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 \( p \left( x' \mid x, u \right) \). We can obtain a linearly-solvable MDP (LMDP) by allowing the controller to specify these probabilities directly:

\[
\begin{align*}
    x' & \sim u \left( \cdot \mid x \right) & \text{controlled dynamics} \\
    x' & \sim p \left( \cdot \mid x \right) & \text{passive dynamics} \\
    p \left( x' \mid x \right) = 0 & \Rightarrow u \left( x' \mid x \right) = 0 & \text{feasible control set } \mathcal{U} \left( x \right)
\end{align*}
\]
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:

$$x' \sim u (\cdot | x)$$ \hspace{1cm} \text{controlled dynamics}

$$x' \sim p (\cdot | x)$$ \hspace{1cm} \text{passive dynamics}

$$p (x' | x) = 0 \Rightarrow u (x' | x) = 0$$ \hspace{1cm} \text{feasible control set $\mathcal{U} (x)$}

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

KL cost over the probability simplex
Simplifying the Bellman equation (first exit)

\[ v(x) = \min_u \left\{ \ell(x, u) + E_{x' \sim p(.|x, u)} [v(x')] \right\} \]

\[ = \min_{u(.|x)} \left\{ q(x) + E_{x' \sim u(.|x)} \left[ \log \frac{u(x'|x)}{p(x'|x)} + \log \frac{1}{\exp(-v(x'))} \right] \right\} \]

\[ = \min_{u(.|x)} \left\{ q(x) + E_{x' \sim u(.|x)} \left[ \log \frac{u(x')}{p(x'|x) \exp(-v(x'))} \right] \right\} \]

The last term is an unnormalized KL divergence...
Simplifying the Bellman equation (first exit)

\[
v(x) = \min_u \left\{ \ell(x, u) + E_{x' \sim p( \cdot | x, u)} \left[ v(x') \right] \right\}
\]

\[
= \min_{u(\cdot | x)} \left\{ q(x) + E_{x' \sim u( \cdot | x)} \left[ \log \frac{u(x'|x)}{p(x'|x)} + \log \frac{1}{\exp(-v(x'))} \right] \right\}
\]

\[
= \min_{u(\cdot | x)} \left\{ q(x) + E_{x' \sim u( \cdot | x)} \left[ \log \frac{u(x')}{p(x'|x) \exp(-v(x'))} \right] \right\}
\]

The last term is an unnormalized KL divergence...

Definitions

**desirability** function \( z(x) \triangleq \exp(-v(x)) \)

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

\[
v(x) = \min_{u(\cdot | x)} \left\{ q(x) - \log \mathcal{P}[z](x) + KL \left( u(\cdot | x) \left\| \frac{p(\cdot | x) z(\cdot)}{\mathcal{P}[z](x)} \right\| \right) \right\}
\]
Linear Bellman equation and optimal control law

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

**Theorem (optimal control law)**

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

The Bellman equation becomes

\[
    v(x) = q(x) - \log \mathcal{P}[z](x)
\]

\[
    z(x) = \exp(-q(x)) \mathcal{P}[z](x)
\]

which can be written more explicitly as

**Theorem (linear Bellman equation)**

\[
    z(x) = \begin{cases} 
        \exp(-q(x)) \sum_{x'} p(x'|x) z(x') & : x \text{ non-terminal} \\
        \exp(-q_T(x)) & : x \text{ terminal}
    \end{cases}
\]
Illustration

\[ z(x') \]
\[ p(x'|x) \]
\[ u^*(x'|x) \sim p(x'|x) \, z(x') \]

*x' : sampled from \( u^*(x'|x) \)*
Summary of results

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

- **first exit** $z = \exp(-q) \mathcal{P}[z]$  
  $z = QPz$

- **finite horizon** $z_k = \exp(-q_k) \mathcal{P}_k[z_{k+1}]$  
  $z_k = Q_k P_k z_{k+1}$

