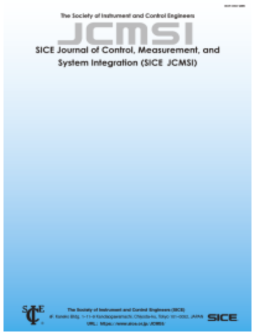


SICE Journal of Control, Measurement, and System Integration

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tmsi20

Kullback–Leibler control for discrete-time nonlinear systems on continuous spaces

Kaito Ito & Kenji Kashima

To cite this article: Kaito Ito & Kenji Kashima (2022) Kullback–Leibler control for discrete-time nonlinear systems on continuous spaces, SICE Journal of Control, Measurement, and System Integration, 15:2, 119-129, DOI: 10.1080/18824889.2022.2095827

To link to this article: https://doi.org/10.1080/18824889.2022.2095827

SICE JOURNAL OF CONTROL, MEASUREMENT, AND SYSTEM INTEGRATION
2022, VOL. 15, NO. 2, 119–129
https://doi.org/10.1080/18824889.2022.2095827

Kullback–Leibler control for discrete-time nonlinear systems on continuous spaces

Kaito Ito and Kenji Kashima

Graduate School of Informatics, Kyoto University, Kyoto, Japan

ABSTRACT

Kullback–Leibler (KL) control enables efficient numerical methods for nonlinear optimal control problems. The crucial assumption of KL control is the full controllability of transition distributions. However, this assumption is often violated when the dynamics evolves in a continuous space. Consequently, applying KL control to problems with continuous spaces requires some approximation, which leads to the loss of the optimality. To avoid such an approximation, in this paper, we reformulate the KL control problem for continuous spaces so that it does not require unrealistic assumptions. The key difference between the original and reformulated KL control is that the former measures the control effort by the KL divergence between controlled and *uncontrolled* transition distributions while the latter replaces the uncontrolled transition by a *noise-driven* transition. We show that the reformulated KL control admits efficient numerical algorithms like the original one without unreasonable assumptions. Specifically, the associated value function can be computed by using a Monte Carlo method based on its path integral representation.

ARTICLE HISTORY
Received 18 January 2022
Revised 4 May 2022
Accepted 13 June 2022

KEYWORDS
Optimal control; Markov decision processes; discrete-time nonlinear systems

1. Introduction

Optimal control theory is a powerful mathematical tool for achieving control objectives while considering, for example, energy efficiency and sparsity of control [1,2]. Optimal control problems arise in a variety of physical, biological, and economic systems, to name a few. Recently, optimal control has also become increasingly important in machine learning [3,4]. It is well known that finding an optimal feedback control law boils down to solving the (Hamilton–Jacobi) Bellman equation [5,6], which suffers from the curse of dimensionality and is difficult to solve in general.

In [7,8], a special class of stochastic optimal control problems was introduced in which the associated Bellman equation can be converted into a linear equation resulting in efficient numerical methods. For continuous state/input spaces and continuous time, the work [7] considers a control-affine diffusion with a quadratic control cost and assumes the noise and control act in the same subspace. Then, the optimal control admits a path integral representation, which can be approximated by forward sampling of an uncontrolled diffusion process. This stochastic control framework is called a path integral control and has many applications, e.g. reinforcement learning [9,10], model predictive control [11], multi-agent systems [12], controllability quantification [13].

For discrete-time cases, the work [8] deals with general dynamics and makes the key assumptions as follows: (A1) the controller can change the distribution of the next state given the current state as desired; (A2) the control cost is quantified by the Kullback–Leibler (KL) divergence between the controlled and uncontrolled state distributions. This formulation is referred to as linearly solvable Markov decision processes (MDPs) or KL control. The KL control framework shares nice properties with the path integral control including a path integral representation of the KL optimal control [14], compositionality of optimal control laws [15], and duality with Bayesian inference [16]. For the connection between the path integral control and KL control, see [17]. Moreover, the special structure of KL control enables the convex formulation of inverse reinforcement learning [18].

However, it should be emphasized that the assumption (A1) of KL control is too restrictive in practice, especially for continuous state spaces. Indeed, as mentioned in [19], even for discrete-time linear systems driven by Gaussian noise, (A1) is violated because the variance of the one step transition distribution given the current state is uncontrollable under the causality of controllers. Therefore, applying KL control to practical problems with continuous spaces requires some approximation, which leads to the loss of the optimality. For instance, if the system of interest is derived from the Euler–Maruyama discretization of a control-affine diffusion, using a smaller time step results in the smaller approximation error [20,21]. However, to the best of

CONTACT Kaito Ito ito.kaito@bode.amp.i.kyoto-u.ac.jp

our knowledge, there is no discussion of approximation in other cases, e.g. the system originally evolves in discrete time.

Contributions: In this paper, we reformulate KL control for continuous state spaces so that its assumption is more realistic than the conventional formulation of KL control. This enables us to apply KL control to discrete-time and continuous space problems without any approximation of dynamics. As a byproduct, we reconsider what the assumption (A1) implies for practical problems. Specifically, we clarify that (A1) essentially requires the controller to know the value of noise to be injected to the system together with control inputs. Moreover, we show that our KL control formulation enjoys the nice properties which the original one has as mentioned above.

Organization: The remainder of this paper is organized as follows. In Section 2, we briefly review KL control. In Section 3, we reformulate KL control for continuous spaces. Section 4 is devoted to the general analysis of the reformulated KL control. In Section 5, we focus on linear systems with a quadratic state cost. In Section 6, numerical examples are presented. Some concluding remarks are given in Section 7.

Notation: Let $\mathbb{R}$ denote the set of real numbers and $\mathbb{Z}_{>0}$ (resp. $\mathbb{Z}_{\geq 0}$) denote the set of positive (resp. non-negative) integers. The set of integers $\{0, 1, \dots, N\}$ is denoted by $[\![N]\!]$. The identity matrix is denoted by I, and its dimension depends on the context. For symmetric matrices $A, B \in \mathbb{R}^{n\times n}$, we write $A \succ B$ if $A - B$ is positive definite. The determinant of a square matrix A is denoted by $\det(A)$. The block diagonal matrix with diagonal entries $\{A_i\}_{i=1}^N, A_i \in \mathbb{R}^{m\times n}$ is denoted by $\mathrm{diag}(A_1, \dots, A_N)$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space where $\mathcal{F}$ is the σ-field on Ω, and $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a probability measure. The space $(\Omega, \mathcal{F}, \mathbb{P})$ is equipped with a natural filtration $\{\mathcal{F}_k\}_{k\geq 0}$. The expectation is denoted by $\mathbb{E}$. The probability density function of a continuous random variable x with respect to the Lebesgue measure is denoted by ρ_x. The support of the density function ρ_x is defined as the smallest closed set S such that $\mathbb{P}(x \in S) = 1$. The conditional density of x given $y = \mathbf{y}$ is denoted by $\rho_{x|y}(\cdot|\mathbf{y})$. Denote by $D_{\mathrm{KL}}\left(\rho_x \| \rho_y\right)$ the KL divergence between probability densities ρ_x and ρ_y. The Dirac delta function is denoted by $\delta(\cdot)$. For an $\mathbb{R}^n$-valued random vector w, $w \sim \mathcal{N}(\mu, \Sigma)$ means that w has a multivariate Gaussian distribution with mean $\mu \in \mathbb{R}^n$ and covariance matrix Σ. When $\Sigma \succ 0$, the density function of $w \sim \mathcal{N}(\mu, \Sigma)$ is denoted by $\mathcal{N}(\cdot|\mu, \Sigma)$.

