A UNIFIED THEORY OF LINEARLY SOLVABLE OPTIMAL CONTROL

KRISHNAMURTHY DVIJOTHAM
EMANUEL TODOROV*

Abstract. We present a unified theory of Linearly Solvable Optimal Control, that is, a class of optimal control problems whose solution reduces to solving a linear equation (for finite state spaces) or a linear integral equation (for continuous state spaces). The framework presented includes all previous work on linearly solvable optimal control as special cases. It includes both standard control problems and risk-sensitive control problems. The degree of risk sensitivity is a parameter of the optimal control problem and can be tuned to achieve the desired trade-off between performance and robustness (to noise/modeling errors). Linearly Solvable Optimal Control problems also possess a number of attractive properties that we explore in this paper. We show that it is possible to construct optimal control laws for new problems by combining the control laws of previously solved optimal control problems. This leads to analytical solutions for a class of non-LQG control problems. Another property is the existence of a path integral representation of the solution to the optimal control problem, which allows us to leverage approximate probabilistic inference techniques to compute optimal control laws. Further, we show that the Inverse Optimal Control problem, that is, the problem of inferring the cost function given trajectories sampled from the optimal control law, can be posed as a convex optimization problem and solved efficiently for problems in this class. We show that the risk sensitive problems can also be viewed as zero sum stochastic games, where the degree of risk averseness grows as the adversary becomes stronger. Under this interpretation, we derive a stochastic maximum principle that characterizes the most likely trajectory of the optimally controlled closed-loop system (that includes both the controller and the adversary).

Key words. Optimal Control, Linear PDEs, Efficient Algorithms, Analytical Solutions

AMS subject classifications. Optimal Control

1. Introduction. [1]

Stochastic optimal control is an elegant framework for specifying and solving control problems. One simply defines a high level cost function encoding the goals of the control task and the framework of stochastic optimal control takes care of the details. Despite this elegance, applications of stochastic optimal control to complex real-world control problems have been rare. This is principally because of the computational complexity of solving optimal control, which scales exponentially with the number of state/control variables. This is a fundamental limitation that Richard Bellman called the curse of dimensionality [3] and we cannot hope to alleviate this problem completely. However, practical control problems often have a lot of structure that can potentially be exploited to obtain an efficient solution. This suggests that a fruitful line of research is to develop a subclass of problems that are more tractable and then try to approximate real control problems with problems of this subclass. One such class of problems is Linear Quadratic Gaussian (LQG) problems where the dynamics are linear, costs are quadratic and noise additive and Gaussian. LQG control is perhaps the most widely used optimal control method in practice. However, the assumption of linear dynamics is limiting. In this paper, we aim to address this limitation and develop a general class of control problems that are computationally tractable and amenable to approximation.

We build on a series of recent interesting results on Linearly Solvable Optimal

*Computer Science and Engineering & Applied Mathematics, University of Washington, Seattle

[1] A preliminary conference version of this work appeared in the proceedings of Uncertainty in Artificial Intelligence (UAI) 2011

Control, i.e., optimal control problems for which the Hamilton Jacobi Bellman (HJB) (in continuous time) [20] or the Bellman equation (BE) (in discrete time) [34] can be made **linear**. We develop a general class of linearly solvable optimal control problems (LCs) that include all previous work on linearly solvable optimal control as special cases. 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 cost-to-go function) over a high-dimensional continuous state space. Nevertheless, the linearity leads to interesting properties that facilitate the development of principled approximation schemes for solving these problems efficiently. LCs include both standard control problems and risk sensitive control problems, which allow one to tune a risk-sensitivity parameter so as to tradeoff performance and robustness. We show that these problems can also be interpreted in a game-theoretic fashion, extending the well known relationships between risk sensitive and game theoretic/robust control [14] [23]. LCs use Rényi divergences [27], a family of divergences between probability measures, as control costs. To the best of our knowledge, this is the first application of Rényi divergences in control. Further, we prove several interesting properties of LCs that can be used to develop efficient approximation methods for solving LCs in high dimensional state spaces.

The paper is organized as follows. In the following section 1.1, we discuss the historical development of the theory of linearly solvable optimal control and relationships between LCs and these previous results. In section 2, we present an intuitive development of LCs , outlining the principal mathematical ideas without going into rigorous proofs. In section 3, we show how all previous work on linearly solvable optimal control (to the best of our knowledge) can be viewed as special cases of LCs, showing that the theory of LCs can be viewed as a unified theory of linearly solvable optimal control. In section 4, we provide a rigorous measure-theoretic development of LCs , providing sufficient conditions for the existence of solutions to LCs for arbitrary compact Borel state spaces.

1.1. Historical perspective. The mathematical trick that makes the HJB or Bellman equation linear is the use of an exponential transformation that maps from cost-to-go functions to so-called "desirability" functions $z = \exp(\kappa v)$. Equivalently, $v = \frac{1}{\kappa}\log(z)$, so the cost-to-go function is a logarithmically transformed desirability function. Logarithmic transformations have a long history, with old roots in physics. Hopf [19] used a logarithmic transformation to show the connection between the Burgers equation and the heat equation. The first application of logarithmic transformations to stochastic optimal control is due to Fleming [13] and Holland [18]. In [15], Fleming and Mitter showed that non-linear filtering corresponds to a stochastic optimal control problem whose HJB equation can be made linear. In [14], these ideas were generalized to risk sensitive stochastic optimal control. A summary of results obtained in this area can be found in [16] (chapter 6). The initial motivation in these works was to use techniques based on optimization and optimal control to solve inference and estimation problems. However, in [20], Kappen realized that this connection between optimal control and estimation can be exploited in the opposite direction, ie, by developing efficient approximation techniques for computing control laws based on algorithms for probabilistic inference. He developed Monte Carlo approximation schemes to approximate the optimal cost-to-go function and compute approximately

optimal control laws efficiently for continuous time optimal control problems [20].

The discrete-time results on linearly solvable optimal control [31] were 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 Linearly Solvable MDP (LMDP) framework. In discrete time, the trick that makes the Bellman equation linear is a well known variational property [10] of the KL divergence and has been used to derive a variational characterization of the Bayesian posterior [25].

The discrete time results for the risk-neutral case were developed first in [31]. This was followed by several papers exploring various properties of the framework and applications. See [5,6,24,40] for continuous time results and [11,12,32–36,41,42] for discrete time results. Applications to robotic control and control of animated characters have been developed in [8, 29, 30, 37]. We generalized these results to the risk sensitive case in a recent conference paper [12] for finite state spaces. Our contribution here is to present a rigorous development of those results for arbitrary Borel state spaces.

The connection to stochastic games presented in section 4.2.2 is not new: Similar tricks have been used to show the existence of solutions to risk sensitive control problems by relating the solution to a discounted stochastic game [23] [9]. However, the results we develop here are slightly different since they directly relate infinite horizon stochastic games with infinite horizon risk sensitive control, without going through vanishing discount limit of discounted stochastic games. Further, we present a maximum principle characterizing the most likely trajectory of the optimally controlled system (section 4.3.4). This is a new result that does not hold in the general stochastic game setting.

2. Linearly Solvable Optimal Control : An Informal Introduction. In this section, we provide an informal introduction to the ideas contained in this paper that would be accessible to a broader audience. In section 4, we provide a rigorous measure theoretic development of these ideas that is applicable to control problems with arbitrary compact Borel state spaces.

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

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

where $\mathbf{x} \in \Re^n$. This equation specifies 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

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

accumulated over time. Suppose that the system is in state $\mathbf{x}_t$ at time t. If time moves forward by h (which is sufficiently small), the probability distribution of the state $\mathbf{x}_{t+h}$ at time $t+h$ is approximately

$$\mathbf{x}_{t+h} \sim \mathbb{P}\,(\mathbf{x}_t, \mathbf{u}_t) = \mathcal{N}(\mathbf{x}_t + h(\mathsf{a}(\mathbf{x}_t) + \mathsf{B}(\mathbf{x}_t)\,\mathbf{u}_t), hCC^T)$$

where we have discretized the stochastic process using a time step h. Thus, one way of thinking of the effect of control is that it changes the distribution of the next state from $\mathcal{N}(\mathbf{x}_t + ha(\mathbf{x}_t), hCC^T)$ to $\mathcal{N}(\mathbf{x}_t + h(\mathsf{a}(\mathbf{x}_t) + \mathsf{B}(\mathbf{x}_t)\,\mathbf{u}_t), HCC^T)$. Thus, we can think of the controlled dynamics $\mathcal{N}(\mathbf{x}_t + h(\mathsf{a}(\mathbf{x}_t) + \mathsf{B}(\mathbf{x}_t)\,\mathbf{u}_t), HCC^T)$ as an alternate way to parameterize our control input.

The linearly solvable optimal control problems we develop in this paper will use control signals that are probability measures over the entire state space and directly specify the distribution of the next state given the current state: $\mathbf{x}' \sim u(\mathbf{x})$. This gives the controller a lot of freedom, and in principle, it could pick the controlled dynamics to be a delta distribution at the goal state (in a goal-reaching task). In order to avoid this unrealistic solution, we impose a cost on picking a controlled density: this is defined as some sort of divergence to the natural or uncontrolled or **passive** dynamics of the system. The control cost measures how different the controlled dynamics is from the uncontrolled dynamics, thereby measuring the "control effort" used. We will also impose the restriction that u has the same support as the uncontrolled dynamics, so that any transition that happened under the controlled dynamics could also have happened under the passive dynamics (but perhaps with a much lower or higher probability). This interchangeability of systematic control and random chance, is a key feature of all linearly solvable optimal control problems. Mathematically, treating control signals as probability distribution and imposing a control cost of this kind, allows us to find the minimizing control in the Bellman equation analytically. Once this is done, for the specific costs we use, the Bellman equation can be exponentiated to get a linear equation which we call the linear Bellman equation. This is the essential intuition behind all the linearly solvable optimal control problems we will describe.

2.2. Markov Decision Processes. A Markov Decision Process (MDP) is characterized by specifying state space $\mathcal{X}$, a control space $\mathcal{U}$, a transition probability kernel $\mathbb{P}(\mathbf{x}'|\mathbf{x}, \mathbf{u}), \mathbf{x}, \mathbf{x}' \in \mathcal{X}, \mathbf{u} \in \mathcal{U}$, which denotes the probability of reaching $\mathbf{x}'$ in one time step after applying control $\mathbf{u}$ in state $\mathbf{x}$, and a cost function $\ell(\mathbf{x}, \mathbf{u})$, denoting the cost of applying control $\mathbf{u}$ in state $\mathbf{x}$. The objective is to design an **optimal feedback policy** $u^* : \mathcal{X} \to \mathcal{U}$ that minimizes the expected accumulated cost

$$\mathop{\mathrm{E}}_{\mathbf{x}_{t+1} \sim \mathbb{P}(\mathbf{x}_t, u(\mathbf{x}_t))} \left[\sum \ell(\mathbf{x}_t, u(\mathbf{x}_t)) \right].$$

The sum is over time, but can be accumulated in different ways (finite horizon, infinite horizon average cost, first exit, discounted infinite horizon). We will not get to the specifics of the formulation in this section. Section 4 contains a rigorous development with all the details.

2.3. Linearly Solvable MDPs (LMDPs). Consider a Markov Decision Process with a finite state space $\mathcal{X}$. In linearly solvable MDPs, the controller is allowed to directly pick the distribution of the next state $\mathbf{x}' \sim u$ given the current state $\mathbf{x}$. However, the choice is penalized for being different from the "passive" or "uncontrolled" dynamics of the system, $P^0(\mathbf{x})$ as $\mathrm{KL}\left(u \parallel P^0(\mathbf{x})\right)$. This acts like a control cost, and puts a penalty on the controller for making the controlled dynamics very different from the uncontrolled dynamics. In addition to this, we allow an arbitrary state cost $\ell(\mathbf{x})$. If we plug this into the Bellman equation for the infinite horizon average cost

MDP, we get

$$\begin{aligned}
v(\mathbf{x}) + \lambda &= \min_u \ell(\mathbf{x}) + \mathrm{KL}\left(u \parallel P^0(\mathbf{x})\right) + \mathop{\mathrm{E}}_u[v] \\
&= \ell(\mathbf{x}) + \min_u \mathop{\mathrm{E}}_u\left[\log\left(\frac{u}{P^0(\mathbf{x})}\right) + v\right] \\
&= \ell(\mathbf{x}) + \min_u \mathop{\mathrm{E}}_u\left[\log\left(\frac{u}{P^0(\mathbf{x})\exp(-v)}\right)\right] \\
&= \ell(\mathbf{x}) - \log\left(\mathop{\mathrm{E}}_{P^0(\mathbf{x})}[\exp(-v)]\right) + \min_u \mathrm{KL}\left(u \parallel \frac{P^0(\mathbf{x})\exp(-v)}{\mathrm{E}_{P^0(\mathbf{x})}[\exp(-v)]}\right) \\
&= \ell(\mathbf{x}) - \log\left(\mathop{\mathrm{E}}_{P^0(\mathbf{x})}[\exp(-v)]\right) \\
\exp(-\lambda)\exp(-v(\mathbf{x})) &= \exp(-\ell(\mathbf{x})) \mathop{\mathrm{E}}_{P^0(\mathbf{x})}[\exp(-v)]
\end{aligned}$$

where the optimal policy is $u^*(\mathbf{x}) = \frac{P^0(\mathbf{x})\exp(-v)}{\mathrm{E}_{P^0(\mathbf{x})}[\exp(-v)]}$, since the KL divergence is minimized when the two distributions are equal. Defining $z = \exp(-v)$, we get the linear Bellman equation

$$\exp(-\lambda)\, z(\mathbf{x}) = \exp(-\ell(\mathbf{x})) \mathop{\mathrm{E}}_{P^0(\mathbf{x})}[z].$$

Since z is inversely related to v, the cost-to-go function, we call it the **desirability function**.