- **average cost** $z = \exp(c-q) \mathcal{P}[z]$  
  $\lambda z = QPz$

- **discounted cost** $z = \exp(-q) \mathcal{P}[z^\alpha]$  
  $z = QPz^\alpha$
Let \( Q = \text{diag}(\exp(-q)) \) and \( P_{xy} = p(y|x) \). Then we have

- **first exit** \( z = \exp(-q) P[z] \quad z = QPz \)
- **finite horizon** \( z_k = \exp(-q_k) P_k [z_{k+1}] \quad z_k = Q_k P_k z_{k+1} \)
- **average cost** \( z = \exp(c-q) P[z] \quad \lambda z = QPz \)
- **discounted cost** \( z = \exp(-q) P[z^\alpha] \quad z = QPz^\alpha \)

In the first exit problem we can also write

\[
\begin{align*}
  z_N &= Q_N N P_N N z_N + b = (I - Q_N N P_N N)^{-1} b \\
  b &\triangleq Q_N N P_{N T} \exp(-q_T)
\end{align*}
\]

where \( N, T \) are the sets of non-terminal and terminal states respectively.

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

$$\mu(x') = \sum_x \frac{p(x'|x)z(x')}{\lambda \exp(q(x))z(x)} \mu(x)$$

$$\lambda r(x') = \sum_x \exp(-q(x)) p(x'|x) r(x)$$

In vector notation this becomes

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

Thus $\mathbf{z}$ and $\mathbf{r}$ are the right and left principal eigenvectors of $QP$, and $\mu = \mathbf{z} \cdot \mathbf{r}$
Comparison to policy and value iteration

- **Optimal control (z iter)**
  - Horizontal position: 
  - Tangential velocity: 

- **Optimal cost-to-go (z iter)**

- **Optimal cost-to-go (policy iter)**

---

Emo Todorov (UW)

AMATH/CSE 579, Winter 2014

Winter 2014 9 / 26
Given a graph and a set $T$ 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
- $q_T(x) = 0$ zero cost at terminal states

For large $\rho$ 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 $\tilde{u} \in \tilde{U} (x)$, transition probabilities $\tilde{p} (x'|x, \tilde{u})$ and costs $\tilde{\ell} (x, \tilde{u})$, 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 $\forall x, \tilde{u} \in \tilde{U} (x)$

$$q (x) + KL (\tilde{p} (\cdot | x, \tilde{u}) \| p (\cdot | x)) = \tilde{\ell} (x, \tilde{u})$$
$$q (x) - \sum_{x'} \tilde{p} (x'|x, \tilde{u}) \log p (x'|x) = \tilde{\ell} (x, \tilde{u}) + \tilde{h} (x, \tilde{u})$$

where $\tilde{h}$ is the entropy of $\tilde{p} (\cdot | x, \tilde{u})$. 
Embedding of traditional MDPs

Given a traditional MDP with controls \( \tilde{u} \in \tilde{U}(x) \), transition probabilities \( \tilde{p}(x'|x,\tilde{u}) \) and costs \( \tilde{\ell}(x,\tilde{u}) \), 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 \( \forall x, \tilde{u} \in \tilde{U}(x) \)

\[
q(x) + KL(\tilde{p}(|x,\tilde{u}) || p(|x)) = \tilde{\ell}(x,\tilde{u})
\]

\[
q(x) - \sum_{x'} \tilde{p}(x'|x,\tilde{u}) \log p(x'|x) = \tilde{\ell}(x,\tilde{u}) + \tilde{h}(x,\tilde{u})
\]

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

\[
q1 + \tilde{P}s = \tilde{b}
\]

\[
\exp(-s)^T1 = 1
\]

\( \tilde{P} \) and \( \tilde{b} = \tilde{\ell} + \tilde{h} \) are known, \( q \) and \( s \) are unknown. Assuming \( \tilde{P} \) is full rank,

\[
y = \tilde{P}^{-1}\tilde{b}, \quad s = y - q1, \quad q = -\log \left( \exp (-y)^T1 \right)
\]
Grid world example

