Distributional Reinforcement Learning

By: Marc G. Bellemare, Will Dabney, Mark Rowland

Citation:
Distributional Reinforcement Learning
By: Marc G. Bellemare, Will Dabney, Mark Rowland
DOI: 10.7551/mitpress/14207.001.0001
ISBN (electronic): 9780262374026
Publisher: The MIT Press
Published: 2023

The MIT Press

8 Statistical Functionals

The development of distributional reinforcement learning in previous chapters has focused on approximating the full return function with parameterized families of distributions. In our analysis, we quantified the accuracy of an algorithm's estimate according to its distance from the true return-distribution function, measured using a suitable probability metric.

Rather than try to approximate the full distribution of the return, we may instead select specific properties of this distribution and directly estimate these properties. Implicitly, this is the approach taken when estimating the expected return. Other common properties of interest include quantiles of the distributions, high-probability tail bounds, and the risk-sensitive objectives described in Chapter 7. In this chapter, we introduce the language of *statistical functionals* to describe such properties.

In some cases, the statistical functional approach allows us to obtain accurate estimates of quantities of interest, in a more straightforward manner. As a concrete example, there is a low-cost dynamic programming procedure to determine the variance of the return distribution.[61] By contrast, categorical and quantile dynamic programming usually under- or overestimate this variance.

This chapter develops the framework of *statistical functional dynamic programming* as a general method for approximately determining the values of statistical functionals. As we demonstrate in Section 8.4, it is in fact possible to interpret both categorical and quantile dynamic programming as operating over statistical functionals. We will see that while some characteristics of the return (including its variance) can be accurately estimated by an iterative procedure, in general, some care must be taken when estimating arbitrary statistical functionals.

61. In fact, the return variance can be determined to machine precision by solving a linear system of equations, similar to what was done in Section 5.1 for the value function.

8.1 Statistical Functionals

A *functional* maps functions to real values. By extension, a *statistical functional* maps probability distributions to the reals. In this book, we view statistical functionals as measuring a particular property or characteristic of a probability distribution. For example, the mapping

$$\nu \mapsto \mathbb{P}_{Z\sim\nu}(Z \geq 0), \quad \nu \in \mathscr{P}(\mathbb{R})$$

is a statistical functional that measures how much probability mass its argument ν puts on the nonnegative reals. Statistical functionals express quantifiable properties of probability distributions such as their mean and variance. The following formalizes this point.

Definition 8.1. A *statistical functional* ψ is a mapping from a subset of probability distributions $\mathscr{P}_\psi(\mathbb{R}) \subseteq \mathscr{P}(\mathbb{R})$ to the reals, written

$$\psi : \mathscr{P}_\psi(\mathbb{R}) \to \mathbb{R}\,.$$

We call the particular scalar $\psi(\nu)$ associated with a probability distribution ν a *functional value* and the set $\mathscr{P}_\psi(\mathbb{R})$ the *domain* of the functional. △

Example 8.2. The *mean functional* maps probability distributions to their expected values. As before, let

$$\mathscr{P}_1(\mathbb{R}) = \{\nu \in \mathscr{P}(\mathbb{R}) : \mathop{\mathbb{E}}_{Z\sim\nu}[|Z|] < \infty\}$$

be the set of distributions with finite first moment. For $\nu \in \mathscr{P}_1(\mathbb{R})$, the mean functional is

$$\mu_1(\nu) = \mathop{\mathbb{E}}_{Z\sim\nu}[Z]\,.$$

The restriction to $\mathscr{P}_1(\mathbb{R})$ is necessary to exclude from the definition distributions without a well-defined mean. △

The purpose of this chapter is to study how functional values of the return distribution can be approximated using dynamic programming procedures and incremental algorithms. In general, we will be interested in a collection of such functionals that exhibit desirable properties: for example, because they can be jointly determined by dynamic programming or because they provide complementary information about the return function. We call such a collection a *distribution sketch*.

Definition 8.3. A distribution sketch (or simply sketch) $\psi : \mathscr{P}_\psi(\mathbb{R}) \to \mathbb{R}^m$ is a vector-valued function specified by a tuple $(\psi_1, \ldots, \psi_m)$ of statistical functionals. Its domain is

$$\mathscr{P}_\psi(\mathbb{R}) = \bigcap_{i=1}^{m} \mathscr{P}_{\psi_i}(\mathbb{R})\,,$$

and it is defined as

$$\psi(\nu) = (\psi_1(\nu), \ldots, \psi_m(\nu)), \quad \nu \in \mathscr{P}_\psi(\mathbb{R}) .$$

Its *image* is

$$I_\psi = \{\psi(\nu) : \nu \in \mathscr{P}_\psi(\mathbb{R})\} \subseteq \mathbb{R}^m .$$

We also extend this notation to return-distribution functions:

$$\psi(\eta) = \big(\psi(\eta(x)) : x \in \mathcal{X}\big), \quad \eta \in \mathscr{P}_\psi(\mathbb{R})^{\mathcal{X}} . \qquad \triangle$$

Example 8.4. The *quantile functionals* are a family of statistical functionals indexed by $\tau \in (0, 1)$ and defined over $\mathscr{P}(\mathbb{R})$. The τ-quantile functional is defined in terms of the inverse cumulative distribution function of its argument (Definition 4.12):

$$\psi^{\mathrm{Q}}_\tau(\nu) = F^{-1}_\nu(\tau) .$$

A finite collection of quantile functionals (say, for $\tau_1, \ldots, \tau_m \in (0, 1)$) constitutes a sketch. $\triangle$

Example 8.5. To prove the convergence of categorical temporal-difference learning (Section 6.10), we introduced the isometry $\mathcal{I} : \mathscr{F}_{\mathrm{C},m} \to \mathbb{R}_{\mathcal{I}}$ defined as

$$\mathcal{I}(\nu) = \Big(F_\nu(\theta_i) : i \in \{1, \ldots, m\}\Big), \tag{8.1}$$

where $(\theta_i)_{i=1}^m$ is the set of locations for the categorical representation. This isometry is also a sketch in the sense of Definition 8.3. If we extend its domain to be $\mathscr{P}(\mathbb{R})$, Equation 8.1 still defines a valid sketch but it is no longer an isometry: it is not possible to recover the distribution ν from its functional values $\mathcal{I}(\nu)$. $\triangle$

8.2 Moments

Moments are an especially important class of statistical functionals. For an integer $p \in \mathbb{N}^+$, the pth moment of a distribution $\nu \in \mathscr{P}_p(\mathbb{R})$ is given by

$$\mu_p(\nu) = \mathop{\mathbb{E}}_{Z \sim \nu}\left[Z^p\right] .$$

In particular, the first moment of ν is its mean, while the variance of ν is the difference between its second moment and squared mean:

$$\mu_2(\nu) - \big(\mu_1(\nu)\big)^2 . \tag{8.2}$$

Moments are ubiquitous in mathematics. They form a natural way of capturing important aspects of a probability distribution, and the infinite sequence of moments $(\mu_p(\nu))_{p=1}^\infty$ uniquely characterizes many probability distributions of interest; see Remark 8.3.

Our goal in this section is to describe a dynamic programming approach to determining the moments of the return distribution. Fix a policy π, and consider a state $x \in \mathcal{X}$ and action $a \in \mathcal{A}$. The pth moment of the return distribution $\eta^\pi(x,a)$ is given by

$$\mathbb{E}_\pi\left[(G^\pi(x,a))^p\right],$$

where as before, $G^\pi(x,a)$ is an instantiation of $\eta^\pi(x,a)$. Although we can also study dynamic programming approaches to learning the pth moment of state-indexed return distributions,

$$\mathbb{E}_\pi\left[(G^\pi(x))^p\right],$$

this is complicated by a potential conditional dependency between the reward R and next state X' due to the action A. One solution is to assume independence of R and X', as we did in Section 5.4. Here, however, to avoid making this assumption, we work with functions indexed by state-action pairs.

To begin, let us fix $m \in \mathbb{N}^+$. The m-moment function M^π is

$$M^\pi(x,a,i) = \mathbb{E}_\pi[(G^\pi(x,a))^i] = \mu_i(\eta^\pi(x,a)), \quad \text{for } i = 1,\dots,m. \tag{8.3}$$

As with value functions, we view M^π as the function (or vector) in $\mathbb{R}^{\mathcal{X}\times\mathcal{A}\times m}$ describing the collection of the first m moments of the random return. In particular, $M^\pi(\cdot,\cdot,1)$ is the usual state-action value function. As elsewhere in the book, to ensure that the expectation in Equation 8.3 is well defined, we assume that all reward distributions have finite pth moments, for $p = 1,\dots,m$. In fact, it is sufficient to assume that this holds for $p = m$ (Assumption 4.29(m)).

