in particular $T^*V^\pi = V^\pi$, so V^π is the fixed point $V^\star$. ∎

8.2 Value iteration, policy iteration, linear programming

Algorithm 8.1: Value iteration

Input: discounted MDP $\mathcal{M}$, accuracy $\varepsilon > 0$

$V \leftarrow 0$; $V' \leftarrow T^*V$

while $\|V' - V\|_\infty > \varepsilon(1-\gamma)/(2\gamma)$ **do**

  $V \leftarrow V'$; $V'(s) \leftarrow \max_a \sum_{s'} P(s' \mid s,a)\big[r(s,a,s') + \gamma V(s')\big]$ for all s

Output: a greedy policy $\pi_{V'}$, which is ε-optimal: $\|V^{\pi_{V'}} - V^\star\|_\infty \le \varepsilon$

Value iteration

The simplest algorithm iterates T^* from any starting point and returns a greedy policy for the last iterate (Algorithm 8.1). Its stopping rule rests on the a posteriori bound (7.3), applied twice (Proposition 8.6). A second bound, not needed here but central to the approximate methods of Chapter 13, compares a value function with its greedy policy: a value function that is close to $V^\star$ has a greedy policy that is close to optimal.

Theorem 8.5 (Performance loss of greedy policies). For every $V \in \mathbb{R}^{\mathcal{S}}$ and every policy π_V greedy with respect to V,

$$\|V^\star - V^{\pi_V}\|_\infty \;\le\; \frac{2\gamma}{1-\gamma}\,\|V^\star - V\|_\infty .$$

Proof. Since $T^*V^\star = V^\star$, $T^{\pi_V}V^{\pi_V} = V^{\pi_V}$ and $T^{\pi_V}V = T^*V$,

$$V^\star - V^{\pi_V} = (T^*V^\star - T^*V) + (T^{\pi_V}V - T^{\pi_V}V^{\pi_V}),$$

whose supremum norm is at most $\gamma\|V^\star - V\| + \gamma\|V - V^{\pi_V}\| \le \gamma\|V^\star - V\| + \gamma(\|V - V^\star\| + \|V^\star - V^{\pi_V}\|)$. Rearranging gives the claim. ∎

Proposition 8.6 (Guarantee of value iteration). Algorithm 8.1 terminates after at most $1 + \ln\big(2\gamma R_{\max}/(\varepsilon(1-\gamma))\big)/\ln(1/\gamma)$ iterations, and its output satisfies $\|V^{\pi_{V'}} - V^\star\|_\infty \le \varepsilon$. Each iteration costs $O(S^2A)$ operations.

Proof. The threshold is half that of Algorithm 7.1 because two objects must be certified, the value and its greedy policy. When the loop stops, $\|T^*V' - V'\|_\infty \le \gamma\|V' - V\|_\infty \le \varepsilon(1-\gamma)/2$. By Proposition 8.2, $\|V' - V^\star\|_\infty \le \varepsilon/2$, and by (7.3) applied to $\pi_{V'}$, for which $T^{\pi_{V'}}V' = T^*V'$, $\|V' - V^{\pi_{V'}}\|_\infty \le \varepsilon/2$; the triangle inequality concludes. Since $\|V_{k+1} - V_k\|_\infty \le \gamma^k\|V_1 - V_0\|_\infty \le \gamma^k R_{\max}$, the stopping condition holds as soon as $\gamma^k R_{\max} \le \varepsilon(1-\gamma)/(2\gamma)$. ∎

Span certificates

The norm $\|V_{k+1} - V_k\|_\infty$ mixes two pieces of information: how far the iterate is from the *level* of $V^\star$, and how far from its *shape*. A greedy policy only sees the shape, since adding a constant to V does not change $\arg\max_a$. The shape is measured by the *span seminorm* $\operatorname{span}(V) \coloneqq \max_s V(s) - \min_s V(s)$, which ignores constants, $\operatorname{span}(V + c) = \operatorname{span}(V)$, and satisfies $\operatorname{span}(V) \le 2\|V\|_\infty$. The bounds below use only the two properties of Proposition 8.2 that do not involve a norm, monotonicity and translation, and therefore hold for every operator that has them.

Proposition 8.7 (Span bounds). Let $T : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ be monotone and satisfy $T(V + c) = TV + \gamma c$ for all V and all constants c, with $\gamma < 1$. Then T is a γ-contraction for $\|\cdot\|_\infty$; let V_T be its fixed point. For $V \in \mathbb{R}^{\mathcal{S}}$, let $m \coloneqq \min_s (TV - V)(s)$ and $M \coloneqq \max_s (TV - V)(s)$.

(i) $V + \dfrac{m}{1-\gamma} \;\le\; TV + \dfrac{\gamma m}{1-\gamma} \;\le\; V_T \;\le\; TV + \dfrac{\gamma M}{1-\gamma} \;\le\; V + \dfrac{M}{1-\gamma}$. In particular the *extrapolated value* $\widetilde{V} \coloneqq TV + \gamma(m + M)/(2(1-\gamma))$ satisfies $\|V_T - \widetilde{V}\|_\infty \le \gamma \operatorname{span}(TV - V)/(2(1-\gamma))$.

(ii) Let T' be a second operator with the same two properties, with $T'U \le TU$ for all U and $T'V = TV$ at the given V, and let $V_{T'}$ be its fixed point. Then $TV + \gamma m/(1-\gamma) \le V_{T'} \le V_T$, hence

$$\|V_T - V_{T'}\|_\infty \;\le\; \frac{\gamma}{1-\gamma} \operatorname{span}(TV - V).$$

Proof. From $V \le V' + \|V - V'\|_\infty$, monotonicity and translation give $TV \le TV' + \gamma\|V - V'\|_\infty$, and symmetrically: T is a γ-contraction, with a fixed point by Theorem D.16. (i) Let $W \coloneqq TV + \gamma m/(1-\gamma)$. Since $TV \ge V + m$, monotonicity and translation give $T(TV) \ge T(V + m) = TV + \gamma m$, hence $TW = T(TV) + \gamma^2 m/(1-\gamma) \ge TV + \gamma m + \gamma^2 m/(1-\gamma) = W$. Iterating, $W \le TW \le T^2 W \le \dots$, and $T^k W \to V_T$, so $W \le V_T$. The upper bound is the same argument with M and the inequalities reversed, and the two outer inequalities are $V + m \le TV$ and $TV \le V + M$, divided by $1 - \gamma$ after adding $\gamma m/(1-\gamma)$ or $\gamma M/(1-\gamma)$. The extrapolated value is the midpoint of the interval. (ii) Since $T'V = TV$, the numbers m, M are the same for T', and (i) applied to T' gives the lower bound on $V_{T'}$. And $V_{T'} = T'V_{T'} \le TV_{T'}$, so the iteration argument of (i) gives $V_{T'} \le V_T$. Subtracting the lower bound on $V_{T'}$ from the upper bound on V_T gives the last display. ∎

With $T = T^*$ and $T' = T^{\pi_V}$ for a policy π_V greedy with respect to V, part (ii) bounds the loss of the greedy policy by $\gamma \operatorname{span}(T^*V - V)/(1-\gamma)$. The factor 2 of Proposition 8.6 disappears, and the bound does not change when a constant is added to V. The bounds go back to MacQueen (1966) and Porteus (1971).

Corollary 8.8 (Span stopping rule). Run value iteration $V_{k+1} = T^*V_k$ from any V_0 and stop at the first k with $\operatorname{span}(V_{k+1} - V_k) \le \varepsilon(1-\gamma)/\gamma$. Then the policy π_{V_k} greedy with respect to V_k, which is computed together with V_{k+1}, is ε-optimal, and $V_{k+1} + \gamma \min_s (V_{k+1} - V_k)(s)/(1-\gamma)$ is a lower bound on $V^\star$ within ε of it. The rule stops no later than the rule of Algorithm 8.1.

Proof. Apply Proposition 8.7 with $T = T^*$, $T' = T^{\pi_{V_k}}$ and $V = V_k$; the hypotheses hold by Proposition 8.2 and $T^* = \max_\pi T^\pi$. The comparison follows from span $\le 2\|\cdot\|_\infty$: the test of Algorithm 8.1 implies this one. ∎

The span of the increments decreases faster than their norm whenever the transition laws overlap. Let $\tau \coloneqq 1 - \min \sum_{s''} \min\{P(s'' \mid s, a), P(s'' \mid s', a')\}$, the minimum being over all pairs $(s, a), (s', a')$. If a is greedy for V at s and a' greedy for V' at s', then $(T^*V - T^*V')(s) \le \gamma \sum_{s''} P(s'' \mid s, a)(V - V')(s'')$ and $(T^*V - T^*V')(s') \ge \gamma \sum_{s''} P(s'' \mid s', a')(V - V')(s'')$. The difference of two probability vectors whose common mass is at least $1 - \tau$ moves a function by at most τ times its span, hence $\operatorname{span}(T^*V - T^*V') \le \gamma\tau \operatorname{span}(V - V')$. So the span of the increments $V_{k+1} - V_k$ decreases at least like $(\gamma\tau)^k$, while their norm decreases only like γ^k.

Proposition 8.7 does not depend on the form of T, and it applies verbatim to the other Bellman operators of the book: the evaluation operator T^π; the robust operator of Theorem 17.12, an infimum of monotone maps with the translation property, whose stopping rule in Algorithm 17.3 can be made span-based; Shapley's operator (18.1), since $\operatorname{val}(M + c) = \operatorname{val}(M) + c$, with player 1's strategies in

Algorithm 8.2: Policy iteration

Input: discounted MDP $\mathcal{M}$, initial policy $\pi_0 \in \Pi_{\mathrm{SD}}$

$k \leftarrow 0$

repeat

  evaluation: $V^{\pi_k} \leftarrow (I - \gamma P^{\pi_k})^{-1} r^{\pi_k}$

  improvement: $\pi_{k+1}(s) \leftarrow$ an element of $\arg\max_a \sum_{s'} P(s' \mid s, a)[r(s, a, s') + \gamma V^{\pi_k}(s')]$, keeping $\pi_k(s)$ whenever it belongs to the arg max

  $k \leftarrow k + 1$

until $\pi_k = \pi_{k-1}$

Output: π_k, optimal

the role of the policies of (ii); the soft operator of Exercise 15.3, whose log-sum-exp is translation-equivariant and is the maximum over π of the regularized evaluation operators; and the nested operators $V \mapsto \max_a \rho\big(r(s, a, S') + \gamma V(S')\big)$ built from a monotone, translation-equivariant ρ such as CVaR_δ (Section 6.10, (6.2)). It does not apply to the recursion of Theorem 8.22, which applies a different operator at each rate β_t and has no fixed point to certify.

Policy iteration

Howard's policy iteration (Algorithm 8.2) alternates exact evaluation of the current policy and greedy improvement.

Theorem 8.9 (Convergence of policy iteration). The values V^{π_k} produced by Algorithm 8.2 are nondecreasing, strictly increasing in at least one state as long as π_k is not optimal, and the algorithm stops after finitely many iterations (at most A^S) with an optimal policy. Moreover $V^{\pi_k} \geq (T^*)^k V^{\pi_0}$ for all k: policy iteration is never slower than value iteration started from the same value.

Proof. Monotonicity and strict improvement are Proposition 8.4; since Π_{SD} is finite and a policy cannot be revisited once strictly improved upon, the algorithm terminates, and by the tie-breaking rule it stops exactly when $\pi_{k+1} = \pi_k$, i.e. when $T^* V^{\pi_k} = V^{\pi_k}$, which is optimality. For the comparison, induct on k: $V^{\pi_{k+1}} \geq T^{\pi_{k+1}} V^{\pi_k} = T^* V^{\pi_k} \geq T^* (T^*)^k V^{\pi_0}$, using the first inequality of the proof of Proposition 8.4 and the monotonicity of T^*. ∎

The bound A^S is very pessimistic. Policy iteration is Newton's method applied to the piecewise-affine equation $V - T^* V = 0$ (Puterman and Brumelle, 1979), and typically converges in a handful of iterations. Each iteration costs a linear solve, $O(S^3)$, plus an improvement step, $O(S^2 A)$; when S is large, the evaluation is itself performed approximately by a few steps of Algorithm 7.1 (*modified* policy iteration).

How many iterations does policy iteration need? ♦

For a fixed discount factor, the number of iterations is bounded by a polynomial in S and A alone, whatever the rewards and the transition probabilities (Ye, 2011; Hansen et al., 2013). The simplest proof uses only the tools above. Write $g_\pi(s) \coloneqq V^\star(s) - Q^\star(s, \pi(s)) \geq 0$ for the *gap* of a deterministic policy π at s: the pair (s, a) is *suboptimal* when $V^\star(s) - Q^\star(s, a) > 0$, and each state has at least one pair that is not, so there are at most $SA - S$ suboptimal pairs.

Theorem 8.10 (Strongly polynomial bound, Scherrer, 2016). Let K be the smallest integer with $\gamma^K < 1-\gamma$, that is $K = \lfloor \ln(1/(1-\gamma))/\ln(1/\gamma) \rfloor + 1$. If π_j is not optimal, there is a state s_j such that $\pi_k(s_j) \neq \pi_j(s_j)$ for every $k \geq j+K$: at least one suboptimal pair is eliminated for good every K iterations. Consequently the policy π_k of Algorithm 8.2 is optimal for every $k \geq (SA-S)K$, and the algorithm stops at the next iteration.

Proof. Gaps and values. Since $T^\pi V^\star(s) = Q^\star(s, \pi(s))$, we have $V^\star - T^\pi V^\star = g_\pi$, and subtracting $V^\pi = T^\pi V^\pi$ gives $V^\star - V^\pi = g_\pi + \gamma P^\pi (V^\star - V^\pi)$, that is $V^\star - V^\pi = (I - \gamma P^\pi)^{-1} g_\pi = \sum_{t \geq 0} \gamma^t (P^\pi)^t g_\pi$. All terms are nonnegative, so $g_\pi \leq V^\star - V^\pi$, and each is at most $\gamma^t \|g_\pi\|_\infty$, so $\|V^\star - V^\pi\|_\infty \leq \|g_\pi\|_\infty / (1-\gamma)$.

Contraction. The induction in the proof of Theorem 8.9, started at π_j instead of π_0, gives $V^\star \geq V^{\pi_{j+k}} \geq (T^*)^k V^{\pi_j}$, hence $\|V^\star - V^{\pi_{j+k}}\|_\infty \leq \gamma^k \|V^\star - V^{\pi_j}\|_\infty$.

Elimination. Let π_j be non-optimal. Then $g_{\pi_j} \neq 0$, since $g_{\pi_j} = 0$ would give $V^{\pi_j} = V^\star$ by the first step; choose s_j with $g_{\pi_j}(s_j) = \|g_{\pi_j}\|_\infty > 0$. For every $k \geq K$, the two previous steps give

$$g_{\pi_{j+k}}(s_j) \leq \|V^\star - V^{\pi_{j+k}}\|_\infty \leq \gamma^k \|V^\star - V^{\pi_j}\|_\infty \leq \frac{\gamma^k}{1-\gamma} g_{\pi_j}(s_j) < g_{\pi_j}(s_j),$$

so $\pi_{j+k}(s_j) \neq \pi_j(s_j)$. The pair $(s_j, \pi_j(s_j))$ is suboptimal and absent from all policies after $j+K$.

