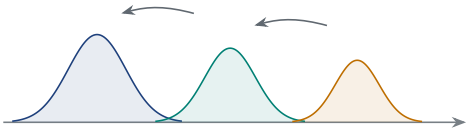
3

Finite Horizon: the Distributional Bellman Equation



Under a Markov policy, the law of the reward-to-go from a state is a mixture of shifted copies of the laws of the reward-to-go from the next states. This identity between probability distributions—the distributional Bellman equation—is the backbone of the book. Taking expectations on both sides gives the classical Bellman equation and the backward induction algorithm for policy evaluation. Optimizing over actions at each step gives Bellman's principle of optimality: an optimal policy exists, is deterministic and Markov, and is computed by a recursion over time. We apply these results to optimal stopping and the secretary problem, to a linear-quadratic control problem, and to the machine replacement problem, where the recursion reveals the threshold structure of the optimal policy.

3.1 Return distributions under a Markov policy

Throughout the chapter, $\mathcal{M} = (\mathcal{S}, \mathcal{A}, (P_t), (r_t), H)$ is a finite-horizon MDP and $\pi \in \Pi_{\mathrm{MR}}$ is a Markov policy. Recall from Section 1.5 that $\mathbb{P}^{\pi}_{t,s}$ denotes the law of the process started at time t in state s, and $\mathbb{P}^{\pi}_{t,s,a}$ the same with the first action imposed. The Markov property (Proposition 1.15) makes the following definition natural.

Definition 3.1 (Return-distribution functions). For $t \in [H]$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$, the *return distributions* of π at time t are

$$\nu^{\pi}_t(s) \coloneqq \mathcal{L}_{\mathbb{P}^{\pi}_{t,s}}\big(W_{t:H}\big), \qquad \nu^{\pi}_t(s,a) \coloneqq \mathcal{L}_{\mathbb{P}^{\pi}_{t,s,a}}\big(W_{t:H}\big),$$

with the terminal convention $\nu^{\pi}_{H+1}(s) = \delta_0$. The collection $\nu^{\pi}_t = (\nu^{\pi}_t(s))_{s\in\mathcal{S}} \in \mathcal{M}_1(\mathbb{R})^{\mathcal{S}}$ is the *return-distribution function* at time t. By Proposition 1.15, $\nu^{\pi}_t(s)$ is also the conditional law of $W_{t:H}$ given $H_t = h_t$ under $\mathbb{P}^{\pi}_{\mu}$, for any history h_t ending in s and of positive probability; and $\nu^{\pi}_1(s) = v^{\pi}(s)$ is the return distribution of Definition 1.17.

The key structural fact is that these distributions are related by a recursion over time. We write $b_r(x) = r + x$ for the shift by r, and $(f)_{\#}\nu$ for the push-forward of ν by a map f—the law of $f(X)$ when $X \sim \nu$—so that $(b_r)_{\#}\nu$ is the law of $r + X$ when $X \sim \nu$ (the notation is collected in Appendix E).

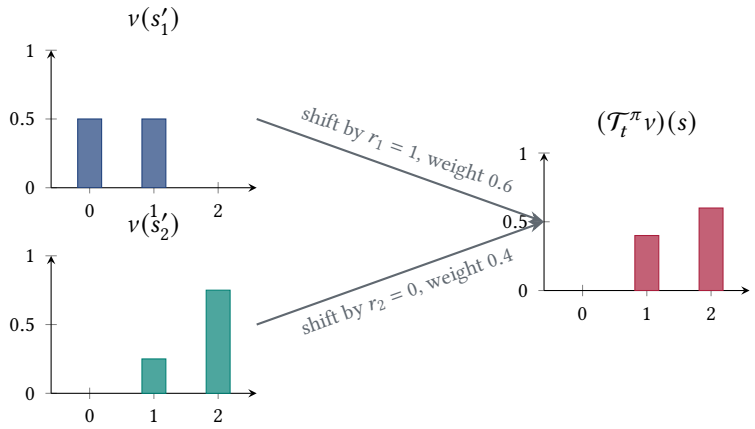


Figure 3.1. The distributional Bellman operator as a mixture of shifts: from state s, the transitions to s'_1 (probability 0.6, reward 1) and s'_2 (probability 0.4, reward 0) shift the next-state return distributions and mix them.

Theorem 3.2 (Distributional Bellman equation). For every $t \in [H]$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$\nu_t^\pi(s,a) = \sum_{s' \in \mathcal{S}} P_t(s' \mid s,a)\ (b_{r_t(s,a,s')})_\# \nu_{t+1}^\pi(s'), \tag{3.1}$$

$$\nu_t^\pi(s) = \sum_{a \in \mathcal{A}} \pi_t(a \mid s)\ \nu_t^\pi(s,a). \tag{3.2}$$

Proof. Under $\mathbb{P}^\pi_{t,s,a}$ we have $W_{t:H} = r_t(s,a,S_{t+1}) + W_{t+1:H}$. Let $g : \mathbb{R} \to \mathbb{R}$ be bounded. Conditioning on S_{t+1},

$$\mathbb{E}^\pi_{t,s,a}\big[g(W_{t:H})\big] = \sum_{s'} P_t(s' \mid s,a)\ \mathbb{E}^\pi_{t,s,a}\big[g\big(r_t(s,a,s') + W_{t+1:H}\big)\,\big|\,S_{t+1} = s'\big].$$

By the restart property of Proposition 1.15, applied to the process started at time t, the conditional law of $(S_{t+1}, A_{t+1}, \dots, S_{H+1})$ given $S_{t+1} = s'$ is $\mathbb{P}^\pi_{t+1,s'}$, so that the conditional law of $W_{t+1:H}$ is $\nu^\pi_{t+1}(s')$ and the inner expectation equals $\int g(r_t(s,a,s') + x)\ \nu^\pi_{t+1}(s')(\,\mathrm{d}x) = \int g\,\mathrm{d}\big((b_{r_t(s,a,s')})_\# \nu^\pi_{t+1}(s')\big)$. As this holds for every bounded g, (3.1) follows. For (3.2), condition on A_t under $\mathbb{P}^\pi_{t,s}$: given $A_t = a$, the process has law $\mathbb{P}^\pi_{t,s,a}$. ∎

In words: the law of the reward-to-go from (t,s) is obtained by drawing an action a from $\pi_t(\cdot \mid s)$ and a next state s' from $P_t(\cdot \mid s,a)$, and then shifting the law of the reward-to-go from $(t+1,s')$ by the reward $r_t(s,a,s')$. The right-hand side of (3.1)–(3.2) defines an operator on return-distribution functions.

Definition 3.3 (Distributional Bellman operator). The *distributional Bellman operator* of π at time t is the map $\mathcal{T}_t^\pi : \mathcal{M}_1(\mathbb{R})^{\mathcal{S}} \to \mathcal{M}_1(\mathbb{R})^{\mathcal{S}}$ defined by

$$(\mathcal{T}_t^\pi \nu)(s) \coloneqq \sum_{a \in \mathcal{A}} \pi_t(a \mid s) \sum_{s' \in \mathcal{S}} P_t(s' \mid s,a)\ (b_{r_t(s,a,s')})_\# \nu(s').$$

Theorem 3.2 states that $\nu_t^\pi = \mathcal{T}_t^\pi \nu_{t+1}^\pi$ for all $t \in [H]$, hence $\nu_1^\pi = \mathcal{T}_1^\pi \mathcal{T}_2^\pi \cdots \mathcal{T}_H^\pi\, \delta_0$, where δ_0 denotes the constant function $s \mapsto \delta_0$.

Two properties of $\mathcal{T}_t^\pi$ follow immediately from its form, a mixture of shifts (Figure 3.1). They will be at the heart of Chapter 4.

Proposition 3.4 (Structure of the distributional operator). For every t, the operator $\mathcal{T}_t^\pi$ is

(i) *mixture-affine*: $\mathcal{T}_t^\pi\big(\lambda\nu+(1-\lambda)\nu'\big) = \lambda\,\mathcal{T}_t^\pi\nu+(1-\lambda)\,\mathcal{T}_t^\pi\nu'$ for $\lambda\in[0,1]$ and $\nu,\nu'\in\mathcal{M}_1(\mathbb{R})^{\mathcal{S}}$ (mixtures taken state by state);

(ii) *monotone for the stochastic order* $\leq_{\mathrm{st}}$, where $\nu\leq_{\mathrm{st}}\nu'$ means $F_\nu\geq F_{\nu'}$ pointwise (Proposition D.15 in Appendix D gives the equivalent forms): if $\nu(s)\leq_{\mathrm{st}}\nu'(s)$ for every s, then $(\mathcal{T}_t^\pi\nu)(s)\leq_{\mathrm{st}}(\mathcal{T}_t^\pi\nu')(s)$ for every s.

