The Bellman equation for state values: a derivation

David Herron

December 2022
v1.0

Contents

1 Introduction and motivation

The concept of *value functions* is fundamental to reinforcement learning (RL). When introducing value functions in Chapter 3 of their RL book, Sutton & Barto (2018) denote the *value* of a state s under policy π as $v_\pi(s)$, and they present the *value function* for any state s under any policy π using the following sequence of equations:

$$\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi[G_t|S_t = s] && (1)\\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] && (2)\\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a) \Big[r + \gamma \mathbb{E}_\pi[G_{t+1}|S_{t+1} = s'] \Big] && (3)\\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a) \Big[r + \gamma v_\pi(s') \Big] \qquad \text{for all } s \in \mathcal{S}. && (4)
\end{aligned}$$

Equation (4) is the *Bellman equation* for the *state value function for policy* π, v_π. This equation expresses the value of a state s in terms of a recursive relationship with respect to the values of its successor states, s'. It is fundamental to RL because every RL model uses some approach for learning an approximation for it. Hence, students of RL will wish to understand how this fascinating recursive equation arises.

The sequence of equations, however, merely announces (4); it is not designed to explain how (4) arises. Three of these four equations arise from definitions alone. Equation (1) is the definition of the *value* of a state s, $v_\pi(s)$. Equation (2) follows from (1) by the definition of G_t, the *return.* Equation (4) follows from (3) because the expectation in (3) is the definition of the value of successor state s', $v_\pi(s')$.

The chain of reasoning needed to understand how equations (3)/(4) arise from (1)/(2) is missing. Between equations (2) and (3) lies a yawning conceptual gap. To the student new to RL, equations (3)/(4) might well (perhaps should) feel like audacious assertions — unexplained results appearing out of nowhere. They felt that way to me. This is despite the fact that all of the concepts and notation appearing within (3)/(4) are carefully introduced, individually, by Sutton & Barto (2018) prior to them presenting equations (3)/(4).

The sense of mystery as to how the Bellman equation for state values arises increases in Chapter 4 of Sutton & Barto (2018). There, as part of introducing the subject of *policy evaluation* (i.e. the problem of estimating/predicting/approximating the state value function), Sutton & Barto (2018) assert that

$$\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi[G_t|S_t = s]\\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s]\\
&= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s] && (5)\\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a) \Big[r + \gamma v_\pi(s') \Big].
\end{aligned}$$

Equation (5) suddenly appears on the scene, but its presence is not discussed in any way. It's not clear how it arises or whether it's a necessary or optional step on the road to equations (3)/(4). Comparison with (2) implies that

$$\mathbb{E}_\pi[v_\pi(S_{t+1})|S_t = s] = \mathbb{E}_\pi[G_{t+1}|S_t = s],$$

but this is non-intuitive. The return, G_{t+1}, is a random variable. Likewise, since S_{t+1} is a random variable, the function of that random variable, $v_\pi(S_{t+1})$, is itself a random variable. But we know nothing of these two random variables conditioned on event $S_t = s$ or of why the expected values of their distributions should be the same. It may be that these two conditioned random variables are 'equal in distribution', or it may be that their distributions are very different but happen to share the same expected value.

If we combine this information from Chapters 3 and 4 of Sutton & Barto (2018), we have the sequence of equations

$$\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi[G_t|S_t = s] && (6)\\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] && (7)\\
&= \dots && (8)\\
&= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s] && (9)\\
&= \dots && (10)\\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a) \Big[r + \gamma \mathbb{E}_\pi[G_{t+1}|S_{t+1} = s'] \Big] && (11)\\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a) \Big[r + \gamma v_\pi(s') \Big] \qquad \forall\ s \in \mathcal{S}. && (12)
\end{aligned}$$

This document fills in the gaps represented by the ellipses in (8) and (10). It provides two separate but related derivations: one primary, one secondary. The primary derivation takes the reader from (6)/(7) straight through to (11)/(12) whilst ignoring (bypassing) (9). The secondary derivation first backtracks a bit to a late-stage point within the primary derivation and then forks to take a new direction that leads to (9).

I developed these derivations for me. I used the exercise as a learning tool — an opportunity to refresh and strengthen and record my grasp of probability. I like the mathematics of probability and feel comfortable with its basics. I am not a sophisticated probability guy for whom large, clever leaps of probabilistic reasoning are helpful or come naturally to mind. Finding maximally terse and eloquent derivations was not my motivation. The derivations I've developed and documented here are detailed and step-by-step, with lots of explanation. It's people like me who are most likely to find these derivations helpful. Sophisticated probability people are likely to find them verbose and tedious. Support for all of the probabilistic reasoning used in my derivations can be found in probability textbooks such as Blitzstein & Hwang (2015).