Counting. If $\pi_0, \pi_K, \ldots, \pi_{(n-1)K}$ are all non-optimal, the pairs eliminated at these n times are distinct: the pair of time iK belongs to π_{iK}, while those of the earlier times $i'K$ are absent from every policy after $(i'+1)K \leq iK$. Hence $n \leq SA - S$, and $\pi_{(SA-S)K}$ is optimal. Once a policy is optimal, the tie-breaking rule keeps it, and the loop stops. ∎

Since $\ln(1/\gamma) \geq 1-\gamma$, $K \leq 1 + \ln(1/(1-\gamma))/(1-\gamma)$: as for value iteration, the price of γ close to 1 is a factor of order $\ln(1/(1-\gamma))/(1-\gamma)$, but the accuracy no longer enters, because policy iteration stops with an exact optimum. The bound is still pessimistic: for the bike store ($\gamma = 0.99$, $K = 459$) it allows $(256-16) \cdot 459 \approx 1.1 \cdot 10^5$ iterations, where three are needed (Example 8.14), but it replaces $A^S = 16^{16} \approx 1.8 \cdot 10^{19}$. The linear dependence on the size of the problem cannot be avoided. In the riverswim of Example 1.5 with $N = 20$ and $\gamma = 0.99$, started from "always L", the first improvement switches states 19 and 20 to R, and each later improvement switches only the next state to the left: twenty improvements, twenty-one evaluations, for a bound of 9 639. The information that swimming right pays travels one state per iteration (the numbers of this paragraph are computed by `figures/scripts/discounted_certificates.py`). A bound of the same kind holds for strategy iteration in turn-based zero-sum games (Hansen et al., 2013); with simultaneous moves, the naive variant that improves both players at once can cycle (Chapter 18).

Linear programming

The monotonicity of T^* characterizes $V^\star$ as the smallest solution of a system of linear inequalities.

Theorem 8.11 (Linear programming formulation). Let μ be a probability vector on $\mathcal{S}$ with $\mu(s) > 0$ for all s. The optimal value function $V^\star$ is the unique optimal solution of the linear program

$$\min_{V \in \mathbb{R}^S} \sum_s \mu(s) V(s) \quad \text{subject to} \quad V(s) \geq \sum_{s'} P(s' \mid s, a) \big[r(s,a,s') + \gamma V(s') \big] \quad \text{for all } (s,a).$$

Proof. The constraints say $V \geq T^* V$. By monotonicity, such a V satisfies $V \geq T^* V \geq (T^*)^2 V \geq \cdots \to V^\star$, so every feasible V dominates $V^\star$ componentwise, and $V^\star$ is feasible ($V^\star = T^* V^\star$). Since $\mu > 0$, any feasible $V \neq V^\star$ has a strictly larger objective. ∎

Remark 8.12 (Feasible starting points). A simplex method, or a projected or barrier first-order method, has to be started from a feasible point, and the proof shows what feasibility means: $V \geq T^*V$. Two such points are explicit.

(i) *The constant upper bound.* With $|r| \leq R_{\max}$, the constant vector $V_0 \equiv R_{\max}/(1-\gamma)$ is feasible: for every (s,a), $\sum_{s'} P(s' \mid s,a)[r(s,a,s') + \gamma V_0(s')] \leq R_{\max} + \gamma R_{\max}/(1-\gamma) = V_0(s)$. More generally, any V with $V \geq T^*V$ stays feasible after adding a nonnegative constant c, since $T^*(V+c) = T^*V + \gamma c \leq V + c$: the feasible set is closed under translation by $\mathbb{R}_+\mathbf{1}$, along which the objective increases by c.

(ii) *The value of a policy.* For any $\pi \in \Pi_{\mathrm{SR}}$, the value V^π of Theorem 7.11 satisfies $V^\pi = T^\pi V^\pi \leq T^*V^\pi$, which is the *wrong* inequality; but $V^\pi + c\mathbf{1}$ is feasible as soon as $c \geq \|T^*V^\pi - V^\pi\|_\infty/(1-\gamma)$, by the same computation as in (i). The advantage of this choice over the constant vector is that the shift c is small when π is nearly greedy for its own value, so the starting point is already close to the optimum; in particular $c = 0$ is admissible exactly when π is optimal.

Running value iteration from a feasible V keeps it feasible ($V \geq T^*V$ implies $T^*V \geq (T^*)^2V$) and decreases the objective at every step: value iteration started from above is itself a feasible descent method for the linear program, and Theorem 8.11 is the statement that it converges to the optimum (Remark 8.13).

By linear programming duality (Proposition C.9; the computation is Exercise 8.2), the dual program maximizes $\sum_{s,a} d(s,a)r(s,a)$, where $r(s,a) = \sum_{s'} P(s' \mid s,a)r(s,a,s')$ is the mean one-step reward (Definition 1.10), over nonnegative d satisfying the flow constraints $\sum_a d(s',a) = \mu(s') + \gamma \sum_{s,a} d(s,a)P(s' \mid s,a)$. The dual variable d ranges over the same kind of object as the unnormalized occupation measures $\tilde{d}^\pi(s,a) = \sum_{t\geq 1} \gamma^{t-1}\mathbb{P}^\pi_\mu(S_t = s, A_t = a)$ of the stationary policies, which are exactly the dual-feasible points (Exercise 8.2). Their normalized version,

$$d^\pi_\mu(s,a) \;\coloneqq\; (1-\gamma)\sum_{t\geq 1}\gamma^{t-1}\,\mathbb{P}^\pi_\mu(S_t = s, A_t = a), \qquad d^\pi_\mu(s) \;\coloneqq\; \sum_a d^\pi_\mu(s,a), \tag{8.1}$$

is the *discounted occupation measure* of π from μ: a probability distribution on $\mathcal{S}\times\mathcal{A}$, with $d^\pi_\mu(s,a) = d^\pi_\mu(s)\pi(a \mid s)$ and, writing $V^\pi(\mu) \coloneqq \sum_s \mu(s)V^\pi(s)$ for the value from the initial law μ, $V^\pi(\mu) = \frac{1}{1-\gamma}\sum_{s,a} d^\pi_\mu(s,a)r(s,a)$—the discounted counterpart of the finite-horizon occupation measures of Exercise 3.2, and the object through which Chapters 13 and 14 measure where a policy goes. An optimal policy is read from an optimal d as $\pi(a \mid s) \propto d(s,a)$ (Exercise 8.2). The linear program has S variables and SA constraints and is solvable in polynomial time. Its dual has S equality constraints, so a basic solution, a vertex of the dual polytope, has at most S nonzero coordinates; when $\mu > 0$, every state carries at least one of them, hence exactly one: vertices are the occupation measures of deterministic policies. The dual is also the natural formulation when linear constraints on the occupation measure are added, which gives the constrained MDPs of Section 8.6, where vertices have a few more nonzero coordinates, and optimal policies randomize.

Remark 8.13 (From above and from below). The two programs approach $V^\star$ from opposite sides. A primal-feasible V is a *super-solution*, $V \geq T^*V$, hence $V \geq V^\star$, and is in general the value of no policy; a dual-feasible d is the occupation measure of a policy π_d, and its objective is $V^{\pi_d}(\mu) \leq V^\star(\mu)$. Every method that keeps primal feasibility therefore produces upper bounds on $V^\star(\mu)$, and every method that keeps dual feasibility produces values of policies; along the simplex method, on either program, and along value iteration started from a point of Remark 8.12, these bounds are moreover monotone. Policy iteration is of the second kind: V^{π_k} increases to $V^\star$ (Theorem 8.9), and the improvement step is a block pivot of the simplex method on the dual,

from the vertex $\tilde{d}^{\pi_k}$ to the vertex $\tilde{d}^{\pi_{k+1}}$, with the action changed in every state at once. The simplex method proper, with Dantzig's rule, changes the action in one state per iteration, the one with the largest advantage $Q^{\pi_k}(s,a) - V^{\pi_k}(s)$; it is a policy iteration with a single switch, and Ye (2011) proves the strongly polynomial bound for it first. Value iteration from an arbitrary V_0 belongs to neither family: its iterates are monotone, and feasible for the primal, only when V_0 is a super-solution; from a sub-solution $V_0 \le T^*V_0$, such as $V_0 = 0$ when the rewards are nonnegative, they increase to $V^\star$ but are the values of no policy either. Finally, weak duality turns any pair into a certificate: for a primal-feasible V and any policy π,

$$V^\pi(\mu) \;\le\; V^\star(\mu) \;\le\; V(\mu),$$

and complementary slackness at the optimum reads $d(s,a) > 0 \Rightarrow V(s) = \sum_{s'} P(s' \mid s,a)[r(s,a,s') + \gamma V(s')]$: an optimal policy only uses actions that are greedy for $V^\star$, which is Theorem 8.3 again.

Example 8.14 (The bike store, solved). For the store of Example 7.4(i) ($M = 15$, $\gamma = 0.99$), policy iteration started from "never order" converges in three iterations, whereas value iteration needs 1 163 iterations to bring $\|V_k - V^\star\|_\infty$ below 10^{-3} (Figure 8.1), and the stopping rule of Algorithm 8.1 with $\varepsilon = 10^{-3}$ fires after 1 232: the factor γ^k is slow when $1/(1-\gamma) = 100$. But the span rule of Corollary 8.8, with the same ε, stops after 10 iterations. The iterate V_{10} is then still 108 away from $V^\star$, yet its greedy policy is already optimal, and the lower bound of the corollary is within 10^{-5} of $V^\star$. The overlap coefficient of the kernel is $\tau = 0.79$, because a demand of at least M bikes, of probability 0.21, empties the store whatever its stock and order, and the observed ratios $\operatorname{span}(V_{k+1} - V_k)/\operatorname{span}(V_k - V_{k-1})$ have median 0.18, much better than the guaranteed $\gamma\tau = 0.79$. Value iteration is slow at locating the *level* of $V^\star$, not at finding its *shape*, and the shape is all that the greedy policy needs. In the riverswim with $\gamma = 0.99$, where the states communicate slowly ($\tau = 1$), the gain is smaller: 89 instead of 1 175 iterations for $N = 4$, and 250 for $N = 20$ (all these numbers are computed by `figures/scripts/discounted_certificates.py`). The optimal policy is not one of the plain order-up-to rules of Example 7.4, which reorder as soon as the stock is below the level: it orders $10 - s$ bikes when the stock s is at most 3, and nothing otherwise—in inventory theory, an (s, S) policy with reorder point 3 and order-up-to level 10 (the classical name, whose letters have nothing to do with our state s and state-space size S; we mostly speak of the reorder point and the order-up-to level), the fixed ordering cost 0.8 being responsible for the gap between the two levels. The optimal value from an empty store is $V^\star(0) = 118.2$, that is, $1/(1-\gamma) = 100$ weeks' worth of a weekly profit of about 1.18.

Example 8.15 (Machine replacement, discounted). Discretizing the wear of Example 1.8 on a grid of step 0.1 with the simplified costs $c(x) = x$ and $C = 4$ of Theorem 3.22 (rather than the values of session H05 quoted in Example 7.4(ii)), exponential increments of mean 1 and $\gamma = 0.9$, value iteration returns a control-limit policy: replace when the wear exceeds $x^\star \approx 2.1$. The stationary threshold is the limit of the thresholds of the *discounted* finite-horizon problems, those of $(T^*)^k 0$, which stabilize after a few steps back from the deadline (Exercise 8.9); the undiscounted thresholds of Theorem 3.22 stabilize at a different value, about 2.0, which belongs to the average-cost problem (Example 9.12).

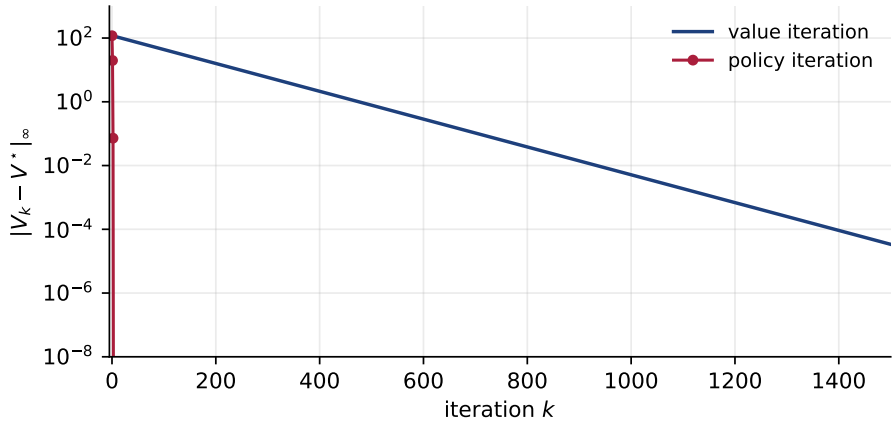


Figure 8.1. Convergence of value iteration and policy iteration on the bike store (Example 8.14, $\gamma = 0.99$): $\|V_k - V^\star\|_\infty$ on a logarithmic scale. Value iteration decays like γ^k; policy iteration (three points) reaches the optimum exactly.

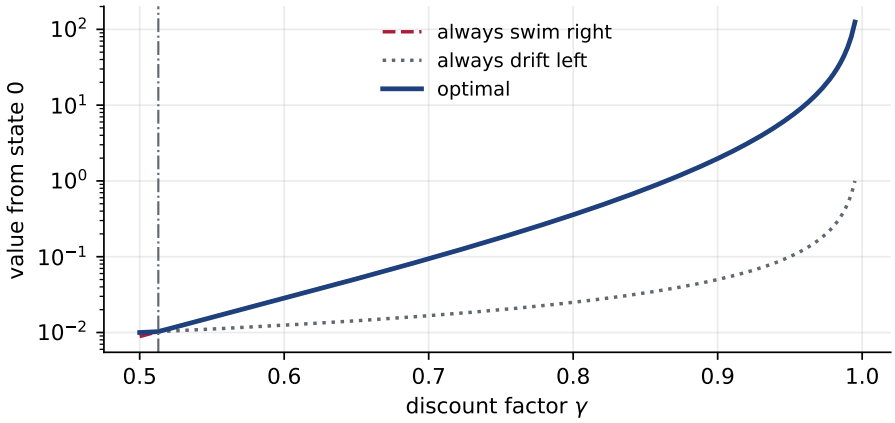


Figure 8.2. Discounted value from state 0 in the riverswim of Example 8.16, as a function of γ (log scale), for the two constant policies and for the optimal one, which coincides with "always swim right" beyond the discount factor marked by the dash-dotted line, above which swimming right is optimal in every state.

Example 8.16 (Riverswim: patience pays). In the riverswim of Example 1.5 with $N = 4$, $p = 0.3$, $q = 0.6$ and the small reward 0.005 for drifting left in state 0, the optimal policy depends on the discount factor (Figure 8.2). For $\gamma \leq 0.51$ the swimmer in state 0 prefers the sure trickle of 0.005 per step to the long struggle upstream, whose first reward arrives after at least four successful moves against the current; for $\gamma \geq 0.52$ she swims right from every state, and her value from state 0 grows from 0.01 at $\gamma = 0.5$ to 2.0 at $\gamma = 0.9$ and 58 at $\gamma = 0.99$. The discount factor is the agent's patience: it decides whether a distant reward is worth a long detour, exactly as the horizon H did in Exercise 3.4. In the learning setting of Chapter 15, the same example shows that an agent that has never seen the reward in state N has no reason to be patient.

8.3 The occupation measure

The dual variable of the linear program deserves a section of its own, because it returns in most of the later chapters. For $\pi \in \Pi_{\mathrm{H}}$ and an initial law μ, the discounted occupation measure d_μ^π of (8.1) is a probability distribution on $\mathcal{S} \times \mathcal{A}$, the law of (S_τ, A_τ) at a geometric time τ of parameter $1-\gamma$ drawn independently of the trajectory (Proposition 7.3). Let $\mathcal{K}_\mu \coloneqq \{d_\mu^\pi : \pi \in \Pi_{\mathrm{H}}\}$ be the set of occupation measures of all history-dependent policies, and let $\mathcal{D}_\mu$ be the polytope of the nonnegative $d \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ satisfying the *flow constraints*

$$\sum_a d(s',a) \;=\; (1-\gamma)\,\mu(s') + \gamma \sum_{s,a} d(s,a)\,P(s' \mid s,a) \qquad \text{for all } s' \in \mathcal{S}, \tag{8.2}$$

the constraints of the dual program after Theorem 8.11, scaled by $1-\gamma$. For $d \in \mathcal{D}_\mu$, write $d(s) \coloneqq \sum_a d(s,a)$ and let $\pi_d \in \Pi_{\mathrm{SR}}$ be the policy $\pi_d(a \mid s) \coloneqq d(s,a)/d(s)$ when $d(s) > 0$, and arbitrary otherwise.

