A new mathematical framework for optimal choice of actions

Emo Todorov

Department of Cognitive Science
University of California San Diego

Problem formulation

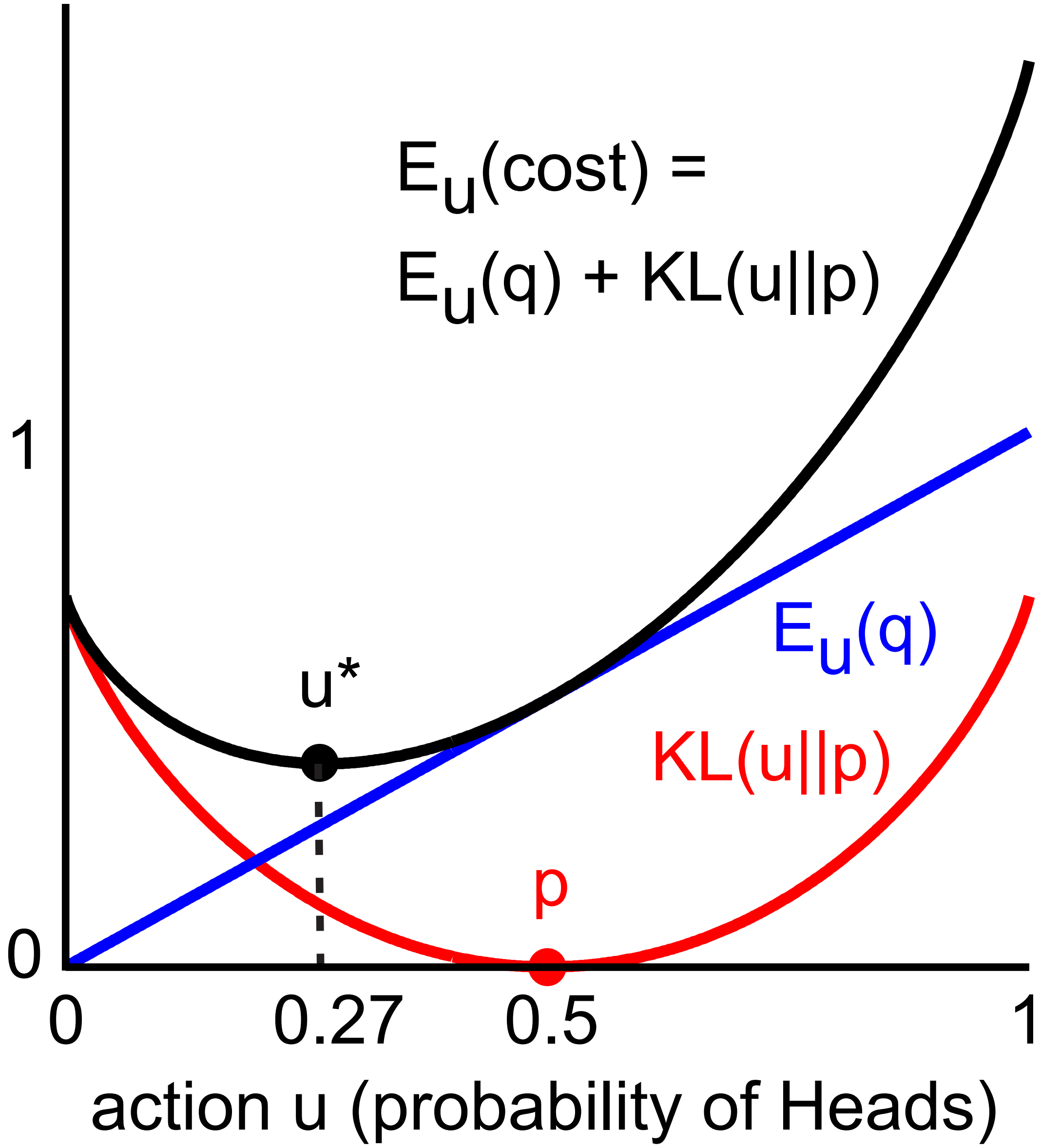
Instead of specifying actions a which determine the state transition probabilities p(x’ | x, a) here the controller specifies the state transition probabilities u(x’ | x) directly.

x	current state
x'	next state
$p\left(x'\vert x\right)$	transition probabilities under passive dynamics
$u\left(x'\vert x\right)$	transition probabilities under controlled dynamics
$q\left(x\right)$	state cost, i.e. cost for being in state x
$KL\left(u\vert\vert p\right)$	control cost, i.e. cost for choosing action $u\left(\cdot\vert x\right)$

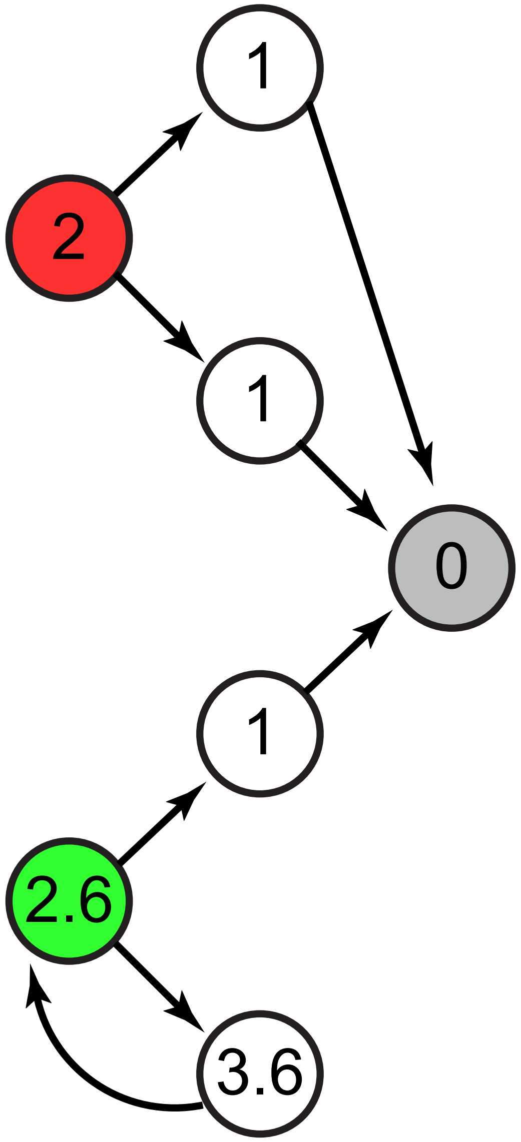
$$KL\left(u\left(\cdot|x\right)||p\left(\cdot|x\right)\right) = E_{x'\sim u(\cdot|x)} \log \frac{u\left(x'|x\right)}{p\left(x'|x\right)}$$

Understanding the KL control cost

biased coins



shortest paths



Reducing the problem to a linear equation

$u^*(x'\|x)$	optimal control law
$v(x)$	optimal cost-to-go function
$z(x) = \exp(-v(x))$	desirability function

Bellman equation:

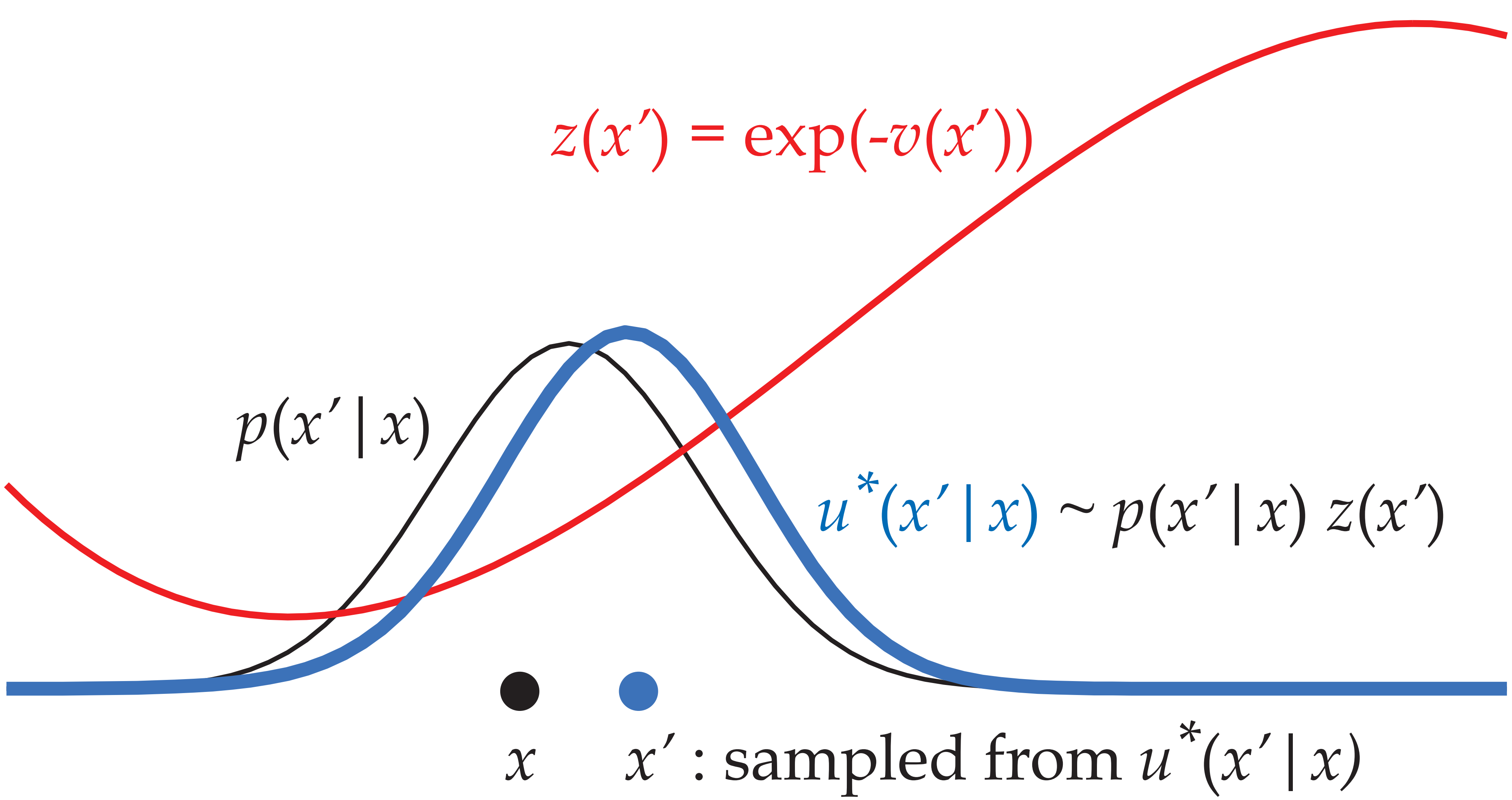
$$\begin{aligned} v(x) &= \min_u \left\{ q(x) + KL(u||p) + E_{x'\sim u(\cdot|x)}\left[v(x')\right] \right\} \\ &= q(x) + \min_u E_{x'\sim u(\cdot|x)}\left[\log\frac{u(x'|x)}{p(x'|x)} + \log\frac{1}{\exp(-v(x))}\right] \\ &= q(x) - \log E_{x'\sim p(\cdot|x)}\left[\exp(-v(x'))\right] + \min_u KL(u||p\exp(-v))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