2.4. Generalization to Risk Sensitive Control. Risk sensitive MDPs [23] are generalization of MDPs that take into account the risk seeking or risk averse nature of the decision maker (controller). Instead of minimizing the standard expected accumulated cost criterion, we minimize

$$\frac{1}{\alpha}\log\left(\mathop{\mathrm{E}}_{\mathbf{x}_{t+1}\sim\mathbb{P}(\mathbf{x}_t, u(\mathbf{x}_t))}\left[\exp\left(\alpha\left(\sum \ell(\mathbf{x}_t, u(\mathbf{x}_t))\right)\right)\right]\right)$$

where α is the degree of risk sensitivity. When $\alpha > 0$, the controller is risk averse, when $\alpha < 0$, the controller is risk seeking and in the limit $\alpha \to 0$, the above objective reduces to the standard MDP objective (known as the risk-neutral case). Risk sensitivity has been discussed in detail in decision theory literature [39]. For our purposes, we will treat it simply as a means to tune the behavior of a controller to trade-off performance and robustness. In fact, the risk sensitive formulation we will use is slightly different from the standard risk sensitive problem: In our formulation, the state cost will be scaled linearly by α, but the control cost will change nonlinearly (but monotonically) as a function of α. We will present a game theoretic interpretation (section 4.2.2) of our results that further justifies the intuition of using α to trade-off performance and robustness. The risk sensitive extension of Linearly Solvable MDPs is formulated with the following objective:

$$\min \frac{1}{\alpha}\log\left(\mathop{\mathrm{E}}_{\mathbf{x}_{t+1}\sim u(\mathbf{x}_t)}\left[\exp\left(\sum_t \alpha\,\ell(\mathbf{x}_t) + \mathbb{D}_\alpha\left(P^0(\mathbf{x}_t) \parallel u(\mathbf{x}_t)\right)\right)\right]\right)$$

where $\mathbb{D}_\alpha$ is the Rényi divergence, informally defined as

$$\mathbb{D}_\alpha(p \parallel q) = \frac{1}{\alpha - 1}\log\left(\int p(\mathbf{x})^\alpha q(\mathbf{x})^{1-\alpha}\,\mathrm{d}\,\mathbf{x}\right)$$

where p, q are probability density functions. The Bellman equation for this problem (in the infinite horizon average cost setting) can be written as

$$\begin{aligned} v(\mathbf{x}) + \lambda &= \min_u \ell(\mathbf{x}) + \frac{\mathbb{D}_\alpha\left(P^0(\mathbf{x}) \parallel u\right)}{\alpha} + \frac{1}{\alpha}\log\left(\mathop{\mathrm{E}}_{\mathbf{x'}\sim u(\mathbf{x})}\left[\exp\left(v(\mathbf{x'})\right)\right]\right) \\ &= \ell(\mathbf{x}) + \min_u \frac{\mathbb{D}_\alpha\left(P^0(\mathbf{x}) \parallel u\right)}{\alpha} + \frac{1}{\alpha}\log\left(\mathop{\mathrm{E}}_{\mathbf{x'}\sim u(\mathbf{x})}\left[\exp\left(v(\mathbf{x'})\right)\right]\right) \\ &= \ell(\mathbf{x}) + \frac{1}{\alpha-1}\log\left(\mathop{\mathrm{E}}_{\mathbf{x'}\sim P^0(\mathbf{x})}\left[\exp\left((\alpha-1)v(\mathbf{x'})\right)\right]\right) \end{aligned}$$

where the minimum is achieved at $u^*(\mathbf{x}) = \frac{P^0(\mathbf{x})(z)^{\frac{1}{\alpha-1}}}{\mathrm{E}_{P^0(\mathbf{x})}\left[(z)^{\frac{1}{\alpha-1}}\right]}$ (this will be proved rigorously in section 4). Like before, we define the **desirability function** $z(\mathbf{x}) = \exp\left((\alpha-1)v(\mathbf{x})\right)$ and get a linear bellman equation

$$\exp\left((\alpha-1)\lambda\right) z(\mathbf{x}) = \exp\left((\alpha-1)\,\ell(\mathbf{x})\right) \mathop{\mathrm{E}}_{P^0(\mathbf{x})}[z].$$

As α increases, the controller becomes increasingly risk averse. As $\alpha \to 0$, $\frac{\mathbb{D}_\alpha}{\alpha} \to \mathrm{KL}$, so that we recover the risk neutral LMDP case. We call this general class of linearly solvable optimal control problems LCs.

3. LCs: A Unified View of Linearly Solvable Control.

3.1. Relationship to LMDPs. Here we show that LCs can be seen as an extension of LMDPs. Suppose μ, μ' are probability measures over $\mathcal{X}$. Then,

$$\begin{aligned} \lim_{\alpha\to 0} \frac{\operatorname{sgn}(\alpha)\,\mathbb{D}_\alpha\left(\mu \parallel \mu'\right)}{\alpha} &= \lim_{\alpha\to 0} \frac{1}{\alpha(\alpha-1)}\log\left(\mathop{\mathrm{E}}_{\mu}\left[\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\right)^{1-\alpha}\right]\right) \\ &= \lim_{\alpha\to 0} \frac{1}{2\alpha-1}\log\left(\mathop{\mathrm{E}}_{\mu}\left[\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\right)^{1-\alpha}\log\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\right)\right]\right) \quad \text{(L'Hopital's rule)} \\ &= \mathop{\mathrm{E}}_{\mu}\left[\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\log\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\right)\right] = \mathop{\mathrm{E}}_{\mu'}\left[\log\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\right)\right] = \mathrm{KL}\left(\mu' \parallel \mu\right). \end{aligned}$$

Thus,

$$\begin{aligned} &\lim_{\alpha\to 0} \Psi^{\alpha}_{\mathbf{x}_{t+1}\sim u(\mathbf{x}_t)}\left[\left(\sum_{t=0}^{T-1}\ell_{\mathrm{t}}(\mathbf{x}_t) + \frac{\operatorname{sgn}(\alpha)}{\alpha}\,\mathbb{D}_\alpha\left(P^0(\mathbf{x}_t) \parallel u(\mathbf{x}_t)\right)\right) + \ell_{\mathrm{f}}(\mathbf{x}_T)\right] \\ &= \mathop{\mathrm{E}}_{\mathbf{x}_{t+1}\sim u(\mathbf{x}_t)}\left[\sum_{t=0}^{T-1}\ell_{\mathrm{t}}(\mathbf{x}_t) + \mathrm{KL}\left(u(\mathbf{x}_t) \parallel P^0(\mathbf{x}_t)\right)\right]. \end{aligned}$$

This is the objective for the Linearly Solvable MDPs (section 2.3) [34]. Taking the limit of the linear Bellman Equation (4.1) as $\alpha \to 0$ gives us the Linear Bellman Equation for LMDPs. Thus, in the risk-neutral limit $\alpha \to 0$, we recover LMDPs as a special case of LCs.

3.2. Relationship to Risk Sensitive Path Integral Control. A Linearly Solvable Risk Sensitive Controlled Diffusion (LRD) is defined by specifying an Ito diffusion process of the following kind:

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

a state cost function $\ell(\mathbf{x})$, a noise level σ, a scaling factor θ and a risk factor α. The Control Problem is to design a feedback policy $u : \mathcal{X} \to \mathcal{U}$ in order to minimize

$$\log\left(\mathrm{E}\exp\left(\alpha\int\left(\ell(\mathbf{x}(t)) + \frac{\theta}{2\sigma^2}u(\mathbf{x}(t))^T u(\mathbf{x}(t))\right)\mathrm{d}\,t\right)\right)$$

where the expectation is under the stochastic process defined by the diffusion. Consider an h-step Euler discretization of the problem:
$P^h(\mathbf{x}' \,|\, \mathbf{x}, \mathbf{u}) = \mathcal{N}(\mathbf{x} + (a(\mathbf{x}) + B(\mathbf{x})\,\mathbf{u})h; \sigma h B(\mathbf{x})B(\mathbf{x})^T)$
Let $z^h(\mathbf{x})$ be the z function for the h-step discretization and

$$P^{0^h}(\mathbf{x}' \,|\, \mathbf{x}) = \mathcal{N}(\mathbf{x} + a(\mathbf{x})h; \sigma h B(\mathbf{x})B(\mathbf{x})^T).$$

The Bellman equation for the LC with risk factor $\alpha' = \theta\alpha$ is

$$z^h(\mathbf{x})\exp\left(h(1-\alpha')(\ell(\mathbf{x}) - \lambda)\right) = \mathop{\mathrm{E}}_{P^{0^h}(\mathbf{x}' \,|\, \mathbf{x})}\left[z^h(\mathbf{x}')\right]$$

$$z^h(\mathbf{x})\frac{\exp\left(h(1-\alpha')(q(\mathbf{x}) - \lambda)\right) - 1}{h} = \mathrm{E}\left[\frac{z^h(\mathbf{x}') - z^h(\mathbf{x})}{h}\right]$$

Taking limits as $h \to 0$, stochastic calculus indicates that [26] that the RHS becomes the generator of the passive Ito diffusion $d\,\mathbf{x} = \mathsf{a}(\mathbf{x})dt + \sigma\,\mathsf{B}(\mathbf{x})d\omega$ applied to $z = \lim_{h\to 0} z^h$. Note here that we're taking two limits simultaneously the function z^h changes with h and so does the stochastic process. Assuming that z^h converges uniformly to a limiting function at a fast enough rate, this result can be stated rigorously. Assuming the result holds, taking the limit gives us

$$z\,(\mathbf{x})\,(\ell(\mathbf{x}) - \lambda) = \frac{\nabla_{\mathbf{x}} z\,(\mathbf{x})^T\,\mathsf{a}(\mathbf{x}) + \frac{\mathrm{tr}\left(\mathsf{B}(\mathbf{x})\mathsf{B}(\mathbf{x})^T\nabla^2 z(\mathbf{x})\right)}{2}}{1 - \theta\alpha}$$

Using the formulas for Rényi divergence between Gaussians, we get $\mathbb{D}_\alpha\left(P^{0^h}(\cdot\,|\,\mathbf{x}) \parallel P^h(\cdot\,|\,\mathbf{x}, \mathbf{u})\right) = \frac{h\alpha'\,\mathbf{u}^T\,\mathbf{u}}{2\sigma^2} = \frac{h\theta\alpha\,\mathbf{u}^T\,\mathbf{u}}{2\sigma^2}$ so that the overall immediate cost for the LCis $h\,\ell(\mathbf{x}) + \frac{h\alpha'\,\mathbf{u}^T\,\mathbf{u}}{2\sigma^2}$, which is exactly the discrete-time Euler approximation of the LRD cost integrated over the time step h. All these results indicate that we're solving exactly the traditional risk sensitive control problem for the given Ito process as $h \to 0$. However, we leave a rigorous derivation for future work and conjecture that the limit exists under suitable regularity assumptions, similar to results obtained in [22]. The linear PDE obtained above coincides with that obtained in [4] where the LRD problem was first considered.

3.2.1. Linearly Solvable Controlled Diffusions (LDs). The above result also relates LCs with the initial work on continuous time linearly solvable optimal control problems by Kappen [20]. We call these problems Linearly Solvable Controlled Diffusions (LDs). LDs can be viewed either as continuous time limits of LMDPs or as risk neutral limit of LRDs. They deal with controlled diffusions just like LRDs, but the objective is just the expected accumulated cost:

$$\mathrm{E}\int_0^T\left(\ell(\mathbf{x}(t)) + \frac{\theta}{2\sigma^2}u(\mathbf{x}(t))^T u(\mathbf{x}(t))\right)\mathrm{d}\,t$$

Since we have already related LCs with LMDPs and LRDs , the relationship to LDs follows.