Proof. The push-forward $\nu\mapsto(b_r)_\#\nu$ is the restriction to probability measures of a linear map on signed measures, and so is a finite mixture of such maps; this gives (i). For (ii), shifts preserve the order since $F_{(b_r)_\#\nu}(x)=F_\nu(x-r)$, and so do mixtures with common weights, since the distribution function of a mixture is the mixture of the distribution functions. ∎

Example 3.5 (A two-state chain). Consider two states G (good) and B (bad), a single action (omitted from the notation, as is the superscript π; $r(s,s')$ denotes the reward of the transition $s\to s'$), and the stationary kernel and rewards

$$P(\mathrm{G}\mid\mathrm{G})=\tfrac{2}{3},\ r(\mathrm{G},\mathrm{G})=1;\qquad P(\mathrm{B}\mid\mathrm{G})=\tfrac{1}{3},\ r(\mathrm{G},\mathrm{B})=0;$$
$$P(\mathrm{G}\mid\mathrm{B})=\tfrac{1}{2},\ r(\mathrm{B},\mathrm{G})=0;\qquad P(\mathrm{B}\mid\mathrm{B})=\tfrac{1}{2},\ r(\mathrm{B},\mathrm{B})=-1.$$

The recursion of Theorem 3.2 gives $\nu_H(\mathrm{G})=\frac{2}{3}\delta_1+\frac{1}{3}\delta_0$ and $\nu_H(\mathrm{B})=\frac{1}{2}\delta_0+\frac{1}{2}\delta_{-1}$, then

$$\nu_{H-1}(\mathrm{G})=\tfrac{2}{3}\,(b_1)_\#\nu_H(\mathrm{G})+\tfrac{1}{3}\,(b_0)_\#\nu_H(\mathrm{B})=\tfrac{4}{9}\delta_2+\tfrac{2}{9}\delta_1+\tfrac{1}{6}\delta_0+\tfrac{1}{6}\delta_{-1},$$

and similarly $\nu_{H-1}(\mathrm{B})=\frac{1}{2}\nu_H(\mathrm{G})+\frac{1}{2}(b_{-1})_\#\nu_H(\mathrm{B})=\frac{1}{3}\delta_1+\frac{1}{6}\delta_0+\frac{1}{4}\delta_{-1}+\frac{1}{4}\delta_{-2}$. The means are $V_H(\mathrm{G})=\frac{2}{3}$, $V_H(\mathrm{B})=-\frac{1}{2}$, $V_{H-1}(\mathrm{G})=\frac{17}{18}$, and indeed $\frac{2}{3}(1+\frac{2}{3})+\frac{1}{3}(0-\frac{1}{2})=\frac{17}{18}$: the mean of the mixture of shifted laws is the average of the shifted means. Since the rewards lie in $\{-1,0,1\}$, each step can extend the support by at most one unit on each side: after k steps the support of $\nu_{H-k}(\mathrm{G})$ is contained in $\{-k-1,\dots,k+1\}$.

The example shows that the distributional recursion is not only a theoretical statement: it is an algorithm, as soon as return distributions can be represented exactly. Here is a less trivial instance, which also illustrates what the mean does not tell.

Example 3.6 (The windy hike: time spent at the bar). Consider the hands-on version of Example 1.4: a hiker collects a reward of 1 at every step she spends at the bar, $r(s,a,s')=\mathbb{1}\{s'=\mathsf{E}\}$, the exit being absorbing. The return W_H is the number of steps spent at the bar before the horizon and takes values in $\{0,\dots,H\}$; it is 0 if the hiker steps into the pit or arrives too late. Take a 4×6 grid, $p=0.1$, $H=14$, and two deterministic policies: the *safe* policy climbs to the top row, walks east, and comes down along the last column; the *risky* policy walks east along the row just above the pit. Figure 3.2 shows their return distributions from S, computed exactly by Algorithm 3.1 (the rewards are integers). The risky policy has twice the expected return (5.02 against 2.43) because it arrives earlier when it arrives at all; but its return is zero with probability 0.34, against 0.22 for the safe policy—and for different reasons: the risky hiker falls into the pit with probability 0.34, the safe one with probability 0.005 only, but she often arrives too late to enjoy the bar. The risky return is also far more dispersed (standard deviation 3.7 against 1.6). The mean ranks the two policies and sides with the child; the distributions explain why the mother may nevertheless prefer the safe path.

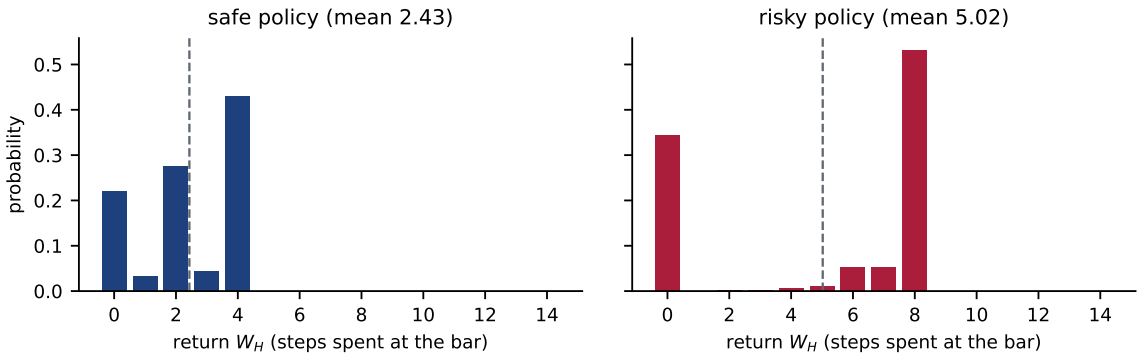


Figure 3.2. Return distributions of the safe and risky policies in the windy hike of Example 3.6 (4×6 grid, $p = 0.1$, $H = 14$), computed by Algorithm 3.1. Dashed lines mark the means.

Algorithm 3.1: Distributional backward induction (finitely supported representation)

Input: MDP $\mathcal{M}$, Markov policy π

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

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

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

    $\nu_t(s) \leftarrow \sum_a \pi_t(a \mid s) \sum_{s'} P_t(s' \mid s, a)\, (b_{r_t(s,a,s')})_{\#}\nu_{t+1}(s')$ ▷ *mixture of shifted atoms*

Output: $(\nu_t)_{t \in [H]}$; in particular $v^\pi = \nu_1$

Proposition 3.7 (Exact distributional evaluation with integer rewards). Suppose that all rewards $r_t(s, a, s')$ are integers in $\{-R_{\max}, \dots, R_{\max}\}$. Then for every t and s, $\nu_t^\pi(s)$ is supported in $\{-kR_{\max}, \dots, kR_{\max}\}$ with $k = H - t + 1$ (in $\{0, \dots, kR_{\max}\}$ when the rewards are nonnegative), and Algorithm 3.1 computes all the return-distribution functions $\nu_1^\pi, \dots, \nu_H^\pi$ exactly in $O(H^2 S^2 A R_{\max})$ arithmetic operations.

Proof. The support statement follows from Theorem 3.2 by backward induction, since shifting by an integer in $\{-R_{\max}, \dots, R_{\max}\}$ and mixing preserve the claimed inclusion. Each $\nu_t(s)$ is represented by a vector of length at most $2HR_{\max} + 1$; computing $(\mathcal{T}_t^\pi \nu)(s)$ requires, for each of the SA pairs (a, s'), shifting such a vector and adding it with a weight, which costs $O(HR_{\max})$ operations. There are S states and H time steps. ∎

Remark 3.8 (Absorbing states and episodes of random length). Many problems end before the horizon: the hike ends in the pit or at the bar, the photo booth session ends when a shot is kept, the secretary problem ends with a hiring. Our formalism handles this with *absorbing* states, from which the process never leaves and collects a reward that does not depend on the action (zero, or a constant such as the "time at the bar"). The horizon H is then only an upper bound on the effective length of an episode, and the return distribution automatically integrates the randomness of this length. When no natural bound exists, one either discounts (Chapter 7) or lets the episode run until absorption, with neither horizon nor discount: this is the *stochastic shortest path* setting of Section 8.6.

