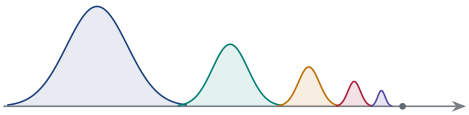
Part III

Infinite Horizon

7

Discounted MDPs: Operators and Contractions



When a decision problem has no natural end—a shop that opens every week, a machine that is maintained indefinitely—the horizon is infinite, and rewards must be discounted for the return to make sense. A long finite horizon already behaves this way: far from the deadline, the optimal decisions of a finite-horizon problem with stationary data stop depending on time, and discounting is the simplest way to make this stationarization exact. Discounting replaces the shift of the finite-horizon recursion by the affine map $x \mapsto r + \gamma x$, and backward induction by the search for a fixed point. The distributional Bellman operator becomes a contraction in Wasserstein distance, the expected one a contraction in supremum norm, and both fixed points are reached by iteration, at a geometric rate governed by γ. This chapter develops the two operators, shows what discounting does to the return distribution—which may become fractal—and to the statistics of Chapter 4: moments survive with powers of γ, exponential moments change rate at each step, and the finite representations of distributions that the projections of distributional reinforcement learning rely on are introduced.

7.1 Far from the horizon

The finite-horizon model of Chapters 1 and 3 allows the kernel and the reward to depend on time. In most of the examples of Chapter 1 they do not: the bike store faces the same demand law every week, the machine wears out by the same law every year, the river flows as it always did. The data (P, r) are *stationary*, and only the horizon H distinguishes one week from another. Backward induction then simplifies. Writing $k = H - t + 1$ for the number of decisions that remain at time t, and $V_k \coloneqq V^\star_{H-k+1}$ for the optimal value with k *steps to go*, the recursion of Theorem 3.15 reads $V_0 = 0$ and

$$V_k(s) \;=\; \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P(s' \mid s, a)\big[r(s, a, s') + V_{k-1}(s')\big], \qquad k = 1, 2, \dots, \tag{7.1}$$

and an optimal decision rule with k steps to go, $d_k(s)$, is any action attaining the maximum: it is *greedy* with respect to V_{k-1}. The map $V_{k-1} \mapsto V_k$ is the Bellman optimality operator T^* of Chapter 8 (Definition 8.1, without discount), so that $V_k = (T^*)^k 0$, and the optimal policy of the H-horizon problem is $(d_H, d_{H-1}, \dots, d_1)$. Neither V_k nor d_k depends on H: the clock that matters runs backward from the deadline, and lengthening the horizon only adds steps at the far end.

What does the rule d_k look like when k is large? Figure 7.1 answers for two examples. In the machine replacement problem with $c(x) = x$ and $C = 4$, the thresholds of Theorem 3.22 are $x^\star_1 = 4$ (replace in the last year only if the maintenance cost reaches the price of a new machine), $x^\star_2 \approx 2.1$, $x^\star_3 \approx 1.8$, and then $x^\star_k = 2.0$ for every $k \geq 4$. In the bike store, with a single week to go the manager

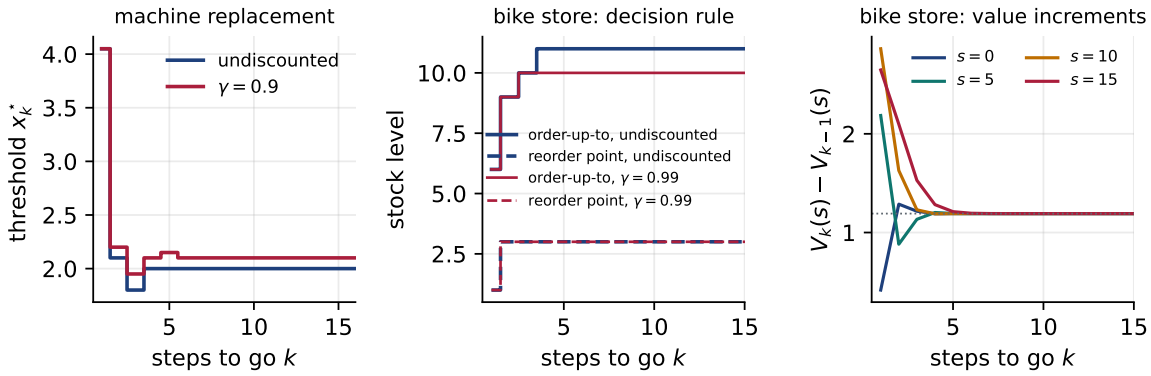


Figure 7.1. Backward induction with stationary data, indexed by the number k of steps to go. Left: the replacement threshold $x_k^\star$ of Theorem 3.22 ($c(x) = x$, $C = 4$, exponential increments of mean 1, wear discretized with step 0.05), undiscounted and with $\gamma = 0.9$. Middle: the reorder point and the order-up-to level of the optimal rule of the bike store (Example 7.4(i)), undiscounted and with $\gamma = 0.99$. Right: the undiscounted increments $V_k(s) - V_{k-1}(s)$ of the bike store for four stock levels; the dotted line is their common limit $g^\star \approx 1.19$, the optimal average weekly profit. In each case a few end effects precede a rule that no longer depends on time.

orders only when the stock is at most 1, and only up to 6 bikes, since whatever is not sold this week is lost; with two weeks to go the reorder point is 3 and the order-up-to level 9; with three weeks, 3 and 10; and from $k = 4$ on the rule is frozen, "order up to 11 when the stock is at most 3", whatever the number of weeks that remain. The value functions themselves do not converge: $V_k(s)$ grows by about 1.19 per additional week. But their increments $V_k(s) - V_{k-1}(s)$ converge, to the same number $g^\star \approx 1.192$ for every stock level s (right panel): an extra week far from the deadline is worth the same whatever the current state. Hence $V_k \approx kg^\star + h^\star$ for a function $h^\star$ of the state alone: the *level* of V_k drifts while its *shape* converges, and the shape is all that the rule d_k sees, since adding a constant to V_{k-1} in (7.1) changes no maximizer.

The phenomenon deserves a name. Seen from far away, the deadline is invisible: the terminal condition $V_0 = 0$ is forgotten by the recursion as the initial law is forgotten by an ergodic Markov chain, every remaining week looks like every other, and the problem *stationarizes*—it becomes, for all practical purposes, the problem of an agent who faces the same situation every week and will do so indefinitely. An optimal policy for such an agent has no reason to depend on time: it is *stationary*, a single decision rule d applied for ever. This is the subject of this part of the book, and there are two ways to make the limit precise.

The first keeps the undiscounted sum and studies the limits directly: the gain $g^\star$, the relative value $h^\star$, and the stationary rules greedy with respect to $h^\star$. This is the average-reward theory of Chapter 9. It is the more faithful route and the more delicate one, because the forgetting that Figure 7.1 exhibits is not guaranteed: it rests on the mixing properties of the chains that the policies induce, and a periodic structure can keep the rules d_k oscillating with the parity of k for ever (Exercise 7.1), in which case neither $V_k - kg^\star$ nor d_k converges. In the replacement problem the undiscounted threshold 2.0 is the one of the average-cost problem (Example 9.12).

The second way is to *discount*: replace the total $\sum_{t \le H} R_t$ by $\sum_{t \le H} \gamma^{t-1} R_t$ for some $\gamma \in (0, 1)$. The recursion (7.1) then carries a factor γ in front of V_{k-1}, and the operator T_γ^* it defines is a γ-contraction for the supremum norm: since $|\max_a x_a - \max_a y_a| \le \max_a |x_a - y_a|$,

$$\big|(T_\gamma^* V)(s) - (T_\gamma^* V')(s)\big| \;\le\; \max_a \gamma \sum_{s'} P(s' \mid s, a)\, |V(s') - V'(s')| \;\le\; \gamma \|V - V'\|_\infty .$$

By Banach's theorem (Theorem D.16) it has a unique fixed point $V_\gamma^\star$, which Chapter 8 identifies with the

optimal value of the infinite-horizon discounted problem, and $V_k = (T_\gamma^*)^k 0$ converges to it at the rate γ^k *whatever the structure of the chains*: the terminal condition is forgotten geometrically, unconditionally, and the decision rules stabilize accordingly.

Proposition 7.1 (Stabilization of the decision rules under discounting). Let $V_k = (T_\gamma^*)^k 0$ be the optimal values of the discounted problems with k steps to go, $V_\gamma^\star$ the fixed point of T_γ^*, $Q_\gamma^\star(s,a) \coloneqq \sum_{s'} P(s' \mid s,a)[r(s,a,s') + \gamma V_\gamma^\star(s')]$, $\mathcal{A}^\star(s) \coloneqq \arg\max_a Q_\gamma^\star(s,a)$ the optimal actions, and

$$\Delta \;\coloneqq\; \min\Big\{ \max_{a'} Q_\gamma^\star(s,a') - Q_\gamma^\star(s,a) \;:\; s \in \mathcal{S},\; a \notin \mathcal{A}^\star(s) \Big\}$$

the *action gap* ($\Delta = +\infty$ if every action is optimal everywhere). Then $\|V_k - V_\gamma^\star\|_\infty \le \gamma^k R_{\max}/(1-\gamma)$ for every $k \ge 0$, and every decision rule d_k greedy with respect to V_{k-1} satisfies $d_k(s) \in \mathcal{A}^\star(s)$ for all s as soon as

$$k \;\ge\; k_0 \;\coloneqq\; \left\lfloor \frac{\ln\big(2R_{\max}/((1-\gamma)\Delta)\big)}{\ln(1/\gamma)} \right\rfloor + 1.$$

In particular, if the optimal action is unique in every state, $\mathcal{A}^\star(s) = \{\pi^\star(s)\}$, the optimal policy of the discounted problem with horizon H applies the stationary rule $\pi^\star$ at every time $t \le H - k_0 + 1$: only the last $k_0 - 1$ decisions feel the deadline.