3.3. A Unified Picture. LCs are the most general class of linearly solvable control problems known, to the best of our knowledge. There are two dimensions along which previous results can be recovered: taking risk neutral limits and taking continuous time limits. As the degree of risk sensitivity decreases ($\alpha \to 0$), we recover Linearly Solvable MDPs (LMDPs) [34] as a special case of LCs. When we view LCs as arising from the time-discretization of Linearly Solvable Risk Sensitive Controlled Diffusions(LRDs) [4] [30], we recover LCs as a continuous time limit ($\mathrm{d}\,t \to 0$). Linearly Solvable Controlled Diffusions(LDs) [20] can be recovered either as the continuous time limit of an LMDP , or as the non-adversarial limit ($\alpha \to 0$) of LRDs. The overall relationships between the various classes of linearly solvable control problems is summarized in the figure below:

$$\begin{array}{ccc} LCs & \xrightarrow[\alpha\to 0]{} & LMDPs \\ \Big\downarrow{\scriptstyle dt\to 0} & & \Big\downarrow{\scriptstyle dt\to 0} \\ LRDs & \xrightarrow{\alpha\to 0} & LDs \end{array}$$

We have given intuitive justifications of these relationships in the preceding sections. We now present a formal measure-theoretic development of LCs. Rigorous proofs of these relationships, however, are left for future work.

4. LCs: A Formal Development. We now present a rigorous measure-theoretic development of the theory of LCs.

4.1. Background and Notation. DEFINITION 1. $\mathrm{sgn}\,(\alpha) = 1$ *if* $\alpha \geq 0$ *and* -1 *otherwise.* Let $(\Omega, \mathcal{F}, \mu)$ be a probability space. Then, we have the following definitions:

DEFINITION 2. *If* μ_1, μ_2 *are two measures on* $(\Omega, \mathcal{F})$*, then we say that* μ_1 *is absolutely continuous with respect to* μ_2 *if*

$$\forall A \in \mathcal{F}, \mu_2(A) = 0 \implies \mu_1(A) = 0.$$

and denote this by $\mu_1 \ll \mu_2$*. Further, if* $\mu_1 \ll \mu_2$ *and* $\mu_2 \ll \mu_1$*, we say that these measures are mutually absolutely continuous.*

DEFINITION 3. *We use* $\mathrm{E}_\mu[f]$ *to denote the expectation of the function* f *under probability measure* μ*. Sometimes, we also write* $\mathrm{E}_{\mathbf{x}\text{'}\sim\mu}[f(\mathbf{x}\text{'})]$*, which denotes the expectation of the random variable* $f(\mathbf{x}\text{'})$ *where* $\mathbf{x}\text{'}$ *is sampled from* μ*. We use* $\mathrm{E}_{\mathbf{x}_{t+1}\sim g(\mathbf{x}_t)}[f(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_T)]$ *to denote the expectation of* f *under trajectories sampled from a Markov chain* $\mathbf{x}\text{'} \sim g(\mathbf{x})$*.*

DEFINITION 4. *Define*

$$\mathrm{Pos}\,\{(\Omega, \mathcal{F}, \mu)\} = \{f : f : \Omega \to \Re^+, \mathop{\mathrm{E}}_{\mu}[f] < \infty\}.$$

For any $f \in \mathrm{Pos}\,\{(\Omega, \mathcal{F}, \mu)\}$*, define a new probability measure on this space as*

$$(\mu \otimes f)(A) = \frac{\mathrm{E}_\mu[f\mathbb{I}[A]]}{\mathrm{E}_\mu[f]} \quad \forall A \in \mathcal{F}\,.$$

Then it can be shown that $(\mu \otimes f) \ll \mu$ *and if* $g \in \mathrm{Pos}\,\{(\Omega, \mathcal{F}, \mu)\}, \mathrm{E}_\mu[fg] < \infty$*, then* $\mathrm{E}_{\mu\otimes f}[g] = \frac{\mathrm{E}_\mu[fg]}{\mathrm{E}_\mu[f]}$*. Also, the Radon Nikodym derivative* $\frac{\mathrm{d}\mu\otimes f}{\mathrm{d}\mu} = \frac{f}{\mathrm{E}_\mu[f]}$ *and since* f *is strictly positive,* $\mu \ll \mu \otimes f$ *so that* $\mu, \mu \otimes f$ *are mutually absolutely continuous.*

DEFINITION 5. *Define*

$$\Psi_{\mu}^{\theta}[f] = \frac{1}{\theta}\log\left(\underset{\mu}{\mathrm{E}}\left[\exp\left(\theta f\right)\right]\right).$$

It can be shown that $\lim_{\theta\to 0}\Psi_{\mu}^{\theta}[f] = \mathrm{E}_{\mu}[f]$*, so we define* $\Psi_{0}^{f}[\mu] = \mathrm{E}_{\mu}[f]$*.*

DEFINITION 6. *If* μ, μ' *are probability measures on* $(\Omega, \mathcal{F})$ *such that* $\mu' \ll \mu, \mu \ll \mu'$ *then define the Rényi divergence of order* α $(\in (-\infty, \infty])$ *between them to be*

$$\mathbb{D}_{\alpha}(\mu \parallel \mu') = \begin{cases} \frac{\mathrm{sgn}(\alpha)}{\alpha-1}\log\left(\mathrm{E}_{\mu}\left[\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\right)^{1-\alpha}\right]\right) & \text{if } \alpha \neq 1 \\ \mathrm{KL}(\mu \parallel \mu') = \mathrm{E}_{\mu}\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\mu'}\right)\right] & \text{if } \alpha = 1 \end{cases}$$

where $\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}$ *is the Radon-Nikodym derivative. This definition generalizes the standard definition Rényi divergence (which is restricted to* $\alpha \geq 0$ *[38]). If* $\alpha < 0$*, it can be shown that* $\mathbb{D}_{\alpha}(\mu \parallel \mu') = \frac{-\alpha}{1-\alpha}\mathbb{D}_{1-\alpha}(\mu' \parallel \mu)$*. This observation allows us to generalize the properties of the standard Rényi divergence proved in [38]:*

- $\mathbb{D}_{\alpha}(\mu \parallel \mu')$ *is continuous as a function of* α *on* $\{\alpha : 0 \leq \alpha \leq 1$ *or* $\mathbb{D}_{\alpha}(\mu \parallel \mu') < \infty\}$*.*
- $\mathbb{D}_{\alpha} > 0\,\forall\alpha$*. If* $\alpha \neq 0$*,* $\mathbb{D}_{\alpha}(\mu \parallel \mu') = 0$ *if and only if* $\mu = \mu'$*.*
- $\mathbb{D}_{\alpha}$ *is decreasing for* $\alpha < 0$ *and increasing for* $\alpha > 0$*.*

4.2. Linearly Solvable Optimal Control Problems (LCs). DEFINITION 7. *A Linearly Solvable Optimal Control Problem (LC) is defined by specifying:*

a A compact metric space $\mathcal{X}$ *equipped with the Borel sigma algebra* $\mathcal{F}[\mathcal{X}]$ *of the Borel subsets of* $\mathcal{X}$*. This is the state space.*
b A passive dynamics $P^0 : \mathcal{X} \to \mathcal{P}[\mathcal{X}] = \{\mu : (\mathcal{X}, \mathcal{F}[\mathcal{X}], \mu)$*is a probability space*$\}$*.*
c A risk parameter $\alpha \in \Re$*.*
*d A problem formulation: Infinite Horizon Average Cost (*IH*), First Exit (*FE*) or Finite Horizon (*FH*).*
e For IH,FE*, a cost function* $\ell : \mathcal{X} \to \Re^{+}$ *and for* FH*, a time dependent cost function* $\ell_{\mathrm{t}} : \mathcal{X} \to \Re^{+}$ *for* FH*.*
f For FH,FE*, a terminal cost function* $\ell_{\mathrm{f}} : \mathcal{X} \to \Re^{+}$*.*
g For FH*, a time horizon* T *and for* FE*, a set of terminal states* $\mathcal{T} \in \mathcal{F}[\mathcal{X}]$*.*

The control problem for LCs can be formulated in 2 ways, all of which lead to the same linear Bellman equation. We present all formulations, since they offer complementary insights.

4.2.1. Risk Sensitive Interpretation. DEFINITION 8. *An optimal risk sensitive solution of an LC is a feedback policy* ${u_{\mathrm{c}}}^* : \mathcal{X} \to \mathcal{P}[\mathcal{X}]$ *that achieves the following extremum*

$$\min_{u:\mathcal{X}\to\mathcal{P}[\mathcal{X}]} \mathcal{J}(\mathbf{x}, u)$$
$$\textit{Subject to } u(\mathbf{x}) \ll P^0(\mathbf{x}), P^0(\mathbf{x}) \ll u(\mathbf{x}) \quad \forall\,\mathbf{x}$$

at every state $\mathbf{x} \in \mathcal{X}$*, where* $\mathcal{J}(\mathbf{x}, u)$ *depends on the problem formulation as follows:*

$$\text{FH:} \mathcal{J}(\mathbf{x}, u) = \Psi^{\alpha}_{\mathbf{x}_{t+1} \sim u(\mathbf{x}_t)} \left[\left(\sum_{t=0}^{T-1} \ell_{\mathrm{t}}(\mathbf{x}_t) + \frac{\operatorname{sgn}(\alpha)}{\alpha} \mathbb{D}_{\alpha} \left(P^0(\mathbf{x}_t) \parallel u(\mathbf{x}_t) \right) \right) + \ell_{\mathrm{f}}(\mathbf{x}_T) \right]$$

$$\text{FE:} \mathcal{J}(\mathbf{x}, u) = \Psi^{\alpha}_{\substack{\mathbf{x}_{t+1} \sim u(\mathbf{x}_t) \\ T_e = \min\{t : \mathbf{x}_t \in \mathcal{T}\}}} \left[\left(\sum_{t=0}^{T_e - 1} \ell_{\mathrm{t}}(\mathbf{x}_t) + \frac{\operatorname{sgn}(\alpha)}{\alpha} \mathbb{D}_{\alpha} \left(P^0(\mathbf{x}_t) \parallel u(\mathbf{x}_t) \right) \right) + \ell_{\mathrm{f}}(\mathbf{x}_{T_e}) \right]$$

$$\text{IH:} \mathcal{J}(\mathbf{x}, u) = \lim_{T \to \infty} \frac{1}{T} \Psi^{\alpha}_{\mathbf{x}_{t+1} \sim u(\mathbf{x}_t)} \left[\left(\sum_{t=0}^{T} \ell(\mathbf{x}_t) + \frac{\operatorname{sgn}(\alpha)}{\alpha} \mathbb{D}_{\alpha} \left(P^0(\mathbf{x}_t) \parallel u(\mathbf{x}_t) \right) \right) \right]$$

where the expectations are under the stochastic dynamics of the system defined by u *and* $\mathbf{x}_0 = \mathbf{x}$*.*

NOTE 1. *The objective above is very similar to that of Risk-Sensitive MDPs [23]. In Risk-Sensitive MDPs, one chooses a feedback policy* $\mathbf{u}_t(\mathbf{x})$ *that minimizes the cost*

$$\Psi^{\alpha}_{\mathbf{x}_{t+1} \sim \mathbb{P}(\mathbf{x}_t, \mathbf{u}_t)} \left[\sum_{t=0}^{T-1} \ell_{\mathrm{t}}(\mathbf{x}_t, \mathbf{u}_t) + \ell_{\mathrm{f}}(\mathbf{x}_T) \right].$$

In traditional Risk-Sensitive MDPs, the control $\mathbf{u}$ *is generally a continuous valued vector* $(\in \Re^n)$ *or a discrete symbol (as in bang-bang control). In LCs, the control input is an entire probability distribution over the state space and* $\mathbb{P}(\mathbf{x}, u) = u$*. For LCs, we have* $\ell_{\mathrm{t}}(\mathbf{x}, u) = \ell_{\mathrm{t}}(\mathbf{x}) + \frac{\operatorname{sgn}(\alpha)}{\alpha} \mathbb{D}_{\alpha} \left(P^0(\mathbf{x}) \parallel u \right)$*, so the control cost depends on* α*. In other words, traditional risk sensitive MDPs minimize linearly scaled exponentiated costs* $\mathrm{E} \exp\left(\sum_t \alpha \, \ell_{\mathrm{t}}(\mathbf{x}_t, \mathbf{u}_t) \right)$ *while LCs minimize*

$$\mathrm{E} \exp\left(\sum_t \alpha \, \ell_{\mathrm{t}}(\mathbf{x}_t) + \operatorname{sgn}(\alpha) \, \mathbb{D}_{\alpha} \left(P^0(\mathbf{x}_t) \parallel u(\mathbf{x}_t) \right) \right),$$

so that the control costs are nonlinearly scaled while the state costs are linearly scaled. However, the nonlinear scaling is still monotonic in α *and always has the same sign as* α*, which makes the resulting behavior similar to traditional risk sensitive control. We also provide a game theoretic interpretation in the next section that further justifies this non-standard formulation of risk sensitivity. Note that the traditional approach of linearly scaling both the state and control costs does not lead to a linearly solvable optimal control problem.*