Remark 3.9 (The cost of exactness). With general real-valued rewards, distinct trajectories typically yield distinct returns and the number of atoms of $\nu_t^\pi(s)$ may be as large as $(SA)^{H-t+1}$

Algorithm 3.2: Backward induction for policy evaluation

Input: MDP $\mathcal{M}$, Markov policy π

$V_{H+1}(s) \leftarrow 0$ for all s

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

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

    $V_t(s) \leftarrow \sum_a \pi_t(a \mid s) \sum_{s'} P_t(s' \mid s, a)\big[r_t(s, a, s') + V_{t+1}(s')\big]$

Output: $(V_t)_{t \in [H]}$

(Exercise 3.9). Algorithm 3.1 remains correct but its cost explodes. Two ways out are explored later: propagate only a finite family of *statistics* of the distributions, which is possible for very specific families (Chapter 4); or propagate *approximate* representations with a fixed number of atoms, at the price of a projection at each step (Chapters 12 and 13).

3.2 Expected values and backward induction

The expectation is a linear functional of the return distribution; applying it to Theorem 3.2 yields the classical Bellman equations.

Definition 3.10 (Value functions). The *value function* and the *state-action value function* of π at time t are

$$V_t^\pi(s) \coloneqq \mathbb{E}\big[\nu_t^\pi(s)\big] = \mathbb{E}_{t,s}^\pi\big[W_{t:H}\big], \qquad Q_t^\pi(s,a) \coloneqq \mathbb{E}\big[\nu_t^\pi(s,a)\big] = \mathbb{E}_{t,s,a}^\pi\big[W_{t:H}\big],$$

with $V_{H+1}^\pi \equiv 0$; and $V^\pi = V_1^\pi$ is the value of π (Definition 1.17).

Corollary 3.11 (Bellman evaluation equations). For every $t \in [H]$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$Q_t^\pi(s,a) = \sum_{s'} P_t(s' \mid s, a)\big[r_t(s,a,s') + V_{t+1}^\pi(s')\big], \qquad V_t^\pi(s) = \sum_a \pi_t(a \mid s)\, Q_t^\pi(s,a).$$

In matrix form, with $P_t^{\pi_t}$ as in Proposition 1.15 and $r_t^{\pi_t}(s) \coloneqq \sum_a \pi_t(a \mid s) r_t(s,a)$,

$$V_t^\pi \;=\; r_t^{\pi_t} + P_t^{\pi_t} V_{t+1}^\pi.$$

Proof. Apply the linear functional $\nu \mapsto \int x\, \nu(\mathrm{d}x)$ to both sides of (3.1) and (3.2), using $\int x\, (b_r)_\# \nu(\mathrm{d}x) = \int (r + x)\, \nu(\mathrm{d}x) = r + \mathbb{E}[\nu]$. ∎

Algorithm 3.2 computes V^π exactly, for all initial states at once, in $O(HS^2A)$ operations, against $O(nH)$ simulator calls for a Monte Carlo estimate of $V^\pi(s)$ for a single s, with an error of order $HR_{\max}/\sqrt{n}$. Backward induction wins whenever the model is known and S^2A is affordable; Monte Carlo remains the method of choice for large or unknown models.

The linearity used in the proof of Corollary 3.11 is available for many other quantities of interest, which are expectations of additive functionals of the trajectory—and therefore value functions for modified rewards—or probabilities of events with a product structure.

Proposition 3.12 (Some computable functionals of the trajectory). Let $B \subset \mathcal{S}$ and $\pi \in \Pi_{\mathrm{MR}}$.

For $t \in [H+1]$ and $s \in \mathcal{S}$ define

$$N_t(s) \coloneqq \mathbb{E}^{\pi}_{t,s}\Big[\sum_{k=t+1}^{H+1} \mathbb{1}\{S_k \in B\}\Big], \qquad q_t(s) \coloneqq \mathbb{P}^{\pi}_{t,s}\big(S_k \notin B \text{ for all } k = t+1, \dots, H+1\big),$$

the expected number of visits to B after time t, and the probability of avoiding B after time t. Then $N_{H+1} = 0$, $q_{H+1} = 1$, and for $t \le H$,

$$N_t(s) = \sum_{s'} P_t^{\pi_t}(s, s')\big[\mathbb{1}\{s' \in B\} + N_{t+1}(s')\big], \qquad q_t(s) = \sum_{s' \notin B} P_t^{\pi_t}(s, s')\, q_{t+1}(s').$$

The probability of reaching B after time t is $1 - q_t(s)$.

Proof. N_t is the value function of π for the reward $\tilde{r}_t(s, a, s') = \mathbb{1}\{s' \in B\}$, and the first recursion is Corollary 3.11. For the second, condition on S_{t+1}: the event requires $S_{t+1} \notin B$, and given $S_{t+1} = s' \notin B$ the remaining constraint has probability $q_{t+1}(s')$ by the restart property. ∎

3.3 Optimal planning

We now turn to the planning problem: find a policy maximizing the expected return. The set of candidates is the whole of Π_{H}—history-dependent, randomized policies. It will turn out that deterministic Markov policies suffice.

For a general policy $\pi \in \Pi_{\mathrm{H}}$ and a history $h_t \in \mathcal{H}_t$, define the *value of the history* $V_t^{\pi}(h_t)$ as the expectation of $W_{t:H}$ under the law of the continuation of h_t by π, that is, under the probability on continuations $(a_t, s_{t+1}, \dots, s_{H+1})$ given by $\prod_{k=t}^{H} \pi_k(a_k \mid h_k) P_k(s_{k+1} \mid s_k, a_k)$. By Theorem 1.14, $V_t^{\pi}(h_t) = \mathbb{E}^{\pi}_{\mu}[W_{t:H} \mid H_t = h_t]$ whenever h_t has positive probability; and $V_1^{\pi}(s) = V^{\pi}(s)$. For a Markov policy, $V_t^{\pi}(h_t)$ depends on h_t only through s_t and coincides with the $V_t^{\pi}(s_t)$ of Definition 3.10, by Proposition 1.15. The same partial-summation argument as in the proof of Theorem 1.14 gives the recursion

$$V_t^{\pi}(h_t) = \sum_{a} \pi_t(a \mid h_t) \sum_{s'} P_t(s' \mid s_t, a)\big[r_t(s_t, a, s') + V_{t+1}^{\pi}(h_t, a, s')\big], \qquad V_{H+1}^{\pi} \equiv 0. \tag{3.3}$$

Definition 3.13 (Optimality). A policy $\pi^\star \in \Pi_{\mathrm{H}}$ is *optimal* if $V^{\pi^\star}(s) \ge V^{\pi}(s)$ for every $s \in \mathcal{S}$ and every $\pi \in \Pi_{\mathrm{H}}$. The *optimal value function* is $V^\star(s) \coloneqq \sup_{\pi \in \Pi_{\mathrm{H}}} V^{\pi}(s)$.

It is not obvious that an optimal policy exists: a policy that is best from one initial state might be poor from another. The following construction settles the question.

Definition 3.14 (Optimal value functions and greedy policies). Define recursively $V_{H+1}^\star \equiv 0$ and, for $t = H, \dots, 1$,

$$Q_t^\star(s, a) \coloneqq \sum_{s'} P_t(s' \mid s, a)\big[r_t(s, a, s') + V_{t+1}^\star(s')\big], \qquad V_t^\star(s) \coloneqq \max_{a \in \mathcal{A}} Q_t^\star(s, a).$$

A deterministic Markov policy π is *greedy* with respect to $(Q_t^\star)$ if $\pi_t(s) \in \arg\max_a Q_t^\star(s, a)$ for all t and s.