As with the standard Bellman equation, from the *state-action random-variable Bellman equation*

$$G^\pi(x,a) = R + \gamma G^\pi(X',A'), \quad X = x, A = a$$

we can derive Bellman equations for the moments of the return distribution. To do so, we raise both sides to the ith power and take expectations with respect to both the random return variables G^π and the random transition $(X = x, A = a, R, X', A')$:

$$\mathbb{E}_\pi[(G^\pi(x,a))^i] = \mathbb{E}_\pi[(R + \gamma G^\pi(X',A'))^i \mid X = x, A = a].$$

From the binomial expansion of the term inside the expectation, we obtain

$$\mathbb{E}_\pi[(G^\pi(x,a))^i] = \mathbb{E}_\pi\left[\sum_{j=0}^{i} \gamma^{i-j}\binom{i}{j} R^j G^\pi(X',A')^{i-j} \,\middle|\, X = x, A = a\right].$$

Since R and $G^\pi(X', A')$ are independent given X and A, we can rewrite the above as

$$M^\pi(x, a, i) = \sum_{j=0}^{i} \gamma^{i-j} \binom{i}{j} \mathbb{E}_\pi[R^j \mid X = x, A = a]\, \mathbb{E}_\pi[M^\pi(X', A', i-j) \mid X = x, A = a],$$

where by convention we take $M^\pi(x', a', 0) = 1$ for all $x' \in \mathcal{X}$ and $a' \in \mathcal{A}$. This is a recursive characterization of the ith moment of a return distribution, analogous to the familiar Bellman equation for the mean. The recursion is cast into the familiar framework of operators with the following definition.

Definition 8.6. Let $m \in \mathbb{N}^+$. The m-moment Bellman operator $T^\pi_{(m)} : \mathbb{R}^{\mathcal{X}\times\mathcal{A}\times m} \to \mathbb{R}^{\mathcal{X}\times\mathcal{A}\times m}$ is given by

$$(T^\pi_{(m)} M)(x, a, i) = \tag{8.4}$$

$$\sum_{j=0}^{i} \gamma^{i-j} \binom{i}{j} \mathbb{E}_\pi[R^j \mid X = x, A = a]\, \mathbb{E}_\pi\left[M(X', A', i-j) \mid X = x, A = a\right]. \quad \triangle$$

The collection of moments $(M^\pi(x, a, i) : (x, a) \in \mathcal{X} \times \mathcal{A}, i = 1, \ldots, m)$ is a fixed point of the operator $T^\pi_{(m)}$. In general, the m-moment Bellman operator is not a contraction mapping with respect to the L^∞ metric (except, of course, for $m = 1$; see Exercise 8.1). However, with a more nuanced analysis, we can still show that $T^\pi_{(m)}$ has a unique fixed point to which the iterates

$$M_{k+1} = T^\pi_{(m)} M_k \tag{8.5}$$

converge.

Proposition 8.7. Let $m \in \mathbb{N}^+$. Under Assumption 4.29(m), M^π is the unique fixed point of $T^\pi_{(m)}$. In addition, for any initial condition $M_0 \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}\times m}$, the iterates of Equation 8.5 converge to M^π. $\triangle$

Proof. We begin by constructing a suitable notion of distance between m-moment functions $\mathbb{R}^{\mathcal{X}\times\mathcal{A}\times m}$. For $M \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}\times m}$, let

$$\|M\|_{\infty,i} = \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} |M(x, a, i)|, \qquad \text{for } i = 1, \ldots, m$$

$$\|M\|_{\infty,<i} = \sup_{j=1,\ldots,i-1} \|M\|_{\infty,j}, \qquad \text{for } i = 2, \ldots, m.$$

Each of $\|\cdot\|_{\infty,i}$ (for $i = 1, \ldots, m$) and $\|\cdot\|_{\infty,<i}$ (for $i = 2, \ldots, m$) is a *semi-norm*; they fulfill the requirements of a norm, except that neither $\|M\|_{\infty,i} = 0$ nor $\|M\|_{\infty,<i} = 0$ implies that $M = 0$. From these semi-norms, we construct the pseudo-metrics

$$(M, M') \mapsto \|M - M'\|_{\infty,i},$$

noting that it is possible for the distance between M and M' to be zero even when M is different from M'.

The structure of the proof is to argue that $T^{\pi}_{(m)}$ is a contraction with modulus γ with respect to $\|\cdot\|_{\infty,1}$ and then to show inductively that it satisfies an inequality of the form

$$\|T^{\pi}_{(m)}M - T^{\pi}_{(m)}M'\|_{\infty,i} \leq C_i\|M - M'\|_{\infty,<i} + \gamma^i\|M - M'\|_{\infty,i}\,, \tag{8.6}$$

for each $i = 2, \ldots, m$, and some constant C_i that depends on $P_{\mathcal{R}}$. Chaining these results together then leads to the convergence statement, and uniqueness follows as an immediate corollary.

To see that $T^{\pi}_{(m)}$ is a contraction with respect to $\|\cdot\|_{\infty,1}$, let $M \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}\times m}$, and write $M_{(i)} = (M(x,a,i) : (x,a) \in \mathcal{X}\times\mathcal{A})$ for the function in $\mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ corresponding to the ith moment function estimates given by M. By inspecting Equation 8.4 with $i = 1$, it follows that

$$(T^{\pi}_{(m)}M)_{(1)} = T^{\pi}M_{(1)},$$

where T^{π} is the usual Bellman operator. Furthermore, $\|M\|_{\infty,1} = \|M_{(1)}\|_{\infty}$, and so the statement that $T^{\pi}_{(m)}$ is a contraction with respect to the pseudo-metric implied by $\|\cdot\|_{\infty,1}$ is equivalent to the contractivity of T^{π} on $\mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ with the respect to the L^{∞} norm, which was shown in Proposition 4.4.

To see that $T^{\pi}_{(m)}$ satisfies the bound of Equation 8.6 for $i > 1$, let $L \in \mathbb{R}$ be such that

$$\left|\,\mathbb{E}[R^i \mid X = x, A = a]\right| \leq L, \quad \text{for all } x, a \in \mathcal{X}\times\mathcal{A} \text{ and } i = 1, \ldots, m.$$

Observe that

$$\begin{aligned}
&\left|(T^{\pi}_{(m)}M)(x,a,i) - (T^{\pi}_{(m)}M')(x,a,i)\right| \\
&\quad= \Bigg|\sum_{j=0}^{i-1}\gamma^{i-j}\binom{i}{j}\mathbb{E}_{\pi}\left[R^j \mid X = x, A = a\right]\times \\
&\qquad\qquad \sum_{\substack{x'\in\mathcal{X}\\ a'\in\mathcal{A}}} P_{\mathcal{X}}(x' \mid x,a)\pi(a' \mid x')(M - M')(x', a', i-j)\Bigg| \\
&\quad\leq \sum_{j=1}^{i-1}\gamma^{i-j}\binom{i}{j}\left|\,\mathbb{E}_{\pi}\left[R^j \mid X = x, A = a\right]\right| \times \|M - M'\|_{\infty,<i} + \gamma^i\|M - M'\|_{\infty,i} \\
&\quad\leq L\sum_{j=1}^{i-1}\gamma^{i-j}\binom{i}{j}\|M - M'\|_{\infty,<i} + \gamma^i\|M - M'\|_{\infty,i} \\
&\quad\leq (2^i - 2)L\|M - M'\|_{\infty,<i} + \gamma^i\|M - M'\|_{\infty,i}\,.
\end{aligned}$$

Taking $C_i = (2^i - 2)L$, we have

$$\|T^\pi_{(m)} M - T^\pi_{(m)} M'\|_{\infty,i} \le C_i \|M - M'\|_{\infty,<i} + \gamma^i \|M - M'\|_{\infty,i}, \text{ for } i = 2, \ldots, m.$$

To chain these results together, first observe that

$$\|M_k - M^\pi\|_{\infty,1} \to 0.$$

We next argue inductively that if, for a given $i < m$, $(M_k)_{k\ge 0}$ converges to M^π in the pseudo-metric induced by $\|\cdot\|_{\infty,<i}$, then also

$$\|M_k - M^\pi\|_{\infty,i} \to 0, \quad \text{and hence}$$
$$\|M_k - M^\pi\|_{\infty,<(i+1)} \to 0.$$

Let $y_k = \|M_k - M^\pi\|_{\infty,<i}$ and $z_k = \|M_k - M^\pi\|_{\infty,i}$. Then the generalized contraction result states that $z_{k+1} \le C_i y_k + \gamma^i z_k$. Taking the limit superior on both sides yields

$$\limsup_{k\to\infty} z_k \le \limsup_{k\to\infty} \left[C_i y_k + \gamma^i z_k\right] = \gamma^i \limsup_{k\to\infty} z_k ,$$

where we have used the result $y_k \to 0$. From this, we deduce $\limsup_{k\to\infty} z_k \le 0$, but since $(z_k)_{k\ge 0}$ is a nonnegative sequence, we therefore have $z_k \to 0$. This completes the inductive step, and we therefore obtain $\|M_k - M^\pi\|_{\infty,i} \to 0$, as required. □

In essence, Proposition 8.7 establishes that the m-moment Bellman operator behaves in a similar fashion to the usual Bellman operator, in the sense that its iterates converge to the fixed point M^π. From here, we may follow the derivations of Chapter 5 to construct a dynamic programming algorithm for learning these moments[62] or those of Chapter 6 to construct the corresponding incremental algorithm (Section 8.8). Although the proof above does not demonstrate the contractive nature of the moment Bellman operator, for $m = 2$, this can be achieved using a different norm and analysis technique (Exercise 8.4).