THEOREM 1. *If the following linear BE can be solved for a measurable function* $z : \mathcal{X} \to (\epsilon, 1)$ *for some* $\epsilon > 0$*,*

$$\begin{aligned}
\text{FH:} z(\mathbf{x}) &= \exp\left((\alpha - 1) \, \ell_{\mathrm{t}}(\mathbf{x}) \right) \mathop{\mathrm{E}}_{P^0(\mathbf{x})} \left[z_{t+1} \right], t = 0, 1, \ldots, T-1, \forall \, \mathbf{x} \\
z_T(\mathbf{x}) &= \exp\left((\alpha - 1) \, \ell_{\mathrm{f}}(\mathbf{x}) \right) \forall \, \mathbf{x} \\
\text{FE:} z(\mathbf{x}) &= \exp\left((\alpha - 1) \, \ell(\mathbf{x}) \right) \mathop{\mathrm{E}}_{P^0(\mathbf{x})} [z] \quad \forall \, \mathbf{x} \notin \mathcal{T} \\
z(\mathbf{x}) &= \exp\left((\alpha - 1) \, \ell_{\mathrm{f}}(\mathbf{x}) \right) \forall \, \mathbf{x} \in \mathcal{T} \\
\text{IH:} z(\mathbf{x}) &= \lambda \exp\left((\alpha - 1) \, \ell(\mathbf{x}) \right) \mathop{\mathrm{E}}_{P^0(\mathbf{x})} [z] \quad \forall \, \mathbf{x}
\end{aligned} \tag{4.1}$$

then the LC has an optimal feedback solution given by $u^*(\mathbf{x}; t) = P^0(\mathbf{x}) \otimes (z_{t+1})^{\frac{1}{1-\alpha}}$ *for* FH *and* $u^*(\mathbf{x}) = P^0(\mathbf{x}) \otimes (z)^{\frac{1}{1-\alpha}}$ *for other formulations. We call* z *the desirability function.*

Proof. We present the proof for IH. If the BE for risk sensitive MDPs ([7] equation (2.4)) has a bounded solution, then it characterizes the optimal feedback policy (see verification theorems in [7] [9]). We will show that if the linear BE has a solution $z : \mathcal{X} \to (\epsilon, 1)$ for some $\epsilon > 0$, then the risk sensitive BE has a solution and its solution can be derived from z.
Suppose the linear BE has a solution $z : \mathcal{X} \to (\epsilon, 1)$ for some $\epsilon > 0$. Consider the function $v(\mathbf{x}) = \frac{1}{\alpha-1} \log(z(\mathbf{x}))$. Taking logarithms of the linear BE and dividing by $\alpha - 1$ gives:

$$v(\mathbf{x}) = \ell(\mathbf{x}) + \frac{1}{\alpha - 1} \log\left(\underset{P^0(\mathbf{x})}{\mathrm{E}} \left[\exp\left((\alpha - 1) v \right) \right] \right) \tag{4.2}$$

and v is measurable and bounded in the interval with endpoints at $\frac{\log(\epsilon)}{\alpha-1}$ and 0. By lemma 2, we know that

$$\operatorname{sgn}(\alpha)\, \mathbb{D}_\alpha \left(P^0(\mathbf{x}) \parallel u \right) + \Psi_u \left[\alpha v \right] \geq - \Psi^{\frac{1-\alpha}{\alpha}}_{P^0(\mathbf{x})} \left[-\alpha v \right] \quad \forall\, u \in \mathcal{P}\left[(\mathcal{X}, \mathcal{F}[\mathcal{X}]) \right].$$

if $\alpha > 0$ and the inequality is reserved if $\alpha < 0$. Since v is bounded, the probability measure $u^* = P^0(\mathbf{x}) \otimes \exp(-v)$ exists and by definition, satisfies $u^* \ll P^0(\mathbf{x}), P^0(\mathbf{x}) \ll u^*$. We have

$$\frac{\mathrm{d}u^*}{\mathrm{d}P^0(\mathbf{x})} = \frac{\exp(-v)}{\mathrm{E}_{P^0(\mathbf{x})}\left[\exp(-v) \right]},$$

$$\mathbb{D}_\alpha \left(P^0(\mathbf{x}) \parallel u^* \right) = \frac{\operatorname{sgn}(\alpha)}{\alpha - 1} \log\left(\left(\underset{P^0(\mathbf{x})}{\mathrm{E}} \left[\exp\left((\alpha - 1) v \right) \right] \right)^{\alpha} \right),$$

$$\Psi_{u^*} \left[\alpha v \right] = \log\left(\underset{P^0(\mathbf{x})}{\mathrm{E}} \left[\exp\left((\alpha - 1) v \right) \right] \right).$$

$$\operatorname{sgn}(\alpha)\, \mathbb{D}_\alpha \left(P^0(\mathbf{x}) \parallel u^* \right) + \Psi_{P^0(\mathbf{x})} \left[(\alpha - 1) v \right] = \frac{\alpha}{\alpha - 1} \Psi_{P^0(\mathbf{x})} \left[(\alpha - 1) v \right] = - \Psi^{\frac{(1-\alpha)}{\alpha}}_{P^0(\mathbf{x})} \left[v \right].$$

Thus, we have

$$-\frac{1}{\alpha} \Psi^{\frac{(1-\alpha)}{\alpha}}_{P^0(\mathbf{x})} \left[-\alpha v \right] = \min_{u \ll P^0(\mathbf{x}), P^0(\mathbf{x}) \ll u} \frac{\operatorname{sgn}(\alpha)}{\alpha}\, \mathbb{D}_\alpha \left(P^0(\mathbf{x}) \parallel u \right) + \frac{1}{\alpha} \Psi_u \left[\alpha v \right].$$

with the minimum attained at $u = u^*$. The LHS is equal to $\Psi^{\alpha-1}_{P^0(\mathbf{x})}[v]$. Thus, (4.2) can be rewritten as

$$v(\mathbf{x}) = \ell(\mathbf{x}) + \min_{u \ll P^0(\mathbf{x}), P^0(\mathbf{x}) \ll u} \frac{\operatorname{sgn}(\alpha)}{\alpha}\, \mathbb{D}_\alpha \left(P^0(\mathbf{x}) \parallel u \right) + \Psi^\alpha_u \left[v \right]$$

or equivalently as

$$v(\mathbf{x}) = \min_{u \ll P^0(\mathbf{x}), P^0(\mathbf{x}) \ll u} \ell(\mathbf{x}) + \frac{\operatorname{sgn}(\alpha)}{\alpha}\, \mathbb{D}_\alpha \left(P^0(\mathbf{x}) \parallel u \right) + \Psi^\alpha_u \left[v \right].$$

This is precisely the Risk Sensitive Bellman Equation ([7] equation (2.4)) with the LC cost plugged in. Thus, the Risk Sensitive Bellman Equation has a bounded

solution and hence the LC has an optimal policy given by $u^* = P^0(\mathbf{x}) \otimes \exp(-v) = P^0(\mathbf{x}) \otimes (z)^{\frac{1}{1-\alpha}}$. □

NOTE 2. *The solution to the Linear BE for* FH *problems always exists and is bounded away from* 0 *as long as the cost* $\ell_{\mathrm{t}}(\mathbf{x})$ *is bounded. This can be easily proven constructively, by starting at time* T *and using dynamic programming backward in time.*

THEOREM 2. *Consider an* IH *LC. Suppose that the passive dynamics is weaker Feller, that is, for every continuous and bounded function* f*, the function* $g(\mathbf{x}) = \mathrm{E}_{P^0(\mathbf{x})}[f]$ *is also continuous and bounded. Suppose also that the state cost* $\ell(\mathbf{x})$ *is continuous. Then, the linear BE has a solution* $z : \mathcal{X} \to (\epsilon, 1)$ *for some* $\epsilon > 0$.

Proof. Consider the cone K_+ of positive continuous measurable functions on $\mathcal{X}$. By the assumptions, the operator defined by $\mathcal{G}[f](\mathbf{x}) = \exp((\alpha-1)\ell(\mathbf{x}))\,\mathrm{E}_{P^0(\mathbf{x})}[f]$ is weakly Feller and leaves the cone K_+ invariant. Also, it is strongly positive in the sense that for any $f \in \mathrm{int}(K_+)$, $\mathcal{G}[f](\mathbf{x}) \in \mathrm{int}(K_+)$. Then, by the Krein-Rutman theorem ([21], chapter 11), we can conclude that the linear operator $\mathcal{G}$ has a unique eigenfunction $z \in \mathrm{int}(K_+)$ corresponding to the leading eigenvalue $\lambda \in \Re^+$. Since z is a continuous function on a compact domain, it must attain its bounds. Also, $z \in \mathrm{int}(K_+)$. The above facts imply that z is bounded below by some number $\eta > 0$. Since any scaled version of z is also an eigenfunction, we can scale z by $1/(M+1)$ (where M is the upper bound on z) to get a function $\tilde{z} : \mathcal{X} \to [\frac{\eta}{M+1}, 1)$. Taking $\epsilon = \frac{\eta}{2*(M+1)}$ gives us the result. □

THEOREM 3. *Consider an* FE *LC. Suppose that the passive dynamics is such that* $\exists T_f < \infty$ *such that* $P(\mathbf{x}_{T_f} \in \mathcal{T} \,|\, \mathbf{x}_0) > \delta > 0 \forall\, \mathbf{x}_0 \in \mathcal{X}$ *under sampling from the passive dynamics* $\mathbf{x}_{t+1} \sim P^0(\mathbf{x}_t)$*. Further, suppose that* $\alpha < 1$*. Then there exists a solution* $z : (\epsilon, 1)$ *satisfying the linear BE for some* $\epsilon > 0$.

Proof. Consider the cone K_+ of positive continuous measurable functions on $\mathcal{X}$. By the assumptions, the operator defined by $\mathcal{G}[f](\mathbf{x}) = \exp((\alpha-1)\ell(\mathbf{x}))\,\mathrm{E}_{P^0(\mathbf{x})}[f - f\mathbb{I}[\mathcal{T}]]$ leaves the cone K_+ invariant. Also, the cone K_+ is solid and normal. Taking f to be the constant function 1, we have $\mathcal{G}[f] \leq \left(\sup_{\mathbf{x}\in\overline{\mathcal{X}\setminus\mathcal{T}}} \exp((\alpha-1)\ell(\mathbf{x}))\right) f$. Since $\overline{\mathcal{X}\setminus\mathcal{T}}$ is compact and $\ell(\mathbf{x})$ is continuous, the supremum is attained and is finite, say γ. Further, since $\ell > 0$,$\alpha < 1$, $\exp((\alpha-1)\ell(\mathbf{x})) < 1$ so that $\gamma < 1$. Thus, there exist $f \in \mathrm{int}(K_+)$ such that $\mathcal{G}[f] \leq \gamma f$, $0 < \gamma < 1$. Hence, the spectral radius of the operator $\mathcal{G}$ is smaller than 1. Hence $(I - \mathcal{G})$ is positively invertible. Define the function $g : \mathcal{X} \setminus \mathcal{T} \mapsto \Re$ as $g(\mathbf{x}) = \mathrm{E}_{P^0(\mathbf{x})}[\mathbb{I}[\mathcal{T}]\exp((\alpha-1)\ell_{\mathrm{f}}(\mathbf{x}))]$. The function defined by $z = (I-\mathcal{G})^{-1} g$ where $\forall\, \mathbf{x} \in \mathcal{X}\setminus\mathcal{T}$ and $z(\mathbf{x}) = \exp((\alpha-1)\ell_{\mathrm{f}}(\mathbf{x}))\,\forall\, \mathbf{x} \in \mathcal{T}$ is the unique solution to the linear BE. Also, from theorem 6, z has a path integral representation that shows that under the assumptions of the theorem, z is bounded above and is bounded below by some $\epsilon > 0$. □

4.2.2. Game Theoretic Interpretation. DEFINITION 9. *In the game theoretic interpretation, an LC has 2 players, the adversary and the controller. It proceeds as follows:*

- *The system is in state* $\mathbf{x}$ *at time* t.
- *The adversary and controller pick distributions* $u_{\mathrm{c}}, u_{\mathrm{a}} \in \mathcal{P}[\mathcal{X}]$ *such that* $u_{\mathrm{c}} \ll P^0(\mathbf{x}) \ll u_{\mathrm{c}}, u_{\mathrm{a}} \ll P^0(\mathbf{x}) \ll u_{\mathrm{a}}$.
- *The controller incurs cost* $\ell(\mathbf{x}, u_{\mathrm{c}}, u_{\mathrm{a}}) = \ell(\mathbf{x}) - \mathbb{D}_{1/\alpha}\left(P^0(\mathbf{x}) \parallel u_{\mathrm{a}}\right) + \mathrm{KL}\left(u_{\mathrm{c}} \parallel u_{\mathrm{a}}\right)$ *while the adversary incurs cost* $-\ell(\mathbf{x}, u_{\mathrm{c}}, u_{\mathrm{a}})$.
- *The system transitions into a state* $\mathbf{x}$' *sampled from the distribution* u_{c} *and time advances to* $t+1$.