Proposition 8.17 (The occupation polytope).

(i) $\mathcal{K}_\mu = \mathcal{D}_\mu$: the occupation measures of the history-dependent policies are exactly the solutions of the flow constraints, a convex compact polytope.

(ii) For every $d \in \mathcal{D}_\mu$, $d = d_\mu^{\pi_d}$. Hence every history-dependent policy has the occupation measure of a stationary policy, and $\mathcal{K}_\mu = \{d_\mu^\pi : \pi \in \Pi_{\mathrm{SR}}\}$.

(iii) If $\mu(s) > 0$ for all s, the vertices of $\mathcal{D}_\mu$ are the occupation measures of the stationary deterministic policies.

Proof. (i) Let $\pi \in \Pi_{\mathrm{H}}$. Whatever the history, S_{t+1} is drawn from $P(\cdot \mid S_t, A_t)$, so $\mathbb{P}_\mu^\pi(S_{t+1} = s') = \sum_{s,a} \mathbb{P}_\mu^\pi(S_t = s, A_t = a) P(s' \mid s,a)$, and $\mathbb{P}_\mu^\pi(S_1 = s') = \mu(s')$; multiplying by $(1-\gamma)\gamma^{t-1}$ and summing over $t \geq 1$ gives (8.2). So $\mathcal{K}_\mu \subset \mathcal{D}_\mu$, and the converse inclusion is (ii). The polytope is closed and, summing (8.2) over s', has $\sum_{s,a} d(s,a) = 1$, hence is bounded. (ii) Let $d \in \mathcal{D}_\mu$ and $\pi = \pi_d$, so that $d(s,a) = d(s)\pi(a \mid s)$ for all (s,a), including where $d(s) = 0$. The flow constraints read $d(s') = (1-\gamma)\mu(s') + \gamma \sum_s d(s) P^\pi(s' \mid s)$, a linear system in $d(\cdot)$ with the unique solution $(1-\gamma)\mu^\top (I - \gamma P^\pi)^{-1}$, which is also solved by the state occupation $d_\mu^\pi(s) = \sum_a d_\mu^\pi(s,a)$ by (i). Hence $d(s) = d_\mu^\pi(s)$ for all s, and $d(s,a) = d(s)\pi(a \mid s) = d_\mu^\pi(s,a)$. (iii) A vertex is a basic solution: its nonzero coordinates correspond to linearly independent columns of the $S \times SA$ matrix of (8.2), hence are at most S. When $\mu > 0$, every state has $d(s) \geq (1-\gamma)\mu(s) > 0$, hence at least one nonzero coordinate, hence exactly one: π_d is deterministic. Conversely, for a deterministic π, the columns indexed by $\{(s, \pi(s))\}$ form the matrix $(I - \gamma P^\pi)^\top$, which is invertible, so d_μ^π is a basic solution. ∎

Part (ii) is the discounted form of the Derman–Strauch theorem (Proposition 5.22). What makes the occupation measure useful is that it carries every expected additive criterion.

Proposition 8.18 (Linearity of expected additive criteria). For $\pi \in \Pi_{\mathrm{H}}$ and every $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,

$$\mathbb{E}_\mu^\pi\Big[\sum_{t\geq 1} \gamma^{t-1} f(S_t, A_t)\Big] \;=\; \frac{1}{1-\gamma} \sum_{s,a} d_\mu^\pi(s,a)\, f(s,a).$$

In particular $V^\pi(\mu) = \frac{1}{1-\gamma}\langle d_\mu^\pi, r\rangle$ for the mean one-step reward $r(s,a)$, and every discounted cost $C_k^\pi(\mu) = \mathbb{E}_\mu^\pi[\sum_t \gamma^{t-1} c_k(S_t, A_t)]$ is $\frac{1}{1-\gamma}\langle d_\mu^\pi, c_k\rangle$. Consequently, any problem whose objective and constraints are expected discounted sums, such as the constrained MDPs of Section 8.6, is a linear program over $\mathcal{D}_\mu$, and its supremum over Π_{H}, when the constraints can be met, is attained by a

stationary policy.

Proof. Fubini's theorem, the sum being absolutely convergent, and the definition (8.1); for the value, $\mathbb{E}^{\pi}_{\mu}[R_t \mid S_t, A_t] = r(S_t, A_t)$. The last claim is Proposition 8.17(ii) and the existence of an optimum of a linear function on a nonempty compact polytope; Theorem 8.26 gives the details for the constrained case. ∎

Remark 8.19 (What the occupation measure does not know). The occupation measure is the sufficient statistic for the mean, and for nothing beyond it. The law of the return W is not a function of d^{π}_{μ}: Exercise 8.3(c) exhibits two policies with the same occupation measure and different return laws, hence different probabilities $\mathbb{P}(W \le u)$, different variances and different β-means. The risk measures, utilities and constraints on the law of the return of Chapters 5 and 6 are therefore neither linear nor even functions on $\mathcal{D}_{\mu}$ (the probability of an *event of the trajectory* is linear, see Section 8.6), which is the structural reason why those criteria needed augmented states rather than a larger linear program. The exponential moments of the return are recovered from a different object, the tilted occupation measures of Chapter 14, which depend on the rate.

A policy evaluated from one pair rather than from an initial law gives the state–action version of the occupation measure, kept unnormalized so that values are read from it without constants.

Definition 8.20 (Successor measure). For $\pi \in \Pi_{\mathrm{SR}}$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$, the *successor measure* of π from (s, a) is

$$M^{\pi}(s, a; s', a') \;\coloneqq\; \sum_{t \ge 1} \gamma^{t-1}\, \mathbb{P}^{\pi}_{s,a}(S_t = s', A_t = a'),$$

where $\mathbb{P}^{\pi}_{s,a}$ conditions on $S_1 = s$ and $A_1 = a$: a measure of mass $1/(1-\gamma)$ on $\mathcal{S} \times \mathcal{A}$, with $M^{\pi}(s, a; s, a) \ge 1$.

Write P_{π} for the kernel on pairs, $P_{\pi}((s, a), (s', a')) \coloneqq P(s' \mid s, a)\pi(a' \mid s')$, whose state marginal is the state chain P^{π} of Theorem 7.11. Conditioning on the first transition gives $M^{\pi} = I + \gamma P_{\pi} M^{\pi}$, that is

$$M^{\pi} \;=\; (I - \gamma P_{\pi})^{-1} \;=\; \sum_{t \ge 0} \gamma^{t} P^{t}_{\pi}, \qquad \text{hence} \qquad Q^{\pi}_{r} \;=\; M^{\pi} r \quad \text{for every } r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}, \tag{8.3}$$

by comparison with the Q-form of the evaluation equation, $Q^{\pi}_{r} = r + \gamma P_{\pi} Q^{\pi}_{r}$. The occupation measure from μ is the average of the successor measures under the first pair, $d^{\pi}_{\mu} = (1-\gamma) \sum_{s,a} \mu(s)\pi(a \mid s) M^{\pi}(s, a; \cdot)$, and the state marginal $\sum_{a'} M^{\pi}(s, a; s', a')$, averaged over $a \sim \pi(\cdot \mid s)$, is the *successor representation* of Dayan (1993). The point of (8.3) is the order of the quantifiers: M^{π} does not depend on the reward, and once it is known, the action values of π for *any* reward are a matrix–vector product. This is the object on which zero-shot reinforcement learning is built (Section 15.5).

Remark 8.21 (Mixtures of policies). Draw an index i with probability α_i before the first step and follow π_i throughout. The law of the trajectory is the mixture of the laws, so the occupation measure of this *mixture policy* is $\sum_i \alpha_i d^{\pi_i}_{\mu}$; and the mixture is realized by a history-dependent policy, which at each step plays the posterior average of the $\pi_i(\cdot \mid S_t)$ given the past (Proposition 8.17(ii) then provides an equivalent stationary policy, which is not the mixture of the π_i as decision rules). Since a compact polytope is the convex hull of its vertices, Proposition 8.17(iii) says that $\mathcal{D}_{\mu}$ is, for $\mu > 0$, the convex hull of the finitely many occupation measures of the deterministic policies, and the mixture construction realizes each point of this hull by a policy. This is what makes a concave objective $F(d)$ tractable although it has no Bellman equation: the Frank–Wolfe method (Proposition C.14) maximizes at each step the linear function $\langle \nabla F(d_k), d \rangle$ over $\mathcal{D}_{\mu}$, a planning

> problem for the reward $\nabla F(d_k)$ solved by a deterministic policy, and mixes that policy into the current one. The visitation entropy of Section 15.4 is the leading example.

Where the occupation measure returns. The concentrability coefficients of Chapter 13 compare the occupation measure of a policy to the sampling distribution, and bound the error of approximate dynamic programming. The performance difference lemma of Chapter 14 (Lemma 14.1) writes $V^{\pi'}(\mu) - V^{\pi}(\mu)$ as the advantage of π averaged under $d_\mu^{\pi'}$, and the Fisher information of the natural gradient is averaged under d_μ^{π}. Off-policy evaluation by density ratios (Line 5) estimates $d_\mu^{\pi}/d^{\mathrm{b}}$ at the level of pairs, and escapes the exponential variance of trajectory-level importance weights. The visitation entropy and the convex MDPs of Chapter 15 are concave functionals on the occupation polytope (in its finite-horizon form $\mathcal{K}_P$ there), and zero-shot reinforcement learning (Section 15.5) learns successor measures. Feature matching and adversarial imitation (Chapter 16) match the occupation measure of the learner to that of the expert. The O-REPS algorithm of Chapter 17 runs mirror descent on the occupation polytope, and the mean-field games of Chapter 18 let the population flow, a linear image of the occupation measure, enter the rewards.

8.4 The distributional optimality operator

The distributional counterpart of T^* propagates full return distributions and selects, at each state, an action with the largest *mean*: $(\mathcal{T}^*\nu)(s) = \sum_{s'} P(s' \mid s, a_\nu(s))(b_{r(s,a_\nu(s),s'),\gamma})_\# \nu(s')$ with $a_\nu(s) \in \arg\max_a \sum_{s'} P(s' \mid s, a)[r(s,a,s') + \gamma \int x \, \mathrm{d}\nu(s')]$—the selected action depends on ν, which is the source of the trouble below. This is Algorithm 5.1 for the mean, in the discounted setting, and its means evolve exactly as value iteration. But $\mathcal{T}^*$ is *not* a contraction for any $\overline{W}_p$: a small change in ν can flip the maximizing action and replace a return distribution by a very different one with almost the same mean (Exercise 8.5). Bellemare et al. (2017) show that when the optimal policy is unique, the iterates $(\mathcal{T}^*)^k \nu_0$ nevertheless converge to the return distribution of the optimal policy, since the greedy actions eventually stabilize; when it is not unique, the iterates may oscillate between the return distributions of different optimal policies, which have the same mean but need not coincide (Example 3.17). All of this concerns the *exact*, unprojected operator. The distributional control algorithms of Chapters 12 and 13 do not inherit it automatically: they combine $\mathcal{T}^*$ with a projection onto a parametric family and with stochastic updates. A quantile projection changes the mean in general, and a categorical projection preserves it only when no mass is clipped at the ends of its support, so the means of projected iterates need not follow value iteration and their greedy actions need not be optimal; the convergence results proved for projected operators concern policy evaluation, not control. Three statements must be kept apart: exact distributional iteration, as above; projected policy evaluation, which converges to the fixed point of a projected operator; and stochastic projected control, for which neither the mean of an optimal policy nor its law is guaranteed in general. What does transfer is the negative part: even exact iteration does not say *which* optimal policy's law is reached when several exist, and, by Theorem 5.5, no change of the selection rule makes these algorithms optimize a non-Bellman-optimizable functional of the return.

8.5 Risk-sensitive planning with β-means

We now take the criterion $U_\beta(W)$ of the discounted return, $\beta \neq 0$, and look for $\sup_{\pi \in \Pi_{\mathrm{H}}} U_\beta(\nu^\pi(s))$, where for $\pi \in \Pi_{\mathrm{H}}$ we write $\nu^\pi(s)$ for the law of W under $\mathbb{P}_s^\pi$ (for non-stationary policies it no longer satisfies a stationary fixed-point equation). In the finite-horizon undiscounted case (Theorem 5.8), backward induction on β-values solved the problem exactly. With discounting, Proposition 7.16 showed that the β-mean of $r + \gamma W'$ involves the $\gamma\beta$-mean of W': the rate changes along the recursion, and the

Algorithm 8.3: Risk-sensitive planning with a risk-neutral tail

Input: discounted MDP $\mathcal{M}$, rate $\beta \neq 0$, number of risk-sensitive steps T'

compute $V^\star$ and a greedy stationary policy $\pi^\star_\infty$ (Algorithm 8.2); $\widehat{U}_{T'+1} \leftarrow V^\star$

for $t = T', \dots, 1$ **do**

  for $s \in \mathcal{S}$, $a \in \mathcal{A}$ **do**

    $\widehat{Q}_t(s,a) \leftarrow \frac{1}{\beta_t} \ln \sum_{s'} P(s' \mid s,a) \exp\big(\beta_t [r(s,a,s') + \gamma \widehat{U}_{t+1}(s')]\big)$ ▷ $\beta_t = \gamma^{t-1}\beta$

  $\widehat{U}_t(s) \leftarrow \max_a \widehat{Q}_t(s,a)$; $\hat{\pi}_t(s) \leftarrow$ any $a \in \arg\max_a \widehat{Q}_t(s,a)$

Output: $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_{T'}, \pi^\star_\infty, \pi^\star_\infty, \dots)$ and the value estimate $\widehat{U}_1$

natural objects are the values at rates

$$\beta_t \coloneqq \gamma^{t-1}\beta, \qquad t \geq 1.$$

For $\pi \in \Pi_{\mathrm{H}}$ and a history h_t, let $U^\pi_{\beta,t}(h_t) \coloneqq U_{\beta_t}(W_{t:\infty})$ under the continuation of h_t; for stationary π, $U^\pi_{\beta,t}(h_t) = U_{\beta_t}(v^\pi(s_t))$. (On notation: $U_\beta(\cdot)$ applied to a law or a random variable is the β-mean functional of Chapter 6; $U^\pi_{\beta,t}$, $U^\star_{\beta,t}$ and the $\widehat{U}_t$ of Algorithm 8.3 are value functions at the rate β_t, the index t always referring to the rate $\gamma^{t-1}\beta$; $U_\beta(s)$ of Proposition 7.16 is $U^\pi_{\beta,1}(s)$.) The tower property and the scaling rule $U_{\beta_t}(\gamma X) = \gamma U_{\beta_{t+1}}(X)$ (Proposition 2.17(iii)) give, exactly as in (5.1),

$$U^\pi_{\beta,t}(h_t) = U_{\beta_t}\Big(r(s_t, A, S') + \gamma\, U^\pi_{\beta,t+1}(h_t, A, S')\Big), \qquad A \sim \pi_t(\cdot \mid h_t),\ S' \sim P(\cdot \mid s_t, A). \tag{8.4}$$

Theorem 8.22 (Risk-sensitive planning with time-dependent rates, Chung and Sobel, 1987; Hau et al., 2023). Let $\beta \neq 0$ and $\Delta \coloneqq 2R_{\max}/(1-\gamma)$.

