Efficient computation of optimal actions
Supplementary notes

Emanuel Todorov

This supplement provides derivations of the results summarized in Table 1 in the main text, derivation of the relationship between the discrete and continuous formulations, and details on the MDP embedding method and the car-on-a-hill simulations.

1 Discrete problems

1.1 Generic formulation

We first recall the general form of a Markov decision process (MDP) and then introduce our new formulation which makes the problem more tractable. Consider a discrete set of states $\mathcal{X}$, a set $\mathcal{U}(x)$ of admissible actions at each state $x \in \mathcal{X}$, a transition probability function $p(x'|x,u)$, an instantaneous cost function $\ell(x,u)$, and (optionally) a final cost function $g(x)$ evaluated at the final state.

The objective of optimal control is to construct a control law $u = \pi_t^*(x)$ which minimizes the expected cumulative cost. Once a control law π is given the dynamics become autonomous, namely $x_{t+1} \sim p(\cdot|x_t, \pi_t(x_t))$. The expectations below are taken over state trajectories sampled from these dynamics. The expected cumulative cost for starting at state x and time t and acting according to π thereafter is denoted $v_t^\pi(x)$. This is called the cost-to-go function. It can be defined in multiple ways, as follows:

$$\begin{array}{ll} \text{first exit} \\ \text{total cost} & v^\pi(x) = \mathrm{E}\left[g\left(x_{t_f}\right) + \sum_{\tau=0}^{t_f-1} \ell\left(x_\tau, \pi\left(x_\tau\right)\right)\right] \\ \text{infinite horizon} \\ \text{average cost} & v^\pi(x) = \lim_{t_f\to\infty} \frac{1}{t_f} \mathrm{E}\left[\sum_{\tau=0}^{t_f-1} \ell\left(x_\tau, \pi\left(x_\tau\right)\right)\right] \\ \text{infinite horizon} \\ \text{discounted cost} & v^\pi(x) = \mathrm{E}\left[\sum_{\tau=0}^{\infty} \alpha^\tau \ell\left(x_\tau, \pi\left(x_\tau\right)\right)\right] \\ \text{finite horizon} \\ \text{total cost} & v_t^\pi(x) = \mathrm{E}\left[g\left(x_{t_f}\right) + \sum_{\tau=t}^{t_f-1} \ell\left(x_\tau, \pi_\tau\left(x_\tau\right)\right)\right] \end{array} \tag{1}$$

In all cases the expectation is taken over trajectories starting at x. Only the finite horizon formulation allows explicit dependence on the time index. All other formulations are time-invariant, which is why the trajectories are initialized at time 0. The final time t_f is predefined in the finite horizon formulation. In the first exit formulation t_f is determined online as the time when a terminal/goal state $x \in \mathcal{A}$ is first reached. We can also think of the latter problem as being infinite horizon, assuming the system can remain forever in a terminal state without incurring extra costs.

The optimal cost-to-go function $v(x)$ is the minimal expected cumulative cost that any control law can achieve starting at state x:

$$v(x) = \min_\pi v^\pi(x) \tag{2}$$

The optimal control law is not always unique but the optimal cost-to-go is. The above minimum is achieved by the same control law(s) for all states x. This follows from Bellman's optimality principle, and has to do with the fact that the optimal action at state x does not depend on how we reached state x. The optimality principle also gives rise to the Bellman equation – which is a self-consistency condition satisfied by the optimal cost-to-go function. Depending on the definition of cumulative cost the Bellman equation takes on different forms, as follows:

$$\begin{array}{ll} \text{first exit} \\ \text{total cost} & v(x) = \min_{u \in \mathcal{U}(x)} \left\{ \ell(x,u) + \mathrm{E}_{x' \sim p(\cdot|x,u)} \left[v(x') \right] \right\}, \quad v(x \in \mathcal{A}) = g(x) \\ \text{infinite horizon} \\ \text{average cost} & c + \widetilde{v}(x) = \min_{u \in \mathcal{U}(x)} \left\{ \ell(x,u) + \mathrm{E}_{x' \sim p(\cdot|x,u)} \left[\widetilde{v}(x') \right] \right\} \\ \text{infinite horizon} \\ \text{discounted cost} & v(x) = \min_{u \in \mathcal{U}(x)} \left\{ \ell(x,u) + \mathrm{E}_{x' \sim p(\cdot|x,u)} \left[\alpha v(x') \right] \right\} \\ \text{finite horizon} \\ \text{total cost} & v_t(x) = \min_{u \in \mathcal{U}(x)} \left\{ \ell(x,u) + \mathrm{E}_{x' \sim p(\cdot|x,u)} \left[v_{t+1}(x') \right] \right\}, \quad v_{t_f}(x) = g(x) \end{array} \tag{3}$$

