Linearly-Solvable Stochastic Optimal Control Problems

Emo Todorov

Applied Mathematics and Computer Science & Engineering

University of Washington

Winter 2014

Problem formulation

In traditional MDPs the controller chooses actions u which in turn specify the transition probabilities $p\left(x'|x,u\right)$. We can obtain a linearly-solvable MDP (LMDP) by allowing the controller to specify these probabilities directly:

$$
\begin{aligned}
&x' \sim u\left(\cdot|x\right) && \text{controlled dynamics} \\
&x' \sim p\left(\cdot|x\right) && \text{passive dynamics} \\
&p\left(x'|x\right) = 0 \Rightarrow u\left(x'|x\right) = 0 && \text{feasible control set } \mathcal{U}\left(x\right)
\end{aligned}
$$

Problem formulation

In traditional MDPs the controller chooses actions u which in turn specify the transition probabilities $p(x'|x,u)$. We can obtain a linearly-solvable MDP (LMDP) by allowing the controller to specify these probabilities directly:

$$\begin{aligned} x' &\sim u(\cdot|x) && \text{controlled dynamics} \\ x' &\sim p(\cdot|x) && \text{passive dynamics} \\ p(x'|x) = 0 &\Rightarrow u(x'|x) = 0 && \text{feasible control set } \mathcal{U}(x) \end{aligned}$$

The immediate cost is in the form

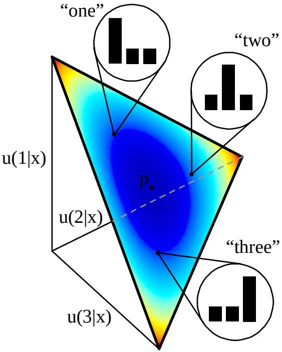
$$\ell(x, u(\cdot|x)) = q(x) + KL(u(\cdot|x)\,||p(\cdot|x))$$

$$KL(u(\cdot|x)\,||p(\cdot|x)) = \sum_{x'} u(x'|x) \log \frac{u(x'|x)}{p(x'|x)} = E_{x' \sim u(\cdot|x)}\left[\log \frac{u(x'|x)}{p(x'|x)}\right]$$

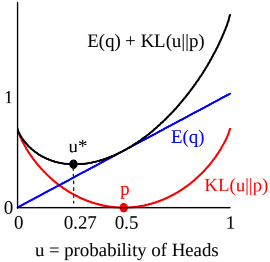
Thus the controller can impose any dynamics it wishes, however it pays a price (KL divergence control cost) for pushing the system away from its passive dynamics.

Understanding the KL divergence cost

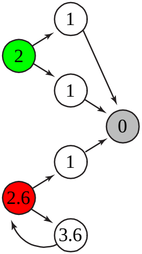
KL cost over the probability simplex



how to bias a coin



benefits of error tolerance



Simplifying the Bellman equation (first exit)

$$
\begin{aligned}
v(x) &= \min_u \left\{ \ell(x,u) + E_{x' \sim p(\cdot|x,u)} \left[v(x') \right] \right\} \\
&= \min_{u(\cdot|x)} \left\{ q(x) + E_{x' \sim u(\cdot|x)} \left[\log \frac{u(x'|x)}{p(x'|x)} + \log \frac{1}{\exp(-v(x'))} \right] \right\} \\
&= \min_{u(\cdot|x)} \left\{ q(x) + E_{x' \sim u(\cdot|x)} \left[\log \frac{u(x')}{p(x'|x)\exp(-v(x'))} \right] \right\}
\end{aligned}
$$

The last term is an unnormalized KL divergence...

Simplifying the Bellman equation (first exit)

$$
\begin{aligned}
v(x) &= \min_u \left\{ \ell(x,u) + E_{x' \sim p(\cdot|x,u)} \left[v(x') \right] \right\} \\
&= \min_{u(\cdot|x)} \left\{ q(x) + E_{x' \sim u(\cdot|x)} \left[\log \frac{u(x'|x)}{p(x'|x)} + \log \frac{1}{\exp(-v(x'))} \right] \right\} \\
&= \min_{u(\cdot|x)} \left\{ q(x) + E_{x' \sim u(\cdot|x)} \left[\log \frac{u(x')}{p(x'|x) \exp(-v(x'))} \right] \right\}
\end{aligned}
$$

The last term is an unnormalized KL divergence...

Definitions

desirability function $\quad z(x) \triangleq \exp(-v(x))$

next-state expectation $\quad \mathcal{P}[z](x) \triangleq \sum_{x'} p(x'|x) z(x')$

$$
v(x) = \min_{u(\cdot|x)} \left\{ q(x) - \log \mathcal{P}[z](x) + KL\left(u(\cdot|x) \,\middle\|\, \frac{p(\cdot|x) z(\cdot)}{\mathcal{P}[z](x)} \right) \right\}
$$

Linear Bellman equation and optimal control law

$KL\left(p_1(\cdot)\,||p_2(\cdot)\right)$ achieves its global minimum of 0 iff $p_1 = p_2$, thus

Theorem (optimal control law)