Having developed my derivations, I can understand why Sutton & Barto (2018) chose not to provide a derivation for the Bellman equation for state values of their own. Their purpose is to teach RL, not probability. And no one derivation could serve all readers well.

2 Concepts and notation

To help this document be self-contained and reduce the reader's reliance on Sutton & Barto (2018) in order to make sense of it, we briefly summarise here the concepts and notation used in the Bellman equation for state values and its derivation.

Time

discrete time steps: $t = 0, 1, 2, 3, \ldots$

States

the set of environmental states: $\mathcal{S}$

the random variable representing the environment's state at time t: $S_t \in \mathcal{S}$

particular states: $s, s' \in \mathcal{S}$

the state at time t (current state): $S_t = s$

the state at time $t+1$ (successor state): $S_{t+1} = s'$

Actions

the set of actions from which the agent may choose: $\mathcal{A}$

the set of actions available in state s: $\mathcal{A}(s)$

the random variable representing the action selected by the agent at time t: $A_t \in \mathcal{A}$

a particular action: $a \in \mathcal{A}(s)$

at time t, after having arrived in state $S_t = s$, the agent selects an action $A_t = a \in \mathcal{A}(s)$

Rewards

the set of numerical rewards: $\mathcal{R} \subset \mathbb{R}$

the random variable representing the reward received by the agent at time t: $R_t \in \mathcal{R}$

a particular reward value: $r \in \mathcal{R}$

as a consequence of arriving in state $S_t = s$ and selecting action $A_t = a$, one time step later the agent receives reward $R_{t+1} = r$ and finds itself in a new (successor) state, $S_{t+1} = s'$

Trajectory of a Markov Decision Process

the Markov Decision Process (MDP) and the agent together give rise to a sequence or trajectory of states, actions and rewards that looks like this:

$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, S_3, A_3, R_4, \ldots$

Dynamics of a Markov Decision Process

$p(s', r|s, a) \doteq Pr\{S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a\} \in [0, 1]$

p specifies a probability distribution for each state-action pair, s and a, such that

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r|s, a) = 1, \qquad \forall\ s \in \mathcal{S}, a \in \mathcal{A}(s)$$

.

p is said to fully characterise the *dynamics* of the environment (the MDP) because, through it, each state-action pair, S_t and A_t, fully determines the probability of each succeeding possible value for S_{t+1} and R_{t+1}. That is, past (historic) states and actions have no influence on what happens next. This is the characteristic (memoryless) *Markov property*, where the present alone determines the future, not the past.

Returns

the agent receives a sequence of rewards according to the trajectory of the Markov Decision Process (MDP): $R_1, R_2, R_3, R_4, \ldots, R_{t-1}, R_t, R_{t+1}, \ldots$

the random variable denoting the *return* at time t: G_t

the *return*, G_t, is a sum of discounted future rewards:

$$\begin{aligned} G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \ldots \\ &= R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \ldots) \\ &= R_{t+1} + \gamma G_{t+1}. \end{aligned}$$

the *discount rate*, $\gamma \in [0, 1]$, is a constant that determines the present value of future rewards

the agent seeks to learn to select actions that maximise the *expected return*

Policies

a *policy*, denoted π, guides/governs the agent's action selection by mapping states s to probabilities of selecting actions a

if the agent is following policy π at time t, then $\pi(a|s)$ is the probability that $A_t = a$ if $S_t = s$; more broadly, $\pi(a|s)$ is a probability distribution over $a \in \mathcal{A}(s)$ for each $s \in \mathcal{S}$:

$$\pi(a|s) \doteq Pr(A_t = a | S_t = s) \in [0, 1]$$

such that $\sum_{a \in \mathcal{A}(s)} \pi(a|s) = 1$

State values

the *value* of a state s under a policy π, denoted $v_\pi(s)$, is defined to be the *expected return* when starting in state s and following policy π thereafter:

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t | S_t = s] \qquad \text{and} \qquad v_\pi(s') \doteq \mathbb{E}_\pi[G_{t+1} | S_{t+1} = s'],$$

where $\mathbb{E}_\pi[\cdot]$ denotes the expected value of a random variable given that the agent follows policy π, and t is any time step

the function v_π is called the *state value function for policy* π

3 The derivation

My derivation relies on the application of several concepts from basic probability theory, including:

- the definitions of expectation and conditional expectation, and their properties such as the LOTUS (law of the unconscious statistician)
- the definition of conditional probability and its generalisations
- the product rule (sometimes called the 'chain rule' or 'general law of multiplication')
- marginalisation (to eliminate random variables)
- reverse marginalisation (to introduce random variables)
- the concept of "extra conditioning"
- the independence of events and its implications for conditional probabilities
- the Markov property (assumption) of MDPs and its implications.