(i) *Finite horizon.* For the criterion $U_\beta(\sum_{t=1}^T \gamma^{t-1} R_t)$, define $U^\star_{\beta,T+1} \equiv 0$ and, for $t = T, \dots, 1$,

$$Q^\star_{\beta,t}(s,a) \coloneqq \frac{1}{\beta_t} \ln \sum_{s'} P(s' \mid s,a)\, e^{\beta_t [r(s,a,s') + \gamma U^\star_{\beta,t+1}(s')]}, \qquad U^\star_{\beta,t}(s) \coloneqq \max_a Q^\star_{\beta,t}(s,a).$$

Then $U^\star_{\beta,1}(s)$ is the optimal value over Π_{H}, and a deterministic Markov policy greedy with respect to $(Q^\star_{\beta,t})_t$ is optimal among all history-dependent policies. It is non-stationary in general: the rate β_t decreases with t.

(ii) *Infinite horizon.* The optimal value $U^\star_\beta(s) \coloneqq \sup_{\pi \in \Pi_{\mathrm{H}}} U_\beta(v^\pi(s))$ is the limit of the finite-horizon optimal values as $T \to \infty$, and Algorithm 8.3, which runs the recursion of (i) for T' steps from the terminal function $V^\star$ and then follows a risk-neutral optimal policy, returns a policy $\hat{\pi}$ with

$$U_\beta\big(v^{\hat{\pi}}(s)\big) \;\geq\; U^\star_\beta(s) - \frac{|\beta| \gamma^{2T'} \Delta^2}{8} \qquad \text{for all } s.$$

Proof. (i) is Theorem 5.8 with the recursion (8.4) in place of (5.1): the induction uses the monotonicity of U_{β_t} and the mixture bound (5.2), both valid at any rate, and the rate β_t is the one prescribed by the recursion. Non-stationarity is illustrated in Exercise 8.6.

(ii) *Truncation.* β-means are 1-Lipschitz for the supremum norm (by translation equivariance and monotonicity, Proposition 2.17(iii)–(iv)), so replacing $W_{t:\infty}$ by its T-step truncation changes $U^\pi_{\beta,t}(h_t)$ by at most $\gamma^T R_{\max}/(1-\gamma)$, uniformly in π and h_t; hence the finite-horizon optimal values of (i) converge to $U^\star_\beta$, and the same truncation argument shows that $\sup_{\pi \in \Pi_{\mathrm{H}}} U^\pi_{\beta,t}(h_t)$ depends on h_t only through s_t

(the finite-horizon optimal values of (i) do, being attained by Markov policies, and the truncation error is uniform in h_t); call it $U^\star_{\beta,t}(s_t)$ (so $U^\star_{\beta,1} = U^\star_\beta$). Passing to the limit in (i), these functions satisfy the infinite system $U^\star_{\beta,t}(s) = \max_a \frac{1}{\beta_t} \ln \sum_{s'} P(s' \mid s, a) e^{\beta_t [r + \gamma U^\star_{\beta,t+1}(s')]}$.

The tail. By Hoeffding's lemma (Lemma 2.10), for any law ν supported in an interval of length Δ and any rate β', $|U_{\beta'}(\nu) - \int x \, \mathrm{d}\nu| \le |\beta'|\Delta^2/8$. Applied at rate $\beta' = \beta_{T'+1} = \gamma^{T'}\beta$ to the return distributions of all policies, this gives $V^\star - \varepsilon_0 \le U^\star_{\beta,T'+1} \le V^\star$ with $\varepsilon_0 \coloneqq |\beta|\gamma^{T'}\Delta^2/8$, and, for the risk-neutral optimal policy $\pi^\star_\infty$, $U_{\beta_{T'+1}}(\nu^{\pi^\star_\infty}(s)) \ge V^\star(s) - \varepsilon_0$. (The display is written for $\beta < 0$. For $\beta > 0$ the inequalities are reversed: the terminal comparison becomes exact, $U^{\hat\pi}_{\beta,T'+1} = U_{\beta_{T'+1}}(\nu^{\pi^\star_\infty}) \ge V^\star = \widehat{U}_{T'+1}$, while ε_0 now appears in $U^\star_{\beta,T'+1} - \varepsilon_0 \le \widehat{U}_{T'+1} \le U^\star_{\beta,T'+1}$; the propagation below then gives $\widehat{U}_1 \ge U^\star_{\beta,1} - \gamma^{T'}\varepsilon_0$ and $U^{\hat\pi}_{\beta,1} \ge \widehat{U}_1$, the same final bound.)

Propagation. Each backward step $u \mapsto \big(s \mapsto \max_a \frac{1}{\beta_t} \ln \sum P \, e^{\beta_t [r + \gamma u(s')]}\big)$ is monotone and γ-Lipschitz for the supremum norm (translation equivariance and monotonicity of β-means, then the maximum). Starting from $U^\star_{\beta,T'+1} \le \widehat{U}_{T'+1} = V^\star \le U^\star_{\beta,T'+1} + \varepsilon_0$, induction gives $U^\star_{\beta,t} \le \widehat{U}_t \le U^\star_{\beta,t} + \gamma^{T'+1-t}\varepsilon_0$ for all $t \le T' + 1$. Finally, the policy $\hat\pi$: at times $t > T'$ it follows $\pi^\star_\infty$, so $U^{\hat\pi}_{\beta,T'+1} \ge V^\star - \varepsilon_0 = \widehat{U}_{T'+1} - \varepsilon_0$; for $t \le T'$, by (8.4) with the greedy action $\hat\pi_t(s)$ and the induction hypothesis $U^{\hat\pi}_{\beta,t+1} \ge \widehat{U}_{t+1} - \gamma^{T'-t}\varepsilon_0$,

$$U^{\hat\pi}_{\beta,t}(s) \ge U_{\beta_t}\Big(r + \gamma\big(\widehat{U}_{t+1}(S') - \gamma^{T'-t}\varepsilon_0\big)\Big) = \widehat{Q}_t(s, \hat\pi_t(s)) - \gamma^{T'+1-t}\varepsilon_0 = \widehat{U}_t(s) - \gamma^{T'+1-t}\varepsilon_0.$$

At $t = 1$: $U_\beta(\nu^{\hat\pi}(s)) = U^{\hat\pi}_{\beta,1}(s) \ge \widehat{U}_1(s) - \gamma^{T'}\varepsilon_0 \ge U^\star_\beta(s) - \gamma^{T'}\varepsilon_0$, and $\gamma^{T'}\varepsilon_0 = |\beta|\gamma^{2T'}\Delta^2/8$. ∎

Part (ii), with the risk-neutral tail and its $\gamma^{2T'}$ bound, is due to Hau et al. (2023); the time-dependent rates of part (i) go back to Chung and Sobel (1987). Three comments.

Remark 8.23 (Reading the theorem).

(i) *Two ways to fade.* Simply truncating the planning horizon at T' and following an arbitrary policy afterwards costs $O(\gamma^{T'})$; the risk-neutral tail costs $O(\gamma^{2T'})$, because the rate has already decayed to $\gamma^{T'}\beta$ when the tail starts, and the remaining risk sensitivity is of that order (Exercise 8.7). To reach an accuracy ε it suffices to take $T' \ge \ln(|\beta|\Delta^2/(8\varepsilon))/(2\ln(1/\gamma))$, so the total cost is $O(S^2 A \ln(1/\varepsilon)/(1-\gamma))$ beyond the risk-neutral solution.

(ii) *Non-stationarity is real.* The optimal risk-sensitive policy is in general not stationary, even for $T = \infty$: the action chosen in state s at time t is greedy for the rate $\gamma^{t-1}\beta$, and moves towards the risk-neutral action as t grows. A risk-averse agent in a discounted world behaves prudently at first and, as the rate applied to the remaining rewards shrinks, more and more like a risk-neutral one. This is a genuinely distributional phenomenon: it reflects the interplay between the discounting of rewards and the non-homogeneity of β-means.

(iii) *Units.* Since $U_\beta(\lambda W) = \lambda U_{\lambda\beta}(W)$, expressing rewards in cents instead of dollars multiplies the effective rate by 100; the bound of the theorem is invariant under a joint rescaling of r and β, and β must be chosen relative to the scale of the returns, as in Remark 5.9(ii). The a priori range $\Delta = 2R_{\max}/(1-\gamma)$ is often very pessimistic; the extreme values of Proposition 7.10, optimized over policies, give a realistic Δ.

Example 8.24 (Prudent bike store owners). Algorithm 8.3 with $T' = 800$ on the bike store (for $\beta < 0$ the Hoeffding step of the proof only involves the returns of $\pi^\star_\infty$, whose supports, computed by Proposition 7.10, have lengths at most $\Delta \approx 876$, and about 720 from the stocks $s \le 10$ that the

store actually visits; the guarantee of the theorem is then about $2 \cdot 10^{-3}$; with $T' = 200$ the bound would be vacuous, and the computed values are observed to exceed those at $T' = 800$ by up to 0.5) gives, for the first decision ($t = 1$, rate β), the following reorder points, order-up-to levels and optimal β-values from an empty store:

β	-0.2	-0.1	-0.05	-0.02	-0.005	0
reorder point	2	2	3	3	3	3
order-up-to level	7	8	9	10	10	10
$U^\star_{\beta,1}(0)$	94.3	104.6	110.9	115.3	117.5	118.2

A more risk-averse owner keeps a smaller stock: the downside she fears is a week of low demand with a full warehouse, and she gives up part of the expected profit ($U^\star_{-0.1,1}(0) = 104.6$ is the certainty equivalent of her random profit, against an expectation of 118.2 for the risk-neutral owner) to avoid it. At later times the rate $\gamma^{t-1}\beta$ is smaller and the same owner orders more: her policy is not stationary.

The entropic value at risk in discounted MDPs ♦

Combining Algorithm 6.1 with Algorithm 8.3 gives an algorithm for $\max_{\pi \in \Pi_{\mathrm{H}}} \mathrm{EVaR}_\delta(v^\pi(s))$ in discounted MDPs. Write $h_\delta(\beta) = U^\star_{-\beta}(s) - \ln(1/\delta)/\beta$ for $\beta > 0$, with $U^\star_{-\beta}$ the infinite-horizon optimal $(-\beta)$-value of Theorem 8.22(ii); the reduction of Proposition 6.29 holds verbatim.

Corollary 8.25 (EVaR planning in discounted MDPs). Let $b_1 < \cdots < b_K$ be the grid of Algorithm 6.1 for the parameters $(\delta, \varepsilon, \Delta)$ (written $\beta_1 < \cdots < \beta_K$ there; the letter β_t is reserved in this section for the decayed rates $\gamma^{t-1}\beta$), and for each k let $\hat\pi_k$ and $\widehat{U}^{(k)}_1$ be returned by Algorithm 8.3 with rate $-b_k$ and $T' \geq \ln(b_K \Delta^2/(8\varepsilon))/(2\ln(1/\gamma))$. Then the policy $\hat\pi_{k^\star}$ with $k^\star \in \arg\max_k \{\widehat{U}^{(k)}_1(s) - \ln(1/\delta)/b_k\}$ satisfies

$$\mathrm{EVaR}_\delta\left(v^{\hat\pi_{k^\star}}(s)\right) \;\geq\; \sup_{\pi \in \Pi_{\mathrm{H}}} \mathrm{EVaR}_\delta\left(v^{\pi}(s)\right) - 2\varepsilon,$$

at a total cost of $O(KS^2AT')$ beyond one risk-neutral solve, with $K = O(\Delta^2 \ln(1/\delta)/\varepsilon^2)$.

Proof. The proof of Theorem 6.31 shows $\sup_\beta h_\delta(\beta) \leq \max_k h_\delta(b_k) + \varepsilon$; it used only the monotonicity of $\beta \mapsto U^\star_{-\beta}$ and Hoeffding's lemma, both valid here. By Theorem 8.22(ii) and its proof, $U^\star_{-b_k} \leq \widehat{U}^{(k)}_1 \leq U^\star_{-b_k} + \varepsilon_k$ and $U_{-b_k}(v^{\hat\pi_k}(s)) \geq \widehat{U}^{(k)}_1(s) - \varepsilon_k$ with $\varepsilon_k = b_k \gamma^{2T'} \Delta^2/8 \leq \varepsilon$. Hence, with $\hat h_k \coloneqq \widehat{U}^{(k)}_1(s) - \ln(1/\delta)/b_k$,

$$\mathrm{EVaR}_\delta(v^{\hat\pi_{k^\star}}(s)) \geq U_{-b_{k^\star}}(v^{\hat\pi_{k^\star}}(s)) - \frac{\ln(1/\delta)}{b_{k^\star}} \geq \hat h_{k^\star} - \varepsilon \geq \max_k \hat h_k - \varepsilon \geq \max_k h_\delta(b_k) - \varepsilon \geq \sup_\beta h_\delta(\beta) - 2\varepsilon.$$

∎

8.6 Constrained MDPs

Many problems come with a second objective that is not traded against the first but bounded: a service level, a budget, a probability of accident. Let $c_1, \ldots, c_K : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be cost functions, $\kappa_1, \ldots, \kappa_K$ thresholds and μ an initial law. For $\pi \in \Pi_{\mathrm{H}}$, the discounted costs are $C^\pi_k(\mu) \coloneqq \mathbb{E}^\pi_\mu\big[\sum_{t \geq 1} \gamma^{t-1} c_k(S_t, A_t)\big]$,

and the *constrained MDP* is the problem

$$\sup_{\pi\in\Pi_{\mathrm{H}}} V^{\pi}(\mu) \quad \text{subject to} \quad C_k^{\pi}(\mu) \le \kappa_k, \qquad k = 1, \dots, K. \tag{8.5}$$

Two features of Theorem 8.3 are lost. The solution depends on μ: the constraint couples the decisions taken at different states, and there is no Bellman equation in the state alone. And deterministic policies are no longer enough. Both facts are read from the dual linear program. For $\pi \in \Pi_{\mathrm{H}}$, write $\tilde{d}^{\pi}(s,a) \coloneqq \sum_{t\ge 1} \gamma^{t-1}\mathbb{P}_{\mu}^{\pi}(S_t = s, A_t = a) = d_{\mu}^{\pi}(s,a)/(1-\gamma)$ for the unnormalized occupation measure, as after Theorem 8.11. By Proposition 8.18,

$$V^{\pi}(\mu) = \sum_{s,a} \tilde{d}^{\pi}(s,a)\, r(s,a), \qquad C_k^{\pi}(\mu) = \sum_{s,a} \tilde{d}^{\pi}(s,a)\, c_k(s,a) :$$

the objective and the constraints are linear functions of the occupation measure.

Theorem 8.26 (Constrained MDPs as linear programs). Assume that some $\pi \in \Pi_{\mathrm{H}}$ satisfies the constraints of (8.5), and consider the linear program

$$\max_{d\ge 0} \sum_{s,a} d(s,a) r(s,a) \quad \text{subject to} \quad \sum_{a} d(s',a) = \mu(s') + \gamma \sum_{s,a} d(s,a) P(s' \mid s,a) \quad \text{for all } s',$$
$$\sum_{s,a} d(s,a) c_k(s,a) \le \kappa_k \quad \text{for all } k.$$

(i) The program has an optimal solution, and its value is the supremum in (8.5). For every optimal d, the stationary policy $\pi_d(a \mid s) \coloneqq d(s,a)/\sum_{a'} d(s,a')$ (any action where the denominator vanishes) is optimal for (8.5) among all history-dependent policies.

(ii) Some optimal solution is a basic solution, and for it the numbers $n_s \coloneqq |\{a : d(s,a) > 0\}|$ satisfy $\sum_{s:\, n_s \ge 1} (n_s - 1) \le K$. In particular there is an optimal stationary policy that randomizes in at most K states.