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

The Bellman equation becomes

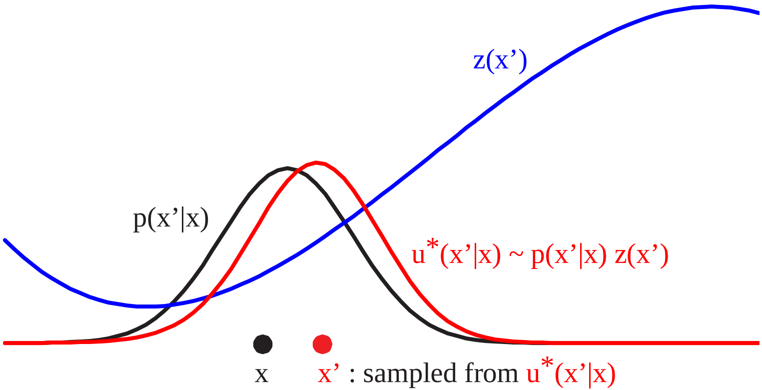
$$\begin{aligned} v(x) &= q(x) - \log \mathcal{P}[z](x) \\ z(x) &= \exp(-q(x))\,\mathcal{P}[z](x) \end{aligned}$$

which can be written more explicitly as

Theorem (linear Bellman equation)

$$z(x) = \begin{cases} \exp(-q(x)) \sum_{x'} p\left(x'|x\right) z\left(x'\right) & : x \textit{ non-terminal} \\ \exp\left(-q_{\mathcal{T}}(x)\right) & : x \textit{ terminal} \end{cases}$$

Illustration



Summary of results

Let $Q = \text{diag}\left(\exp\left(-\mathbf{q}\right)\right)$ and $P_{xy} = p\left(y|x\right)$. Then we have

first exit	$z = \exp\left(-q\right)\mathcal{P}\left[z\right]$	$\mathbf{z} = QP\mathbf{z}$
finite horizon	$z_k = \exp\left(-q_k\right)\mathcal{P}_k\left[z_{k+1}\right]$	$\mathbf{z}_k = Q_k P_k \mathbf{z}_{k+1}$
average cost	$z = \exp\left(c-q\right)\mathcal{P}\left[z\right]$	$\lambda\mathbf{z} = QP\mathbf{z}$
discounted cost	$z = \exp\left(-q\right)\mathcal{P}\left[z^{\alpha}\right]$	$\mathbf{z} = QP\mathbf{z}^{\alpha}$

Summary of results

Let $Q = \text{diag}\left(\exp\left(-\mathbf{q}\right)\right)$ and $P_{xy} = p\left(y|x\right)$. Then we have

first exit	$z = \exp\left(-q\right)\mathcal{P}\left[z\right]$	$\mathbf{z} = QP\mathbf{z}$
finite horizon	$z_k = \exp\left(-q_k\right)\mathcal{P}_k\left[z_{k+1}\right]$	$\mathbf{z}_k = Q_k P_k \mathbf{z}_{k+1}$
average cost	$z = \exp\left(c-q\right)\mathcal{P}\left[z\right]$	$\lambda\mathbf{z} = QP\mathbf{z}$
discounted cost	$z = \exp\left(-q\right)\mathcal{P}\left[z^{\alpha}\right]$	$\mathbf{z} = QP\mathbf{z}^{\alpha}$

In the first exit problem we can also write

$$
\begin{aligned}
\mathbf{z}_{\mathcal{N}} &= Q_{\mathcal{N}\mathcal{N}}P_{\mathcal{N}\mathcal{N}}\mathbf{z}_{\mathcal{N}} + \mathbf{b} = \left(I - Q_{\mathcal{N}\mathcal{N}}P_{\mathcal{N}\mathcal{N}}\right)^{-1}\mathbf{b} \\
\mathbf{b} &\triangleq Q_{\mathcal{N}\mathcal{N}}P_{\mathcal{N}\mathcal{T}}\exp\left(-\mathbf{q}_{\mathcal{T}}\right)
\end{aligned}
$$

where $\mathcal{N},\mathcal{T}$ are the sets of non-terminal and terminal states respectively.

In the average cost problem $\lambda = -\log\left(c\right)$ is the principal eigenvalue.

Stationary distribution under the optimal control law

Let $\mu(x)$ denote the stationary distribution under the optimal control law $u^*(\cdot|x)$ in the average cost problem. Then

$$\mu(x') = \sum_x u^*(x'|x)\,\mu(x)$$

Recall that

$$u^*(x'|x) = \frac{p(x'|x)\,z(x')}{\mathcal{P}[z](x)} = \frac{p(x'|x)\,z(x')}{\lambda \exp(q(x))\,z(x)}$$

Defining $r(x) \triangleq \mu(x)/z(x)$, we have

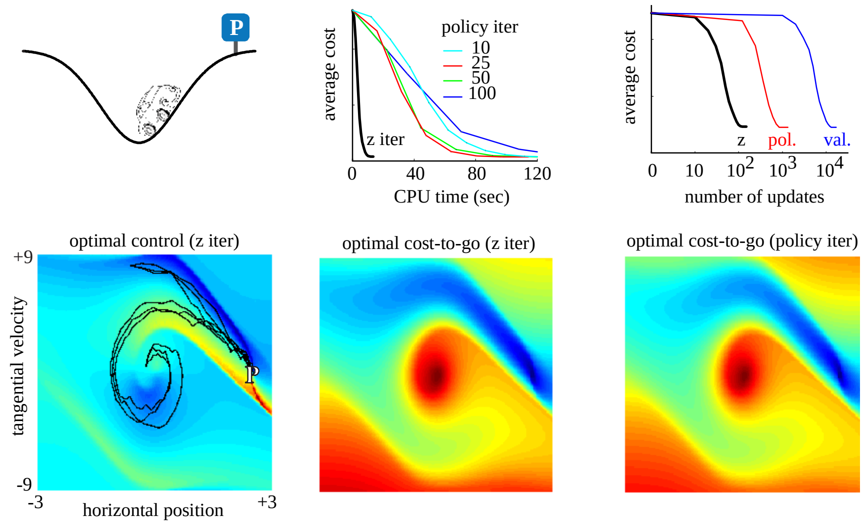
$$\begin{aligned} \mu(x') &= \sum_x \frac{p(x'|x)\,z(x')}{\lambda \exp(q(x))\,z(x)}\mu(x) \\ \lambda r(x') &= \sum_x \exp(-q(x))\,p(x'|x)\,r(x) \end{aligned}$$

