Classical mechanics as reinforcement learning

Matthew Weiss

September 3, 2026

Last time on the Postdoc Blog, we discussed the *Bellman optimality equation*,

$$V(s) = \max_a \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s') \right], \tag{0.1}$$

where $P(s'|s,a)$ is the probability of the environment transitioning to state s' given that the agent takes action a in state s, $R(s,a)$ is the reward for the agent if they perform a in state s, and $V(s)$ is the optimal state value: it is the expected total future return obtained from starting in state s and acting optimally thereafter [1, 2]. Here γ is the temporal discounting factor.

Now let's suppose that $\gamma = 1$ and the state transition is deterministic. We are left with

$$V(s) = \max_a \Big[R(s,a) + V(s') \Big], \tag{0.2}$$

where s' is the state achieved by performing a in state s. Our objective: to take the continuum limit of this equation. To this end, take the state to be $s = (x,t)$, where $x \equiv x(t)$ and t are real valued: t will be a time coordinate and $x(t)$ will be a trajectory of in one dimension (for simplicity). The agent's choice of action in each little moment of time will be represented by a function $u \equiv u(x,t)$, and the goal is to find the optimal u and the corresponding trajectory. We'll model the response of the environment to the agent's action by some function $f(x,t,u)$, imposing that

$$\frac{dx(t)}{dt} = f(x,t,u). \tag{0.3}$$

In other words, in each little time step dt, the agent takes a little action u, and the environment responds by curving the trajectory through the state space: in a little interval dt, the state transitions from $s = (x,t) \to s' = (x + dx, t + dt)$. Upon choosing an action, the agent receives a tiny little reward from the environment modeled by some function $r(x,t,u)$,

$$\frac{dR(x,t,u)}{dt} = r(x,t,u). \tag{0.4}$$

Putting this all together, we can write the Bellman optimality equation as

$$V(x,t) = \max_u \Big[r(x,u)dt + V(x + dx, t + dt) \Big]. \tag{0.5}$$

To first order, we have

$$V(x + dx, t + dt) = V(x,t) + dV, \tag{0.6}$$

where

$$dV = \nabla V \cdot ds = \begin{pmatrix} \frac{\partial V}{\partial x} & \frac{\partial V}{\partial t} \end{pmatrix} \begin{pmatrix} dx \\ dt \end{pmatrix} = \frac{\partial V}{\partial x} dx + \frac{\partial V}{\partial t} dt \tag{0.7}$$

$$= \frac{\partial V}{\partial x} f(x,t,u) dt + \frac{\partial V}{\partial t} dt = \left(\frac{\partial V}{\partial x} f(x,t,u) + \frac{\partial V}{\partial t} \right) dt. \tag{0.8}$$

Plugging this back into the Bellman equation, we have

$$V(x,t) = \max_u \left[r(x,t,u)dt + V(x,t) + \left(\frac{\partial V}{\partial x} f(x,t,u) + \frac{\partial V}{\partial t} \right) dt \right], \tag{0.9}$$

so that subtracting $V(x,t)$ from both sides and dividing through by dt, we obtain

$$\frac{\partial V}{\partial t} + \max_u \left[r(x,t,u) + \frac{\partial V}{\partial x} f(x,t,u) \right] = 0. \tag{0.10}$$

This is our continuum generalization of the Bellman equation.

To find the maximum, let's take the derivative with respect to u and set it to zero.

$$\frac{\partial}{\partial u} \left[r(x,t,u) + \frac{\partial V}{\partial x} f(x,t,u) \right] = \frac{\partial r(x,t,u)}{\partial u} + \frac{\partial V}{\partial x} \frac{\partial f(x,t,u)}{\partial u} = 0. \tag{0.11}$$

We conclude that at an extremum

$$\frac{\partial V}{\partial x} = -\left(\frac{\partial f(x,t,u)}{\partial u} \right)^{-1} \frac{\partial r(x,t,u)}{\partial u} \tag{0.12}$$

$$\frac{\partial V}{\partial t} = \left(\frac{\partial f(x,t,u)}{\partial u} \right)^{-1} \frac{\partial r(x,t,u)}{\partial u} f(x,t,u) - r(x,t,u), \tag{0.13}$$

where it is safe to take the reciprocal since an f that doesn't depend on u is not really worth considering.