Optimal control law: $u^*(x'|x) \propto p(x'|x)\, z(x')$

Desirability function: $z(x) = \exp(-q(x))\, E_{x'\sim p(\cdot|x)} z(x')$

Vector notation: $\mathbf{z} = QP\mathbf{z}$

Relationship between desirability and control


$z(x') = \exp(-v(x'))$
$p(x' \mid x)$
$u^*(x' \mid x) \sim p(x' \mid x)\, z(x')$
x
x' : sampled from $u^*(x' \mid x)$

Summary of results

Let $\mathcal{G}$ denote expectation under the passive dynamics: $\mathcal{G}[z](x) = E_{x' \sim p(\cdot|x)}[z(x')]$

Results for different performance criteria:

first-exit total cost	$z = \exp(-q)\,\mathcal{G}[z]$	$z(x)$ given on the boundary
finite-horizon total cost	$z_t = \exp(-q_t)\,\mathcal{G}_t[z_{t+1}]$	$z(x)$ given at the final time
infinite-horizon average cost	$z = \exp(c - q)\,\mathcal{G}[z]$	unknown average cost c
infinite-horizon discounted cost	$z = \exp(-q)\,\mathcal{G}[z^\alpha]$	discount factor α

Shortest paths

$p(x'|x) =$ random walk on the graph

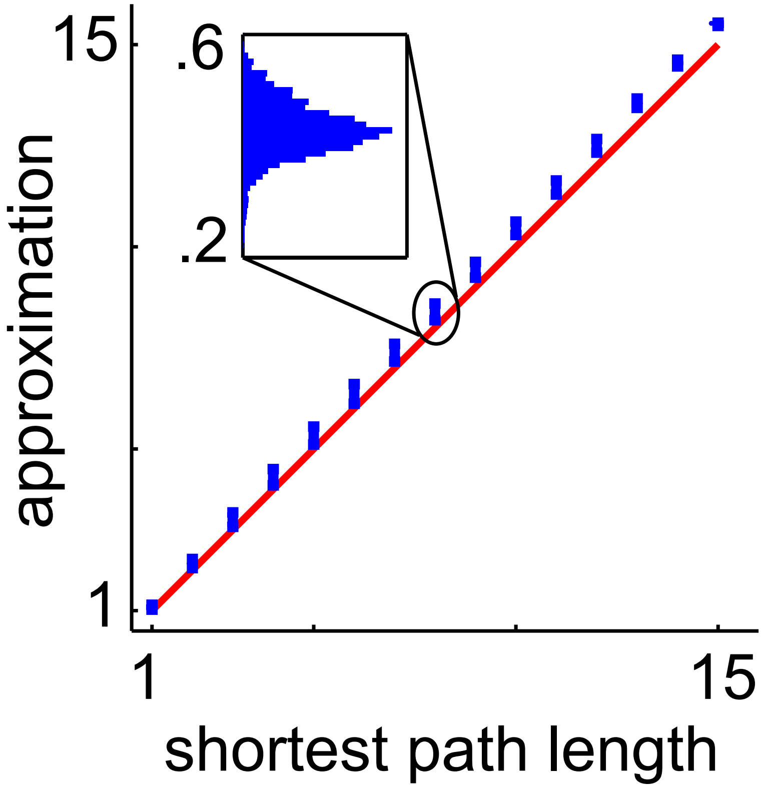
$$q\left(x\right) = \begin{cases} 0, & x \text{ terminal} \\ \rho, & x \text{ non-terminal} \end{cases}$$

For large ρ the optimal cost-to-go $v_\rho\left(x\right)$ is dominated by the state cost, thus:

$$\lim_{\rho\to\infty} \frac{v_\rho\left(x\right)}{\rho} = \text{shortest path from } x \text{ to a terminal state}$$

Performance on the graph of internet routers as of 2003: 190914 nodes, 609066 edges, data from caida.org.

When the approximation was rounded down to the nearest integer, all shortest paths were recovered exactly.



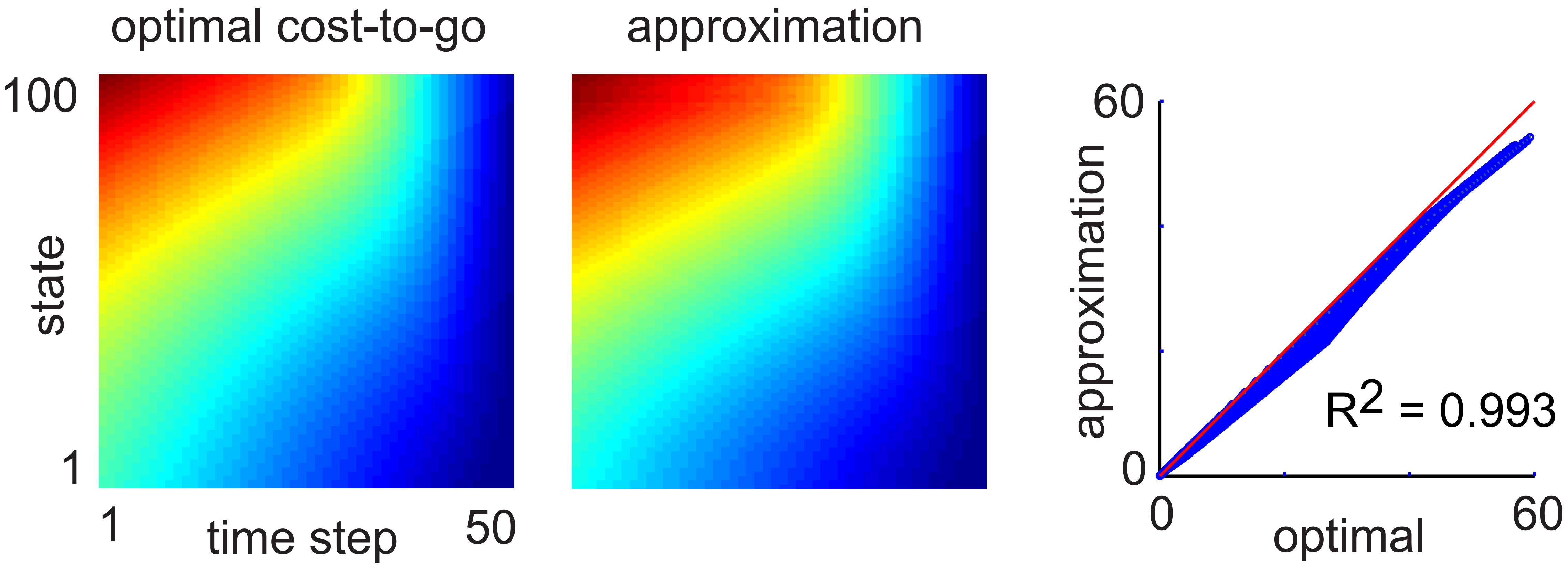
Embedding of traditional MDPs

Consider a traditional MDP with actions a, transition probabilities $\widetilde{p}\left(x'|x,a\right)$ and costs $\widetilde{\ell}\left(x,a\right)$. This MDP can be embedded in our family by choosing $q\left(x\right)$ and $p\left(x'|x\right)$ such that for every $\left(x,a\right)$ we have

$$q\left(x\right) + KL\left(\widetilde{p}\left(\cdot|x,a\right)||p\left(\cdot|x\right)\right) = \widetilde{\ell}\left(x,a\right)$$

Computing q and p requires solving a linear equation at every x.

machine repair example



optimal cost-to-go

approximation

Off-policy reinforcement learning

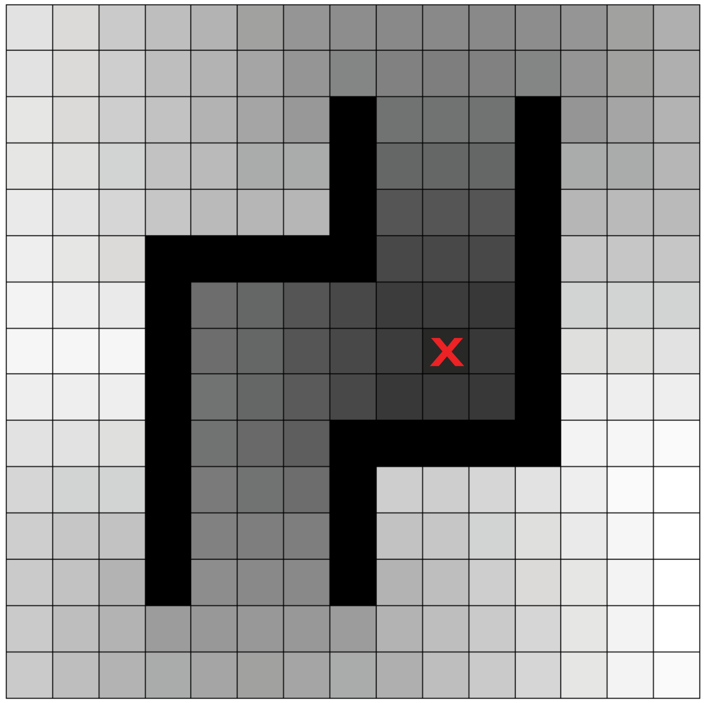
The linear Bellman equation $z(x) = \exp(-q(x))\, E_{x' \sim p(\cdot|x)} z(x')$ yields the following stochastic approximation method (Z-learning):