Algorithm 3.3: Backward induction for optimal planning

Input: MDP $\mathcal{M}$

$V_{H+1}(s) \leftarrow 0$ for all s
for $t = H, \dots, 1$ **do**
  for $s \in \mathcal{S}$ **do**
    for $a \in \mathcal{A}$ **do**
      $Q_t(s,a) \leftarrow \sum_{s'} P_t(s' \mid s,a)\big[r_t(s,a,s') + V_{t+1}(s')\big]$
    $V_t(s) \leftarrow \max_a Q_t(s,a)$; $\quad \pi_t^\star(s) \leftarrow$ any $a \in \arg\max_a Q_t(s,a)$

Output: optimal policy $\pi^\star$, optimal values (V_t)

Theorem 3.15 (Bellman's principle of optimality).

(i) For every policy $\pi \in \Pi_{\mathrm{H}}$, every $t \in [H+1]$ and every history $h_t \in \mathcal{H}_t$, $V_t^\pi(h_t) \le V_t^\star(s_t)$.

(ii) Every greedy policy $\pi^\star$ satisfies $V_t^{\pi^\star} = V_t^\star$ for all t. In particular $\pi^\star$ is optimal, $V^\star = V_1^\star$, and $\max_{\pi \in \Pi_{\mathrm{H}}} V^\pi(s) = \max_{\pi \in \Pi_{\mathrm{MD}}} V^\pi(s)$ for every s.

Proof. (i) By backward induction on t. For $t = H+1$ both sides vanish. Assume the claim for $t+1$. By (3.3) and the induction hypothesis applied to the history (h_t, a, s'), whose last state is s',

$$\begin{aligned} V_t^\pi(h_t) &\le \sum_a \pi_t(a \mid h_t) \sum_{s'} P_t(s' \mid s_t, a)\big[r_t(s_t,a,s') + V_{t+1}^\star(s')\big] \\ &= \sum_a \pi_t(a \mid h_t)\, Q_t^\star(s_t,a) \le \max_a Q_t^\star(s_t,a) = V_t^\star(s_t). \end{aligned}$$

(ii) For a greedy $\pi^\star$, both inequalities above are equalities: by induction $V_{t+1}^{\pi^\star} = V_{t+1}^\star$, and $\pi_t^\star(s_t)$ attains the maximum. Hence $V_t^{\pi^\star} = V_t^\star$; with (i) at $t = 1$ this shows that $\pi^\star$ is optimal and that $V_1^\star = V^\star$. The last claim holds because $\pi^\star \in \Pi_{\mathrm{MD}}$. ∎

Algorithm 3.3 computes an optimal policy in $O(HS^2A)$ operations, to be compared with the A^{SH} policies of Π_{MD}. Three remarks on the theorem.

Remark 3.16 (What the proof uses).

(i) *Neither memory nor randomization helps.* The bound in (i) holds for every history-dependent randomized policy, and is attained by a deterministic Markov one. This is specific to the expected-return criterion; for other functionals of the return distribution, optimal policies may need memory or randomization (Chapter 5).

(ii) *The proof uses two properties of the expectation only:* it is *linear* in the return distribution, which is what makes the recursion (3.3) possible, and *monotone in the continuation values*: the operator $V \mapsto \sum_{s'} P_t(s' \mid s,a)[r_t(s,a,s') + V(s')]$ is nondecreasing for the pointwise order, which is what allows the replacement of V_{t+1}^π by the larger $V_{t+1}^\star$ inside the sum. In distributional terms, this monotonicity is that of the expectation for the stochastic order—$\mathbb{E}[\nu] \le \mathbb{E}[\nu']$ whenever $\nu \le_{\mathrm{st}} \nu'$—combined with Proposition 3.4(ii); this is the form in which Chapters 4 and 5, devoted to identifying the functionals that share these two properties, will need it.

(iii) *Optimal policies are not unique*, and different optimal policies may have very different return distributions: see Example 3.17.

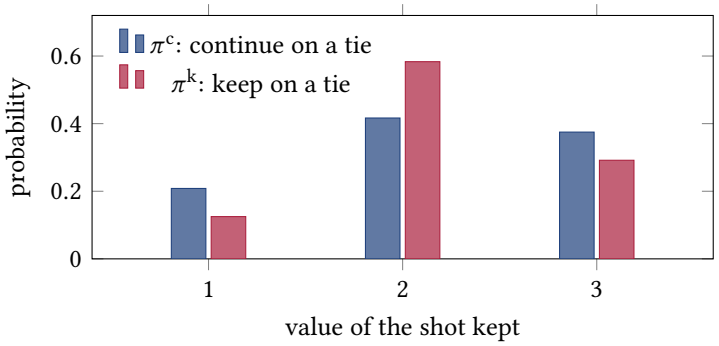


Figure 3.3. Return distributions of the two optimal policies of Example 3.17. Both have expected return 13/6.

Example 3.17 (The photo booth, solved). Take $H = 3$ shots with values $\{1, 2, 3\}$ and probabilities $(\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$. Write $V_t^\star(x)$ for the optimal value when the t-th shot, of value x, has just been taken and the session is still going on, and $c_t \coloneqq \mathbb{E}[V_{t+1}^\star(X)]$ for the value of continuing. The last shot must be kept: $V_3^\star(x) = x$ and $c_2 = \mathbb{E}[X] = \frac{5}{3}$. Then $V_2^\star(x) = \max(x, c_2)$, so $V_2^\star(1) = \frac{5}{3}$, $V_2^\star(2) = 2$, $V_2^\star(3) = 3$ and $c_1 = \frac{1}{2} \cdot \frac{5}{3} + \frac{1}{3} \cdot 2 + \frac{1}{6} \cdot 3 = 2$. Finally $V_1^\star(x) = \max(x, 2)$: the customer keeps the first shot if its value is 3, discards it if its value is 1, and is *indifferent* if its value is 2. The optimal expected value is $\mathbb{E}[V_1^\star(X)] = \frac{13}{6}$. Compare with the naive rule "stop only at a 3", whose expected value is $(1 - (\frac{5}{6})^2) \cdot 3 + (\frac{5}{6})^2 \cdot \frac{5}{3} = \frac{56}{27} \approx 2.07 < \frac{13}{6} \approx 2.17$.

The tie at $x = 2$ produces two optimal policies. Let π^{c} continue and π^{k} keep when the first shot is a 2. Both have expected return $\frac{13}{6}$, but their return distributions differ:

$$\nu^{\pi^{\mathrm{c}}} = \tfrac{5}{24}\delta_1 + \tfrac{5}{12}\delta_2 + \tfrac{3}{8}\delta_3, \qquad \nu^{\pi^{\mathrm{k}}} = \tfrac{1}{8}\delta_1 + \tfrac{7}{12}\delta_2 + \tfrac{7}{24}\delta_3.$$

A customer who would be disappointed by a value below 2 prefers π^{k} ($\mathbb{P}(W \geq 2) = \frac{7}{8}$ against $\frac{19}{24}$); one who hopes for a 3 prefers π^{c} (Figure 3.3). The expected-return criterion is blind to the difference.

3.4 Optimal stopping and the secretary problem

In an *optimal stopping* problem, the agent observes a random sequence and decides, at each time, whether to stop and collect a reward depending on the current observation, or to continue. It is an MDP with two actions, *stop* and *continue*, in which stopping leads to an absorbing state where nothing more happens. The photo booth is an instance; the secretary problem of Example 1.7 is a more subtle one, because the reward is not observed at the time of stopping.

Recall the setting: the t-th candidate interviewed has rank $\sigma(t)$ among all H (rank 1 for the best), where σ is uniformly distributed on the permutations of $[H]$; at time t the employer observes the relative ranks $\mathrm{rk}_1, \dots, \mathrm{rk}_t$, where $\mathrm{rk}_k \in [k]$ is the rank of the k-th candidate among the first k, and in particular whether the current candidate is the best so far, $I_t = \mathbb{1}\{\mathrm{rk}_t = 1\}$.

Lemma 3.18 (Relative ranks are independent). The relative ranks $\mathrm{rk}_1, \dots, \mathrm{rk}_H$ are independent, rk_t being uniform on $[t]$; in particular the indicators $I_1, \dots, I_H$ are independent, with $\mathbb{P}(I_t = 1) = 1/t$. Consequently, for $t \in [H]$, the probability that a candidate who is the best among the first t is the best of all H is $\mathbb{P}(I_{t+1} = \cdots = I_H = 0) = t/H$.