Proof. From $V_\gamma^\star = T_\gamma^* V_\gamma^\star$, $\|V_\gamma^\star\|_\infty \le R_{\max} + \gamma\|V_\gamma^\star\|_\infty$, so $\|V_\gamma^\star\|_\infty \le R_{\max}/(1-\gamma)$, and the contraction gives $\|V_k - V_\gamma^\star\|_\infty = \|(T_\gamma^*)^k 0 - (T_\gamma^*)^k V_\gamma^\star\|_\infty \le \gamma^k \|V_\gamma^\star\|_\infty$. Let $Q_k(s,a) \coloneqq \sum_{s'} P(s' \mid s,a)[r(s,a,s') + \gamma V_{k-1}(s')]$ be the quantity that d_k maximizes; then $|Q_k(s,a) - Q_\gamma^\star(s,a)| \le \gamma\|V_{k-1} - V_\gamma^\star\|_\infty \le \gamma^k R_{\max}/(1-\gamma)$, which is less than $\Delta/2$ when $k \ge k_0$. For such k, if $a \notin \mathcal{A}^\star(s)$ and $a^\star \in \mathcal{A}^\star(s)$, then $Q_k(s,a) < Q_\gamma^\star(s,a) + \Delta/2 \le Q_\gamma^\star(s,a^\star) - \Delta/2 < Q_k(s,a^\star)$: a does not maximize $Q_k(s,\cdot)$. The last claim follows since d_k is then forced to be $\pi^\star$, and $(d_H, \dots, d_1)$ is optimal by Theorem 3.15. ∎

The guarantee is uniform over all MDPs with the same $R_{\max}$, γ and Δ, and correspondingly pessimistic. In the bike store with $\gamma = 0.99$, the rule freezes from $k = 3$ on (middle panel of Figure 7.1), whereas $\Delta \approx 1.4 \cdot 10^{-3}$ and $R_{\max} = 10.5$ give $k_0 = 1\,419$; in the replacement problem with $\gamma = 0.9$, it freezes from $k = 6$ while $k_0 = 103$. The reason is the one just seen: the bound tracks the *level* of V_k, which converges at the rate γ^k, while the rule only needs its *shape*, which converges much faster. Chapter 8 returns to this gap with the span seminorm (Corollary 8.8), and Exercise 7.1 shows a case where k_0 is of the right order.

The infinite-horizon discounted problem is the problem "with infinitely many steps to go": its value $V_\gamma^\star = \lim_k V_k$ carries no end effect at all, and its optimal policies are stationary (Theorem 8.3). This is what we mean by *stationarization by discounting*, and the model of this chapter and the next. Its price is a change of criterion, which the next section examines from three sides (a preference for the present, an effective horizon, a random horizon). Its reward is unconditional forgetting at an explicit rate, hence contractions and fixed points in place of inductions, and a theory that holds for every finite MDP. The undiscounted limit is recovered as $\gamma \to 1$: the number k_0 of end-effect steps then grows like $\ln(1/(1-\gamma))/(1-\gamma)$, and the discounted optimal policies converge to a policy that is optimal for the average reward (Blackwell's theorem, Theorem 9.9). A third way, where the horizon is random and forgetting comes from absorption rather than from discounting, is the subject of Section 8.7.

7.2 Infinite horizon and discounting

A *discounted MDP* is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$: finite state and action spaces, a *stationary* kernel $P(s' \mid s,a)$ and reward function $r(s,a,s')$ with $|r| \le R_{\max}$, and a discount factor $\gamma \in (0,1)$. Time

runs over $t = 1, 2, \dots$ without end. Policies are defined as in Chapter 1, with histories of arbitrary length: Π_{H} denotes the history-dependent randomized policies, Π_{MR} the Markov ones, $\pi = (\pi_t)_{t\geq 1}$, and we single out the *stationary* policies, whose decision rule does not depend on time: Π_{SR} (stationary randomized, $\pi_t = \pi$ for all t) and $\Pi_{\mathrm{SD}} \subset \Pi_{\mathrm{SR}}$ (stationary deterministic, $\pi : \mathcal{S} \to \mathcal{A}$). The law $\mathbb{P}_\mu^\pi$ of the infinite trajectory $(S_1, A_1, S_2, A_2, \dots)$ is given by the Ionescu-Tulcea theorem (Theorem A.4); its finite-dimensional marginals are the product formula (1.1), so that the Markov and restart properties of Proposition 1.15 hold verbatim, with the conditional law of $(S_t, A_t, S_{t+1}, \dots)$ given $H_t = h_t$ equal to $\mathbb{P}_{t,s_t}^\pi$ for Markov policies—and equal to $\mathbb{P}_{s_t}^\pi$, the law from s_t at time 1, for stationary ones.

Definition 7.2 (Discounted return). The *discounted return* and the reward-to-go from time t are

$$W \coloneqq \sum_{t=1}^{\infty} \gamma^{t-1} R_t, \qquad W_{t:\infty} \coloneqq \sum_{k=t}^{\infty} \gamma^{k-t} R_k, \qquad R_t = r(S_t, A_t, S_{t+1}),$$

so that $W = W_{1:\infty}$ and $W_{t:\infty} = R_t + \gamma W_{t+1:\infty}$. The series converge absolutely, and $|W_{t:\infty}| \leq R_{\max}/(1-\gamma)$. For $\pi \in \Pi_{\mathrm{SR}}$ the *return distribution* and the *value function* are $\nu^\pi(s) \coloneqq \mathcal{L}_{\mathbb{P}_s^\pi}(W)$ and $V^\pi(s) \coloneqq \mathbb{E}_s^\pi[W]$, and $\nu^\pi(s,a)$, $Q^\pi(s,a)$ are defined likewise with the first action imposed.

For a stationary policy, the restart property gives that the conditional law of $W_{t:\infty}$ given $H_t = h_t$ is $\nu^\pi(s_t)$, whatever t: return distributions no longer carry a time index. This is the reward of stationarity, and the reason why the recursions of this chapter are fixed-point equations rather than backward inductions.

The discount factor admits three readings. First, a *preference for the present*: a reward received t steps from now is worth γ^{t-1} times the same reward received now, as with an interest rate $\gamma^{-1} - 1$. Second, an *effective horizon*: the tail $\sum_{t>T} \gamma^{t-1} R_t$ is at most $\gamma^T R_{\max}/(1-\gamma)$ in absolute value, negligible for T of order $\ln(1/(\varepsilon(1-\gamma)))/(1-\gamma)$, so that $1/(1-\gamma)$ plays the role of the horizon H: with $\gamma = 0.99$, what happens after 917 weeks (about eighteen years) changes the discounted total by at most one hundredth of the largest weekly reward, and 917 is the smallest integer for which the bound $\gamma^T R_{\max}/(1-\gamma)$ certifies it, the exact threshold being $\ln(10^4)/\ln(1/0.99) \approx 916.4$ (Exercise 7.3). Third, a *random horizon*.

Proposition 7.3 (Discounting as a geometric horizon). Let τ be independent of the trajectory with $\mathbb{P}(\tau \geq t) = \gamma^{t-1}$ for $t \geq 1$ (a geometric variable: at each step the process continues with probability γ). Then

$$\mathbb{E}_s^\pi\Big[\sum_{t=1}^{\tau} R_t\Big] = \mathbb{E}_s^\pi\Big[\sum_{t=1}^{\infty} \gamma^{t-1} R_t\Big] = V^\pi(s).$$

The *laws* of $\sum_{t\leq\tau} R_t$ and of W differ in general: for a constant reward $r \equiv 1$, $W = 1/(1-\gamma)$ is deterministic while $\sum_{t\leq\tau} 1 = \tau$ has variance $\gamma/(1-\gamma)^2$.

Proof. $\mathbb{E}[\sum_{t\leq\tau} R_t] = \sum_t \mathbb{E}[R_t \mathbb{1}\{\tau \geq t\}] = \sum_t \mathbb{E}[R_t]\mathbb{P}(\tau \geq t)$ by independence and Fubini (the sum converges absolutely since $|R_t| \leq R_{\max}$). ∎

The discounted criterion is thus, *in expectation*, the undiscounted criterion of an agent who may be interrupted at any time with probability $1-\gamma$; but an agent interested in the distribution of her return must decide which of the two she means, and this book takes W. How much the two laws differ is measured by Proposition 7.15 below.