MDP cost-to-go

LMDP cost-to-go

R² = 0.986

Emo Todorov (UW)
AMATH/CSE 579, Winter 2014
Winter 2014 13 / 26
Machine repair example

\[ R^2 = 0.993 \]

Performance on MDP

<table>
<thead>
<tr>
<th>Control Policy</th>
<th>MDP</th>
<th>LMDP</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>1.00</td>
<td>1.01</td>
<td>1.64</td>
</tr>
</tbody>
</table>

Emo Todorov (UW)
Continuous-time limit

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

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

Let \( z = \lim_{h \downarrow 0} z^{(h)} \). The limit yields \( z (x) = z (x) \),
Continuous-time limit

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

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

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

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

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

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

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

$$\mathcal{L} [z](x) = a(x)^T z_x(x) + \frac{1}{2} \text{trace} \left( \Sigma(x) z_{xx}(x) \right)$$
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?
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?

\[ dx = (a(x) + B(x)u)\,dt + C(x)\,d\omega \]

\[ \ell(x, u) = q(x) + \frac{1}{2}u^TR(x)u \]

Recall that for such problems we have $u^* = -R^{-1}B^Tv_x$ and

\[ 0 = q + a^Tv_x + \frac{1}{2} \text{tr} \left( C^Tz_{xx} \right) - \frac{1}{2}v_x^TR^{-1}B^Tv_x \]

Define $z(x) = \exp(-v(x))$ and write the PDE in terms of $z$:

\[ v_x = -\frac{z_x}{z}, \quad v_{xx} = -\frac{z_{xx}}{z} + \frac{z_xz_x^T}{z^2} \]

\[ 0 = q - \frac{1}{z} \left( a^Tz_x + \frac{1}{2} \text{tr} \left( C^Tz_{xx} \right) + \frac{1}{2z}z_x^TR^{-1}B^Tz_x - \frac{1}{2z}z_x^TCC^Tz_x \right) \]

Now if $CC^T = BR^{-1}B^T$, we obtain the linear HJB equation $qz = \mathcal{L}[z]$. 
The KL divergence between two Gaussians with means $\mu_1, \mu_2$ and common full-rank covariance $\Sigma$ is $\frac{1}{2} (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)$.

Using Euler discretization of the controlled diffusion, the passive and controlled dynamics have means $x + ha, x + ha + hBu$ and covariance $hCC^T$. Thus the KL divergence control cost is

$$\frac{1}{2} hu^T B^T (hCC^T)^{-1} hBu = \frac{h}{2} u^T B^T (BR^{-1}B^T)^{-1} Bu = \frac{h}{2} u^T Ru$$

This is the quadratic control cost accumulated over time $h$.

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

<table>
<thead>
<tr>
<th></th>
<th>discrete time</th>
<th>continuous time</th>
</tr>
</thead>
<tbody>
<tr>
<td>first exit</td>
<td>$\exp(q)z = \mathcal{P}[z]$</td>
<td>$qz = \mathcal{L}[z]$</td>
</tr>
<tr>
<td>finite horizon</td>
<td>$\exp(q_k)z_k = \mathcal{P}<em>k[z</em>{k+1}]$</td>
<td>$qz - z_t = \mathcal{L}[z]$</td>
</tr>
<tr>
<td>average cost</td>
<td>$\exp(q-c)z = \mathcal{P}[z]$</td>
<td>$(q-c)z = \mathcal{L}[z]$</td>
</tr>
<tr>
<td>discounted cost</td>
<td>$\exp(q)z = \mathcal{P}[z^\alpha]$</td>
<td>$z \log(z^\alpha) = \mathcal{L}[z]$</td>
</tr>
</tbody>
</table>
Summary of results

discrete time : continuous time :

first exit \[ \exp (q) z = \mathcal{P} [z] \quad q z = \mathcal{L} [z] \]

finite horizon \[ \exp (q_k) z_k = \mathcal{P}_k [z_{k+1}] \quad q z - z_t = \mathcal{L} [z] \]

average cost \[ \exp (q - c) z = \mathcal{P} [z] \quad (q - c) z = \mathcal{L} [z] \]

discounted cost \[ \exp (q) z = \mathcal{P} [z^\alpha] \quad z \log (z^\alpha) = \mathcal{L} [z] \]