2. Brief introduction of KL control

Here, we briefly review KL control [8]. Let $\mathbb{X} \subseteq \mathbb{R}^n$ be a state space and $\mathbb{U} \subseteq \mathbb{R}^m$ an input space. Throughout the paper, $\mathbb{X}$ and $\mathbb{U}$ are assumed to be Borel measurable. Consider an MDP with a transition density function $\rho_{x_{k+1}|x_k,u_k}$ where $\{x_k\}$ is an $\mathbb{X}$-valued state process and $\{u_k\}$ is a $\mathbb{U}$-valued control process. In this section, we implicitly assume the existence of probability density functions. Nevertheless, we can apply the same argument for discrete random variables by replacing densities by probabilities. Let ρ_{x_0} be the density of the initial state x_0. Denote by $\rho^{\pi_k}_{k+1}(\cdot|x)$ the conditional density of x_{k+1} given $x_k = x$ induced by a stochastic policy (control law) $\pi_k(\cdot|x) := \rho_{u_k|x_k}(\cdot|x)$. Let $\rho^0_{k+1}(\cdot|x) := \rho_{x_{k+1}|x_k,u_k}(\cdot|x, 0)$ be the transition density for the uncontrolled dynamics. Note that when we assume the existence of the density functions $\rho^{\pi_k}_{k+1}(\cdot|x), \pi_k(\cdot|x)$, the state space $\mathbb{X}$ and the input space $\mathbb{U}$ must have positive Lebesgue measure in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. Hence, we assume that $\mathbb{X}$ and $\mathbb{U}$ have positive measure when dealing with densities. Then, the KL control problem is formulated as follows.

Problem 2.1: Find a policy $\pi = \{\pi_k\}_{k=0}^{N-1}$ that solves

$$\underset{\pi}{\text{minimize}} \quad \mathbb{E}\left[\ell_N(x_N) + \sum_{k=0}^{N-1}\{\ell_k(x_k) \right.$$
$$\left. + D_{\mathrm{KL}}\left(\rho^{\pi_k}_{k+1}(\cdot|x_k)\|\rho^0_{k+1}(\cdot|x_k)\right)\}\right], \tag{1}$$

where $\ell_k : \mathbb{R}^n \to \mathbb{R}$ is the running cost ($k = 0, \dots, N-1$) and terminal cost ($k = N$) for the state, respectively, and $N \in \mathbb{Z}_{>0}$ is the terminal time. ♢

Here, we assume that the infimum of (1) is finite. The KL divergence measures the difference between two probability distributions. Hence, Problem 2.1 penalizes the deviation of the transition density $\rho^{\pi_k}_{k+1}(\cdot|x_k)$ from the uncontrolled transition density $\rho^0_{k+1}(\cdot|x_k)$. Denote the support of $\rho^0_{k+1}(\cdot|x)$ by $\mathbb{X}^0_{x,k+1}$.

Now, we introduce the most important assumption of KL control.

Assumption 2.2: For any $(k, x) \in [\![N-1]\!] \times \mathbb{X}$ and any density $\check{\rho}$ whose support is given by $\mathbb{X}^0_{x,k+1}$, there exists a policy π_k such that $\check{\rho}(x') = \rho^{\pi_k}_{k+1}(x'|x)$ for almost all $x' \in \mathbb{X}^0_{x,k+1}$.

The above assumption says that the controller can change the transition density $\rho^{\pi_k}_{k+1}(\cdot|x)$ as desired. Under this assumption, the Bellman equation for (1) becomes linear by an exponential transformation:

$$z(k, x) = \exp(-\ell_k(x))\mathcal{A}_{\rho^0_{k+1}}[z](k, x),$$
$$(k, x) \in [\![N-1]\!] \times \mathbb{X}, \tag{2}$$
$$z(N, x) = \exp(-\ell_N(x)),\ x \in \mathbb{X}, \tag{3}$$

where $\mathcal{A}_{\rho^0_{k+1}}[z](k, x) := \int_{\mathbb{X}} z(k+1, x')\rho^0_{k+1}(x'|x)\mathrm{d}x'$. The solution of (2), (3) is given by the so-called desirability function $z(k, x) := \exp(-v(k, x))$, and the value function v associated with (1) is defined by

$$v(k,x) := \inf_{\{\pi_s\}_{s=k}^{N-1}} \mathbb{E}\left[\ell_N(x_N) + \sum_{s=k}^{N-1} \left\{ \ell_s(x_s) + D_{\mathrm{KL}}\left(\rho_{s+1}^{\pi_s}(\cdot|x_s) \| \rho_{s+1}^{0}(\cdot|x_s) \right) \right\} \middle| \; x_k = x \right],$$

$$(k,x) \in [\![N-1]\!] \times \mathbb{X},$$

$$v(N,x) := \ell_N(x), \; x \in \mathbb{X}.$$

In particular, a policy $\{\pi_k^*\}$ satisfying

$$\rho_{k+1}^{\pi_k^*}(x'|x) = \frac{\rho_{k+1}^{0}(x'|x) z(k+1,x')}{\mathcal{A}_{\rho_{k+1}^0}[z](k,x)}, \quad \forall x', x \in \mathbb{X}, \; \forall k \in [\![N-1]\!] \tag{4}$$

is an optimal policy of Problem 2.1, and its existence is ensured by Assumption 2.2. It is remarkable that an optimal transition density can be written analytically given the desirability function unlike the conventional MDPs [5]. However, as mentioned in the Introduction, Assumption 2.2 is typically violated for continuous state spaces, and there is no policy satisfying (4). To see this, as an example, we consider a linear system driven by Gaussian noise:

$$x_{k+1} = Ax_k + Bu_k + w_k, \quad w_k \sim \mathcal{N}(0, \Sigma), \tag{5}$$

where $\{w_k\}$ is an independent sequence and $\Sigma \succ 0$. Then, the transition density $\rho_{k+1}^{\pi_k}(\cdot|x)$ cannot be shaped to the density $\mathcal{N}(\cdot|Ax, \Sigma')$ where $\Sigma \succ \Sigma'$. This is because any causal controller π_k cannot decrease the variance of x_{k+1} due to the noise w_k. A similar argument applies to general nonlinear stochastic systems. That is, it is impossible to decrease the uncertainty of the state at time $k+1$ due to the noise at time k. As a result, the assumption of controllability of the transition distributions is violated in general, and the Bellman equation cannot be linearized.

Then, does the KL control framework work well for deterministic systems without probabilistic uncertainty? To answer this, let us consider a general nonlinear system of the form:

$$x_{k+1} = f(x_k, u_k), \quad f : \mathbb{X} \times \mathbb{U} \to \mathbb{X}.$$

Then, the reference distribution $\rho_{k+1}^0(x'|x)$ is given by the uncontrolled transition distribution $\delta(x' - f(x,0))$. Hence, to make the cost $D_{\mathrm{KL}}(\rho_{k+1}^{\pi_k}(\cdot|x_k) \| \rho_{k+1}^0(\cdot|x_k))$ finite, we are only allowed to choose the trivial transition distribution $\rho_{k+1}^{\pi_k}(x'|x_k) = \delta(x' - f(x_k, 0))$. Otherwise, the KL divergence diverges to infinity. Of course, the above situation is meaningless.

3. Reformulation of KL control for continuous spaces

In the previous section, we have observed that the KL control framework has severe problems for both stochastic and deterministic general systems. Then, are there situations in which the linearization of the Bellman equation is possible and for which a meaningful solution exists? In this section, we answer this question. The same notation as in Section 2 is employed. We consider general deterministic nonlinear systems of the form:

$$x_{k+1} = f(x_k, u_k), \quad k \in \mathbb{Z}_{\geq 0}, \tag{6}$$

$$x_0 \sim \rho_{x_0}, \tag{7}$$

where $\{x_k\}$ is an $\mathbb{X}$-valued state process, $\{u_k\}$ is a $\mathbb{U}$-valued control process, and $f : \mathbb{X} \times \mathbb{U} \to \mathbb{X}$. The extension of the results in this paper to the time-varying case $x_{k+1} = f_k(x_k, u_k)$ is straightforward. Next, we introduce the associated noise-driven dynamics:

$$\bar{x}_{k+1} = f(\bar{x}_k, w_k), \quad k \in \mathbb{Z}_{\geq 0}, \tag{8}$$

$$\bar{x}_0 \sim \rho_{x_0}, \tag{9}$$

where $\{w_k\}$ is a sequence of independent (not necessarily identically distributed) random variables, and w_k has the density function ρ_{w_k} with the support $\mathbb{W}$. Denote the conditional density of $\bar{x}_{k+1}$ given $\bar{x}_k = x$ by $\bar{\rho}_{k+1}(\cdot|x)$. Now, we are ready to state our problem.