Example 7.4 (Three discounted problems).

(i) *The bike store* of Example 1.3 runs week after week: in the hands-on version (session H04), $M = 15$, the weekly demand is geometric with mean 9 (truncated at M), the selling price is 1, the purchase cost 0.5 per bike plus 0.8 per order, the storage cost 0.3 per bike in stock at the start of the week, and $\gamma = 0.99$. A natural family of stationary policies are the *order-up-to* rules: order enough to bring the stock to a level m whenever it is below m.

(ii) *The machine replacement* of Example 1.8 with a fixed replacement cost and a maintenance cost growing with the wear: in session H05, $\gamma = 0.9$, the replacement costs 150, the annual maintenance costs $4x$ at wear x plus surcharges on some wear ranges, and the wear increases by an exponential amount of mean 1 each year. With a nondecreasing maintenance cost, the discounted counterparts of the finite-horizon thresholds of Theorem 3.22 converge to a single stationary threshold when the limiting comparison of the two actions crosses zero at a single wear level (Exercise 8.9; Example 8.15, for the simplified costs $c(x) = x$, $C = 4$ of Theorem 3.22); the surcharges of H05 make the cost non-monotone, and the optimal replacement region is then a union of intervals rather than a half-line.

(iii) *The riverswim* of Example 1.5 with a discount factor: the swimmer collects rewards for ever, and the question is whether the far-away large reward is worth the effort, which depends on γ.

7.3 The distributional Bellman equation

Discounting replaces the shift b_r by the *bootstrap map* $b_{r,\gamma}(x) \coloneqq r + \gamma x$, which shifts and *contracts*. Everything else is as in Chapter 3.

Theorem 7.5 (Distributional Bellman equation, discounted case). For $\pi \in \Pi_{\mathrm{SR}}$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$\nu^\pi(s,a) = \sum_{s' \in \mathcal{S}} P(s' \mid s,a)\ (b_{r(s,a,s'),\gamma})_{\#}\nu^\pi(s'), \qquad \nu^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s)\, \nu^\pi(s,a). \tag{7.2}$$

Proof. Under $\mathbb{P}^\pi_{s,a}$, $W = r(s,a,S_2) + \gamma W_{2:\infty}$, and by the restart property the conditional law of $W_{2:\infty}$ given $S_2 = s'$ is $\nu^\pi(s')$; the law of $r + \gamma W_{2:\infty}$ given $S_2 = s'$ is therefore $(b_{r,\gamma})_{\#}\nu^\pi(s')$. Mix over s', then over a, as in the proof of Theorem 3.2. ∎

Definition 7.6 (Distributional Bellman operator). The *distributional Bellman operator* of $\pi \in \Pi_{\mathrm{SR}}$ is the map $\mathcal{T}^\pi : \mathcal{M}_1(\mathbb{R})^{\mathcal{S}} \to \mathcal{M}_1(\mathbb{R})^{\mathcal{S}}$,

$$(\mathcal{T}^\pi \nu)(s) \coloneqq \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s,a)\ (b_{r(s,a,s'),\gamma})_{\#}\nu(s').$$

Theorem 7.5 says that ν^π is a *fixed point* of $\mathcal{T}^\pi$: $\nu^\pi = \mathcal{T}^\pi \nu^\pi$.

A fixed-point equation may have several solutions or none; the following theorem, due to Bellemare et al. (2017), settles the matter. Recall the Wasserstein distances $\mathcal{W}_p$ and their supremum extension $\overline{\mathcal{W}}_p(\nu,\nu') = \max_s \mathcal{W}_p(\nu(s),\nu'(s))$ from Appendix D, and let $\mathcal{P}_\gamma \subset \mathcal{M}_1(\mathbb{R})^{\mathcal{S}}$ be the set of return-distribution functions whose components are supported in $I_\gamma \coloneqq [-R_{\max}/(1-\gamma), R_{\max}/(1-\gamma)]$.

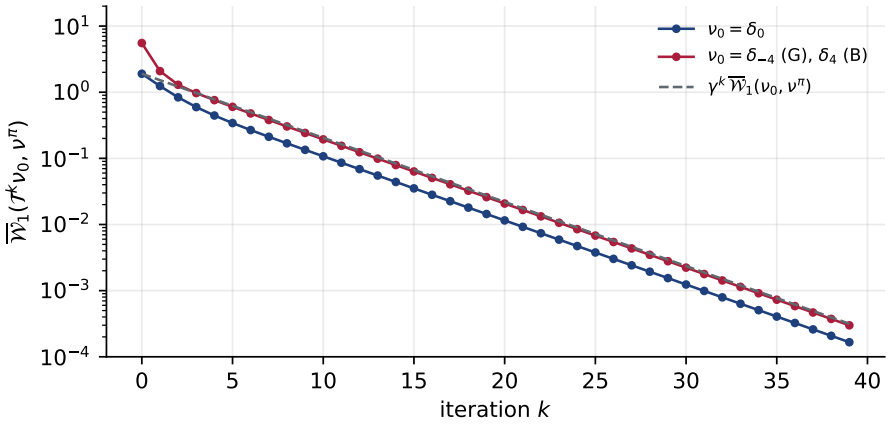


Figure 7.2. Distance $\overline{W}_1((\mathcal{T}^\pi)^k \nu_0, \nu^\pi)$ between the iterates of the distributional Bellman operator and its fixed point, for the two-state chain of Example 3.5 with $\gamma = 0.8$ and two initial conditions (log scale). The dashed line is the bound $\gamma^k \overline{W}_1(\nu_0, \nu^\pi)$ of Theorem 7.7.

Theorem 7.7 (Contraction and fixed point). For every $p \in [1, \infty)$ and $\nu, \nu' \in \mathcal{M}_1(\mathbb{R})^{\mathcal{S}}$ with finite p-th moments,

$$\overline{W}_p(\mathcal{T}^\pi \nu, \mathcal{T}^\pi \nu') \;\le\; \gamma\, \overline{W}_p(\nu, \nu').$$

Consequently ν^π is the unique fixed point of $\mathcal{T}^\pi$ in $\mathcal{P}_\gamma$ (and the unique fixed point among return-distribution functions with finite p-th moments), $\mathcal{T}^\pi$ maps $\mathcal{P}_\gamma$ to itself, and for every $\nu_0 \in \mathcal{P}_\gamma$ the iterates $\nu_k \coloneqq (\mathcal{T}^\pi)^k \nu_0$ satisfy

$$\overline{W}_p(\nu_k, \nu^\pi) \;\le\; \gamma^k\, \overline{W}_p(\nu_0, \nu^\pi) \;\le\; \gamma^k\, \frac{2R_{\max}}{1-\gamma}.$$

Proof. By Lemma D.7(i), $\mathcal{W}_p((b_{r,\gamma})_\# \mu, (b_{r,\gamma})_\# \mu') = \gamma \mathcal{W}_p(\mu, \mu')$ for any μ, μ'; by Lemma D.7(ii), a mixture with common weights satisfies $\mathcal{W}_p(\sum q_i \mu_i, \sum q_i \mu_i')^p \le \sum q_i \mathcal{W}_p(\mu_i, \mu_i')^p \le \max_i \mathcal{W}_p(\mu_i, \mu_i')^p$. Applied to $(\mathcal{T}^\pi \nu)(s)$, a mixture over (a, s') of the push-forwards of $\nu(s')$ and $\nu'(s')$ by the same maps, this gives $\mathcal{W}_p((\mathcal{T}^\pi \nu)(s), (\mathcal{T}^\pi \nu')(s)) \le \gamma \max_{s'} \mathcal{W}_p(\nu(s'), \nu'(s'))$. If $\nu \in \mathcal{P}_\gamma$, each $(\mathcal{T}^\pi \nu)(s)$ is a mixture of laws supported in $r + \gamma I_\gamma \subset I_\gamma$ (since $|r| + \gamma R_{\max}/(1-\gamma) \le R_{\max}/(1-\gamma)$), so $\mathcal{T}^\pi \mathcal{P}_\gamma \subset \mathcal{P}_\gamma$. On laws supported in the compact interval I_γ, $\mathcal{W}_p$ metrizes weak convergence and the set of such laws is weakly compact, so $(\mathcal{P}_\gamma, \overline{W}_p)$ is a complete metric space (Proposition D.6); Banach's fixed-point theorem (Theorem D.16) gives a unique fixed point in $\mathcal{P}_\gamma$, which is ν^π by Theorem 7.5, and the geometric convergence of the iterates. Uniqueness among laws with finite p-th moments follows from the contraction inequality applied to two fixed points. Finally $\mathcal{W}_p(\nu_0(s), \nu^\pi(s)) \le \operatorname{diam} I_\gamma = 2R_{\max}/(1-\gamma)$ for $\nu_0 \in \mathcal{P}_\gamma$. ∎

The theorem is the engine of the infinite-horizon theory (Figure 7.2): the return distributions can be computed by *iterating* the operator from any starting point—say $\nu_0 = \delta_0$—as many times as the desired accuracy requires, about $\ln(2/\varepsilon)/(1-\gamma)$ iterations for accuracy $\varepsilon R_{\max}/(1-\gamma)$ (the bound $\gamma^k\, 2R_{\max}/(1-\gamma)$ of the theorem, with $\ln(1/\gamma) \ge 1-\gamma$). Three remarks temper the enthusiasm.