NOTE 3. *An LC game is a special case of a zero-sum stochastic game [2]. The passive dynamics is meant to encode the autonomous behavior of the system in the absence of controls. The cost function consists of three terms. The first is a state cost* $\ell(\mathbf{x})$ *encoding desirable states the controller would like to be in. The second term is a control cost for the adversary* $-\mathbb{D}_{1/\alpha}\left(P^0(s) \parallel u_{\mathrm{a}}\right)$ *which penalizes how much the adversary changes the uncontrolled dynamics (the negative sign is because the adversary is maximizing the cost). This cost decreases monotonically with* α*, so the adversary becomes more powerful as* α *increases. The third term is a control cost for the controller,* $\mathrm{KL}\left(u_{\mathrm{c}} \parallel u_{\mathrm{a}}\right)$ *which measures the additional change in the system dynamics due to the controller. Thus, as* α *increases, we expect the adversary to dominate and the controller's strategy to change so as to counter a stronger adversary. The final controlled dynamics is completely determined by the controller's choice, but the controller has to pay a cost for deviating from the choice of the adversary and the adversary has to pay a cost for deviating from the passive dynamics.*

DEFINITION 10. *A* saddle point equilibrium *of an LC is a pair of feedback policies* ${u_{\mathrm{c}}}^*, {u_{\mathrm{a}}}^* : \mathcal{X} \to \mathcal{P}\left[\mathcal{X}\right]$ *that achieve the following extremum*

$$\min_{u_{\mathrm{c}}} \max_{u_{\mathrm{a}}} \mathcal{J}(\mathbf{x}, u_{\mathrm{c}}, u_{\mathrm{a}}) = \max_{u_{\mathrm{a}}} \min_{u_{\mathrm{c}}} \mathcal{J}(\mathbf{x}, u_{\mathrm{c}}, u_{\mathrm{a}})$$

at every state $\mathbf{x} \in \mathcal{X}$*, where* $\mathcal{J}(\mathbf{x}, u_{\mathrm{c}}, u_{\mathrm{a}})$ *depends on the problem formulation:*

$$\textit{Finite Horizon (FH):}\ \mathcal{J}(\mathbf{x}, u_{\mathrm{c}}, u_{\mathrm{a}}) = \mathop{\mathrm{E}}_{u_{\mathrm{c}}, u_{\mathrm{a}}, \mathbf{x}_0 = \mathbf{x}}\left[\sum_{t=0}^{T-1} \ell(\mathbf{x}_t, u_{\mathrm{c}}(\mathbf{x}_t), u_{\mathrm{a}}(\mathbf{x}_t)) + \ell_{\mathrm{f}}(\mathbf{x}_T)\right]$$

$$\textit{First Exit (FE):}\ \mathcal{J}(\mathbf{x}, u_{\mathrm{c}}, u_{\mathrm{a}}) = \mathop{\mathrm{E}}_{u_{\mathrm{c}}, u_{\mathrm{a}}, \mathbf{x}_0 = \mathbf{x}}\left[\sum_{t=0}^{T-1} \ell(\mathbf{x}_t, u_{\mathrm{c}}(\mathbf{x}_t), u_{\mathrm{a}}(\mathbf{x}_t)) + \ell_{\mathrm{f}}(\mathbf{x}_T)\right]$$

$$T = \min\{t : s_t \in \mathcal{T}\} - 1$$

$$\textit{Infinite Horizon (IH):}\ \mathcal{J}(\mathbf{x}, u_{\mathrm{c}}, u_{\mathrm{a}}) = \lim_{T \to \infty} \frac{1}{T} \mathop{\mathrm{E}}_{u_{\mathrm{c}}, u_{\mathrm{a}}, \mathbf{x}_0 = \mathbf{x}}\left[\sum_{t=0}^{T} \ell(\mathbf{x}_t, u_{\mathrm{c}}(\mathbf{x}_t), u_{\mathrm{a}}(\mathbf{x}_t))\right]$$

where the expectations are under the stochastic dynamics of the system defined by $P^0, u_{\mathrm{c}}, u_{\mathrm{a}}$*.*

THEOREM 4. *If* $\alpha <$*, the LC always has a saddle point equilibrium provided* (4.1) *has a continuous solution* $z : \mathcal{X} \to (\epsilon, 1)$ *for some* $\epsilon > 0$*. The saddle point equilibrium is given by*

$${u_{\mathrm{a}}}^*(\mathbf{x}; t) = P^0(\mathbf{x}) \otimes (z_{t+1})^{\frac{\alpha}{\alpha - 1}},\ {u_{\mathrm{c}}}^*(\mathbf{x}; t) = {u_{\mathrm{a}}}^*(\mathbf{x}; t) \otimes (z_{t+1})^{\frac{1}{1-\alpha}}$$

for FH *and*

$${u_{\mathrm{a}}}^*(\mathbf{x}) = P^0(\mathbf{x}) \otimes (z)^{\frac{\alpha}{\alpha - 1}},\ {u_{\mathrm{c}}}^*(\mathbf{x}) = {u_{\mathrm{a}}}^*(\mathbf{x}) \otimes (z)^{\frac{1}{1-\alpha}}$$

for IH,FE*.*

Proof. We do the proof for the IH case, the proof for the other cases is similar. The verification theorem in [17] (theorem 3.4) says that if v is a continuous bounded solution to the Shapeley equation [28]

$$\begin{aligned}\lambda + v(\mathbf{x}) &= \min_{u_{\mathrm{c}} \ll P^0(\mathbf{x}) \ll u_{\mathrm{c}}} \max_{u_{\mathrm{a}} \ll P^0(\mathbf{x}) \ll u_{\mathrm{a}}} \left[\ell(\mathbf{x}) - \mathbb{D}_{1/\alpha}\left(P^0(\mathbf{x}) \parallel u_{\mathrm{a}}\right) + \mathrm{KL}\left(u_{\mathrm{c}} \parallel u_{\mathrm{a}}\right) + \mathop{\mathrm{E}}_{u_{\mathrm{c}}}[v]\right] \\ &= \max_{u_{\mathrm{a}} \ll P^0(\mathbf{x}) \ll u_{\mathrm{a}}} \min_{u_{\mathrm{c}} \ll P^0(\mathbf{x}) \ll u_{\mathrm{c}}} \left[\ell(\mathbf{x}) - \mathbb{D}_{1/\alpha}\left(P^0(\mathbf{x}) \parallel u_{\mathrm{a}}\right) + \mathrm{KL}\left(u_{\mathrm{c}} \parallel u_{\mathrm{a}}\right) + \mathop{\mathrm{E}}_{u_{\mathrm{c}}}[v]\right]\end{aligned}$$

then the saddle point equilibrium is attained for the feedback policies setting $u_{\mathrm{c}}, u_{\mathrm{a}}$ to the values attaining the extremum in the Shapely equation above. We are given that the linear BE (4.1) has a solution, say z. Consider the function $v = \frac{1}{\alpha-1}\log(z)$. This is bounded and continuous function since z is continuous,bounded above and below away from 0. Using theorem 9, we know that

$$\begin{aligned}&\max_{u_{\mathrm{a}} \ll P^0(\mathbf{x}) \ll u_{\mathrm{a}}} \min_{u_{\mathrm{c}} \ll P^0(\mathbf{x}) \ll u_{\mathrm{c}}} \left[\ell(\mathbf{x}) - \mathbb{D}_{1/\alpha}\left(P^0(\mathbf{x}) \parallel u_{\mathrm{a}}\right) + \mathrm{KL}\left(u_{\mathrm{c}} \parallel u_{\mathrm{a}}\right) + \mathop{\mathrm{E}}_{u_{\mathrm{c}}}[v]\right] \\ &= \ell(\mathbf{x}) - \Psi_{P^0(\mathbf{x})}^{\frac{\frac{1}{\alpha}-1}{\frac{1}{\alpha}}}[-v] = \ell(\mathbf{x}) + \Psi_{P^0(\mathbf{x})}^{1-\alpha}[v]\end{aligned}$$

and the same result holds if the min, max are interchanged. Since z satisfies the linear BE, we know that v satisfies

$$v(\mathbf{x}) = \ell(\mathbf{x}) + \frac{1}{\alpha-1}\log\left(\mathop{\mathrm{E}}_{P^0(\mathbf{x})}\left[\exp\left((\alpha-1)v\right)\right]\right) = \ell(\mathbf{x}) + \Psi_{P^0(\mathbf{x})}^{1-\alpha}[v].$$

Combining the results above, we have that v is a bounded continuous solution to the Shapely equation and hence by the verification theorem (theorem 3.4 in [17]), we have the result. □

4.3. LCs: Properties and Efficient Algorithms. For control problems with discrete (finite) state spaces, the linearity implies that the optimal control problem reduces to solving a linear equation (in the first exit setting), a linear eigenvalue problem (in the infinite horizon average cost setting) and a set of matrix-vector multiplications (in the finite horizon setting). Also, for most control problems of practical interest, the linear systems involved are very sparse and this can be exploited further to get very efficient algorithms. We can easily solve problems with millions of discrete states in time of the order of minutes. For continuous state spaces, the linearity makes the Bellman equation amenable to cost-to-go function approximation techniques. It also leads to other interesting properties that we describe in the following subsections.

4.3.1. Compositionality of Optimal Control Laws. THEOREM 5. *Consider a set of* FH *or* FE *LC problems with the different terminal costs* ${\ell_{\mathrm{f}}}^{(i)}(\mathbf{x}), i = 1, 2, \ldots, k$ *but the same running cost* $\ell(\mathbf{x})$ *(or* $\ell_{\mathrm{t}}(\mathbf{x})$*), passive dynamics* $P^0(\mathbf{x})$ *and risk parameter* α*. Suppose that for each LC, the optimal desirability function* $z^{(i)}$ *satisfying the linear BE exists and is bounded away from* 0*. Then, the optimal desirability for a the LC with terminal cost given by*

$$\ell_{\mathrm{f}}(\mathbf{x}) = \frac{1}{\alpha-1}\log\left(\sum_{i=1}^{k} w_i \exp\left((\alpha-1)\,{\ell_{\mathrm{f}}}^{(i)}(\mathbf{x})\right)\right), w_i > 0$$

is

$$z(\mathbf{x}) = \sum_{i=1}^{k} w_i\, z^{(i)}(\mathbf{x}).$$

z, ℓ *are additionally time dependent in the* FH *case.*

Proof. This is a direct consequence of linearity. Let $\beta = \alpha - 1$. For the FE problem, the linear BE is given by

$$z^{(i)}(\mathbf{x}) = \exp\left(\beta\, \ell(\mathbf{x})\right) \left(\int (1 - \mathbb{I}\left[\mathcal{T}\right]) \, z^{(i)} \, \mathrm{d}\, P^0(\mathbf{x}) + \int \mathbb{I}\left[\mathcal{T}\right] \exp\left(\beta\, {\ell_{\mathrm{f}}}^{(i)}(\mathbf{x})\right) \mathrm{d}\, P^0(\mathbf{x}) \right) \mathbf{x} \notin \mathcal{T}$$

$$z^{(i)}(\mathbf{x}) = \exp\left(\beta\, {\ell_{\mathrm{f}}}^{(i)}(\mathbf{x})\right) \mathbf{x} \in \mathcal{T}$$

Thus,

$$\sum_i w_i \, z^{(i)}(\mathbf{x}) = \sum_i w_i \exp\left(\beta\, {\ell_{\mathrm{f}}}^{(i)}(\mathbf{x})\right) \mathbf{x} \in \mathcal{T}$$

$$\sum_i w_i \, z^{(i)}(\mathbf{x}) = \sum_i w_i \exp\left(\beta\, \ell(\mathbf{x})\right) \left(\int (1 - \mathbb{I}\left[\mathcal{T}\right]) \, z^{(i)} \, \mathrm{d}\, P^0(\mathbf{x}) + \int \mathbb{I}\left[\mathcal{T}\right] \exp\left(\beta\, {\ell_{\mathrm{f}}}^{(i)}(\mathbf{x})\right) \mathrm{d}\, P^0(\mathbf{x}) \right)$$

$$= \exp\left(\beta\, \ell(\mathbf{x})\right) \int (1 - \mathbb{I}\left[\mathcal{T}\right]) \left(\sum_i w_i \, z^{(i)} \right) \mathrm{d}\, P^0(\mathbf{x})$$

$$+ \exp\left(\beta\, \ell(\mathbf{x})\right) \int \mathbb{I}\left[\mathcal{T}\right] \left(\sum_i w_i \exp\left(\beta\, {\ell_{\mathrm{f}}}^{(i)}(\mathbf{x})\right) \right) \mathrm{d}\, P^0(\mathbf{x})$$

Thus $z = \sum_i w_i \, z^{(i)}$ satisfies the linear BE for the new terminal cost. Also, since $w_i > 0$ and k is finite, z is also bounded away from 0.

