CHAPTER 6

LINEARLY SOLVABLE OPTIMAL CONTROL

K. Dvijotham[1] and E. Todorov[2]

[1]Computer Science & Engineering, University of Washington, Seattle
[2] Computer Science & Engineering and Applied Mathematics, University of Washington, Seattle

6.1 ABSTRACT

We summarize the recently-developed framework of linearly-solvable stochastic optimal control. Using an exponential transformation, the (Hamilton-Jacobi) Bellman equation for such problems can be made linear, giving rise to efficient numerical methods. Extensions to game theory are also possible and lead to linear Isaacs equations. The key restriction that makes a stochastic optimal control problem linearly-solvable is that the noise and the controls must act in the same subspace. Apart from being linearly solvable, problems in this class have a number of unique properties including: path-integral interpretation of the exponentiated value function; compositionality of optimal control laws; duality with Bayesian inference; trajectory-based Maximum Principle for stochastic control. Development of a general class of more easily solvable problems tends to accelerate progress – as linear systems theory has done. The new framework may have similar impact in fields where stochastic optimal control is relevant.

6.2 INTRODUCTION

Optimal control is of interest in many fields of science and engineering [4, 21], and is arguably at the core of robust-yet-efficient animal behavior [23, 26]. Apart from the fact that "optimal" tends to be good even when it is not exactly optimal, this approach to control engineering is appealing because one can in principle define a high-level cost function specifying the task goal, and leave the hard work of synthesizing a controller to numerical optimization software. This leads to better automation, especially when compared to the manual designs often used in engineering practice. Yet optimizing controllers for real-world tasks is very challenging even numerically, and the present book explores the state-of-the-art approaches to overcoming this challenge.

One of the most productive lines of attack when it comes to solving hard problems is to identify restricted problem formulations that can be solved efficiently, and use these restricted formulations to approximate (perhaps iteratively) the harder problem. An example is the field of numerical optimization, where the only multivariate function we know how to optimize analytically is the quadratic – and so we model every other function as being locally quadratic. This is the key idea behind all second-order methods. The situation is similar in optimal control and control theory in general, where the only systems we truly understand are linear – and so we often approximate many other systems as being linear, either locally or globally. An example of an optimal control method relying on iterative linearizations of the dynamics (and quadratizations of the cost) is the iterative LQG method [34].

This general approach to solving hard problems relies on having restricted problem formulations that are computationally tractable. For too long, linear systems theory has remained pretty much the only item on the menu. Recently, we and others have developed a restricted class of stochastic optimal control problems that are linearly-solvable [14, 27]. The dynamics in such problems can be non-linear (and even non-smooth), the costs can be non-quadratic, and the noise can be non-Gaussian. Yet the problem reduces to solving a linear equation – which is a minimized and exponentially-transformed Bellman equation. To be sure, this is not nearly as tractable as an LQG problem, because the linear equation is question is a functional equation characterizing a scalar function (the exponent of the value function) over a high-dimensional continuous state space. Nevertheless solving such problems is much easier computationally than solving generic optimal control problems.

The key restriction that makes a stochastic optimal control problem linearly-solvable is that the noise and the controls are interchangeable, i.e. anything that the control law can accomplish could also happen by chance (however small the probability may be) and vice versa. The control cost associated with a given outcome is inversely related to the probability of the same outcome under the passive/uncontrolled dynamics. The form of this control cost is fixed, while the state cost can be arbitrary.

Apart from being linearly-solvable, problems in this class have unique properties that enable specialized numerical algorithms. These can be summarized as follows:

- The solution can be expressed as an expectation/path-integral, which enables sampling approximations. This yields a model-free reinforcement learning method which only estimates the value function, as opposed to the much larger Q-function estimated in Q-learning;
- The most likely trajectory under the optimally-controlled stochastic dynamics coincides with the optimal trajectory in a related deterministic problem, giving rise to the first trajectory-based Maximum Principle for stochastic control;
- The state density under the optimal controls coincides with the Bayesian posterior in a related inference problem, giving rise to a general duality between Bayesian inference and stochastic optimal control;
- The optimal solutions to first-exit and finite-horizon problems with identical dynamics and running cost, but different final costs, can be used as control primitives: they can be combined analytically so as to yield provably-optimal solutions to new problems;
- Bellman residual minimization reduces to a linear algebraic equation;
- Natural policy gradient for linearly-parameterized policies is possible by estimating only the value function, as opposed to the Q-function;
- Inverse optimal control, i.e. the problem of inferring the cost from state space trajectories of the optimally controlled system, reduces to an unconstrained convex optimization problem and does not require solving the forward problem;
- Extensions to risk-sensitive and game theoretic control yield linear Isaacs equations.

6.2.1 Notation

Before we proceed, we summarize notational conventions that will be used throughout this chapter. Let S be a set, $\mathcal{P}\left[S\right]$ the set of probability distributions over S, and $S^{\mathbf{R}^+}$ the set of positive real-valued functions on S. For any $p \in \mathcal{P}\left[S\right]$, let $H\left[p\right] = \mathrm{E}_p\left[-\log(p)\right]$ denote the entropy. If f is a real-valued function on S, the expectation of f under p is denoted $\mathrm{E}_p\left[f\right] = \sum_s p(s)f(s)$. Define the function

$$\Psi_p^{\alpha}\left[f\right] = \alpha^{-1}\log\left(\mathop{\mathrm{E}}_{p}\left[\exp(\alpha f)\right]\right), \Psi_p\left[f\right] = \Psi_p^{1}\left[f\right].$$

One can prove that in the limit $\alpha \to 0$ this is just the expectation, so we define $\Psi_\pi^0\left[f\right] = \mathrm{E}_\pi\left[f\right]$. Given two positive functions $p, q \in S^{\mathbf{R}^+}$, define the distribution

$$(p \otimes q)(s) = (q \otimes p)(s) = \frac{p(s)q(s)}{\sum_{s\in S} p(s)q(s)}.$$

We will use the shorthand notation Pol for policy, Dyn for dynamics, Co for cost and OP for optimal policy. In general, we will use boldface for vectors or discrete symbols, and italics for scalar valued functions.

6.2.2 Markov Decision Processes (MDPs)

Markov Decision Processes (MDPs) are a widely used framework for specifying and solving optimal control problems. MDPs are formally defined by specifying:

- A state space $\mathcal{X}$. We use $\mathbf{x}$ to denote states, $\mathbf{x} \in \mathcal{X}$. This could be continuous (subset of $\Re^n$), discrete (set of nodes in a graph) or a mixture of both.
- An action space $\mathcal{U}(\mathbf{x})$ for each state. Actions are denoted by $\mathbf{u}$. We denote policies by the same letter $\mathbf{u}(\mathbf{x}) \in \mathcal{U}(\mathbf{x})$.
- A stochastic dynamics $\mathbb{P}\left(\mathbf{x}, \mathbf{u}\right)$, which is the probability distribution over the next state given the current state $\mathbf{x}$ and action $\mathbf{u} \in \mathcal{U}(\mathbf{x})$.
- An immediate cost function $\ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u})$.

At any time t, an action $\mathbf{u}$ is chosen depending on the current state and the system transitions into a new state sampled from the stochastic dynamics. The objective of the control is to minimize the expected cost accumulated over time. The precise notion of accumulation can vary, giving rise to different problem formulations as follows. Finite Horizon (FH) problems are specified by a horizon T, a running cost $\ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u})$ and a terminal cost $\ell_{\mathrm{f}}(\mathbf{x}, \mathbf{u})$. First exit (FE) problems are specified by a set of terminal states $\mathcal{T}$, a running cost $\ell(\mathbf{x}, \mathbf{u})$ and a terminal cost $\ell_{\mathrm{f}} : \mathcal{T} \to \Re$. Infinite Horizon Average Cost (IH) problems are specified just by a running cost $\ell(\mathbf{x}, \mathbf{u})$, and Infinite Horizon Discounted Cost problems are specified by a running cost $\ell(\mathbf{x}, \mathbf{u})$ and a discount factor γ. Discounted cost problems are very popular in Reinforcement Learning [23], however we do not consider them here as they do not lead to linear Bellman equations. All other problem formulations lead to linear Bellman equations.

The optimal cost-to-go function (or optimal value function) $v_t\left(\mathbf{x}\right)$ is defined as the expected cumulative cost for starting at state $\mathbf{x}$ at time t and acting optimally thereafter. This function is characterized by the Bellman equation (BE):

$$v_t\left(\mathbf{x}\right) = \min_{\mathbf{u}} \ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u}) + \mathop{\mathrm{E}}_{\mathbb{P}(\mathbf{x}, \mathbf{u})} \left[v_{t+1}\right] \tag{6.1}$$

$$\mathbf{u}^*\left(\mathbf{x}; t\right) = \operatorname*{argmin}_{\mathbf{u}} \ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u}) + \mathop{\mathrm{E}}_{\mathbb{P}(\mathbf{x}, \mathbf{u})} \left[v_{t+1}\right]$$

$\mathbf{u}^*\left(\cdot; t\right)$ is called the *optimal policy*.