Remark 7.8 (Limits of the contraction).

(i) *The metric matters.* The contraction fails in total variation and in relative entropy (Appendix D.4): the bootstrap map moves the atoms without merging them, and two distinct Dirac masses stay at total-variation distance 1 for ever (Exercise 7.6). It holds for the Cramér

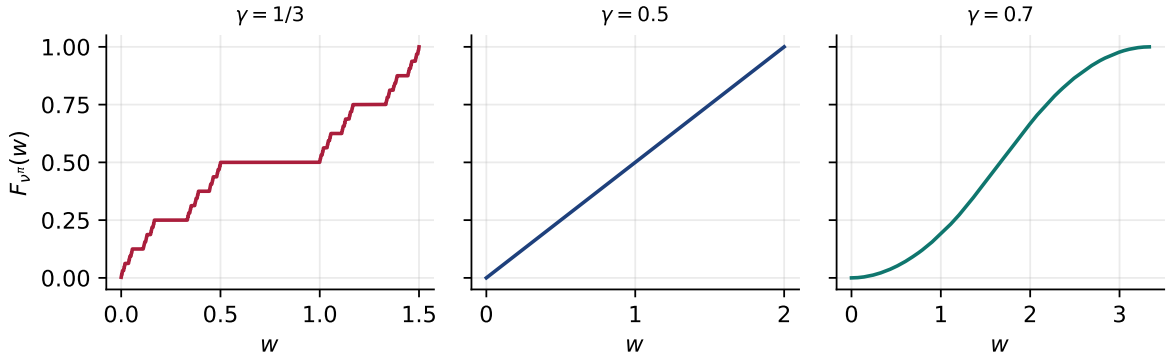


Figure 7.3. Distribution function of the discounted return of Example 7.9 (i.i.d. fair Bernoulli rewards) for three discount factors: a devil's staircase for $\gamma = 1/3$, the uniform law on $[0, 2]$ for $\gamma = 1/2$, an apparently smooth law for $\gamma = 0.7$ (absolute continuity holds for almost every $\gamma > 1/2$). The first two cases are Figure 2.3 of Bellemare et al. (2023); the third is the regime $\gamma > 1/2$.

distance ℓ_2 (Definition D.8) with modulus $\sqrt{\gamma}$ (Lemma D.9), and for $\mathcal{W}_\infty$ with modulus γ (Exercise 7.4).

(ii) *The iterates cannot be represented exactly.* Starting from δ_0, the iterate $\nu_k(s)$ has up to $(SA)^k$ atoms, and the fixed point typically has infinitely many (Example 7.9). Each iteration must therefore be followed by a *projection* on a finite family of laws (Section 7.6), at the price of a bias that the contraction keeps under control.

(iii) *The rate is γ, not better.* For the risk-neutral value the same rate holds, so that nothing is lost by propagating distributions instead of values; but for $\gamma = 0.99$, several hundred iterations are needed in both cases, and the direct solution of a linear system (Theorem 7.11) is preferable when S is moderate.

Example 7.9 (Bernoulli convolutions: fractal return distributions, after Bellemare et al., 2023, Examples 2.10–2.11 ♦). Let $\mathcal{S} = \{0, 1\}$ with a single action, $P(s' \mid s) = \frac{1}{2}$ for all s, s' and $r(s, a, s') = s'$: the rewards R_t are i.i.d. Bernoulli($\frac{1}{2}$), and $W = \sum_t \gamma^{t-1} R_t$ has the same law ν from both states. The fixed-point equation reads $\nu = \frac{1}{2}(b_{0,\gamma})_\# \nu + \frac{1}{2}(b_{1,\gamma})_\# \nu$: the law is a mixture of two contracted copies of itself, a *self-similar* measure (Figure 7.3). For $\gamma = \frac{1}{2}$, $W = 2\sum_t R_t 2^{-t}$ is twice a number whose binary digits are i.i.d. fair bits: ν is the uniform law on $[0, 2]$. For $\gamma < \frac{1}{2}$, the two copies γK and $1 + \gamma K$ of the support $K \subset [0, 1/(1-\gamma)]$ are disjoint, so K is a Cantor set of Lebesgue measure zero and ν has no density and no atoms: its distribution function is a devil's staircase. For $\gamma > \frac{1}{2}$ the copies overlap and ν has a density for almost every γ (Solomyak, 1995), but not for all: for $\gamma = (\sqrt{5} - 1)/2$ it is singular (Erdős, 1939); see Peres et al. (2000). The mean $1/(2(1-\gamma))$ and the variance $1/(4(1-\gamma^2))$ are elementary in all cases (Exercise 7.5). The example is a warning: the return distribution of a discounted MDP is a rich object, which no finite recursion represents exactly—in contrast with the lattice-supported returns of Proposition 3.7.

Proposition 7.10 (Extreme values). For $\pi \in \Pi_{\mathrm{SR}}$, the functions $\underline{x}^\pi(s) = \operatorname{ess\,inf} \nu^\pi(s)$ and

$\overline{x}^{\pi}(s) = \operatorname{ess\,sup} v^{\pi}(s)$ are the unique solutions of

$$\underline{x}(s) = \min\big\{r(s,a,s') + \gamma\,\underline{x}(s') : \pi(a \mid s)P(s' \mid s,a) > 0\big\},$$
$$\overline{x}(s) = \max\big\{r(s,a,s') + \gamma\,\overline{x}(s') : \pi(a \mid s)P(s' \mid s,a) > 0\big\},$$

and the corresponding $(\min,+)$ and $(\max,+)$ operators are γ-contractions for the supremum norm.

Proof. The support of $(b_{r,\gamma})_{\#}\mu$ is $r + \gamma \operatorname{supp}\mu$ and the support of a mixture with positive weights is the union of the supports; taking infima in (7.2) gives the equations. The map $x \mapsto (\min_{(a,s')}[r + \gamma x(s')])_s$ is a γ-contraction because $|\min_i u_i - \min_i u'_i| \le \max_i |u_i - u'_i|$; uniqueness follows. ∎

7.4 The expected Bellman operator

Integrating (7.2) against the identity gives the classical equations. Write $r^{\pi}(s) \coloneqq \sum_{a,s'} \pi(a \mid s)P(s' \mid s,a)r(s,a,s')$ for the expected one-step reward and $P^{\pi}(s,s') \coloneqq \sum_a \pi(a \mid s)P(s' \mid s,a)$ for the transition matrix of the state chain under π; V^{π} and r^{π} are vectors indexed by $\mathcal{S}$.

Theorem 7.11 (Bellman evaluation equation and its solution). For $\pi \in \Pi_{\mathrm{SR}}$, V^{π} is the unique solution of $V = r^{\pi} + \gamma P^{\pi} V$, namely

$$V^{\pi} = (I - \gamma P^{\pi})^{-1} r^{\pi} = \sum_{t=0}^{\infty} \gamma^t (P^{\pi})^t r^{\pi},$$

and $Q^{\pi}(s,a) = \sum_{s'} P(s' \mid s,a)[r(s,a,s') + \gamma V^{\pi}(s')]$. The *Bellman operator* $T^{\pi}V \coloneqq r^{\pi} + \gamma P^{\pi} V$ on $\mathbb{R}^{\mathcal{S}}$ is affine, monotone ($V \le V'$ implies $T^{\pi}V \le T^{\pi}V'$) and a γ-contraction for the supremum norm: $\|T^{\pi}V - T^{\pi}V'\|_{\infty} \le \gamma\|V - V'\|_{\infty}$. Its iterates satisfy $\|(T^{\pi})^k V_0 - V^{\pi}\|_{\infty} \le \gamma^k \|V_0 - V^{\pi}\|_{\infty}$, and for every V,

$$\|V - V^{\pi}\|_{\infty} \;\le\; \frac{\|T^{\pi}V - V\|_{\infty}}{1-\gamma}. \tag{7.3}$$

Proof. Taking means in (7.2) (the mean of $(b_{r,\gamma})_{\#}\mu$ is $r + \gamma \int x \,\mathrm{d}\mu$) gives $V^{\pi} = r^{\pi} + \gamma P^{\pi} V^{\pi}$. Since P^{π} is stochastic, $\|\gamma P^{\pi} V\|_{\infty} \le \gamma\|V\|_{\infty}$, so the Neumann series $\sum_t (\gamma P^{\pi})^t$ converges in operator norm to $(I - \gamma P^{\pi})^{-1}$ and the solution is unique. Monotonicity holds because P^{π} has nonnegative entries, and the contraction property is $\|\gamma P^{\pi}(V - V')\|_{\infty} \le \gamma\|V - V'\|_{\infty}$. For (7.3): $\|V - V^{\pi}\| \le \|V - T^{\pi}V\| + \|T^{\pi}V - T^{\pi}V^{\pi}\| \le \|V - T^{\pi}V\| + \gamma\|V - V^{\pi}\|$. ∎

The series expresses $V^{\pi}(s)$ as the discounted sum of the expected rewards along the state chain, $\sum_t \gamma^t \mathbb{E}^{\pi}_s[r^{\pi}(S_{t+1})]$—the definition again, read through the Markov property. Equation (7.3) turns the iteration into an algorithm with a certificate (Algorithm 7.1): stopping when two successive iterates differ by less than $\varepsilon(1-\gamma)/\gamma$ guarantees that the last iterate is within ε of V^{π} (Exercise 7.10). Solving the linear system directly costs $O(S^3)$ operations; each iteration costs $O(S^2)$ (or $O(S \cdot \text{support of } P)$), and $\ln(R_{\max}/(\varepsilon(1-\gamma)))/(1-\gamma)$ iterations suffice.