In vector notation this becomes

$$\lambda \mathbf{r} = (QP)^{\mathsf{T}}\,\mathbf{r}$$

Thus $\mathbf{z}$ and $\mathbf{r}$ are the right and left principal eigenvectors of QP, and $\boldsymbol{\mu} = \mathbf{z}.*\mathbf{r}$

Comparison to policy and value iteration



Application to deterministic shortest paths

Given a graph and a set $\mathcal{T}$ of goal states, define the first-exit LMDP

$p(x'\|x)$	random walk on the graph
$q(x) = \rho > 0$	constant cost at non-terminal states
$q_{\mathcal{T}}(x) = 0$	zero cost at terminal states

For large ρ the optimal cost-to-go $v^{(\rho)}$ is dominated by the state costs, since the KL divergence control costs are bounded. Thus we have

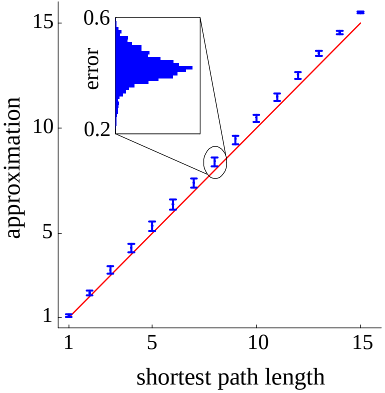
Theorem

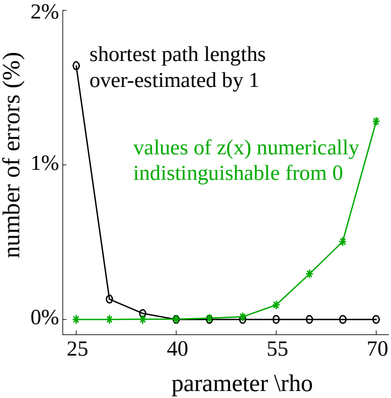
The length of the shortest path from state x to a goal state is

$$\lim_{\rho \to \infty} \frac{v^{(\rho)}(x)}{\rho}$$

Internet example

Performance on the graph of Internet routers as of 2003 (data from caida.org)
There are 190914 nodes and 609066 undirected edges in the graph.





Embedding of traditional MDPs

Given a traditional MDP with controls $\widetilde{u} \in \widetilde{\mathcal{U}}(x)$, transition probabilities $\widetilde{p}(x'|x,\widetilde{u})$ and costs $\widetilde{\ell}(x,\widetilde{u})$, we can construct and LMDP such that the controls corresponding to the MDPs transition probabilities have the same costs as in the MDP. This is done by constructing p and q such that for $\forall x, \widetilde{u} \in \widetilde{\mathcal{U}}(x)$

$$
\begin{aligned}
q(x) + KL(\widetilde{p}(\cdot|x,\widetilde{u})\,||p(\cdot|x)) &= \widetilde{\ell}(x,\widetilde{u}) \\
q(x) - \textstyle\sum_{x'} \widetilde{p}\left(x'|x,\widetilde{u}\right) \log p\left(x'|x\right) &= \widetilde{\ell}(x,\widetilde{u}) + \widetilde{h}(x,\widetilde{u})
\end{aligned}
$$

where $\widetilde{h}$ is the entropy of $\widetilde{p}(\cdot|x,\widetilde{u})$.

Embedding of traditional MDPs

Given a traditional MDP with controls $\widetilde{u} \in \widetilde{\mathcal{U}}(x)$, transition probabilities $\widetilde{p}(x'|x,\widetilde{u})$ and costs $\widetilde{\ell}(x,\widetilde{u})$, we can construct and LMDP such that the controls corresponding to the MDPs transition probabilities have the same costs as in the MDP. This is done by constructing p and q such that for $\forall x, \widetilde{u} \in \widetilde{\mathcal{U}}(x)$

$$\begin{aligned} q(x) + KL(\widetilde{p}(\cdot|x,\widetilde{u}) \,||\, p(\cdot|x)) &= \widetilde{\ell}(x,\widetilde{u}) \\ q(x) - \textstyle\sum_{x'} \widetilde{p}\left(x'|x,\widetilde{u}\right) \log p\left(x'|x\right) &= \widetilde{\ell}(x,\widetilde{u}) + \widetilde{h}(x,\widetilde{u}) \end{aligned}$$

where $\widetilde{h}$ is the entropy of $\widetilde{p}(\cdot|x,\widetilde{u})$. The construction is done separately for every x. Suppressing x, vectorizing over $\widetilde{u}$ and defining $s = -\log p$,