Problem 3.1: Find a policy $\pi = \{\pi_k\}_{k=0}^{N-1}$ that solves

$$\begin{aligned} \underset{\pi}{\text{minimize}} \quad & \mathbb{E}\left[\ell_N(x_N) + \sum_{k=0}^{N-1} \left\{ \ell_k(x_k) + D_{\mathrm{KL}}\left(\rho_{k+1}^{\pi_k}(\cdot|x_k) \| \bar{\rho}_{k+1}(\cdot|x_k) \right) \right\} \right] \\ \text{subject to} \quad & \text{(6) and (7).} \end{aligned} \tag{10}$$

We emphasize that Problem 3.1 employs *noise-driven* dynamics ($u_k = w_k$) as a reference transition density $\bar{\rho}_{k+1}(\cdot|x_k)$ while Problem 2.1 employs *uncontrolled* dynamics ($u_k = 0$). This allows us to choose $\rho_{k+1}^{\pi_k}(\cdot|x)$ other than a delta function despite the deterministic dynamics (6). Consequently, Problem 3.1 admits a non-trivial solution. Note that for a deterministic policy $u_k = K(x_k)$, i.e. $\pi_k(u|x) = \delta(u - K(x))$, $D_{\mathrm{KL}}\left(\rho_{k+1}^{\pi_k}(\cdot|x_k) \| \bar{\rho}_{k+1}(\cdot|x_k)\right)$ is infinite because $\rho_{k+1}^{\pi_k}(\cdot|x_k)$ is not absolutely continuous with respect to $\bar{\rho}_{k+1}(\cdot|x_k)$. Therefore, an optimal policy for Problem 3.1 must be stochastic. This is in contrast to the conventional optimal control problems without the KL divergence cost whose optimal policy is deterministic [5].

For $x \in \mathbb{X}$, let $f_x(u) := f(x, u)$ and $\mathbb{X}_x := \{f_x(u) : u \in \mathbb{U}\}$. Recall that since we are dealing with densities, we assume that $\mathbb{X}$ and $\mathbb{U}$ have positive Lebesgue measure in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. In addition, we assume the following conditions.

Assumption 3.2: (i) $\mathbb{W} \subseteq \mathbb{U}$;
(ii) $m = n$;
(iii) For all $x \in \mathbb{X}$, $f_x : \mathbb{U} \to \mathbb{X}_x$ is bijective, and it and its inverse are both continuously differentiable;
(iv) For all $k \in [\![N]\!]$, $\mathbb{E}[\ell_k(\bar{x}_k)]$ is finite.

Assumptions 3.2-(ii),(iii) with the fact that u_k given $x_k = x$ has the density ensure the existence of the density $\rho_{k+1}^{\pi_k}(\cdot|x)$; see, e.g. [22, Chapter 6, Theorem 5]. Note that when $\mathbb{U}$ is an open set, Assumption 3.2-(iii) means that for all $x \in \mathbb{X}$, f_x is a C^1-diffeomorphism. In addition, Assumption 3.2-(i) implies that there exists a feasible control process that replicates a given noise process. Consequently, the transition density $\rho_{k+1}^{\pi_k}(\cdot|x)$ can be shaped to a desired form with the support $\bar{\mathbb{X}}_x := \{f_x(w) : w \in \mathbb{W}\}$ by an appropriate policy; see the proof of Theorem 4.1 in the next section. Conversely, if the set difference $\mathbb{W} \backslash \mathbb{U}$ has positive Lebesgue measure, then the support of $\rho_{k+1}^{\pi_k}(\cdot|x)$ cannot be $\bar{\mathbb{X}}_x$ for any π_k under the injectivity of $f(x, w)$ with respect to $w \in \mathbb{U} \cup \mathbb{W}$. Therefore, Assumption 3.2-(i) with the injectivity can be seen as playing the same role as Assumption 2.2. Lastly, Assumption 3.2-(iv) is a technical assumption that ensures there exists a policy that makes (10) finite. For instance, if ℓ_k is bounded for all $k \in [\![N]\!]$, Assumption 3.2-(iv) is satisfied.

Remark 3.1: Consider the control-affine case $f(x, u) = \bar{f}(x) + g(x)u$ where $\bar{f} : \mathbb{R}^n \to \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}^{n\times m}$. Then, Assumptions 3.2-(ii),(iii) imply that, for all $x \in \mathbb{R}^n$, $g(x)$ is square and invertible. Note that when $n < m$ and $g(x)$ has full row rank for all $x \in \mathbb{R}^n$, we can introduce an auxiliary system

$$\tilde{x}_{k+1} = \tilde{f}(\tilde{x}_k) + \tilde{g}(\tilde{x}_k)u_k \tag{11}$$

where $\tilde{x}_k \in \mathbb{R}^{m-n}$, $\tilde{f} : \mathbb{R}^{m-n} \to \mathbb{R}^{m-n}$, $\tilde{g} : \mathbb{R}^{m-n} \to \mathbb{R}^{(m-n)\times m}$, such that the combined system

$$\begin{bmatrix} x_{k+1} \\ \tilde{x}_{k+1} \end{bmatrix} = \begin{bmatrix} \bar{f}(x_k) \\ \tilde{f}(\tilde{x}_k) \end{bmatrix} + \begin{bmatrix} g(x_k) \\ \tilde{g}(\tilde{x}_k) \end{bmatrix} u_k \tag{12}$$

satisfies Assumptions 3.2-(ii),(iii). That is, $[g(x)^\top \ \tilde{g}(\tilde{x})^\top]^\top$ is invertible for all $[x^\top \ \tilde{x}^\top]^\top \in \mathbb{R}^m$. When the state cost function ℓ_k does not depend on $\tilde{x}_k$, the introduction of the auxiliary system (11) is explicitly relevant only for the KL divergence cost of (10).

4. General analysis of KL control for continuous spaces

In this section, we characterize the value function and the optimal control of Problem 3.1 and then reconsider the implication of Assumption 2.2 for Problem 2.1.

4.1. Characterization of the value function and optimal control

Define the value function associated with (10) as follows:

$$V(k, x) := \inf_{\{\pi_s\}_{s=k}^{N-1}} \mathbb{E}\left[\ell_N(x_N) + \sum_{s=k}^{N-1} \{\ell_s(x_s) + D_{\mathrm{KL}}\left(\rho_{s+1}^{\pi_s}(\cdot|x_s) \| \bar{\rho}_{s+1}(\cdot|x_s)\right)\} \middle| \ x_k = x\right],$$
$$(k, x) \in [\![N-1]\!] \times \mathbb{X},$$
$$V(N, x) := \ell_N(x), \ x \in \mathbb{X}.$$

Then the optimal value for Problem 3.1 is given by $\mathbb{E}[V(0, x_0)]$. Also, define the desirability function

$$Z(k, x) := \exp(-V(k, x)). \tag{13}$$

Similarly to the conventional optimal control, the desirability function or, equivalently, the value function plays a crucial role in our problem.

Theorem 4.1: *Suppose that Assumption 3.2 holds. Then, the unique optimal policy $\pi^* = \{\pi_k^*\}$ for Problem 3.1 is given by*

$$\pi_k^*(u|x) := \frac{\rho_{w_k}(u)Z(k+1, f(x, u))}{\mathcal{A}_{\bar{\rho}_{k+1}}[Z](k, x)},$$
$$k \in [\![N-1]\!], \ u \in \mathbb{U}, \ x \in \mathbb{X}. \tag{14}$$

In addition, the desirability function Z satisfies

$$Z(k, x) = \exp(-\ell_k(x))\mathcal{A}_{\bar{\rho}_{k+1}}[Z](k, x),$$
$$(k, x) \in [\![N-1]\!] \times \mathbb{X}, \tag{15}$$
$$Z(N, x) = \exp(-\ell_N(x)), \ x \in \mathbb{X}. \tag{16}$$

Proof: By the dynamic programming principle [23, Chapter 3], the value function V satisfies the Bellman equation

$$V(k, x) = \ell_k(x) + \inf_{\pi_k}\Big\{D_{\mathrm{KL}}\left(\rho_{k+1}^{\pi_k}(\cdot|x) \| \bar{\rho}_{k+1}(\cdot|x)\right) + \mathcal{A}_{\rho_{k+1}^{\pi_k}}[V](k, x)\Big\},$$
$$(k, x) \in [\![N-1]\!] \times \mathbb{X}, \tag{17}$$
$$V(N, x) = \ell_N(x), \quad x \in \mathbb{X}. \tag{18}$$