Proof. (This is Exercise 1.7.) The map $\sigma \mapsto (\mathrm{rk}_1, \dots, \mathrm{rk}_H)$ from the permutations of $[H]$ to $[1] \times [2] \times$

$\cdots \times [H]$ is a bijection: both sets have $H!$ elements, and the sequence of relative ranks determines the permutation (insert the candidates one at a time at the position given by their relative rank). Since σ is uniform, $(\mathrm{rk}_1, \dots, \mathrm{rk}_H)$ is uniform on a product set, so its coordinates are independent and rk_t is uniform on $[t]$. As $I_t = \mathbb{1}\{\mathrm{rk}_t = 1\}$, the claim on the I_t follows; the last one is $\prod_{k=t+1}^{H}(1 - \frac{1}{k}) = \prod_{k=t+1}^{H} \frac{k-1}{k} = \frac{t}{H}$. ∎

The problem is therefore an MDP with time-dependent dynamics, state $I_t \in \{0, 1\}$ (plus an absorbing state once the employer has stopped), independent transitions $I_{t+1} \sim \mathrm{Ber}(\frac{1}{t+1})$, and actions *stop* or *continue*. Stopping at time t in state 1 yields the random reward $\mathbb{1}\{\sigma(t) = 1\} = I_t \prod_{k=t+1}^{H}(1 - I_k)$, whose conditional expectation given the past is t/H by Lemma 3.18; stopping in state 0 yields 0; at time H the employer must stop. This reward depends on the future, which Definition 1.10 does not allow; but since we maximize an expectation, we may replace the random reward by its conditional expectation t/H, which turns the problem into a standard MDP: state space $\{0, 1\} \cup \{\mathrm{stop}\}$, kernels $P_t(1 \mid x, \mathrm{continue}) = \frac{1}{t+1}$ and $P_t(0 \mid x, \mathrm{continue}) = \frac{t}{t+1}$ for $x \in \{0, 1\}$, $P_t(\mathrm{stop} \mid x, \mathrm{stop}) = 1$, stop absorbing, and rewards $r_t(1, \mathrm{stop}, \mathrm{stop}) = t/H$, all other rewards being 0. Both problems have the same expected return under every policy (tower property), hence the same optimal policies (and since the return is a Bernoulli variable, its distribution is anyway determined by its mean). Observing all the relative ranks rather than the I_t does not help: the future indicators $I_{t+1}, \dots, I_H$ are independent of $(\mathrm{rk}_1, \dots, \mathrm{rk}_t)$ by Lemma 3.18, and the reward of stopping at t is a function of I_t and of them, so that the additional information acts on the process only as an external randomization, which cannot improve on the optimal deterministic Markov policy (Theorem 3.15).

Theorem 3.19 (Solution of the secretary problem). Let $r^\star := \min\left\{t \in [H] : \sum_{k=t}^{H-1} \frac{1}{k} \le 1\right\}$ (with the empty sum equal to 0; in this theorem and its proof we follow the classical notation of Ferguson (1989), in which r denotes a time threshold and $v(r)$ a success probability, not a reward and a value). An optimal policy is the *threshold rule*: reject the first $r^\star - 1$ candidates, then hire the first candidate who is the best so far (if none appears, the last candidate is hired and the attempt fails). Its success probability is

$$v(r^\star) = \frac{r^\star - 1}{H} \sum_{t=r^\star}^{H} \frac{1}{t-1} \quad \text{if } r^\star \ge 2, \qquad v(1) = \frac{1}{H}.$$

As $H \to \infty$, $r^\star/H \to 1/e$ and $v(r^\star) \to 1/e \approx 0.368$.

Proof. Structure of the optimal policy. Let c_t denote the optimal expected reward when the employer decides to continue at time $t < H$ (and $c_H := 0$, since continuing is impossible), and $V_t^\star(x)$ the optimal value at time t in state $x \in \{0, 1\}$. By Theorem 3.15, $V_t^\star(0) = c_t$ (stopping yields $0 \le c_t$), $V_t^\star(1) = \max(t/H, c_t)$, and since I_{t+1} is independent of the past, $c_{t-1} = \frac{1}{t} V_t^\star(1) + (1 - \frac{1}{t}) V_t^\star(0) = \frac{1}{t}\max(\frac{t}{H}, c_t) + (1 - \frac{1}{t}) c_t \ge c_t$. Thus $t \mapsto c_t$ is nonincreasing while $t \mapsto t/H$ is increasing: there is a first time $r' := \min\{t : t/H \ge c_t\}$ (it exists since $1 \ge c_H = 0$), before which continuing is strictly better and from which on stopping at a relative best is optimal. This is the threshold rule with threshold r'; it remains to see that $r' = r^\star$.

Value of a threshold rule. The rule with threshold $r \ge 2$ succeeds if for some $t \ge r$, candidate t is the first relative best after time $r - 1$ and is the overall best. By Lemma 3.18,

$$v(r) = \sum_{t=r}^{H} \mathbb{P}\big(I_r = \cdots = I_{t-1} = 0\big)\mathbb{P}(I_t = 1)\mathbb{P}\big(I_{t+1} = \cdots = I_H = 0\big) = \sum_{t=r}^{H} \frac{r-1}{t-1} \cdot \frac{1}{t} \cdot \frac{t}{H} = \frac{r-1}{H} \sum_{t=r}^{H} \frac{1}{t-1},$$

using $\prod_{k=r}^{t-1} \frac{k-1}{k} = \frac{r-1}{t-1}$; and $v(1) = \mathbb{P}(I_1 = 1)\mathbb{P}(I_2 = \cdots = I_H = 0) = \frac{1}{H}$.

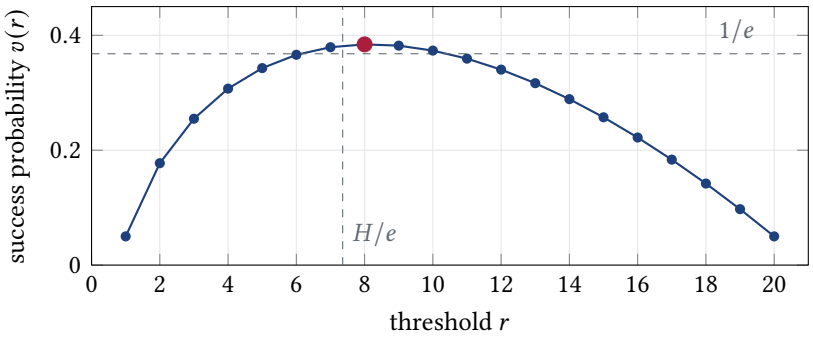


Figure 3.4. Success probability $v(r)$ of the threshold rules for $H = 20$ candidates. The optimal threshold is $r^\star = 8$, with $v(r^\star) \approx 0.384$; the dashed lines mark the asymptotic values $H/e \approx 7.4$ and $1/e \approx 0.368$.

Identification of r'. For $t \geq r' - 1$, the optimal continuation from time t is the threshold rule with threshold $t+1$ applied to the remaining candidates, whose success probability, by the independence of the I_k, is $c_t = v(t+1) = \frac{t}{H} \sum_{k=t}^{H-1} \frac{1}{k}$. The stopping condition $t/H \geq c_t$ thus reads $\sum_{k=t}^{H-1} \frac{1}{k} \leq 1$ for such t; since the sum is decreasing in t and the condition fails at $t = r' - 1$, it fails for all smaller t as well. Hence r' is the first t at which $\sum_{k=t}^{H-1} \frac{1}{k} \leq 1$, that is, $r' = r^\star$.

Asymptotics. Comparing sums and integrals, $\ln \frac{H}{t} \leq \sum_{k=t}^{H-1} \frac{1}{k} \leq \ln \frac{H-1}{t-1}$ for $2 \leq t \leq H$. Hence $\sum_{k=t}^{H-1} \frac{1}{k} \leq 1$ implies $t \geq H/e$, and it holds as soon as $t-1 \geq (H-1)/e$: therefore $H/e \leq r^\star \leq (H-1)/e+2$, so $r^\star/H \to 1/e$. Finally $v(r^\star) = \frac{r^\star-1}{H} \sum_{k=r^\star-1}^{H-1} \frac{1}{k}$ is squeezed between $\frac{r^\star-1}{H} \ln \frac{H}{r^\star-1}$ and $\frac{r^\star-1}{H} \ln \frac{H-1}{r^\star-2}$, both of which tend to $\frac{1}{e} \ln e = \frac{1}{e}$. ∎

The optimal rule is remarkable in two respects: it is very simple, and its success probability does not vanish as the number of candidates grows. Even among a thousand candidates, observing the first 368 and hiring the next candidate who is the best so far gives a 37% chance of hiring the very best. The value $1/e$ is the price of observing only comparisons: when the candidates' values are observed, i.i.d. with a known continuous law, the best success probability is about 0.58 for large H, a limit computed by Gilbert and Mosteller (1966) (Exercise 3.5). Figure 3.4 shows the success probability of the threshold rules for $H = 20$.