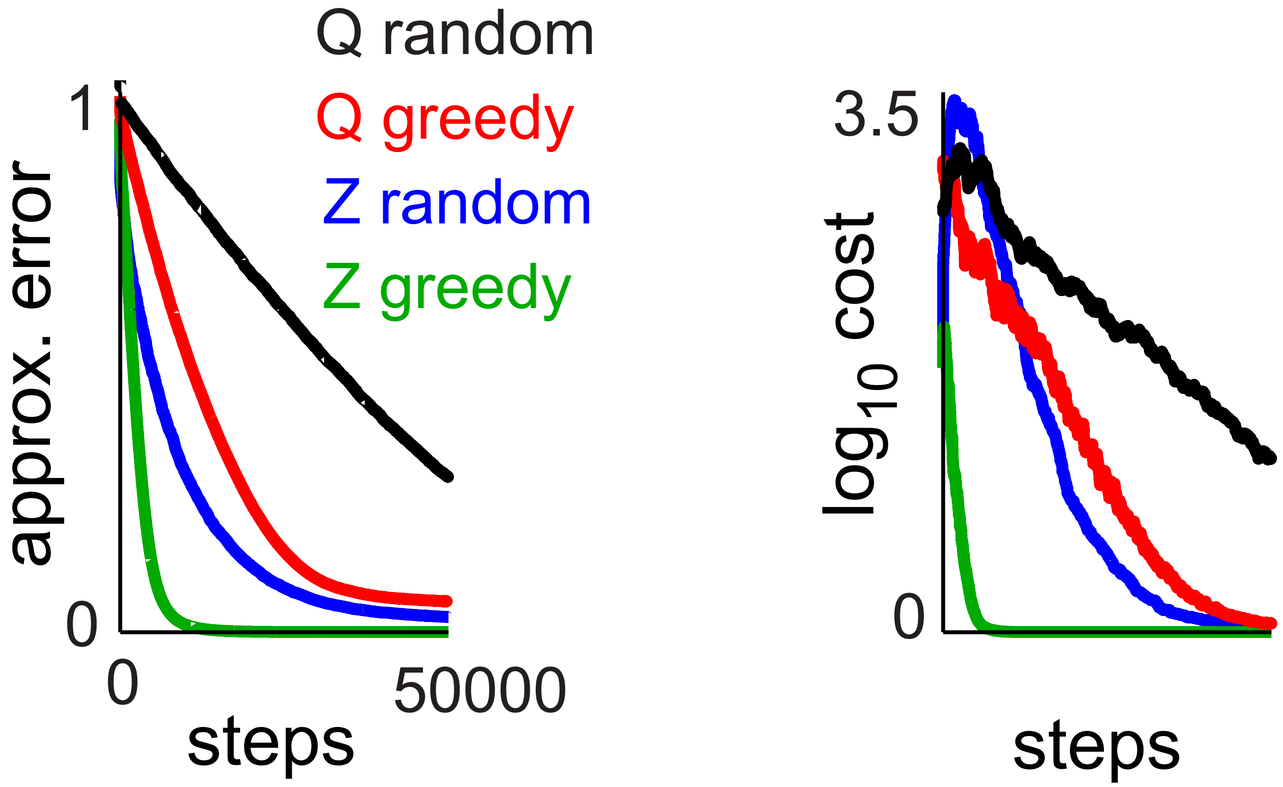
$$\widehat{z}(x_t) \leftarrow \eta_t \exp(-q_t)\, \widehat{z}(x_{t+1}) + (1 - \eta_t)\, \widehat{z}(x_t)$$

This learning rule is simpler and more efficient than Q-learning:

$$\widehat{Q}(x_t, a_t) \leftarrow \eta_t \left(\ell_t + \min_{a'} \widehat{Q}\left(x_{t+1}, a'\right)\right) + (1 - \eta_t)\, \widehat{Q}(x_t, a_t)$$

grid world example





Continuous analog

Control-affine diffusions with control-quadratic cost rate:

dynamics: $d\mathbf{x} = \mathbf{a}(\mathbf{x})\,dt + B(\mathbf{x})\,(\mathbf{u}dt + \sigma d\boldsymbol{\omega})$

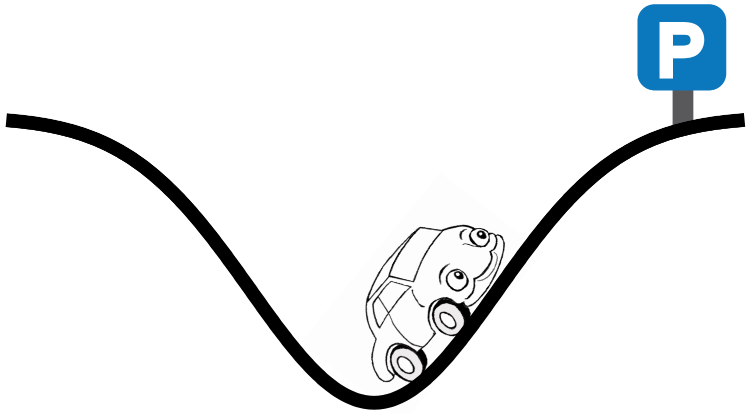
cost rate: $\ell(\mathbf{x}, \mathbf{u}) = q(\mathbf{x}) + \frac{1}{2\sigma^2}\|\mathbf{u}\|^2$

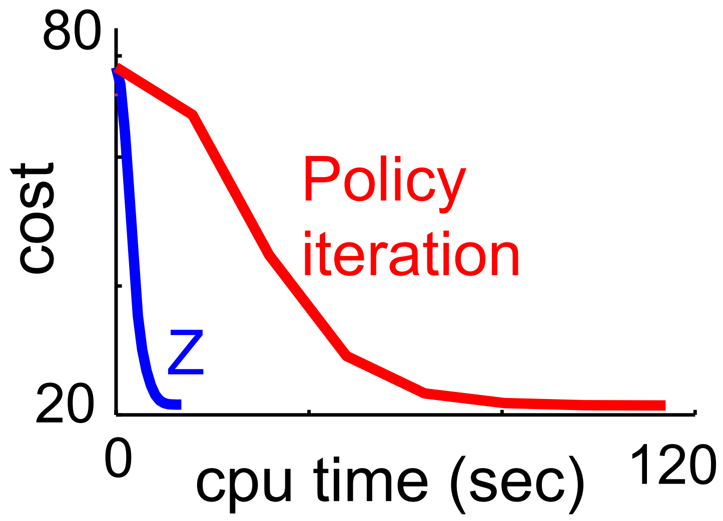
The optimal control law is $\mathbf{u}^*(\mathbf{x}) = -\sigma^2 B(\mathbf{x})^\mathsf{T} v_\mathbf{x}(\mathbf{x})$. The minimized HJB equations expressed in terms of the desirability function $z = \exp(-v)$ are

first-exit total cost:	$0 = \mathcal{L}[z] - qz$
finite-horizon total cost:	$-z_t = \mathcal{L}[z] - qz$
infinite-horizon average cost:	$-\lambda z = \mathcal{L}[z] - qz$
infinite-horizon discounted cost:	$\log(z^\alpha)\, z = \mathcal{L}[z] - qz$

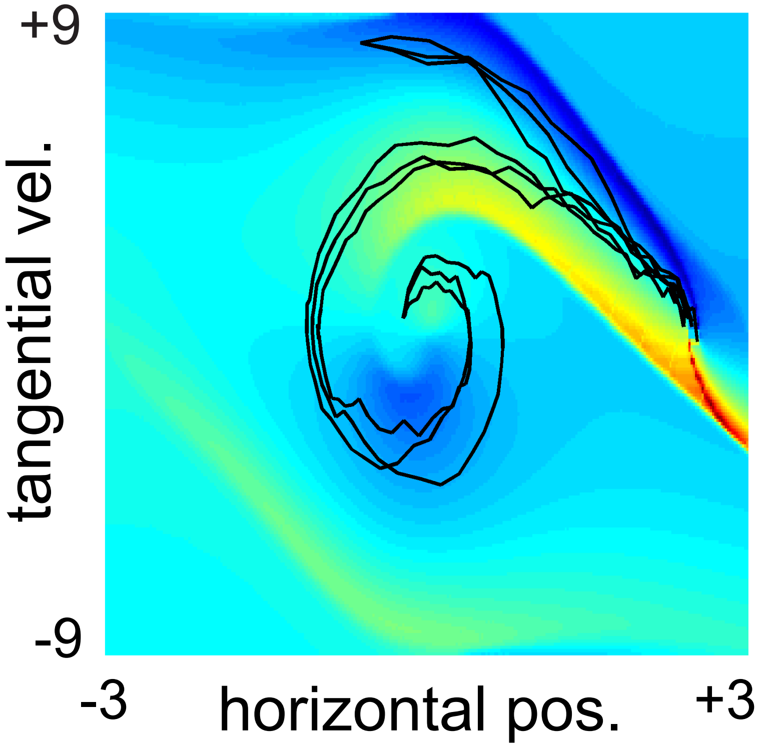
$\mathcal{L}$ is the generator of the uncontrolled diffusion: $\mathcal{L}[z] = \mathbf{a}^\mathsf{T} z_\mathbf{x} + \frac{\sigma^2}{2}\operatorname{tr}\left(BB^\mathsf{T} z_{\mathbf{xx}}\right)$

Novel discretization of continuous problems

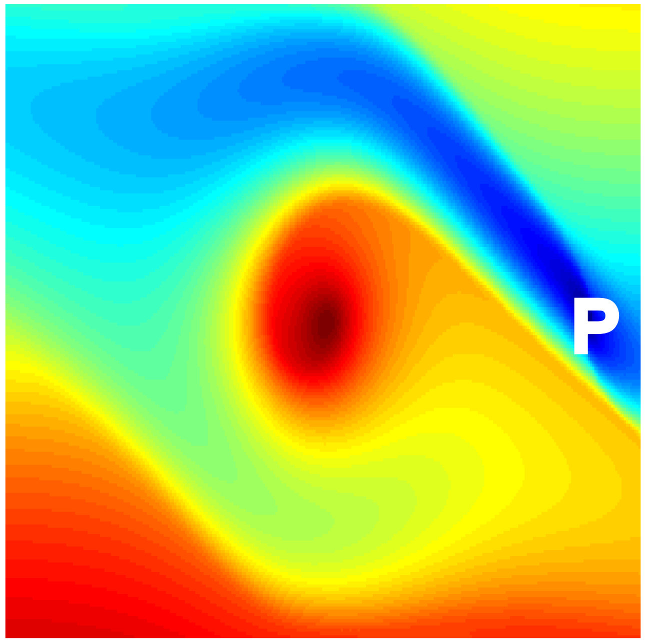




optimal control law



optimal cost-to-go



Function approximation methods

Define the function approximator

$$\widehat{z}\left(\mathbf{x};\boldsymbol{\theta},\mathbf{w}\right)=\sum_i w_i f_i\left(\mathbf{x}\right)$$