For most control problems of practical interest, solving the Bellman equation is computationally intractable. This is because one needs to store the value function at each state $\mathbf{x}$ and the number of states could be very large (infinite if $\mathcal{X}$ is a continuous domain). This has led to a variety of approximation schemes. Many of these rely on solving the BE approximately. However, getting such schemes to work often requires a lot of problem-specific tuning, and even then may not scale to genuinely hard problems. Part of the difficulty is the highly nonlinear nature of the BE which is a result of the $\min_{\mathbf{u}}$ term. A key advantage of linearly-solvable MDPs (see below) is that the minimization over actions can be done analytically given the value function. The minimized Bellman equation can then be made linear by exponentiating the value function.

6.3 LINEARLY SOLVABLE OPTIMAL CONTROL PROBLEMS

6.3.1 Probability shift: A an alternative view of control

Conventionally, we think of control signals as quantities that modify the system behavior in some pre-specified manner. In our framework it is more convenient to work with a somewhat different notion of control, which is nevertheless largely equivalent to the conventional notion, allowing us to model problems of practical interest. To motivate this alternative view, consider a control-affine diffusion:

$$d\mathbf{x} = (\mathsf{a}(\mathbf{x}) + \mathsf{B}(\mathbf{x})\,\mathbf{u})\,dt + \mathsf{C}(\mathbf{x})d\omega$$

This is a stochastic differential equation specifying the infinitesimal change in the state $\mathbf{x}$, caused by a passive/uncontrolled drift term $\mathsf{a}(\mathbf{x})$, a control input $\mathbf{u}$ scaled by a control gain $\mathsf{B}(\mathbf{x})$, and Brownian motion noise with amplitude $\mathsf{C}(\mathbf{x})$. Subject to this system dynamics, the controller seeks to minimize a cost function of the form

$$\ell(\mathbf{x}) + \frac{1}{2}\,\mathbf{u}^T\,\mathbf{u}$$

In terms of MDPs, the transition probability may be written as

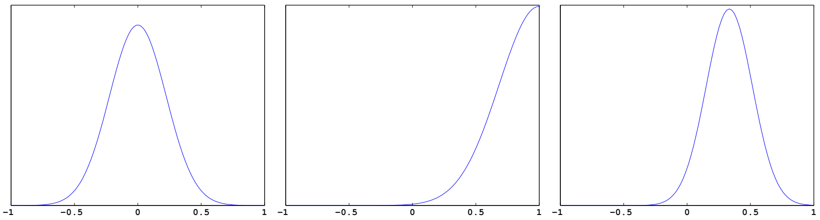
$$\mathbb{P}\,(\mathbf{x}, \mathbf{u}) = \mathcal{N}(\mathbf{x} + \delta(\mathsf{a}(\mathbf{x}) + \mathsf{B}(\mathbf{x})\,\mathbf{u}), \Sigma)$$

where we have discretized time using a time step δ. Thus, one way of thinking of the effect of control is that it changes the distribution of the next state from $\mathcal{N}(\mathbf{x} + \delta a(\mathbf{x}), \Sigma)$ to $\mathcal{N}(\mathbf{x} + \delta(\mathsf{a}(\mathbf{x}) + \mathsf{B}(\mathbf{x})\,\mathbf{u}), \Sigma)$. In other words, the controller shifts probability mass from one region of the state space to another. More generally, we can think of the system as having an uncontrolled dynamics which gives a distribution p over future states. The controller acts by modifying this distribution by probability shift to get a new distribution: $u \otimes p = \frac{pu}{\mathrm{E}_p[u]}$. This causes the probability mass in p to shift towards areas where u is large (figure 6.3.1). The controllers in our framework will act on the system dynamics by performing such probability shifts. The control signals will be positive scalar functions over the state space, rather than vectors or discrete symbols.

6.3.2 Linearly-solvable Markov Decision Processes (LMDPs)

Here we introduce the framework of linearly-solvable optimal control in discrete time. Such problems, called LMDPs, can be viewed in two mathematically equivalent ways. We shall describe both, since they both offer useful perspectives and illustrate the relationship to traditional MDPs in complementary ways.

In traditional MDPs the controller chooses a control signal or action $\mathbf{u}$ which determines the distribution of the next state $\mathbf{x}$' $\sim \mathbb{P}\,(\mathbf{x}, \mathbf{u})$. In LMDPs, we assume that there is an uncontrolled or passive dynamics $\Pi^0(\mathbf{x})$ for each state $\mathbf{x}$ that gives the distribution of the next state. The controller can change this distribution by picking a probability shift $u \in \mathcal{X}^{\mathbf{R}^+}$. This causes the distribution of the next state to change:



(a) p (b) u (c) $u \otimes p$

Figure 6.1 Probability Shift

$\mathbf{x}$' $\sim u \otimes \Pi^0(\mathbf{x})$. However, the controller must pay a price for doing so, given by the KL divergence between the controlled distribution $u \otimes \Pi^0(\mathbf{x})$ and the uncontrolled distribution $\Pi^0(\mathbf{x})$, which is a measure of the amount of change in the dynamics due to the controller. The Bellman equation for LMDPs is nonlinear in terms of the value function, but using an exponential transformation $z_t = \exp(-v_t)$ yields a linear equation in z. We call this the desirability function, since it is inversely related to the cost-to-go. The desirability function also gives the optimal shift policy $u^*(\mathbf{x};t) = z_{t+1}$, so the optimal controller is always trying to shift the uncontrolled dynamics towards more desirable states. The key results and their analogs for traditional MDPs are summarized in the following table:

	MDPs	LMDPs
Pol	$\mathbf{u} : \mathcal{X} \to \mathcal{U}$	$u : \mathcal{X} \to \mathcal{X}^{\mathbf{R}^+}$
Dyn	$\mathbf{x} \xrightarrow{\mathbf{u}} \mathbf{x}' \sim \mathbb{P}(\mathbf{x}, \mathbf{u}(\mathbf{x}))$	$\mathbf{x} \xrightarrow{u} \mathbf{x}' \sim u(\mathbf{x}) \otimes \Pi^0(\mathbf{x})$
Co	$\ell_t(\mathbf{x}, \mathbf{u}(\mathbf{x}))$	$\ell_t(\mathbf{x})+$ $\mathrm{KL}\left(u(\mathbf{x}) \otimes \Pi^0(\mathbf{x}) \parallel \Pi^0(\mathbf{x})\right)$
BE	$v_t(\mathbf{x}) = \min_{\mathbf{u}} \ell_t(\mathbf{x}, \mathbf{u}) + \mathop{\mathrm{E}}_{\mathbb{P}(\mathbf{x}, \mathbf{u}(\mathbf{x}))}[v_{t+1}]$	$z_t(\mathbf{x}) = \exp(-\ell_t(\mathbf{x})) \mathop{\mathrm{E}}_{\Pi^0(\mathbf{x})}[z_{t+1}]$
OP	$\mathbf{u}^*(\mathbf{x};t) =$ $\operatorname*{argmin}_{\mathbf{u}} \ell_t(\mathbf{x}, \mathbf{u}) + \mathop{\mathrm{E}}_{\mathbb{P}(\mathbf{x}, \mathbf{u}(\mathbf{x}))}[v_{t+1}]$	$u^*(\mathbf{x};t) = z_{t+1}$

6.3.3 An alternate view of LMDPs

In the alternate view, LMDPs are almost the same as traditional MDPs with deterministic dynamics and stochastic policies, except for two differences: we impose an additional cost that encourages policies with high entropy, and we compute the cost based not on the action that happened to be sampled from the stochastic policy, but by taking an expectation over all actions that could have been sampled. In this view, the relation between traditional deterministic MDPs and LMDPs is summarized as:

	Deterministic MDPs with Stochastic Policies	LMDPs
Pol	$u : \mathcal{X} \to \mathcal{P}\left[\mathcal{U}\right]$	$u : \mathcal{X} \to \mathcal{P}\left[\mathcal{U}\right]$
Dyn	$\mathbf{u} \sim u(\mathbf{x})$ $\mathbf{x}\text{'} = \mathrm{f}(\mathbf{x}, \mathbf{u})$	$\mathbf{u} \sim u(\mathbf{x})$ $\mathbf{x}\text{'} = \mathrm{f}(\mathbf{x}, \mathbf{u})$
Co	$\ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u})$	$\mathop{\mathrm{E}}_{\mathbf{u} \sim u(\mathbf{x})} \left[\ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u})\right] - H(u(\mathbf{x}))$
BE	$v_t\left(\mathbf{x}\right) =$ $\min_{u(\mathbf{x})} \mathop{\mathrm{E}}_{u(\mathbf{x})} \left[\ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u}) + v_{t+1}\left(\mathrm{f}(\mathbf{x}, \mathbf{u})\right)\right]$	$z_t\left(\mathbf{x}\right) =$ $\sum_{\mathbf{u}} \exp\left(-\ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u})\right) z_{t+1}\left(\mathrm{f}(\mathbf{x}, \mathbf{u})\right)$
OP	$u^*\left(\mathbf{x}; t\right) = \delta(\mathbf{u}^*)$ $\mathbf{u}^* =$ $\mathop{\mathsf{argmin}}_{\mathbf{u}} \ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u}) + v_{t+1}\left(\mathrm{f}(\mathbf{x}, \mathbf{u})\right)$	$u^*\left(\mathbf{x}; t\right) = z_{t+1}$

We can rewrite the BE for LMDPs in this interpretation as:

$$v_t\left(\mathbf{x}\right) = -\log\left(\sum_{\mathbf{u}} \exp\left(-\ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u}) - v_{t+1}\left(\mathrm{f}(\mathbf{x}, \mathbf{u})\right)\right)\right)$$

The relationships between MDPs and LMDPs is now clear: the hard minimum in the Bellman equation for MDPs is replaced by a soft minimum for LMDPs, namely $-\log(\sum(\exp(-\ldots)))$. If we replace the cost $\ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u})$ by a scaled version $\gamma\,\ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u})$, as γ increases we move closer and closer to the hard minimum, and in the limit $\gamma \to \infty$ we recover the Bellman equation for MDPs. Thus any deterministic MDP can be obtained as a limit of LMDPs.