We begin by writing

$$\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi[G_t|S_t = s] && \text{by definition of state value}\\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] && \text{by definition of } G_t\\
&= \mathbb{E}_\pi[R_{t+1}|S_t = s] + \mathbb{E}_\pi[\gamma G_{t+1}|S_t = s] && (13)\\
&= \mathbb{E}_\pi[R_{t+1}|S_t = s] + \gamma\mathbb{E}_\pi[G_{t+1}|S_t = s], && (14)
\end{aligned}$$

where (13) and (14) are justified by the linearity of expectation and, by extension, of conditional expectation.

To keep things simple (and to conserve horizontal space), from here onward we work with the two expectation components of the RHS of (14) individually and separately. For clarity, we name these two components *expectation component 1* and *expectation component 2*, respectively. At the end of the derivation, we bring the separately derived expansions of these two expectation components together to construct the final derivation results. Throughout the remainder of the derivation, we also employ a space-conserving notation convention whereby events, such as $S_t = s$ and $R_{t+1} = r$, are represented by the values of the random variables alone, such as, in these two examples, s and r, respectively.

3.1 Expansion of expectation component 1

We begin our expansion of *expectation component 1* by writing

$$\begin{aligned}
\mathbb{E}_\pi[R_{t+1}|S_t = s] &= \sum_r \; p(r|s)\; r && \text{by definition of expectation} \\
&= \sum_r r\; p(r|s) \\
&= \sum_r r \sum_{s'} p(s', r|s) && \text{by reverse marginalisation} \\
&= \sum_r r \sum_{s'} \sum_a p(s', r, a|s) && \text{by reverse marginalisation} \\
&= \sum_a \sum_{s'} \sum_r p(s', r, a|s)\; r. && (15)
\end{aligned}$$

Now, in probability theory the concept of conditional probability is defined using two random variables (or events), one being conditioned and the other doing the conditioning. It is known that this definition generalises to three random variables (or events) in the following way:

$$P(A|B) = \frac{P(A,B)}{P(B)} \qquad \Longrightarrow \qquad P(A,B|C) = \frac{P(A,B,C)}{P(C)}. \tag{16}$$

It is also known that to any valid probabilistic relation one can apply "extra conditioning", so long as it is done consistently to all terms in the relation. Hence, we can introduce an additional random variable, D, to the RHS of (16) by writing

$$P(A,B|C) = \frac{P(A,B,C)}{P(C)} \qquad \Longrightarrow \qquad P(A,B|C,D) = \frac{P(A,B,C|D)}{P(C|D)}, \tag{17}$$

which, via algebra, yields the result

$$P(A,B,C|D) = P(A,B|C,D)P(C|D). \tag{18}$$

Using result (18), we can take the probability expression in (15) and re-express it as follows

$$p(s', r, a|s) = p(s', r|a, s)p(a|s) = p(s', r|s, a)p(a|s) = p(s', r|s, a)\pi(a|s). \tag{19}$$

Notice, here, that we also took the opportunity to introduce the use of the notation π by representing $p(a|s)$ as $\pi(a|s)$, per the convention used by Sutton & Barto (2018).

We can now return to (15) and continue our expansion by writing

$$\begin{aligned}
\mathbb{E}_\pi[R_{t+1}|S_t = s] &= \sum_a \sum_{s'} \sum_r p(s', r, a|s)\; r \\
&= \sum_a \sum_{s'} \sum_r p(s', r|s, a)\pi(a|s)\; r && \text{by (19)} \\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)\; r. && (20)
\end{aligned}$$

This concludes our expansion of *expectation component 1* in the RHS of (14).