8.3 Bellman Closedness

In preceding chapters, our approach to distributional reinforcement learning considered approximations of the return distributions that could be tractably manipulated by algorithms. The m-moment Bellman operator, on the other hand, is not directly applied to probability distributions – compared to say, a m-categorical distribution, there is no immediate procedure for drawing a sample from a collection of m moments. Compared to the categorical and

62. When the reward distributions take on a finite number of values, in particular, the expectations of Definition 8.6 can be implemented as sums.

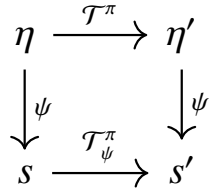

Figure 8.1
A sketch is Bellman closed if there is an operator $\mathcal{T}^\pi_\psi$ such that in the diagram above, the composite functions $\psi \circ \mathcal{T}^\pi$ and $\mathcal{T}^\pi_\psi \circ \psi$ coincide.

quantile projected operators, however, the m-moment operator yields an error-free dynamic programming procedure – with sufficiently many iterations and under some finiteness assumptions, we can determine the moments of the return function to any degree of accuracy. The concept of *Bellman closedness* formalizes this idea.

Definition 8.8. A sketch $\psi = (\psi_1, \ldots, \psi_m)$ is *Bellman closed* if, whenever its domain $\mathscr{P}_\psi(\mathbb{R})^{\mathcal{X}}$ is closed under the distributional Bellman operator:

$$\eta \in \mathscr{P}_\psi(\mathbb{R})^{\mathcal{X}} \implies \mathcal{T}^\pi \eta \in \mathscr{P}_\psi(\mathbb{R})^{\mathcal{X}},$$

there is an operator $\mathcal{T}^\pi_\psi : I^{\mathcal{X}}_\psi \to I^{\mathcal{X}}_\psi$ such that

$$\psi(\mathcal{T}^\pi \eta) = \mathcal{T}^\pi_\psi \psi(\eta) \quad \text{for all } \eta \in \mathscr{P}_\psi(\mathbb{R})^{\mathcal{X}}.$$

The operator $\mathcal{T}^\pi_\psi$ is said to be the Bellman operator for the sketch ψ. △

As was demonstrated in the preceding section, the collection of the m first moments $(\mu_1, \ldots, \mu_m)$ is a Bellman-closed sketch. Its associated operator is the m-moment operator $T^\pi_{(m)}$.

When a sketch ψ is Bellman closed, the operator T^π_ψ mirrors the application of the distributional Bellman operator to the return-distribution function η; see Figure 8.1. The concept of Bellman closedness is related to that of a diffusion-free projection (Chapter 5), and we will in fact establish an equivalence between the two in Section 8.4. In addition, Bellman-closed sketches are particularly interesting from a computational perspective because they support an exact dynamic programming procedure, as the following establishes.

Proposition 8.9. Let $\psi = (\psi_1, \ldots, \psi_m)$ be a Bellman-closed sketch and suppose that $\mathscr{P}_\psi(\mathbb{R})^{\mathcal{X}}$ is closed under $\mathcal{T}^\pi$. Then for any initial condition $\eta_0 \in \mathscr{P}_\psi(\mathbb{R})^{\mathcal{X}}$, and sequences $(\eta_k)_{k\geq 0}$, $(s_k)_{k\geq 0}$ defined by

$$\eta_{k+1} = \mathcal{T}^\pi \eta_k\,, \quad s_0 = \psi(\eta_0)\,, \quad s_{k+1} = \mathcal{T}^\pi_\psi s_k\,,$$

we have, for $k \geq 0$,

$$s_k = \psi(\eta_k)\,.$$

In addition, the functional values $s^\pi = \psi(\eta^\pi)$ of the return-distribution function are a fixed point of the operator $\mathcal{T}^\pi_\psi$. △

Proof. Both parts of the result follow immediately from the definition of the operator $\mathcal{T}^\pi_\psi$. First suppose that $s_k = \psi(\eta_k)$, for some $k \geq 0$. Then note that

$$s_{k+1} = \mathcal{T}^\pi_\psi s_k = \mathcal{T}^\pi_\psi \psi(\eta_k) = \psi(\mathcal{T}^\pi \eta_k) = \psi(\eta_{k+1})\,.$$

Thus, by induction, the first statement is proven. For the second statement, we have

$$s^\pi = \psi(\eta^\pi) = \psi(\mathcal{T}^\pi \eta^\pi) = \mathcal{T}^\pi_\psi \psi(\eta^\pi) = \mathcal{T}^\pi_\psi s^\pi\,. \qquad \square$$

Of course, dynamic programming is only feasible if the operator $\mathcal{T}^\pi_\psi$ can itself be implemented in a computationally tractable manner. In the case of the m-moment operator, we know this is possible under similar assumptions as were made in Chapter 5.

Proposition 8.9 illustrates how, when the sketch ψ is Bellman closed, we can do away with probability distributions and work exclusively with functional values. However, many sketches of interest fail to be Bellman closed, as the following examples demonstrate.

Example 8.10 (The median functional). A median of a distribution ν is its 0.5-quantile $F^{-1}_\nu(0.5)$.[63] Perhaps surprisingly, there is in general no way to determine the median of a return distribution based solely on the medians at the successor states. To see this, consider a state x that leads to state y_1 with probability $1/3$ and to state y_2 with probability $2/3$, with zero reward. The following are two scenarios in which the median returns at y_1 and y_2 are the same, but the median at x is different (see Figure 8.2):

63. As usual, there might be multiple values of z for which $\mathbb{P}_{Z\sim\nu}(Z \leq z) = 0.5$; recall that F^{-1} takes the smallest such value.

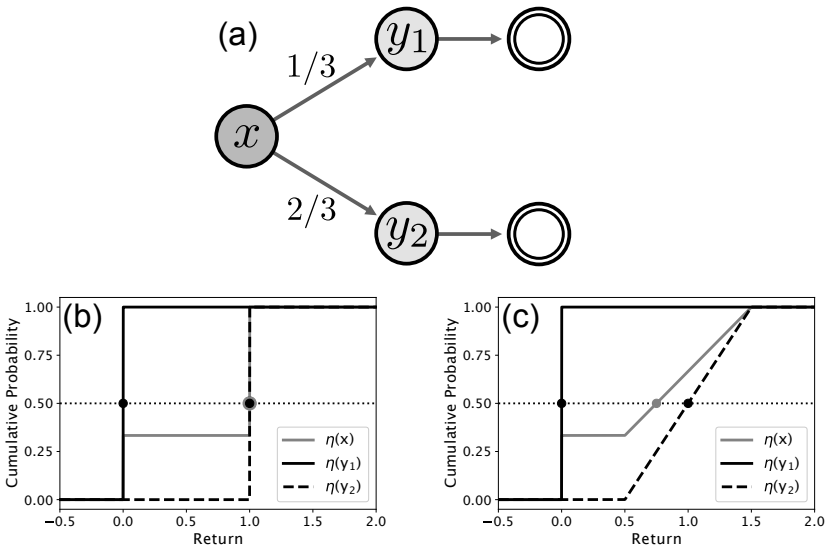

Figure 8.2
Illustration of Example 8.10. **(a)** A Markov decision process in which state x leads to states y_1 and y_2 with probability $1/3$ and $2/3$, respectively. **(b)** Case 1, in which the median of $\eta(x)$ matches the median of $\eta(y_2)$. **(c)** Case 2, in which the median of $\eta(x)$ differs from the median of $\eta(y_2)$.

Case 1. The return distributions at y_1 and y_2 are Dirac deltas at 0 and 1, respectively, and these are also the medians of these distributions. The median at x is also 1.

Case 2. The return distributions at y_1 and y_2 are a Dirac delta at 0 and the uniform distribution on [0.5, 1.5], respectively, and have the same medians as in Case 1. However, the median at x is now 0.75. △

Example 8.11 (**At-least functionals**)**.** For $\nu \in \mathscr{P}(\mathbb{R})$ and $z \in \mathbb{R}$, let us define the *at-least functional*

$$\psi_{\geq z}(\nu) = \mathbb{1}\{\mathbb{P}_{Z\sim\nu}(Z \geq z) > 0\}\,,$$

measuring whether ν assigns positive probability to values in $[z, \infty)$. Now consider a state x that deterministically leads to y, with no reward, and suppose that there is a single action a available. The statement "it is possible to obtain a return of at least 10 at state y" corresponds to

$$\psi_{\geq 10}(\eta^\pi(y)) = 1\,. \tag{8.7}$$

If Equation 8.7 holds, can we deduce whether or not a return of at least 10 is possible at state x? The answer is no. Suppose that $\gamma = 0.9$, and consider the following two situations:

Case 1. $\eta^\pi(y) = \delta_{10}$. Then $\psi_{\geq 10}(\eta^\pi(y)) = 1$, $\eta^\pi(x) = \delta_9$ and $\psi_{\geq 10}(\eta^\pi(x)) = 0$.