The relationship between the two interpretations can be understood as follows. Define a passive dynamics with support only on the states immediately reachable from $\mathbf{x}$ under some action $\mathbf{u}$:

$$\Pi^0(\mathrm{f}(\mathbf{x}, \mathbf{u}) |\, \mathbf{x}) \propto \exp\left(-\ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u})\right)$$

For states not immediately reachable from $\mathbf{x}$, the probability under the passive dynamics is 0. Given any control (probability shift) $u \in \mathcal{X}^{\mathbf{R}^+}$, we have:

$$\begin{aligned}\mathrm{KL}\left(u \otimes \Pi^0(\mathbf{x}) \parallel \Pi^0(\mathbf{x})\right) &= -H\left[u \otimes \Pi^0(\mathbf{x})\right] + \mathop{\mathrm{E}}_{u \otimes \Pi^0(\mathbf{x})}\left[-\log\left(\Pi^0(\mathbf{x})\right)\right] \\ &= -H\left[u \otimes \Pi^0(\mathbf{x})\right] + \mathop{\mathrm{E}}_{u \otimes \Pi^0(\mathbf{x})}\left[\ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u})\right] - \ell_{\mathrm{t}}(\mathbf{x})\end{aligned}$$

where $\ell_{\mathrm{t}}(\mathbf{x}) = -\log\left(\sum_{\mathbf{u}} \exp\left(-\ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u})\right)\right)$. Thus, the alternate interpretation is equivalent to the original interpretation with passive dynamics proportional to $\exp\left(-\ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u})\right)$ and cost function $-\log\left(\sum_{\mathbf{u}} \exp\left(-\ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u})\right)\right)$.

6.3.4 Other Problem Formulations

Thus far we focused on the FH problem formulation. We can obtain linearly-solvable problems with other problem formulations as well. The corresponding BEs are

$$\begin{aligned}
\text{FE} \quad & z\left(\mathbf{x}\right) = \exp\left(-\,\ell(\mathbf{x})\right) \underset{\Pi^0(\mathbf{x})}{\mathrm{E}}\left[z\right] \text{ if } \mathbf{x} \notin \mathcal{T} \\
& z\left(\mathbf{x}\right) = \exp\left(-\,\ell_{\mathrm{f}}(\mathbf{x})\right) \text{ if } \mathbf{x} \in \mathcal{T}
\end{aligned}$$

$$\text{IH} \quad z\left(\mathbf{x}\right) = \exp\left(c - \ell(\mathbf{x})\right) \underset{\Pi^0(\mathbf{x})}{\mathrm{E}}\left[z\right],\ c \text{ is the Optimal Average Cost}$$

In the IH case the linear BE becomes an eigenvalue problem, with eigenvalue $\exp(-c)$ where c is the average cost. It can be shown that the solution to the optimal control problem corresponds to the principal eigenpair.

6.3.5 Applications

We now give some examples of how commonly occurring control problems can be modeled as LMDPs.

Shortest paths: Consider the shortest path problem defined on a graph. We can view this as an MDP with nodes corresponding to states and edges corresponding to actions. A stochastic version of this problem is one where the action does not take you directly where you intend, but possibly to the end of one of the other outgoing edges from that node. We can define an LMDP with passive dynamics at a node to be the uniform distribution over all nodes reachable in one step. The cost is a constant cost per unit time and the problem is a FE problem with the goal state as the state to which the shortest path is being computed. By scaling up the constant cost by ρ, in the limit as $\rho \to \infty$ we recover the traditional deterministic shortest paths problem. This yields an efficient approximation algorithm for the shortest paths problem, by solving an LMDPs with sufficiently large ρ, see [30].

Discretizing continuous problems: We can construct efficient solutions to problems with continuous state spaces and continuous time, provided the state space can be discretized to a reasonable size (LMDPs can easily handle problems with millions of discrete states). We consider a simple problem that has been a standard benchmark in the Reinforcement Learning literature, the mountain-car problem. In this problem, the task is to get a car to drive down from a hill into a valley and park on another hill on the other side of the valley. The control variable is the acceleration of the car, and the state consists of the position and velocity of the car. We impose limits on all these quantities and discretize the state space to within those limits. The dynamics is completely determined by gravity and the shape of the hill. We plot results in figure 6.2 comparing the LMDP discretization and a iterative solution of the LMDP to a standard MDP discretization and using policy/value iteration to solve that. It can be seen that the LMDP solution converges faster to the optimal policy. See [30].

6.3.6 Linearly-solvable controlled diffusions (LDs)

Although the focus of this chapter is on discrete-time problems (i.e. LMDPs), here we summarize related results in continuous time. The linearly-solvable optimal control problems in continuous time are control-affine diffusions with dynamics

$$\mathrm{d}\,\mathbf{x} = \mathsf{a}(\mathbf{x})\,\mathrm{d}\,t + \mathsf{B}(\mathbf{x})\,\mathbf{u}\,\mathrm{d}\,t + \sigma\,\mathsf{B}(\mathbf{x})\,\mathrm{d}\,\omega$$

and cost rate

$$\ell_{\mathrm{t}}(\mathbf{x}) + \frac{1}{2\sigma^2}\|\mathbf{u}\|^2$$

The unusual aspects of this problem are that: (i) the noise and the control act in the same subspace spanned by the columns of $\mathsf{B}(\mathbf{x})$; (ii) the control cost is scaled by σ^{-2}, thus increasing the noise in the dynamics makes the controls cheaper.

For problems in this class one can show that the optimal control law is

$$\mathbf{u}^*\left(\mathbf{x};t\right) = \frac{\sigma^2}{z_t\left(\mathbf{x}\right)}\mathsf{B}(\mathbf{x})^T\frac{\partial z_t\left(\mathbf{x}\right)}{\partial \mathbf{x}}$$

and the Hamilton-Jacobi-Bellman (HJB) equation expressed in terms of z becomes linear and is given by

$$\frac{\partial z_t\left(\mathbf{x}\right)}{\partial t} = \ell_{\mathrm{t}}(\mathbf{x})\,z_t\left(\mathbf{x}\right) - \mathcal{L}\left[z_t\right]\left(\mathbf{x}\right) \tag{6.2}$$

Here $\mathcal{L}$ is a 2nd-order linear differential operator known as the generator of the passive dynamics:

$$\mathcal{L}\left[f\right]\left(\mathbf{x}\right) = a(\mathbf{x})^T\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} + \frac{\sigma^2}{2}\mathrm{tr}\left(\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}^T}\,\mathsf{B}(\mathbf{x})\mathsf{B}(\mathbf{x})^T\right) \tag{6.3}$$

This operator computes expected directional derivatives of functions along trajectories of the passive dynamics. We call problems of this kind linearly solvable controlled diffusions (LDs).

6.3.7 Relationship between discrete and continuous-time problems

If we take the first view of LMDPs that uses the notion of a stochastic passive dynamics, we can interpret the above linearly solvable diffusion as a continuous-time limit of LMDPs. This can be done by discretizing the time axis of the diffusion process with time step h using the Euler approximation:

$$\mathbf{x}(t+h) = \mathbf{x}(t) + h\,\mathsf{a}(\mathbf{x}) + h\,\mathsf{B}(\mathbf{x})\,\mathbf{u} + \epsilon$$

where $\epsilon \sim \mathcal{N}\left(0, h\sigma^2\,\mathsf{B}(\mathbf{x})\mathsf{B}(\mathbf{x})^T\right)$. The covariance is scaled by h since for Brownian noise the standard deviation grows as the square root of time. The discrete-time cost becomes $h\,\ell_{\mathrm{t}}(\mathbf{x}) + h\frac{1}{2\sigma^2}\mathbf{u}^T\,\mathbf{u}$. We will now construct an LMDP that resembles