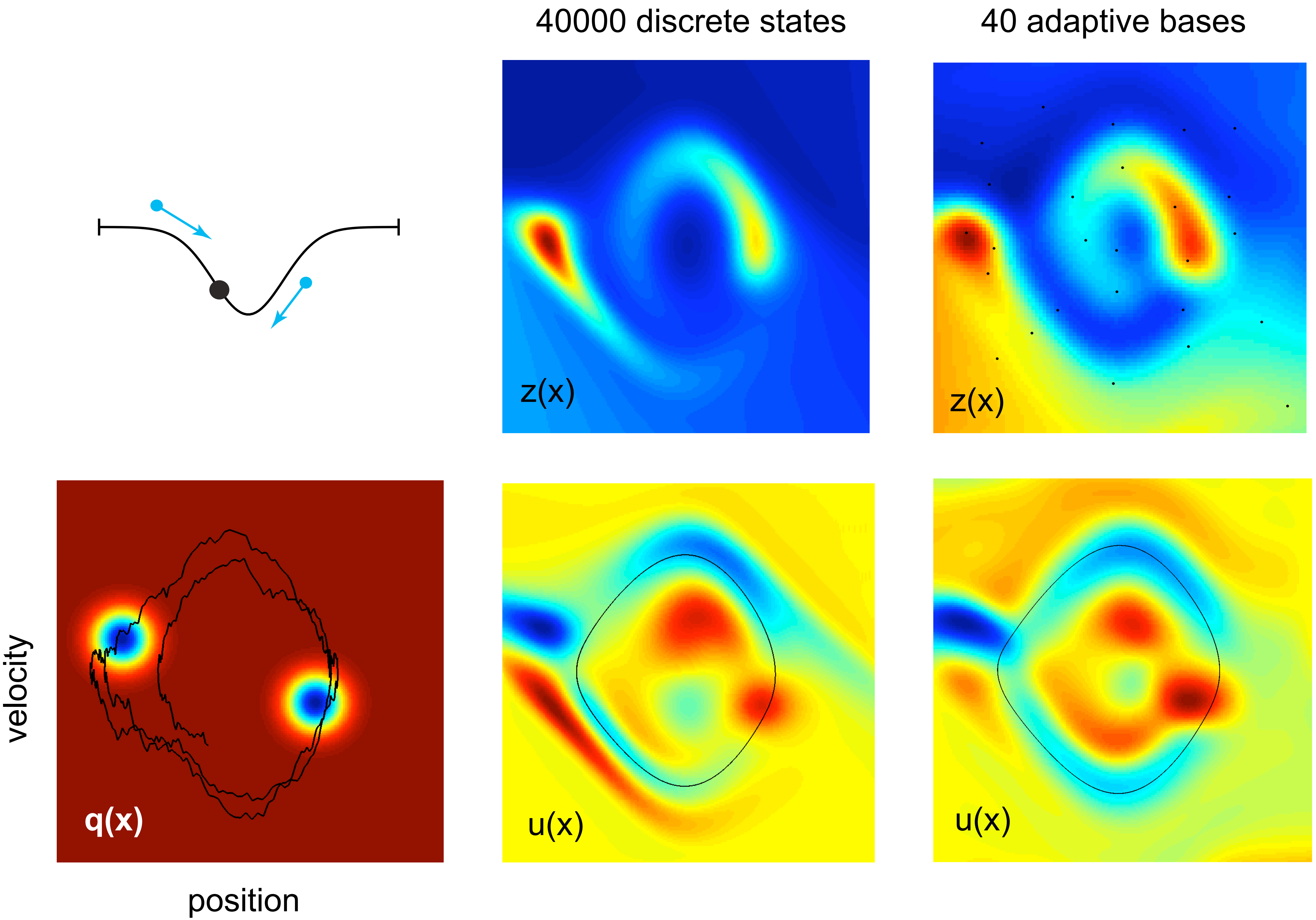
where f_i are Gaussians with means and Covariances contained in the parameter vector $\boldsymbol{\theta}$. Choose a set of collocation states $\{\mathbf{x}_n\}$ at which the Bellman equations will be enforced. Define the matrices $F\left(\boldsymbol{\theta}\right), G\left(\boldsymbol{\theta}\right)$ with elements $F_{ni}=f_i\left(\mathbf{x}_n\right)$ and $G_{ni}=\exp\left(-q\left(\mathbf{x}_n\right)\right)\mathcal{G}\left[f_i\right]\left(\mathbf{x}_n\right)$. The problem becomes

first-exit total cost: $F\left(\boldsymbol{\theta}\right)\mathbf{w}=G\left(\boldsymbol{\theta}\right)\mathbf{w}+\mathbf{b}$

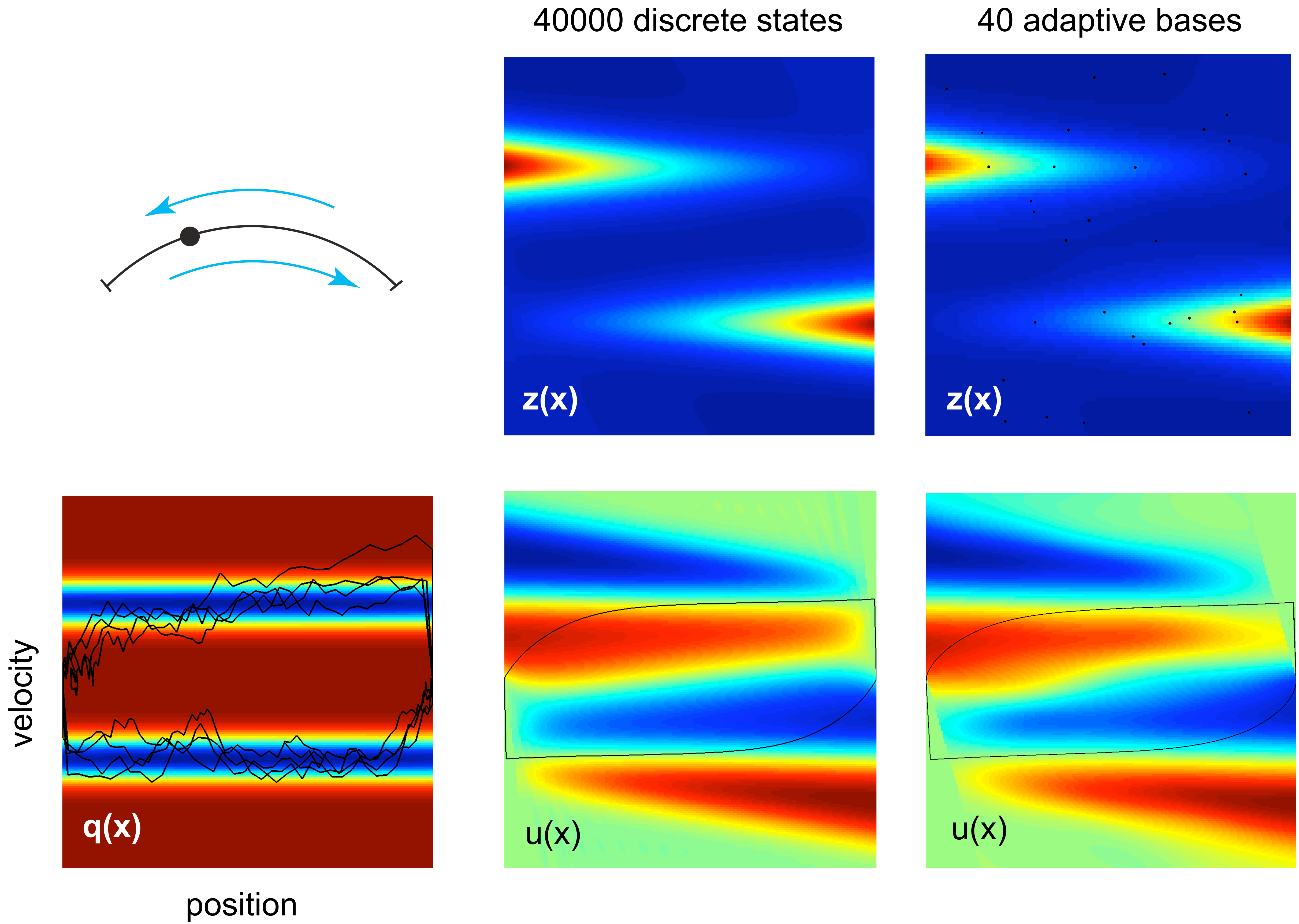
infinite-horizon average cost: $\lambda F\left(\boldsymbol{\theta}\right)\mathbf{w}=G\left(\boldsymbol{\theta}\right)\mathbf{w}$

Solving for $\lambda, \mathbf{w}$ is a linear problem. $\boldsymbol{\theta}$ can be optimized using Gauss-Newton. The collocation set can span the entire state space, or just the region where good solutions are expected. Automatic improvement of the collocation set is also possible (resembling Differential Dynamic Programming).

Example: car on a hill



Example: inverted pendulum



Compositionality of optimal control laws

Consider a *composite* first-exit problem with final/boundary cost in the form

$$b(\mathbf{x}) = -\log\left(\sum\nolimits_k w_k \exp\left(-b_k(\mathbf{x})\right)\right)$$

where b_k are the final costs for *component* problems whose solutions $z_k(\mathbf{x})$ we already have. Then the solution to the composite problem is simply