In addition, if a policy π_k achieves the minimum of the right-hand side of (17), it is an optimal policy. Note that

$$D_{\mathrm{KL}}\left(\rho_{k+1}^{\pi_k}(\cdot|x) \| \bar{\rho}_{k+1}(\cdot|x)\right) + \mathcal{A}_{\rho_{k+1}^{\pi_k}}[V](k, x)$$
$$= \int_{\mathbb{X}} \rho_{k+1}^{\pi_k}(x'|x) \log \frac{\rho_{k+1}^{\pi_k}(x'|x)}{\bar{\rho}_{k+1}(x'|x)Z(k+1, x')} \mathrm{d}x'$$
$$= D_{\mathrm{KL}}\left(\rho_{k+1}^{\pi_k}(\cdot|x) \| \rho_{k+1}^{*}(\cdot|x)\right) - \log \mathcal{A}_{\bar{\rho}_{k+1}}[Z](k, x), \tag{19}$$

where we defined

$$\rho^*_{k+1}(x'|x) := \frac{\bar{\rho}_{k+1}(x'|x)Z(k+1,x')}{\mathcal{A}_{\bar{\rho}_{k+1}}[Z](k,x)}, \quad x', x \in \mathbb{X}. \tag{20}$$

The second term in the right-hand side of (19) does not depend on π_k. Therefore, if a policy π_k satisfies

$$\rho^{\pi_k}_{k+1}(x'|x) = \rho^*_{k+1}(x'|x), \quad \forall x, x' \in \mathbb{X}, \tag{21}$$

it is an optimal policy at time k. For any $x \in \mathbb{X}$, by Assumption 3.2 and the change of variables $x' = f_x(u)$ for $\pi_k(u|x)$ [22, Chapter 6, Theorem 5], we obtain

$$\rho^{\pi_k}_{k+1}(x'|x) = \pi_k\left(f_x^{-1}(x')|x\right)\left|\det\left(\mathrm{J}_{f_x^{-1}}(x')\right)\right|,$$

where $\mathrm{J}_{f_x^{-1}}$ denotes the Jacobian matrix of the inverse function f_x^{-1}. Similarly, we have

$$\bar{\rho}_{k+1}(x'|x) = \rho_{w_k}\left(f_x^{-1}(x')\right)\left|\det\left(\mathrm{J}_{f_x^{-1}}(x')\right)\right|.$$

Therefore, π^*_k defined in (14) is a unique policy satisfying (21). As a result, the Bellman Equation (17) can be simplified as

$$V(k,x) = \ell_k(x) - \log \mathcal{A}_{\bar{\rho}_{k+1}}[Z](k,x), \tag{22}$$

which completes the proof. ■

From Theorem 4.1, similarly to the conventional optimal control, Problem 3.1 boils down to calculating the desirability function Z. A notable difference between them is that thanks to the linearity of (15), the desirability function for KL control admits the path integral representation.

Corollary 4.2: *Suppose that Assumption 3.2 holds. Then, the desirability function Z satisfies*

$$Z(k,x) = \mathbb{E}\left[\exp\left(-\sum_{s=k}^{N}\ell_s(\bar{x}_s)\right)\middle|\, \bar{x}_k = x\right],$$
$$(k,x) \in [\![N]\!] \times \mathbb{X}, \tag{23}$$

where $\{\bar{x}_s\}$ is a solution of (8).

Proof: By using (15), (16), and induction on k, we immediately obtain the desired result. ■

The path integral representation (23) motivates us to compute the desirability function by sampling approximations. In particular, if one can simulate sample paths of $\{\bar{x}_k\}$, the sampling approximations of (23) do not require the knowledge of f. Hence, (23) enables model-free approaches for obtaining the optimal policy.

Next, we consider the discrete input space $\mathbb{U} = \{u^{(i)}\}_{i=1}^r$, $u^{(i)} \in \mathbb{R}^m$, $r \in \mathbb{Z}_{>0} \cup \{\infty\}$. In this case, density functions must be replaced by probabilities such as a policy $\Pi_k(u^{(i)}|x) := \mathbb{P}(u_k = u^{(i)}|x_k = x)$. Then, we obtain the following.

Corollary 4.3: *Suppose that Assumptions 3.2-(i),(iv) hold. Then, for Problem 3.1 with $\mathbb{U} = \{u^{(i)}\}_{i=1}^r$, there exists a policy $\{\Pi^*_k\}$ such that for all $k \in [\![N-1]\!]$, $x \in \mathbb{X}$, $x' \in \bar{\mathbb{X}}_x = \{f_x(w) : w \in \mathbb{W}\}$, it holds*

$$\sum_{i: f(x,u^{(i)})=x'} \Pi^*_k(u^{(i)}|x) = \frac{\mathbb{P}\left(f(x,w_k) = x'\right)Z(k+1,x')}{\sum_{x'' \in \bar{\mathbb{X}}_x}\mathbb{P}\left(f(x,w_k) = x''\right)Z(k+1,x'')}. \tag{24}$$

Here, the desirability function Z satisfies (15) *and* (16) *where $\mathcal{A}_{\bar{\rho}_{k+1}}[Z](k,x)$ is replaced by $\sum_{x' \in \bar{\mathbb{X}}_x} Z(k+1,x')$ $\mathbb{P}(f(x,w_k) = x')$ and admits the representation* (23). *In addition, $\{\Pi^*_k\}$ is an optimal policy for Problem* 3.1. *Furthermore, if for all $x \in \mathbb{X}$, $f_x : \mathbb{U} \to \bar{\mathbb{X}}_x$ is bijective, a policy satisfying* (24) *is the unique optimal policy.*

Proof: Note that

$$\mathbb{P}(x_{k+1} = x'|x_k = x) = \sum_{i: f(x,u^{(i)})=x'} \Pi^*_k(u^{(i)}|x), \tag{25}$$

$$\mathbb{P}(\bar{x}_{k+1} = x'|\bar{x}_k = x) = \mathbb{P}\left(f(x,w_k) = x'\right). \tag{26}$$

Then, by the same argument as in the proof of Theorem 4.1, we obtain the existence and optimality of $\{\Pi^*_k\}$ satisfying (24). Especially when for all $x \in \mathbb{X}$, f_x is bijective, $\{u \in \mathbb{U} : f(x,u) = x'\}$ is a singleton for all $x \in \mathbb{X}$, $x' \in \bar{\mathbb{X}}_x$, which leads to the uniqueness of the optimal policy. ■

The above result clarifies that in Assumption 3.2, the condition (i) $\mathbb{W} \subseteq \mathbb{U}$ plays a crucial role in making optimal control problems linearly solvable while the bijectivity of f_x ensures the uniqueness of the optimal policy. Note that Corollary 4.3 does not assume $m = n$.

4.2. Reconsideration of the controllability assumption of transition densities

Now, let us go back to the original formulation of KL control (Problem 2.1) and reconsider the implication of Assumption 2.2 for stochastic systems. In the rest of this section, the control-affine system is considered:

$$x_{k+1} = \bar{f}(x_k) + g(x_k)(u_k + w_k), \; w_k \sim \rho_{w_k}, \tag{27}$$

where $\{w_k\}$ is a sequence of independent random variables. For simplicity, let $\mathbb{X} = \mathbb{R}^n$, $\mathbb{U} = \mathbb{R}^m$. Note that the continuous-time counterpart of (27) is often considered in the path integral control. Similarly to the linear system (5), for the above system, causal controllers $\pi_k(u_k|x_k)$ cannot satisfy (4), and therefore the associated Bellman equation cannot be linearized. To gain deeper insight into Assumption 2.2 that ensures the existence of a policy satisfying (4), we shall introduce an atypical assumption.

Assumption 4.4: The control input u_k is allowed to depend on w_k.

This assumption means that the causality of controllers can be violated. Now the decision variables for Problem 2.1 are replaced by $\pi_{\mathrm{w},k}(\cdot|x,w) := \rho_{u_k|x_k,w_k}(\cdot|x,w)$, $k \in [\![N-1]\!]$. Then, we have the following result.

Theorem 4.5: *Suppose that Assumptions 3.2-(ii),(iii) and 4.4 hold for $f_x(u) = \bar{f}(x) + g(x)u$. Then, the unique optimal policy for Problem 2.1 is given by*

$$\pi^*_{\mathrm{w},k}(u|x,w) := \frac{\rho_{w_k}(u+w)z\left(k+1, \bar{f}(x)+g(x)u\right)}{\mathcal{A}_{\rho^0_{k+1}}[z](k,x)},$$
$$k \in [\![N-1]\!],\ u,x \in \mathbb{R}^n,\ w \in \mathbb{W}. \tag{28}$$

In addition, the desirability function z satisfies (2) *and* (3).