Proof. (i) The flow constraints are those of (8.2) scaled by $1/(1-\gamma)$, so by Proposition 8.17 the points satisfying them are exactly the $\tilde{d}^{\pi}$, $\pi \in \Pi_{\mathrm{H}}$, and each of them is $\tilde{d}^{\pi_d}$ for the stationary policy π_d. *Every policy gives a feasible point:* by the display above, $\tilde{d}^{\pi}$ satisfies the cost constraints when π does, with objective $V^{\pi}(\mu)$, so the value of the program is at least the supremum in (8.5). *Every feasible point is a stationary policy:* a feasible d equals $\tilde{d}^{\pi_d}$, so π_d has value $\sum d\, r$ and costs $\sum d\, c_k \le \kappa_k$. *Existence.* The feasible set is nonempty by assumption, closed, and bounded, since $\sum_{s,a} d(s,a) = 1/(1-\gamma)$. A linear function attains its maximum on such a polytope, at a vertex.

(ii) With slack variables $u_k \ge 0$, the cost constraints become equalities $\sum d\, c_k + u_k = \kappa_k$, and the program has $S+K$ equality constraints. A vertex is a basic solution: the columns of the constraint matrix that correspond to its nonzero coordinates are linearly independent. Let $Z \coloneqq \{s : d(s) > 0\}$. For $s' \notin Z$ the flow constraint reads $0 = \mu(s') + \gamma \sum_{s,a} d(s,a) P(s' \mid s,a)$, so $P(s' \mid s,a) = 0$ whenever $d(s,a) > 0$. The column of a variable $d(s,a) > 0$, namely $\big(e_s - \gamma P(\cdot \mid s,a),\, c_1(s,a), \dots, c_K(s,a)\big)$, therefore vanishes in the rows $s' \notin Z$, and so do the columns of the slacks. All these columns live in a space of dimension $|Z| + K$, so there are at most $|Z| + K$ of them: $\sum_{s\in Z} n_s \le |Z| + K$. Outside Z the policy π_d can be chosen deterministic. ∎

When $K = 0$ this is the vertex statement after Theorem 8.11. The result is classical (Altman, 1999). The bound concerns stationary policies; with history-dependent policies the need for randomization can sometimes be circumvented and sometimes not (Exercise 8.3).

Probabilities of failure are linear. Let $B \subset \mathcal{S}$ be an absorbing set of bad states and $c(s,a) \coloneqq \mathbb{1}\{s \notin B\}P(B \mid s,a)$. Since B is entered at most once, $C^\pi(\mu) = \sum_t \gamma^{t-1}\mathbb{P}^\pi_\mu(S_t \notin B, S_{t+1} \in B)$ is, by Proposition 7.3, the probability of entering B before an independent geometric deadline τ with $\mathbb{P}(\tau \geq t) = \gamma^{t-1}$. A chance constraint on an *event of the trajectory* is thus a linear constraint on $\tilde{d}^\pi$. A constraint on the *law of the return*, such as $\mathbb{P}(W \leq u) \leq \delta$, is of a different nature: the occupation measure does not even determine the law of W (Exercise 8.3(c)), and such constraints call for the augmented state of Chapter 5.

Example 8.27 (A safe hike before nightfall). Take the windy hike of Example 1.4 without its horizon H: night falls at each step with probability 0.1, and the hikers' expected time at the bar before nightfall is their discounted value with $\gamma = 0.9$ (Proposition 7.3). With B the sink and the cost above, C^π is the probability of falling into the pit before nightfall. The safe route along the top row has value 2.93 and falls with probability 0.002; the route along the pit has value 3.37 and falls with probability 0.23; the unconstrained optimum walks along the third row, with value 3.75 and a fall probability of 0.032. With the constraint $C^\pi \leq \delta = 0.01$, the linear program of Theorem 8.26 has value 3.34 and randomizes in a single cell, the first cell of the third row, which the hikers reach with discounted occupation 0.90: there they go north with probability 0.67, towards the top row, and east with probability 0.33, along the third row. The best stationary deterministic policy satisfying the constraint has value 3.17 only. The multiplier of the constraint is $\lambda^\star = 29.1$, the marginal value of risk: allowing one more fall in a thousand hikes buys 0.029 of expected time at the bar. The two policies that take one of the two actions in that cell, and coincide with the randomized one elsewhere, have values 3.16 and 3.69 and fall probabilities 0.004 and 0.022; both are optimal for the reward $r - \lambda^\star c$ (computed by `figures/scripts/cmdp_hike.py`).

Lagrangian, front and randomization ♦. The last observation is general. By linear programming duality (Propositions C.8 and C.9), for $K = 1$ and a policy satisfying the constraint strictly, the value $F(\kappa)$ of (8.5) is

$$F(\kappa) = \min_{\lambda \geq 0} \left\{ \max_{\pi \in \Pi_{\mathrm{SD}}} V^\pi_{r-\lambda c}(\mu) + \lambda\kappa \right\},$$

where $V^\pi_{r-\lambda c}$ is the value for the reward $r - \lambda c$. For each λ the inner problem is an ordinary MDP, solved by deterministic policies; as λ varies, its optimal policy is piecewise constant, exactly as $\beta \mapsto \pi^\star_\beta$ on the optimality front of Definition 6.33, with λ in the role of the rate. F is concave, nondecreasing and piecewise linear in κ, with supergradient $\lambda^\star$ at κ. At the minimizing $\lambda^\star$, the optimal solution $d^\star$ of the constrained problem maximizes the Lagrangian (Proposition C.8(iii)), a linear function, over the flow polytope. When $d^\star$ randomizes between two actions in one state, it lies inside the segment joining the occupation measures of the two deterministic policies obtained by fixing either action there (the points of the flow polytope supported by the pairs that $d^\star$ uses form a polytope of dimension one), so both endpoints maximize the Lagrangian too. When the constraint is tight, one of them typically violates it and the other satisfies it, and the constrained optimum is the coin toss between them that makes the constraint tight. A scalarized front gives the candidates; the constraint is met by randomizing between two neighbours.

This is in sharp contrast with the criteria of Chapters 5 and 6. For β-means, EVaR and CVaR (the latter on the augmented state), deterministic policies suffice (Theorems 5.8 and 6.30 and Corollary 5.19). The reason is the mixture bound (5.2): a criterion ρ with $\rho(\sum_i p_i \nu_i) \leq \max_i \rho(\nu_i)$ never gains from a coin toss. It holds for β-means; for CVaR_δ, a supremum of linear functionals of the law, hence convex; and for EVaR, a supremum of β-means shifted by constants. The constrained criterion, $\mathbb{E}[W]$ if the expected cost is at most κ and $-\infty$ otherwise, violates it: mixing a policy that is too costly with one that is too cautious can be feasible and better than every feasible deterministic policy. Randomization is the price of a hard constraint, as it was the price of an adversary in Chapter 17. The Lagrangian also

underlies the learning versions of the problem, which alternate policy-gradient steps on $r - \lambda c$ with gradient steps on λ (Chapter 14; the penalized form of Chapter 16 is the same relaxation).

8.7 Episodes of random length

A total of rewards over an infinite horizon can be finite for three reasons. The horizon may be finite (Chapter 3); the rewards may be discounted (Chapter 7); or the episode may *end*, at a random time that the policy influences: the hiker reaches the bar, the game of tennis is won, the machine is scrapped. This section treats the third case, with $\gamma = 1$ and no horizon. It is the natural infinite-horizon setting for the return distribution: the law of the return exists without any discount, the moments and the exponential moments are Bellman closed at a *fixed* rate, and, unlike in the discounted case, risk-sensitive optimal policies are stationary.

Model. Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r)$ be a finite MDP with a *terminal* state $\Delta \in \mathcal{S}$, absorbing with zero reward: $P(\Delta \mid \Delta, a) = 1$ and $r(\Delta, a, \Delta) = 0$ for all a (several terminal states are merged into one). Write $\mathcal{S}^\circ := \mathcal{S} \setminus \{\Delta\}$. From $S_1 = s \in \mathcal{S}^\circ$, the *length* of the episode and its *return* are

$$N := \inf\{t \geq 1 : S_{t+1} = \Delta\} \in \{1, 2, \dots\} \cup \{\infty\}, \qquad W := \sum_{t=1}^{N} R_t, \quad R_t = r(S_t, A_t, S_{t+1}).$$

A stationary policy π is *proper* if $\mathbb{P}_s^\pi(N < \infty) = 1$ for every $s \in \mathcal{S}^\circ$. Its *transient kernel* Q^π is the restriction of P^π to $\mathcal{S}^\circ \times \mathcal{S}^\circ$, a substochastic matrix; $(Q^\pi)^n \mathbf{1}(s) = \mathbb{P}_s^\pi(N > n)$, since the n-step paths that stay in $\mathcal{S}^\circ$ are those of the episodes still running after n steps.

Lemma 8.28 (Proper policies). For $\pi \in \Pi_{\mathrm{SR}}$, the following are equivalent: (a) π is proper; (b) Δ is reachable from every state of the chain P^π; (c) $\rho(Q^\pi) < 1$, where ρ denotes the spectral radius. Then $I - Q^\pi$ is invertible, $\mathbb{E}_s^\pi[N] = ((I - Q^\pi)^{-1}\mathbf{1})(s)$, all the moments $\mathbb{E}_s^\pi[N^k]$ are finite, and $\max_s \mathbb{P}_s^\pi(N > n)^{1/n} \to \rho(Q^\pi)$.

Proof. (a)⇒(b) is clear, and (b)⇒(a) with a geometric tail $\mathbb{P}_s^\pi(N > kS) \leq (1-p)^k$ is Proposition A.6. Since $\|(Q^\pi)^n\|_\infty = \max_s \mathbb{P}_s^\pi(N > n)$, Gelfand's formula $\|(Q^\pi)^n\|_\infty^{1/n} \to \rho(Q^\pi)$ gives the last claim, and the geometric tail gives (c). Conversely, if $\rho(Q^\pi) < 1$ then $(Q^\pi)^n \to 0$, which is (a). The Neumann series $\sum_n (Q^\pi)^n = (I - Q^\pi)^{-1}$ converges, and $\mathbb{E}_s^\pi[N] = \sum_{n \geq 0} \mathbb{P}_s^\pi(N > n)$; the geometric tail makes every moment finite. ∎

Two classical assumptions make the theory work; the first is the easier one.

(T) *Transient model:* every $\pi \in \Pi_{\mathrm{SD}}$ is proper.

(C) *Shortest path with costs:* some $\pi \in \Pi_{\mathrm{SD}}$ is proper, and every step costs at least $c > 0$: $r(s, a, s') \leq -c$ for all $s \in \mathcal{S}^\circ$, a, s'.

A finite-horizon MDP, made stationary by adding the time to the state, is a transient model; so is a discounted MDP read as a killed one (Proposition 7.3): an extra terminal state is reached with probability $1 - \gamma$ at every step, whatever the policy. In both, the length of the episode does not depend on the policy; in a stochastic shortest path it does. Under (C), an improper policy has an infinite expected cost from some state, since it runs for ever with positive probability; (C) is a special case of the assumptions of Bertsekas and Tsitsiklis (1991) (some policy is proper, and improper policies have infinite cost from some state), sufficient for our examples.

Evaluation: the law of the return and its statistics

For a proper π, $|W| \le R_{\max} N$, so W is finite almost surely and has finite moments of every order by Lemma 8.28. Let $\nu^\pi(s) \coloneqq \mathcal{L}_{\mathbb{P}_s^\pi}(W)$ for $s \in \mathcal{S}^\circ$ and $\nu^\pi(\Delta) \coloneqq \delta_0$. Conditioning on (A_1, S_2) and using the restart property exactly as in Theorem 7.5 gives the distributional Bellman equation with $\gamma = 1$,

$$\nu^\pi(s) = \sum_{a,s'} \pi(a \mid s) P(s' \mid s, a)\, (b_{r(s,a,s')})_{\#} \nu^\pi(s') \quad (s \in \mathcal{S}^\circ), \qquad \nu^\pi(\Delta) = \delta_0 : \tag{8.6}$$

ν^π is a fixed point of the operator $\mathcal{T}^\pi$ of Definition 7.6 with $\gamma = 1$, acting on the return-distribution functions with $\nu(\Delta) = \delta_0$. Integrating x^i against (8.6) gives the moment recursion of Proposition 7.13 with $\gamma = 1$, in which the coefficient of $M^{(i)}$ is now the transient kernel: given the lower moments, $M^{(i)} = Q^\pi M^{(i)} +$ (known terms), a linear system with matrix $I - Q^\pi$, invertible by Lemma 8.28. The moments are Bellman closed *without discount*. In particular, with $M^{(i)}(\Delta) = 0$,

$$\begin{aligned} V^\pi &= (I - Q^\pi)^{-1} r^\pi, \qquad \sigma^2 = (I - Q^\pi)^{-1} c^\pi, \\ c^\pi(s) &\coloneqq \sum_{a,s'} \pi(a \mid s) P(s' \mid s, a) \big[r(s,a,s') + V^\pi(s') - V^\pi(s) \big]^2, \end{aligned} \tag{8.7}$$

by the law of total variance as in (7.4). What replaces the factor γ in the contraction arguments is the expected time to absorption.

Proposition 8.29 (Weighted contractions). Let $b : \mathcal{S}^\circ \to [1, \infty)$ and Π_b the set of decision rules π such that

$$\sum_a \pi(a \mid s) \sum_{s' \in \mathcal{S}^\circ} P(s' \mid s, a)\, b(s') \;\le\; b(s) - 1 \qquad \text{for all } s \in \mathcal{S}^\circ, \tag{8.8}$$

and $\rho_b \coloneqq 1 - 1/\|b\|_\infty < 1$. Let $\|V\|_b \coloneqq \max_{s \in \mathcal{S}^\circ} |V(s)|/b(s)$.

(i) For $\pi \in \Pi_b$, $T^\pi V \coloneqq r^\pi + Q^\pi V$ is a ρ_b-contraction for $\|\cdot\|_b$ on $\mathbb{R}^{\mathcal{S}^\circ}$. If (8.8) holds for every action, i.e. for every deterministic rule, then so is $T^* V(s) \coloneqq \max_a \sum_{s'} P(s' \mid s, a)\,[r(s,a,s') + V(s')]$ (with $V(\Delta) = 0$).

(ii) For $\pi \in \Pi_b$ and return-distribution functions ν, ν' with finite p-th moments and $\nu(\Delta) = \nu'(\Delta) = \delta_0$,

$$\max_{s \in \mathcal{S}^\circ} \frac{\mathcal{W}_p\big((\mathcal{T}^\pi \nu)(s), (\mathcal{T}^\pi \nu')(s)\big)^p}{b(s)} \;\le\; \rho_b \max_{s \in \mathcal{S}^\circ} \frac{\mathcal{W}_p\big(\nu(s), \nu'(s)\big)^p}{b(s)} :$$

with $d_{b,p}(\nu, \nu') \coloneqq \max_{s \in \mathcal{S}^\circ} \mathcal{W}_p(\nu(s), \nu'(s))/b(s)^{1/p}$, the operator $\mathcal{T}^\pi$ is a $\rho_b^{1/p}$-contraction for $d_{b,p}$, in particular a ρ_b-contraction for $d_{b,1}$, the weighted supremum-$\mathcal{W}_1$ distance. Hence ν^π is its unique fixed point among such functions, and $(\mathcal{T}^\pi)^k \nu_0 \to \nu^\pi$ geometrically in these distances.

(iii) A proper π belongs to Π_b for $b = b^\pi \coloneqq (I - Q^\pi)^{-1}\mathbf{1}$, its expected time to absorption. Under (T), the maximal expected time to absorption $b^\star(s) \coloneqq \max_{\pi \in \Pi_{\mathrm{SD}}} b^\pi(s)$ solves $b^\star(s) = 1 + \max_a \sum_{s' \in \mathcal{S}^\circ} P(s' \mid s, a) b^\star(s')$, so that every policy belongs to $\Pi_{b^\star}$; moreover $\mathbb{E}_s^\sigma[N] \le b^\star(s)$ and $\mathbb{P}_s^\sigma(N > k) \le \rho_{b^\star}^k \|b^\star\|_\infty$ for every $\sigma \in \Pi_{\mathrm{H}}$.