Example 7.12 (The two-state chain, discounted). For the chain of Example 3.5 ($P(\mathrm{G} \mid \mathrm{G}) = \frac{2}{3}$, $P(\mathrm{G} \mid \mathrm{B}) = \frac{1}{2}$; reward 1 on G $\to$ G, -1 on B $\to$ B, 0 otherwise) with $\gamma = 0.8$, the equations $V(\mathrm{G}) = \frac{2}{3} + \frac{4}{5}(\frac{2}{3}V(\mathrm{G}) + \frac{1}{3}V(\mathrm{B}))$ and $V(\mathrm{B}) = -\frac{1}{2} + \frac{4}{5}(\frac{1}{2}V(\mathrm{G}) + \frac{1}{2}V(\mathrm{B}))$ give $V(\mathrm{G}) = \frac{20}{13} \approx 1.54$ and $V(\mathrm{B}) = \frac{5}{26} \approx 0.19$. The fixed point v^{π} has these means; its variances, 2.74 and 2.88, are computed

Algorithm 7.1: Iterative policy evaluation

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

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

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

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

Output: V', with $\|V' - V^\pi\|_\infty \le \varepsilon$

in Example 7.14.

For history-dependent policies $\pi \in \Pi_{\mathrm{H}}$, we define as in Section 3.3 the *value of a history*, $V^\pi(h_t) \coloneqq \mathbb{E}[W_{t:\infty} \mid H_t = h_t]$, the expectation of the reward-to-go under the continuation of h_t by π; the recursion $V^\pi(h_t) = \sum_{a,s'} \pi_t(a \mid h_t) P(s' \mid s_t, a)[r(s_t,a,s') + \gamma V^\pi(h_t,a,s')]$ holds by the tower property, and for $\pi \in \Pi_{\mathrm{SR}}$ the restart property gives $V^\pi(h_t) = V^\pi(s_t)$. It is with these history values that Chapter 8 proves that stationary deterministic policies are optimal among all policies, through the truncated version of the recursion.

7.5 Statistics of the discounted return

What becomes of the Bellman-closed statistics of Chapter 4 under discounting? The bootstrap map acts on a moment of order i through $(r+\gamma x)^i$, and the factor γ^j in front of x^j is harmless: moments remain closed. It acts on an exponential moment through $e^{\beta(r+\gamma x)} = e^{\beta r} e^{(\gamma\beta)x}$, and the change of rate is not harmless.

Proposition 7.13 (Moments of the discounted return). For $\pi \in \Pi_{\mathrm{SR}}$ and $i \ge 1$, the moments $M^{(i)}(s) = \int x^i \, \mathrm{d}\nu^\pi(s)$ satisfy

$$M^{(i)}(s) = \sum_{a,s'} \pi(a \mid s) P(s' \mid s,a) \sum_{j=0}^{i} \binom{i}{j} r(s,a,s')^{i-j} \gamma^j \, M^{(j)}(s'), \qquad M^{(0)} \equiv 1.$$

Given $M^{(0)}, \dots, M^{(i-1)}$, the map $M^{(i)} \mapsto$ right-hand side is affine and a γ^i-contraction for the supremum norm, so the first K moments are the unique solution of a triangular system of K linear equations of size S, or the limit of the joint iteration. In particular the variance $\sigma^2(s)$ of $\nu^\pi(s)$ is the unique solution of

$$\sigma^2 = \gamma^2 P^\pi \sigma^2 + c^\pi, \qquad c^\pi(s) \coloneqq \sum_{a,s'} \pi(a \mid s) P(s' \mid s,a)\big[r(s,a,s') + \gamma V^\pi(s') - V^\pi(s)\big]^2, \tag{7.4}$$

that is, $\sigma^2 = (I - \gamma^2 P^\pi)^{-1} c^\pi$.

Proof. Integrate x^i against (7.2) using $\int x^i \, \mathrm{d}((b_{r,\gamma})_\# \mu) = \mathbb{E}(r+\gamma X)^i = \sum_j \binom{i}{j} r^{i-j} \gamma^j \mathbb{E} X^j$ for $X \sim \mu$. The coefficient of $M^{(i)}(s')$ is $\gamma^i P^\pi(s,s')$, whence the contraction. For the variance, condition on (A_1, S_2) as in Corollary 4.2: the conditional variance of $W = R_1 + \gamma W_{2:\infty}$ is $\gamma^2 \sigma^2(S_2)$ and its conditional mean is $R_1 + \gamma V^\pi(S_2)$; the law of total variance gives (7.4), and $I - \gamma^2 P^\pi$ is invertible. ∎

Example 7.14 (Variance in the two-state chain). With $\gamma = 0.8$ and the values of Example 7.12, the one-step targets from G are $1+\gamma V(\mathrm{G})$ and $\gamma V(\mathrm{B})$, and $c^\pi(\mathrm{G}) = \frac{2}{3}(1+\frac{4}{5}\cdot\frac{20}{13} - \frac{20}{13})^2 + \frac{1}{3}(\frac{4}{5}\cdot\frac{5}{26} - \frac{20}{13})^2 \approx$

0.96; similarly $c^\pi(\mathrm{B}) \approx 1.08$. Solving (7.4) gives $\sigma^2(\mathrm{G}) \approx 2.74$ and $\sigma^2(\mathrm{B}) \approx 2.88$, in agreement with the variances of the fixed point computed numerically on a fine grid. The variance of the return is about three times the variance created at each step: the factor $1/(1-\gamma^2) \approx 2.8$ accumulates the dispersion of the future, discounted twice.

The geometric horizon of Proposition 7.3 and the discount give the same values but different laws; the variances show by how much. Write $K \coloneqq \sum_{t \le \tau} R_t$ for the *killed return*, $\sigma_K^2(s)$ for its variance from s, and $\sigma_W^2 = \sigma^2$ for the variance of W, given by (7.4).

Proposition 7.15 (Variance of the killed return). For $\pi \in \Pi_{\mathrm{SR}}$, with c^π as in (7.4),

$$\begin{aligned} \sigma_K^2 &= \gamma P^\pi \sigma_K^2 + c^\pi + \gamma(1-\gamma) P^\pi (V^\pi)^2, \quad \text{that is,} \\ \sigma_K^2 &= (I - \gamma P^\pi)^{-1} \left[c^\pi + \gamma(1-\gamma) P^\pi (V^\pi)^2 \right] \;\ge\; (I - \gamma^2 P^\pi)^{-1} c^\pi = \sigma_W^2, \end{aligned}$$

where $(V^\pi)^2$ is the componentwise square.

Proof. $|K| \le R_{\max}\tau$, so K has a finite variance. Write $K = R_1 + BK'$ with $B \coloneqq \mathbb{1}\{\tau \ge 2\}$, a Bernoulli($\gamma$) variable independent of the trajectory, and $K' \coloneqq \sum_{t=2}^{\tau} R_t$. Given (A_1, S_2) and $B = 1$, the geometric variable $\tau - 1$ has the law of τ (memorylessness) and the trajectory restarts from S_2, so K' has the law of K from S_2. Hence $\mathbb{E}[K \mid A_1, S_2, B] = R_1 + BV^\pi(S_2)$ and $\mathbb{V}\mathrm{ar}(K \mid A_1, S_2, B) = B\sigma_K^2(S_2)$, and the law of total variance gives $\sigma_K^2(s) = \gamma(P^\pi \sigma_K^2)(s) + \mathbb{E}\left[(R_1 + BV^\pi(S_2) - V^\pi(s))^2\right]$. Split $R_1 + BV^\pi(S_2) - V^\pi(s) = [R_1 + \gamma V^\pi(S_2) - V^\pi(s)] + (B-\gamma)V^\pi(S_2)$: the cross term vanishes because B is independent of (A_1, S_2) with mean γ, and $\mathbb{E}(B-\gamma)^2 = \gamma(1-\gamma)$; this is the recursion. For the inequality, $(I - \gamma P^\pi)^{-1} = \sum_t \gamma^t (P^\pi)^t \ge \sum_t \gamma^{2t} (P^\pi)^t = (I - \gamma^2 P^\pi)^{-1}$ entrywise, and the vectors c^π and $P^\pi(V^\pi)^2$ are nonnegative. ∎