3.5 A linear-quadratic control problem ♦

We end with an example where the state and action spaces are the real line, and where backward induction can be carried out in closed form because all value functions are quadratic.

Example 3.20 (Steering a rocket). A rocket must reach the position $\bar{s}$ at time $H+1$. Its position evolves as $S_{t+1} = S_t + A_t + U_{t+1}$, where $A_t \in \mathbb{R}$ is the thrust chosen at time t and the perturbations $U_2, U_3, \ldots$ are i.i.d. with $\mathbb{E}[U_{t+1}] = 0$ and $\mathbb{V}\mathrm{ar}[U_{t+1}] = \sigma^2$, U_{t+1} being independent of the past (the noise is indexed as in Proposition 1.12). Each unit of thrust is costly, and so is missing the target: the rewards are $r_t(s, a, s') = -a^2$ for $t < H$ and $r_H(s, a, s') = -a^2 - \kappa(s' - \bar{s})^2$, with $\kappa > 0$. Here $\mathcal{S} = \mathcal{A} = \mathbb{R}$. A simplified model of the heating of a building, where S_t is the temperature and A_t the heating power, has the same structure (Exercise 3.7).

In Theorem 3.15, the finiteness of $\mathcal{A}$ was used to guarantee that the maxima exist, and the finiteness of $\mathcal{S}$ to make all sums finite and every function measurable. Here the maxima are attained because the maximand is a strictly concave quadratic function, the sums over s' become integrals with respect to the law of U_{t+1}, and all the value functions computed below are polynomials, hence measurable with

finite expectations; the induction is otherwise unchanged, and the optimality it establishes holds among all history-dependent policies. These are the three points to check whenever the planning theorem is used on a general state space (attained maxima, finite expectations, measurable value functions); Section 3.6 checks them again for a second continuous example.

Theorem 3.21 (Linear-quadratic control). In Example 3.20, let $k = H - t + 1$ be the number of remaining decisions at time t, and define

$$\kappa_k \coloneqq \frac{\kappa}{1 + k\kappa}, \qquad d_k \coloneqq \sigma^2 \sum_{j=0}^{k-1} \kappa_j = \sigma^2 \kappa \sum_{j=0}^{k-1} \frac{1}{1 + j\kappa}.$$

Then the optimal value functions and the optimal policy are

$$V_t^\star(s) = -\kappa_k \, (\bar{s} - s)^2 - d_k, \qquad \pi_t^\star(s) = \kappa_k \, (\bar{s} - s) = \frac{\kappa \, (\bar{s} - s)}{1 + (H - t + 1)\kappa}.$$

Proof. It is convenient to attach the terminal cost to time $H + 1$: set $V_{H+1}^\star(s) = -\kappa(\bar{s} - s)^2$, which is $-\kappa_0(\bar{s} - s)^2 - d_0$ with $\kappa_0 = \kappa$ and $d_0 = 0$, and use rewards $-a^2$ at all times. Assume $V_{t+1}^\star(s) = -\kappa_{k-1}(\bar{s} - s)^2 - d_{k-1}$ for some $k \geq 1$. Then

$$Q_t^\star(s, a) = \mathbb{E}\big[V_{t+1}^\star(s + a + U_{t+1})\big] - a^2 = -\kappa_{k-1}\big[(\bar{s} - s - a)^2 + \sigma^2\big] - d_{k-1} - a^2,$$

using $\mathbb{E}[U_{t+1}] = 0$ and $\mathbb{E}[U_{t+1}^2] = \sigma^2$. This is a strictly concave quadratic function of a, maximized at the unique point where $2\kappa_{k-1}(\bar{s} - s - a) - 2a = 0$, that is, at $a^\star = \frac{\kappa_{k-1}}{1+\kappa_{k-1}}(\bar{s} - s)$. Substituting, $\bar{s} - s - a^\star = \frac{1}{1+\kappa_{k-1}}(\bar{s} - s)$ and

$$V_t^\star(s) = -\left[\frac{\kappa_{k-1}}{(1 + \kappa_{k-1})^2} + \frac{\kappa_{k-1}^2}{(1 + \kappa_{k-1})^2}\right](\bar{s} - s)^2 - \kappa_{k-1}\sigma^2 - d_{k-1} = -\frac{\kappa_{k-1}}{1 + \kappa_{k-1}}(\bar{s} - s)^2 - \big(d_{k-1} + \kappa_{k-1}\sigma^2\big).$$

Hence $\kappa_k = \frac{\kappa_{k-1}}{1+\kappa_{k-1}}$, i.e. $\frac{1}{\kappa_k} = \frac{1}{\kappa_{k-1}} + 1$, which gives $\frac{1}{\kappa_k} = \frac{1}{\kappa} + k$; and $d_k = d_{k-1} + \kappa_{k-1}\sigma^2$. The optimal action is $a^\star = \kappa_k(\bar{s} - s)$ since $\frac{\kappa_{k-1}}{1+\kappa_{k-1}} = \kappa_k$, and it is unique at every step, so the optimal policy is unique. ∎

The optimal thrust is a fraction of the remaining distance to the target, and this fraction increases as the deadline approaches: correct gently at first, decisively at the end (Figure 3.5). Two features deserve attention. The optimal policy does not depend on σ^2, nor on the law of the perturbations beyond its first two moments: this is the *certainty equivalence* principle of linear-quadratic control, and it fails as soon as the cost is not quadratic. And the noise degrades the value by d_k, which grows only logarithmically with the number of remaining steps, since $\sum_{j=0}^{k-1} \frac{\kappa}{1+j\kappa} \leq \kappa + \ln\big(1 + (k-1)\kappa\big)$. The general linear-quadratic problem, with vector states, linear dynamics $S_{t+1} = FS_t + GA_t + U_{t+1}$ and quadratic costs, is solved by the same induction; the coefficients κ_k are then matrices satisfying the *Riccati recursion*, see Anderson and Moore (1990) and Bertsekas (2017, Chap. 3).

3.6 Structured policies: the machine replacement problem

Backward induction computes optimal policies; it also reveals their *structure*, which is often what a practitioner needs—a rule of thumb rather than a table. The secretary problem gave a first instance (a threshold on time). The machine replacement problem of Example 1.8 gives another: a threshold on the state.

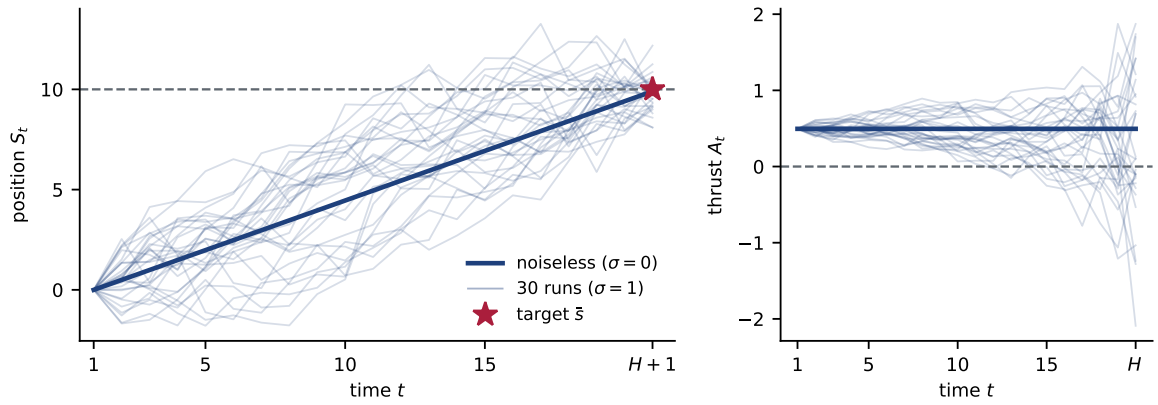


Figure 3.5. The rocket of Example 3.20 under the optimal policy of Theorem 3.21, with $H = 20$, $S_1 = 0$, $\bar{s} = 10$, $\kappa = 5$ and Gaussian perturbations. Left: positions; right: thrusts. Without noise, the optimal thrust is constant, equal to $\kappa(\bar{s} - S_1)/(1 + H\kappa) \approx 0.50$, and the trajectory is a straight line. With $\sigma = 1$, the early deviations are corrected gently and the late ones decisively; the last perturbation U_{H+1} cannot be corrected, so that the final position keeps a standard deviation close to σ.