In the average cost formulation $\widetilde{v}(x)$ has the meaning of a differential cost-to-go function, while c is the average cost which does not depend on the starting state. In the discounted cost formulation the constant $\alpha < 1$ is the exponential discount factor. In all formulations the Bellman equation involves minimization over the action set $\mathcal{U}(x)$. For generic MDPs such minimization requires exhaustive search. Our goal is to construct a class of MDPs for which this exhaustive search can be replaced with an analytical solution.

1.2 Restricted formulation where the Bellman equation is linear

In the traditional MDP formulation the controller chooses symbolic actions u which in turn specify transition probabilities $p(x'|x,u)$. In contrast, we allow the controller to choose transition probabilities $u(x'|x)$ directly, thus

$$p\left(x'|x,u\right) = u\left(x'|x\right) \tag{4}$$

The actions $u(\cdot|x)$ are real-valued vectors with non-negative elements which sum to 1. To prevent direct transitions to goal states, we define the passive or uncontrolled dynamics $p(x'|x)$ and require the actions to be compatible with it in the following sense:

$$\text{if } \; p\left(x'|x\right) = 0 \;\; \text{then we require } \;\; u\left(x'|x\right) = 0 \tag{5}$$

Since $\mathcal{U}(x)$ is now a continuous set, we can hope to perform the minimization in the Bellman equation analytically. Of course this also requires a proper choice of cost function $\ell(x,u)$ and in particular proper dependence of ℓ on u:

$$\ell(x,u) = q(x) + \mathrm{KL}\left(u(\cdot|x) \,||\, p(\cdot|x)\right) = q(x) + \mathrm{E}_{x' \sim u(\cdot|x)} \left[\log \frac{u(x'|x)}{p(x'|x)} \right] \tag{6}$$

$q(x)$ can be an arbitrary function. At terminal states $q(x) = g(x)$. Thus, as far as the state cost is concerned, we have not introduced any restrictions. The control cost however must equal the Kullback-Leibler (KL) divergence between the controlled and passive dynamics. This is a natural way to measure how "large" the action is, that is, how much it pushes the system away from its default behavior.

With these definitions we can proceed to solve for the optimal actions given the optimal costs-to-go. In all forms of the Bellman equation the minimization that needs to be performed is

$$\min_{u\in\mathcal{U}(x)}\left\{q\left(x\right)+\mathrm{E}_{x'\sim u(\cdot|x)}\left[\log\frac{u\left(x'|x\right)}{p\left(x'|x\right)}\right]+\mathrm{E}_{x'\sim u(\cdot|x)}\left[w\left(x'\right)\right]\right\} \tag{7}$$

where $w\left(x'\right)$ is one of $v\left(x'\right)$, $\widetilde{v}\left(x'\right)$, $\alpha v\left(x'\right)$, $v_{t+1}\left(x'\right)$. Below we give the derivation for $w=v$; the other cases are identical. The u-dependent expression being minimized in (7) is

$$\begin{aligned}\mathrm{E}_{x'\sim u(\cdot|x)}\left[\log\frac{u\left(x'|x\right)}{p\left(x'|x\right)}\right]+\mathrm{E}_{x'\sim u(\cdot|x)}\left[v\left(x'\right)\right] &= \mathrm{E}_{x'\sim u(\cdot|x)}\left[\log\frac{u\left(x'|x\right)}{p\left(x'|x\right)}+v\left(x'\right)\right] \\ &= \mathrm{E}_{x'\sim u(\cdot|x)}\left[\log\frac{u\left(x'|x\right)}{p\left(x'|x\right)}+\log\frac{1}{\exp\left(-v\left(x'\right)\right)}\right] \\ &= \mathrm{E}_{x'\sim u(\cdot|x)}\left[\log\frac{u\left(x'|x\right)}{p\left(x'|x\right)\exp\left(-v\left(x'\right)\right)}\right]\end{aligned} \tag{8}$$