3.2 Expansion of expectation component 2

Now we turn our attention to expanding *expectation component 2* of the RHS of (14). We begin by writing

$$\begin{aligned}
\gamma\mathbb{E}_\pi[G_{t+1}|S_t = s] &= \gamma \sum_{g_{t+1}} g_{t+1}\ p(g_{t+1}|s) && \text{by def. of expectation} \\
&= \gamma \sum_{g_{t+1}} g_{t+1} \sum_{s'} p(s', g_{t+1}|s) && \text{by rev. marg.} \\
&= \gamma \sum_{g_{t+1}} g_{t+1} \sum_{s'} \sum_{r} p(s', r, g_{t+1}|s) && \text{by rev. marg.} \\
&= \gamma \sum_{g_{t+1}} g_{t+1} \sum_{s'} \sum_{r} \sum_{a} p(s', r, a, g_{t+1}|s) && \text{by rev. marg..} \qquad (21)
\end{aligned}$$

To make it clear how it is we can take (21) further, we first generalise the conditional probability and "extra conditioning" arguments given in (17) and (18) by introducing a fifth variable. We can assert that

$$P(A, B, C|D) = \frac{P(A, B, C, D)}{P(D)} \quad \Longrightarrow \quad P(A, B, C|D, E) = \frac{P(A, B, C, D|E)}{P(D|E)}, \qquad (22)$$

which implies that

$$P(A, B, C, D|E) = P(A, B, C|D, E)P(D|E). \qquad (23)$$

Picking up from (21), we can thus write

$$\begin{aligned}
\gamma\mathbb{E}_\pi[G_{t+1}|S_t = s] &= \gamma \sum_{g_{t+1}} g_{t+1} \sum_{s'} \sum_{r} \sum_{a} p(s', r, a, g_{t+1}|s) \\
&= \gamma \sum_{g_{t+1}} g_{t+1} \sum_{s'} \sum_{r} \sum_{a} p(s', r, a|g_{t+1}, s)p(g_{t+1}|s), \qquad (24)
\end{aligned}$$

where (24) is justified by (23).

The two probability components in (24) need further re-expression. For clarity, we give them names. We refer to $p(g_{t+1}|s)$ as *probability component A* and to $p(s', r, a|g_{t+1}, s)$ as *probability component B*. We develop their re-expressions individually, in turn.

3.2.1 Re-expression of probability component A

It is helpful to develop the re-expression of probability component A of (24), $p(g_{t+1}|s)$, in two stages. Each stage benefits from a bit of preliminary discussion.

Stage 1 To motivate and set up the first stage, we discuss a general rule in probability theory sometimes called the "product rule" (also variously called in the literature the "chain rule", the "general law of multiplication", and other names). Blitzstein & Hwang (2015) describe this rule using theorems but without assigning any name. The rule describes a general approach to factoring (expressing) joint probabilities in terms of products of conditional and marginal probabilities. The most basic form of the product rule flows directly (via algebra) from the definition of conditional probability. The two statements (definitions)

$$P(A|B) = \frac{P(A,B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(A,B)}{P(A)}$$

together imply that

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A). \tag{25}$$

Statement (25) of the product rule shows that with 2 events there are $2! = 2$ possible ways to factor the joint probability of those events into products of conditional and marginal probabilities. In general, with n events there are $n!$ unique ways to factor the joint probability of the events into products of conditional and marginal probabilities. For example, if we have a joint probability with 3 events, $P(A,B,C)$, one of the $3! = 6$ ways to factor it is as follows:

$$P(A,B,C) = P(A|B,C)P(B|C)P(C). \tag{26}$$

Now we're ready to implement the first stage in re-expressing *probability component A* of (24). We do this by writing

$$\begin{aligned}
p(g_{t+1}|s) &= \sum_{s'} p(g_{t+1}, s'|s) && \text{by rev. marginalisation} && (27)\\
&= \sum_{s'} \frac{p(g_{t+1}, s', s)}{p(s)} && \text{by def. conditional prob.} && (28)\\
&= \sum_{s'} \frac{p(g_{t+1}|s', s)p(s'|s)p(s)}{p(s)} && \text{by product rule, per (26)} && (29)\\
&= \sum_{s'} p(g_{t+1}|s', s)p(s'|s) && && (30)\\
&= p(g_{t+1}|s', s)\sum_{s'} p(s'|s) && && (31)\\
&= p(g_{t+1}|s', s), && && (32)
\end{aligned}$$

where (32) is justified by marginalisation, since $\sum_{s'} p(s'|s) = 1$.