□

NOTE 4. *This result was presented for the risk neutral case in [33]. It was stated in the conference paper [12] for the risk sensitive case without formal proof. Here, we derive it for arbitrary Borel spaces.*

NOTE 5. *This gives a set of non-LQG problems for which we can compute analytical solutions. If we have a* FH *problem with linear dynamics and quadratic costs, the optimal desirability function can be computed analytically. Now, we can compose the terminal costs* $\ell_{\mathrm{f}}(\mathbf{x}) = \frac{1}{\alpha - 1} \log\left(\sum_i w_i \exp\left((\alpha - 1) \frac{\mathbf{x}^T Q_f^{(i)} \mathbf{x}}{2} \right) \right)$ *to get a non quadratic terminal cost. By compositionality, the optimal desirability function for this problem is available analytically by linearly combining those for the constituent problems. The corresponding control policy can be seen as interpolating between the optimal linear feedback policies for each of the constituent problems.*

4.3.2. Path Integral Representation. THEOREM 6. *If there exists a solution* z *to the linear BE for an* FH *or* FE *LC problem, then it has the path integral representation*

$$z\left(\mathbf{x}\right) = \mathop{\mathrm{E}}_{\substack{\mathbf{x}_{\tau+1} \sim P^0(\mathbf{x}_\tau), \mathbf{x}_0 = \mathbf{x} \\ T_f = \min \tau : \mathbf{x}_\tau \in \mathcal{T}}} \left[\exp\left((\alpha - 1) \left(\left(\sum_{\tau=0}^{\min(T_f - 1, 0)} \ell(\mathbf{x}_\tau) \right) + \ell_{\mathrm{f}}(\mathbf{x}_{T_f}) \right) \right) \right]$$

for FE *and*

$$z_t\left(\mathbf{x}\right) = \mathop{\mathrm{E}}_{\mathbf{x}_{\tau+1} \sim P^0(\mathbf{x}_\tau), \mathbf{x}_0 = \mathbf{x}} \left[\exp\left((\alpha - 1) \left(\left(\sum_{\tau=t}^{T-1} \ell_\tau(\mathbf{x}_\tau) \right) + \ell_{\mathrm{f}}(s_T) \right) \right) \right]$$

for FH.

Proof. For FH the result is easily proven by induction. For FE if $\mathbf{x} \in \mathcal{T}$, $T_f = 0$, and from the formula above, $z(\mathbf{x}) = \exp((\alpha-1)\ell_{\mathrm{f}}(\mathbf{x}))$. If $\mathbf{x} \notin \mathcal{T}$, then

$$\begin{aligned}
\mathop{\mathrm{E}}_{\mathbf{x}'\sim P^0(\mathbf{x})}[z(\mathbf{x}')] &= \mathop{\mathrm{E}}_{\mathbf{x}'\sim P^0(\mathbf{x})}\left[\mathop{\mathrm{E}}_{\substack{\mathbf{x}_{\tau+1}\sim P^0(\mathbf{x}_\tau),\mathbf{x}_0=\mathbf{x}'\\ T_f=\min\tau:\mathbf{x}_\tau\in\mathcal{T}}}\left[\exp\left((\alpha-1)\sum_{\tau=0}^{\min(T_f-1,0)}\ell(\mathbf{x}_\tau)+\ell_{\mathrm{f}}(\mathbf{x}_{T_f})\right)\right]\right]\\
&= \mathop{\mathrm{E}}_{\substack{\mathbf{x}_{\tau+1}\sim P^0(\mathbf{x}_\tau),\mathbf{x}_0=\mathbf{x}\\ T_f=\min\tau:\mathbf{x}_\tau\in\mathcal{T}}}\left[\exp\left((\alpha-1)\sum_{\tau=1}^{\min(T_f-1,0)}\ell(\mathbf{x}_\tau)+\ell_{\mathrm{f}}(\mathbf{x}_{T_f})\right)\right]
\end{aligned}$$

Multiplying throughout by $\exp((\alpha-1)\ell(\mathbf{x}))$, we see that the z function defined through the formula above satisfies the linear BE. □

NOTE 6. *The path integral formulation was developed for continuous time problems by Kappen in [4, 20]. For the LMDP case, it was developed in [34]. It was stated in the conference paper [12] without formal proof for finite state spaces. Here, we derive it for arbitrary Borel spaces.*

4.3.3. Inverse Optimal Control. The inverse optimal control problem is the problem of inferring the state cost $\ell(\mathbf{x})$ of an LC given the risk parameter α, the passive dynamics P^0 and state transitions sampled from the optimal control law $\mathcal{D} = \{\mathbf{x}_i, \mathbf{x}'_i\}_{i=1}^N$ where $\mathbf{x}'_i \sim u^*(\mathbf{x}_i)$.

THEOREM 7. *Consider an LC with a finite state space $\mathcal{X} = \{1, 2, \ldots, n\}$ and let $\mathcal{F}[\mathcal{X}]$ be the power set of $\mathcal{X}$. Let $v(\mathbf{x};\theta)$ parameterized family of candidate cost-to-gofunctions. The log-likelihood of observing a set of transitions $\{\mathbf{x}_i, \mathbf{x}'_i\}$ under the optimally controlled system (assuming each transition is independent) given θ is*

$$\begin{aligned}
L(\theta) &= \sum_i \log\left(u^*(\mathbf{x}'_i \,|\, \mathbf{x}_i;\theta)\right)\\
&= \sum_i \log\left(P^0(\mathbf{x}'_i \,|\, \mathbf{x}_i)\exp\left(-v(\mathbf{x}';\theta)\right)\right) - \log\left(\mathop{\mathrm{E}}_{\mathbf{x}'\sim P^0(\mathbf{x}_i)}\left[\exp\left(-v(\mathbf{x}';\theta)\right)\right]\right)\\
&= c + \sum_i -v(\mathbf{x}'_i;\theta) - \log\left(\mathop{\mathrm{E}}_{\mathbf{x}'\sim P^0(\mathbf{x}_i)}\left[\exp\left(-v(\mathbf{x}';\theta)\right)\right]\right)
\end{aligned}$$

where c is a constant independent of θ. The maximum likelihood estimate of the state cost is then

$$\ell^*(\mathbf{x}) = v(\mathbf{x};\theta^*) - \frac{1}{\alpha-1}\log\left(\mathop{\mathrm{E}}_{P^0(\mathbf{x})}\left[\exp\left((\alpha-1)v(\theta^*)\right)\right]\right).$$

Further, the maximum likelihood estimate $\theta^* = \max_\theta L(\theta)$ can be found efficiently by solving a convex optimization problem if the parametrization is linear $v(\mathbf{x};\theta) = f(\mathbf{x})^T\theta$.

NOTE 7. *This result was first presented in [11] for the LMDP case and extended to the risk sensitive case in [12]. We restate the result in this section for completeness. This method is computationally more efficient than previous methods for inverse optimal control [1] [43], where one needs to solve the forward optimal control problem repeatedly, which is computationally demanding for large state spaces.*

4.3.4. Characterizing the Most Likely Trajectory under Optimal Control. THEOREM 8. *Suppose the state space $\mathcal{X}$ is finite. Consider the game theoretic interpretation of LCs. The most likely state trajectory of the optimally controlled stochastic system (that includes both the controller and the adversary) starting at $\mathbf{x}_0$ is the one that minimizes the following objective:*

$$\min_{\mathbf{x}_1,\ldots,\mathbf{x}_T} \sum_{t=0}^{T-1} -\log\left(P^0(\mathbf{x}_{t+1}\,|\,\mathbf{x}_t)\right) + (\alpha-1)\left(\sum_{t=0}^{T-1}\ell_{\mathrm{t}}(\mathbf{x}_t) + \ell_{\mathrm{f}}(\mathbf{x}_T)\right)$$

Proof. The optimally controlled system has the dynamics of the adversary (interpretation 2) and is given by $u^*(\mathbf{x};t) = P^0(\mathbf{x}) \otimes z_{t+1}$. So the probability of a trajectory $\mathbf{x}_0,\ldots,\mathbf{x}_T$ is given by

$$\prod_{t=0}^{T-1} \frac{P^0(\mathbf{x}_t)\, z_{t+1}(\mathbf{x}_{t+1})}{\mathrm{E}_{P^0(\mathbf{x}_t)}[z_{t+1}]}$$

From the linear BE we have $\mathrm{E}_{P^0(\mathbf{x})}[z_{t+1}] = \exp((\alpha-1)\,\ell_{\mathrm{t}}(\mathbf{x}))\, z_t(\mathbf{x})$. The above expression then becomes

$$\prod_{t=0}^{T-1} \frac{P^0(\mathbf{x}_{t+1}\,|\,\mathbf{x}_t)\, z_{t+1}(\mathbf{x}_{t+1})}{\exp((\alpha-1)\,\ell_{\mathrm{t}}(\mathbf{x}_t))\, z_t(\mathbf{x}_t)} =$$
$$\frac{1}{z_0(\mathbf{x}_0)}\left(\prod_{t=0}^{T-1} P^0(\mathbf{x}_{t+1}\,|\,\mathbf{x}_t)\exp((1-\alpha)\,\ell_{\mathrm{t}}(\mathbf{x}_t))\right) z_T(\mathbf{x}_T) =$$
$$\frac{1}{z_0(\mathbf{x}_0)}\left(\prod_{t=0}^{T-1} P^0(\mathbf{x}_{t+1}\,|\,\mathbf{x}_t)\exp((1-\alpha)\,\ell_{\mathrm{t}}(\mathbf{x}_t))\right)\exp((\alpha-1)\,\ell_{\mathrm{f}}(\mathbf{x}_T))$$

The most likely trajectory maximizes the RHS above, or equivalently, minimizes its negative logarithm:

$$-\log(z_0(\mathbf{x}_0)) + \sum_{t=0}^{T-1} -\log\left(P^0(\mathbf{x}_{t+1}\,|\,\mathbf{x}_t)\right) + (\alpha-1)\left(\sum_{t=0}^{T-1}\ell_{\mathrm{t}}(\mathbf{x}_t)\right) + \ell_{\mathrm{f}}(\mathbf{x}_T)$$

Since $\mathbf{x}_0$ is fixed, the first term can be dropped from the objective giving the required result. □

NOTE 8. *For the risk neutral case, this result was derived in [36] and the risk sensitive extension was stated without formal proof in [12]. Here we prove it formally for finite state spaces.*

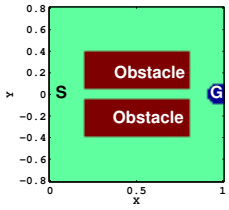
NOTE 9. *The result, appropriately phrased, should extend to non-finite state spaces under suitable assumptions. We do not consider those technicalities here. This result gives us intuition about the behavior of the system under the optimal policies in a game theoretic setting: When $\alpha < 1$, the most likely trajectory minimizes state costs while when $\alpha > 1$, the most likely trajectory maximizes state costs. This corresponds to the fact that when $\alpha < 1$, the controller "wins" the game with high probability while when $\alpha < 1$, the adversary "wins" with high probability. This is reflected in the control policy, for $\alpha < 1$, the controller plans so as to minimize long term costs being slightly conservative to counter effects of the adversary. When $\alpha > 1$, the controller has to give up and settle for minimizing the damage caused by the adversary.*

5. Numerical Examples. In this section, we present numerical examples that illustrate the effect of the parameter α. We demonstrate that changing α can have nontrivial effects on the optimal control policy, and the effects changes depending on the type of control problem being solved.