Case 2. $\eta^\pi(y) = \delta_{20}$. Then $\psi_{\geq 10}(\eta^\pi(y)) = 1$ still. However, $\eta^\pi(x) = \delta_{18}$ and $\psi_{\geq 10}(\eta^\pi(x)) = 1$. △

What goes wrong in the examples above is that we do not have sufficient information about the return distribution at the successor states to compute the functional values for the return distribution of state x. Consequently, we cannot use an iterative procedure to determine the functional values of η^π, at least not without error.

As it turns out, m-moment sketches are somewhat special in being Bellman closed. As the following theorem establishes, any sketch whose functionals are expectations of functions must encode the same information as a moment sketch.

Theorem 8.12. Let $\psi = (\psi_1, \ldots, \psi_m)$ be a sketch. Suppose that ψ is Bellman closed and that for each $i = 1, \ldots, m$, there is a function $f_i : \mathbb{R} \to \mathbb{R}$ for which

$$\psi_i(\nu) = \mathop{\mathbb{E}}_{Z\sim\nu}[f_i(Z)]\,.$$

Then, ψ is equivalent to the first n-moment functionals for some $n \leq m$, in the sense that there are real-valued coefficients (b_{ij}) and (c_{ij}) such that for any $\nu \in \mathscr{P}_\psi(\mathbb{R}) \cap \mathscr{P}_m(\mathbb{R})$,

$$\psi_i(\nu) = \sum_{j=1}^{n} b_{ij}\mu_j(\nu) + b_{i0}\,, \quad i = 1, \ldots, m\,;$$

$$\mu_j(\nu) = \sum_{i=1}^{m} c_{ij}\psi_i(\nu) + c_{0j}\,, \quad j = 1, \ldots, n\,. \qquad \triangle$$

The proof is somewhat lengthy and is given in Remark 8.2 at the end of the chapter.

As a corollary, we may deduce that any sketch that can be expressed as an invertible function of the first m moments is also Bellman closed. More precisely, if ψ' is a sketch that is an invertible transformation of the sketch ψ corresponding to the first m moments, say $\psi' = h \circ \psi$, then ψ' is Bellman closed with corresponding Bellman operator $h \circ T^\pi_\psi \circ h^{-1}$. Thus, for example, we may deduce that the sketch corresponding to the mean and variance functionals is Bellman closed, since the mean and variance are expressible as an invertible function of the mean and uncentered second moment. On the other hand, many

other statistical functionals (including quantile functionals) are not covered by Theorem 8.12. In the latter case, this is because there is no function $f : \mathbb{R} \to \mathbb{R}$ whose expectation for an arbitrary distribution ν recovers the τth quantile of ν (Exercise 8.5). Still, as established in Example 8.10, quantile sketches are not Bellman closed.

8.4 Statistical Functional Dynamic Programming

When a sketch ψ is not Bellman closed, we lack an operator $\mathcal{T}_\psi^\pi$ that emulates the combination of the distributional Bellman operator and this sketch. This precludes a dynamic programming approach that bootstraps its functional value estimates directly from the previous estimates. However, approximate dynamic programming with arbitrary statistical functionals is still possible if we introduce an additional *imputation step* ι that reconstructs plausible probability distributions from functional values. As we will now see, this allows us to apply the distributional Bellman operator to the reconstructed distributions and then extract the functional values of the resulting return function estimate.

Definition 8.13. An *imputation strategy* for the sketch $\psi : \mathscr{P}_\psi(\mathbb{R}) \to \mathbb{R}^m$ is a function $\iota : I_\psi \to \mathscr{P}_\psi(\mathbb{R})$. We say that it is *exact* if for any valid functional values $(s_1, \ldots, s_m) \in I_\psi$, we have

$$\psi_i(\iota(s_1, \ldots, s_m)) = s_i, \quad i = 1, \ldots, m\,.$$

Otherwise, we say that it is *approximate*.

By extension, we write $\iota(s) \in \mathscr{P}_\psi(\mathbb{R})^X$ for the return-distribution function corresponding to the collection of functional values $s \in I_\psi^X$. △

In other words, if ι is an exact imputation strategy for the sketch $\psi = (\psi_1, \ldots, \psi_m)$, then for any valid values $s_1, \ldots, s_m$ of the functionals $\psi_1, \ldots, \psi_m$, we have that $\iota(s_1, \ldots, s_m)$ is a probability distribution with the required values under each functional. In a certain sense, ι is a pseudo-inverse to the vector-valued map $\psi : \nu \mapsto (\psi_1(\nu), \ldots, \psi_m(\nu))$. Note that a true inverse to ψ does not exist, as ψ generally does not capture all aspects of the distribution ν.

Once an imputation strategy has been selected, it is possible to write down an approximate dynamic programming algorithm for the functional values under consideration. An abstract framework is given in Algorithm 8.1. In effect, such an algorithm recursively computes the iterates

$$s_{k+1} = \psi(\mathcal{T}^\pi \iota(s_k)) \tag{8.8}$$

from an initial $s_0 \in I_\psi^X$. Procedures that implement the iterative process described by Equation 8.8 are referred to as *statistical functional dynamic programming* (SFDP) algorithms. When the sketch ψ is Bellman closed and its imputation

strategy ι is exact, the sequence of iterates $(s_k)_{k\geq 0}$ converges to $\psi(\eta^\pi)$, so long as ψ is continuous (with respect to a Wasserstein metric).

Algorithm 8.1: Statistical functional dynamic programming

Algorithm parameters: statistical functionals $\psi_1, \ldots, \psi_m$,
imputation strategy ι,
initial functional values $((s_i(x))_{i=1}^m : x \in \mathcal{X})$,
desired number of iterations K

for $k = 1, \ldots, K$ **do**
  ▷ `Impute distributions`
  $\eta \leftarrow (\iota(s_1(x), \ldots, s_m(x)) : x \in \mathcal{X})$
  ▷ `Apply distributional Bellman operator`
  $\tilde{\eta} \leftarrow \mathcal{T}^\pi \eta$
  foreach state $x \in \mathcal{X}$ **do**
    for $i = 1, \ldots, m$ **do**
      ▷ `Update statistical functional values`
      $s_i(x) \leftarrow \psi_i(\tilde{\eta}(x))$
    end for
  end foreach
end for
return $((s_i(x))_{i=1}^m : x \in \mathcal{X})$

Example 8.14. For the quantile functionals $(\psi_{\tau_i}^Q)_{i=1}^m$ with $\tau_i = \frac{2i-1}{2m}$ for $i = 1, \ldots, m$, an exact imputation strategy is

$$(q_1, \ldots, q_m) \mapsto \frac{1}{m} \sum_{i=1}^m \delta_{q_i} . \tag{8.9}$$

This follows because the $\frac{2i-1}{2m}$-quantile of $\frac{1}{m} \sum_{i=1}^m \delta_{q_i}$ is precisely q_i.

Note that when $\tau_1, \ldots, \tau_m \in (0, 1)$ are arbitrary levels with quantile values $(q_1, \ldots, q_m)$, however, it is generally not true that Equation 8.9 is an exact imputation strategy for the corresponding quantile functionals. △

Example 8.15. Categorical dynamic programming can be interpreted as an SFDP algorithm. Indeed, the parameters $p_1, \ldots, p_m$ found by the categorical

projection correspond to the values of the following statistical functionals:

$$\psi_i^C(\nu) = \mathop{\mathbb{E}}_{Z\sim\nu}[h_i(\varsigma_m^{-1}(Z-\theta_i)], \quad i = 1,\dots,m \tag{8.10}$$

where $(h_i)_{i=1}^m$ are the triangular and half-triangular kernels defining the categorical projection on $(\theta_i)_{i=1}^m$ (Section 5.6). An exact imputation strategy in this case is the function that returns the unique distribution supported on $(\theta_i)_{i=1}^m$ that matches the estimated functional values $p_i = \psi_i^C(\nu)$, $i = 1,\dots,m$:

$$(p_1,\dots,p_m) \mapsto \sum_{i=1}^m p_i \delta_{\theta_i}\,. \qquad \triangle$$

Mathematically, an exact imputation strategy always exists, because we defined imputation strategies in terms of valid functional values. However, there is no guarantee that an efficient algorithm exists to compute the application of this strategy to arbitrary functional values. In practice, we may favor *approximate strategies* with efficient implementations. For example, we may map functional values to probability distributions from a representation $\mathscr{F}$ by optimizing some notion of distance between functional values. The optimization process may not yield an exact match in $\mathscr{F}$ (one may not even exist) but can often be performed efficiently.