The latter expression resembles KL divergence between u and $p\exp\left(-v\right)$, except that $p\exp\left(-v\right)$ is not normalized to sum to 1. In order to obtain a proper a KL divergence we introduce the normalization term

$$\mathcal{G}\left[z\right]\left(x\right)=\sum\nolimits_{x'}p\left(x'|x\right)z\left(x'\right)=\mathrm{E}_{x'\sim p(\cdot|x)}\left[z\left(x'\right)\right] \tag{9}$$

where the *desirability* function z is defined as

$$z\left(x\right)=\exp\left(-v\left(x\right)\right) \tag{10}$$

Now we multiply and divide the denominator on the last line of (8) by $\mathcal{G}\left[z\right]\left(x\right)$. The derivation proceeds as follows:

$$\begin{aligned}\mathrm{E}_{x'\sim u(\cdot|x)}\left[\log\frac{u\left(x'|x\right)}{p\left(x'|x\right)z\left(x'\right)}\right] &= \mathrm{E}_{x'\sim u(\cdot|x)}\left[\log\frac{u\left(x'|x\right)}{p\left(x'|x\right)z\left(x'\right)\mathcal{G}\left[z\right]\left(x\right)/\mathcal{G}\left[z\right]\left(x\right)}\right] \\ &= \mathrm{E}_{x'\sim u(\cdot|x)}\left[-\log\mathcal{G}\left[z\right]\left(x\right)+\log\frac{u\left(x'|x\right)}{p\left(x'|x\right)z\left(x'\right)/\mathcal{G}\left[z\right]\left(x\right)}\right] \\ &= -\log\mathcal{G}\left[z\right]\left(x\right)+\mathrm{KL}\left(u\left(\cdot|x\right)\left\|\frac{p\left(\cdot|x\right)z\left(\cdot\right)}{\mathcal{G}\left[z\right]\left(x\right)}\right.\right)\end{aligned} \tag{11}$$

Thus the minimization involved in the Bellman equation takes the form

$$\min_{u\in\mathcal{U}(x)}\left\{q\left(x\right)-\log\mathcal{G}\left[z\right]\left(x\right)+\mathrm{KL}\left(u\left(\cdot|x\right)\left\|\frac{p\left(\cdot|x\right)z\left(\cdot\right)}{\mathcal{G}\left[z\right]\left(x\right)}\right.\right)\right\} \tag{12}$$

The first two terms do not depend on u. KL divergence achieves its global minimum of 0 if and only if the two probability distributions are equal. Thus the optimal action is

$$u^{*}\left(x'|x\right)=\frac{p\left(x'|x\right)z\left(x'\right)}{\mathcal{G}\left[z\right]\left(x\right)} \tag{13}$$

We can now drop the min operator, exponentiate the Bellman equations and write them in terms

of z as follows:

$$\begin{array}{lll} \text{first exit} \\ \text{total cost} & z(x) = \exp(-q(x))\,\mathcal{G}[z](x) & \mathbf{z} = QP\mathbf{z} \\ \text{infinite horizon} \\ \text{average cost} & \exp(-c)\,\widetilde{z}(x) = \exp(-q(x))\,\mathcal{G}[\widetilde{z}](x) & \exp(-c)\,\widetilde{\mathbf{z}} = QP\widetilde{\mathbf{z}} \\ \text{infinite horizon} \\ \text{discounted cost} & z(x) = \exp(-q(x))\,\mathcal{G}[z^{\alpha}](x) & \mathbf{z} = QP\mathbf{z}^{\alpha} \\ \text{finite horizon} \\ \text{total cost} & z_t(x) = \exp(-q(x))\,\mathcal{G}[z_{t+1}](x) & \mathbf{z}_t = QP\mathbf{z}_{t+1} \end{array} \tag{14}$$

The third column gives the matrix form of these equations. The elements of the function $z(x)$ are assembled into the n-dimensional column vector $\mathbf{z}$, the passive dynamics $p(x'|x)$ are expressed as the n-by-n matrix P where the row index corresponds to x and the column index to x', and Q is the n-by-n diagonal matrix with elements $\exp(-q(x))$ along its main diagonal. In the average cost formulation it can be shown that $\lambda = \exp(-c)$ is the principal eigenvalue. In the discounted cost formulation we have used the fact that $\exp(-\alpha v) = \exp(-v)^{\alpha} = z^{\alpha}$.