Killing loses twice: once in the resolvent, γ instead of γ^2 (the dispersion created at step t is no longer damped by $\gamma^{2(t-1)}$ but only by the survival probability γ^{t-1}), and once in the *kill noise* $\gamma(1-\gamma)P^\pi(V^\pi)^2$, the variance of stopping or not. For a constant reward $c^\pi = 0$ and the formula returns $\gamma/(1-\gamma)^2$. In the two-state chain of Example 7.12 ($\gamma = 0.8$), both returns have mean 1.54 from G, but $\sigma_K^2(\mathrm{G}) = 6.15$ against $\sigma_W^2(\mathrm{G}) = 2.74$ (Example 7.14): 4.98 is the resolvent applied to c^π, and 1.18 the kill noise (`geometric_horizon_variance.py`, which checks these values by simulation). For Monte Carlo evaluation this has two consequences. To estimate V^π, simulating killed episodes gives an unbiased estimator at an expected cost of $1/(1-\gamma)$ steps per episode, and truncating W at a horizon M gives a variance close to σ_W^2 and a bias at most $\gamma^M R_{\max}/(1-\gamma)$ at a cost of M steps; which is cheaper is a matter of budget (on the two-state chain the products variance×cost are 31 for killing and 82 for truncation at $M = 30$, whose bias bound is 0.006). To estimate the *law* of the return, only truncation is valid: $|W - \sum_{t \le M} \gamma^{t-1} R_t| \le \gamma^M R_{\max}/(1-\gamma)$ almost surely, so the truncated returns are within that distance of ν^π in $\mathcal{W}_\infty$, whereas the killed returns sample the law of K, and their empirical quantiles, CVaR or β-means estimate those of K, not of W.

Proposition 7.16 (Exponential moments change rate). For $\pi \in \Pi_{\mathrm{SR}}$ and $\beta \in \mathbb{R}$, let $Z_\beta(s) \coloneqq \int e^{\beta x} \, \mathrm{d}\nu^\pi(s)$. Then

$$Z_\beta(s) = \sum_{a,s'} \pi(a \mid s) P(s' \mid s, a) \, e^{\beta r(s,a,s')} \, Z_{\gamma\beta}(s'), \tag{7.5}$$

and, in terms of the β-values $U_\beta(s) \coloneqq U_\beta(\nu^\pi(s))$, with $(A, S') \sim \pi(\cdot \mid s) \otimes P(\cdot \mid s, \cdot)$,

$$U_\beta(s) = U_\beta\Big(r(s, A, S') + \gamma \, U_{\gamma\beta}(S')\Big). \tag{7.6}$$

Proof. $\int e^{\beta x}\,\mathrm{d}((b_{r,\gamma})_{\#}\mu) = e^{\beta r}\int e^{\gamma\beta x}\,\mathrm{d}\mu$ gives (7.5). For (7.6), condition on (A_1, S_2): the conditional β-mean of $W = R_1 + \gamma W_{2:\infty}$ is $R_1 + U_\beta(\gamma W_{2:\infty} \mid A_1, S_2) = R_1 + \gamma U_{\gamma\beta}(v^\pi(S_2))$ by translation and scaling (Proposition 2.17(iii)), the conditional law of $W_{2:\infty}$ given (A_1, S_2) being $v^\pi(S_2)$; apply the tower property (Proposition 2.17(vii)). ∎

The exponential moment at rate β thus calls for the exponential moment at rate $\gamma\beta$, which calls for the one at rate $\gamma^2\beta$, and so on: no single exponential moment, and no finite family of them, is Bellman closed in the discounted setting (Remark 4.13); the only finite Bellman-closed families of utilities are spanned by moments (Rowland et al., 2019). The countable family $(Z_{\gamma^k\beta})_{k\geq 0}$ is closed, however, and since $\gamma^k\beta \to 0$ it is nearly risk-neutral after a number of steps of order $1/(1-\gamma)$: the recursion (7.6) can be run backward from a horizon T' at which $U_{\gamma^{T'}\beta}$ is replaced by the expected value V^π, at a controlled cost. This is the basis of risk-sensitive planning in discounted MDPs (Chapter 8), following Chung and Sobel (1987) and Hau et al. (2023): a risk-averse agent in a discounted world is, whether she likes it or not, less and less averse about the more and more distant future.

7.6 Finite representations and projections

Remark 7.8(ii) left the distributional iteration with a practical obstacle: the iterates of $\mathcal{T}^\pi$ have exponentially many atoms and its fixed point typically infinitely many, so that return distributions cannot be iterated exactly. Distributional reinforcement learning therefore works with a finite-dimensional class $C \subset \mathcal{M}_1(\mathbb{R})$ and a projection $\Pi : \mathcal{M}_1(\mathbb{R}) \to C$ (applied componentwise to return-distribution functions), and iterates $\Pi\mathcal{T}^\pi$ instead of $\mathcal{T}^\pi$; the questions are whether the projected iteration still converges, and how far its limit is from v^π. Two classes dominate the practice.

- The *categorical* representation (Bellemare et al., 2017): laws supported on a fixed grid $z_1 < \cdots < z_N$, $C = \{\sum_i p_i\delta_{z_i}\}$. The categorical projection Π_{C} sends a Dirac mass δ_y with $z_i \leq y \leq z_{i+1}$ to $\frac{z_{i+1}-y}{z_{i+1}-z_i}\delta_{z_i} + \frac{y-z_i}{z_{i+1}-z_i}\delta_{z_{i+1}}$ (masses outside the grid go to the nearest endpoint), and extends by linearity. It preserves the mean of laws inside the grid, and it is the orthogonal projection on C for the Cramér distance ℓ_2 (Rowland et al., 2018): ℓ_2 is the $L^2(\,\mathrm{d}x)$ distance between distribution functions (Definition D.8), so that, through $F \mapsto F - \mathbb{1}_{[0,\infty)}$, the laws supported in a bounded interval embed isometrically in the Hilbert space $L^2(\,\mathrm{d}x)$, where C—whose distribution functions are the nondecreasing step functions with jumps on the grid—is a closed convex set.
- The *quantile* representation (Dabney et al., 2018b): laws with N equally weighted atoms at arbitrary positions, $C = \{\frac{1}{N}\sum_i \delta_{\theta_i}\}$. The quantile projection Π_{Q} places the atoms at the quantiles $F^{-1}((2i-1)/(2N))$ (Exercise 4.12), and minimizes $\mathcal{W}_1$ over C. When such a quantile is not unique (a flat piece of F at that level), any point between the lower and the upper quantile minimizes $\mathcal{W}_1$ as well; we fix the convention of the generalized inverse F^{-1}, the lower quantile, so that Π_{Q} is a well-defined map.

The projected operator $\Pi\mathcal{T}^\pi$ is analyzed by a general principle, which has nothing to do with distributions and will serve again for approximate dynamic programming in Chapter 13.

Proposition 7.17 (Projected fixed points). Let $(\mathcal{X}, d)$ be a complete metric space, $T : \mathcal{X} \to \mathcal{X}$ a γ'-contraction ($\gamma' < 1$) with fixed point $x^\star$, $C \subset \mathcal{X}$ a closed subset and $\Pi : \mathcal{X} \to C$ a nonexpansive map, $d(\Pi x, \Pi x') \leq d(x, x')$. Then ΠT is a γ'-contraction of C with a unique fixed point $\hat{x} \in C$, the iterates $(\Pi T)^k x_0$ converge to it geometrically, and

$$d(\hat{x}, x^\star) \;\leq\; \frac{d(\Pi x^\star, x^\star)}{1-\gamma'}.$$

Proof. $d(\Pi Tx, \Pi Tx') \le d(Tx, Tx') \le \gamma' d(x, x')$, and Banach's theorem (Theorem D.16) on the complete space (C, d) gives the fixed point and the geometric convergence. Then $d(\hat{x}, x^\star) \le d(\Pi T\hat{x}, \Pi Tx^\star) + d(\Pi x^\star, x^\star) \le \gamma' d(\hat{x}, x^\star) + d(\Pi x^\star, x^\star)$, using $Tx^\star = x^\star$. ∎

Here $\mathcal{X} = \mathcal{P}_\gamma$ is the set of return-distribution functions supported in I_γ, C^S is the class of representable functions, $\hat{x} = \hat{v}^\pi$ is the projected fixed point, and the metric d depends on the representation; the price of the representation is in both cases the projection error of the true fixed point, amplified by the horizon factor $1/(1-\gamma')$. The hypotheses of the proposition are checked as follows.