Stage 2 The second stage of re-expressing *probability component A* of (24), $p(g_{t+1}|s)$, involves taking (32) one step further. Doing this requires discussion of the *Markov property.*

To facilitate careful consideration of the probability component $p(g_{t+1}|s', s)$ in (32), it helps to relax our space-conserving notation convention for a moment and write it fully, as

$$P(G_{t+1} = g_{t+1} \mid S_{t+1} = s', S_t = s), \tag{33}$$

so that its three events can be interpreted precisely.

The expression in (33) refers to the probability that the *return* random variable,

$$G_{t+1} = R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \dots,$$

will take on a certain value given that the events $S_{t+1} = s'$ and $S_t = s$ have both been observed. That is, time steps t and $t+1$ have occurred, and the value of random variable G_{t+1} depends on rewards yet to received in time steps $t + 2$, $t + 3$, $t + 4$, etc.. Or, in other words, we can think of time steps $t + 2$, $t + 3$, $\dots$, (and, thus, the value of random variable G_{t+1}) as representing the *future*, time step $t + 1$ as representing the *present*, and time step t as representing the *past*. But we know that MDPs assume that the *Markov property* always applies, which means that the *present* alone determines the *future* and that the *past* is irrelevant with respect to what happens next. Hence, we can confidently assert that the *future* event $G_{t+1} = g_{t+1}$ and the *past*, conditioning event $S_t = s$ are (statistically) *independent* of one another. By definition, if events A and B are *independent*, then $P(A|B) = P(A)$. Thus, by the property of independence, we are permitted to drop the (redundant) conditioning event $S_t = s$ and assert that

$$P(G_{t+1} = g_{t+1} \mid S_{t+1} = s', S_t = s) \; = \; P(G_{t+1} = g_{t+1} \mid S_{t+1} = s').$$

Thus, we may implement the second stage of, and conclude, our re-expression of *probability component A* in (24) by picking up from (32) and writing

$$\begin{aligned} p(g_{t+1}|s) &= p(g_{t+1}|s', s) \\ &= p(g_{t+1}|s') \qquad \text{by independence due to Markov property.} \end{aligned} \tag{34}$$

3.2.2 Re-expression of probability component B

Now we turn to re-expressing *probability component B* in (24), $p(s', r, a|g_{t+1}, s)$. Relaxing our space-conserving notation convention once again in order to fully express its events, this component becomes

$$P(S_{t+1} = s', R_{t+1} = r, A_t = a \mid G_{t+1} = g_{t+1}, S_t = s).$$

As explained earlier, the event $G_{t+1} = g_{t+1}$ represents a sum of discounted rewards pertaining to time steps $t + 2$, $t + 3$, etc.. As such, it is a *future* event relative to the latest (*present*) time step $t + 1$ and its associated events $S_{t+1} = s'$ and $R_{t+1} = r$, and relative to the *past* time step, t, and its associated event, $A_t = a$. In other words, because it is in the *future*, event $G_{t+1} = g_{t+1}$ is (intuitively) *independent* of the three events it is conditioning, whether we consider those events individually or jointly. Thus, by the property of *independence*, we may drop the conditioning event $G_{t+1} = g_{t+1}$ from *probability component B* and write, in our space-conserving notation, that

$$p(s', r, a|g_{t+1}, s) = p(s', r, a|s). \tag{35}$$

Note that our assertion of the *independence* of events in this case has nothing to do with the Markov property. It flows from intuition, based on the notion of *time* alone. Another way of expressing the observation we are making here (our reasoning) is to say simply that *conditioning on a future event is redundant.*

3.2.3 Concluding expansion of expectation component 2

Having re-expressed *probability components* A and B of equation (24), in (34) and (35), respectively, we now continue our expansion of expectation component 2 from (14) by picking up from where we left off in (24). We do so by writing

$$
\begin{aligned}
\gamma\mathbb{E}_\pi[G_{t+1}|S_t = s] &= \gamma \sum_{g_{t+1}} g_{t+1} \sum_{s'}\sum_{r}\sum_{a} p(s', r, a|g_{t+1}, s)p(g_{t+1}|s) \\
&= \gamma \sum_{g_{t+1}} g_{t+1} \sum_{s'}\sum_{r}\sum_{a} p(s', r, a|s)p(g_{t+1}|s) \qquad \text{by (35)} \qquad (36) \\
&= \gamma \sum_{g_{t+1}} g_{t+1} \sum_{s'}\sum_{r}\sum_{a} p(s', r, a|s)p(g_{t+1}|s') \qquad \text{by (34)} \qquad (37) \\
&= \sum_{s'}\sum_{r}\sum_{a} p(s', r, a|s)\gamma \sum_{g_{t+1}} g_{t+1}p(g_{t+1}|s') \qquad (38) \\
&= \sum_{s'}\sum_{r}\sum_{a} p(s', r, a|s)\gamma\mathbb{E}_\pi[G_{t+1}|S_{t+1} = s'], \qquad (39)
\end{aligned}
$$

where (39) is justified by the definition of conditional expectation.