The optimal control law in the average cost, discounted cost and finite horizon cost formulations is again in the form (13), but z is replaced with $\widetilde{z}$ or z_{t+1} or z^{α} respectively.

2 Continuous problems

2.1 Generic formulation

As in the discrete case, we first summarize the generic problem formulation and then introduce a new formulation which makes it more tractable. Consider a controlled Ito diffusion of the form

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, \mathbf{u})\,dt + F(\mathbf{x}, \mathbf{u})\,d\omega \tag{15}$$

where $\mathbf{x} \in \mathbb{R}^{n_{\mathbf{x}}}$ is a state vector, $\mathbf{u} \in \mathbb{R}^{n_{\mathbf{u}}}$ is a control vector, $\omega \in \mathbb{R}^{n_{\omega}}$ is standard multidimensional Brownian motion, $\mathbf{f}$ is the deterministic drift term and F is the diffusion coefficient. If a control law $\mathbf{u} = \pi(\mathbf{x})$ is given the above dynamics become autonomous. We will need the 2nd-order linear differential operator $\mathcal{L}_{(\mathbf{u})}$ defined as

$$\mathcal{L}_{(\mathbf{u})}[v] = \mathbf{f}^{\mathsf{T}} v_{\mathbf{x}} + \frac{1}{2}\,\mathrm{trace}\left(FF^{\mathsf{T}} v_{\mathbf{xx}}\right) \tag{16}$$

This is called the generator of the stochastic process (15). It is normally defined for autonomous dynamics, however the same notion applies for controlled dynamics as long as we make $\mathcal{L}$ dependent on $\mathbf{u}$. The generator equals the expected directional derivative along the state trajectories. In the absence of noise (i.e. when $F = 0$) we have $\mathcal{L}_{(\mathbf{u})}[v] = \mathbf{f}^{\mathsf{T}} v_{\mathbf{x}}$ which is the familiar directional derivative. The trace term is common in stochastic calculus and reflects the noise contribution.

Let $g(\mathbf{x})$ be a final cost evaluated at the final time t_f which is either fixed or defined as a first exist time as before, and let $\ell(\mathbf{x}, \mathbf{u})$ be a cost rate. The expected cumulative cost resulting from

control law π can be defined in the following ways:

$$\begin{array}{ll}
\text{first exit} & \\
\text{total cost} & v^{\pi}(\mathbf{x}) = \mathrm{E}\left[g(\mathbf{x}(t_f)) + \int_0^{t_f} \ell(\mathbf{x}(\tau), \pi(\mathbf{x}(\tau)))\, d\tau\right] \\
\text{infinite horizon} & \\
\text{average cost} & v^{\pi}(\mathbf{x}) = \lim_{t_f \to \infty} \frac{1}{t_f} \mathrm{E}\left[\int_0^{t_f} \ell(\mathbf{x}(\tau), \pi(\mathbf{x}(\tau)))\, d\tau\right] \\
\text{infinite horizon} & \\
\text{discounted cost} & v^{\pi}(\mathbf{x}) = \mathrm{E}\left[\int_0^{\infty} \exp(-\alpha\tau)\, \ell(\mathbf{x}(\tau), \pi(\mathbf{x}(\tau)))\, d\tau\right] \\
\text{finite horizon} & \\
\text{total cost} & v^{\pi}(\mathbf{x}, t) = \mathrm{E}\left[g(\mathbf{x}(t_f)) + \int_t^{t_f} \ell(\mathbf{x}(\tau), \pi(\mathbf{x}(\tau), \tau))\, d\tau\right]
\end{array} \tag{17}$$