- *Categorical:* $d = \bar{\ell}_2$, the supremum Cramér distance, for which $(\mathcal{P}_\gamma, \bar{\ell}_2)$ is complete (on laws supported in I_γ, ℓ_2 and $\mathcal{W}_1$ define the same topology by Lemma D.9, and C^S, a finite-dimensional simplex, is closed); $\mathcal{T}^\pi$ is a $\sqrt{\gamma}$-contraction (Lemma D.9), and Π_C is non-expansive, being an orthogonal projection onto a closed convex set of a Hilbert space (Rowland et al., 2018). The projection error is at most $\sqrt{\Delta z/2}$ for a grid of spacing Δz covering the support (Exercise 7.11), so that $\bar{\ell}_2(\hat{v}^\pi, v^\pi) \le \sqrt{\Delta z/2}/(1-\sqrt{\gamma})$.
- *Quantile:* $d = \overline{\mathcal{W}}_\infty$, for which $\mathcal{P}_\gamma$ is complete and C^S closed (a uniform limit of quantile functions of laws supported in I_γ is again one, and the quantile functions of C are the step functions constant on each $((i-1)/N, i/N]$, a uniformly closed set); $\mathcal{T}^\pi$ is a γ-contraction (Exercise 7.4) and Π_Q is non-expansive (Dabney et al., 2018b), so that $\overline{\mathcal{W}}_\infty(\hat{v}^\pi, v^\pi) \le \overline{\mathcal{W}}_\infty(\Pi_Q v^\pi, v^\pi)/(1-\gamma)$. The projection error in $\mathcal{W}_\infty$ is not small for arbitrary laws (an atom of small mass can be displaced by the whole range), but when the return distributions have L-Lipschitz quantile functions it is of order L/N, and the fixed-point error of order $L/((1-\gamma)N)$ (Bellemare et al., 2023, Chap. 5).

In both cases the factor $1/(1-\gamma')$ is the price of accumulating a projection error at every iteration, exactly as the factor H of Proposition 4.27. Propagating m exact moments and reconstructing (Theorem 4.21) avoids this factor: the moments of the discounted return are computed exactly by Proposition 7.13, and any law with these moments is within $6\Delta/m$ of $v^\pi(s)$ in $\mathcal{W}_1$, where $\Delta = 2R_{\max}/(1-\gamma)$ is the length of I_γ (the range of Chapter 8), whatever γ (Bringas Miranda et al., 2026). The learning versions of these recursions, driven by samples rather than by the known kernel, are the subject of Chapter 12.

Example 7.18 (The two projections at work). Take the two-state chain of Example 7.12 and compute the exact return distribution $v^\pi(\mathrm{G})$ by iterating $\mathcal{T}^\pi$ on a grid fine enough to be a reference (8001 atoms): its mean is $20/13 = 1.5385$ and its variance 2.742, the values of Example 7.14. Figure 7.4 (left) shows its distribution function together with the fixed points of the two projected iterations at $N = 11$. They approximate it in complementary ways: the categorical fixed point is a law on a fixed grid, so its distribution function jumps at prescribed places and by varying amounts; the quantile fixed point has prescribed jumps of size $1/N$ at places it chooses. Neither is the projection of v^π: each is the fixed point of a perturbed recursion, and Proposition 7.17 is what bounds the difference.

The right panel measures it. Both errors fall like $1/N$: the Cramér error of the categorical fixed point goes from 0.62 at $N = 3$ to 0.0075 at $N = 161$, and the $\mathcal{W}_1$ error of the quantile fixed point from 1.16 to 0.013. For the quantile representation the rate $1/N$ is the one predicted above, only the constant is pessimistic; for the categorical one the observation beats the guarantee: Proposition 7.17 gives $\sqrt{\Delta z/2}/(1-\sqrt{\gamma})$ with $\Delta z = 10/(N-1)$, which decays only like $N^{-1/2}$—the dashed line. The gap is not slack in the proposition but in the input: the bound charges the worst one-step projection error at every iteration and lets the errors accumulate in the worst direction, whereas here successive projections of a smooth law partly cancel. The lesson is the usual one about error propagation, already met with the factor H of Proposition 4.27 and to be met again in Chapter 13: the factor $1/(1-\gamma')$ is real in the worst case and pessimistic on any particular

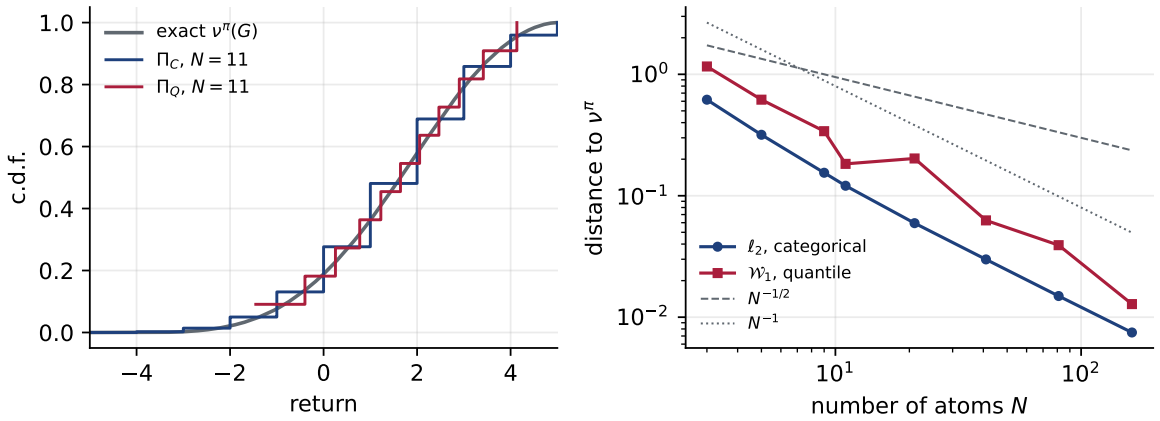


Figure 7.4. The two finite representations on the two-state chain (Example 7.18). Left: distribution function of the exact return distribution and of the fixed points of the two projected iterations, with $N = 11$ atoms. Right: distance from the projected fixed point to ν^π against N, in Cramér distance for the categorical representation and in $\mathcal{W}_1$ for the quantile one, with the reference slopes $N^{-1/2}$ and N^{-1}.

problem.

Distributional viewpoint

Discounting turns the shift $x \mapsto r + x$ into the affine contraction $x \mapsto r + \gamma x$, and the distributional Bellman equation into a fixed-point equation whose operator contracts Wasserstein distances by γ (Theorem 7.7). The expected operator inherits the contraction in supremum norm, and its fixed point is a linear system. The return distribution of a discounted MDP is a rich object—a self-similar measure, fractal for small γ (Example 7.9)—which no finite recursion represents exactly; projections onto finite classes are controlled by the contraction (Proposition 7.17), moments are propagated exactly with factors γ^j, and exponential moments lose their closure: the rate β becomes $\gamma\beta$ at each step, so that risk sensitivity fades with the horizon and risk-sensitive planning becomes non-stationary (Chapter 8).

Hands-on

Session H04 (`retailStore.ipynb`) sets up the bike store of Example 7.4(i) as a discounted MDP: the notebook computes the transition matrix and expected rewards of a policy, evaluates it by solving the linear system of Theorem 7.11, and implements value iteration with the stopping rule of Algorithm 7.1; its final assignment, a *distributional* value iteration approximating the discounted return distribution, is the projected iteration of Section 7.6 (categorical or quantile), whose convergence is Proposition 7.17. Session H05 (`replacement.ipynb`) has a continuous wear level: the first task is to discretize it into a finite discounted MDP, to which this chapter applies. Both notebooks also contain the learning algorithms of Chapter 12.

Exercises

Exercise 7.1 (Parity and the horizon). [limits] Three states c, a, b. From a the only move leads to b and from b to a; entering b yields the reward 2, entering a the reward 0. From the initial state c two actions are available: enter a (reward 0) or enter b at a cost 1 (reward $2-1=1$). (a) Without discount, compute the total reward of each first action with k steps to go, and show that the optimal first decision alternates with the parity of k for ever; what do the increments $V_k(\mathrm{a})-V_{k-1}(\mathrm{a})$ do? (b) With a discount factor $\gamma<1$, show that entering b is strictly better for every odd k, and for even k if and only if $\gamma^k<(1-\gamma)/2$: the rule stabilizes from $k\approx\ln(2/(1-\gamma))/\ln(1/\gamma)$ on ($k=29$ for $\gamma=0.9$). Compare with the bound k_0 of Proposition 7.1. (c) Compute the two infinite-horizon discounted values from c and their difference. Which policy is optimal for every $\gamma<1$? Expand the two values in powers of $1-\gamma$ and explain why the undiscounted problem cannot decide between them.

Exercise 7.2 (Geometric horizon). [verify] (a) Prove the inequality $\sigma_K^2\geq\sigma_W^2$ of Proposition 7.15 directly, without the recursion (*hint:* condition on the trajectory). (b) Show that the two laws coincide if and only if $R_t=0$ almost surely for all $t\geq 2$. (c) An agent stops with probability $1-\gamma$ at each step *and* discounts with factor γ': express its expected return as a discounted return. (d) Solve the recursion of Proposition 7.15 for the two-state chain and check that $\sigma_K^2(\mathrm{G})=80/13\approx 6.15$; which of the two losses dominates?

Exercise 7.3 (Effective horizon). [verify] (a) Show that $|W-\sum_{t\leq T}\gamma^{t-1}R_t|\leq\gamma^T R_{\max}/(1-\gamma)$ and that this is at most $\varepsilon R_{\max}$ as soon as $T\geq\ln(1/(\varepsilon(1-\gamma)))/\ln(1/\gamma)$; compare $1/\ln(1/\gamma)$ with $1/(1-\gamma)$. (b) For $\gamma=0.99$ and $\varepsilon=0.01$, how many weeks does the bike store owner effectively plan for?