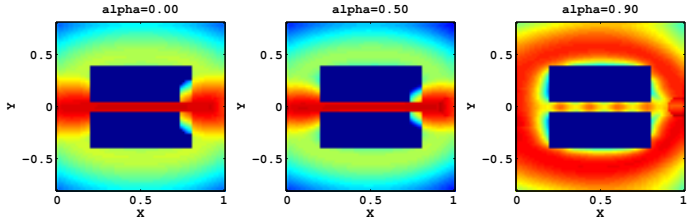
5.1. Crossing a narrow bridge. Consider the problem of driving across a narrow bridge over a pond on a windy day where there is a risk of being blown off of the bridge. The state space consists of the position of the car $\mathbf{x} = (x, y)$ and the controller specifies the velocity (u_x, u_y). The continuous time dynamics of the system is given by

$$\dot{x} = u_x + \sigma \, \mathrm{d}\, \omega, \dot{y} = u_y + \sigma \, \mathrm{d}\, \omega.$$

The state cost reflects a high cost for falling into the pond and a constant small cost elsewhere, that encourages getting to the goal as soon as possible. The situation is depicted in figure 5.1.



(a) Terrain



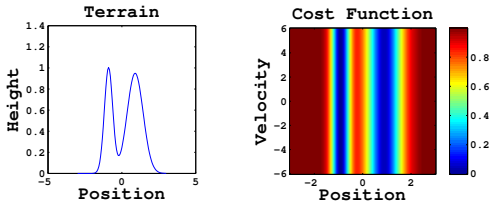
(b) Logarithm of Distribution of states (averaged over time) under optimally control vs α (Red Large, Blue Small)

FIG. 5.1. *Driving across a narrow bridge*

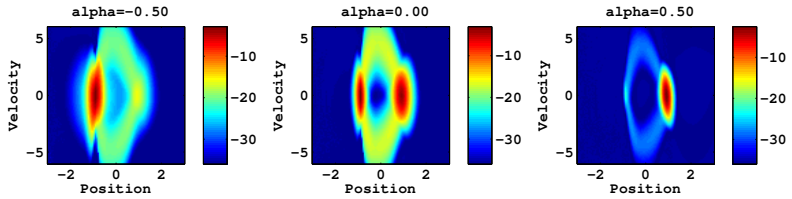
We solve this problem by discretizing time (with time step Δ) to get an LRD. We define a passive dynamics $P^0(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \Delta\sigma^2 I)$. This says that the state remains unchanged on average under 0 control (0 velocity) and is perturbed only by noise. The $\Delta\sigma^2 I$ covariance term reflects the fact that the variance of noise for Brownian motion grows linearly with time. When α is small (risk-neutral), the optimal control strategy is to cross the bridge and get to the other side. However, as α becomes large (risk-seeking), the optimal strategy is to play it safe and go around the pond (figure 5.1b).

5.2. Driving up hills. Consider the problem of driving on a slippery hilly terrain, with a cost function that encourages driving up as high as possible. We consider an example with 2 hills, where one hill is higher but steeper than the other (figure 5.2a). When $\alpha = -.5$ (risk-seeking), the optimally controlled distribution peaks at

the higher/steeper hill. When $\alpha = .5$ (risk-averse), the optimally controlled distribution peaks at the lower/less-steep hill. When $\alpha = 0$ (risk-neutral), the probability is shared between both hills.



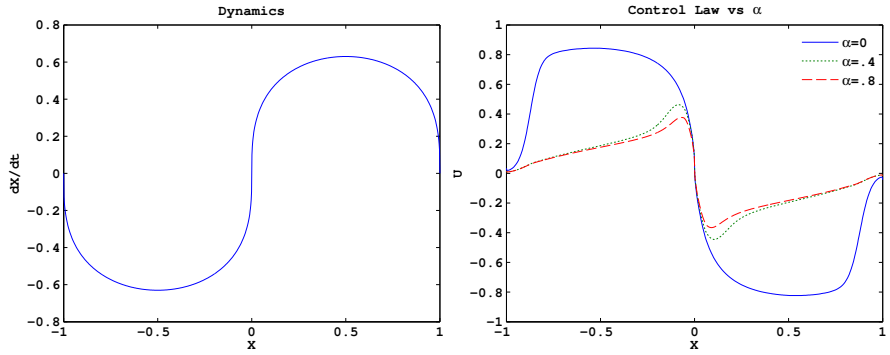
(a) Terrain



(b) Logarithm of Stationary distribution under optimally control vs α (Red Large, Blue Small)

FIG. 5.2. *Driving up hills*

5.3. Stabilizing an unstable system. Consider an unstable system on $\Re$. The continuous time deterministic dynamics is plotted in 5.3a. We consider a system that has this nominal deterministic dynamics plus zero mean Brownian noise: $\mathrm{d}\,\mathbf{x} = f(\mathbf{x}) + \sigma\,\mathrm{d}\,\omega$. We can visualize the solution by constructing a "mean" control law: $u_c(\mathbf{x}) = \left(\mathrm{E}_{\mathbf{x}'\sim u_c{}^*(\mathbf{x})}\left[\mathbf{x}'\right] - \mathrm{E}_{\mathbf{x}'\sim P^0(\mathbf{x})}\left[\mathbf{x}'\right]\right)$. This has the interpretation that the control input is the expected change in the system state due to the controller. The control law constructed this way for various values of α is plotted in figure 5.3b. We see that the controls become smaller as α increases. This is reminiscent of several examples in robust control, where high gains could lead to instability when one has imperfect state estimation and robust control policies tend to have lower gains.



(a) Unstable Dynamics (b) Optimal Control Law vs α

FIG. 5.3. *Controlling an unstable system*

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 paper we presented a unified theory of linearly solvable control, with rigorous derivations of all the results for arbitrary compact Borel state spaces. 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 [29] and character animation [8] 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.

Appendix A. Proofs.

THEOREM 9. *Let* $(\Omega, \mathcal{F}[\Omega], \nu)$ *be a probability space. If* $\alpha > 1$ *and* $f : \Omega \mapsto \Re$ *is a bounded measurable function, then the problem*

$$\begin{aligned} &\min_{\mu} \max_{\mu'} -\mathbb{D}_\alpha(\nu \parallel \mu') + \mathrm{KL}(\mu \parallel \mu') + \mathop{\mathrm{E}}_{\mu}[f] \\ &\textit{Subject to } \ \mu, \mu' \in \mathcal{P}[\Omega] \\ &\qquad\qquad \mu \ll \nu \ll \mu, \mu' \ll \nu \ll \mu' \end{aligned}$$

has optimum $-\Psi_\nu^{\frac{\alpha-1}{\alpha}}[-f]$ *with the optimum attained at* $\mu'^* = \nu \otimes \exp\left(\frac{f}{\alpha}\right), \mu^* = \mu'^* \otimes \exp(-f)$. *The results hold even if the order of the* min, max *are reversed.*

Proof. Let μ, μ' be any measures satisfying the constraints of the problem. Then by lemma 3, we have $\mathbb{D}_\alpha(\nu \parallel \mu') - \mathrm{KL}(\mu \parallel \mu') \geq \frac{\alpha \,\mathrm{KL}(\mu\|\nu)}{1-\alpha}$. This bound is attained when $\mu' = \mu \otimes \left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)^{\frac{\alpha}{1-\alpha}}$ (assuming this measure exists). Also, it is easy to see that this choice for μ' satisfies the constraints. So we have

$$\max_{\mu'} -\mathbb{D}_\alpha(\nu \parallel \mu') + \mathrm{KL}(\mu \parallel \mu') = -\min_{\mu'} \mathbb{D}_\alpha(\nu \parallel \mu') - \mathrm{KL}(\mu \parallel \mu') = \frac{\alpha}{1-\alpha}\,\mathrm{KL}(\mu \parallel \nu)$$

We are left with $\frac{\alpha}{1-\alpha}\,\mathrm{KL}(\mu \parallel \nu) + \mathrm{E}_\mu[f]$, which can be lower bounded using lemma 1:

$$\frac{\alpha}{1-\alpha}\left(\mathrm{KL}(\mu \parallel \nu) + \mathop{\mathrm{E}}_{\mu}\left[\frac{(1-\alpha)f}{\alpha}\right]\right) \geq \frac{\alpha}{\alpha-1}\,\Psi_\nu\left[\frac{(1-\alpha)f}{\alpha}\right] = -\,\Psi_\nu^{\frac{\alpha-1}{\alpha}}[-f]$$

Since f is bounded, the probability measure $\mu = \nu \otimes \exp\left(\frac{(1-\alpha)f}{\alpha}\right)$ exists and attains the lower bound above. Thus, the minimum is attained at the measure $\mu^* = \nu \otimes \exp\left(\frac{(1-\alpha)f}{\alpha}\right)$ which satisfies the constraints and the minimum value is $-\,\Psi_\nu^{\frac{\alpha-1}{\alpha}}[f]$. Given this choice for μ, we can verify that the measure $\mu \otimes \left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)^{\frac{\alpha}{1-\alpha}}$ exists (since f is bounded). Thus, the optimal value of the overall minimax problem is $-\,\Psi_\nu^{\frac{\alpha-1}{\alpha}}[-f]$ and the saddle point is given by $\mu^* = \nu \otimes \exp\left(\frac{(1-\alpha)f}{\alpha}\right), \mu'^* = \mu^* \otimes \left(\frac{\mathrm{d}\mu^*}{\mathrm{d}\nu}\right)^{\frac{\alpha}{1-\alpha}} = \nu \otimes \exp\left(\frac{f}{\alpha}\right)$. This can be rewritten as in the statement of the theorem, $\mu'^* = \nu \otimes \exp\left(\frac{f}{\alpha}\right)$ and $\mu^* = \mu'^* \otimes \exp(-f)$.

If we switch the max and min, we can write $\min_\mu \mathrm{KL}(\mu \parallel \mu') + \mathrm{E}_\mu[f] = -\,\Psi_{\mu'}[-f]$

with the minimum attained at $\mu^* = \mu' \otimes \exp(-f)$ (this measure exists since f is bounded) by lemma 1. We are then left with

$$\max_{\mu'} -\mathbb{D}_\alpha(\nu \parallel \mu') - \Psi_{\mu'}[\exp(-f)] = -\min_{\mu'} \mathbb{D}_\alpha(\nu \parallel \mu') + \Psi_{\mu'}[-f]$$

By lemma 2, we know that $\mathbb{D}_\alpha(\nu \parallel \mu') + \Psi_{\mu'}[-f] \geq -\Psi_\nu^{\frac{1-\alpha}{\alpha}}[f]$ with the minimum attained at $\mu'^* = \nu \otimes \exp\left(\frac{f}{\alpha}\right)$. □

LEMMA 1. *Let $\mu, \nu \in \mathcal{P}[(\Omega, \mathcal{F}[\Omega])], \mu \ll \nu, \nu \ll \mu$ and $f : \Omega \to \Re$ be a bounded measurable function. Then,* $\mathrm{KL}(\mu \parallel \nu) + \mathrm{E}_\mu[f] \geq -\Psi_\nu[-f]$.

Proof. The objective can be rewritten as

$$\mathop{\mathrm{E}}_{\mu}\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) + f\right] = \mathop{\mathrm{E}}_{\mu}\left[-\log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right) + f\right] = \mathop{\mathrm{E}}_{\mu}\left[-\log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\exp(-f)\right)\right].$$

By Jensen's inequality, since $-\log$ is convex, the RHS is larger than $-\log\left(\mathrm{E}_\mu\left[\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\exp(-f)\right]\right) = -\log\left(\mathrm{E}_\nu[\exp(-f)]\right)$, establishing the result. □

LEMMA 2. *Let $\mu, \nu \in \mathcal{P}[(\Omega, \mathcal{F}[\Omega])], \mu \ll \nu, \nu \ll \mu$ and $f : \Omega \to \Re$ be a bounded measurable function. Then,*

$$\operatorname{sgn}(\alpha)\,\mathbb{D}_\alpha(\nu \parallel \mu) + \Psi_\mu[f] \geq \Psi_\nu^{\frac{\alpha-1}{\alpha}}[f] \;\; \textit{if } \alpha > 0$$

$$\operatorname{sgn}(\alpha)\,\mathbb{D}_\alpha(\nu \parallel \mu) + \Psi_\mu[f] \leq \Psi_\nu^{\frac{\alpha-1}{\alpha}}[f] \;\; \textit{if } \alpha > 0$$

Proof. Letting $g = \exp(f)$, the LHS can be rewritten as

$$\frac{\log\left(\mathrm{E}_\nu\left[\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)^{1-\alpha}\right]\right)}{\alpha - 1} + \log\left(\mathop{\mathrm{E}}_{\mu}[g]\right) = \log\left(\left(\mathop{\mathrm{E}}_{\nu}\left[\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)^{1-\alpha}\right]\right)^{\frac{1}{\alpha-1}} \mathop{\mathrm{E}}_{\nu}\left[\frac{\mathrm{d}\mu}{\mathrm{d}\nu}g\right]\right).$$