$$z(\mathbf{x}) = \sum\nolimits_k w_k z_k(\mathbf{x})$$

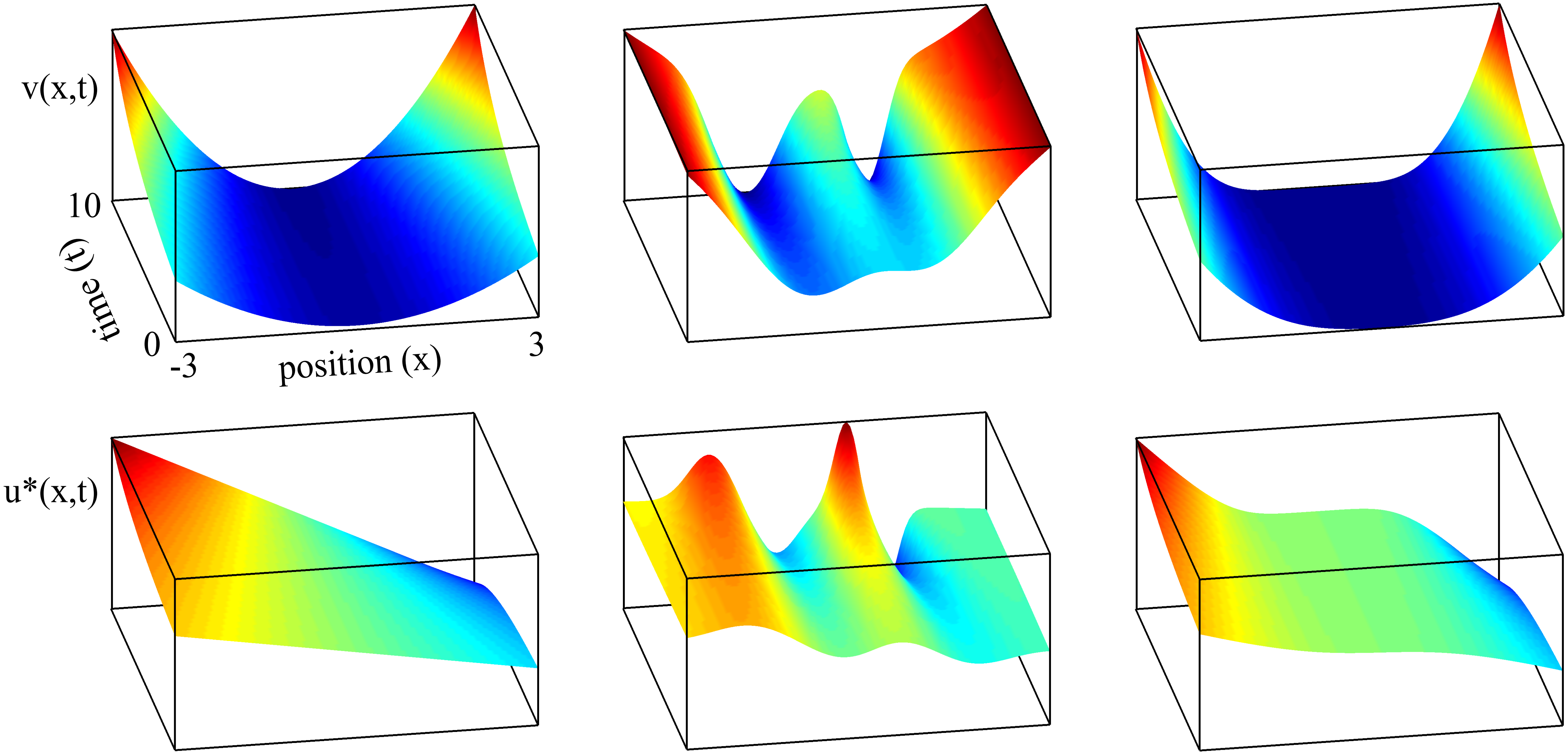
More generally, consider the "Green's function" relating the vectors $\mathbf{z}_I$ and $\mathbf{z}_B$ of desirabilities at interior and boundary states:

$$\begin{aligned} \mathbf{z}_I &= M\mathbf{z}_I + N\mathbf{z}_B \\ \mathbf{z}_I &= (I - M)^{-1} N\mathbf{z}_B \end{aligned}$$

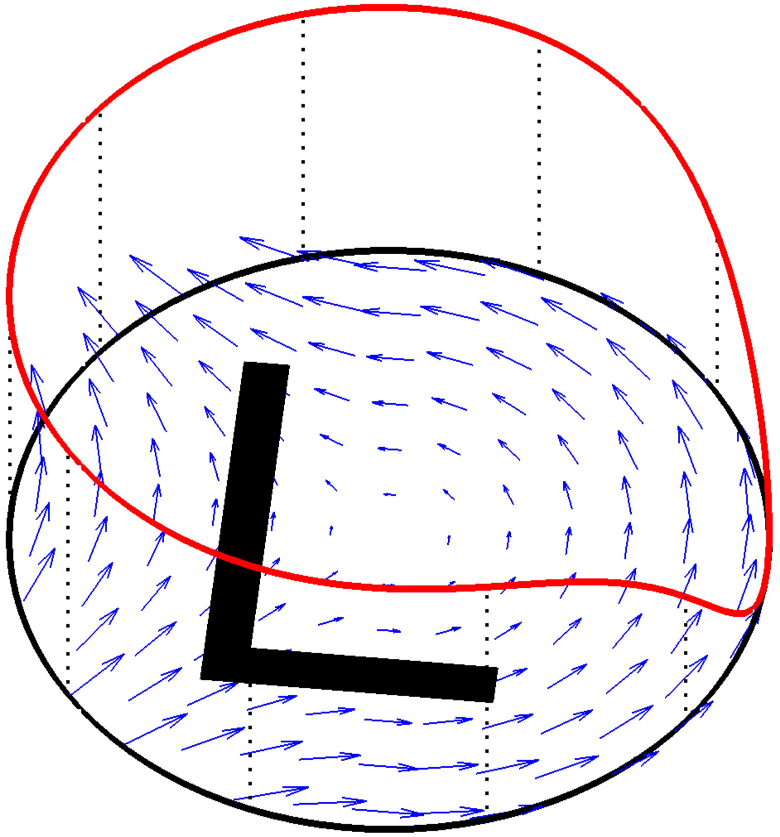
Replace $\exp(-b_k(\mathbf{x}))$ with the left singular vectors of $(I - M)^{-1} N$.

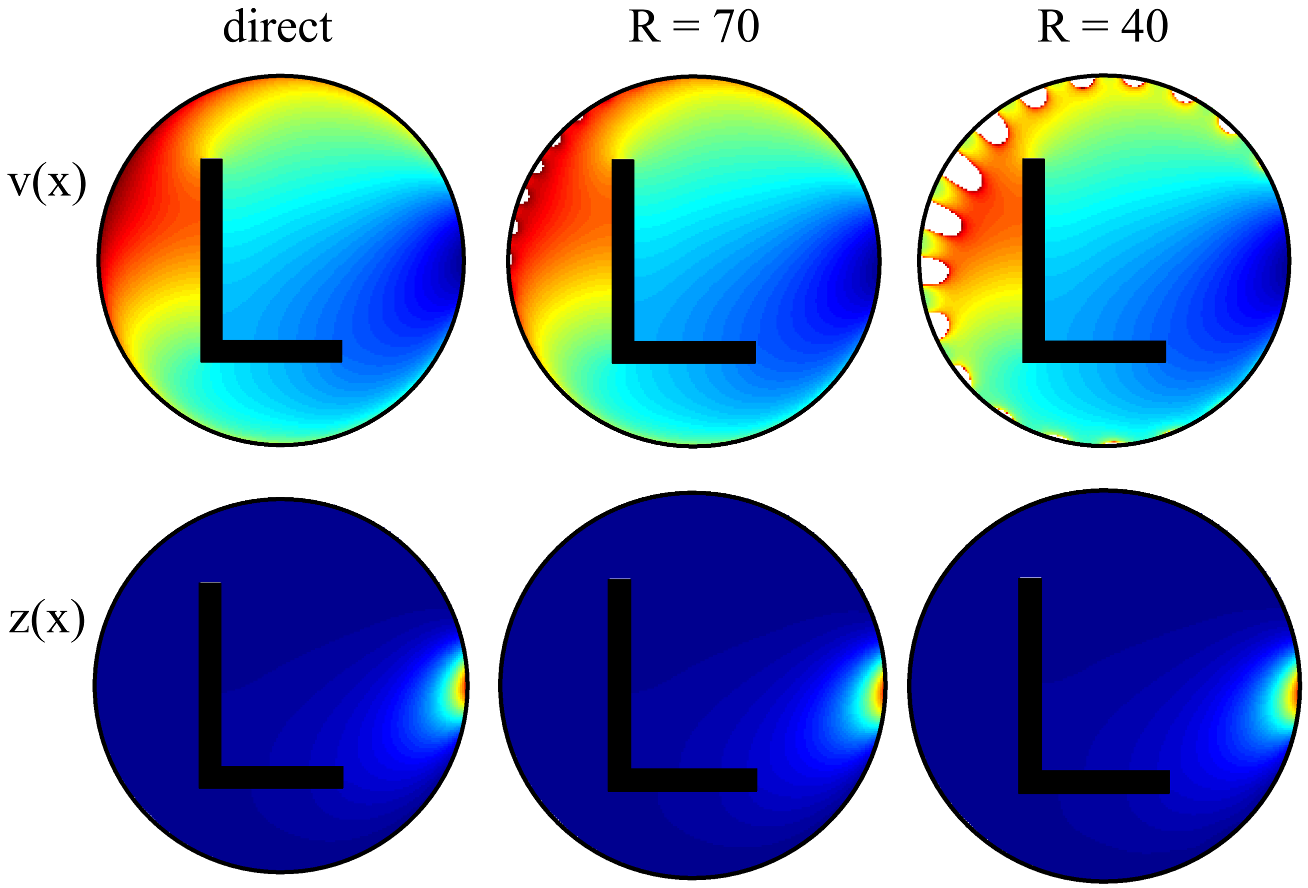
This can provide a universal basis for approximating any composite cost.

Linear-quadratic primitives

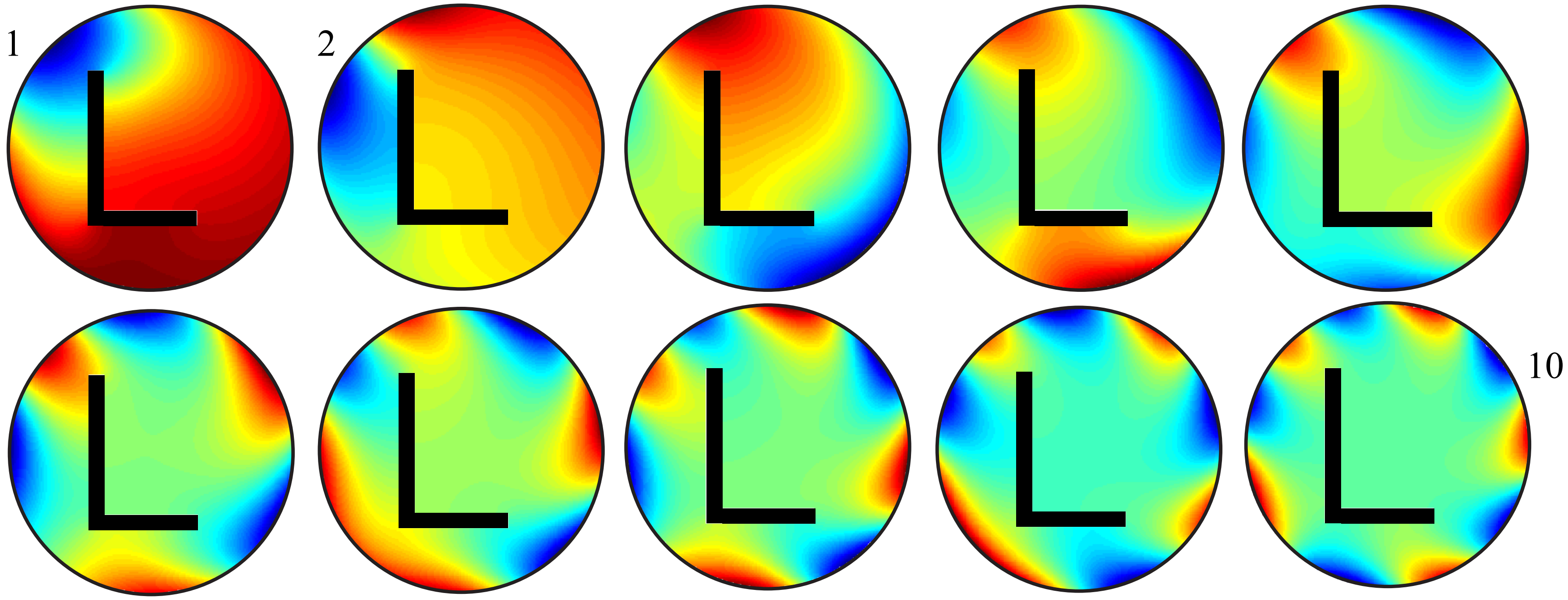


Singular vectors of Green’s function





top 10 singular vectors:



Duality with Bayesian inference

In the finite horizon we can show that the optimal control law is

$$u_t^*\left(\mathbf{x}'|\mathbf{x}\right) = \exp\left(-q_t\left(\mathbf{x}\right)\right) p\left(\mathbf{x}'|\mathbf{x}\right) \frac{z_{t+1}\left(\mathbf{x}'\right)}{z_t\left(\mathbf{x}\right)}$$

Thus the probability of a trajectory $\mathbf{x}_0, \mathbf{x}_1, \cdots \mathbf{x}_{t_f}$ is

$$\prod_t u_t^*\left(\mathbf{x}_{t+1}|\mathbf{x}_t\right) = \frac{z_{t_f}\left(\mathbf{x}_{t_f}\right)}{z_0\left(\mathbf{x}_0\right)} \prod_t \exp\left(-q_t\left(\mathbf{x}_t\right)\right) p\left(\mathbf{x}_{t+1}|\mathbf{x}_t\right)$$

The latter expression equals the probability of a (hidden) trajectory in a partially observed system, with dynamics p and emission probabilities which satisfy

$$p_{\mathbf{y}}\left(\mathbf{y}_t|\mathbf{x}\right) = \exp\left(-q_t\left(\mathbf{x}\right)\right)$$

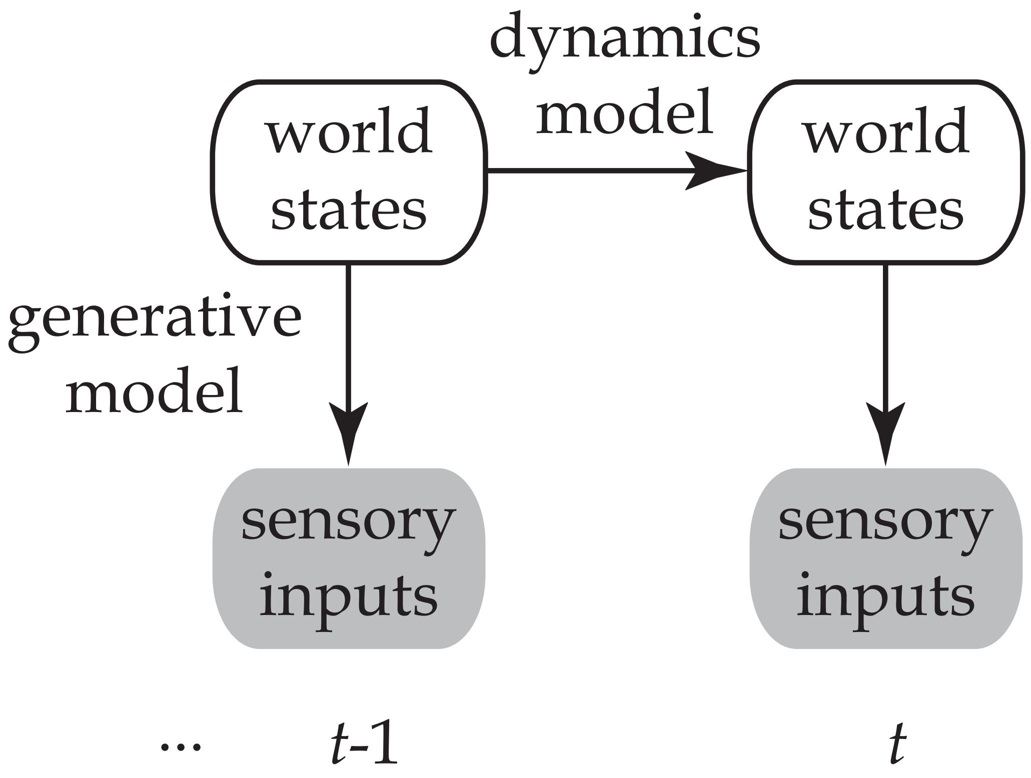
Thus the state distribution under the optimal control law corresponds to the posterior distribution in a Bayesian estimation problem. The desirability function corresponds to the backward filtering density.

Summary of duality results

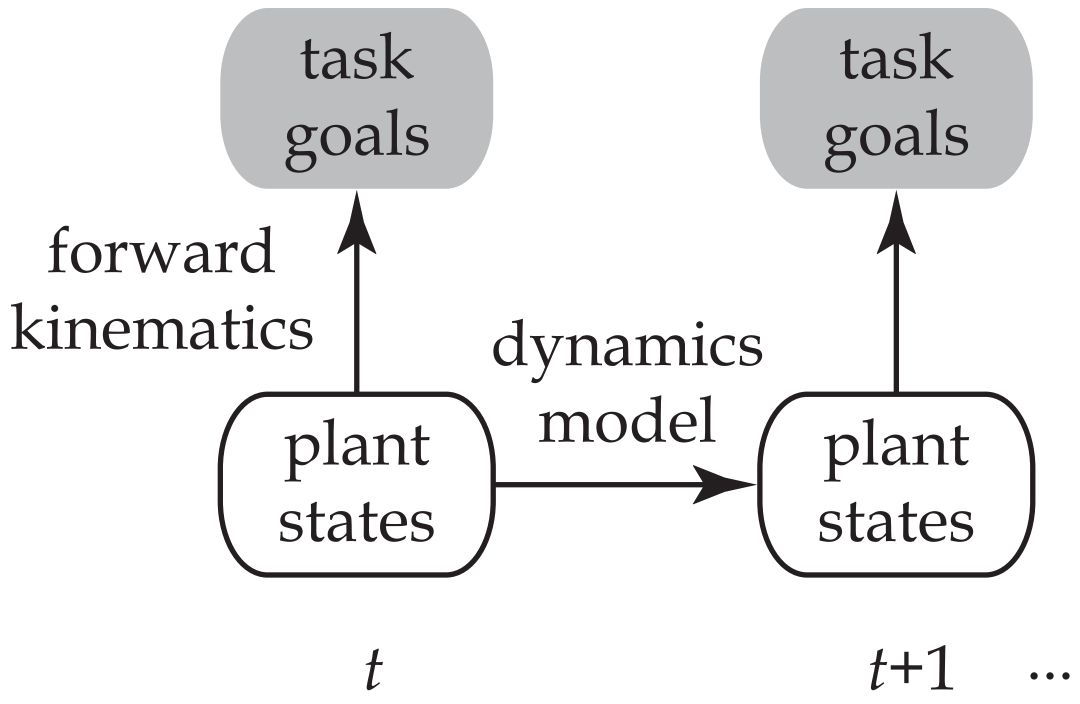
	continuous	discrete
control	$d\mathbf{x} = \mathbf{a}(\mathbf{x})\,dt + B(\mathbf{x})(\mathbf{u}dt + d\boldsymbol{\omega})$ $\ell(\mathbf{x}, \mathbf{u}, t) = q(\mathbf{x}, t) + \frac{1}{2}\|\mathbf{u}\|^2$	$\mathbf{x}_{t+1} \sim u(\cdot\vert\mathbf{x}_t)$ $\ell_t(\mathbf{x}, u) = q_t(\mathbf{x}) + KL(u\vert\vert p)$
	is dual to	is dual to
estimation	$d\mathbf{x} = \mathbf{a}(\mathbf{x})\,dt + B(\mathbf{x})\,d\boldsymbol{\omega}$ $d\mathbf{y} = \mathbf{h}(\mathbf{x})\,dt + d\boldsymbol{\nu}$	$\mathbf{x}_{t+1} \sim p(\cdot\vert\mathbf{x}_t)$ $\mathbf{y}_t \sim p_{\mathbf{y}}(\cdot\vert\mathbf{x}_t)$
	when	when
	$q(\mathbf{x}, t) = \frac{1}{2}\|\mathbf{h}(\mathbf{x})\|^2 - \mathbf{h}(\mathbf{x})^\mathsf{T}\,\dot{\mathbf{y}}(t)$	$q_t(\mathbf{x}) = -\log(p_{\mathbf{y}}(\mathbf{y}_t\vert\mathbf{x}))$

Belief networks for estimation and control

estimation



control



Characterizing the most likely trajectory

The probability of a given trajectory under the optimal control law is

$$\prod_t \exp\left(-q_t\left(\mathbf{x}_t\right)\right) p\left(\mathbf{x}_{t+1} | \mathbf{x}_t\right)$$

Maximizing probability is equivalent to minimizing negative log-probability:

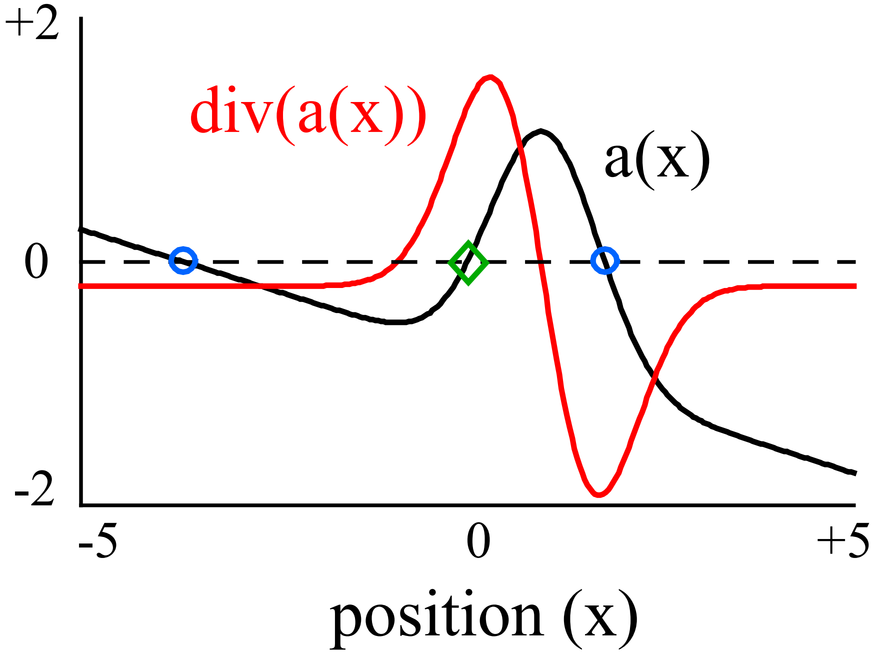
$$\sum_t q_t\left(\mathbf{x}_t\right) - \log\left(p\left(\mathbf{x}_{t+1} | \mathbf{x}_t\right)\right)$$

This can be interpreted as the total cost for a deterministic optimal control problem with control cost $-\log\left(p\left(\mathbf{x}_{t+1} | \mathbf{x}_t\right)\right)$.