Proof: Note that

$$\rho^0_{k+1}(x'|x) = \frac{1}{|\det\left(g(x)\right)|}\rho_{w_k}\left(\left(g(x)\right)^{-1}(x' - \bar{f}(x))\right) \tag{29}$$

and under a policy $\pi_{\mathrm{w},k}$,

$$\rho_{x_{k+1}|x_k,w_k}(x'|x,w)$$
$$= \frac{1}{|\det\left(g(x)\right)|}\pi_{\mathrm{w},k}\left(\left(g(x)\right)^{-1}(x' - \bar{f}(x)) - w\middle|\, x, w\right). \tag{30}$$

Also, we have

$$\begin{aligned}\rho^{\pi_{\mathrm{w},k}}_{k+1}(x'|x) &= \int_{\mathbb{W}} \rho_{x_{k+1},w_k|x_k}(x',w|x)\mathrm{d}w\\ &= \int_{\mathbb{W}} \rho_{x_{k+1}|x_k,w_k}(x'|x,w)\rho_{w_k|x_k}(w|x)\mathrm{d}w\\ &= \int_{\mathbb{W}} \rho_{x_{k+1}|x_k,w_k}(x'|x,w)\rho_{w_k}(w)\mathrm{d}w. \end{aligned} \tag{31}$$

Then, it is straightforward to check that (4) is satisfied for $\pi_{\mathrm{w},k} = \pi^*_{\mathrm{w},k}$. ■

This theorem shows that for the stochastic system (27), Assumption 4.4 for the noncausality of policies plays the same role as Assumption 2.2. In particular, the noncausality enables the controller to cancel the noise w_k. Combining this with $\mathbb{W} \subseteq \mathbb{U} = \mathbb{R}^n$, the controller can shape the transition density $\rho^{\pi_{k,\mathrm{w}}}_{k+1}(x'|x)$ to a desired form with the support $\bar{\mathbb{X}}_x$. Of course, the noncausality is unrealistic for practical applications. This clarifies that the reformulated KL control is much more realistic for systems on continuous spaces than the original formulation of KL control.

5. Linear quadratic Gaussian setting

In this section, we focus on a linear system ($f(x,u) = Ax + Bu$) with $\mathbb{U} = \mathbb{R}^m$, a quadratic cost

$$\ell_k(x) = \frac{1}{2}x^\top Q_k x, \quad Q_k \succ 0,\ k = 0,\dots,N, \tag{32}$$

and Gaussian noise $w_k \sim \mathcal{N}(0,\Sigma_k)$, $\Sigma_k \succ 0$. Assume that $m = n$ and B is invertible. Then Assumption 3.2 is satisfied. Now, we calculate the optimal policy for Problem 3.1 analytically. First, for $k = N-1$, we have

$$\begin{aligned}&\pi^*_{N-1}(u|x) \propto \mathcal{N}(u|0,\Sigma_{N-1})Z(N, Ax+Bu)\\ &\quad\propto \exp\left(-\frac{1}{2}\left(u^\top \Sigma^{-1}_{N-1}u\right.\right.\\ &\qquad \left.\left. + \ (Ax+Bu)^\top Q_N(Ax+Bu)\right)\right)\\ &\quad\propto \exp\left(-\frac{1}{2}\left[u + (\Sigma^{-1}_{N-1} + B^\top Q_N B)^{-1}B^\top Q_N Ax\right]^\top\right.\\ &\qquad \times (\Sigma^{-1}_{N-1} + B^\top Q_N B)\\ &\qquad \left.\times \left[u + (\Sigma^{-1}_{N-1} + B^\top Q_N B)^{-1}B^\top Q_N Ax\right]\right),\end{aligned} \tag{33}$$

which means that

$$\begin{aligned}&\pi^*_{N-1}(u|x)\\ &\quad= \mathcal{N}\big(u\big|-(\Sigma^{-1}_{N-1} + B^\top Q_N B)^{-1}B^\top Q_N Ax,\\ &\qquad (\Sigma^{-1}_{N-1} + B^\top Q_N B)^{-1}\big).\end{aligned} \tag{34}$$

On the other hand,

$$\begin{aligned}&Z(N-1,x)\\ &\quad= \exp\left(-\frac{1}{2}x^\top Q_{N-1}x\right)\\ &\qquad \times \int_{\mathbb{R}^n} \mathcal{N}(x'|Ax, B\Sigma_{N-1}B^\top)Z(N,x')\mathrm{d}x'\\ &\quad= [\det(I + Q_N B\Sigma_{N-1}B^\top)]^{-1/2}\\ &\qquad \times \exp\left(-\frac{1}{2}x^\top(Q_{N-1}\right.\\ &\qquad + A^\top(I - (I + Q_N B\Sigma_{N-1}B^\top)^{-1})\\ &\qquad \left.\times\ (B\Sigma_{N-1}B^\top)^{-1}A)x\right),\end{aligned} \tag{35}$$

where we used the formula

$$\begin{aligned}&\mathbb{E}\left[\exp\left(-\frac{1}{2}x^\top Qx\right)\right]\\ &\quad= [\det(I + Q\Sigma)]^{-1/2}\\ &\qquad \times \exp\left(-\frac{1}{2}\mu^\top(I - (I+Q\Sigma)^{-1})\Sigma^{-1}\mu\right)\end{aligned}$$

for $Q \succ 0$ and $x \sim \mathcal{N}(\mu,\Sigma)$, $\Sigma \succ 0$. Note that

$$(I - (I + Q_N B\Sigma_{N-1}B^\top)^{-1})(B\Sigma_{N-1}B^\top)^{-1}$$

$$\begin{aligned}
&= (I - \Sigma_B^{-1}(I + Q_N^{-1}\Sigma_B^{-1})^{-1}Q_N^{-1})\Sigma_B^{-1} \\
&= (Q_N^{-1} + B\Sigma_{N-1}B^\top)^{-1} \\
&= Q_N - Q_N B(\Sigma_{N-1}^{-1} + B^\top Q_N B)^{-1}B^\top Q_N,
\end{aligned}$$

where $\Sigma_B := B\Sigma_{N-1}B^\top$. Substituting this into (35), we obtain

$$\begin{aligned}
Z(N-1, x) = {} & [\det(I + Q_N B\Sigma_{N-1}B^\top)]^{-1/2} \\
& \times \exp\left(-\frac{1}{2}x^\top P_{N-1}x\right), \qquad (36)
\end{aligned}$$

$$\begin{aligned}
P_{N-1} &:= Q_{N-1} + A^\top P_N A \\
&\quad - A^\top P_N B(\Sigma_{N-1}^{-1} + B^\top P_N B)^{-1}B^\top P_N A, \\
P_N &:= Q_N. \qquad (37)
\end{aligned}$$

By applying the same argument as above for $k = N-2, \dots, 0$, we obtain the following result.

Theorem 5.1: *Assume that $m = n$ and B is invertible. Then, the optimal policy $\pi^* = \{\pi_k^*\}$ for Problem 3.1 with $f(x,u) = Ax + Bu$, $\mathbb{U} = \mathbb{R}^n$, $w_k \sim \mathcal{N}(0, \Sigma_k)$, $\Sigma_k \succ 0$, and (32) is given by*

$$\begin{aligned}
\pi_k^*(u|x) = \mathcal{N}\big(u\big| & -(\Sigma_k^{-1} + B^\top P_{k+1}B)^{-1}B^\top P_{k+1}Ax, \\
& (\Sigma_k^{-1} + B^\top P_{k+1}B)^{-1}\big), \\
& k \in [\![N-1]\!],\ u, x \in \mathbb{R}^n \qquad (38)
\end{aligned}$$

where P_k is a solution of the Riccati difference equation

$$\begin{aligned}
P_k &= Q_k + A^\top P_{k+1}A \\
&\quad - A^\top P_{k+1}B(\Sigma_k^{-1} + B^\top P_{k+1}B)^{-1}B^\top P_{k+1}A, \\
&\quad k \in [\![N-1]\!], \qquad (39) \\
P_N &= Q_N. \qquad (40)
\end{aligned}$$

The desirability function is given by

$$\begin{aligned}
Z(k, x) = {} & \left(\prod_{s=k+1}^{N} [\det(I + P_s B\Sigma_{s-1}B^\top)]^{-1/2}\right) \\
& \times \exp\left(-\frac{1}{2}x^\top P_k x\right). \qquad (41)
\end{aligned}$$

The mean of the optimal policy (38) coincides with the LQ optimal controller [1]. In other words, the optimal policy is the LQ optimal feedback controller perturbed by additive Gaussian noise with zero mean and covariance matrix $(\Sigma_k^{-1} + B^\top P_{k+1}B)^{-1}$.