As in the discrete case, the optimal cost-to-go function is

$$v(\mathbf{x}) = \inf_{\pi} v^{\pi}(\mathbf{x}) \tag{18}$$

This function satisfies the Hamilton-Jacobi-Bellman (HJB) equation. The latter has different forms depending on the definition of cumulative cost, as follows:

$$\begin{array}{ll}
\text{first exit} & \\
\text{total cost} & 0 = \min_{\mathbf{u}} \left\{\ell(\mathbf{x}, \mathbf{u}) + \mathcal{L}_{(\mathbf{u})}[v](\mathbf{x})\right\}, \quad v(\mathbf{x} \in \mathcal{A}) = g(\mathbf{x}) \\
\text{infinite horizon} & \\
\text{average cost} & c = \min_{\mathbf{u}} \left\{\ell(\mathbf{x}, \mathbf{u}) + \mathcal{L}_{(\mathbf{u})}[\widetilde{v}](\mathbf{x})\right\} \\
\text{infinite horizon} & \\
\text{discounted cost} & \alpha v(\mathbf{x}) = \min_{\mathbf{u}} \left\{\ell(\mathbf{x}, \mathbf{u}) + \mathcal{L}_{(\mathbf{u})}[v](\mathbf{x})\right\} \\
\text{finite horizon} & \\
\text{total cost} & -v_t(\mathbf{x}, t) = \min_{\mathbf{u}} \left\{\ell(\mathbf{x}, \mathbf{u}) + \mathcal{L}_{(\mathbf{u})}[v](\mathbf{x}, t)\right\}, \quad v(\mathbf{x}, t_f) = g(\mathbf{x})
\end{array} \tag{19}$$

Unlike the discrete case where the minimization over u had not been done analytically before, in the continuous case there is a well-known family of problems where analytical minimization is possible. These are problems with control-affine dynamics and control-quadratic costs:

$$\begin{aligned}
d\mathbf{x} &= (\mathbf{a}(\mathbf{x}) + B(\mathbf{x})\,\mathbf{u})\, dt + C(\mathbf{x})\, d\omega \\
\ell(\mathbf{x}, \mathbf{u}) &= q(\mathbf{x}) + \frac{1}{2}\mathbf{u}^{\mathsf{T}} R(\mathbf{x})\,\mathbf{u}
\end{aligned} \tag{20}$$

For such problems the quantity $\ell(\mathbf{x}, \mathbf{u}) + \mathcal{L}_{(\mathbf{u})}[v](\mathbf{x})$ becomes quadratic in $\mathbf{u}$, and so the optimal control law can be found analytically given the gradient of the optimal cost-to-go:

$$\mathbf{u}^*(\mathbf{x}) = -R(\mathbf{x})^{-1} B(\mathbf{x})^{\mathsf{T}} v_{\mathbf{x}}(\mathbf{x}) \tag{21}$$

Substituting this optimal control law, the right hand side of all four HJB equations takes the form

$$q - \frac{1}{2} v_{\mathbf{x}}^{\mathsf{T}} B R^{-1} B^{\mathsf{T}} v_{\mathbf{x}} + \mathbf{a}^{\mathsf{T}} v_{\mathbf{x}} + \frac{1}{2} \operatorname{tr}\left(C C^{\mathsf{T}} v_{\mathbf{xx}}\right) \tag{22}$$

where the dependence on $\mathbf{x}$ (and t when relevant) has been suppressed for clarity. The latter expression is nonlinear in the unknown function v. Our goal is to make it linear.

2.2 Restricted formulation where the HJB equation is linear

As in the discrete case, linearity is achieved by defining the desirability function

$$z(\mathbf{x}) = \exp(-v(\mathbf{x})) \tag{23}$$

and rewriting the HJB equations in terms of z. To do so we need to express v and its derivatives in terms of z and its derivatives:

$$v = -\log(z), \quad v_{\mathbf{x}} = -\frac{z_{\mathbf{x}}}{z}, \quad v_{\mathbf{xx}} = -\frac{z_{\mathbf{xx}}}{z} + \frac{z_{\mathbf{x}} z_{\mathbf{x}}^{\mathsf{T}}}{z^2} \tag{24}$$

The last equation is key, because it contains the term $z_{\mathbf{x}} z_{\mathbf{x}}^{\mathsf{T}}$ which will cancel the nonlinearity present in (22). Substituting (24) in (22), using the properties of the trace operator and rearranging yields

$$q - \frac{1}{z}\left(\mathbf{a}^{\mathsf{T}} z_{\mathbf{x}} + \frac{1}{2}\operatorname{tr}\left(CC^{\mathsf{T}} z_{\mathbf{xx}}\right) + \frac{1}{2z} z_{\mathbf{x}}^{\mathsf{T}} B R^{-1} B^{\mathsf{T}} z_{\mathbf{x}} - \frac{1}{2z} z_{\mathbf{x}}^{\mathsf{T}} CC^{\mathsf{T}} z_{\mathbf{x}}\right) \tag{25}$$