As already shown in the expansion of expectation component 1, we may reuse (18) and (19) to write

$$p(s', r, a|s) = p(s', r|s, a)\pi(a|s),$$

and, hence, we may continue our expansion of expectation component 2 from (39), and conclude it, by writing

$$
\begin{aligned}
\gamma\mathbb{E}_\pi[G_{t+1}|S_t = s] &= \sum_{s'}\sum_{r}\sum_{a} p(s', r, a|s)\gamma\mathbb{E}_\pi[G_{t+1}|S_{t+1} = s'] \\
&= \sum_{s'}\sum_{r}\sum_{a} p(s', r|s, a)\pi(a|s)\gamma\mathbb{E}_\pi[G_{t+1}|S_{t+1} = s'] \qquad (40) \\
&= \sum_{a} \pi(a|s) \sum_{s'}\sum_{r} p(s', r|s, a)\ \gamma\mathbb{E}_\pi[G_{t+1}|S_{t+1} = s']. \qquad (41)
\end{aligned}
$$

3.3 Concluding the primary derivation

Having completed our expansions of *expectation components 1* and *2* from (14), we may now combine the results, from (20) and (41), respectively, and realise our primary objective of deriving the Bellman equation for state values expressed in equations (11) and (12). To provide full context, we resume our derivation from the very beginning, with equation (6). We may now write that

$$\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi[G_t|S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1}|S_t = s] + \gamma\mathbb{E}_\pi[G_{t+1}|S_t = s] \\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)\; r \;\; + && \text{by (20)} \\
&\quad \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)\; \gamma\mathbb{E}_\pi[G_{t+1}|S_{t+1} = s'] && \text{by (41)} \\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)\Big[r + \gamma\mathbb{E}_\pi[G_{t+1}|S_{t+1} = s']\Big] && (42) \\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)\Big[r + \gamma v_\pi(s')\Big]. && (43)
\end{aligned}$$

The reader is invited to confirm that, as desired, equation (42) is identical to (11) and equation (43) is identical to (12). This concludes our derivation of the Bellman equation for state values that is used so extensively within RL.

3.4 Concluding the secondary derivation

To realise our secondary objective of finding a justifying derivation for the intermediate equation (9) that Sutton & Barto (2018) claim lies on the road to the Bellman equation for state values, we need an alternate expression for *expectation component 2* of (14). We can find the alternate expression we need if we return (backtrack) to equation (39) and take it in a new direction. Picking up from (39), we may write

$$\begin{aligned}
\gamma\mathbb{E}_\pi[G_{t+1}|S_t = s] &= \sum_{s'} \sum_r \sum_a p(s', r, a|s)\gamma\mathbb{E}_\pi[G_{t+1}|S_{t+1} = s'] \\
&= \sum_{s'} \sum_r p(s', r|s)\gamma\mathbb{E}_\pi[G_{t+1}|S_{t+1} = s'] && \text{by marg.} \\
&= \sum_{s'} p(s'|s)\gamma\mathbb{E}_\pi[G_{t+1}|S_{t+1} = s'] && \text{by marg.} \\
&= \sum_{s'} p(s'|s)\gamma v_\pi(s') && \text{by def. of state value} \\
&= \gamma \sum_{s'} p(s'|s) v_\pi(s') \\
&= \gamma\mathbb{E}_\pi[v_\pi(S_{t+1})|S_t = s] && (44) \\
&= \mathbb{E}_\pi[\gamma v_\pi(S_{t+1})|S_t = s], && (45)
\end{aligned}$$

where: 1) (44) is justified by the definition of conditional expectation and the property whereby the LOTUS (law of the unconscious statistician) can be applied to that definition, and 2) (45) is justified by the linearity of expectation, since γ is a constant.