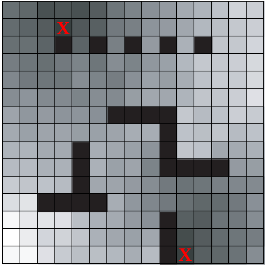
$$\begin{aligned} q\mathbf{1} + \widetilde{P}\mathbf{s} &= \widetilde{\mathbf{b}} \\ \exp(-\mathbf{s})^{\mathsf{T}}\mathbf{1} &= 1 \end{aligned}$$

$\widetilde{P}$ and $\widetilde{\mathbf{b}} = \widetilde{\ell} + \widetilde{\mathbf{h}}$ are known, q and $\mathbf{s}$ are unknown. Assuming $\widetilde{P}$ is full rank,

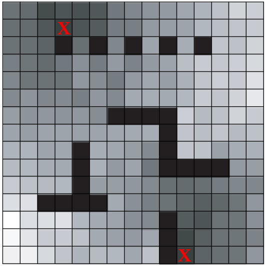
$$\mathbf{y} = \widetilde{P}^{-1}\widetilde{\mathbf{b}}, \quad \mathbf{s} = \mathbf{y} - q\mathbf{1}, \quad q = -\log\left(\exp(-\mathbf{y})^{\mathsf{T}}\mathbf{1}\right)$$

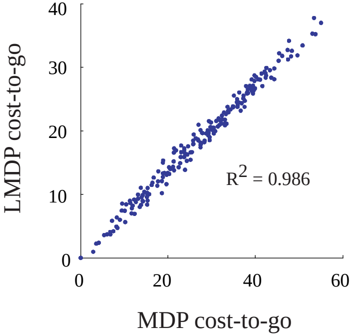
Grid world example

MDP
cost-to-go

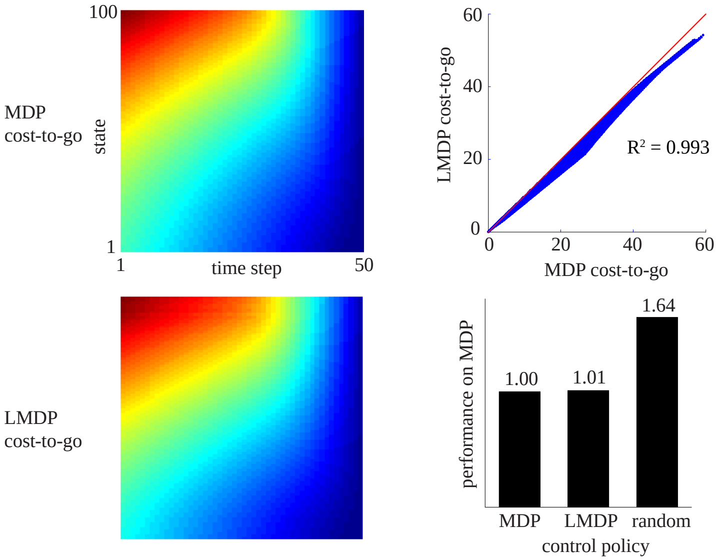


LMDP
cost-to-go





Machine repair example



Continuous-time limit

Consider a continuous-state discrete-time LMDP where $p^{(h)}(\mathbf{x}'|\mathbf{x})$ is the h-step transition probability of some continuous-time stochastic process, and $z^{(h)}(\mathbf{x})$ is the LMDP solution. The linear Bellman equation (first exit) is

$$z^{(h)}(\mathbf{x}) = \exp(-hq(\mathbf{x}))\, E_{\mathbf{x}' \sim p^{(h)}(\cdot|\mathbf{x})}\left[z^{(h)}(\mathbf{x}')\right]$$

Let $z = \lim_{h \downarrow 0} z^{(h)}$. The limit yields $z(\mathbf{x}) = z(\mathbf{x})$,

Continuous-time limit

Consider a continuous-state discrete-time LMDP where $p^{(h)}(\mathbf{x}'|\mathbf{x})$ is the h-step transition probability of some continuous-time stochastic process, and $z^{(h)}(\mathbf{x})$ is the LMDP solution. The linear Bellman equation (first exit) is

$$z^{(h)}(\mathbf{x}) = \exp(-hq(\mathbf{x}))\, E_{\mathbf{x}'\sim p^{(h)}(\cdot|\mathbf{x})}\left[z^{(h)}(\mathbf{x}')\right]$$

Let $z = \lim_{h\downarrow 0} z^{(h)}$. The limit yields $z(\mathbf{x}) = z(\mathbf{x})$, but we can rearrange as

$$\lim_{h\downarrow 0} \frac{\exp(hq(\mathbf{x})) - 1}{h} z^{(h)}(\mathbf{x}) = \lim_{h\downarrow 0} \frac{E_{\mathbf{x}'\sim p^{(h)}(\cdot|\mathbf{x})}\left[z^{(h)}(\mathbf{x}')\right] - z^{(h)}(\mathbf{x})}{h}$$

Recalling the definition of the generator $\mathcal{L}$, we now have

$$q(\mathbf{x})\, z(\mathbf{x}) = \mathcal{L}[z](\mathbf{x})$$

If the underlying process is an Ito diffusion, the generator is

$$\mathcal{L}[z](\mathbf{x}) = \mathbf{a}(\mathbf{x})^{\mathsf{T}} z_{\mathbf{x}}(\mathbf{x}) + \frac{1}{2}\operatorname{trace}(\Sigma(\mathbf{x})\, z_{\mathbf{xx}}(\mathbf{x}))$$

Linearly-solvable controlled diffusions

Above z was defined as the continuous-time limit to LMDP solutions $z^{(h)}$.
But is z the solution to a continuous-time problem, and if so, what problem?