Example 8.16. Let $\psi_1^{\mathrm{c}},\dots,\psi_m^{\mathrm{c}}$ be the categorical functionals from Equation 8.10. Suppose we are given the corresponding functional values $p_1,\dots,p_m$ of a probability distribution ν:

$$p_i = \psi_i^{\mathrm{c}}(\nu), \quad i = 1,\dots,m\,.$$

An approximate imputation strategy for these functionals is to find the n-quantile distribution (n possibly different from m)

$$\nu_\theta = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i}$$

that best fits p_i according to the loss

$$\mathcal{L}(\theta) = \sum_{i=1}^m \left|p_i - \psi_i^{\mathrm{c}}(\nu_\theta)\right|. \tag{8.11}$$

Exercise 8.7 asks you to demonstrate that this strategy is approximate for $m > 2$. Although in this context, we know of an exact imputation strategy based on categorical distributions, this illustrates that it is possible to impute distributions from a different representation. $\triangle$

8.5 Relationship to Distributional Dynamic Programming

In Chapter 5, we introduced distributional dynamic programming (DDP) as a class of methods that operates over return-distribution functions. In fact, every statistical functional dynamic programming is also a DDP algorithm (but not the other way around; see Exercise 8.8). This relationship is established by considering the implied representation

$$\mathscr{F} = \{\iota(s) : s \in I_\psi\} \subseteq \mathscr{P}(\mathbb{R})$$

and the projection $\Pi_\mathscr{F} = \iota \circ \psi$ (see Figure 8.3).

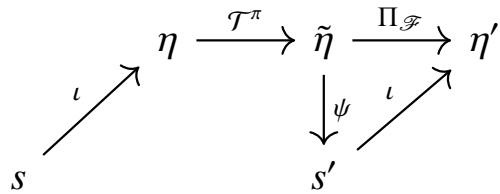

Figure 8.3
The interpretation of SFDP algorithms as distributional dynamic programming algorithms. Traversing along the diagram from η to η' corresponds to dynamic programming implementing a projected Bellman operator, while the path from s to s' corresponds to statistical functional dynamic programming (SFDP).

From this correspondence, we may establish the relationship between Bellman closedness and the notion of a diffusion-free projection developed in Chapter 5.

Proposition 8.17. Let ψ be a Bellman-closed sketch. Then for any choice of exact imputation strategy $\iota : I_\psi \to \mathscr{P}_\psi(\mathbb{R})$, the projection operator $\Pi_\mathscr{F} = \iota\psi$ is diffusion-free. △

Proof. We may directly check the diffusion-free property (omitting parentheses for conciseness):

$$\Pi_\mathscr{F}\mathcal{T}^\pi\Pi_\mathscr{F} = \iota\psi\mathcal{T}^\pi\iota\psi \overset{(a)}{=} \iota\mathcal{T}^\pi_\psi\psi\iota\psi \overset{(b)}{=} \iota\mathcal{T}^\pi_\psi\psi \overset{(a)}{=} \iota\psi\mathcal{T}^\pi = \Pi_\mathscr{F}\mathcal{T}^\pi\,,$$

where steps marked (a) follow from the identity $\psi\mathcal{T}^\pi = \mathcal{T}^\pi_\psi\psi$, and (b) follows from the identity $\psi\iota\psi = \psi$ for any exact imputation strategy ι for ψ. □

Imputation strategies formalize how one might interpret functional values as parameters of a probability distribution. Naturally, the chosen imputation

strategy affects the approximation artifacts from distributional dynamic programming, the rate of convergence, and whether the algorithm converges at all.

Compared with representation-based algorithms of the style introduced in Chapter 5, working with statistical functionals allows us to design the projection $\Pi_{\mathscr{F}}$ in two separate pieces: a sketch ψ and an imputation strategy ι. In particular, this makes it possible to learn statistical functionals that would be difficult to directly capture in a probability distribution representation. As the next section demonstrates, this allows us to create new kinds of distributional reinforcement learning algorithms.

8.6 Expectile Dynamic Programming

Expectiles form a family of statistical functionals parameterized by a level $\tau \in (0, 1)$. They extend the notion of the mean of a distribution ($\tau = 0.5$) similar to how quantiles extend the notion of a median. Expectiles have classically found application in econometrics and finance as a form of risk measure (see the bibliographical remarks for further details). Based on the principles of statistical functional dynamic programming, *expectile dynamic programming*[64] uses an approximate imputation strategy in order to iteratively estimate the expectiles of the return function.

Definition 8.18. For a given $\tau \in (0, 1)$, the *τ-expectile* of a distribution $\nu \in \mathscr{P}_2(\mathbb{R})$ is

$$\psi^{\mathrm{E}}_\tau(\nu) = \operatorname*{arg\,min}_{z \in \mathbb{R}} \mathrm{ER}_\tau(z; \nu)\,, \tag{8.12}$$

where

$$\mathrm{ER}_\tau(z; \nu) = \mathop{\mathbb{E}}_{Z \sim \nu} \left[|\mathbb{1}_{\{Z < z\}} - \tau| \times (Z - z)^2 \right] \tag{8.13}$$

is the *expectile loss*. △

The loss appearing in Definition 8.18 is strongly convex (Boyd and Vandenberghe 2004) and bounded below by 0. As a consequence, Equation 8.12 has a unique minimizer for a given τ; this verifies that the corresponding expectile is uniquely defined.

To understand the relationship to the mean functional and develop some intuition for the statistical property than an expectile encodes, observe that the mean of a distribution $\nu \in \mathscr{P}_2(\mathbb{R})$ can be expressed as

$$\mu_1(\nu) = \operatorname*{arg\,min}_{z \in \mathbb{R}} \mathop{\mathbb{E}}_{Z \sim \nu} \left[(Z - z)^2 \right].$$

64. The incremental analogue is called *expectile temporal-difference learning* (Rowland et al. 2019).

Similar to how a quantile is derived from a loss that weights errors asymmetrically (depending on whether the realization from Z is smaller or greater than z), the expectile loss for $\tau \in (0, 1)$ is the asymmetric version of the above. For τ greater than $1/2$, one can think of the expectile as an "optimistic" summary of the distribution – a value that emphasizes outcomes that are greater than the mean. Conversely, for τ smaller than $1/2$, the corresponding expectile is in a sense "pessimistic."

Expectile dynamic programming (EDP) estimates the values of a finite set of expectile functionals with values $0 < \tau_1 < \cdots < \tau_m < 1$. For a distribution $\nu \in \mathscr{P}_2(\mathbb{R})$, let us write

$$e_i = \psi^{\mathrm{E}}_{\tau_i}(\nu)\,.$$

Given the collection of expectile values $e_1, \ldots, e_m$, EDP uses an imputation strategy that outputs an n-quantile probability distribution that approximately has these expectile values.[65]

The imputation strategy finds a suitable reconstruction by finding a solution to a root-finding problem. To begin, this strategy outputs a n-quantile distribution $\hat{\nu}$, with n possibly different from m:

$$\hat{\nu} = \frac{1}{n}\sum_{j=1}^{n} \delta_{\theta_j}\,.$$

Following Definition 8.13, for this imputation to be exact, the expectiles of $\hat{\nu}$ at $\tau_1, \ldots, \tau_m$ should be equal to $e_1, \ldots, e_m$:

$$\psi^{\mathrm{E}}_{\tau_i}(\hat{\nu}) = e_i, \quad i = 1, \ldots, m\,.$$

This constraint implies that the derivatives of the expectile loss, instantiated with $\tau_1, \ldots, \tau_m$ and evaluated with $\hat{\nu}$, should all be 0:

$$\partial_z \,\mathrm{ER}_{\tau_i}\Big(z; \hat{\nu}\Big)\Big|_{z=e_i} = 0, \quad i = 1, \ldots, m\,. \tag{8.14}$$

Written out in full for the choice of $\hat{\nu}$ above, these derivatives take the form

$$\partial_z \,\mathrm{ER}_{\tau_i}\Big(z; \hat{\nu}\Big)\Big|_{z=e_i} = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{2}(e_i - \theta_j)|\mathbb{1}_{\{\theta_j < e_i\}} - \tau_i|, \qquad i = 1, \ldots, m\,.$$

An alternative to the root-finding problem expressed in Equation 8.14 is the following optimization problem:

$$\text{minimise} \sum_{i=1}^{m} \left(\partial_z \,\mathrm{ER}_{\tau_i}\Big(z; \hat{\nu}\Big)\Big|_{z=e_i}\right)^2\,. \tag{8.15}$$

65. Of course, this particular form for the imputation strategy is a design choice; the reader is invited to consider what other imputation strategies might be sensible here.

A practical implementation of this imputation strategy therefore applies an optimization algorithm to the objective in Equation 8.15, or a root-finding method to Equation 8.14, viewed as functions of $\theta_1, \ldots, \theta_n$. Because the optimization algorithm may return a solution that does not exactly satisfy Equation 8.14, this method is an approximate (rather than exact) imputation strategy. It can be used in the *impute distributions* step of Algorithm 8.1, yielding a dynamic programming algorithm that aims to approximately learn return-distribution expectiles. If the root-finding algorithm is always able to find $\hat{\nu}$ exactly satisfying Equation 8.14, then the imputation strategy is exact in this instance; otherwise, it is approximate. A specific implementation is explored in detail in Exercise 8.10.

8.7 Infinite Collections of Statistical Functionals

Thus far, our treatment of statistical functionals has focused on finite collections of statistical functionals – what we call a sketch. From a computational standpoint, this is sensible since, to implement an SFDP algorithm, one needs to be able to operate on individual functional values. On the other hand, in Section 8.3, we saw that many sketches are not Bellman closed and must be combined with an imputation strategy in order to perform dynamic programming. An alternative, which we will study in greater detail in Chapter 10, is to implicitly parameterize an *infinite* family of statistical functionals.