With this in hand, let us at last specify how the environment reacts to the agent's action. In general, this will depend on the control problem one wishes to solve. But let us make a very special choice. Take

$$f(x,t,u) = u, \tag{0.14}$$

so that since $f(x,t,u) = dx/dt$, we have

$$u = \frac{dx}{dt}. \tag{0.15}$$

In other words, in each tiny moment of time, the agent gets to choose the velocity of its trajectory. For the reward, we take

$$r(x,t,u) = -\frac{1}{2} m u^2 + U(x). \tag{0.16}$$

With these particular choices, we have

$$\frac{\partial V}{\partial x} = mu \tag{0.17}$$

$$\frac{\partial V}{\partial t} = -\frac{1}{2} m u^2 - U(x). \tag{0.18}$$

Those familiar with classical mechanics will have noticed that since $u = dx/dt$ is the velocity, we can interpret $\frac{1}{2}mu^2$ as the *kinetic energy* of a particle, $U(x)$ as specifying a *potential*, $mu = p$ as the *momentum*, $\frac{1}{2}mu^2 + U(x)$ as the total energy, or *Hamiltonian* H, and $\frac{1}{2}mu^2 - U(x)$ as the *Lagranian* L, the difference between kinetic and potential energy [3, 4]. We conclude

$$L(x,t,u) = -r(x,t,u) \qquad \frac{\partial V}{\partial x} = p \qquad \frac{\partial V}{\partial t} = -H(x,t,u). \tag{0.19}$$

In fact, these latter two relations tells us that $V \equiv S$, the classical *action*, which is usually defined as the integral of the Lagranian along a trajectory,

$$S = \int_{t_i}^{t_f} L dt. \tag{0.20}$$

In particular, $\partial S/\partial t = -H$ is known as the *Hamilton-Jacobi equation*. To further cement the connection, notice that we can now rewrite

$$\frac{\partial V}{\partial t} + \max_u \left[r(x,t,u) + \frac{\partial V}{\partial x} f(x,t,u) \right] = 0. \tag{0.21}$$

as

$$\frac{\partial S}{\partial t} + \max_u \left[pu - L(x,t,u) \right] = 0, \tag{0.22}$$

where the second term is precisely the *Legendre transform* of the Lagranian [5] which as is well known yields the Hamiltonian

$$H(x,t,p) = \max_u \left[pu - L(x,u) \right], \tag{0.23}$$

of the system.

But even if we don't know any classical mechanics, we can solve the problem at hand directly. The trick is to take some more partial derivatives. On the one hand,

$$\frac{\partial^2 V}{\partial x \partial t} = \frac{\partial}{\partial t}\left[mu \right] = m\frac{\partial u}{\partial t}. \tag{0.24}$$

On the other hand,

$$\frac{\partial^2 V}{\partial t \partial x} = \frac{\partial}{\partial x}\left[-\frac{1}{2}mu^2 - U(x) \right] = -mu\frac{\partial u}{\partial x} - \frac{dU(x)}{dx}, \tag{0.25}$$

where we've used

$$\frac{\partial u^2}{\partial x} = \frac{\partial u}{\partial x}u + u\frac{\partial u}{\partial x} = 2u\frac{\partial u}{\partial x}. \tag{0.26}$$

But since partial derivatives commute, we can set these two expressions equal to each other, which leads to

$$-\frac{dU(x)}{dx} = m\left[\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} \right]. \tag{0.27}$$

We can simplify the expression in the brackets by noticing that

$$du = \nabla u \cdot ds = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial t} \end{pmatrix} \begin{pmatrix} dx \\ dt \end{pmatrix} = \frac{\partial u}{\partial x}dx + \frac{\partial u}{\partial t}dt, \tag{0.28}$$

from which, dividing by dt, we conclude that

$$\frac{du}{dt} = \frac{\partial u}{\partial x}\frac{dx}{dt} + \frac{\partial u}{\partial t}. \tag{0.29}$$