Exercise 7.4 (Contraction for $\mathcal{W}_\infty$). [verify] Let $\mathcal{W}_\infty(\mu,\mu')=\sup_{u\in(0,1)}|F_\mu^{-1}(u)-F_{\mu'}^{-1}(u)|$. Show that $\mathcal{W}_\infty((b_{r,\gamma})_\#\mu,(b_{r,\gamma})_\#\mu')=\gamma\mathcal{W}_\infty(\mu,\mu')$ and that $\mathcal{W}_\infty(\sum_i q_i\mu_i,\sum_i q_i\mu_i')\leq\max_i\mathcal{W}_\infty(\mu_i,\mu_i')$ (*hint:* couple the mixtures through a common index), and deduce that $\mathcal{T}^\pi$ is a γ-contraction for $\overline{\mathcal{W}}_\infty$.

Exercise 7.5 (Bernoulli convolutions, after Bellemare et al., 2023, Exercises 2.6–2.7). [verify] In Example 7.9: (a) prove that ν is the uniform law on $[0,2]$ when $\gamma=\frac12$; (b) for $\gamma<\frac12$, show that the support K satisfies $K=\gamma K\cup(1+\gamma K)$ with disjoint pieces, deduce that its Lebesgue measure is zero, and that ν has no atoms; (c) compute the mean and the variance for every γ; (d) show that $U_\beta(\nu)=\sum_{t\geq1}\gamma^{t-1}U_{\gamma^{t-1}\beta}(\mathrm{Ber}(\frac12))$ and interpret with Proposition 7.16.

Exercise 7.6 (No contraction in total variation). [limits] Consider a single state with a single action and reward 1, so that $\mathcal{T}^\pi\mu=(b_{1,\gamma})_\#\mu$. Show that $\|(\mathcal{T}^\pi)^k\delta_0-(\mathcal{T}^\pi)^k\delta_\varepsilon\|_{\mathrm{TV}}=1$ for all k and $\varepsilon>0$, while $\mathcal{W}_1((\mathcal{T}^\pi)^k\delta_0,(\mathcal{T}^\pi)^k\delta_\varepsilon)=\gamma^k\varepsilon$. What does the sequence $(\mathcal{T}^\pi)^k\delta_0$ converge to, and in which sense?

Exercise 7.7 (Variance of the discounted return). [verify] (a) Derive (7.4) from the second-moment recursion of Proposition 7.13. (b) Verify the numbers of Example 7.14. (c) Show that

$\sigma^2(s) = \sum_{t\ge 0} \gamma^{2t} \mathbb{E}_s^\pi[c^\pi(S_{t+1})]$ and compare with (4.2). (d) For the Bernoulli convolution, check that (7.4) gives $1/(4(1-\gamma^2))$.

Exercise 7.8 (Exponential moments and rates). [limits] (a) Show that for $\beta \neq 0$ the utility $\nu \mapsto \int e^{\beta x}\, \mathrm{d}\nu$ is not Bellman closed for the discounted operator: exhibit two laws with the same exponential moment at rate β whose push-forwards by $b_{0,\gamma}$ have different exponential moments at rate β. (b) Show that the family $\{\int e^{\gamma^k \beta x}\, \mathrm{d}\nu : k \ge 0\}$ is closed. (c) Suppose the rates $\beta_t = \gamma^{t-1}\beta$ (the notation of Section 8.5) are used for $t \le T'$ and $U_{\beta_{T'+1}}$ is replaced by V^π in (7.6). Using Hoeffding's lemma (Lemma 2.10), bound the error committed on $U_\beta(s)$ in terms of $\gamma^{T'}$, β and $R_{\max}/(1-\gamma)$.

Exercise 7.9 (Stochastic shortest paths: an m-step argument). [verify] Let $\gamma = 1$, and suppose that a goal state s_{goal} is absorbing with zero reward and that, under π, it is reached from every state within m steps with probability at least $\eta > 0$ (a reaching probability in the sense of Proposition 3.12, with $B = \{s_{\mathrm{goal}}\}$ and horizon m). (a) Show that $(P^\pi)^m$ restricted to $\mathcal{S} \setminus \{s_{\mathrm{goal}}\}$ has row sums at most $1-\eta$, that the Neumann series $\sum_t (P^\pi)^t r^\pi$ converges, and that $(T^\pi)^m$ is a $(1-\eta)$-contraction for the supremum norm on $\mathbb{R}^{\mathcal{S}\setminus\{s_{\mathrm{goal}}\}}$, although T^π itself may only be non-expansive. (b) Show that the expected time to the goal satisfies $b^\pi \le m/\eta$, and compare the modulus $(1-\eta)^{1/m}$ per step obtained in (a) with the modulus $1 - 1/\|b^\pi\|_\infty$ of the weighted norm of Proposition 8.29, which needs no m-step argument. Random-length episodes are treated in Section 8.7.

Exercise 7.10 (Q-functions and a posteriori bounds). [verify] (a) Write the fixed-point equation for Q^π on $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and show that its operator is a γ-contraction. (b) Prove the stopping rule of Algorithm 7.1: if $\|V' - V\|_\infty \le \varepsilon(1-\gamma)/\gamma$ with $V' = T^\pi V$, then $\|V' - V^\pi\|_\infty \le \varepsilon$. (c) Bound the number of iterations needed from $V_0 = 0$.

Exercise 7.11 (The categorical projection). [verify] Let $z_1 < \cdots < z_N$ be a grid of spacing Δz covering the support of μ. (a) Show that Π_{C} preserves the mean. (b) Show that $\mathcal{W}_1(\Pi_{\mathrm{C}}\mu, \mu) \le \Delta z/2$ (*hint:* it suffices to treat Dirac masses and to use Lemma D.7(ii)), and deduce $\ell_2(\Pi_{\mathrm{C}}\mu, \mu) \le \sqrt{\Delta z/2}$ from Lemma D.9. (c) Admitting that Π_{C} is non-expansive for ℓ_2 (Rowland et al., 2018), deduce from Proposition 7.17 a bound on $\overline{\ell}_2(\hat{\nu}^\pi, \nu^\pi)$, then on $\overline{\mathcal{W}}_1(\hat{\nu}^\pi, \nu^\pi)$ using Lemma D.9, and compare its dependence on γ with the bound $6\Delta/m$ of Theorem 4.21.

Exercise 7.12 (Numerical study: the bike store). [implement] Build the discounted MDP of Example 7.4(i) (states: stock left, 0 to M; actions: order sizes; geometric demand truncated at M). Evaluate the order-up-to policies with levels $m \in \{5, 8, 10, 12\}$ by solving the linear system and by Algorithm 7.1 with $\varepsilon = 10^{-3}$; report the numbers of iterations, and check the iterative values against the linear solve, which is the exact reference. Compute the variance of the discounted return of the best of these policies by (7.4), and its extreme values by Proposition 7.10.

Exercise 7.13 (Numerical study: the two projections). [implement] Run `projections.py` to reproduce Figure 7.4. The reference ν^π is the discretized fixed point on 8001 atoms of Example 7.18, a numerical reference, not the exact law: its own discretization error must remain well below

the errors measured. (a) Check that the categorical fixed point has the exact mean whenever the grid covers the support, and that the quantile fixed point does not; explain both facts from the definitions of the projections. (b) Measure the one-step projection error $\bar{\ell}_2(\Pi_{\mathrm{C}}\nu^\pi, \nu^\pi)$ and compare with $\sqrt{\Delta z/2}$ and with the observed fixed-point error: which of the two inequalities in the proof of Proposition 7.17 is loose? (c) Repeat with $\gamma = 0.95$ and $\gamma = 0.99$ and check how the error grows with $1/(1-\sqrt{\gamma})$ (the support I_γ widens, and at $\gamma = 0.99$ each fixed point needs of the order of $\ln(10^4)/(1-\gamma) \approx 900$ iterations). (d) Replace the grid of the categorical representation by one concentrated where the mass is, and show that the error improves although the theory does not.

Bibliographical notes

Discounted dynamic programming was formalized by Bellman (1957b) and given its modern contraction-based treatment by Blackwell (1962); Puterman (1994, Chap. 6) and Bertsekas (2017) are the standard references, and Bertsekas and Tsitsiklis (1991) treat the stochastic shortest path problems of Section 8.7. The construction of the trajectory law is the Ionescu-Tulcea theorem (Appendix A).

The distributional Bellman operator and its contraction in Wasserstein distance are due to Bellemare et al. (2017); the Cramér-distance analysis of the categorical projection to Rowland et al. (2018), the quantile representation to Dabney et al. (2018b), and the monograph of Bellemare et al. (2023) develops the theory systematically, including the fixed-point error bounds quoted in Section 7.6. The variance recursion is due to Sobel (1982); that exponential utilities are not Bellman closed under discounting, and the remedy of time-dependent rates, go back to Chung and Sobel (1987), and are the starting point of Hau et al. (2023). Bernoulli convolutions are a classical topic of harmonic analysis and fractal geometry, from Erdős (1939) to Solomyak (1995); Peres et al. (2000) survey the first sixty years.