Now we see that the nonlinear terms cancel when $CC^{\mathsf{T}} = BR^{-1}B^{\mathsf{T}}$, which holds when

$$C(\mathbf{x}) = B(\mathbf{x})\sqrt{R(\mathbf{x})^{-1}} \tag{26}$$

The s.p.d. matrix square root of R^{-1} is uniquely defined because R is s.p.d. Note that in the main text we assumed $R = I/\sigma^2$ and so $C = B\sigma$. The present derivation is more general. However the noise and controls still act in the same subspace, and the noise amplitude and control cost are still inversely related.

Assuming condition (26) is satisfied, the right hand side of all four HJB equations takes the form

$$q - \frac{1}{z}\mathcal{L}_{(0)}[z] \tag{27}$$

where $\mathcal{L}_{(0)}$ is the generator of the passive dynamics (corresponding to $\mathbf{u} = 0$). We will omit the subscript (0) for clarity. For this class of problems the generator of the passive dynamics is

$$\mathcal{L}[z] = \mathbf{a}^{\mathsf{T}} z_{\mathbf{x}} + \frac{1}{2}\operatorname{tr}\left(CC^{\mathsf{T}} z_{\mathbf{xx}}\right) \tag{28}$$

Multiplying by $-z \neq 0$ we now obtain the transformed HJB equations

$$\begin{array}{ll}
\text{first exit} \\ \text{total cost} & 0 = \mathcal{L}[z] - qz \\
\text{infinite horizon} \\ \text{average cost} & -c\widetilde{z} = \mathcal{L}[\widetilde{z}] - q\widetilde{z} \\
\text{infinite horizon} \\ \text{discounted cost} & z\log(z^{\alpha}) = \mathcal{L}[z] - qz \\
\text{finite horizon} \\ \text{total cost} & -z_t = \mathcal{L}[z] - qz
\end{array} \tag{29}$$

As in the discrete case, these equations are linear in all but the discounted cost formulation.

3 Relation between the discrete and continuous formulations

Here we show how the continuous formulation can be obtained from the discrete formulation. This will be done by first making the state space of the MDP continuous (which merely replaces the sums with integrals), and then taking a continuous-time limit. Recall that the passive dynamics in the continuous formulation are

$$d\mathbf{x} = \mathbf{a}(\mathbf{x})\,dt + C(\mathbf{x})\,d\omega \tag{30}$$

Let $p_{(h)}(\cdot|\mathbf{x})$ denote the transition probability distribution of (30) over a time interval h. We can now define an MDP in our class with passive dynamics $p_{(h)}(\cdot|\mathbf{x})$ and state cost $hq(\mathbf{x})$. Let $z_{(h)}(\mathbf{x})$ denote the desirability function for this MDP, and suppose the following limit exists:

$$s(\mathbf{x}) = \lim_{h\to 0} z_{(h)}(\mathbf{x}) \tag{31}$$

The linear Bellman equation for the above MDP is

$$z_{(h)}(\mathbf{x}) = \exp(-hq(\mathbf{x}))\, E_{\mathbf{x}'\sim p_{(h)}(\cdot|\mathbf{x})}\left[z_{(h)}(\mathbf{x}')\right] \tag{32}$$

Our objective now is to take the continuous-time limit $h \to 0$ and recover the PDE

$$qs = \mathcal{L}[s] \tag{33}$$

A straightforward limit in (32) yields the trivial result $s = s$ because $p_{(0)}(\cdot|\mathbf{x})$ is the Dirac delta function centered at $\mathbf{x}$. However we can rearrange (32) as follows:

$$\frac{\exp(hq(\mathbf{x})) - 1}{h} z_{(h)}(\mathbf{x}) = \frac{E_{\mathbf{x}'\sim p_{(h)}(\cdot|\mathbf{x})}\left[z_{(h)}(\mathbf{x}') - z_{(h)}(\mathbf{x})\right]}{h} \tag{34}$$