Recall the setting: the wear $S_t \in \mathbb{R}_+$ increases each year by an independent nonnegative amount X_{t+1} with a fixed law; keeping the machine costs $c(S_t)$, replacing it costs $C + c(0)$ and resets the wear to 0 before the year's deterioration. We assume only that $c \geq 0$ is nondecreasing. Since replacing every year is always possible, $-H(C + c(0)) \leq V_t^\star \leq 0$, so that all the expectations below are finite; the value functions are measurable because they are monotone (this is part of the induction); and the maximizations are over two actions. The proof of Theorem 3.15 therefore applies to this continuous state space, for the same three reasons as in Section 3.5.

Theorem 3.22 (Control-limit policies, Derman, 1963; see Puterman, 1994, Thm. 4.6). In Example 1.8, for every $t \in [H]$ the optimal value function $V_t^\star$ is nonincreasing in the wear, and the set $\mathcal{X}_t^{\text{rep}} \coloneqq \{x : Q_t^\star(x, \text{keep}) \leq Q_t^\star(x, \text{replace})\}$ of wear levels at which replacing is optimal is an upper set of $\mathbb{R}_+$: with $x_t^\star \coloneqq \inf \mathcal{X}_t^{\text{rep}} \in [0, \infty]$, it is either $[x_t^\star, \infty)$ or $(x_t^\star, \infty)$. Replacing is optimal at time t whenever $S_t > x_t^\star$, keeping is optimal whenever $S_t < x_t^\star$, and the policy "at time t, replace if and only if $S_t \in \mathcal{X}_t^{\text{rep}}$" is optimal. If moreover $x \mapsto Q_t^\star(x, \text{keep})$ is continuous (for instance if c is continuous and the increments have a density) and $0 < x_t^\star < \infty$, both actions are optimal at $x_t^\star$, so that "replace if and only if $S_t > x_t^\star$" is optimal too.

Proof. By backward induction on t, we prove that $V_t^\star$ is nonincreasing. This holds for $V_{H+1}^\star \equiv 0$. Assume it for $t + 1$. The two state-action values are

$$Q_t^\star(x, \text{keep}) = -c(x) + \mathbb{E}\big[V_{t+1}^\star(x + X)\big], \qquad Q_t^\star(x, \text{replace}) = -C - c(0) + \mathbb{E}\big[V_{t+1}^\star(X)\big] \coloneqq \bar{Q}_t,$$

where X has the law of the yearly increment. The first one is nonincreasing in x, as the sum of $-c$ and of the expectation of a nonincreasing function of $x + X$; the second one does not depend on x. Hence $V_t^\star(x) = \max(Q_t^\star(x, \text{keep}), \bar{Q}_t)$ is nonincreasing, which completes the induction. Moreover the set $\{x : Q_t^\star(x, \text{keep}) \leq \bar{Q}_t\}$ on which replacing is optimal is an upper set of $\mathbb{R}_+$ (if x belongs to it, so does every $x' \geq x$); letting $x_t^\star$ be its infimum (with $x_t^\star = \infty$ if it is empty), replacing is optimal for $x > x_t^\star$ and keeping for $x < x_t^\star$; at $x = x_t^\star$, replacing is optimal if the set is closed and keeping otherwise, which is the rule "replace if and only if $S_t \in \mathcal{X}_t^{\text{rep}}$". The boundary convention matters: with a single decision left, $C = 1$ and $c(x) = 2\mathbb{1}\{x \geq 1\}$, one has $\mathcal{X}_H^{\text{rep}} = [1, \infty)$, and at $x = 1$ replacing (cost 1) is strictly better than keeping (cost 2), so that the strict rule "replace if and only if $x > 1$" is not optimal. Under the

continuity assumption, finally, $x \mapsto Q_t^\star(x, \text{keep}) - \bar{Q}_t$ is continuous and nonincreasing, positive on $[0, x_t^\star)$ and nonpositive on $(x_t^\star, \infty)$, hence zero at $x_t^\star$. (If c is continuous and the increment X has a density, then $x \mapsto \mathbb{E}[V_{t+1}^\star(x + X)]$ is continuous, since $V_{t+1}^\star$ is bounded and translation is continuous in L^1.) ∎

The theorem reduces the search for an optimal policy from a function of the wear to H numbers. For instance, with exponential increments of mean 1, $c(x) = x$ and $C = 4$, backward induction on a fine grid gives $x_H^\star = 4$ when a single year remains—replace only if the maintenance cost reaches the price of a new machine—then $x_{H-1}^\star \approx 2.1$, $x_{H-2}^\star \approx 1.8$, and $x_t^\star \approx 2.0$ for all earlier times: the thresholds quickly stabilize, as they should when the remaining horizon is long. The limit is the threshold of the long-run average cost (Example 9.12); the discounted problem of Chapter 8 has a slightly different threshold (≈ 2.1 for $\gamma = 0.9$). Similar monotonicity arguments underlie many structured policies of operations research: order-up-to levels in inventory control, acceptance thresholds in revenue management, and the threshold rules of optimal stopping. The proof above uses one property of the expectation only, its monotonicity; Proposition 5.11 and Corollary 5.12 isolate the argument, which therefore holds for the risk-sensitive criteria of Chapters 5 and 6, and Example 5.13 shows the thresholds moving in an unexpected direction under risk aversion.

Distributional viewpoint

The Bellman equation is first of all an identity between probability measures (Theorem 3.2): the law of the reward-to-go is a mixture of shifted copies of the laws of the reward-to-go from the next states, $v_t^\pi = \mathcal{T}_t^\pi v_{t+1}^\pi$. The classical equation for value functions is its image under the linear functional "expectation" (Corollary 3.11). The distributional recursion is an algorithm too (Algorithm 3.1), exact when the rewards live on a lattice; in general it faces the growth of the support of the return distribution (Remark 3.9). The planning theorem (Theorem 3.15) rests on two properties of the expectation, linearity and monotonicity, and its conclusion—optimal policies are deterministic and Markov—is tied to the expected-return criterion. Two optimal policies may have very different return distributions (Example 3.17); which one to prefer is a question the mean cannot answer.

Hands-on

The notebook `photoBooth.ipynb` (session H02) asks for an agent solving the photo booth for arbitrary numbers of shots, probabilities and values: backward induction as in Example 3.17. Session H03 (*windy hike around the pit*) is the windy hike of Examples 1.4 and 3.6: it asks for backward recursions computing statistics of the return beyond its mean—the criteria of the two hikers—and prepares Chapters 4 and 5. The planning part of `retailStore.ipynb` (session H04) can be run in the finite-horizon setting with Algorithms 3.2 and 3.3.

Exercises

Exercise 3.1 (Maximal attainable return). [verify] Under a fixed Markov policy π, let $M_t^\pi(s)$ be the largest value of $W_{t:H}$ that has positive probability under $\mathbb{P}_{t,s}^\pi$. Show that $M_{H+1}^\pi = 0$ and

$$M_t^\pi(s) = \max\left\{r_t(s, a, s') + M_{t+1}^\pi(s') : \ \pi_t(a \mid s) > 0, \ P_t(s' \mid s, a) > 0\right\},$$

a Bellman equation in the $(\max, +)$ algebra. Give the analogous recursion for the largest return attainable with positive probability under *some* policy, and explain why the resulting policy is

deterministic and Markov. What about the *smallest* return?