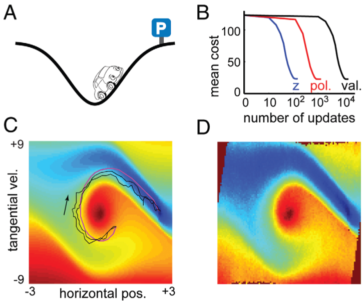


Figure 6.2 Continuous problems. Comparison of our MDP approximation and a traditional MDP approximation on a continuous car-on-a-hill problem. (A) Terrain, (B) Z iteration (ZI) (blue), policy iteration (PI) (red), and value iteration (VI) (black) converge to control laws with identical performance; ZI is 10 times faster than PI and 100 times faster than VI. Horizontal axis is on log-scale. (C) Optimal cost-to-go for our approximation. Blue is small, red is large. The two black curves are stochastic trajectories resulting from the optimal control law. The thick magenta curve is the most likely trajectory of the optimally controlled stochastic system. (D) The optimal cost-to-go is inferred from observed state transitions by using our algorithm for inverse optimal control. Figure taken from [30].

this time-discretized LD. To do this, we define the passive dynamics at state $\mathbf{x}$ to be the Euler approximation of the distribution of $\mathbf{x}(t+h)$ given $\mathbf{x}(t) = \mathbf{x}$:

$$\Pi^0(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} + h\,\mathsf{a}(\mathbf{x}), h\sigma^2\,\mathsf{B}(\mathbf{x})\mathsf{B}(\mathbf{x})^T\right).$$

This converges to the continuous time LD dynamics with $\mathbf{u} = 0$ as $h \to 0$. Now, consider a family of probability shifts $u^{\mathbf{u}}$ parameterized by $\mathbf{u}$ such that

$$u^{\mathbf{u}} \otimes \Pi^0(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} + h\,\mathsf{a}(\mathbf{x}) + h\,\mathsf{B}(\mathbf{x})\,\mathbf{u}, h\sigma^2\,\mathsf{B}(\mathbf{x})\mathsf{B}(\mathbf{x})^T\right).$$

This distribution is the Euler discretization of the LD dynamics under control $\mathbf{u}$. It can be shown that $\mathrm{KL}\left(u^{\mathbf{u}} \otimes \Pi^0(\mathbf{x}) \parallel \Pi^0(\mathbf{x})\right) = h\frac{1}{2\sigma^2}\mathbf{u}^T\,\mathbf{u}$. Thus, for every $\mathbf{u}$, there is a probability shift $u^{\mathbf{u}}$ that matches the Euler approximation of the LD dynamics under control $\mathbf{u}$ and also matches the time-discretized control cost. We define the state cost to be $h\,\ell_t(\mathbf{x})$. This LMDP is very close to the MDP corresponding to the time discretized LD, the only difference being that we allow probability shifts that are not equal to $u^{\mathbf{u}}$ for any $\mathbf{u}$. However, it turns out that this extra freedom does not change the optimal control law, at least in the limit $h \to 0$. The BE corresponding to this LMDP is:

$$z_t(\mathbf{x}) = \exp\left(-h\,\ell_t(\mathbf{x})\right) \underset{\mathcal{N}\left(\mathbf{x} + h\,\mathsf{a}(\mathbf{x}), h\sigma^2\,\mathsf{B}(\mathbf{x})\mathsf{B}(\mathbf{x})^T\right)}{\mathrm{E}}\left[z_{t+h}\right]$$

It can be shown that after some algebra and taking the limit $h \to 0$, we recover the linear HJB equation (6.2).

6.3.8 Historical perspective

Linearly-solvable optimal control is a rich mathematical framework that has recently received a lot of attention, following Kappen's work on control-affine diffusions in continuous time [14], and our work on Markov decision processes in discrete time [27]. Both groups have since then obtained many additional results: see [36, 17, 6, 5] and [28, 31, 30, 29, 8, 32, 33, 9, 38, 39] respectively. Other groups have also started to use and further develop this framework [35, 7, 24, 25].

The initial studies [14, 27] were done independently, yet they both built upon the same earlier results which we discuss here. For over 30 years these earlier results had remained a curious mathematical fact, that was never actually used to solve control problems – which, unfortunately, is not uncommon in control theory.

In continuous time, the trick that makes the HJB equation linear is

$$v_{xx} - v_x v_x^T = -\frac{z_{xx}}{z}, \text{ where } z = \exp(-v)$$

Applying this exponential (or logarithmic) transformation to 2nd-order PDEs has a long history in Physics [12, 11]. Its first application to control was due to Fleming and Mitter, who showed that non-linear filtering corresponds to a stochastic optimal control problem whose HJB equation can be made linear [10]. Kappen generalized this idea, and noted that the solution to the resulting linear PDE is also a path integral – which yields sampling approximations to the optimal value function [14].

Our work [27] was motivated by the same earlier results but in a more abstract way: we asked, are there classes of linearly-solvable optimal control problems involving arbitrary dynamics? This led to the LMDP framework summarized here. In discrete time, the trick that makes the Bellman equation linear is

$$\min_q \{\mathrm{KL}(q \parallel p) + \mathop{\mathrm{E}}_q [v]\} = -\log \mathop{\mathrm{E}}_p [\exp(-v)]$$

where the minimum is achieved at $q^* = \exp(-v) \otimes p$. We introduced this trick in [27], although it turned out to have been used earlier to derive a variational characterization of the Bayesian posterior [18]. Indeed if p is a prior and v is a negative log-likelihood, then the above q^* is a Bayesian posterior.

6.4 EXTENSION TO RISK-SENSITIVE CONTROL AND GAME THEORY

6.4.1 Game Theoretic Control : Competitive Games

Here we briefly introduce the notion of game theoretic control or robust control [3]. In this setting, the system can be influenced by another agent (adversary) in addition to the controller. The controller needs to design a strategy that achieves the control objective in spite of the adversarial disturbances. We shall focus on the simplest case of two-player zero-sum dynamic games, where the adversary is trying to maximize the same cost that the controller is trying to minimize. The game proceeds as follows: 1) The adversary and controller pick actions $\mathbf{u}_{\mathrm{a}}, \mathbf{u}_{\mathrm{c}}$ respectively. 2) The controller pays cost $\ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u}_{\mathrm{c}}, \mathbf{u}_{\mathrm{a}})$ and adversary pays $-\ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u}_{\mathrm{c}}, \mathbf{u}_{\mathrm{a}})$. 3) The system transitions to state $\mathbf{x}\text{'} \sim \mathbb{P}\left(\mathbf{x}\text{'}|\mathbf{x}, \mathbf{u}_{\mathrm{c}}, \mathbf{u}_{\mathrm{a}}\right)$. The solution to such a game can be formulated using the Bellman-Isaacs equations:

$$v_t\left(\mathbf{x}\right) = \max_{\mathbf{u}_{\mathrm{a}} \in \mathcal{U}_{\mathrm{a}}(\mathbf{x}, u_{\mathrm{c}})} \min_{\mathbf{u}_{\mathrm{c}} \in \mathcal{U}(\mathbf{x})} \ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u}_{\mathrm{c}}, \mathbf{u}_{\mathrm{a}}) + \mathop{\mathrm{E}}_{\mathbb{P}(\mathbf{x}, \mathbf{u}_{\mathrm{c}}, \mathbf{u}_{\mathrm{a}})} [v_{t+1}]$$

We call such problems Markov Games or MGs. If the $\min, \max$ can be interchanged without changing the optimal policies for either the controller or the adversary, we say that the game has a *saddle-point equilibrium*. If not, then it matters which player plays first and we have corresponding *upper* and *lower* value functions.

We have recently discovered a class of linearly-solvable Markov games (LMGs), where the Bellman-Isaacs equation can be made linear as explained below. But first, we need to introduce a class of divergence measures between probability distributions that will play a key role in LMGs.

6.4.2 Rényi divergence

Rényi divergences are a generalization of the KL divergence. For distributions $p, q \in \mathcal{P}\left[\mathcal{X}\right]$, the Rényi divergence of order α is defined as

$$\mathbb{D}_{\alpha}\left(p \parallel q\right) = \frac{\mathrm{sign}(\alpha)}{\alpha - 1} \log\left(\mathop{\mathrm{E}}_{p}\left[\left(\frac{q}{p}\right)^{1-\alpha}\right]\right)$$

For any fixed p, q, it is known that $\mathbb{D}_{\alpha}$ is always non-negative, decreasing for $\alpha < 0$, and increasing for $\alpha > 0$. It is also known that $\lim_{\alpha \to 1} \mathbb{D}_{\alpha}\left(p \parallel q\right) = \mathrm{KL}\left(p \parallel q\right)$.

6.4.3 Linearly Solvable Markov Games (LMGs)

An LMG proceeds as follows:

The system in state $\mathbf{x}$ at time t.

The adversary picks controls $u_{\mathrm{a}} \in \mathcal{X}^{\mathbf{R}^{+}}$.

The controller picks controls $u_{\mathrm{c}} \in \mathcal{X}^{\mathbf{R}^{+}}$.