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

\[
\mathcal{P} [z] (x) = E_{x' \sim p(\cdot \mid x)} [z (x')]
\]

\[
\mathcal{L} [z] (x) = \lim_{h \downarrow 0} \frac{E_{x' \sim p^{(h)}(\cdot \mid x)} [z (x')] - z (x)}{h} = \lim_{h \downarrow 0} \frac{\mathcal{P}^{(h)} [z] (x) - z (x)}{h}
\]

\[
\mathcal{P}^{(h)} [z] (x) = z (x) + h \mathcal{L} [z] (x) + o (h^2)
\]
Path-integral representation

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

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

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

\[
= \cdots
\]

\[
= E_{x_0 = x}^{x_{k+1} \sim p(\cdot|x_k)} \left[ \exp \left( -qT(x_{t_{\text{first}}}) - \sum_{k=0}^{t_{\text{first}}-1} q(x_k) \right) \right]
\]

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

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

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

\[
\begin{align*}
    z(x) &= \exp(-q(x)) \mathbb{E}_{x' \sim p(\cdot|x)}[z(x')]
    \\
    &= \exp(-q(x)) \mathbb{E}_{x' \sim p(\cdot|x)}[\exp(-q(x')) \mathbb{E}_{x'' \sim p(\cdot|x')}[z(x'')]]
    \\
    &= \ldots
    \\
    &= \mathbb{E}_{x_0 = x, x_{k+1} \sim p(\cdot|x_k)}[\exp(-q_T(x_{t_{\text{first}}}) - \sum_{k=0}^{t_{\text{first}}-1} q(x_k))]\exp(q_T(x_0))
\end{align*}
\]

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

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

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

\[
z(x) = \mathbb{E}_{dx = a(x)dt + C(x)d\omega}^{x(0) = x}[\exp(-q_T(x(t_{\text{first}})) - \int_0^{t_{\text{first}}} q(x(t))dt)]
\]
Model-free learning

The solution to the linear Bellman equation

\[ z(x) = \exp(-q(x)) \mathbb{E}_{x' \sim p(\cdot|x)} [z(x')] \]

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

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

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 convergence can be slow:

\[ \hat{z}(x) = \frac{1}{\# \text{ trajectories starting at } x} \sum \exp(-\text{trajectory cost}) \]
Model-free learning

The solution to the linear Bellman equation

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

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 convergence can be slow:

\[ \hat{z}(x) = \frac{1}{\# \text{ trajectories starting at } x} \sum \exp(-\text{trajectory cost}) \]

Faster convergence is obtained using temporal difference learning:

\[ \hat{z}(x_n) \leftarrow (1 - \beta) \hat{z}(x_n) + \beta \exp(-q_n) \hat{z}(x'_n) \]

The learning rate \(\beta\) 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 $\{x_n\}_{n=1}^{N}$ 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 \( \{x_n\}_{n=1}^{N} \) 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.
Let $\hat{u}(x'|x)$ denote the *greedy* control law, i.e. the control law which would be optimal if the current approximation $\hat{z}(x)$ were the exact desirability function. Then we can sample from $\hat{u}$ rather than $p$ and use importance sampling:

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

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

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

\[
\hat{z}(x_n) \leftarrow (1 - \beta) \hat{z}(x_n) + \beta \frac{p(x'_n|x_n)}{\hat{u}(x'_n|x_n)} \exp(-q_n) \hat{z}(x'_n)
\]