Linearly-solvable controlled diffusions

Above z was defined as the continuous-time limit to LMDP solutions $z^{(h)}$. But is z the solution to a continuous-time problem, and if so, what problem?

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

Recall that for such problems we have $\mathbf{u}^* = -R^{-1}B^{\mathsf{T}}v_{\mathbf{x}}$ and

$$0 = q + \mathbf{a}^{\mathsf{T}}v_{\mathbf{x}} + \frac{1}{2}\operatorname{tr}\left(CC^{\mathsf{T}}v_{\mathbf{xx}}\right) - \frac{1}{2}v_{\mathbf{x}}^{\mathsf{T}}BR^{-1}B^{\mathsf{T}}v_{\mathbf{x}}$$

Define $z\left(\mathbf{x}\right) = \exp\left(-v\left(\mathbf{x}\right)\right)$ and write the PDE in terms of z:

$$v_{\mathbf{x}} = -\frac{z_{\mathbf{x}}}{z}, \quad v_{\mathbf{xx}} = -\frac{z_{\mathbf{xx}}}{z} + \frac{z_{\mathbf{x}}z_{\mathbf{x}}^{\mathsf{T}}}{z^2}$$

$$0 = q - \frac{1}{z}\left(\mathbf{a}^{\mathsf{T}}z_{\mathbf{x}} + \frac{1}{2}\operatorname{tr}\left(CC^{\mathsf{T}}z_{\mathbf{xx}}\right) + \frac{1}{2z}z_{\mathbf{x}}^{\mathsf{T}}BR^{-1}B^{\mathsf{T}}z_{\mathbf{x}} - \frac{1}{2z}z_{\mathbf{x}}^{\mathsf{T}}CC^{\mathsf{T}}z_{\mathbf{x}}\right)$$

Now if $CC^{\mathsf{T}} = BR^{-1}B^{\mathsf{T}}$, we obtain the linear HJB equation $qz = \mathcal{L}\left[z\right]$.

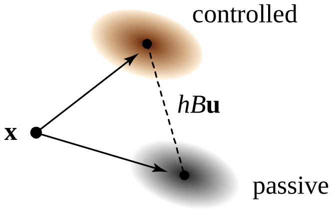
Quadratic control cost and KL divergence

The KL divergence between two Gaussians with means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and common full-rank covariance Σ is $\frac{1}{2}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^\mathsf{T} \Sigma^{-1} \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)$.

Using Euler discretization of the controlled diffusion, the passive and controlled dynamics have means $\mathbf{x} + h\mathbf{a}$, $\mathbf{x} + h\mathbf{a} + hB\mathbf{u}$ and covariance hCC^T. Thus the KL divergence control cost is

$$\frac{1}{2} h\mathbf{u}^\mathsf{T} B^\mathsf{T} \left(hCC^\mathsf{T}\right)^{-1} hB\mathbf{u} = \frac{h}{2}\mathbf{u}^\mathsf{T} B^\mathsf{T} \left(BR^{-1}B^\mathsf{T}\right)^{-1} B\mathbf{u} = \frac{h}{2}\mathbf{u}^\mathsf{T} R\mathbf{u}$$

This is the quadratic control cost accumulated over time h.



Here we used $CC^\mathsf{T} = BR^{-1}B^\mathsf{T}$ and assumed that B is full rank. If B is rank-defficient, the same result holds but the Gaussians are defined over the subspace spanned by the columns of B.

Summary of results

	discrete time :	continuous time :
first exit	$\exp(q)\, z = \mathcal{P}[z]$	$qz = \mathcal{L}[z]$
finite horizon	$\exp(q_k)\, z_k = \mathcal{P}_k[z_{k+1}]$	$qz - z_t = \mathcal{L}[z]$
average cost	$\exp(q - c)\, z = \mathcal{P}[z]$	$(q - c)\, z = \mathcal{L}[z]$
discounted cost	$\exp(q)\, z = \mathcal{P}[z^{\alpha}]$	$z \log(z^{\alpha}) = \mathcal{L}[z]$

Summary of results

	discrete time :	continuous time :
first exit	$\exp(q)\, z = \mathcal{P}[z]$	$qz = \mathcal{L}[z]$
finite horizon	$\exp(q_k)\, z_k = \mathcal{P}_k[z_{k+1}]$	$qz - z_t = \mathcal{L}[z]$
average cost	$\exp(q - c)\, z = \mathcal{P}[z]$	$(q - c)\, z = \mathcal{L}[z]$
discounted cost	$\exp(q)\, z = \mathcal{P}[z^\alpha]$	$z \log(z^\alpha) = \mathcal{L}[z]$

The relation between $\mathcal{P}[z]$ and $\mathcal{L}[z]$ is

$$\begin{aligned}
\mathcal{P}[z](\mathbf{x}) &= E_{\mathbf{x}' \sim p(\cdot|\mathbf{x})}\left[z(\mathbf{x}')\right] \\
\mathcal{L}[z](\mathbf{x}) &= \lim_{h\downarrow 0} \frac{E_{\mathbf{x}' \sim p^{(h)}(\cdot|\mathbf{x})}[z(\mathbf{x}')] - z(\mathbf{x})}{h} = \lim_{h\downarrow 0} \frac{\mathcal{P}^{(h)}[z](\mathbf{x}) - z(\mathbf{x})}{h} \\
\mathcal{P}^{(h)}[z](\mathbf{x}) &= z(\mathbf{x}) + h\mathcal{L}[z](\mathbf{x}) + o\left(h^2\right)
\end{aligned}$$