Exercise 3.2 (Occupation measures). [verify] For $\pi \in \Pi_{\mathrm{MR}}$ and $\mu \in \mathcal{M}_1(\mathcal{S})$, let $d_t^\pi(s,a) = \mathbb{P}_\mu^\pi(S_t = s, A_t = a)$. Show the forward recursion $d_{t+1}^\pi(s') = \sum_{s,a} d_t^\pi(s,a) P_t(s' \mid s,a)$ and $d_t^\pi(s,a) = \pi_t(a \mid s) \sum_{a'} d_t^\pi(s,a')$, and that $V^\pi(\mu) = \sum_{t,s,a} d_t^\pi(s,a)\, r_t(s,a)$. Deduce that $\max_\pi V^\pi(\mu)$ is the value of a linear program over the nonnegative vectors (d_t) satisfying the flow constraints $\sum_a d_1(s,a) = \mu(s)$ and $\sum_a d_{t+1}(s',a) = \sum_{s,a} d_t(s,a) P_t(s' \mid s,a)$. How does one recover a policy from an optimal solution?

Exercise 3.3 (Q-form of the distributional equation). [verify] With the convention $\nu_{H+1}^\pi(s,a) \coloneqq \delta_0$, show that the state-action return distributions satisfy

$$\nu_t^\pi(s,a) = \sum_{s'} P_t(s' \mid s,a)\, (b_{r_t(s,a,s')})_{\#}\Big(\sum_{a'} \pi_{t+1}(a' \mid s')\, \nu_{t+1}^\pi(s',a') \Big),$$

and define the corresponding operator on $\mathcal{M}_1(\mathbb{R})^{\mathcal{S}\times\mathcal{A}}$. Check that it enjoys the two properties of Proposition 3.4.

Exercise 3.4 (Riverswim). [implement] Take $N = 4$, $p = 0.3$, $q = 0.6$ in Example 1.5, with its reward $r(s,a,s') = \mathbb{1}\{s' = N\}$ (a reward of 1 for every transition into state N, including from N itself), and add the small reward 0.005 of the original formulation for the action L in state 0. For $H = 3$, $H = 6$ and $H = 12$, compute the optimal policy and the optimal value from state 0 (by computer; backward induction is exact here and instantaneous). From which horizon on is it optimal to swim to the right from the leftmost state at the first step? Compute the return distribution of the optimal policy for $H = 6$ and comment on its shape.

Exercise 3.5 (The full-information secretary problem). [verify] Suppose that the candidates of Example 1.7 have i.i.d. values $U_1, \dots, U_H$, uniform on $[0,1]$, the best candidate being the one with the largest value, and that the employer observes the values themselves. The goal is still to hire the best candidate.

(i) For $H = 2$, show that hiring the first candidate if and only if $U_1 \geq \frac{1}{2}$ succeeds with probability $\frac{3}{4}$, whereas Theorem 3.19 gives $\frac{1}{2}$. Why does this not contradict Lemma 3.18?

(ii) Let $M_t = \max_{s \leq t} U_s$ and $I_t = \mathbb{1}\{U_t = M_t\}$. Show that the problem is an MDP with state (M_t, I_t) (plus an absorbing state once the employer has stopped), and that hiring at time t a best-so-far candidate of value u succeeds with probability u^k, where $k = H - t$ is the number of candidates left.

(iii) Show that dismissing her and then hiring the first later candidate of value larger than u succeeds with probability $g_k(u) = \sum_{j=1}^k \binom{k}{j} (1-u)^j u^{k-j}/j$. Deduce that $u^k \geq g_k(u)$ if and only if $u \geq b_k$, where $b_k \in (0,1)$ is defined by $\sum_{j=1}^k \binom{k}{j} \big(b_k^{-1} - 1\big)^j / j = 1$ for $k \geq 1$, and $b_0 \coloneqq 0$; show that b_k is nondecreasing in k.

(iv) Show by induction on k that an optimal policy hires at time t the first best-so-far candidate with $U_t \geq b_{H-t}$. *Hint:* if $u \geq b_k$, every later best-so-far candidate has a value larger than $u \geq b_{k'}$, where $k' < k$ is the number of candidates left after her.

(v) Compute b_1, b_2, b_3 and, by computer, the optimal success probability for $H = 20$; compare

with Figure 3.4.

Exercise 3.6 (Selling to the best offer). [verify] In the setting of Exercise 3.5, suppose that the employer wants to maximize the expected value of the hired candidate instead (in the classical version of Cayley and Moser, a seller receives successive offers for a house and wants to maximize the expected price obtained; see the bibliographical notes). Show that $V_t^\star(u) = \max(u, c_t)$ with $c_H = 0$ and $c_{t-1} = \mathbb{E}[\max(U, c_t)] = \frac{1+c_t^2}{2}$. Show that c_t increases as t decreases, that $1 - c_{H-k} \sim 2/k$ as $k \to \infty$, and interpret.

Exercise 3.7 (Heating). [verify] The temperature of a building evolves as $S_{t+1} = (1-\lambda)S_t + \lambda A_t + U_{t+1}$, where A_t is the temperature of the heating fluid, $\lambda \in (0,1)$ is the inertia, and the U_{t+1} are i.i.d. centered with variance σ^2, as in Example 3.20; the rewards are $-a^2$ at each step and $-\kappa(S_{H+1} - \bar{s})^2$ at the end. Adapt the proof of Theorem 3.21 to compute the optimal policy and value.

Exercise 3.8 (Deterministic MDPs and shortest paths). [verify] Suppose all kernels are Dirac masses: $P_t(\cdot \mid s, a) = \delta_{f_t(s,a)}$. Show that $v_t^\pi(s)$ is a Dirac mass for every deterministic π, and that the planning problem is a longest-path problem in a layered graph with $H+1$ layers. Compare the complexity of Algorithm 3.3 with that of the Bellman–Ford algorithm.

Exercise 3.9 (Growth of the support). [limits] Build an MDP with a single state, two actions and time-dependent rewards, together with a randomized Markov policy, such that the return distribution $v_1^\pi(s)$ has 2^H distinct atoms. *Hint:* use rewards equal to powers of 2. Show conversely that, for any MDP and any Markov policy, the number of atoms of $v_t^\pi(s)$ is at most $(SA)^{H-t+1}$.

Exercise 3.10 (Mean and variance are not enough). [limits] In the photo booth of Example 3.17, compute the variances of the two optimal return distributions. Show that no choice of three values $v_1 < v_2 < v_3$ can make them equal (*hint:* two distinct probability vectors on three points with the same mean have different second moments—consider the Vandermonde determinant). Then build a one-step MDP with two actions whose return distributions have the same mean *and* the same variance but are different (for instance $\frac{1}{2}(\delta_{-1} + \delta_1)$ and $\frac{1}{8}(\delta_{-2} + 6\delta_0 + \delta_2)$), and conclude that no finite family of moments can, in general, decide which of two optimal policies a risk-sensitive agent should prefer.

Bibliographical notes

Backward induction and the principle of optimality are due to Bellman (1957b); the finite-horizon theory in the generality of Theorem 3.15, including the sufficiency of deterministic Markov policies among history-dependent ones, is in Puterman (1994, Chap. 4). The measure-theoretic difficulties that arise on general state spaces are treated by Bertsekas and Shreve (1978).

The distributional Bellman equation appears, in the discounted setting, in Morimura et al. (2010) and Bellemare et al. (2017); the operator viewpoint and the structural properties of Proposition 3.4 are developed in Rowland et al. (2018) and in the monograph of Bellemare et al. (2023). The finite-horizon formulation used here follows Marthe et al. (2023).

Optimal stopping is a classical chapter of probability theory; Ferguson (2006) is an accessible introduction. The solution of the secretary problem is due to Lindley (1961); Gilbert and Mosteller (1966) solve many variants of the *best-choice* problem, in which the employer maximizes the probability of hiring the very best candidate; the full-information version of Exercise 3.5, with its limiting success probability ≈ 0.580, is theirs. Ferguson (1989) recounts the history. Exercise 3.6, where the observed values themselves are maximized in expectation, is a different problem: it is the one posed by Cayley in 1875 and given its smooth solution by Moser (1956), treated in Ferguson (2006, Sect. 2.4). Linear-quadratic control goes back to Kalman (1960); Anderson and Moore (1990) and Bertsekas (2017, Chap. 3) are standard references, and the latter discusses certainty equivalence.