This is great because $u = dx/dt$, and so

$$-\frac{dU(x)}{dx} = m\frac{du}{dt} = m\frac{d^2x}{dt^2}. \tag{0.30}$$

But the negative derivative of a potential is a *conservative force*, and the second derivative of position with respect to time is the *acceleration*. So what we have derived is nothing other than

$$F = ma, \tag{0.31}$$

that is, Newton's second law.

For this reason, the equation

$$\frac{\partial V}{\partial t} + \max_u \left[r(x,t,u) + \frac{\partial V}{\partial x} f(x,t,u) \right] = 0 \tag{0.32}$$

is sometimes known as the *Hamilton-Jacobi-Bellman* equation. As we have seen, the equations of classical mechanics can be derived from the Bellman optimality equation in the continuum limit, where the agent's choice of action is their instantaneous velocity, and where their reward balances what physicists would call kinetic and potential energies. In this setting, Bellman optimality is just the principle of least action. Instead of calculating the action by integrating the Lagrangian over a trajectory and then varying it to find an extremum, the very same dynamics falls out naturally from the recursivity of the Bellman value equation. Classical mechanics is just a special case of a Bellman control problem.

A coupling of parting notes. More generally, the solutions of classical mechanics satisfy the principle of *stationary* action: in fact, our derivation is compatible with this, since all we did was take a first derivative and set it to 0. We have focused on the single variable case, but the extension to multiple spatial coordinates is straightforward. Moreover, we can consider the case that the response of the environment to the agent's choices is nondeterministic. To do this, one specifies

$$dx = f(x,t,u)dt + \sigma(x,t,u)dW_t, \tag{0.33}$$

where dW_t is a Wiener increment [3, 6]. Then the more general multidimensional stochastic Hamilton-Jacobi-Bellman equation is

$$\frac{\partial V}{\partial t} + \max_u \left[r + \nabla V \cdot f + \frac{1}{2}\mathrm{tr}\Big(\sigma\sigma^T \nabla^2 V\Big)\right] = 0, \tag{0.34}$$

where both f and σ are vector valued and $\nabla^2 V$ is the Hessian of V.

One final parting note: suppose the agent chooses their velocity at each moment with reward

$$r(x,t,u) = -\frac{1}{2} g_{ij} u^i u^j, \tag{0.35}$$

where g_{ij} is a metric on a Riemannian manifold [7]. This is the generalization of $r(x,t,u) = -\frac{1}{2}mu^2$ to a curved space. Applying the Bellman optimality principle yields geodesic trajectories on the manifold. There are some interesting subleties generalizing this to the Lorentzian manifolds of general relativity [8, 9], but in just the same way, inertial motion itself becomes a series of optimal choices.

References

[1] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, New Jersey (1957).

[2] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., MIT Press, Cambridge, Massachusetts (2018).

[3] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, 2nd ed., Springer, New York (2006). doi:10.1007/0-387-31071-1.

[4] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations*, Birkhäuser, Boston (1997). doi:10.1007/978-0-8176-4755-1.

[5] H. Goldstein, C. P. Poole, and J. L. Safko, *Classical Mechanics*, 3rd ed., Addison-Wesley, San Francisco (2002).

[6] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, 6th ed., Springer, Berlin (2003).

[7] J. M. Lee, *Introduction to Riemannian Manifolds*, 2nd ed., Graduate Texts in Mathematics, Vol. 176, Springer, Cham (2018). doi:10.1007/978-3-319-91755-9.

[8] B. O'Neill, *Semi-Riemannian Geometry: With Applications to Relativity*, Pure and Applied Mathematics, Vol. 103, Academic Press, New York (1983).

[9] S. M. Carroll, *Spacetime and Geometry: An Introduction to General Relativity*, 2nd ed., Cambridge University Press, Cambridge (2019).