Path-integral representation

We can unfold the linear Bellman equation (first exit) as

$$
\begin{aligned}
z(x) &= \exp(-q(x))\, E_{x' \sim p(\cdot|x)}\left[z(x')\right] \\
&= \exp(-q(x))\, E_{x' \sim p(\cdot|x)}\left[\exp(-q(x'))\, E_{x'' \sim p(\cdot|x')}\left[z(x'')\right]\right] \\
&= \cdots \\
&= E^{x_0 = x}_{x_{k+1} \sim p(\cdot|x_k)}\left[\exp\left(-q_{\mathcal{T}}\left(x_{t_{\text{first}}}\right) - \sum_{k=0}^{t_{\text{first}}-1} q(x_k)\right)\right]
\end{aligned}
$$

This is a path-integral representation of z. Since $KL(p||p) = 0$, we have

$$
\exp\left(E_{\text{optimal}}\left[-\text{total cost}\right]\right) = z(x) = E_{\text{passive}}\left[\exp(-\text{total cost})\right]
$$

Path-integral representation

We can unfold the linear Bellman equation (first exit) as

$$\begin{aligned} z(x) &= \exp(-q(x))\, E_{x'\sim p(\cdot|x)}\left[z(x')\right] \\ &= \exp(-q(x))\, E_{x'\sim p(\cdot|x)}\left[\exp(-q(x'))\, E_{x''\sim p(\cdot|x')}\left[z(x'')\right]\right] \\ &= \cdots \\ &= E^{x_0=x}_{x_{k+1}\sim p(\cdot|x_k)}\left[\exp\left(-q_{\mathcal{T}}\left(x_{t_{\text{first}}}\right) - \sum_{k=0}^{t_{\text{first}}-1} q(x_k)\right)\right] \end{aligned}$$

This is a path-integral representation of z. Since $KL(p||p) = 0$, we have

$$\exp\left(E_{\text{optimal}}\left[-\text{total cost}\right]\right) = z(x) = E_{\text{passive}}\left[\exp(-\text{total cost})\right]$$

In continuous problems, the Feynman-Kac theorem states that the unique positive solution z to the parabolic PDE $qz = \mathbf{a}^{\mathsf{T}} z_{\mathbf{x}} + \frac{1}{2}\,\text{tr}\left(CC^{\mathsf{T}} z_{\mathbf{xx}}\right)$ has the same path-integral representation:

$$z(\mathbf{x}) = E^{\mathbf{x}(0)=\mathbf{x}}_{d\mathbf{x}=\mathbf{a}(\mathbf{x})dt+C(\mathbf{x})d\omega}\left[\exp\left(-q_{\mathcal{T}}\left(\mathbf{x}\left(t_{\text{first}}\right)\right) - \int_0^{t_{\text{first}}} q(\mathbf{x}(t))\, dt\right)\right]$$

Model-free learning

The solution to the linear Bellman equation

$$z(x) = \exp(-q(x))\, E_{x' \sim p(\cdot|x)} \left[z(x') \right]$$

can be approximated in a model-free way given samples $(x_n, x_n', q_n = q(x_n))$ obtained from the **passive dynamics** $x_n' \sim p(\cdot|x_n)$.

Model-free learning

The solution to the linear Bellman equation

$$z(x) = \exp(-q(x))\, E_{x' \sim p(\cdot|x)}\left[z(x')\right]$$

can be approximated in a model-free way given samples $(x_n, x_n', q_n = q(x_n))$ obtained from the **passive dynamics** $x_n' \sim p(\cdot|x_n)$.

One possibility is a Monte Carlo method based on the path integral representation, although covergence can be slow:

$$\widehat{z}(x) = \frac{1}{\substack{\#\text{ trajectories} \\ \text{starting at } x}} \sum \exp(-\text{ trajectory cost})$$

Model-free learning

The solution to the linear Bellman equation

$$z(x) = \exp(-q(x))\, E_{x' \sim p(\cdot|x)}\left[z(x')\right]$$

can be approximated in a model-free way given samples $(x_n, x'_n, q_n = q(x_n))$ obtained from the **passive dynamics** $x'_n \sim p(\cdot|x_n)$.

One possibility is a Monte Carlo method based on the path integral representation, although covergence can be slow:

$$\widehat{z}(x) = \frac{1}{\begin{array}{c}\text{\# trajectories}\\ \text{starting at } x\end{array}} \sum \exp(-\text{ trajectory cost})$$

Faster convergence is obtained using *temporal difference* learning:

$$\widehat{z}(x_n) \leftarrow (1-\beta)\,\widehat{z}(x_n) + \beta \exp(-q_n)\,\widehat{z}(x'_n)$$

The learning rate β should decrease over time.

Importance sampling

The expectation of a function $f(x)$ under a distribution $p(x)$ can be approximated as

$$E_{x \sim p(\cdot)}[f(x)] \approx \frac{1}{N} \sum_n f(x_n)$$

where $\{x_n\}_{n=1 \cdots N}$ are i.i.d. samples from $p(\cdot)$.

However, if $f(x)$ has interesting behavior in regions where $p(x)$ is small, convergence can be slow, i.e. we may need a very large N to obtain an accurate approximation. In the case of Z learning, the passive dynamics may rarely take the state to regions with low cost.