Many (though not all) infinite families of functionals provide a lossless encoding of probability distributions and are consequently Bellman closed – that is, knowing the values taken on by these functionals is equivalent to knowing the distribution itself. We encode this property with the following definition.

Definition 8.19. Let Ψ be a set of statistical functionals. We say that Ψ *characterizes* probability distributions over the real numbers if, for each $\nu \in \mathscr{P}(\mathbb{R})$, there is a unique collection of functional values $(\psi(\nu) : \psi \in \Psi)$. $\triangle$

The following families of statistical functionals all characterize probability distributions over $\mathbb{R}$.

The cumulative distribution function. The functionals mapping distributions ν to the probabilities $\mathbb{P}_{Z\sim\nu}(Z \leq z)$, indexed by $z \in \mathbb{R}$. Closely related are *upper-tail probabilities*,

$$\nu \mapsto \mathbb{P}_{Z\sim\nu}(Z \geq z)\,,$$

and the quantile functionals

$$\nu \mapsto F_\nu^{-1}(\tau)\,,$$

indexed by $\tau \in (0, 1)$.

The characteristic function. Functionals of the form

$$\nu \mapsto \mathop{\mathbb{E}}_{Z\sim\nu}[e^{\mathrm{i}uZ}] \in \mathbb{C}\,,$$

indexed by $u \in \mathbb{R}$ (and where $\mathrm{i}^2 = -1$). The corresponding collection of statistical values is the *characteristic function* of ν, denoted χ_ν.

Moments and cumulants. The infinite collection of moment functionals $(\mu_p)_{p=1}^\infty$ does not unconditionally characterize the distribution ν: there are distinct distributions that have the same sequence of moments. However, if the sequence of moments does not grow too quickly, uniqueness is restored. In particular, a sufficient condition for uniqueness is that the underlying distribution ν has a *moment-generating function*

$$u \mapsto \mathop{\mathbb{E}}_{Z\sim\nu}[e^{uZ}]\,,$$

which is finite in an open neighborhood of $u = 0$; see Remark 8.3 for further details. Under this condition, the moment-generating function itself also characterizes the distribution, as does the *cumulant-generating function*, defined as the logarithm of the moment-generating function,

$$u \mapsto \log\left(\mathop{\mathbb{E}}_{Z\sim\nu}[e^{uZ}]\right).$$

The *cumulants* $(\kappa_p)_{p=1}^\infty$ are defined through a power series expansion of the cumulant-generating function

$$\log\left(\mathop{\mathbb{E}}_{Z\sim\nu}[e^{uZ}]\right) = \sum_{p=1}^{\infty} \frac{\kappa_p u^p}{p!}\,.$$

Under the condition that the moment-generating function is finite in an open neighborhood of the origin, the sequences of cumulants and moments are determined by one another, and so the sequence of cumulants is another characterization of the distribution under this condition.

Example 8.20. Consider the return-variable Bellman equation

$$G^\pi(x,a) \overset{\mathcal{D}}{=} R + \gamma G^\pi(X', A')\,, \quad X = x\,, A = a\,.$$

If for each $u \in \mathbb{R}$ we apply the functional $\nu \mapsto \mathop{\mathbb{E}}_{Z\sim\nu}[e^{\mathrm{i}uZ}]$ to the distribution of the random variables on each side, we obtain the *characteristic function Bellman equation*:

$$\begin{aligned}
\chi_{\eta^\pi(x,a)}(u) &= \mathbb{E}_\pi[e^{\mathrm{i}u(R+\gamma G^\pi(X',A'))} \mid X = x, A = a] \\
&= \mathbb{E}_\pi\left[e^{\mathrm{i}uR} \mid X = x, A = a\right] \mathbb{E}_\pi\left[e^{\mathrm{i}\gamma u G^\pi(X',A')} \mid X = x, A = a\right] \\
&= \chi_{P_{\mathcal{R}}(\cdot|x,a)}(u)\, \mathbb{E}_\pi\left[\chi_{\eta^\pi(X',A')}(\gamma u) \mid X = x, A = a\right].
\end{aligned}$$

This is a different kind of distributional Bellman equation in which the addition of independent random variables corresponds to a multiplication of their characteristic functions. The equation highlights that the characteristic function of ν evaluated at u depends on the next-state characteristic functions evaluated at γu. This shows that for a set $S \subseteq \mathbb{R}$, the sketch $(\nu \mapsto \chi_\nu(u) : u \in S)$ cannot be Bellman closed unless S is infinite or $S = \{0\}$. Exercise 8.12 asks you to give a theoretical analysis of a dynamic programming approach based on characteristic functions. △

Another way to understand collections of statistical functionals that are characterizing (in the sense of Definition 8.19) is to interpret them in light of our definition of a probability distribution representation (Definition 5.2). Recall that a representation $\mathscr{F}$ is a collection of distributions indexed by a parameter θ:

$$\mathscr{F} = \{\nu_\theta \in \mathscr{P}(\mathbb{R}) : \theta \in \Theta\}.$$

Here, the functional values associated with the set of statistical functionals Ψ correspond to the (infinite-dimensional) parameter θ, so that

$$\mathscr{F}_\Psi = \mathscr{P}(\mathbb{R}).$$

This clearly implies that $\mathscr{F}_\Psi$ is closed under the distributional Bellman operator $\mathcal{T}^\pi$ (Section 5.3) and hence that approximation-free distributional dynamic programming is (mathematically) possible with $\mathscr{F}_\Psi$.

8.8 Moment Temporal-Difference Learning*

In Section 8.2, we introduced the m-moment Bellman operator, from which an exact dynamic programming algorithm can be derived. A natural follow-up is to apply the tools of Chapter 6 to derive an incremental algorithm for learning the moments of the return-distribution function from samples. Here, an algorithm that incrementally updates an estimate $M \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}\times m}$ of the m first moments of the return function can be directly obtained through the unbiased estimation approach, as the corresponding operator can be written as an expectation. Given a sample transition (x, a, r, x', a'), the unbiased estimation approach yields the update rule (for $i = 1, \ldots, m$)

$$M(x, a, i) \leftarrow (1-\alpha)M(x, a, i) + \alpha\left[\sum_{j=0}^{i} \gamma^{i-j}\binom{i}{j} r^j M(x', a', i-j)\right], \qquad (8.16)$$

where again we take $M(\cdot, \cdot, 0) = 1$ by convention.

Unlike the TD and CTD algorithms analyzed in Chapter 6, this algorithm is derived from an operator, $T^\pi_{(m)}$, which is not a contraction in a supremum-norm over states. As a result, the theory developed in Chapter 6 cannot immediately

be applied to demonstrate convergence of this algorithm under appropriate conditions. With some care, however, a proof is possible; we now give an overview of what is needed.

The proof of Proposition 8.7 demonstrates that the behavior of $T^{\pi}_{(m)}$ is closely related to that of a contraction mapping. Specifically, the behavior of $T^{\pi}_{(m)}$ in updating the estimates of ith moments of returns is contractive if the lower moment estimates are sufficiently close to their correct values. To turn these observations into a proof of convergence, an inductive argument on the moments being learnt must be made, as in the proof of Proposition 8.7. Further, the approach of Chapter 6 needs to be extended to deal with a vanishing bias term in the update to account for this "near-contractivity" of $T^{\pi}_{(m)}$; to this end one may, for example, begin from the analysis of Bertsekas and Tsitsiklis (1996, Proposition 4.5).

Before moving on, let us remark that in practice, we are likely to be interested in centered moments such as the variance ($m = 2$); these take the form

$$\mathbb{E}_{\pi}\Big[\Big(\sum_{t=0}^{\infty}\gamma^t R_t - Q^{\pi}(x,a)\Big)^m \,\Big|\, X_0 = x, A_0 = a\Big],$$

These can be derived from their uncentered counterparts; for example, the variance of the return distribution $\eta^{\pi}(x,a)$ is obtained from the first two uncentered moments via Equation 8.2.

It is also possible to perform dynamic programming on centered moments directly, as was shown in the context of the mean and variance in Section 5.4 (Exercise 8.14 asks you to derive the Bellman operators for the more general case of the first m centered moments). Given in terms of state-action pairs, the Bellman equation for the return variances $\bar{M}^{\pi}(\cdot,\cdot,2) \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ is

$$\bar{M}^{\pi}(x,a,2) = \mathrm{Var}_{\pi}(R \mid X = x, A = a) + \gamma^2\Big(\mathrm{Var}_{\pi}(Q^{\pi}(X',A') \mid X = x, A = a) + \mathbb{E}_{\pi}[\bar{M}^{\pi}(X',A',2) \mid X = x, A = a]\Big); \tag{8.17}$$

contrast with Equation 5.20.

One challenge with deriving an incremental algorithm for learning the variance directly is that unbiasedly estimating some of the variance terms on the right-hand side requires multiple samples. For example, an unbiased estimator of

$$\mathrm{Var}_{\pi}(Q^{\pi}(X',A') \mid X = x, A = a)$$

in general requires two independent realizations of X', A' for a given source state-action pair x, a. Consequently, unbiased estimation of the corresponding operator application with a single transition is not feasible in this case. Despite the fact that the first m centered and uncentered moments of a probability

distribution can be recovered from one another, there is a distinct advantage associated with working with uncentered moments when learning from samples.