First suppose $\alpha < 0$. Then, using Holder's inequality, we have

$$\mathop{\mathrm{E}}_{\nu}\left[\frac{\mathrm{d}\mu}{\mathrm{d}\nu}g\right] \leq \left(\mathop{\mathrm{E}}_{\nu}\left[\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)^{1-\alpha}\right]\right)^{\frac{1}{1-\alpha}} \left(\mathop{\mathrm{E}}_{\nu}\left[(g)^{\frac{1-\alpha}{-\alpha}}\right]\right)^{\frac{-\alpha}{1-\alpha}} \implies$$

$$\mathop{\mathrm{E}}_{\nu}\left[\frac{\mathrm{d}\mu}{\mathrm{d}\nu}g\right]\left(\mathop{\mathrm{E}}_{\nu}\left[\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)^{1-\alpha}\right]\right)^{\frac{1}{\alpha-1}} \leq \left(\mathop{\mathrm{E}}_{\nu}\left[(g)^{\frac{1-\alpha}{-\alpha}}\right]\right)^{\frac{-\alpha}{1-\alpha}} = \left(\mathop{\mathrm{E}}_{\nu}\left[\exp\left(\frac{(\alpha-1)f}{\alpha}\right)\right]\right)^{\frac{\alpha}{\alpha-1}}.$$

Taking log on both sides, we have the result. The measure $\nu \otimes (g)^{\frac{\alpha-1}{\alpha}}$ exists since f is bounded. Also note that $h(x) = x^{\frac{1}{1-\alpha}}$ is convex if $\alpha > 0, \alpha \neq 1$. The first term inside the log can be bounded as follows:

$$\begin{aligned}
\left(\mathop{\mathrm{E}}_{\nu}\left[\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)^{1-\alpha}\right]\right)^{\frac{1}{1-\alpha}} &= \left(\mathop{\mathrm{E}}_{\nu \otimes g^{\frac{\alpha-1}{\alpha}}}\left[g^{\frac{1-\alpha}{\alpha}}\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)^{1-\alpha}\right]\right)^{\frac{1}{1-\alpha}} \left(\mathop{\mathrm{E}}_{\nu}\left[g^{\frac{\alpha-1}{\alpha}}\right]\right)^{\frac{1}{1-\alpha}} \\
&\leq \mathop{\mathrm{E}}_{\nu \otimes g^{\frac{\alpha-1}{\alpha}}}\left[g^{\frac{1}{\alpha}}\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right]\left(\mathop{\mathrm{E}}_{\nu}\left[g^{\frac{\alpha-1}{\alpha}}\right]\right)^{\frac{1}{1-\alpha}} \text{ (Jensen's Inequality)} \\
&= \mathop{\mathrm{E}}_{\nu}\left[g^{\frac{\alpha-1}{\alpha}+\frac{1}{\alpha}}\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right]\left(\mathop{\mathrm{E}}_{\nu}\left[g^{\frac{\alpha-1}{\alpha}}\right]\right)^{\frac{1}{1-\alpha}-1} = \mathop{\mathrm{E}}_{\nu}\left[g\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right]\left(\mathop{\mathrm{E}}_{\nu}\left[g^{\frac{\alpha-1}{\alpha}}\right]\right)^{\frac{\alpha}{1-\alpha}}
\end{aligned}$$

Rewriting the last inequality, we get $\left(\mathrm{E}_\nu\left[g^{\frac{\alpha-1}{\alpha}}\right]\right)^{\frac{\alpha}{\alpha-1}} \leq \mathrm{E}_\mu[g]\left(\mathrm{E}_\nu\left[\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)^{1-\alpha}\right]\right)^{\frac{1}{1-\alpha}}$. Taking log on both sides gives the result.
□

LEMMA 3. *Let $\mu, \mu', \nu \in \mathcal{P}[(\Omega, \mathcal{F}[\Omega])], \mu' \ll \nu, \nu \ll \mu', \mu \ll \mu', \mu' \ll \mu$ and $f : \Omega \to \Re$ be a bounded measurable function, $\alpha > 1$. Then, $\mathbb{D}_\alpha(\nu \parallel \mu') - \mathrm{KL}(\mu \parallel \mu') \geq \frac{\alpha}{1-\alpha}\mathrm{KL}(\mu \parallel \nu)$.*

Proof. By definition,

$$\mathrm{KL}(\mu \parallel \mu') = \mathop{\mathrm{E}}_{\mu}\left[-\log\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\right)\right], \mathbb{D}_\alpha(\nu \parallel \mu') = \frac{\log\left(\mathrm{E}_\nu\left[\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\nu}\right)^{1-\alpha}\right]\right)}{\alpha-1}.$$

Now, $\mathrm{E}_\nu\left[\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\nu}\right)^{1-\alpha}\right] = \mathrm{E}_\mu\left[\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\nu}\right)^{1-\alpha}\right] = \mathrm{E}_\mu\left[\frac{\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\nu}\right)^{1-\alpha}}{\frac{\mathrm{d}\mu'}{\mathrm{d}\nu}}\right] = \mathrm{E}_\mu\left[\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\nu}\right)^{-\alpha}\right]$.
This gives us

$$\mathbb{D}_\alpha(\nu \parallel \mu') - \mathrm{KL}(\mu \parallel \mu') = \frac{\log\left(\mathrm{E}_\mu\left[\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\nu}\right)^{-\alpha}\right]\right)}{\alpha-1} + \mathop{\mathrm{E}}_{\mu}\left[\log\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\right)\right]$$

When $\alpha > 1$, $\frac{\log}{\alpha-1}$ is concave and by Jensen's inequality the first term is $\geq \frac{\mathrm{E}_\mu\left[\log\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\nu}\right)^{-\alpha}\right)\right]}{\alpha-1} = \frac{\mathrm{E}_\mu\left[-\alpha\log\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\nu}\right)+\log\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\right)\right]}{\alpha-1}$. This implies:

$$\mathbb{D}_\alpha(\nu \parallel \mu') - \mathrm{KL}(\mu \parallel \mu') \geq \frac{\alpha}{\alpha-1}\mathop{\mathrm{E}}_{\mu}\left[\log\left(\frac{\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}}{\frac{\mathrm{d}\mu'}{\mathrm{d}\nu}}\right)\right] = \frac{\alpha}{\alpha-1}\mathop{\mathrm{E}}_{\mu}\left[\log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)\right] = \frac{\alpha\,\mathrm{KL}(\mu \parallel \nu)}{1-\alpha}.$$

□

REFERENCES

[1] P. Abbeel and A.Y. Ng, *Apprenticeship learning via inverse reinforcement learning*, in Proceedings of the twenty-first international conference on Machine learning, ACM, 2004, p. 1.

[2] T. Başar and G.J. Olsder, *Dynamic noncooperative game theory*, Society for Industrial Mathematics, 1999.

[3] R.E. Bellman and Rand Corporation, *Dynamic programming*, Rand Corporation research study, Princeton University Press, 1957.

[4] B. Van Den Broek, W. Wiegerinck, and B. Kappen, *Risk sensitive path integral control*, in Uncertainty in AI, 2010. Proceedings of the 2010, 2010.

[5] J. Broek, W. Wiegerinck, and Kappen H., *Stochastic optimal control of state constrained systems*, International Journal of Control, (2011), pp. 1–9.

[6] J. Broek, W. Wiegerinck, and H. Kappen, *Risk sensitive path integral control*, Uncertainty in Artificial Intelligence, (2010).

[7] R. Cavazos-Cadena, *Optimality equations and inequalities in a class of risk-sensitive average cost markov decision chains*, Mathematical Methods of Operations Research, 71 (2010), pp. 47–84.

[8] M. Da Silva, F. Durand, and J. Popović, *Linear Bellman combination for control of character animation*, ACM Transactions on Graphics (TOG), 28 (2009), pp. 1–10.

[9] G.B. Di Masi and L. Stettner, *Risk-sensitive control of discrete-time markov processes with infinite horizon*, SIAM Journal on Control and Optimization, 38 (1999), p. 61.

[10] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, Wiley, New York, NY, 1997.

[11] K. Dvijotham and E. Todorov, *Inverse optimal control with linearly solvable MDPs*, International Conference on Machine Learning, (2010).
[12] Krishnamurthy Dvijotham and Emanuel Todorov, *A unifying framework for linearly solvable control*, in UAI, Fabio Gagliardi Cozman and Avi Pfeffer, eds., AUAI Press, 2011, pp. 179–186.
[13] W.H. Fleming, *Exit probabilities and optimal stochastic control*, Applied Mathematics & Optimization, 4 (1977), pp. 329–346.
[14] W. Fleming and W. McEneaney, *Risk sensitive optimal control and differential games*, Stochastic Theory and Adaptive Control, (1992), pp. 185–197.
[15] W. Fleming and S. Mitter, *Optimal control and nonlinear filtering for nondegenerate diffusion processes*, Stochastics, 8 (1982), pp. 226–261.
[16] W.H. Fleming and H.M. Soner, *Controlled Markov processes and viscosity solutions*, vol. 25, Springer Verlag, 2006.
[17] M. K. Ghosh and A. Bagchi, *Stochastic games with average payoff criterion*, Applied Mathematics & Optimization, 38 (1998), pp. 283–301. 10.1007/s002459900092.
[18] C. Holland, *A new energy characterization of the smallest eigenvalue of the Schrödinger equation*, Comm Pure Appl Math, 30 (1977), pp. 755–765.
[19] E. Hopf, *The partial differential equation ut+ uux= μxx*, Communications on Pure and Applied mathematics, 3 (1950), pp. 201–230.
[20] H. Kappen, *Linear theory for control of nonlinear stochastic systems*, Physical Review Letters, 95 (2005).
[21] M.A. Krasnoselski, E.A. Lifshits, and A.V. Sobolev, *Positive linear systems: the method of positive operators*, vol. 5, Heldermann, 1989.
[22] H.J. Kushner and P. Dupuis, *Numerical methods for stochastic control problems in continuous time*, vol. 24, Springer Verlag, 2001.
[23] 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).
[24] T. Mensink, J. Verbeek, and H. Kappen, *EP for efficient stochastic control with obstacles*, ECAI, (2010).
[25] S. Mitter and N. Newton, *A variational approach to nonlinear estimation*, SIAM J Control Opt, (2003), pp. 1813–1833.
[26] B.K. Øksendal, *Stochastic differential equations: an introduction with applications*, Springer Verlag, 2003.
[27] A. Renyi, *On measures of entropy and inf*, in Proc. 4th Berkley symp. Stat. and Prob, vol. 1, pp. 541–561.
[28] L.S. Shapley, *Stochastic games*, Proceedings of the National Academy of Sciences of the United States of America, 39 (1953), p. 1095.
[29] E. Theodorou, J. Buchli, and S. Schaal, *Learning policy improvements with path integrals*, AISTATS, (2010).
[30] E. A. Theodorou, *Iterative Path Integral Stochastic Optimal Control: Theory and Applications to Motor Control*, PhD thesis, University of Southern California, 2011.
[31] E. Todorov, *Linearly-solvable Markov decision problems*, Advances in Neural Information Processing Systems, (2006).
[32] Emanuel Todorov, *General duality between optimal control and estimation*, in CDC, IEEE, 2008, pp. 4286–4292.
[33] E. Todorov, *Compositionality of optimal control laws*, in NIPS, Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, eds., Curran Associates, Inc., 2009, pp. 1856–1864.
[34] ———, *Efficient computation of optimal actions*, PNAS, 106 (2009), pp. 11478–11483.
[35] Emanuel Todorov, *Policy gradients in linearly-solvable mdps*, in NIPS, John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta, eds., Curran Associates, Inc., 2010, pp. 2298–2306.
[36] E. Todorov, *Finding the most likely trajectories of optimally-controlled stochastic systems*, in World Congress of the International Federation of Automatic Control (IFAC), 2011.
[37] M. Toussaint, *Robot trajectory optimization using approximate inference*, International Conference on Machine Learning, 26 (2009), pp. 1049–1056.
[38] Tim van Erven, *When Data Compression and Statistics Disagree*, PhD thesis, CWI, 2010.
[39] J. Von Neumann, O. Morgenstern, A. Rubinstein, and H.W. Kuhn, *Theory of games and economic behavior*, Princeton Univ Pr, 2007.
[40] 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).
[41] M. Zhong and E. Todorov, *Aggregation methods for linearly-solvable MDPs*, IFAC World Congress, (2011).
[42] ———, *Moving least-squares approximations for linearly-solvable stochastic optimal control problems*, Journal of Control Theory and Applications, (2011).
[43] B.D. Ziebart, A. Maas, J.A. Bagnell, and A.K. Dey, *Maximum entropy inverse reinforcement learning*, in Proc. AAAI, 2008, pp. 1433–1438.