Importance sampling

The expectation of a function $f(x)$ under a distribution $p(x)$ can be approximated as

$$E_{x \sim p(\cdot)}[f(x)] \approx \frac{1}{N} \sum_n f(x_n)$$

where $\{x_n\}_{n=1 \cdots N}$ are i.i.d. samples from $p(\cdot)$.

However, if $f(x)$ has interesting behavior in regions where $p(x)$ is small, convergence can be slow, i.e. we may need a very large N to obtain an accurate approximation. In the case of Z learning, the passive dynamics may rarely take the state to regions with low cost.

Importance sampling is a general (unbiased) method for speeding up convergence. Let $q(x)$ be some other distribution which is better "adapted" to $f(x)$, and let $\{x_n\}$ now be samples from $q(\cdot)$. Then

$$E_{x \sim p(\cdot)}[f(x)] \approx \frac{1}{N} \sum_n \frac{p(x_n)}{q(x_n)} f(x_n)$$

This is essential for particle filters.

Greedy Z learning

Let $\widehat{u}\left(x'|x\right)$ denote the *greedy* control law, i.e. the control law which would be optimal if the current approximation $\widehat{z}\left(x\right)$ were the exact desirability function. Then we can sample from $\widehat{u}$ rather than p and use importance sampling:

$$\widehat{z}\left(x_n\right) \leftarrow \left(1-\beta\right)\widehat{z}\left(x_n\right) + \beta \frac{p\left(x'_n|x_n\right)}{\widehat{u}\left(x'_n|x_n\right)} \exp\left(-q_n\right)\widehat{z}\left(x'_n\right)$$

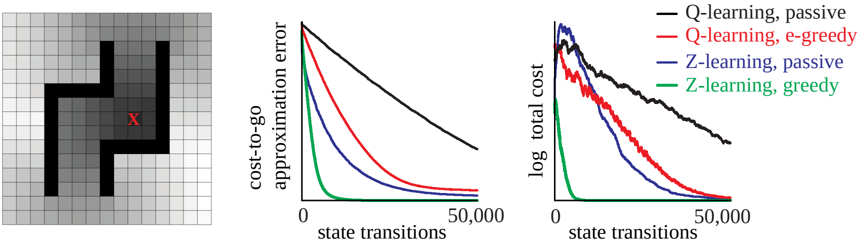
We now need access to the model $p\left(x'|x\right)$of the passive dynamics.

Greedy Z learning

Let $\widehat{u}(x'|x)$ denote the *greedy* control law, i.e. the control law which would be optimal if the current approximation $\widehat{z}(x)$ were the exact desirability function. Then we can sample from $\widehat{u}$ rather than p and use importance sampling:

$$\widehat{z}(x_n) \leftarrow (1-\beta)\,\widehat{z}(x_n) + \beta \frac{p(x'_n|x_n)}{\widehat{u}(x'_n|x_n)} \exp(-q_n)\,\widehat{z}(x'_n)$$

We now need access to the model $p(x'|x)$ of the passive dynamics.



Maximum principle for the most likely trajectory

Recall that for finite-horizon LMDPs we have

$$u_k^*\left(x'|x\right) = \exp\left(-q\left(x\right)\right) p\left(x'|x\right) \frac{z_{k+1}\left(x'\right)}{z_k\left(x\right)}$$

The probability that the optimally-controlled stochastic system initialized at state x_0 generates a given trajectory $x_1, x_2, \cdots x_T$ is

$$\begin{aligned} p^*\left(x_1, x_2, \cdots x_T | x_0\right) &= \prod_{k=0}^{T-1} u_k^*\left(x_{k+1}|x_k\right) \\ &= \prod_{k=0}^{T-1} \exp\left(-q\left(x_k\right)\right) p\left(x_{k+1}|x_k\right) \frac{z_{k+1}\left(x_{k+1}\right)}{z_k\left(x_k\right)} \\ &= \frac{\exp\left(-q_T\left(x_T\right)\right)}{z_0\left(x_0\right)} \prod_{k=0}^{T-1} \exp\left(-q\left(x_k\right)\right) p\left(x_{k+1}|x_k\right) \end{aligned}$$

Maximum principle for the most likely trajectory

Recall that for finite-horizon LMDPs we have

$$u_k^*\left(x'|x\right) = \exp\left(-q\left(x\right)\right) p\left(x'|x\right) \frac{z_{k+1}\left(x'\right)}{z_k\left(x\right)}$$

The probability that the optimally-controlled stochastic system initialized at state x_0 generates a given trajectory $x_1, x_2, \cdots x_T$ is

$$\begin{aligned} p^*\left(x_1, x_2, \cdots x_T | x_0\right) &= \prod_{k=0}^{T-1} u_k^*\left(x_{k+1}|x_k\right) \\ &= \prod_{k=0}^{T-1} \exp\left(-q\left(x_k\right)\right) p\left(x_{k+1}|x_k\right) \frac{z_{k+1}\left(x_{k+1}\right)}{z_k\left(x_k\right)} \\ &= \frac{\exp\left(-q_T\left(x_T\right)\right)}{z_0\left(x_0\right)} \prod_{k=0}^{T-1} \exp\left(-q\left(x_k\right)\right) p\left(x_{k+1}|x_k\right) \end{aligned}$$

Theorem (LMDP maximum principle)