In the above, we have analysed the desirability function based on the backward equation (15). Hence, the obtained representation (41) contains the solution of the backward Riccati difference equation. For comparison, we calculate the desirability function based on the forward representation (23). Let $\bar{x}_{k+1:N} := [\bar{x}_{k+1}^\top \cdots \bar{x}_N^\top]^\top$ and

$$\bar{A}_k := \left[A^\top\ (A^2)^\top\ \cdots\ (A^k)^\top\right]^\top, \qquad (42)$$

$$\Sigma_{k+1:N} := \mathrm{diag}(\Sigma_{k+1}, \dots, \Sigma_N), \qquad (43)$$

$$Q_{k+1:N} := \mathrm{diag}(Q_{k+1}, \dots, Q_N), \qquad (44)$$

$$L_k := \begin{bmatrix} B & 0 & \cdots & \cdots & 0 \\ AB & B & \ddots & & \vdots \\ A^2B & AB & B & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ A^{k-1}B & A^{k-2}B & \cdots & AB & B \end{bmatrix}. \qquad (45)$$

Then, the conditional distribution of $\bar{x}_{k+1:N}$ given $\bar{x}_k = x$ is $\mathcal{N}(\bar{A}_{N-k}x, L_{N-k}\Sigma_{k+1:N}L_{N-k}^\top)$. By Corollary 4.2,

$$\begin{aligned}
Z(k, x) &= \exp\left(-\frac{1}{2}\|x\|_{Q_k}^2\right) \\
&\qquad \mathbb{E}\left[\exp\left(-\frac{1}{2}\|\bar{x}_{k+1:N}\|_{Q_{k+1:N}}^2\right)\middle|\ \bar{x}_k = x\right] \\
&= [\det(I + Q_{k+1:N}L_{N-k}\Sigma_{k+1:N}L_{N-k}^\top)]^{-1/2} \\
&\quad \times \exp\Big(-\frac{1}{2}x^\top\big(Q_k + \bar{A}_{N-k}^\top(Q_{k+1:N}^{-1} \\
&\qquad + L_{N-k}\Sigma_{k+1:N}L_{N-k}^\top)^{-1}\bar{A}_{N-k}\big)x\Big), \qquad (46)
\end{aligned}$$

where $\|x\|_Q := (x^\top Qx)^{1/2}$ for $Q \succ 0$. The fact that the desirability function can be expressed in two different ways (41) and (46) is similar to the fact that the value function for the LQR problem

$$\begin{aligned}
& V_{\mathrm{LQR}}(k, x) \\
& := \inf_{\{u_s\}} \frac{1}{2}\|x_N\|_{Q_N}^2 + \sum_{s=k}^{N-1}\frac{1}{2}\left(\|x_s\|_{Q_s}^2 + \|u_s\|_{\Sigma_s^{-1}}^2\right) \\
& \qquad \text{subj. to } x_{s+1} = Ax_s + Bu_s,\ s \in [k, N-1],\ x_k = x
\end{aligned}$$

can be written in the following two ways:

$$V_{\mathrm{LQR}}(k, x) = \begin{cases} \frac{1}{2}x^\top P_k x, \\ \frac{1}{2}x^\top\big(Q_k + \bar{A}_{N-k}^\top(Q_{k+1:N}^{-1} \\ \quad + L_{N-k}\Sigma_{k+1:N}L_{N-k}^\top)^{-1}\bar{A}_{N-k}\big)x. \end{cases} \qquad (47)$$

6. Numerical examples

In this section, we illustrate the reformulated KL control through two examples.

6.1. Linear quadratic case

Consider the linear quadratic case where

$$A = 0.85, \quad B = 0.10, \quad Q_k = 3.0, \quad \Sigma_k = 1.5, \forall k \qquad (48)$$

and a finite horizon $N = 30$. First, for comparison we compute the associated value function in two ways: by using the explicit expression (41) and by using a Monte Carlo method based on the path integral representation (23). For the Monte Carlo method, we generate S sample paths $\{\bar{x}_k^{(i)}\}_{k=0}^{N}, i = 1, \dots, S$ with $\bar{x}_0 = x, w_k \sim \mathcal{N}(0, \Sigma_k)$ and compute

$$-\log\left[\frac{1}{S}\sum_{i=1}^{S}\exp\left(-\sum_{s=0}^{N}\ell_s(\bar{x}_s^{(i)})\right)\right]$$

to approximate $V(0, x)$. As shown in Figure 1, $V(0, x)$ is well approximated by the Monte Carlo estimate with 3000 samples. The computation time for each x is about 0.025 s, 0.24 s, and 0.71 s for $S = 100, 1000, 3000$, respectively, with MATLAB on MacBook Pro with Apple M1 Pro. Note that the Monte Carlo simulations can be easily parallelized. Next, three samples of the optimal state and control processes $\{x_k\}, \{u_k\}$ for different (Q_k, Σ_k) are shown in Figure 2. As can be seen, as Σ_k increases, the absolute mean and variance of the optimal control get larger. This is because for larger Σ_k, the cost of shifting the transition distribution $\rho_{k+1}^{\pi_k}(\cdot|x_k)$ from the reference distribution $\mathcal{N}(\cdot|Ax_k, B\Sigma_k B^\top)$ becomes smaller, while the cost of reducing the variance of the transition distribution becomes larger. In Figures 2(c,d), the values of Q_k/Σ_k^{-1} coincide. Therefore the mean values of the optimal policies (38) for the two cases also coincide although the control process in Figure 2(d) has smaller variance than in Figure 2(c). On the other hand, for the LQR problem

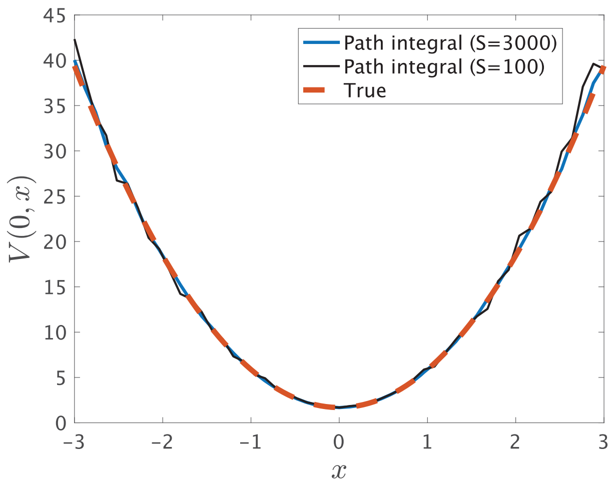


Figure 1. Monte Carlo estimates of the value function $V(0, x)$ (red, dashed) with $S = 100$ (black) and $S = 3000$ (blue).