The limit of the left hand side now yields qs. The limit of the right hand side closely resembles the generator of the passive dynamics (i.e. the expected directional derivative). If we had s instead of $z_{(h)}$ that limit would be exactly $\mathcal{L}[s]$ and we would recover (33). The same result is obtained by assuming that $z_{(h)}$ converges to s sufficiently rapidly and uniformly, so that

$$E_{\mathbf{x}'\sim p_{(h)}(\cdot|\mathbf{x})}\left[z_{(h)}(\mathbf{x}') - z_{(h)}(\mathbf{x})\right] = E_{\mathbf{x}'\sim p_{(h)}(\cdot|\mathbf{x})}\left[s(\mathbf{x}') - s(\mathbf{x})\right] + o(h^2) \tag{35}$$

Then the limit of (34) yields (33).

4 Embedding of traditional MDPs

In the main text we outlined a method for embedding traditional MDPs. The details are provided here. Denote the symbolic actions in the traditional MDP with a, the transition probabilities with $\widetilde{p}(x'|x,a)$ and the costs with $\widetilde{\ell}(x,a)$. We seek an MDP within our class such that for each (x,a) the action $u^a(\cdot|x) = \widetilde{p}(\cdot|x,a)$ has cost $\ell(x,u^a) = \widetilde{\ell}(x,a)$. In other words, for each symbolic action in the traditional MDP we want a corresponding continuous action with the same cost and transition probability distribution. The above requirement for proper embedding is equivalent to

$$q(x) + \sum\nolimits_{x'} \widetilde{p}(x'|x,a) \log \frac{\widetilde{p}(x'|x,a)}{p(x'|x)} = \widetilde{\ell}(x,a), \quad \forall x \in \mathcal{X}, a \in \widetilde{\mathcal{U}}(x) \tag{36}$$

This system of $\left|\widetilde{\mathcal{U}}(x)\right|$ equations has to be solved separately for each x, where $\widetilde{p}, \widetilde{\ell}$ are given and q, p are unknown. Let us fix x and define the vectors $\mathbf{m}, \mathbf{b}$ and the matrix D with elements

$$\begin{aligned} m_{x'} &= \log p\left(x'|x\right) \\ b_a &= \widetilde{\ell}\left(x,a\right) - \sum\nolimits_{x'} \widetilde{p}\left(x'|x,a\right) \log \widetilde{p}\left(x'|x,a\right) \\ D_{ax'} &= \widetilde{p}\left(x'|x,a\right) \end{aligned} \tag{37}$$

Then the above system of equations becomes linear:

$$q\mathbf{1} - D\mathbf{m} = \mathbf{b} \tag{38}$$

$D, \mathbf{b}$ are given, $q, \mathbf{m}$ are unknown, $\mathbf{1}$ is a column vector of 1's. In addition we require that p be a normalized probability distribution, which holds when

$$\sum\nolimits_{x'} \exp\left(m_{x'}\right) = 1 \tag{39}$$

The latter equation is nonlinear but nevertheless the problem can be made linear. Since D is a stochastic matrix, we have $D\mathbf{1} = \mathbf{1}$. Then equation (38) is equivalent to

$$D\left(q\mathbf{1} - \mathbf{m}\right) = \mathbf{b} \tag{40}$$

which can be solved for $\mathbf{c} = q\mathbf{1} - \mathbf{m}$ using a linear solver. For any q the vector $\mathbf{m} = q\mathbf{1} - \mathbf{c}$ is a solution to (38). Thus we can choose q so as to make $\mathbf{m}$ satisfy (39), namely

$$q = -\log \sum\nolimits_{x'} \exp\left(-c_{x'}\right) \tag{41}$$

If D is row-rank-defficient the solution $\mathbf{c}$ is not unique, and we should be able to exploit the freedom in choosing $\mathbf{c}$ to improve the approximation of the traditional MDP. If D is column-rank-defficient then an exact embedding cannot be constructed. However this is unlikely to occur in practice because it essentially means that the number of symbolic actions is greater than the number of possible next states.