The system transitions into a state $\mathbf{x}\text{'} \sim u_{\mathrm{c}} \otimes u_{\mathrm{a}} \otimes \Pi^{0}(\mathbf{x})$

The cost function is

$$\begin{aligned} \ell_{\mathrm{t}}(\mathbf{x}, u_{\mathrm{c}}, u_{\mathrm{a}}) \quad &= \ell_{\mathrm{t}}(\mathbf{x}) \\ &+ \mathrm{KL}\left(u_{\mathrm{c}} \otimes u_{\mathrm{a}} \otimes \Pi^{0}(\mathbf{x}) \parallel u_{\mathrm{a}} \otimes \Pi^{0}(\mathbf{x})\right) \text{ (Control Cost)} \\ &- \mathbb{D}_{\frac{1}{\alpha}}\left(\Pi^{0}(\mathbf{x}) \parallel u_{\mathrm{a}} \otimes \Pi^{0}(\mathbf{x})\right) \text{ (Control Cost for Adversary)} \end{aligned}$$

We focus on competitive games and require that $\alpha > 0, \alpha \neq 1$. Also, the dynamics of the game is such that the adversary plays first, so the controller has a chance to respond to the adversarial disturbance. Thus, it is a maximin problem where we work with the lower value function. Later, we describe the case $\alpha < 0$ which leads to cooperative games.

The differences between standard MGs and LMGs can be summarized as follows:

	MGs	LMGs
Pol	$\mathbf{u}_{\mathrm{c}} : \mathcal{X} \times \mathcal{U}_{\mathrm{a}} \to \mathcal{U}$ $\mathbf{u}_{\mathrm{a}} : \mathcal{X} \to \mathcal{U}_{\mathrm{a}}$	$u_{\mathrm{c}} : \mathcal{X} \times \mathcal{X}^{\mathbf{R}^+} \to \mathcal{X}^{\mathbf{R}^+}$ $u_{\mathrm{a}} : \mathcal{X} \to \mathcal{X}^{\mathbf{R}^+}$
Dyn	$\mathbf{u}_{\mathrm{a}} = \mathbf{u}_{\mathrm{a}}(\mathbf{x}), \mathbf{u}_{\mathrm{c}} = \mathbf{u}_{\mathrm{c}}(\mathbf{x}, \mathbf{u}_{\mathrm{a}})$ $\mathbf{x} \xrightarrow{u_{\mathrm{c}}, u_{\mathrm{a}}} \mathbf{x}' \sim \mathbb{P}\left(\mathbf{x}' \mid \mathbf{x}, \mathbf{u}_{\mathrm{c}}, \mathbf{u}_{\mathrm{a}}\right)$	$u_{\mathrm{a}} = u_{\mathrm{a}}(\mathbf{x}), u_{\mathrm{c}} = u_{\mathrm{c}}(\mathbf{x}, u_{\mathrm{a}})$ $\mathbf{x} \xrightarrow{u_{\mathrm{c}}, u_{\mathrm{a}}} \mathbf{x}' \sim u_{\mathrm{c}} \otimes u_{\mathrm{a}} \otimes \Pi^0(\mathbf{x})$
Co	$\ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u}_{\mathrm{c}}, \mathbf{u}_{\mathrm{a}})$	$\ell_{\mathrm{t}}(\mathbf{x}) - \mathbb{D}_{\frac{1}{\alpha}}\left(\Pi^0(\mathbf{x}) \Vert u_{\mathrm{a}} \otimes \Pi^0(\mathbf{x})\right)$ $+ \mathrm{KL}\left(u_{\mathrm{c}} \otimes u_{\mathrm{a}} \otimes \Pi^0(\mathbf{x}) \Vert u_{\mathrm{a}} \otimes \Pi^0(\mathbf{x})\right)$
BE	$v_t(\mathbf{x}) = \max_{\mathbf{u}_{\mathrm{a}}} \min_{\mathbf{u}_{\mathrm{c}}} \ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u}_{\mathrm{c}}, \mathbf{u}_{\mathrm{a}})$ $+ \mathrm{E}_{\mathbb{P}(\mathbf{x}, \mathbf{u}_{\mathrm{c}}, \mathbf{u}_{\mathrm{a}})}[v_{t+1}]$	$z_t(\mathbf{x}) = \mathcal{Q}_{\mathrm{t}}(\mathbf{x}) \underset{\Pi^0(\mathbf{x})}{\mathrm{E}}[z_{t+1}]$ $z_t(\mathbf{x}) = \exp\left((\alpha - 1) v_t(\mathbf{x})\right)$ $\mathcal{Q}_{\mathrm{t}}(\mathbf{x}) = \exp\left((\alpha - 1)\, \ell_{\mathrm{t}}(\mathbf{x})\right)$
OP	${\mathbf{u}_{\mathrm{c}}}^*(\mathbf{x}, \mathbf{u}_{\mathrm{a}}; t) = \underset{\mathbf{u}_{\mathrm{c}}}{\mathsf{argmin}}\, \ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u}_{\mathrm{c}}, \mathbf{u}_{\mathrm{a}})$ $+ \mathrm{E}_{\mathbb{P}(\mathbf{x}, \mathbf{u}_{\mathrm{c}}, \mathbf{u}_{\mathrm{a}})}[v_{t+1}]$	${u_{\mathrm{c}}}^*(\mathbf{x}, u_{\mathrm{a}}; t) = z_{t+1}^{\frac{1}{1-\alpha}}$

6.4.3.1 LMDPs as a special case of LMGs: As $\alpha \to 0$, we recover the LMDP Bellman equation. We can explain this by looking at the cost function. It is known that $\lim_{\alpha \to 0} \mathbb{D}_{1/\alpha}(p \parallel q) \to \log\left(\sup_{\mathbf{x}} p(\mathbf{x})/q(\mathbf{x})\right)$. For this cost, the optimal strategy for the adversary is to always leave the passive dynamics unchanged, that is ${u_{\mathrm{a}}}^*(\mathbf{x}; t) = 1$. Intuitively, this says that the control cost for the adversary is high enough and the optimal strategy for him is to do nothing. Thus the problem reduces to the LMDP setting.

6.4.3.2 Effect of α : As α increases, the relative control cost of the controller with respect to the adversary increases, so, effectively, the adversary becomes more powerful. This makes the controller more conservative (or risk-averse), since it is fighting a stronger adversary.

6.4.3.3 Cooperative LMGs: We have also derived a cooperative LMG where two agents collaborate to accomplish the same control task. The game proceeds similar to a competitive game, however now both agents pay the same cost and are trying to minimize it in collaboration. The cost function for cooperative LMGs (for both agents) is:

$$\ell_{\mathrm{t}}(\mathbf{x}) + \mathbb{D}_{1/\alpha}\left(u_{\mathrm{a}} \otimes \Pi^0(\mathbf{x}) \parallel \Pi^0(\mathbf{x})\right) + \mathrm{KL}\left(u_{\mathrm{c}} \otimes u_{\mathrm{a}} \otimes \Pi^0(\mathbf{x}) \parallel u_{\mathrm{a}} \otimes \Pi^0(\mathbf{x})\right)$$

where $\alpha < 0$. As $|\alpha|$ gets bigger, the control cost for the helper gets smaller and the helper contributes more towards accomplishing the control task while the controller contributes less. The resulting BE is similar to the competitive case:

$$z_t\left(\mathbf{x}\right) = \exp\left(\left(\alpha - 1\right) \ell_{\mathrm{t}}(\mathbf{x})\right) \mathop{\mathrm{E}}_{\Pi^0(\mathbf{x})}\left[z_{t+1}\right]$$

$$z_t\left(\mathbf{x}\right) = \exp\left(\left(\alpha - 1\right) v_t\left(\mathbf{x}\right)\right)$$

In this case, again we can recover LMDPs by taking $\alpha \to 0$ and making the control cost for the helper effectively large enough that he always chooses not to change the passive dynamics.

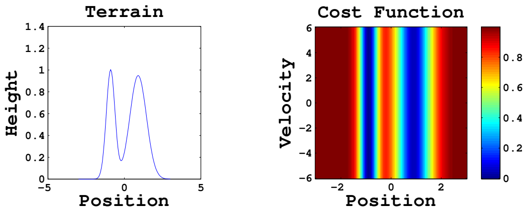


Figure 6.3 Terrain and Cost Function for LMG example

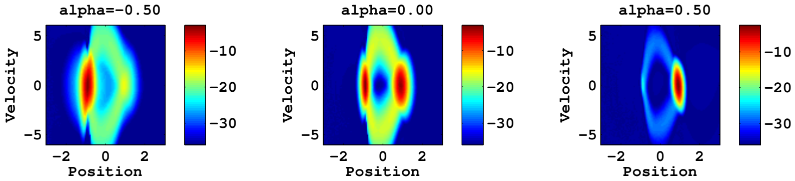


Figure 6.4 Logarithm of Stationary Distribution under Optimal Control vs α

6.4.3.4 Examples: We illustrate the effect of α with a simple control problem that requires one to drive up as high as possible on a hilly terrain. The cost function encourages one to drive up to the highest point, but the highest point is the peak of a steep hill, so that even a small perturbation from the adversary can push one downhill quickly. On the other hand, there is a shorter but less steep hill, where the adversary cannot have as much of an effect. The problem is formulated in the IH setting, so we are looking for a control strategy that achieves low average cost over a very long horizon. The terrain and cost function are plotted in figure 6.3. The stationary distributions over $\mathcal{X}$ under optimal control for different values of α are plotted in 6.4. It can be seen that when $\alpha < 0$ (cooperative case), the controller places more probability on the riskier but more rewarding option (steeper/higher hill) but when

$\alpha > 0$, the controller is more conservative and chooses the safer but less rewarding option (shorter/less steep hill). In the LMDP case, the solution splits its probability more or less evenly between the two options.