Our alternate expression for *expectation component 2* of (14), given in (44)/(45), shows that, as we surmised in Section 1, the assertion by Sutton & Barto (2018) of equation (9) does indeed imply that

$$\mathbb{E}_\pi[v_\pi(S_{t+1})|S_t = s] = \mathbb{E}_\pi[G_{t+1}|S_t = s].$$

Further, our derivation of (44)/(45) shows that, as we surmised, the equality of the expected values of these two distributions is highly non-intuitive.

We may now use our alternate expression for *expectation component 2* of equation (14), given in (45), to conclude our derivation of intermediate equation (9). We do so by writing

$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}_\pi[G_t|S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1}|S_t = s] + \gamma\mathbb{E}_\pi[G_{t+1}|S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1}|S_t = s] + \mathbb{E}_\pi[\gamma v_\pi(S_{t+1})|S_t = s] \qquad \text{by (45)} \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s]. \end{aligned} \tag{46}$$

The reader is invited to confirm that, as desired, equation (46) is identical to equation (9). This concludes our secondary derivation, that of intermediate equation (9).

3.5 Addressing a potential source of confusion

Here I address a certain aspect of my derivation that at one point caused me some confusion and anxiety. I don't want readers to suffer the same concerns.

Subsection 3.2.1 shows that $p(g_{t+1}|s) = p(g_{t+1}|s')$, as given in (34). The second stage of the argument deployed there reasons that $p(g_{t+1}|s', s) = p(g_{t+1}|s')$ because of the Markov property and statistical independence. Specifically, we reason that event $S_t = s$ is in the *past* and that event $G_{t+1} = g_{t+1}$ is in the *future* and that, due to the MDP assumption that the Markov property always holds, these two events are therefore *independent* and hence we can drop the conditioning event $S_t = s$ from the expression $p(g_{t+1}|s', s)$.

Not long after having convinced myself of the soundness of this reasoning, I started to worry that perhaps the same reasoning must therefore mean that $p(g_{t+1}|s) = p(g_{t+1})$. If this were true, then my derivation falls apart because the conclusion in (34) that $p(g_{t+1}|s) = p(g_{t+1}|s')$ must then progress to the conclusion that $p(g_{t+1}) = p(g_{t+1}|s')$, which is false, because events $G_{t+1} = g_{t+1}$ and $S_{t+1} = s'$ are clearly not independent. For a time, I feared that my derivation led to a fatal contradiction.

Happily, I was soon able to convince myself that all was OK. It is *not* the case that $p(g_{t+1}|s) = p(g_{t+1})$. Yes, event $S_t = s$ is in the *past* and event $G_{t+1} = g_{t+1}$ is in the *future*, but here the *present* (event $S_{t+1} = s'$) is not represented. Because the present has not been observed, the future event is not independent of the past event. A dependency remains, so the conditioning event $S_t = s$ cannot be dropped from the probability expression $p(g_{t+1}|s)$.

4 Alternate derivations

4.1 Cross Validated

One question/discussion on Cross Validated contains several contributions relating to deriving the Bellman equation for state values, per Sutton & Barto (2018). All of the contributions there are much more terse than mine, most of them dramatically so. So most of them don't help me much. But there is an exception. One contribution is *remarkably* close to my approach — almost like a condensed version of my derivation. It thus provides strong corroboration for my approach that I value highly. But that contribution also differs from my approach in interesting ways that illustrate the flexibility of probabilistic reasoning and provide an instructive alternate perspective. None of the contributions in that question/discussion, however, address the derivation of the intermediate equation (9) from Sutton & Barto (2018).

The author of the contribution that is similar to my derivation writes, as I do, that

$$p(g_{t+1}|s) = \sum_{s'}\sum_{r}\sum_{a} p(s', r, a, g_{t+1}|s),$$

by reverse marginalisation. But he then reasons differently from me. His reasoning makes a leap that was too big for me, at first. Here I discuss that leap by breaking it down and recording my analysis. The author writes that

$$\begin{aligned} p(s', r, a, g_{t+1}|s) &= p(g_{t+1}|s', r, a, s)p(s', r, a|s) && (47) \\ &= p(g_{t+1}|s', r, a, s)p(s', r|a, s)\pi(a|s) && (48) \\ &= p(g_{t+1}|s')p(s', r|a, s)\pi(a|s). && (49) \end{aligned}$$

The author justifies (49) by referring to the Markov property and pointing out that all three of the conditioning events $R_{t+1} = r$, $A_t = a$ and $S_t = s$, in (48), are independent of the future event $G_{t+1} = g_{t+1}$, and can therefore be dropped from $p(g_{t+1}|s', r, a, s)$. I get that part. The leap I couldn't get is (47).