*The most likely trajectory under p^* coincides with the optimal trajectory for a deterministic finite-horizon problem with final cost $q_T\left(x\right)$, dynamics $x' = f\left(x, u\right)$ where f can be **arbitrary**, and immediate cost $\ell\left(x, u\right) = q\left(x\right) - \log p\left(f\left(x, u\right), x\right)$.*

Trajectory probabilities in continuous time

There is no formula for the probability of a trajectory under the Ito diffusion $d\mathbf{x} = \mathbf{a}(\mathbf{x}) + Cd\boldsymbol{\omega}$. However the relative probabilities of two trajectories $\boldsymbol{\varphi}(t)$ and $\boldsymbol{\psi}(t)$ can be defined using the Onsager-Machlup functional:

$$OM[\boldsymbol{\varphi}(\cdot), \boldsymbol{\psi}(\cdot)] \triangleq \lim_{\varepsilon \to 0} \frac{p\left(\sup_t |\mathbf{x}(t) - \boldsymbol{\varphi}(t)| < \varepsilon\right)}{p\left(\sup_t |\mathbf{x}(t) - \boldsymbol{\psi}(t)| < \varepsilon\right)}$$

Trajectory probabilities in continuous time

There is no formula for the probability of a trajectory under the Ito diffusion $d\mathbf{x} = \mathbf{a}(\mathbf{x}) + C d\boldsymbol{\omega}$. However the relative probabilities of two trajectories $\boldsymbol{\varphi}(t)$ and $\boldsymbol{\psi}(t)$ can be defined using the Onsager-Machlup functional:

$$OM[\boldsymbol{\varphi}(\cdot), \boldsymbol{\psi}(\cdot)] \triangleq \lim_{\varepsilon \to 0} \frac{p(\sup_t |\mathbf{x}(t) - \boldsymbol{\varphi}(t)| < \varepsilon)}{p(\sup_t |\mathbf{x}(t) - \boldsymbol{\psi}(t)| < \varepsilon)}$$

It can be shown that

$$OM[\boldsymbol{\varphi}(\cdot), \boldsymbol{\psi}(\cdot)] = \exp\left(\int_0^T L(\boldsymbol{\psi}(t), \dot{\boldsymbol{\psi}}(t)) - L(\boldsymbol{\varphi}(t), \dot{\boldsymbol{\varphi}}(t))\, dt\right)$$

where

$$L[\mathbf{x}, \mathbf{v}] \triangleq \frac{1}{2}(\mathbf{a}(\mathbf{x}) - \mathbf{v})^{\mathsf{T}} \left(CC^{\mathsf{T}}\right)^{-1} (\mathbf{a}(\mathbf{x}) - \mathbf{v}) + \frac{1}{2}\operatorname{div}(\mathbf{a}(\mathbf{x}))$$

We can then fix $\boldsymbol{\psi}(t)$ and define the relative probability of a trajectory as

$$p_{OM}(\boldsymbol{\varphi}(\cdot)) = \exp\left(-\int_0^T L(\boldsymbol{\varphi}(t), \dot{\boldsymbol{\varphi}}(t))\, dt\right)$$

Continuous-time maximum principle

It can be shown that the trajectory maximizing $p_{OM}(\cdot)$ under the optimally-controlled stochastic dynamics for the problem

$$\begin{aligned} d\mathbf{x} &= \mathbf{a}(\mathbf{x}) + B(\mathbf{u}dt + \sigma d\boldsymbol{\omega}) \\ \ell(\mathbf{x},\mathbf{u}) &= q(\mathbf{x}) + \frac{1}{2\sigma^2}\|\mathbf{u}\|^2 \end{aligned}$$

coincides with the optimal trajectory for the deterministic problem

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{a}(\mathbf{x}) + B\mathbf{u} \\ \ell(\mathbf{x},\mathbf{u}) &= q(\mathbf{x}) + \frac{1}{2\sigma^2}\|\mathbf{u}\|^2 + \frac{1}{2}\operatorname{div}(\mathbf{a}(\mathbf{x})) \end{aligned}$$

Continuous-time maximum principle

It can be shown that the trajectory maximizing $p_{OM}(\cdot)$ under the optimally-controlled stochastic dynamics for the problem

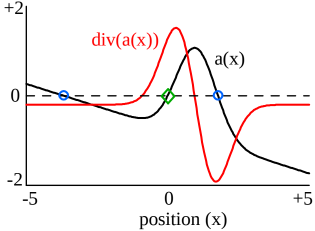
$$\begin{aligned} d\mathbf{x} &= \mathbf{a}(\mathbf{x}) + B(\mathbf{u}dt + \sigma d\boldsymbol{\omega}) \\ \ell(\mathbf{x}, \mathbf{u}) &= q(\mathbf{x}) + \frac{1}{2\sigma^2}\|\mathbf{u}\|^2 \end{aligned}$$

coincides with the optimal trajectory for the deterministic problem

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{a}(\mathbf{x}) + B\mathbf{u} \\ \ell(\mathbf{x}, \mathbf{u}) &= q(\mathbf{x}) + \frac{1}{2\sigma^2}\|\mathbf{u}\|^2 + \frac{1}{2}\operatorname{div}(\mathbf{a}(\mathbf{x})) \end{aligned}$$

Example:

$$\begin{aligned} dx &= (a(x) + u)\,dt + \sigma d\omega \\ \ell(x, u) &= \frac{1}{2\sigma^2}u^2 \end{aligned}$$



Example