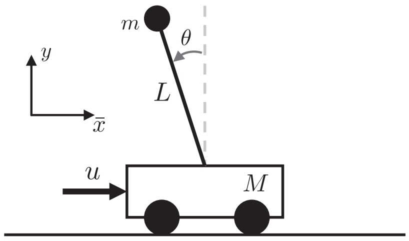
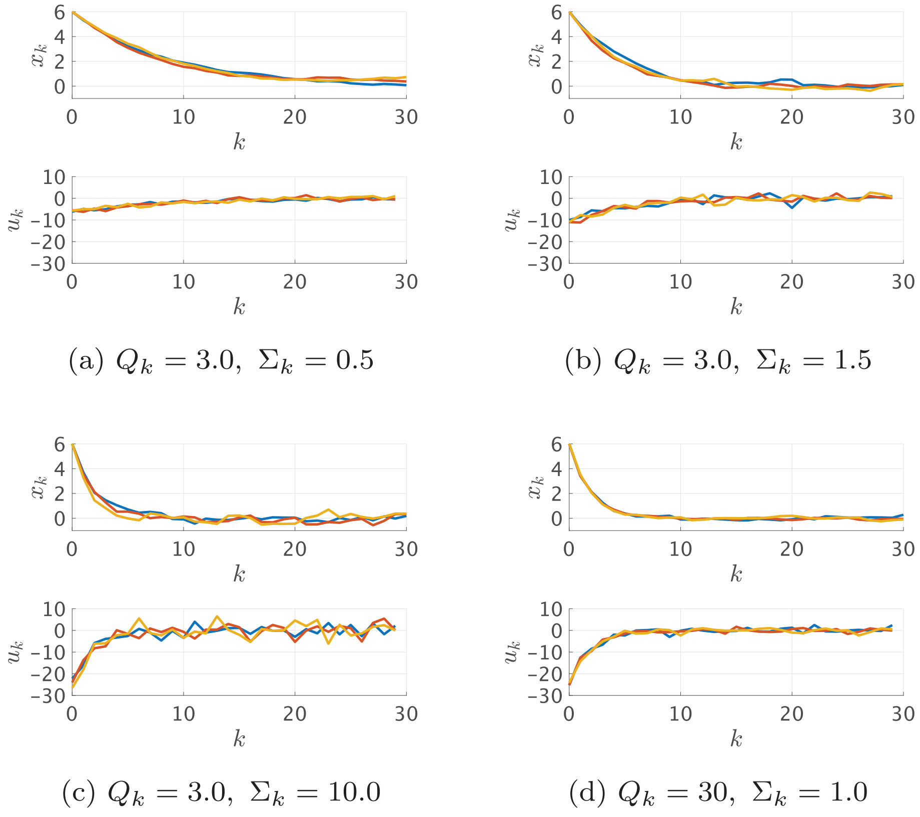


Figure 3. Cart-pole pendulum.



(a) $Q_k = 3.0,\ \Sigma_k = 0.5$

(b) $Q_k = 3.0,\ \Sigma_k = 1.5$

(c) $Q_k = 3.0,\ \Sigma_k = 10.0$

(d) $Q_k = 30,\ \Sigma_k = 1.0$

Figure 2. Three samples of the optimal state and control processes $\{x_k\}, \{u_k\}$ for different (Q_k, Σ_k): (a) $Q_k = 3.0,\ \Sigma_k = 0.5$. (b) $Q_k = 3.0,\ \Sigma_k = 1.5$. (c) $Q_k = 3.0,\ \Sigma_k = 10.0$ and (d) $Q_k = 30,\ \Sigma_k = 1.0$.

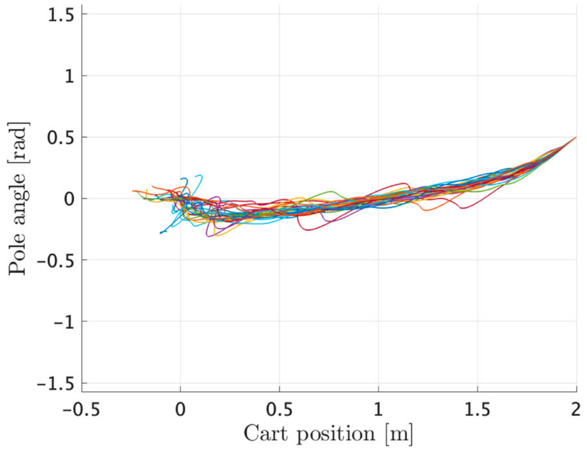


Figure 4. 30 sample paths of the optimal state process $\{(\bar{x}_k, \theta_k)\}$.

whose cost is given by

$$\frac{1}{2}Qx_N^2 + \sum_{k=0}^{N-1}\frac{1}{2}(Qx_k^2 + \Sigma^{-1}u_k^2),$$

the optimal control depends on Q, Σ only via Q/Σ^{-1}. This is in clear contrast to KL control.

6.2. Cart-pole pendulum

We now proceed to apply our result to a nonlinear system. Specifically, we consider the cart-pole inverted pendulum in Figure 3. The system consists of a cart of mass $M = 1.0\,\mathrm{kg}$ moving horizontally, a massless rod of length $L = 0.5\,\mathrm{m}$ attached to the cart and rotating around a pivot point in the $\bar{x}y$-plane only, and a point mass $m = 0.1\,\mathrm{kg}$ at the end of the rod. The input u is the horizontal force applied to the cart to maintain the pendulum in a balanced and upright position. Here, we neglect the influence of friction. Let $\bar{x}, \theta$ be the position of the cart and the angle of the rod ($\theta = 0$ for the upright position and $\theta = \pi$ for the downward position of the pendulum), respectively.

We then have the following continuous-time model of the cart-pole system:

$$\begin{aligned}\ddot{\bar{x}} &= \frac{-mL(\dot{\theta})^2\sin\theta + mg\sin\theta\cos\theta + u}{M + m\sin^2\theta}\\ &=: h_1(\theta, \dot{\theta}, u),\end{aligned} \qquad (49)$$

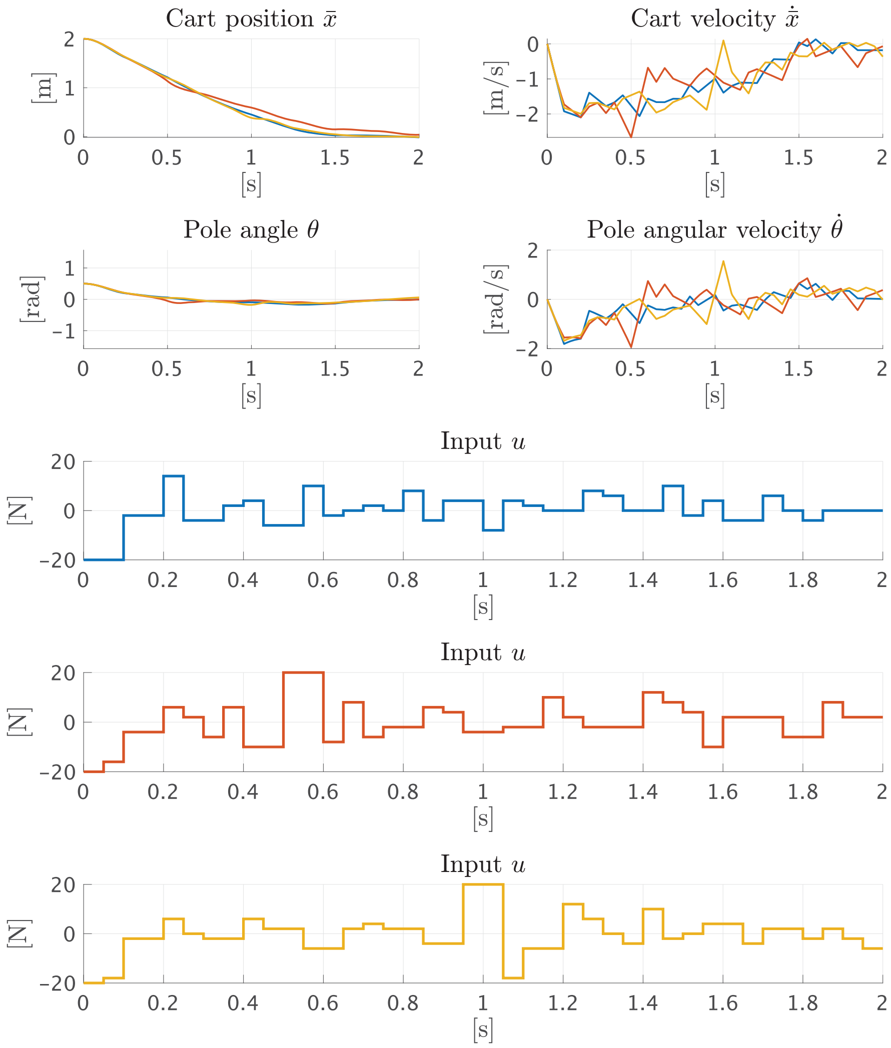


Figure 5. Three sample paths of the optimal state and control processes for the cart-pole pendulum. The same color indicates the correspondence between the sample paths of the state process and the control process.