Proof. (i) $|(Q^\pi V)(s)| \le \sum_{s'} Q^\pi(s,s') b(s') \|V\|_b \le (b(s) - 1)\|V\|_b$, and $(b(s) - 1)/b(s) \le \rho_b$. For T^*, apply this to each action and use $|\max_a x_a - \max_a y_a| \le \max_a |x_a - y_a|$. (ii) Shifting by r is an isometry of $\mathcal{W}_p$ (Lemma D.7(i) with $\gamma = 1$), and the mixture bound of Lemma D.7(ii) gives, the

transitions to Δ contributing $\mathcal{W}_p(\delta_0,\delta_0)=0$, $\mathcal{W}_p((\mathcal{T}^\pi v)(s),(\mathcal{T}^\pi v')(s))^p \le \sum_{a,s'\in\mathcal{S}^\circ}\pi(a\mid s)P(s'\mid s,a)\,b(s')\,d \le (b(s)-1)\,d$, with d the right-hand maximum without ρ_b; divide by $b(s)$. Two fixed points are at distance $d\le\rho_b d$, hence equal, and $d((\mathcal{T}^\pi)^k v_0,v^\pi)\le\rho_b^k\,d(v_0,v^\pi)$ since v^π is a fixed point. (iii) $b^\pi=\mathbf{1}+Q^\pi b^\pi$ is (8.8) with equality. Under (T), run policy iteration for the reward +1 per step: if π' is greedy for b^π (keeping $\pi(s)$ when it is greedy), then $\mathbf{1}+Q^{\pi'}b^\pi\ge b^\pi$, hence $b^\pi\le(T^{\pi'})^k b^\pi\to b^{\pi'}$ as π' is proper, with strict improvement where π' differs from π. No policy is visited twice, so the iteration stops at a $\pi^\star$ whose $b^{\pi^\star}$ solves the displayed equation, and $b^{\pi^\star}\ge b^\pi$ for all π by the same comparison, so $b^{\pi^\star}=b^\star$. Finally, for $\sigma\in\Pi_{\mathrm{H}}$ and $\mathfrak{F}_t=\sigma(S_1,A_1,\dots,S_t)$, (8.8) for every action gives $\mathbb{E}[b(S_{t+1})\mathbb{1}\{N\ge t+1\}\mid\mathfrak{F}_t]\le(b(S_t)-1)\mathbb{1}\{N\ge t\}\le\rho_b\,b(S_t)\mathbb{1}\{N\ge t\}$; summing the first inequality over $t\le k$ gives $\mathbb{E}\sum_{t\le k}\mathbb{1}\{N\ge t\}\le b(s)$, and iterating the second, $\mathbb{P}(N>k)\le\mathbb{E}[b(S_{k+1})\mathbb{1}\{N>k\}]\le\rho_b^k b(s)$. ∎

The weights $b^\star$ are the ones used for Q-learning in episodic problems (Section 12.4): under (T), T^* is a contraction for $\|\cdot\|_{b^\star}$, although not in general for $\|\cdot\|_\infty$ (a state with a self-loop of probability close to 1 makes the supremum-norm modulus close to 1). The modulus $1-1/\|b\|_\infty$ is the counterpart of γ, and $\|b\|_\infty$ that of the effective horizon $1/(1-\gamma)$: for a killed discounted MDP, $b^\star\equiv 1/(1-\gamma)$ and $\rho_{b^\star}=\gamma$. In (ii) the distance $\mathcal{W}_\infty$ is lost: there is no contraction for $\overline{\mathcal{W}}_\infty$, since a shift by 1 of ν survives on the event $\{N>k\}$, of positive probability for every k (Exercise 8.11).

Example 8.30 (The windy hike without horizon). Take the 4×6 hike of Example 1.4 with $p=0.1$, remove the horizon, and change the rewards: every step costs 1, the bar is the terminal state, and the pit sends the hiker back to the start S, as in the cliff walking of Sutton and Barto (2018, Example 6.6). The return is $W=-N$, minus the number of steps to the bar. Assumption (C) holds, (T) does not: walking into a wall for ever is an improper policy. Three stationary policies cross the grid along row 2 (the risky path along the pit, 7 steps without wind), row 3 (9 steps) and row 4 (the safe path, 11 steps); a gust pushes the hiker one row down, into the pit from row 2. Their laws (Figure 8.3, left, computed by iterating (8.6)) have means $\mathbb{E}N=9.82$, 10.73 and 12.86 and standard deviations 4.41, 2.45 and 2.27, which (8.7) reproduces. The risky path is the risk-neutral optimum (value iteration gives $V^\star(\mathrm{S})=-9.82$). It arrives in 7 steps with probability 0.53, but it pays for each fall with a new start: its tail is the heaviest, $\mathbb{P}(N>30)=4\cdot10^{-3}$ against at most $2.3\cdot10^{-4}$ for the other two, and $\rho(Q^\pi)=0.81$, 0.69 and 0.65 respectively. The expected absorption time $\|b^\pi\|_\infty=9.82$ gives the risky policy the evaluation modulus $\rho_{b^\pi}=0.90$ of Proposition 8.29.

Control in expectation

Theorem 8.31 (Optimality for random-length episodes). Assume (T) or (C), and let $V^\star(s)\coloneqq\sup_{\sigma\in\Pi_{\mathrm{H}}}\mathbb{E}_s^\sigma[W]$ for $s\in\mathcal{S}^\circ$ (under (C), $W\in[-\infty,0]$ and expectations may be $-\infty$).

(i) $V^\star$ is finite and is the unique solution in $\mathbb{R}^{\mathcal{S}^\circ}$ of the optimality equation $V=T^*V$ (with $V(\Delta)=0$).

(ii) Every stationary deterministic policy greedy with respect to $V^\star$ is proper and optimal among all history-dependent policies.

(iii) Value iteration $V_{k+1}=T^*V_k$ converges to $V^\star$ from every V_0; under (T), $\|V_k-V^\star\|_{b^\star}\le\rho_{b^\star}^k\|V_0-V^\star\|_{b^\star}$.

(iv) Policy iteration started from a proper policy visits only proper policies and stops after

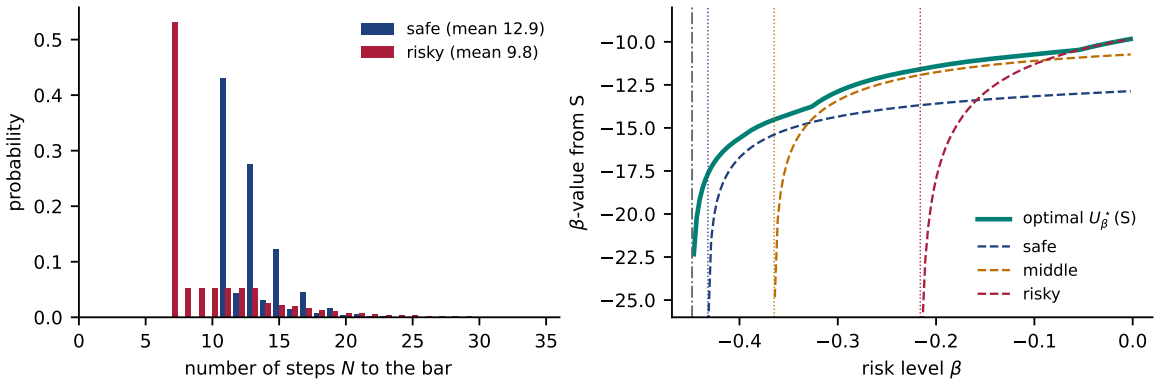


Figure 8.3. The windy hike without horizon (Examples 8.30 and 8.35). Left: law of the number of steps N to the bar under the safe path (row 4) and the risky path (row 2). Right: optimal β-value $U_\beta^\star(\mathrm{S})$ (solid) and β-values of the three fixed paths (dashed), with their critical levels $\ln \rho(Q^\pi)$ (dotted) and the critical level $\beta_c^\star = -0.448$ of the problem (dash-dotted), below which every policy has $U_\beta = -\infty$.

finitely many steps with an optimal one.

Proof. Let $U_k \coloneqq (T^*)^k 0$: by Theorem 3.15, applied to the k-step problem, it is the optimal expected value of the truncated return $\sum_{t\le k\wedge N} R_t$ over Π_{H}, and $(T^*)^k u$ is the same with the terminal reward $u(S_{k+1})$ on $\{N > k\}$.

Under (T). By Proposition 8.29(iii), for every $\sigma \in \Pi_{\mathrm{H}}$, W is integrable and $|\mathbb{E}_s^\sigma[W] - \mathbb{E}_s^\sigma[\sum_{t\le k\wedge N} R_t]| \le R_{\max} \sum_{j\ge k} \mathbb{P}_s^\sigma(N > j) \le R_{\max}\rho^k \|b^\star\|_\infty^2$, with $\rho = \rho_{b^\star}$, uniformly in σ; hence $V^\star = \lim_k U_k$, which is the unique fixed point of the contraction T^* (Proposition 8.29(i)). A greedy π satisfies $T^\pi V^\star = V^\star$, so $V^\pi = V^\star$ by uniqueness of the fixed point of T^π. This gives (i)–(iii); (iv) is Theorem 8.9 verbatim, all policies being proper.

Under (C). Since rewards are ≤ 0, $T^*0 \le 0$ and U_k is nonincreasing; for a proper π_0, $U_k \ge (T^{\pi_0})^k 0 \ge V^{\pi_0} > -\infty$; and $U_k \ge V^\star$, because dropping the rewards after time k can only increase a total of nonpositive rewards. So $U_k \downarrow U_\infty$, finite, with $U_\infty = T^* U_\infty$ by continuity of T^*. Let π be greedy for U_∞. Then $U_\infty = (T^\pi)^k U_\infty = \mathbb{E}_s^\pi[\sum_{t\le k\wedge N} R_t + U_\infty(S_{k+1})\mathbb{1}\{N > k\}] \le -c\,\mathbb{E}_s^\pi[k \wedge N]$, so $\mathbb{E}_s^\pi[N] \le -U_\infty(s)/c < \infty$: π is proper, and letting $k \to \infty$ (dominated convergence, $|\sum_{t\le k\wedge N} R_t| \le R_{\max} N$) gives $U_\infty = V^\pi \le V^\star$. Hence $V^\star = U_\infty = V^\pi$ is a solution. If $U = T^*U$, a greedy π_U is proper by the same argument (otherwise $\mathbb{P}_s^{\pi_U}(N > k) \ge \eta > 0$ for some s and all k, and $U(s) \le -ck\eta + \|U\|_\infty \to -\infty$), so $U = V^{\pi_U} \le V^\star$; and $U \ge (T^\pi)^k U \to V^\pi = V^\star$. This proves (i)–(ii). For (iii), by monotonicity it suffices to take $V_0 = \pm M\mathbf{1}$ with $M \ge 0$. From below, $(T^*)^k(-M\mathbf{1}) \ge (T^\pi)^k(-M\mathbf{1}) \to V^\star$ and $(T^*)^k(-M\mathbf{1}) \le U_k \to V^\star$. From above, let σ_k attain $(T^*)^k(M\mathbf{1})(s)$ and $q_k = \mathbb{P}_s^{\sigma_k}(N > k)$; then $V^\star(s) \le U_k(s) \le (T^*)^k(M\mathbf{1})(s) \le -ckq_k + Mq_k$, so $q_k \le |V^\star(s)|/(ck - M) \to 0$, and $(T^*)^k(M\mathbf{1})(s) \le U_k(s) + Mq_k \to V^\star(s)$. For (iv), if π is proper and π' greedy for V^π, then $V^\pi \le (T^{\pi'})^k V^\pi$, and the bound used for π_U shows that π' is proper; then $V^{\pi'} = \lim_k (T^{\pi'})^k V^\pi \ge V^\pi$, the values increase, and termination at a fixed point of T^* follows as in Theorem 8.9. ∎

Without (C), policy iteration started from an improper policy may fail at the first step: the evaluation system $V = r^\pi + Q^\pi V$ is then singular or has a solution that is not the value (Exercise 8.11). Under the general assumptions of Bertsekas and Tsitsiklis (1991), the conclusions of the theorem hold as well, but T^* is no longer a weighted-norm contraction.

Exponential moments and the critical risk level ♦

For $\beta \in \mathbb{R}$ and a stationary π, let $M_\beta^\pi(s,s') \coloneqq \sum_a \pi(a \mid s) P(s' \mid s,a) e^{\beta r(s,a,s')}$ for $s, s' \in \mathcal{S}^\circ$, the tilted kernel of Theorem 4.15 restricted to the transient states, and $m_\beta^\pi(s) \coloneqq \sum_a \pi(a \mid s) P(\Delta \mid s,a) e^{\beta r(s,a,\Delta)}$. With $\gamma = 1$ the rate does not change along the recursion (Proposition 7.16 with $\gamma = 1$): $Z_\beta(s) \coloneqq \mathbb{E}_s^\pi[e^{\beta W}]$ satisfies $Z_\beta = M_\beta^\pi Z_\beta + m_\beta^\pi$, a single exponential moment is Bellman closed, and the only question is whether it is finite.

Proposition 8.32 (Finiteness of exponential moments). Let π be proper and $\beta \in \mathbb{R}$. Then $\mathbb{E}_s^\pi[e^{\beta W}] < \infty$ for every $s \in \mathcal{S}^\circ$ if and only if $\rho(M_\beta^\pi) < 1$, and then $Z_\beta = (I - M_\beta^\pi)^{-1} m_\beta^\pi$. If the rewards are nonpositive on $\mathcal{S}^\circ$, $\rho(M_\beta^\pi)$ is continuous and nonincreasing in β, so that the set of β with finite exponential moments is an interval $(\beta_c(\pi), \infty)$, $\beta_c(\pi) \in [-\infty, 0)$: the *critical risk level* of π. With a unit cost per step, $W = -N$, $M_\beta^\pi = e^{-\beta} Q^\pi$ and $\beta_c(\pi) = \ln \rho(Q^\pi)$.

Proof. Splitting according to the value of N, $\mathbb{E}_s^\pi[e^{\beta W} \mathbb{1}\{N = n+1\}] = ((M_\beta^\pi)^n m_\beta^\pi)(s)$, a sum over the paths of n steps in $\mathcal{S}^\circ$ followed by a step into Δ, so that $Z_\beta = \sum_{n \geq 0} (M_\beta^\pi)^n m_\beta^\pi$, a series of nonnegative terms. If $\rho(M_\beta^\pi) < 1$ it converges to $(I - M_\beta^\pi)^{-1} m_\beta^\pi$. Conversely, suppose that $z \coloneqq Z_\beta$ is finite; $z > 0$ since π is proper. Then $z = m_\beta^\pi + M_\beta^\pi z$ and, iterating, $(M_\beta^\pi)^n z = z - \sum_{j<n} (M_\beta^\pi)^j m_\beta^\pi \to 0$; since $z > 0$, every column of $(M_\beta^\pi)^n$ tends to 0, hence $\rho(M_\beta^\pi) < 1$. If $r \leq 0$, the entries of M_β^π are continuous and nonincreasing in β, and so is the spectral radius of a nonnegative matrix (it is monotone in the entries, Seneta, 2006, Chap. 1, and continuous); $\rho(M_0^\pi) = \rho(Q^\pi) < 1$ gives $\beta_c(\pi) < 0$. ∎

This is the Perron root of Chapter 9 again. In Theorem 9.17, $\Lambda_\pi(\beta) = \ln \lambda_\pi(\beta)$ was the growth rate of $\mathbb{E}[e^{\beta W_T}]$ for a chain that never stops; here $\ln \rho(M_\beta^\pi)$ is the growth rate of $\mathbb{E}[e^{\beta W_T} \mathbb{1}\{N > T\}]$, the contribution of the episodes still running, and the total exponential moment is finite exactly when this rate is negative. For a risk-averse agent ($\beta < 0$) facing costs, the critical level is a breakdown point: the β-mean, $(1/\beta) \ln Z_\beta$, decreases to $-\infty$ as $\beta \downarrow \beta_c(\pi)$, because the geometric tail $\rho(Q^\pi)^n$ of the episode length eventually beats the exponential penalty $e^{|\beta| n}$.