5 Car-on-a-hill simulation

The continuous control problem illustrated in **Fig. 5** in the main text is as follows. x_1 and x_2 denote the horizontal position and tangential velocity of the car. The state vector is $\mathbf{x} = [x_1, x_2]^{\mathsf{T}}$. The hill elevation over position x_1 is

$$s\left(x_1\right) = 2 - 2\exp\left(-x_1^2/2\right) \tag{42}$$

The slope is $s'\left(x_1\right) = 2x_1 \exp\left(-x_1^2/2\right)$ and the angle relative to the horizontal plane is $\operatorname{atan}\left(s'\left(x_1\right)\right)$. The tangential acceleration reflects the effects of gravity, damping, control signal u and noise. The dynamics are

$$\begin{aligned} dx_1 &= x_2 \cos\left(\operatorname{atan}\left(s'\left(x_1\right)\right)\right) dt \\ dx_2 &= -g \operatorname{sgn}\left(x_1\right) \sin\left(\operatorname{atan}\left(s'\left(x_1\right)\right)\right) dt - \beta x_2 dt + u dt + d\omega \end{aligned} \tag{43}$$

$g = 9.8$ is the gravitational constant and $\beta = 0.5$ is the damping coefficient. The cost rate is $\ell\left(\mathbf{x}, u\right) = q + \frac{1}{2}u^2$ with $q = 5$. The goal states are all states such that $|x_1 - 2.5| < 0.05$ and

$|x_2| < 0.2$. This cost model encodes the task of parking at horizontal position 2.5 in minimal time and with minimal control energy. The constant q determines the relative importance of time and energy. Some error tolerance is needed because the dynamics are stochastic. This continuous problem is in the form given in the main text, so it can be approximated with an MDP in our class. It can also be approximated with a traditional MDP. Both approximations use the same state space discretization: a 101-by-101 grid spanning $x_1 \in [-3, +3]$, $x_2 \in [-9, +9]$. The traditional MDP also uses discretization of the control space: a 101 point grid spanning $u \in [-30, +30]$. The passive dynamics p are constructed by discretizing the time axis (with time step $h = 0.05$) and defining probabilistic transitions among discrete states so that the mean and variance of the continuous-state dynamics are preserved. The noise distribution is discretized at 9 points spanning ± 3 standard deviations in the x_2 direction, that is, $\left[-3\sqrt{h}, +3\sqrt{h}\right]$. The controlled dynamics are obtained from p by shifting in the x_2 direction. For each value of u the set of possible next states is a 2-by-9 sub-grid, except at the edges of the grid where non-existent states are removed and the distribution is normalized to sum to 1.

(a) – Schematic illustration of the problem.

(b) – Comparison of Z-iteration (blue), policy iteration (red) and value iteration (black). The vertical axis shows the empirical performance of the control policies. It was found by initializing 10 trajectories in each discrete state and sampling until the goal was reached or until the trajectory length exceeded 500 steps. The sampling was done in discrete time and continuous space, using the nearest discrete state to determine the control signal. The horizontal axis (note the log scale) shows the number of updates for each method. One update involves a computation of the form $A\mathbf{v} + \mathbf{b}$ for policy and value iteration, and $A\mathbf{z}$ for Z-iteration. The computation of minima in policy and value iteration is not counted, thus our method has an even bigger advantage than what is shown in the figure. The evaluation step in policy iteration was done with an iterative linear solver which was terminated at 20 iterations (or when convergence was reached) because complete evaluation slows down policy iteration. The value of 20 was manually optimized. Recall that the evaluation step in policy iteration, as well as the linear problem that needs to be solved in our formulation, can also be handled with a direct solver. Then our method becomes equivalent to a single evaluation step in policy iteration. Policy iteration using a direct solver converged in 10 iterations, thus our method was more than 10 times faster.

(c) – The optimal cost-to-go function computed by Z-iteration (blue is small; red is large). Also shown are two stochastic trajectories generated by the optimal controller (black). The magenta curve is the most likely trajectory of the optimally-controlled stochastic system. It is computed by solving the corresponding deterministic problem via dynamic programming applied to the discretization. Note that we could also solve a continuous deterministic problem (given in the main text) and recover the same trajectory.

(d) – The optimal cost-to-go $v(x)$ inferred from a dataset generated by the optimal controller. The dataset contained 20 state transitions per state: the system was initialized 20 times in each discrete state and the next discrete state was sampled from the optimal controller. The pixels shown in brown correspond to states where none of the transitions landed. The cost-to-go at those states cannot be inferred. The inference procedure is based on the diagonal Gauss-Newton method applied to the function $L(\mathbf{v})$ in the main text.