$$\ddot{\theta} = \frac{1}{L}(h_1(\theta,\dot{\theta},u)\cos\theta + g\sin\theta) =: h_2(\theta,\dot{\theta},u), \tag{50}$$

where $g = 9.8\,\mathrm{m/s^2}$ is the gravitational acceleration. By the Euler method, we obtain the discrete-time system:

$$x_{k+1} = f(x_k,u_k) = \begin{bmatrix} \bar{x}_k + \tau\dot{\bar{x}}_k \\ \dot{\bar{x}}_k + \tau h_1(\theta_k,\dot{\theta}_k,u_k) \\ \theta_k + \tau\dot{\theta}_k \\ \dot{\theta}_k + \tau h_2(\theta_k,\dot{\theta}_k,u_k) \end{bmatrix}, \tag{51}$$

where $x_k = [\bar{x}_k\ \dot{\bar{x}}_k\ \theta_k\ \dot{\theta}_k]^\top$ and τ is a step size. Here, we consider the discrete input space $\mathbb{U} = \{2i\ \mathrm{N}\}_{i=-10}^{10}$. For a cost function, let

$$\ell_k(x_k) = q_1|\bar{x}_k| + q_2|\dot{\bar{x}}_k| + q_3|\theta_k| + q_4|\dot{\theta}_k|$$

with $q_1 = 11.5\,\mathrm{m}^{-1}$, $q_2 = 3.0\ \mathrm{s/m}$, $q_3 = 11.5\,\mathrm{rad}^{-1}$, $q_4 = 3.0\,\mathrm{s/rad}$. In addition, the noise w_k for the reference transition distribution is designed to follow a discretized Gaussian distribution

$$\mathbb{P}(w_k = \mathrm{w}) \propto \exp\left(-\frac{1}{2\sigma^2}\mathrm{w}^2\right), \quad \mathrm{w} \in \mathbb{W} = \mathbb{U} \tag{52}$$

with $\sigma = 5.0$ N. The initial state is given by $\bar{x}_0 = 2.0\,\mathrm{m}$, $\dot{\bar{x}}_0 = 0\,\mathrm{m/s}$, $\theta_0 = 0.5\,\mathrm{rad}$, $\dot{\theta}_0 = 0\,\mathrm{rad/s}$.

Suppose that the state value at the current time k is $x_k = x$. Then by Corollary 4.3, the optimal policy at time k is given by

$$\Pi_k^*(u|x) \propto \mathbb{P}(w_k = u)\, Z\left(k+1, f(x,u)\right), \ u \in \mathbb{U}, \tag{53}$$

where the desirability function $Z(k+1, f(x,u))$ for each $u \in \mathbb{U}$ can be computed by the Monte Carlo method based on (23). In this example, we use 20, 000 samples for the sampling approximation of Z.

Figure 4 shows 30 sample paths of the optimal state process in the $\bar{x}\theta$-plane. The sampling time for simulating the cart-pole system (49), (50) is $\tau = 0.1$ ms while the sampling time for determining control inputs is $\tau = 0.05$ s. The optimal policy balances the pendulum around the upright position while the cart-pole system fluctuates around the origin due to the stochasticity of the policy. The detailed behaviour of the optimal state and control processes is illustrated in Figure 5. One can see that the cart and pole velocity shows large fluctuations while as time evolves, their mean values approach zero. If one takes larger values of q_2, q_4, their fluctuations are reduced.

7. Conclusion

In this paper, we reformulated KL control to make its assumption reasonable for continuous spaces and remove the approximation of dynamics. Then, we analysed the associated optimal control via the desirability function. In particular, we showed that the reformulated KL control admits sampling approximations of the desirability function. We emphasize that the Bellman equation for the infinite horizon KL control can also be linearized by the same argument as in the finite horizon case, and the associated inverse reinforcement learning can be formulated as a convex optimization [18]. In addition, we revisited the original KL control and clarified that the assumption of controllability of transition densities implies the noncausality of controllers. For linear systems with a quadratic state cost and Gaussian noise, we derived the optimal policy analytically. Lastly, we illustrated our KL control via numerical examples. Future work will focus on weakening Assumptions 3.2-(ii),(iii) by analysing the problem without using densities.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported in part by JSPS (Japan Society for the Promotion of Science) KAKENHI [grant number JP21J14577, JP21H04875] and by JST, ACT-X [grant number JPMJAX2102].

Notes on contributors

Kaito Ito received the Bachelor's degree in Engineering and the Master's degree in Informatics from Kyoto University in 2017 and 2019, respectively. He is currently a Ph.D. student at Kyoto University. His research interests include stochastic control, machine learning, and privacy protection for dynamical systems. He is a student member of IEEE.

Kenji Kashima received his Doctoral degree in Informatics from Kyoto University in 2005. He was with Tokyo Institute of Technology, Universität Stuttgart, Osaka University, before he joined Kyoto University in 2013, where he is currently an Associate Professor. His research interests include control and learning theory for complex (large scale, stochastic, networked) dynamical systems, as well as its interdisciplinary applications. He received Humboldt Research Fellowship (Germany), IEEE CSS Roberto Tempo Best CDC Paper Award, Pioneer Award of SICE Control Division, and so on. He is a Senior Member of IEEE and Member of ISCIE and IEICE.

References

[1] Lewis FL, Vrabie D, Syrmos VL. Optimal control. Hoboken, NJ: John Wiley & Sons; 2012.

[2] Ito K, Ikeda T, Kashima K. Sparse optimal stochastic control. Automatica. 2021;125:109438.

[3] Liu GH, Theodorou EA. Deep learning theory review: an optimal control and dynamical systems perspective. arXiv preprint arXiv:190810920. 2019.

[4] Recht B. A tour of reinforcement learning: the view from continuous control. Ann Rev Control Robot Autonom Syst. 2019;2(1):253–279.

[5] Hernández-Lerma O, Lasserre JB. Discrete-time Markov control processes: basic optimality criteria. Vol. 30. New York: Springer-Verlag New York; 1996.

[6] Yong J, Zhou XY. Stochastic controls: Hamiltonian systems and HJB equations. Vol. 43. New York: Springer Science & Business Media; 1999.

[7] Kappen HJ. Linear theory for control of nonlinear stochastic systems. Phys Rev Lett. 2005;95(20):Article ID 200201.

[8] Todorov E. Linearly-solvable Markov decision problems. In: Advances in Neural Information Processing Systems; 2006. p. 1369–1376.

[9] Theodorou E, Buchli J, Schaal S. A generalized path integral control approach to reinforcement learning. J Mach Learn Res. 2010;11:3137–3181.

[10] Theodorou E, Buchli J, Schaal S. Reinforcement learning of motor skills in high dimensions: a path integral approach. In: 2010 IEEE International Conference on Robotics and Automation. IEEE; 2010. p. 2397–2403.

[11] Williams G, Aldrich A, Theodorou EA. Model predictive path integral control: from theory to parallel computation. J Guid Control Dyn. 2017;40(2):344–357.

[12] Van Den Broek B, Wiegerinck W, Kappen B. Graphical model inference in optimal control of stochastic multiagent systems. J Artific Intel Res. 2008;32:95–122.

[13] Kashima K. Noise response data reveal novel controllability Gramian for nonlinear network dynamics. Sci Rep. 2016;6(1):Article ID 27300.

[14] Todorov E. Efficient computation of optimal actions. Proc Nat Acad Sci. 2009;106(28):11478–11483.

[15] Todorov E. Compositionality of optimal control laws. Adv Neural Inf Process Syst. 2009;22:1856–1864.

[16] Todorov E. General duality between optimal control and estimation. In: 2008 47th IEEE Conference on Decision and Control. IEEE; 2008. p. 4286–4292.

[17] Theodorou EA, Todorov E. Relative entropy and free energy dualities: connections to path integral and KL control. In: 2012 IEEE 51st IEEE Conference on Decision and Control (CDC). IEEE; 2012. p. 1466–1473.

[18] Dvijotham K, Todorov E. Inverse optimal control with linearly-solvable MDPs. In: ICML. 2010.

[19] Rawlik K, Toussaint M, Vijayakumar S. On stochastic optimal control and reinforcement learning by approximate inference. In: Proceedings of Robotics: Science and Systems; 2012.

[20] Todorov E. Eigenfunction approximation methods for linearly-solvable optimal control problems. In: 2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning. IEEE; 2009. p. 161–168.

[21] Zhong M, Todorov E. Aggregation methods for linearly-solvable Markov decision process. In: Proceedings of the World Congress of the International Federation of Automatic Control. Elsevier; 2011. p. 11220–11225.

[22] Roussas GG. An introduction to probability and statistical inference. San Diego, California: Elsevier; 2015.

[23] Kushner HJ. Introduction to stochastic control. New York: Holt, Rinehart, and Winston; 1971.