8.9 Technical Remarks

Remark 8.1. Theorem 8.12 illustrates how dynamic programming over functional values must incur some approximation error, unless the underlying sketch is Bellman closed. One way to avoid this error is to augment the state space with additional information: for example, the return accumulated so far. We in fact took this approach when optimizing the conditional value-at-risk (CVaR) of the return in Chapter 7; in fact, risk measures are statistical functionals that may also take on the value $-\infty$ (see Definition 7.14). △

Remark 8.2 (**Proof of Theorem 8.12**)**.** It is sufficient to consider a pair of states, x and y, such that x deterministically transitions to y with reward r. Because ψ is Bellman closed, we can identify an associated Bellman operator T^π_ψ. For a given return function η whose state-indexed collection of functional values is $s = \psi(\eta)$, let us write $(T^\pi_\psi s)_i(x)$ for the ith functional value at state x, for $i = 1, \ldots, m$. By construction and definition of the operator T^π_ψ, $(T^\pi_\psi s)_i(x)$ is a function of the functional values at y as well as the reward r and discount factor γ, and so we may write

$$(T^\pi_\psi s)_i(x) = g_i(r, \gamma, \psi_1(\eta(y)), \ldots, \psi_m(\eta(y)))$$

for some function g_i. We next argue that g_i is affine[66] in the inputs $\psi_1(\eta(y)), \ldots, \psi_m(\eta(y))$. This is readily observed as each functional $\psi_1, \ldots, \psi_m$ is affine in its input distribution,

$$\begin{aligned}\psi_i(\alpha\nu + (1-\alpha)\nu') &= \mathop{\mathbb{E}}_{Z\sim\alpha\nu+(1-\alpha)\nu'}[f_i(Z)] \\ &= \alpha \mathop{\mathbb{E}}_{Z\sim\nu}[f_i(Z)] + (1-\alpha) \mathop{\mathbb{E}}_{Z\sim\nu'}[f_i(Z)] \\ &= \alpha\psi_i(\nu) + (1-\alpha)\psi_i(\nu'),\end{aligned}$$

and

$$(T^\pi_\psi s)_i(x) = \mathop{\mathbb{E}}_{Z\sim\eta(y)}[f_i(r + \gamma Z)]$$

is also affine as a function of η. This affineness would be contradicted if g_i were not also affine. Hence, there exist functions $\beta_i : \mathbb{R} \times [0, 1) \to \mathbb{R}$ for $i = 1, \ldots, m$

66. Recall that a function $h : M \to M'$ between vector spaces M and M' is affine if for $u_1, u_2 \in M$, $\lambda \in (0, 1)$, we have $h(\lambda u_1 + (1-\lambda)u_2) = \lambda h(u_1) + (1-\lambda)h(u_2)$.

such that

$$g_i(r, \gamma, \psi_1(\eta(y)), \ldots, \psi_m(\eta(y))) = \beta_0(r, \gamma) + \sum_{i=1}^{m} \beta_i(r, \gamma)\psi_i(\eta(y)) ,$$

and therefore

$$\mathop{\mathbb{E}}_{Z\sim\eta(y)}[f_i(r + \gamma Z)] = \mathop{\mathbb{E}}_{Z\sim\eta(y)}\left[\sum_{j=0}^{m} \beta_i(r, \gamma) f_j(Z)\right],$$

where $f_0(z) = 1$. Taking $\eta(y)$ to be a Dirac delta δ_z then gives the following identity:

$$f_i(r + \gamma z) = \sum_{j=0}^{m} \beta_i(r, \gamma) f_j(z) .$$

We therefore have that the finite-dimensional function space spanned by $f_0, f_1, \ldots, f_m$ (where f_0 is the constant function equal to 1) is closed under translation (by $r \in \mathbb{R}$) and scaling (by $\gamma \in [0, 1)$). Engert (1970) shows that the only finite-dimensional subspaces of measurable functions closed under translation are contained in the span of finitely many functions of the form $z \mapsto z^\ell \exp(\lambda z)$, with $\ell \in \mathbb{N}$ and $\lambda \in \mathbb{C}$. Since we further require closure under scaling by $\gamma \in [0, 1)$, we deduce that we must have $\lambda = 0$ in any such function, and the subspace must be equal to the space spanned by the first n monomials (and the constant function).

To conclude, since each monomial $z \mapsto z^i$ for $i = 1, \ldots, n$ is expressible as a linear combination of $f_0, \ldots, f_m$, the corresponding expectations $\mathbb{E}_{Z\sim\nu}[Z^i]$ are expressible as linear combinations of the expectations $\mathbb{E}_{Z\sim\nu}[f_j(Z)]$, for any distribution ν. The converse also holds, and so we conclude that the sketch ψ encodes the same distributional information as the first n moments. △

Remark 8.3. The question of whether a distribution is characterized by its sequence of moments has been a subject of study in probability theory for over a century. The sufficient condition on the moment-generating function described in Section 8.8 means that the characteristic function of such a distribution can be written as a power series with scaled moments as coefficients, ensuring uniqueness of the distribution; see, for example, Billingsley (2012) for a detailed discussion. Lin (2017) gives a survey of known sufficient conditions for characterization, as well as examples where characterization does not hold. △

8.10 Bibliographical Remarks

8.1. Statistical functionals are a core notion in statistics; see, for example, the classic text by van der Vaart (2000). In reinforcement learning, specific

functionals such as moments, quantiles, and CVaR have been of interest for risk-sensitive control (more on this in the bibliographical remarks of Chapter 7). Chandak et al. (2021) consider the problem of off-policy Monte Carlo policy evaluation of arbitrary statistical functionals of the return distribution.

8.2, 8.8. Sobel (1982) gives a Bellman equation for return-distribution moments for state-indexed value functions with deterministic policies. More recent work in this direction includes that of Lattimore and Hutter (2012), Azar et al. (2013), and Azar et al. (2017), who make use of variance estimates in combination with Bernstein's inequality to improve the efficiency of exploration algorithms, as well as the work of White and White (2016), who use estimated return variance to set trace coefficients in multistep TD learning methods. Sato et al. (2001), Tamar et al. (2012), Tamar et al. (2013), and Prashanth and Ghavamzadeh (2013) further develop methods for learning the variance of the return. Tamar et al. (2016) show that the operator $T^{\pi}_{(2)}$ is a contraction under a weighted norm (see Exercise 8.4), develop an incremental algorithm with a proof of convergence using the ODE method, and study both dynamic programming and incremental algorithms under linear function approximation (the topic of Chapter 9).

8.3–8.5. The notion of Bellman closedness is due to Rowland et al. (2019), although our presentation here is a revised take on the idea. The noted connection between Bellman closedness and diffusion-free representations and the term "statistical functional dynamic programming" are new to this book.

8.6. The expectile dynamic programming algorithm is new to this book but is directly derived from expectile temporal-difference learning (Rowland et al. 2019). Expectiles themselves were introduced by Newey and Powell (1987) in the context of testing in econometric regression models, with the asymmetric squared loss defining expectiles already appearing in Aigner et al. (1976). Expectiles have since found further application as risk measures, particularly within finance (Taylor 2008; Kuan et al. 2009; Bellini et al. 2014; Ziegel 2016; Bellini and Di Bernardino 2017). Our presentation here focuses on the asymmetric squared loss, requiring a finite second-moment assumption, but an equivalent definition allows expectiles to be defined for all distributions with a finite first moment (Newey and Powell 1987).

8.7. The study of characteristic functions in distributional reinforcement learning is due to Farahmand (2019), who additionally provides error propagation analysis for the *characteristic value iteration* algorithm, in which value iteration is carried out with characteristic function representations of return distributions. Earlier, Mandl (1971) studied the characteristic function of the return in Markov decision processes with deterministic immediate rewards and policies. Chow et al. (2015) combine a state augmentation method (see Chapter 7) with an

infinite-dimensional Bellman equation for CVaR values to learn a CVaR-optimal policy. They develop an implementable version of the algorithm by tracking finitely many CVaR values and using linear interpolation for the remainder, an approach related to the imputation strategies described earlier in the chapter. Characterization via the quantile function has driven the success of several large-scale distributional reinforcement learning algorithms (Dabney et al. 2018a; Yang et al. 2019), and is the subject of further study in Chapter 10.