6.4.4 Linearly Solvable Differential Games (LDGs)

In this section we consider differential games (DGs) which are continuous-time versions of MGs. A differential game is described by a stochastic differential equation

$$\mathrm{d}\,\mathbf{x} = \big(\mathsf{a}(\mathbf{x}) + \mathsf{B}(\mathbf{x})\,\mathbf{u}_{\mathrm{c}} + \sqrt{\alpha}\,\mathsf{B}(\mathbf{x})\,\mathbf{u}_{\mathrm{a}}\big)\,\mathrm{d}\,t + \sigma\,\mathsf{B}(\mathbf{x})\,\mathrm{d}\,\omega$$

The infinitesimal generator $\mathcal{L}\left[\cdot\right]$ for the uncontrolled process $(\mathbf{u}_{\mathrm{c}}, \mathbf{u}_{\mathrm{a}} = 0)$ can be defined similarly to (6.3). We also define a cost rate

$$\ell_{\mathrm{t}}(\mathbf{x}, \mathbf{u}_{\mathrm{c}}, \mathbf{u}_{\mathrm{a}}) = \underbrace{\ell_{\mathrm{t}}(\mathbf{x})}_{\text{State Cost}} + \underbrace{\frac{1}{2\sigma^2}{\mathbf{u}_{\mathrm{c}}}^T\,\mathbf{u}_{\mathrm{c}}}_{\text{Control Cost for Controller}} - \underbrace{\frac{1}{2\sigma^2}{\mathbf{u}_{\mathrm{a}}}^T\,\mathbf{u}_{\mathrm{a}}}_{\text{Control Cost for Adversary}}$$

Like LMGs, these are two-player zero-sum games, where the controller is trying to minimize the cost function while the adversary tries to maximize the same cost. It can be shown that the optimal solution to differential games based on diffusion processes is characterized by a nonlinear PDE known as the Isaacs equation [3]. However, for the kinds of differential games we described here, the Isaacs equation expressed in terms of $z_t = \exp\left((\alpha - 1)v_t\right)$ becomes linear and is given by:

$$\begin{aligned}
\frac{\partial z_t\left(\mathbf{x}\right)}{\partial t} &= (1-\alpha)\,\ell_{\mathrm{t}}(\mathbf{x})\,z_t\left(\mathbf{x}\right) - \mathcal{L}\left[z_t\right]\left(\mathbf{x}\right) \\
{\mathbf{u}_{\mathrm{c}}}^*\left(\mathbf{x}; t\right) &= \frac{\sigma^2}{(\alpha-1)\,z_t\left(\mathbf{x}\right)}\mathsf{B}(\mathbf{x})^T\frac{\partial\, z_t\left(\mathbf{x}\right)}{\partial\,\mathbf{x}} \\
{\mathbf{u}_{\mathrm{a}}}^*\left(\mathbf{x}; t\right) &= \frac{-\sqrt{\alpha}\sigma^2}{(\alpha-1)\,z_t\left(\mathbf{x}\right)}\mathsf{B}(\mathbf{x})^T\frac{\partial\, z_t\left(\mathbf{x}\right)}{\partial\,\mathbf{x}}
\end{aligned}$$

When $\alpha = 0$, the adversarial control $\mathbf{u}_{\mathrm{a}}$ has no effect and we recover LDs. As α increases, the adversary's power increases and the control policy becomes more conservative.

There is a relationship between LDGs and LMGs. LDGs can be derived as the continuous time limit of LMGs that solve time-discretized versions of differential games. This relationship is analogous to the one between LMDPs and LDs.

6.4.4.1 Connection to Risk-Sensitive Control Both LMGs and LDGs can be interpreted in an alternate manner, as solving a sequential decision making problem with an alternate objective: Instead of minimizing expected total cost, we minimize the expectation of the exponential of the total cost:

$$\underset{\mathbf{x}_{t+1}\sim u_{\mathrm{c}}(\mathbf{x}_t)\,\otimes\,\Pi^0(\mathbf{x}_t)}{\mathrm{E}}\left[\exp\left(\sum_{t=0}^{T}\alpha\,\ell_{\mathrm{t}}(\mathbf{x}_t) + \mathbb{D}_{\alpha}\left(u_{\mathrm{c}}(\mathbf{x}_t)\otimes\Pi^0(\mathbf{x}_t)\parallel\Pi^0(\mathbf{x}_t)\right)\right)\right]$$

This kind of objective is used in risk-sensitive control [16] and it has been shown that this problem can also be solved using dynamic programming giving rise to a risk-sensitive Bellman equation. It turns out that for this objective, the Bellman equation is exactly the same as that of an LMG. The relationship between risk-sensitive control and game theoretic or robust control has been studied extensively in the literature [3], and it also shows up in the context of linearly solvable control problems.

6.4.5 Relationships among the different formulations

Linearly Solvable Markov Games (LMGs) are the most general class of linearly solvable control problems, to the best of our knowledge. As the adversarial cost increases ($\alpha \to 0$), we recover Linearly Solvable MDPs (LMDPs) as a special case of LMGs. When we view LMGs as arising from the time-discretization of Linearly Solvable Differential Games (LDGs), we recover LDGs as a continuous time limit ($\mathrm{d}\,t \to 0$). Linearly Solvable Controlled Diffusions(LDs) can be recovered either as the continuous time limit of an LMDP , or as the non-adversarial limit ($\alpha \to 0$) of LDGs. The overall relationships between the various classes of linearly solvable control problems is summarized in the figure below:

$$\begin{tikzcd}
LMGs \arrow[r, "\alpha\to 0"'] \arrow[d, "dt\to 0"] & LMDPs \arrow[d, "dt\to 0"] \\
LDGs \arrow[r, "\alpha\to 0"] & LDs
\end{tikzcd}$$

6.5 PROPERTIES AND ALGORITHMS

6.5.1 Sampling approximations and path-integral control

For LMDPs , it can be shown that the FH desirability function equals the expectation

$$z_0\left(\mathbf{x}_0\right) = \underset{\mathbf{x}_{t+1}\sim\Pi^0(\mathbf{x}_t)}{\mathrm{E}}\left[\exp\left(-\ell_\mathrm{f}\left(\mathbf{x}_T\right) - \sum\nolimits_{t=1}^{T-1} \ell_\mathrm{t}\left(\mathbf{x}_t\right)\right)\right]$$

over trajectories $\mathbf{x}_1 \cdots \mathbf{x}_T$ sampled from the passive dynamics starting at $\mathbf{x}_0$. This is also known as a path-integral. It was first used in the context of linearly-solvable controlled diffusions [14] to motivate sampling approximations. This is a model-free method for Reinforcement Learning [23], however unlike Q-learning (the classic model-free method) which learns a Q-function over the state-action space, here we only learn a function over the state space. This makes model-free learning in the LMDP setting much more efficient [30].

One could sample directly from the passive dynamics, however the passive dynamics are very different from the optimally-controlled dynamics that we are trying to learn. Faster convergence can be obtained using importance sampling:

$$z_0\left(\mathbf{x}_0\right) = \underset{\mathbf{x}_{t+1}\sim\Pi^1(\mathbf{x}_t)}{\mathrm{E}}\left[\exp\left(-\ell_\mathrm{f}\left(\mathbf{x}_T\right) - \sum\nolimits_{t=1}^{T-1} \ell_\mathrm{t}\left(\mathbf{x}_t\right)\right)\frac{p^0\left(\mathbf{x}_1\cdots\mathbf{x}_T \,|\, \mathbf{x}_0\right)}{p^1\left(\mathbf{x}_1\cdots\mathbf{x}_T \,|\, \mathbf{x}_0\right)}\right]$$

Here $\Pi^1\left(\mathbf{x}_{t+1} \mid \mathbf{x}_t\right)$ is a proposal distribution and p^0, p^1 denote the trajectory probabilities under Π^0, Π^1. The proposal distribution would ideally be Π^*, the optimally controlled distribution, but since we do not have access to it, we use the approximation based on our latest estimate of the function z . We have observed that importance sampling speeds up convergence substantially [30]. Note however that in order to evaluate the importance weights p^0/p^1, one needs a model of the passive dynamics.