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

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

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

$$p^* (x_1, x_2, \cdots x_T|x_0) = \prod_{k=0}^{T-1} u_k^* (x_{k+1}|x_k)$$

$$= \prod_{k=0}^{T-1} \exp (-q(x_k)) p(x_{k+1}|x_k) \frac{z_{k+1}(x_{k+1})}{z_k(x_k)}$$

$$= \exp (-q_T(x_T)) \prod_{k=0}^{T-1} \exp (-q(x_k)) p(x_{k+1}|x_k)$$

$$= \frac{\exp (-q_T(x_T))}{z_0(x_0)} \prod_{k=0}^{T-1} \exp (-q(x_k)) p(x_{k+1}|x_k)$$
Maximum principle for the most likely trajectory

Recall that for finite-horizon LMDPs we have

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

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

\[
p^* (x_1, x_2, \ldots, x_T | x_0) = \prod_{k=0}^{T-1} u_k^* (x_{k+1} | x_k) = \prod_{k=0}^{T-1} \exp (-q(x_k)) p(x_{k+1} | x_k) \frac{z_{k+1} (x_{k+1})}{z_k (x_k)} = \frac{\exp (-q_T(x_T))}{z_0 (x_0)} \prod_{k=0}^{T-1} \exp (-q(x_k)) p(x_{k+1} | x_k)
\]

**Theorem (LMDP maximum principle)**

The most likely trajectory under \( p^* \) coincides with the optimal trajectory for a deterministic finite-horizon problem with final cost \( q_T (x) \), dynamics \( x' = f (x, u) \) where \( f \) can be arbitrary, and immediate cost \( \ell (x, u) = q (x) - \log p (f(x, u), x) \).
There is no formula for the probability of a trajectory under the Ito diffusion \( dx = a(x) + Cd\omega \). However, the relative probabilities of two trajectories \( \varphi(t) \) and \( \psi(t) \) can be defined using the Onsager-Machlup functional:

\[
OM[\varphi(\cdot), \psi(\cdot)] \triangleq \lim_{\varepsilon \to 0} \frac{p(\sup_t |x(t) - \varphi(t)| < \varepsilon)}{p(\sup_t |x(t) - \psi(t)| < \varepsilon)}
\]
Trajectory probabilities in continuous time

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

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

It can be shown that

$$OM[\varphi(\cdot), \psi(\cdot)] = \exp\left(\int_{0}^{T} L(\psi(t), \dot{\psi}(t)) - L(\varphi(t), \dot{\varphi}(t)) dt\right)$$

where

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

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

$$p_{OM}(\varphi(\cdot)) = \exp\left(-\int_{0}^{T} L(\varphi(t), \dot{\varphi}(t)) dt\right)$$
Continuous-time maximum principle

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

$$dx = a(x) + B(u dt + \sigma d\omega)$$

$$\ell(x, u) = q(x) + \frac{1}{2\sigma^2} \|u\|^2$$

coincides with the optimal trajectory for the deterministic problem

$$\dot{x} = a(x) + Bu$$

$$\ell(x, u) = q(x) + \frac{1}{2\sigma^2} \|u\|^2 + \frac{1}{2} \text{div}(a(x))$$
Continuous-time maximum principle

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

$$dx = a(x) + B(u dt + \sigma d\omega)$$

$$\ell(x, u) = q(x) + \frac{1}{2\sigma^2} \|u\|^2$$

coincides with the optimal trajectory for the deterministic problem

$$\dot{x} = a(x) + Bu$$

$$\ell(x, u) = q(x) + \frac{1}{2\sigma^2} \|u\|^2 + \frac{1}{2} \text{div}(a(x))$$

Example:

$$dx = (a(x) + u) dt + \sigma d\omega$$

$$\ell(x, u) = \frac{1}{2\sigma^2} u^2$$
Example

\[
\begin{align*}
\mu(x,t) & \\
r(x,t) & \\
z(x,t) & \\
\end{align*}
\]

\[
\sigma = 0.6
\]

\[
\begin{align*}
\sigma = 1.2
\end{align*}
\]