In the diffusion case, the most probable trajectory equals the optimal trajectory for a deterministic problem with modified cost rate:

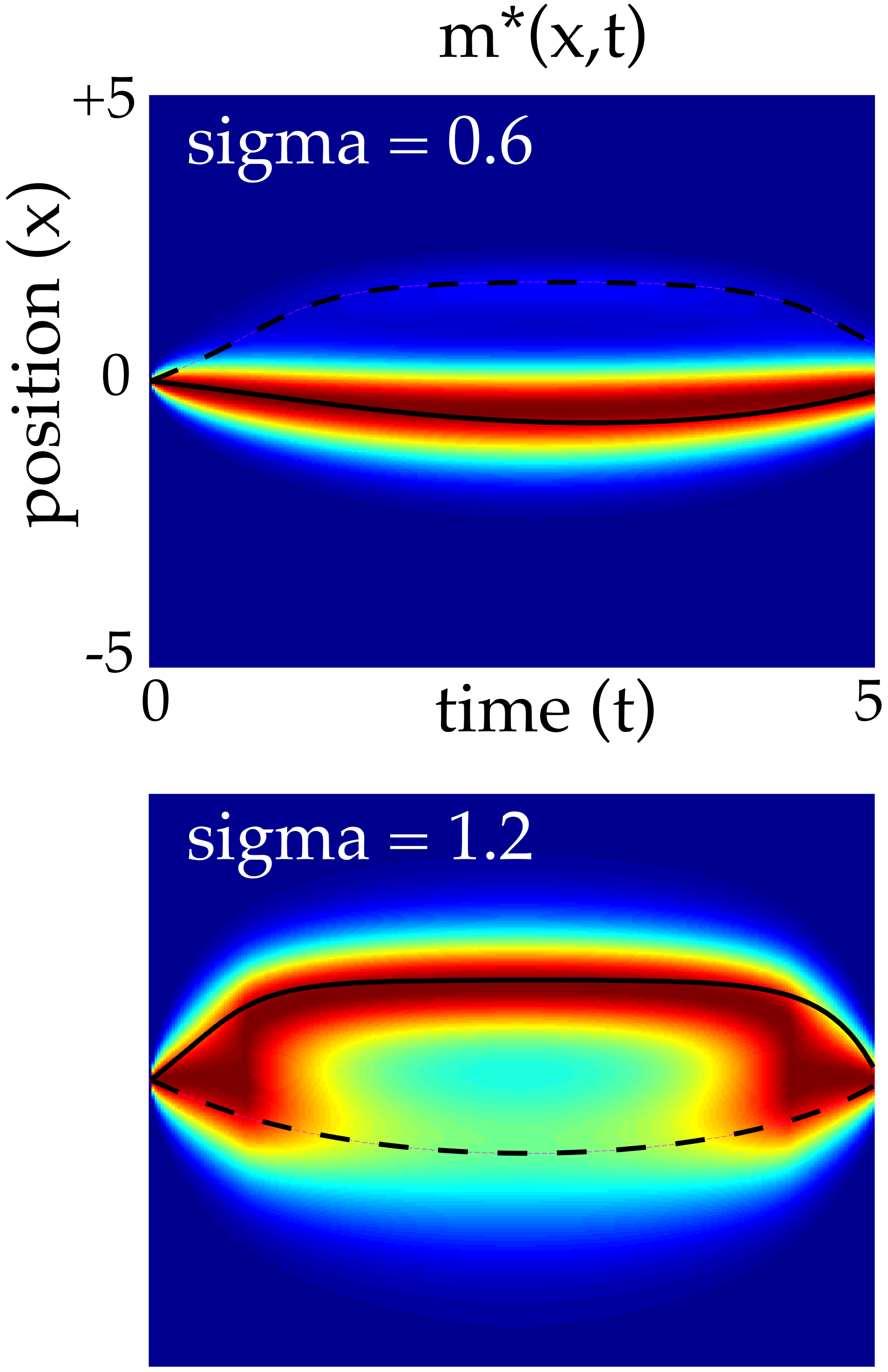
$$\widetilde{\ell}\left(\mathbf{x}, \mathbf{u}\right) = \ell\left(\mathbf{x}, \mathbf{u}\right) + \frac{1}{2}\log\left(\det\left(\Sigma\left(\mathbf{x}\right)\right)\right) + \frac{\Delta t}{2}\operatorname{div}\left(\mathbf{a}\left(\mathbf{x}\right)\right)$$

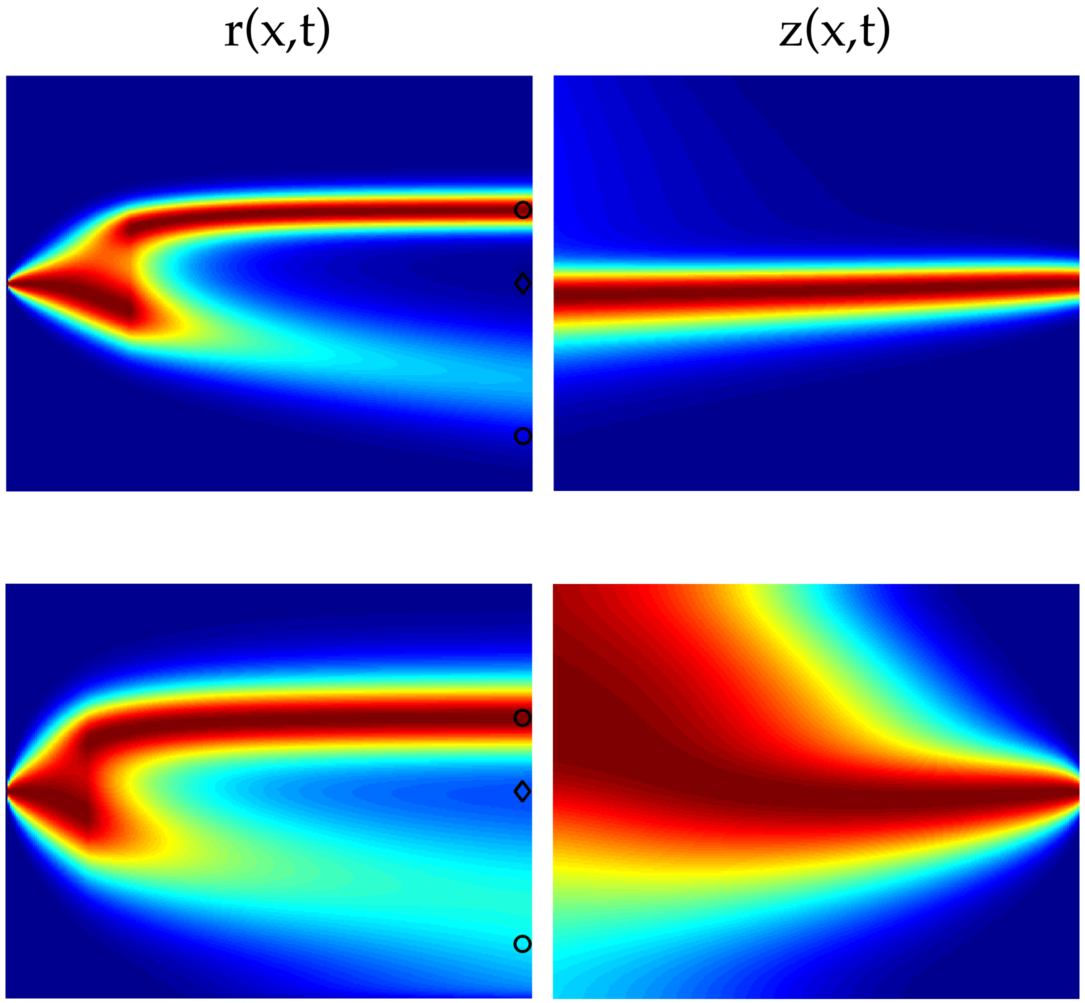
Example



$$dx = a(x)\, dt + u\, dt + \text{sigma}\, dw$$

$$q(x) = 5\, x^2$$





Inverse optimal control

Suppose we are given a dataset of state transitions $\mathcal{D} = \{x_n, x'_n\}_{n=1\cdots N}$ generated by an optimally-controlled system. Our goal is to infer the cost $q(x)$ for which the system is optimal. The passive dynamics $p(x'|x)$ are known.

This can be done by inferring $v(x)$, computing $z(x) = \exp(-v(x))$, and substituting in the linear Bellman equation to obtain $q(x)$.