6.5.2 Residual minimization via function approximation

A general class of methods for approximate dynamic programming is to represent the value function with a function approximator, and tune its parameters by minimizing the Bellman residual. In the LMDP setting such methods reduce to linear algebraic equations. Consider the function approximator

$$\widehat{z}\left(\mathbf{x}; w, \theta\right) = \sum\nolimits_i w_i f_i\left(\mathbf{x}; \theta\right) \tag{6.4}$$

where w are linear weights while θ are location and shape parameters of the bases f. The reason for separating the linear and non-linear parameters is that the former can be computed efficiently by linear solvers. Choose a set of "collocation" states $\{\mathbf{x}_n\}$ where the residual will be evaluated. Defining the matrices F and G with elements

$$\begin{aligned} F_{ni} &= f_i\left(\mathbf{x}_n\right) \\ G_{ni} &= \exp\left(-\ell\left(\mathbf{x}_n\right)\right) \underset{\Pi^0(\mathbf{x}_n)}{\mathrm{E}}\left[f_i\right] \end{aligned}$$

the linear Bellman equation (in the IH case) reduces to

$$\lambda F\left(\theta\right) w = G\left(\theta\right) w$$

One can either fix θ and only optimize λ, w using a linear solver, or alternatively implement an outer loop in which θ is also optimized – using a general-purpose method such as Newton's method or conjugate gradient descent. When the bases are localized (e.g. Gaussians), the matrices F, G are sparse and diagonally-dominant, which speeds up the computation [31]. This approach can be easily extended to the LMG case.

6.5.3 Natural policy gradient

The residual in the Bellman equation is not monotonically related to the performance of the corresponding control law. Thus many researchers have focused on policy gradient methods that optimize control performance directly [37, 22, 13]. The remarkable finding in this literature is that, if the policy is parameterized linearly and the Q-function for the current policy can be approximated, then the gradient of the average cost is easy to compute.

Within the LMDP framework, we have shown [32] that the same gradient can be computed by estimating only the value function. This yields a significant improvement in terms of computational efficiency. The result can be summarized as follows.

Let $g\left(\mathbf{x}\right)$ denote a vector of bases, and define the control law

$$u^{(s)}(\mathbf{x}) = \exp\left(-s^{\mathsf{T}}g(\mathbf{x})\right)$$

This coincides with the optimal control law when $s^{\mathsf{T}}g\left(\mathbf{x}\right)$ equals the optimal value function $v\left(\mathbf{x}\right)$. Now let $v^{(s)}\left(\mathbf{x}\right)$ denote the value function corresponding to control law $u^{(s)}$, and let $v\left(\mathbf{x}\right) = r^{\mathsf{T}}g\left(\mathbf{x}\right)$ be an approximation to $v\left(\mathbf{x}\right)$, obtained by sampling from the optimally controlled dynamics $u^{(s)} \otimes \Pi^0$ and following a procedure described in [32]. Then it can be shown that the natural gradient [2] of the average cost with respect to the Fisher information metric is simply $s - r$. Note that these results do not extend to the LMG case since the policy-specific Bellman equation is nonlinear in this case.

6.5.4 Compositionality of optimal control laws

One way to solve hard control problems is to use suitable primitives [20, 15]. The only previously known primitives that preserve optimality were Options [20], which provide temporal abstraction. However what makes optimal control hard is space rather than time, i.e. the curse of dimensionality. The LMDP framework for the first time provided a way to construct spatial primitives, and combine them into provably-optimal control laws [29, 7]. This result is specific to FE and FH formulations. Consider a set of LMDPs (indexed by k) which have the same dynamics and running cost, and differ only by their final costs ${\ell_{\mathrm{f}}}^{(k)}\left(\mathbf{x}\right)$. Let the corresponding desirability functions be $z^{(k)}\left(\mathbf{x}\right)$. These will serve as our primitives. Now define a new (composite) problem whose final cost can be represented as

$$\ell_{\mathrm{f}}\left(\mathbf{x}\right) = -\log\left(\sum\nolimits_k w_k \exp\left(-\,{\ell_{\mathrm{f}}}^{(k)}\left(\mathbf{x}\right)\right)\right)$$

for some constants w_k. Then the composite desirability function is

$$z\left(\mathbf{x}\right) = \sum\nolimits_k w_k\, z^{(k)}\left(\mathbf{x}\right)$$

and composite optimal control law is

$$u^*\left(\mathbf{x}\right) = \sum\nolimits_k w_k\, {u^*}^{(k)}\left(\mathbf{x}\right)$$

One application of these results is to use LQG primitives – which can be constructed very efficiently by solving Riccati equations. The composite problem has linear dynamics, Gaussian noise and quadratic cost rate, however the final cost no longer has to be quadratic. Instead it can be the log of any Gaussian mixture. This represents a substantial extension to the LQG framework. These results can also be applied in infinite-horizon problems where they are no longer guaranteed to yield optimal solutions, but nevertheless may yield good approximations in challenging tasks such as those studied in Computer Graphics [7]. These results extend to the LMG case as well, by simply defining the final cost as $\ell_{\mathrm{f}}(\mathbf{x}) = \frac{1}{\alpha-1}\log\left(\sum_k w_k \exp\left((\alpha-1){\ell_{\mathrm{f}}}^{(k)}(\mathbf{x})\right)\right)$.

6.5.5 Stochastic Maximum Principle

Pontryagin's Maximum Principle is one of the two pillars of optimal control theory (the other being dynamic programming and the Bellman equation). It applies to deterministic problems, and characterizes locally-optimal trajectories as solutions to an ODE. In stochastic problems it seemed impossible to characterize isolated trajectories, because noise makes every trajectory dependent on its neighbors. There exist results called stochastic maximum principles, however they are PDEs that characterize global solutions, and in our view are closer to the Bellman equation than the Maximum Principle.

The LMDP framework provided the first trajectory-based maximum principle for stochastic control. In particular, it can be shown that the probability of a trajectory $\mathbf{x}_1 \cdots \mathbf{x}_T$ starting from $\mathbf{x}_0$ under the optimal control law is

$$p^*\left(\mathbf{x}_1, \cdots \mathbf{x}_T \mid \mathbf{x}_0\right) = \frac{\exp\left(-\,\ell_{\mathrm{f}}\left(\mathbf{x}_T\right)\right)}{z_0\left(\mathbf{x}_0\right)} \exp\left(-\sum\nolimits_{t=1}^{T-1} \ell_t\left(\mathbf{x}_t\right)\right) p^0\left(\mathbf{x}_1 \cdots \mathbf{x}_T \mid \mathbf{x}_0\right)$$

Note that $z_0\left(\mathbf{x}_0\right)$ acts as a partition function. Computing z_0 for all $\mathbf{x}_0$ would be equivalent to solving the problem globally. However in FH formulations where $\mathbf{x}_0$ is known, $z_0\left(\mathbf{x}_0\right)$ is merely a normalization constant. Thus we can characterize the *most likely* trajectory under the optimal control law, without actually knowing what the optimal control law is. In terms of negative log-probabilities, the most likely trajectory is the minimizer of

$$J\left(\mathbf{x}_1, \cdots \mathbf{x}_T \mid \mathbf{x}_0\right) = \ell_f\left(\mathbf{x}_T\right) + \sum\nolimits_{t=0}^{T-1} \ell_t\left(\mathbf{x}_t\right) - \log \Pi^0\left(\mathbf{x}_{t+1} \mid \mathbf{x}_t\right)$$

Interpreting $-\log \Pi^0\left(\mathbf{x}_{t+1} \mid \mathbf{x}_t\right)$ as a control cost, J becomes the total cost for a deterministic optimal control problem [33].

Similar results are also obtained in continuous time, where the relation between the stochastic and deterministic problems is particularly simple. Consider a FH problem with dynamics and cost rate

$$\begin{aligned} \mathrm{d}\,\mathbf{x} &= \mathsf{a}\left(\mathbf{x}\right)\mathrm{d}\,t + \mathsf{B}\left(\mathbf{x}\right)\left(u\,\mathrm{d}\,t + \sigma\,\mathrm{d}\,\omega\right) \\ \ell\left(\mathbf{x}, u\right) &= \ell\left(\mathbf{x}\right) + \frac{1}{2\sigma^2}\left\|u\right\|^2 \end{aligned}$$

It can be shown that the most likely trajectory under the optimally-controlled stochastic dynamics coincides with the optimal trajectory for the deterministic problem

$$\begin{aligned} \dot{\mathbf{x}} &= \mathsf{a}\left(\mathbf{x}\right) + \mathsf{B}\left(\mathbf{x}\right) u \\ \ell\left(\mathbf{x}, u\right) &= \ell\left(\mathbf{x}\right) + \frac{1}{2\sigma^2}\left\|u\right\|^2 + \frac{1}{2}\operatorname{div}\mathsf{a}\left(\mathbf{x}\right) \end{aligned} \tag{6.5}$$

The extra divergence cost pushes the deterministic dynamics away from states where the drift $\mathsf{a}\left(\mathbf{x}\right)$ is unstable. Note that the latter cost still depends on σ, and so the solution to the deterministic problem reflects the noise amplitude in the stochastic problem [33]. The maximum principle does extend to the LMG case and it characterizes the mostly likely trajectory of the closed loop system that includes both the controller and the adversary. For the discrete-time problem, the maximum principle reduces to minimizing

$$J^{\alpha}\left(\mathbf{x}_1, \cdots \mathbf{x}_T \,|\, \mathbf{x}_0\right) = (1-\alpha)\ell_f\left(\mathbf{x}_T\right) + \sum\nolimits_{t=0}^{T-1} (1-\alpha)\ell_t\left(\mathbf{x}_t\right) - \log \Pi^0\left(\mathbf{x}_{t+1} \,|\, \mathbf{x}_t\right)$$

Thus, when $\alpha < 1$, the most likely trajectory is trying to minimize accumulated state costs, while when $\alpha > 1$, the most likely trajectory is trying to maximize state costs. This gives us the interpretation that the controller "wins" the game for $\alpha < 1$ while the adversary "wins" the game for $\alpha > 1$.