Risk-sensitive control is stationary ♦

For $\beta \neq 0$ and $\sigma \in \Pi_{\mathrm{H}}$, let $U_\beta^\sigma(s) \coloneqq U_\beta(W)$ under $\mathbb{P}_s^\sigma$, with the convention $U_\beta^\sigma(s) = -\infty$ when $\beta < 0$ and $\mathbb{E}_s^\sigma[e^{\beta W}] = \infty$, and $U_\beta^\star \coloneqq \sup_{\sigma \in \Pi_{\mathrm{H}}} U_\beta^\sigma$. The *risk-sensitive optimality equation*

$$U(s) \;=\; \max_{a \in \mathcal{A}} U_\beta\big(r(s,a,S') + U(S')\big), \qquad S' \sim P(\cdot \mid s,a), \qquad U(\Delta) = 0, \tag{8.9}$$

reads, for $\varphi = e^{\beta U}$, $\varphi(s) = \mathsf{M}_\beta\big(\sum_{s'} P(s' \mid s, \cdot) e^{\beta r(s,\cdot,s')} \varphi(s')\big)$ with $\varphi(\Delta) = 1$, where M_β, applied to a vector indexed by the actions, is its maximum for $\beta > 0$ and its minimum for $\beta < 0$, as in Chapter 12; write $G_\beta \varphi$ for its right-hand side on $\mathcal{S}^\circ$.

Theorem 8.33 (Stationary risk-sensitive optimal policies). Let $\beta \neq 0$.

(i) *The easy sign.* Suppose $\beta r(s,a,s') \leq 0$ for all $s \in \mathcal{S}^\circ$ (a risk-averse agent collecting gains, or a risk-seeking one paying costs), and (T) or (C). Then G_β is a contraction, for $\|\cdot\|_{b^\star}$ under (T) and for $\|\cdot\|_\infty$ with modulus $e^{-|\beta| c}$ under (C); (8.9) has a unique solution $U_\beta^\star$, and every stationary deterministic policy greedy for it is optimal among all history-dependent policies.

(ii) *The hard sign.* Under (C), let $\beta < 0$ and let $\beta_c^\star$ be the smallest critical level $\beta_c(\pi)$ of

Proposition 8.32 over the proper $\pi \in \Pi_{\mathrm{SD}}$. If $\beta > \beta_c^\star$, then $U_\beta^\star$ is finite, it is the unique solution of (8.9) with $U \le 0$, the iteration of G_β from $\varphi \equiv 1$ increases to $e^{\beta U_\beta^\star}$, and every stationary deterministic policy greedy for $U_\beta^\star$ is proper and optimal among all history-dependent policies. If $\beta \le \beta_c^\star$, then for some state s, $U_\beta^\sigma(s) = -\infty$ for every $\sigma \in \Pi_{\mathrm{H}}$.

Proof. (i) In $G_\beta\varphi(s)$ the coefficient of $\varphi(s')$, $s' \in \mathcal{S}^\circ$, is $P(s' \mid s,a)e^{\beta r} \le P(s' \mid s,a)$, and under (C) also $\le e^{-|\beta|c}P(s' \mid s,a)$, so the contraction follows as in Proposition 8.29(i). Let $\varphi^\star$ be the fixed point and π a greedy policy: $\varphi^\star$ is also the fixed point of the affine contraction $\varphi \mapsto M_\beta^\pi\varphi + m_\beta^\pi$, which is $\sum_n (M_\beta^\pi)^n m_\beta^\pi = \mathbb{E}_s^\pi[e^{\beta W}]$ (with $e^{\beta W} = 0$ on $\{N = \infty\}$ when $\beta > 0$). It is positive: under (T) because π is proper, under (C) because $G_\beta \ge G_\beta^{\pi_0}$ for a proper π_0, whence $\varphi^\star \ge Z_\beta^{\pi_0} > 0$. For $\sigma \in \Pi_{\mathrm{H}}$, let $X_t \coloneqq e^{\beta\sum_{j<t}R_j}\varphi^\star(S_t)$, with $\varphi^\star(\Delta) = 1$ (the rewards vanish after absorption). By the definition of G_β, X is a supermartingale if $\beta > 0$ and a submartingale if $\beta < 0$; and $X_{k+1} = e^{\beta W}$ on $\{N \le k\}$, while $0 \le X_{k+1} \le \|\varphi^\star\|_\infty$ on $\{N > k\}$ since $e^{\beta R_j} \le 1$. For $\beta > 0$, $\varphi^\star(s) \ge \mathbb{E}X_{k+1} \ge \mathbb{E}[e^{\beta W}\mathbb{1}\{N \le k\}] \to \mathbb{E}_s^\sigma[e^{\beta W}]$. For $\beta < 0$, $\varphi^\star(s) \le \mathbb{E}[e^{\beta W}\mathbb{1}\{N \le k\}] + \|\varphi^\star\|_\infty \mathbb{P}_s^\sigma(N > k)$, and $\mathbb{P}_s^\sigma(N > k) \to 0$ under (T) by Proposition 8.29(iii) (under (C), $\beta r \le 0$ forces $\beta > 0$). In both cases $U_\beta^\sigma(s) \le \frac{1}{\beta}\ln\varphi^\star(s) = U_\beta^\pi(s)$.

(ii) Here $e^{\beta r} \ge e^{|\beta|c} > 1$ on $\mathcal{S}^\circ$, so $G_\beta\mathbf{1} \ge \mathbf{1}$ and $\varphi_k \coloneqq G_\beta^k\mathbf{1}$ is nondecreasing; by Theorem 5.8 applied to the k-step problem, $\varphi_k(s)$ is the minimum over Π_{H} of $\mathbb{E}[e^{\beta\sum_{t\le k\wedge N}R_t}]$, which is at most $\mathbb{E}_s^\sigma[e^{\beta W}]$ for every σ (adding steps adds costs). Let $\varphi_\infty = \lim_k \varphi_k \in [1,\infty]^{\mathcal{S}^\circ}$, so that $\varphi_\infty \le \mathbb{E}_s^\sigma[e^{\beta W}]$ for all σ. If $\beta > \beta_c^\star$, some $\pi_0 \in \Pi_{\mathrm{SD}}$ has $\rho(M_\beta^{\pi_0}) < 1$, so $\varphi_\infty \le Z_\beta^{\pi_0} < \infty$ by Proposition 8.32, and $\varphi_\infty = G_\beta\varphi_\infty$ by continuity. For π greedy for φ_∞, $\varphi_\infty = (M_\beta^\pi)^k\varphi_\infty + \sum_{j<k}(M_\beta^\pi)^j m_\beta^\pi$. Since $\varphi_\infty \ge \mathbf{1}$ and $M_\beta^\pi \ge e^{|\beta|c}Q^\pi$ entrywise, $(Q^\pi)^k\mathbf{1} \le e^{-|\beta|ck}\|\varphi_\infty\|_\infty \to 0$: π is proper; and $\varphi_\infty \ge \sum_{j<k}(M_\beta^\pi)^j m_\beta^\pi \to Z_\beta^\pi$. Hence $Z_\beta^\pi \le \varphi_\infty \le \mathbb{E}_s^\sigma[e^{\beta W}]$ for all σ: π is optimal and $U_\beta^\star = \frac{1}{\beta}\ln\varphi_\infty$. If $\psi \ge \mathbf{1}$ is another solution, then $\psi = G_\beta^k\psi \ge \varphi_k \to \varphi_\infty$, and $\psi \le M_\beta^\pi\psi + m_\beta^\pi$ gives $\psi \le (M_\beta^\pi)^k\psi + \sum_{j<k}(M_\beta^\pi)^j m_\beta^\pi \to Z_\beta^\pi = \varphi_\infty$, as $\rho(M_\beta^\pi) < 1$. If $\beta \le \beta_c^\star$ and φ_∞ were finite, the same argument would produce a proper $\pi \in \Pi_{\mathrm{SD}}$ with $Z_\beta^\pi \le \varphi_\infty < \infty$, hence $\rho(M_\beta^\pi) < 1$ and $\beta_c^\star \le \beta_c(\pi) < \beta$, a contradiction: so $\varphi_\infty(s) = \infty$ for some s, and $\mathbb{E}_s^\sigma[e^{\beta W}] \ge \varphi_\infty(s) = \infty$ for every σ. ∎

The theorem is the counterpart, and the opposite, of Theorem 8.22 and Exercise 8.6: with $\gamma = 1$, β-means compose without changing the rate, the optimality equation (8.9) is a genuine fixed-point equation, and its greedy policy is stationary. Part (ii) is in the spirit of the monotone arguments of Denardo and Rothblum (1979) and Patek (2001), who treat exponential utilities in transient and terminating models; Su et al. (2025) prove the stationarity of ERM- and EVaR-optimal policies in transient models with rewards of any sign, beyond the contraction regime of (i).

Remark 8.34 (Discounting, not the infinite horizon, breaks stationarity). Read the example of Exercise 8.6 as a killed problem rather than a discounted one: the episode stops with probability 0.2 after each step, and the return is the undiscounted sum K of the rewards collected. The model is transient ($b^\star \equiv 5$), the rewards are nonnegative, and Theorem 8.33(i) applies for every $\beta < 0$: in $\varphi = \min_a \mathbb{E}[e^{\beta r_a}](0.2 + 0.8\varphi)$ the factor $0.2 + 0.8\varphi$ is common to both actions, so the optimal policy plays at every step the action with the smaller $\mathbb{E}[e^{\beta r_a}]$. At $\beta = -2$ this is the sure reward 0.3, for ever, with $U_{-2} = 0.816$ against 0.785 for the random action (`hike_ssp.py`). For the discounted return W of the same problem, the optimal policy plays the sure action first and the random one at the second step (Exercise 8.6(b)). The two criteria have the same expectation (Proposition 7.3) and differ only through the law of the return: an infinite horizon with a random end keeps the risk-sensitive problem stationary, and it is the discount applied to the rewards, which turns the

rate β into $\gamma^{t-1}\beta$, that destroys stationarity. A consequence for the entropic value at risk: in a random-length problem, the reduction of Proposition 6.29 only involves stationary deterministic policies, one per rate β, and the EVaR-optimal policy can be taken among them (Su et al., 2025).

Example 8.35 (Prudent hikers without horizon). In the hike of Example 8.30, the costs are positive and $\beta < 0$: Theorem 8.33(ii) applies. The critical levels of the three paths are $\ln \rho(Q^\pi) = -0.216$ (row 2), -0.365 (row 3) and -0.432 (row 4). Solving (8.9) for a range of β (`hike_ssp.py`, monotone iteration from $\varphi \equiv 1$ finished by policy iteration) gives three regimes (Figure 8.3, right): the risky path along the pit is optimal for $\beta \geq -0.056$, the middle path for $-0.326 \leq \beta \leq -0.056$, and below, the path along the top row. Off that path the optimal actions still change with β; near the critical level, when the wind pushes the hiker down to rows 2 or 3 in column 5, she walks east to the exit column instead of climbing back. That policy has $\rho(Q^\pi) = 0.639$, the smallest of all, and the problem breaks down at $\beta_c^\star = \ln 0.639 = -0.448$: beyond, every policy, stationary or not, has an infinite exponential cost. Each of these optimal policies is stationary, whereas in the finite-horizon hike of Example 5.10 the optimal decision depends on the time left.

Distributional viewpoint

For the expected return, discounting changes nothing essential: the optimal value is the fixed point of a contraction, greedy policies are optimal, stationary and deterministic, and three algorithms compute them. For β-means, discounting changes everything: the rate β applied to the reward-to-go decays like γ^{t-1}, optimal policies are non-stationary, and the fixed-point formulation is lost—but a backward recursion over decreasing rates, closed by a risk-neutral tail, is exact up to $|\beta|\gamma^{2T'}\Delta^2/8$ (Theorem 8.22), and the entropic value at risk follows by a grid of rates (Corollary 8.25). The operator that propagates distributions and selects greedily for the mean is not a contraction: distributional control converges to the distribution of *an* optimal policy, not to a canonical one. Episodes of random length (Section 8.7) show where the trouble comes from: without discount, a β-mean keeps its rate along the recursion, the expected absorption time replaces $1/(1-\gamma)$ in the contraction modulus, and risk-sensitive optimal policies are stationary as long as the exponential moments are finite (Theorem 8.33). It is the discount applied to the rewards, not the infinite horizon, that breaks stationarity.

The fourth property announced with Table 5.2, the contraction of the discounted iteration, sorts the statistics again (Table 8.1). The mean and the support extrema keep a contracting scalar recursion for control as well as for evaluation, the moments for evaluation only; the whole law contracts for the Wasserstein distances only as long as no action is selected; the exponential moment loses its fixed rate; quantiles and CVaR still have no recursion of their own.

Table 8.1. The fourth property: what discounting ($\gamma < 1$) does to each statistic, for the evaluation of a stationary policy $\pi \in \Pi_{\mathrm{SR}}$ and for control.

Statistic	Evaluation	Control
Mean	T^π is a γ-contraction for $\|\cdot\|_\infty$ (Theorem 7.11)	T^* is a γ-contraction for $\|\cdot\|_\infty$ (Proposition 8.2); greedy policies are optimal (Theorem 8.3)
Moments, variance	given the lower moments, the map of $M^{(i)}$ is a γ^i-contraction; $\sigma^2 = (I - \gamma^2 P^\pi)^{-1} c^\pi$ (Proposition 7.13)	no valid greedy optimization, as in finite horizon (Table 5.2)
Whole law	$\mathcal{T}^\pi$ is a γ-contraction for $\overline{\mathcal{W}}_p$, $p \in [1, \infty]$ (Theorem 7.7, Exercise 7.4), not in total variation (Remark 7.8)	$\mathcal{T}^*$ is a contraction for no $\overline{\mathcal{W}}_p$ (Exercise 8.5); its iterates reach the law of *an* optimal policy (Section 8.4)
Exponential moment, β-mean	no fixed-rate recursion: Z_β calls for $Z_{\gamma\beta}$ (Proposition 7.16 and Remark 4.13); the countable family $(Z_{\gamma^k\beta})_{k\geq 0}$ is closed	no fixed-point formulation; optimal policies are non-stationary; backward recursion over the rates $\beta_t = \gamma^{t-1}\beta$ with a risk-neutral tail (Theorem 8.22)
Quantile, CVaR	no closed recursion; the projected operator $\Pi_Q \mathcal{T}^\pi$ is a γ-contraction for $\overline{\mathcal{W}}_\infty$, with a fixed point that only approximates v^π (Proposition 7.17)	no greedy recursion; projected distributional control carries no guarantee (Section 8.4)
Support extrema	the $(\min, +)$ and $(\max, +)$ operators are γ-contractions for $\|\cdot\|_\infty$ (Proposition 7.10)	adding a maximum over actions keeps a γ-contraction (same proof)

Hands-on

Session H04 (`retailStore.ipynb`) implements value iteration with the stopping rule of Algorithm 8.1, policy iteration with a linear solve, and the extraction of greedy policies; the plot of $Q^\star(s, \cdot)$ shows the optimal (s, S) structure of Example 8.14. Session H05 (`replacement.ipynb`) asks first for a discretization of the continuous wear level and a planning algorithm—value or policy iteration on the discretized MDP; since its maintenance cost is not monotone, the optimal policy is not the threshold rule of Example 8.15 but replaces on several disjoint wear ranges—before turning to function approximation and fitted Q-learning (Chapter 13). The risk-sensitive recursion of Algorithm 8.3 is a small modification of value iteration in either notebook: replace the average over demands (or wear increments) by a β_t-mean, computed in the log domain.