I think (47) can be described as a generalisation of what I refer to as the "product rule" — a particular application of the rule whereby multiple variables are shifted across the conditioning symbol, $(|)$, all at once, rather than one at a time in series. But I haven't seen this done before so I want to walk through this one step at a time to better learn the pattern at play. To do so, I've chosen to use generalised random variables.

We wish to show that

$$P(A, B, C, D|E) = P(D|A, B, C, E)P(A, B, C|E).$$

Going step-by-step, shifting only one variable at a time, we can write

$$\begin{aligned}
P(A,B,C,D|E) &= P(A,B,D|C,E)P(C|E) && \text{shift C} \quad (50)\\
&= P(A,D|B,C,E)P(B|C,E)P(C|E) && \text{shift B} \quad (51)\\
&= P(D|A,B,C,E)P(A|B,C,E)P(B|C,E)P(C|E) && \text{shift A} \quad (52)\\
&= P(D|A,B,C,E)\frac{P(A,B,C,E)}{P(B,C,E)}\frac{P(B,C,E)}{P(C,E)}\frac{P(C,E)}{P(E)} && \quad (53)\\
&= P(D|A,B,C,E)\frac{P(A,B,C,E)}{P(E)} && \quad (54)\\
&= P(D|A,B,C,E)P(A,B,C|E). && \quad (55)
\end{aligned}$$

So the leap the author makes in (47) looks sound! And we have learned more about the "product rule" and how it can be generalised.

4.2 Beware of derivations based on the LOIE

I have come across *purported* derivations of the Bellman equation for state values that are succinct and that achieve their succinctness through what they *claim* to be an application of the *Law of Iterated Expectations* (LOIE). I believe the validity of such derivations may be highly suspect, for two reasons: 1) they may apply the LOIE in contexts to which it does not pertain, and 2) they may use invalid forms of the LOIE.

The LOIE, also called *Adam's law* by Blitzstein & Hwang (2015), asserts that

$$E[E[Y|X]] = E[Y]. \qquad (56)$$

Here, both X and Y are random variables, and hence $E[Y|X]$ is a random variable. This is very different from $E[Y|X=x]$, where $X=x$ is an event and hence $E[Y|X=x]$ is a real number. Taking an expectation of a random variable, as in $E[E[Y|X]]$, makes sense; taking an expectation of a real number, as in $E[E[Y|X=x]]$, does not, since $E[c]=c$.

Blitzstein & Hwang (2015) also define another form of Adam's law (the LOIE) that involves "extra conditioning". They start with (56) and condition everything on a third random variable, Z, giving

$$E[E[Y|X,Z]|Z] = E[Y|Z]. \qquad (57)$$

Note how Z is introduced in three places, not just two: it conditions the random variable Y on both sides of the equality sign, and it also conditions the random variable $E[Y|X,Z]$.

One *purported* derivation of the Bellman equation for state values I encountered (the URL for which no longer exists) cited the LOIE to justify the following assertion:

$$E[E[G_{t+1}|S_{t+1}]|S_t] = E[G_{t+1}|S_t]. \qquad (58)$$

Equation (58) is invalid, for two reasons. First, structurally it is malformed. It looks like the author is trying to use the form of the LOIE given in (57), but the "extra conditioning" with S_t has not been applied properly. There should be three instances of S_t, not two. Second, the author is attempting to invoke the LOIE in an inappropriate context. This becomes clear when he defines some of his notation. At one point he explains that his use of S_t really means $S_t = s$ (i.e. that S_t denotes an event, not a random variable). But, as I have explained, the LOIE is defined only in terms of random variables, not events.

Equation (58) was key to the author's purported derivation, so the fact that it is badly flawed means the whole derivation was fatally flawed. The moral of the story is: be wary of derivations of the Bellman equation for state values that rely on the LOIE. I'm not claiming that it's impossible to use the LOIE correctly to derive the Bellman equation for state values. It may well be feasible to do so. I'm just cautioning people that I've seen the LOIE be misused, resulting in purported Bellman equation derivations that are specious.

References

Blitzstein, J. & Hwang, J. (2015). Introduction to Probability, CRC Press.

Sutton, R. & Barto, A. (2018). Reinforcement Learning: An Introduction, 2nd edition, The MIT Press.