6.5.6 Inverse optimal control

Consider the problem of getting a robot to perform locomotion or manipulation. Designing optimal controllers for these tasks is a computationally daunting task but biological systems accomplish these tasks with ease. Given this, a promising approach to designing controllers is to learn from biological systems and apply the same principles to robotic systems. There are reasons to believe that biological systems are optimal or near optimal, having been shaped by the processes of evolution and learning [26]. This motivates the problem of inverse optimal control, that is, inferring the control law and cost function given state space trajectories of the optimally controlled system. Traditionally, this has been done [1, 19, 40] by guessing a cost function, solving the (forward) optimal control problem and adjusting the cost function so that the resulting optimal behavior matches the observed behavior. However this approach defeats one of the main motivations of studying inverse optimal control – which is to leverage observed behavior of biological systems to design controllers without having to solve optimal control problems from scratch. We present an efficient algorithm that circumvents this problem, by using the framework of LMDPs to infer state cost functions given the passive dynamics Π^0 and state trajectories of the optimally-controlled system. Given a set of observed state transitions $\{(\mathbf{x}_n, \mathbf{x}\text{'}_n)\}$, the log-likelihood of the data up to a constant offset is

$$\sum_n -v\left(\mathbf{x}\text{'}_n;\theta\right) - \log\left(\mathop{\mathrm{E}}_{\Pi^0(\mathbf{x}_n)}\left[\exp\left(-v\left(\cdot;\theta\right)\right)\right]\right)$$

where $v\left(\mathbf{x};\theta\right)$ is a parameterized value function. We choose θ by maximizing the above log-likelihood, yielding an optimal estimate $v\left(\cdot;\theta^*\right)$ of the value function within our parametric family. Once we have inferred the value function, we can recover the cost function using $\ell(\mathbf{x}) = v\left(\mathbf{x};\theta^*\right) + \log\left(\mathrm{E}_{\Pi^0(\mathbf{x})}\left[\exp\left(-v\left(\cdot;\theta^*\right)\right)\right]\right)$. When we use a linear parametrization, $v\left(\mathbf{x};\theta\right) = \theta^T f(\mathbf{x})$, the likelihood maximization problem is a convex optimization problem and can be solved efficiently. However, in order to cope with high dimensional continuous state spaces, one needs to be able to adapt the features $f(\mathbf{x})$ as well, and we describe a non convex optimization approach to do this in [8]. Provided we know the risk parameter α, we can extend these results in a straightforward manner to LMGs.

6.6 CONCLUSIONS AND FUTURE WORK

Linearly-solvable optimal control is an exciting new development in control theory and has been the subject of many papers over the past few years. In this chapter we have attempted to provide a unified treatment of the developments in this area. The work so far has been mostly aimed at understanding the framework and its properties. We are now at a stage where the framework is mature and well understood and can lead to the development of algorithms that scale to hard real-world control problems from various application domains. Impressive results in robotics [24] and character animation [7] have recently been obtained. We feel that the surface has barely been scratched in terms of developing more efficient numerical methods for stochastic optimal control.

REFERENCES

1. P. Abbeel and A. Ng. Apprenticeship learning via inverse reinforcement learning. *International Conference on Machine Learning*, 21, 2004.
2. S.I. Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.
3. T. Başar and P. Bernhard. *H-infinity optimal control and related minimax design problems: a dynamic game approach*. Birkhauser, 1995.
4. D. Bertsekas. *Dynamic Programming and Optimal Control (2nd Ed)*. Athena Scientific, Bellmont, MA, 2001.
5. J. Broek, W. Wiegerinck, and Kappen H. Stochastic optimal control of state constrained systems. *International Journal of Control*, pages 1–9, 2011.
6. J. Broek, W. Wiegerinck, and H. Kappen. Risk sensitive path integral control. *Uncertainty in Artificial Intelligence*, 2010.
7. M. Da Silva, F. Durand, and J. Popović. Linear Bellman combination for control of character animation. *ACM Transactions on Graphics (TOG)*, 28(3):1–10, 2009.
8. K. Dvijotham and E. Todorov. Inverse optimal control with linearly-solvable MDPs. In Johannes Fürnkranz and Thorsten Joachims, editors, *International Conference on Machine Learning*, pages 335–342. Omnipress, 2010.
9. K. Dvijotham and E. Todorov. A unifying framework for linearly solvable control. In *Proceedings of the Proceedings of the Twenty-Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-11)*, pages 179–186, Corvallis, Oregon, 2011. AUAI Press.
10. W. Fleming and S. Mitter. Optimal control and nonlinear filtering for nondegenerate diffusion processes. *Stochastics*, 8:226–261, 1982.
11. C. Holland. A new energy characterization of the smallest eigenvalue of the Schrödinger equation. *Comm Pure Appl Math*, 30:755–765, 1977.
12. B. Hopf. The partial differential equation $u_t + uu_x = \mu u_{xx}$. *Comm Pure Appl Math*, 3:201–230, 1950.

13. Peters J. and Schaal S. Natural actor-critic. *Neurocomputing*, 71(7-9):1180 – 1190, 2008.
14. H.J. Kappen. Linear theory for control of nonlinear stochastic systems. *Physical Review Letters*, 95(20):200201, 2005.
15. S. Mahadevan and M. Maggioni. Proto-value functions: A laplacian framework for learning representation and control in markov decision processes. *Journal of Machine Learning Research*, 8:2169–2231, 2007.
16. S.I. Marcus, E. Fernández-Gaucherand, D. Hernández-Hernandez, S. Coraluppi, and P. Fard. Risk sensitive Markov decision processes. *Systems and Control in the Twenty-First Century*, 29, 1997.
17. T. Mensink, J. Verbeek, and H. Kappen. EP for efficient stochastic control with obstacles. *ECAI*, 2010.
18. S. Mitter and N. Newton. A variational approach to nonlinear estimation. *SIAM J Control Opt*, 42:1813–1833, 2003.
19. A.Y. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 663–670. Morgan Kaufmann Publishers Inc., 2000.
20. D. Precup, R.S. Sutton, and S. Singh. Multi-time models for temporally abstract planning. In *Advances in Neural Information Processing Systems 11*, 1998.
21. R. Stengel. *Optimal Control and Estimation*. Dover, New York, 1994.
22. R. Sutton, D. Mcallester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, 2000.
23. R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge MA, 1998.
24. E. Theodorou, J. Buchli, and S. Schaal. Reinforcement learning of motor skills in high dimensions: A path integral approach. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 2397–2403. IEEE, 2010.
25. E. A. Theodorou. *Iterative Path Integral Stochastic Optimal Control: Theory and Applications to Motor Control*. PhD thesis, University of Southern California, 2011.
26. E. Todorov. Optimality principles in sensorimotor control. *Nature Neuroscience*, 7(9):907–915, 2004.
27. E. Todorov. Linearly-solvable Markov decision problems. *Advances in neural information processing systems*, 19:1369, 2007.
28. E. Todorov. General duality between optimal control and estimation. *IEEE Conference on Decision and Control*, 47:4286–4292, 2008.
29. E. Todorov. Compositionality of optimal control laws. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1856–1864, 2009.
30. E. Todorov. Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences*, 106(28):11478, 2009.
31. E. Todorov. Eigenfunction approximation methods for linearly-solvable optimal control problems. In *IEEE International Symposium on Adaptive Dynamic Programming and Reinforcemenet Learning*, 2009.

32. E. Todorov. Policy gradients in linearly-solvable mdps. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2298–2306, 2010.

33. E. Todorov. Finding the Most Likely Trajectories of Optimally-Controlled Stochastic Systems. In *World Congress of the International Federation of Automatic Control (IFAC)*, 2011.

34. E. Todorov and W. Li. A generalized iterative LQG method for locally-optimal feedback control of constrained nonlinear stochastic systems. *American Control Conference*, pages 300–306, 2005.

35. M. Toussaint. Robot trajectory optimization using approximate inference. *International Conference on Machine Learning*, 26:1049–1056, 2009.

36. W. Wiegerinck, B. Broek, and H. Kappen. Stochastic optimal control in continuous space-time multi-AgentSystems. *22nd annual conference on Uncertainty in Artificial Intelligence*, 2006.

37. R. Williams. Simple statistical gradient following algorithms for connectionist reinforcement learning. *Machine Learning*, pages 229–256, 1992.

38. M. Zhong and E. Todorov. Aggregation methods for linearly-solvable MDPs. *IFAC World Congress*, 2011.

39. M. Zhong and E. Todorov. Moving least-squares approximations for linearly-solvable stochastic optimal control problems. *Journal of Control Theory and Applications*, 9:451–463, 2011.

40. B.D. Ziebart, A. Maas, J.A. Bagnell, and A.K. Dey. Maximum entropy inverse reinforcement learning. In *Proc. AAAI*, pages 1433–1438, 2008.