It can be shown that the negative log-likelihood is

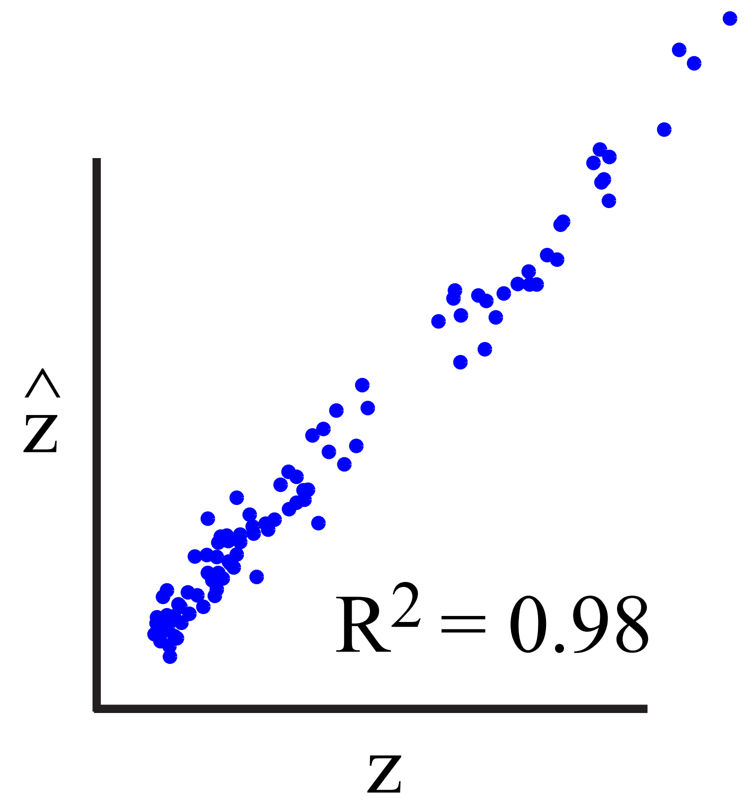
$$L(v(\cdot)) = \sum_x a(x)\, v(x) + \sum_x b(x) \log \sum_{x'} p\left(x'|x\right) \exp\left(-v\left(x'\right)\right)$$

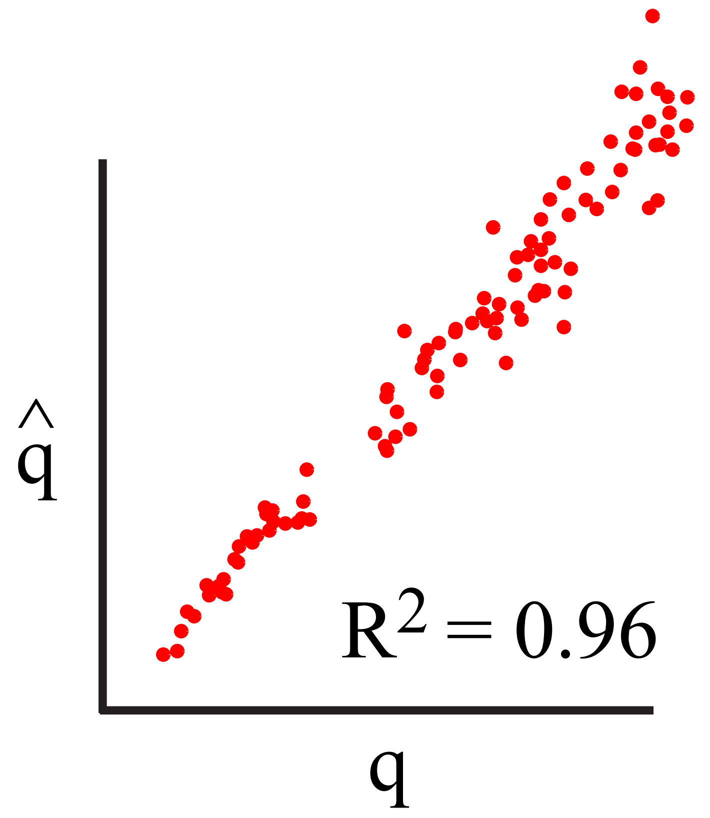
where $b(\cdot)$ and $a(\cdot)$ are the histograms of x_n and x'_n respectively.

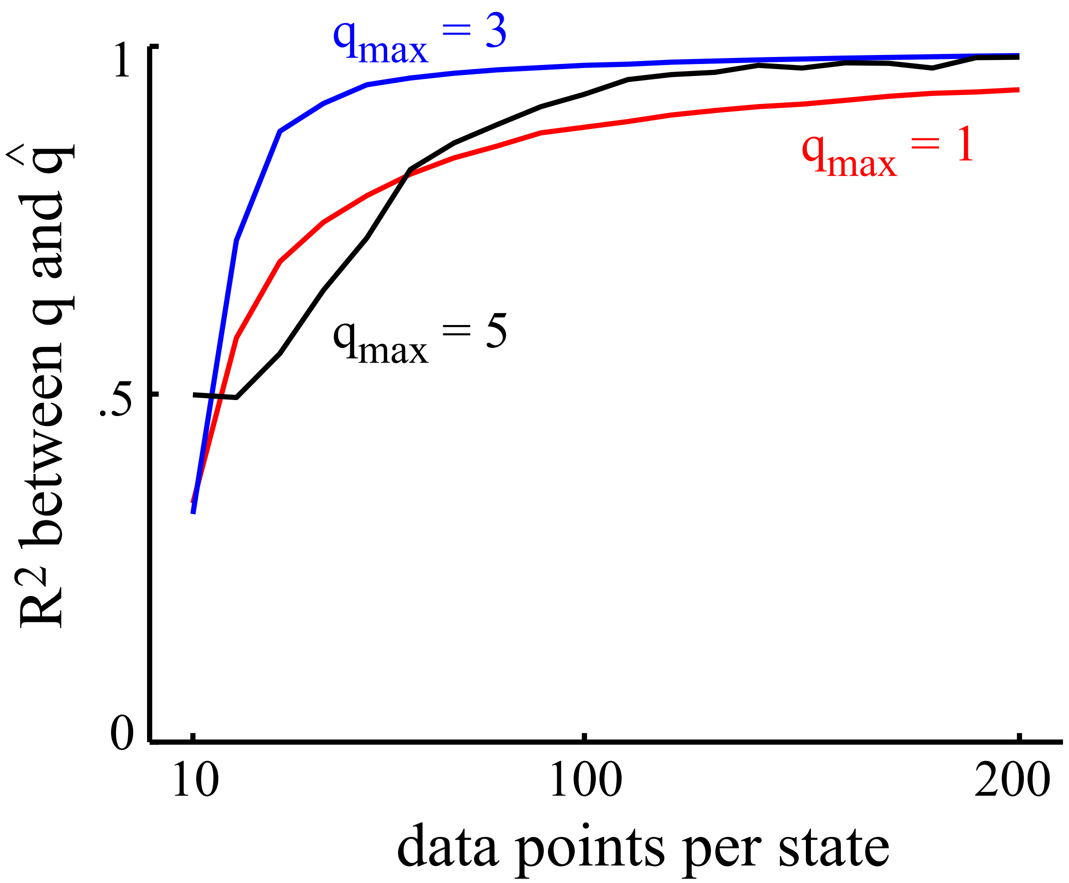
The function L is **convex** in v. Numerically, the Hessian H turns out to be diagonally dominant. This yields an efficient quasi-Newton method:

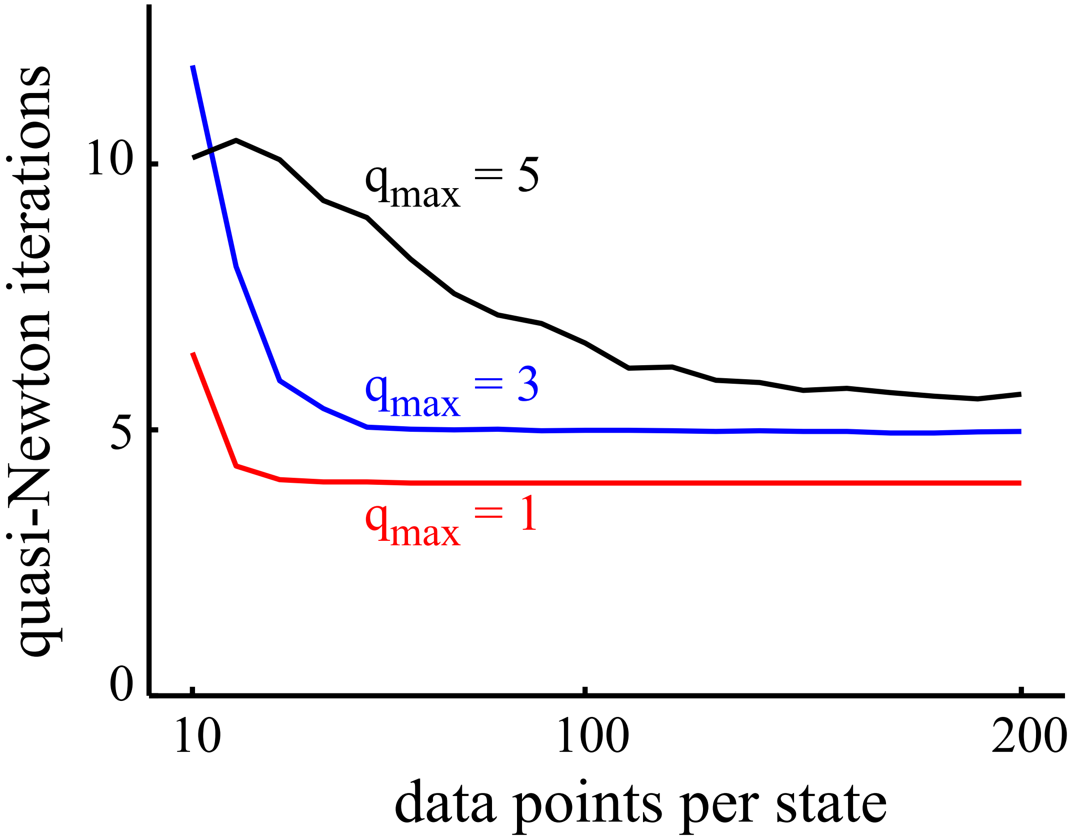
$$\mathbf{v} \leftarrow \mathbf{v} - \mathsf{grad}\ ./diag(H)$$

Example: randomly-generated problems









Papers available online

E. Todorov (2009) A new mathematical framework for optimal choice of actions. Submitted to *Proceedings of the National Academic of Science*

E. Todorov (2009) Compositionality of optimal control laws. Submitted to *Robotics Science and Systems*

E. Todorov (2009) Classic maximum principles and estimation control dualities for nonlinear stochastic systems. Submitted to *IEEE Transactions on Automatic Control*

E. Todorov (2009) Efficient algorithms for inverse optimal control. Submitted to *International Conference on Machine Learning*

E. Todorov (2008) Eigenfunction approximation methods for linearly-solvable optimal control problems. To appear in *IEEE Adaptive Dynamic Programming and Reinforcement Learning*

E. Todorov (2008) Parallels between sensory and motor information processing. To appear in *The Cognitive Neurosciences IV*, M. Gazzaniga (ed)

E. Todorov (2008) General duality between optimal control and estimation. *IEEE Conference on Decision and Control*

E. Todorov (2006) Linearly-solvable Markov decision problems. *Advances in Neural Information Processing Systems*