Exercises

Exercise 8.1 (Policy iteration). [verify] (a) Show that policy iteration never visits the same policy twice and deduce the bound A^S on the number of iterations. (b) Show that the improvement step of Algorithm 8.2 is the Newton step for the equation $F(V) := V - T^*V = 0$: if π is greedy for V, the Newton iterate from V is V^π (*hint:* on the region where π stays greedy, $F(V) = V - T^\pi V$ is affine). (c) Modified policy iteration replaces the exact evaluation by m applications of T^π started from the previous value; show that it still converges to $V^\star$ when started from V_0 with $T^*V_0 \geq V_0$.

Exercise 8.2 (The dual linear program). [verify] (a) Write the dual of the program of Theorem 8.11 and check that it is the one stated after the theorem. (b) Check directly, without Proposition 8.18, that for every $\pi \in \Pi_{\mathrm{SR}}$ the unnormalized occupation measure $\tilde{d}^\pi(s, a) = \sum_t \gamma^{t-1} \mathbb{P}^\pi_\mu(S_t = s, A_t = a)$ is dual feasible and that its objective is $\sum_s \mu(s) V^\pi(s)$. (c) Deduce from

Proposition 8.17(ii) that an optimal dual solution d yields an optimal policy π_d, and from (iii) that the simplex method returns a deterministic one. Compare with Exercise 3.2.

Exercise 8.3 (A constrained MDP with one state). [limits] One state, two actions: a has reward 1 and cost 1, b has reward 0 and cost 0; the constraint is $C^\pi \le \kappa/(1-\gamma)$ with $\kappa \in (0,1)$. (a) Write the program of Theorem 8.26, show that the optimal value is $\kappa/(1-\gamma)$, attained by the stationary policy that plays a with probability κ, and that no stationary deterministic policy is optimal. (b) A deterministic policy, even history-dependent, is here a fixed set $\mathcal{T} \subset \mathbb{N}^*$ of times at which a is played. Show that for $\gamma \ge \frac{1}{2}$ every $x \in [0, 1/(1-\gamma)]$ is of the form $\sum_{t\in\mathcal{T}} \gamma^{t-1}$ (*hint:* choose the times greedily and show that the remainder after step t is at most $\gamma^t/(1-\gamma)$), so that some deterministic policy is optimal. (c) Show that the policies of (a) and (b) have the same occupation measure but different return laws, so that $\mathbb{P}(W \le u)$ is not a function of the occupation measure. (d) Let $\gamma = \frac{1}{3}$ and $\kappa = \frac{1}{2}$: show that the best deterministic policy has value $\frac{1}{2}$, against $\frac{3}{4}$ for the randomized one. (e) Compute $\min_{\lambda\ge 0}\{\max_\pi V^\pi_{r-\lambda c} + \lambda\kappa/(1-\gamma)\}$ and the minimizing $\lambda^\star$, and check which policies maximize the Lagrangian at $\lambda^\star$.

Exercise 8.4 (Performance loss in Q-form). [verify] For $Q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ let $\pi_Q(s) \in \arg\max_a Q(s,a)$. Show that $\|Q^\star - Q^{\pi_Q}\|_\infty \le \frac{2\gamma}{1-\gamma}\|Q^\star - Q\|_\infty$, and give a two-state example showing that the factor $2\gamma/(1-\gamma)$ in Theorem 8.5 cannot be improved in general (Singh and Yee, 1994).

Exercise 8.5 (The distributional optimality operator is not a contraction). [limits] Consider three states. From state 1, action a leads to state 2 and action b to state 3, both with zero reward; states 2 and 3 are absorbing with a single action, state 2 yielding reward 0 or 2 with probability $\frac{1}{2}$ at each step (encoded by two states as in Example 7.9) and state 3 yielding reward 1 at each step; $\gamma \in (0,1)$. (a) Compute the return distributions $\nu^{(2)}$ and $\nu^{(3)}$ from states 2 and 3 (Example 7.9) and show that $\mathcal{W}_1(\nu^{(2)}, \nu^{(3)}) \ge c_\gamma > 0$ although both have mean $1/(1-\gamma)$. (b) Let ν and ν' coincide with the true return distributions at states 1, 2 and differ at state 3: $\nu(3) = \delta_{(1+\varepsilon)/(1-\gamma)}$, $\nu'(3) = \delta_{(1-\varepsilon)/(1-\gamma)}$. Show that $\overline{\mathcal{W}}_1(\nu,\nu') = 2\varepsilon/(1-\gamma)$ while $\mathcal{T}^*$ selects b at state 1 for ν and a for ν', so that $\mathcal{W}_1((\mathcal{T}^*\nu)(1), (\mathcal{T}^*\nu')(1)) \ge \gamma c_\gamma - O(\varepsilon)$: no contraction inequality can hold. (c) When both actions are optimal, as here, what do the iterates $(\mathcal{T}^*)^k\delta_0$ do under a fixed tie-breaking rule, and under a rule that alternates?

Exercise 8.6 (Non-stationarity of risk-sensitive optimal policies). [limits] Consider one state and two actions whose one-step rewards have laws $\delta_{0.3}$ and $\frac{1}{2}(\delta_0 + \delta_1)$ (random rewards are encoded by splitting the state as in Example 7.9; we keep the lighter description), with $\gamma = 0.8$ and a two-step horizon ($T = 2$ in Theorem 8.22(i)). (a) Show that $U_\beta(\frac{1}{2}(\delta_0+\delta_1)) < 0.3$ iff $\beta < \beta^\star \approx -1.8$. (b) Take $\beta = -2$: show that the optimal policy chooses the sure action at $t = 1$ (rate -2) and the random action at $t = 2$ (rate -1.6). (c) Explain why no stationary policy can be optimal for U_{-2} in the infinite-horizon version of this example. Compare with Remark 8.34, where the same example is read as a killed problem.

Exercise 8.7 (Truncation versus risk-neutral tail). [verify] Suppose the recursion of Algorithm 8.3 is stopped at T' and *any* policy is followed afterwards, with the terminal function $\widehat{U}_{T'+1} = 0$. Show that the loss is at most $2\gamma^{T'}R_{\max}/(1-\gamma)$, and compare with the bound of

Theorem 8.22(ii) for $\gamma = 0.99$, $|\beta| = 0.1$, $R_{\max} = 15$: how large must T' be in each case to guarantee a loss of 1?

Exercise 8.8 (Details of the EVaR corollary). [verify] (a) Check the claim, in the proof of Corollary 8.25, that $\beta \mapsto U^\star_{-\beta}(s)$ is nonincreasing and that Hoeffding's lemma applies at every rate, so that the three-case argument of Theorem 6.31 goes through. (b) Show that the value $b_1 = 8\varepsilon/\Delta^2$ of the first grid point can be replaced by $8\varepsilon/\Delta_\star^2$, where $\Delta_\star$ is the maximal range of the return over all policies, computed from Proposition 7.10 optimized over actions ($(\max, +)$ and $(\min, +)$ recursions), and evaluate the gain on the bike store.

Exercise 8.9 (Control-limit structure, discounted). [verify] In the discounted machine replacement problem (maintenance cost c nondecreasing, replacement cost C, i.i.d. nonnegative increments), work on a finite grid $\mathcal{G} = \{0, \delta, \dots, \bar{x}\}$ of wear levels, as in Example 8.15: the next wear $x + X$ is capped at $\bar{x}$ and rounded to $\mathcal{G}$ by a nondecreasing map, so that the finite-state results of this chapter apply (the continuous model, with $c(x) = x$ unbounded, would need the theory of general state spaces, not developed here). Show that $V^\star$ is nonincreasing in x, that the set $\mathcal{X}^{\text{rep}} = \{x \in \mathcal{G} : Q^\star(x, \text{keep}) \le Q^\star(x, \text{replace})\}$ on which replacing is optimal is an upper set of $\mathcal{G}$, $\{x \in \mathcal{G} : x \ge x^\star\}$ with $x^\star = \min \mathcal{X}^{\text{rep}}$ (the open case $(x^\star, \infty)$ of Theorem 3.22 needs a continuum of wear levels), and that the stationary policy "replace if and only if the wear lies in $\mathcal{X}^{\text{rep}}$" is optimal (adapt the proof of Theorem 3.22). Let $x_k^\star$ be the corresponding thresholds of the discounted problems with k steps to go, those of $(T^*)^k 0$. Show that $x_k^\star = x^\star$ for all large k if the limiting comparison $D(x) = Q^\star(x, \text{keep}) - Q^\star(x, \text{replace})$ does not vanish on $\mathcal{G}$, and that $x_k^\star \in \{x^\star, x^\star + \delta\}$ eventually if it vanishes at one grid point only: convergence up to one grid step (*hint:* D is nonincreasing, $(T^*)^k 0 \to V^\star$ uniformly, and consider the smallest nonzero value of $|D|$ on $\mathcal{G}$). Explain why, if D vanishes on the grid points of a whole interval $[a, b]$, the uniform convergence of the values only confines the limit points of $x_k^\star$ to $[a, b + \delta]$, and the action choices converge only away from it. Check numerically the value $x^\star \approx 2.1$ of Example 8.15.

Exercise 8.10 (Numerical study: the bike store). [implement] Reproduce Figure 8.1 and the (s, S) policy of Example 8.14. Vary the fixed ordering cost $C_0 \in \{0, 0.8, 3\}$ and the storage cost λ (notation of Example 1.3), and describe how the reorder point and the order-up-to level move. The risk-neutral reference is exact: policy iteration terminates with an optimal policy of the finite MDP. Then run Algorithm 8.3 for $\beta = -0.1$ (its values carry the certified error of Theorem 8.22(ii)) and report the order-up-to levels at $t = 1$ and at $t = 100$: is the policy stationary? Explain, from the decay of the effective rate $\beta\gamma^{t-1}$, what happens for $\beta = -0.2$ and for $\beta = -0.05$ without rerunning the algorithm.

Exercise 8.11 (A geometric episode). [limits] One nonterminal state s, reward -1 per step; the action a ends the episode with probability $1 - q$ and returns to s with probability $q \in (0, 1)$. (a) Show that N is geometric, that $\mathbb{E}[e^{\theta N}] < \infty$ if and only if $\theta < \ln(1/q)$, and compute $U_\beta(s)$ for $\beta > \ln q$; check that $\beta_c = \ln \rho(Q) = \ln q$ (Proposition 8.32) and that $U_\beta(s) \to -\infty$ as $\beta \downarrow \ln q$. (b) Let $v(s) = \delta_0$ and $v'(s) = \delta_1$. Show that $\mathcal{W}_\infty((\mathcal{T}^\pi)^k v(s), (\mathcal{T}^\pi)^k v'(s)) = 1$ for every k, while $\mathcal{W}_1$ decreases like q^k: the contraction of Proposition 8.29(ii) has no $\mathcal{W}_\infty$ version. (c) Add an action b that returns to s with probability 1 and reward -1, so that (C) still holds. Show that the evaluation system $V = r^\pi + Q^\pi V$ of "always b" has no solution, and explain why policy iteration must start

from a proper policy. With reward 0 for b instead, (C) fails: show that the system has infinitely many solutions and that T^* has the fixed points $V(s) = v$ for every $v \geq -1/(1-q)$, so that the optimality equation no longer characterizes $V^\star$.

Exercise 8.12 (A tennis game with deuce). [verify] The server wins each point independently with probability p, $q = 1 - p$. A game is won by the first player to reach four points with a lead of two; from 3–3 (deuce), a player who wins one point has the advantage, and wins the game if she also wins the next point, while losing it returns to deuce. (a) Model the game as a transient model without control (scores (i, j), $0 \leq i, j \leq 3$, two advantage states, one terminal state reached by either player's win, reward 1 on the transitions that win the game for the server) and check (T). (b) Show that the value from $(0, 0)$, the probability that the server wins, is $p^4(1 + 4q + 10q^2) + 20p^3q^3\, p^2/(p^2 + q^2)$. (c) Show that the number N of points satisfies $\mathbb{P}(N = 4) = p^4 + q^4$, $\mathbb{P}(N = 5) = 4pq(p^3 + q^3)$, $\mathbb{P}(N = 6) = 10p^2q^2(p^2 + q^2)$, $\mathbb{P}(N = 6 + 2k) = 20p^3q^3(2pq)^{k-1}(p^2 + q^2)$ for $k \geq 1$, and $\mathbb{P}(N \text{ odd}, N > 5) = 0$; deduce $\mathbb{E}N$, and compare with $(I - Q)^{-1}\mathbf{1}$ of Lemma 8.28. (d) Show that $\rho(Q) = \sqrt{2pq}$ and that $\mathbb{E}[e^{\theta N}] < \infty$ iff $\theta < \frac{1}{2}\ln(1/(2pq))$. Which value of p makes the games longest, in mean and in tail? (e) For $p = 0.6$, check the values 0.736 (winning probability), 6.48 ($\mathbb{E}N$), 2.59 (standard deviation of N, by the moment recursion with $\gamma = 1$) and 0.367 (critical rate) with `tennis_deuce.py`.

Bibliographical notes

Value iteration is Bellman's method of successive approximations (Bellman, 1957b), policy iteration is due to Howard (1960), and the linear programming formulation to Manne (1960) and d'Epenoux (1963); the occupation polytope, its vertices and the state–action frequencies of history-dependent policies are treated in Puterman (1994, Sec. 6.9), and the successor representation of Definition 8.20 is due to Dayan (1993). The existence of stationary deterministic optimal policies among history-dependent ones is due to Blackwell (1962). Puterman (1994, Chap. 6) gives a complete account, including modified policy iteration, the comparison of policy and value iteration and the interpretation of policy iteration as Newton's method (Puterman and Brumelle, 1979); Ye (2011) proves the strong polynomiality, for a fixed discount factor, of policy iteration and of the simplex method on the dual program, which is a policy iteration switching one state per step (Remark 8.13), Hansen et al. (2013) improve the bound and extend it to turn-based games, and the elimination argument of Theorem 8.10 is that of Scherrer (2016). The span bounds and the extrapolation of Proposition 8.7 are due to MacQueen (1966) and Porteus (1971); Puterman (1994, Secs. 6.6–6.7) treats them in detail, with their use to eliminate suboptimal actions. Constrained MDPs, with the occupation-measure program and the bound on the number of randomizations, are the subject of Altman (1999). The performance-loss bound of Theorem 8.5 appears in Singh and Yee (1994) and is the basic tool of approximate dynamic programming (Bertsekas and Tsitsiklis, 1996).

Exponential utilities in discounted MDPs were studied by Howard and Matheson (1972) and Chung and Sobel (1987), who introduced the time-dependent rates and noticed the loss of stationarity; the risk-neutral tail with its $\gamma^{2T'}$ bound and the extension to the entropic value at risk are due to Hau et al. (2023), whose numerical experiments include the machine replacement and inventory problems of this chapter. Nested risk measures in discounted MDPs are developed by Ruszczyński (2010). The non-contractivity of the distributional optimality operator and the convergence of its iterates under uniqueness of the optimal policy are in Bellemare et al. (2017) and Bellemare et al. (2023, Chap. 7).

Stochastic shortest paths are analysed by Bertsekas and Tsitsiklis (1991); the weighted-norm contraction of Proposition 8.29 for transient models is in Bertsekas and Tsitsiklis (1996, Sect. 2.2).

Exponential utilities in transient and terminating models were studied by Denardo and Rothblum (1979) and Patek (2001); Su et al. (2025) prove that ERM- and EVaR-optimal policies are stationary in transient total-reward MDPs with rewards of any sign, and contrast this with the discounted case.