8.11 Exercises

Exercise 8.1. Consider the m-moment Bellman operator $T^\pi_{(m)}$ (Definition 8.6). For $M \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}\times m}$, define the norm

$$\|M\|_{\infty,\text{MAX}} = \max_{i\in\{1,\dots,m\}} \sup_{\substack{x\in\mathcal{X}\\ a\in\mathcal{A}}} M(x,a,i).$$

By means of a counterexample, show that $T^\pi_{(m)}$ is not a contraction mapping in the metric induced by $\|\cdot\|_{\infty,\text{MAX}}$. △

Exercise 8.2. Let $\varepsilon > 0$. Determine a bound on the computational cost (in $O(\cdot)$ notation) of performing iterative policy evaluation with the m-moment Bellman operator to obtain an approximation $\hat{M}^\pi$ such that

$$\max_{i\in\{1,\dots,m\}} \sup_{\substack{x\in\mathcal{X}\\ a\in\mathcal{A}}} |\hat{M}^\pi(x,a,i) - M^\pi(x,a,i)| < \varepsilon.$$

You may find it convenient to refer to the proof of Proposition 8.7. △

Exercise 8.3. Equation 5.2 gives the value function V^π as the solution of the linear system of equations

$$V = r^\pi + \gamma P^\pi V.$$

Provide the analogous linear system for the moment function M^π. △

Exercise 8.4. The purpose of this exercise is to show that $T^\pi_{(2)}$ is a contraction mapping on $\mathbb{R}^{\mathcal{X}\times\mathcal{A}\times 2}$ in a *weighted* L^∞ norm, as shown by Tamar et al. (2016). Let $M \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}\times 2}$ be a moment function estimate (specifically, for the first two moments). For each $\alpha \in (0,1)$, define the α-weighted norm on $\mathbb{R}^{\mathcal{X}\times\mathcal{A}\times 2}$ by

$$\|M\|_\alpha = \alpha\|M_{(1)}\|_\infty + (1-\alpha)\|M_{(2)}\|_\infty ,$$

where $M_{(i)} = M(\cdot,\cdot,i) \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$. For any $M, M' \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}\times 2}$, show that

$$\begin{aligned}\|T^\pi_{(2)}(M-M')\|_\alpha \le{}& \alpha\|\gamma P^\pi(M_{(1)} - M'_{(1)})\|_\infty \\ &+ (1-\alpha)\|2\gamma C_r P^\pi(M_{(1)} - M'_{(1)}) + \gamma^2 P^\pi(M_{(2)} - M'_{(2)})\|_\infty ,\end{aligned}$$

where P^π is the state-action transition operator, defined by

$$(P^\pi Q)(x,a) = \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} P^\pi(x'|x,a)\pi(a'|x')Q(x',a')\,,$$

and C_r is the diagonal reward operator

$$(C_r Q)(x,a) = \mathbb{E}[R \,|\, X = x, A = a]Q(x,a)\,.$$

Writing $\lambda \geq 0$ for the Lipschitz constant of $C_r P^\pi$ with respect to the L^∞ metric, deduce that

$$\begin{aligned}&\|T^\pi_{(2)}(M - M')\|_\alpha \\ \leq&(\alpha\gamma + 2(1-\alpha)\gamma\lambda)\|M_{(1)} - M'_{(1)}\|_\infty + (1-\alpha)\gamma^2\|M_{(2)} - M'_{(2)}\|_\infty\,.\end{aligned}$$

Hence, deduce that there exist parameters $\alpha \in (0,1), \beta \in [0,1)$ such that

$$\|T^\pi_{(2)}(M - M')\|_\alpha \leq \beta\|M - M'\|_\alpha\,,$$

as required. △

Exercise 8.5. Consider the median functional

$$\nu \mapsto F_\nu^{-1}(0.5)\,.$$

Show that there does not exist a function $f : \mathbb{R} \to \mathbb{R}$ such that, for any $\nu \in \mathbb{R}$,

$$\mathop{\mathbb{E}}_{Z\sim\nu}[f(Z)] = F_\nu^{-1}(0.5)\,.$$ △

Exercise 8.6. Consider the subset of probability distributions endowed with a probability density f_ν. Repeat the preceding exercise for the differential entropy functional

$$\nu \mapsto -\int_{z\in\mathbb{R}} f_\nu(z)\log\left(f_\nu(z)\right)\mathrm{d}z\,.$$ △

Exercise 8.7. For the imputation strategy of Example 8.16:

(i) show that for $m = 2$, the imputation strategy is exact, for any $n \in \mathbb{N}^+$.
(ii) show that for $m > 2$, this imputation strategy is inexact. *Hint.* Find a distribution ν for which $\psi_i^{\mathrm{c}}(\iota(p_1,\ldots,p_m)) \neq p_i$, for some $i = 1,\ldots,m$. △

Exercise 8.8. In Section 8.4, we argued that every statistical functional dynamic programming algorithm is a distributional dynamic programming algorithm. Explain why the converse is false. Under what circumstances may we favor either an algorithm that operates on statistical functionals or one that operates on probability distribution representations? △

Exercise 8.9. Consider an imputation strategy ι for a sketch ψ. We say the (ψ,ι) pair is *mean-preserving* if, for any probability distribution $\nu \in \mathscr{P}_\psi(\mathbb{R})$,

$$\nu' = \iota\psi(\nu)$$

satisfies

$$\mathop{\mathbb{E}}_{Z\sim\nu'}[Z] = \mathop{\mathbb{E}}_{Z\sim\nu}[Z]\,.$$

Show that in this case, the operator

$$\psi \circ \mathcal{T}^{\pi} \circ \iota$$

is also mean-preserving. △

Exercise 8.10. Using your favorite numerical computation software, implement the expectile imputation strategy described in Section 8.6. Specifically:

(i) Implement a procedure for approximately determining the expectile values $e_1, \ldots, e_m$ of a given distribution. *Hint.* An incremental approach in the style of quantile regression, or a binary search approach, will allow you to deal with continuous distributions.

(ii) Given a set of expectile values, $e_1, \ldots, e_m$, implement a procedure that imputes an n-quantile distribution

$$\frac{1}{n}\sum_{i=1}^{n}\delta_{\theta_i}$$

by minimizing the objective given in Equation 8.15.

Test your implementation on discrete and continuous distributions, and compare it with the best m-quantile approximation of those distributions. Is one method better suited to discrete distributions than the other? More generally, when might one method be preferred over the other? △

Exercise 8.11. Formulate a variant of expectile dynamic programming that imputes n-quantile distributions and whose (possibly approximate) imputation strategy is mean-preserving in the sense of Exercise 8.9. △

Exercise 8.12 (*)**.** This exercise applies the line of reasoning from Chapter 4 to characteristic functions and is based on Farahmand (2019). For a probability distribution ν, recall that its characteristic function χ_ν is

$$\chi_\nu(u) = \mathop{\mathbb{E}}_{Z\sim\nu}\left[e^{iuZ}\right].$$

Now, for $p \in [1, \infty)$, define the probability metric

$$d_{1,p}(\nu, \nu') = \int_{u\in\mathbb{R}} \frac{|\chi_\nu(u) - \chi_{\nu'}(u)|}{|u|^p}\mathrm{d}u$$

and its supremum extension to return functions

$$\overline{d}_{1,p}(\eta, \eta') = \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} d_{1,p}(\eta(x,a), \eta'(x,a))\,.$$

(i) Determine a subset $\mathscr{P}_{\chi,p}(\mathbb{R}) \subseteq \mathscr{P}(\mathbb{R})$ on which $d_{1,p}$ is a proper metric.

(ii) Provide assumption(s) under which the return function η^π lies in $\mathscr{P}_{\chi,p}(\mathbb{R})$.

(iii) Prove that for $p \geq 2$, the distributional Bellman operator is a contraction mapping in $d_{1,p}$, with modulus γ^{p-1}. △

Exercise 8.13 (*)**.** Consider the probability metric

$$d_{2,2}(\nu, \nu') = \Big(\int_{u \in \mathbb{R}} \frac{(\chi_\nu(u) - \chi_{\nu'}(u))^2}{u^2} \mathrm{d}u \Big)^{1/2} .$$

Show that $d_{2,2}$ is the Cramér distance ℓ_2. *Hint.* Use the Parseval–Plancherel identity. △

Exercise 8.14. Let $m \in \mathbb{N}^+$. Derive a Bellman operator on $\mathbb{R}^{\mathcal{X} \times \mathcal{A} \times m}$ whose unique fixed point $\tilde{M}^\pi$ is the collection of centered moments:

$$\tilde{M}^\pi(x, a, i) = \mathbb{E}\left[(G^\pi(x, a) - Q^\pi(x, a))^i \right], \quad i = 1, \ldots, m . \qquad \triangle$$

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in LaTeX by the authors.

Library of Congress Cataloging-in-Publication Data

Names: Bellemare, Marc G., author. | Dabney, Will, author. | Rowland, Mark (Research scientist), author.

Title: Distributional reinforcement learning / Marc G. Bellemare, Will Dabney, Mark Rowland.

Description: Cambridge, Massachusetts : The MIT Press, [2023] | Series: Adaptive computation and machine learning | Includes bibliographical references and index.

Identifiers: LCCN 2022033240 (print) | LCCN 2022033241 (ebook) | ISBN 9780262048019 (hardcover) | ISBN 9780262374019 (epub) | ISBN 9780262374026 (pdf)

Subjects: LCSH: Reinforcement learning. | Reinforcement learning–Statistical methods.

Classification: LCC Q325.6 .B45 2023 (print) | LCC Q325.6 (ebook) | DDC 006.3/1–dc23/eng20221102

LC record available at https://lccn.loc.gov/2022033240

LC ebook record available at https://lccn.loc.gov/2022033241