An Information-based Learning Approach to Dual Control

Tansu Alpcan, *Senior Member, IEEE,* and Iman Shames

***Abstract*—Dual control aims to concurrently learn and control an unknown system. However, actively learning the system conflicts directly with any given control objective for it will disturb the system during exploration. This paper presents a receding horizon approach to dual control, where a multi-objective optimization problem is solved repeatedly subject to constraints representing system dynamics. Balancing a standard finite horizon control objective, a knowledge gain objective is defined to explicitly quantify the information acquired when learning the system dynamics. Measures from information theory such as entropy-based uncertainty, Fisher information, and relative entropy are studied and used to quantify knowledge gained as a result of the control actions. The resulting iterative framework is applied to Markov Decision Processes and discrete-time nonlinear systems. Thus, the broad applicability and usefulness of the presented approach is demonstrated in diverse problem settings. The framework is illustrated with multiple numerical examples.**

***Index Terms*—Dual control, information theory, nonlinear systems, active learning, black-box systems.**

I. Introduction

MANY real world systems are controlled with only limited amount of information. In some cases, acquiring extensive information on system characteristics may be simply infeasible due to prohibitive costs or observation limitations. In others, the observed system may be so non-stationary that by the time the information is obtained, it is already outdated due to system's fast-changing nature. Therefore, the only option left to the controller is to develop a strategy for collecting information efficiently and estimate the system within a chosen modeling framework in order to achieve a given control objective [1].

When a controller aims to concurrently learn and control an unknown system, the objective of actively learning the system by disturbing it for exploration conflicts directly with the given control objective. Therefore, active learning plays a crucial role in this problem setting. This paper focuses on quantifying knowledge obtained during active learning using concepts from information theory such as mutual information, Fisher information and relative entropy which is also known as Kullback-Leibler divergence. Thus, knowledge as a measurable quantity is explicitly integrated to the decision process. Since this knowledge can be quantified only within a model, Gaussian process regression (GPR) [2] are used as learning frameworks in this paper for modeling and estimating unknown dynamical systems. Note that *knowledge* and *knowledge gain* are used in this paper instead of *information* in order to avoid any confusion with the specific meaning of "information" in information and communications theories.

This paper presents a receding horizon approach to dual control, where a multi-objective optimization problem subject to constraints representing at-the-time-known system dynamics is solved repeatedly. Specifically, an iterative algorithm is proposed where at each step the control action is chosen to maximize not only a given finite horizon control objective but also a knowledge objective which helps learning the system dynamics (constraints of the optimization problem) through exploration. One of the main contributions of the paper is the explicit and quantitative representation of this knowledge acquisition using measures from information theory. Thus, dual control is formulated as a multi-objective problem which directly incorporates a knowledge objective. The framework developed is applied to dual control of Markov Decision Processes (MDPs) under a Dirichlet prior as well as dual control of nonlinear systems using GPR to illustrate its broad applicability to diverse problems and settings.

Literature Overview

The *dual control* problem has a long history in the control literature [3], [4]. Conventional (non-dual) adaptive controllers do not actively learn a system by varying the control input for exploration. Although some dual adaptive controllers introduce a (small) perturbation signal to their control for learning, there is often limited guidelines on how to choose the perturbation signal and especially when to use probing [5]. Decision making with limited information is also related to search theory. Information (theory) has been used in this context for decades [6], [7] and the topic has been more recently revisited in [8].

Learning plays an important role in the presented framework, especially *regression*, which is a classical pattern recognition (statistical/machine learning) method [9], [10]. The book [11] provides important and valuable insights into the relationship between information theory, inference, and learning. Another relevant topic is Bayesian inference [11], [12], which is in the foundation of the presented framework. The book [2] presents a comprehensive treatment of GPs.

The area of active learning or experiment design focuses on data scarcity in machine learning and makes use of Shannon information theory among other criteria [13]. The paper [14] discusses objective functions which measure the expected

T. Alpcan and I. Shames are with the Department of Electrical and Electronic Engineering, the University of Melbourne, Parkville, VIC, 3010 Australia e-mail: (tansu.alpcan, iman.shames)@unimelb.edu.au

This work was supported in part by the ARC Discovery Project, DP140100819 and a McKenzie Fellowship.

informativeness of candidate measurements within a Bayesian learning framework. The subsequent study [15] investigates active learning for GP regression using variance as a (heuristic) confidence measure for test point rejection.

Iterative learning control is concerned with tracking a reference trajectory defined over a finite time duration, and is applied to systems which perform this action repeatedly. Therefore, it differs significantly from the problem setup considered in this paper [16]. Extremum seeking (ES) is an optimal control approach that deals with situations when the system model is not available to the designer but the input and output signals are measured similar to the case here. An extremum seeking controller dynamically searches for the optimizing inputs in real time. Unlike the framework proposed here, ES does not try to model the system explicitly and relies on a gradient descent approach [17] or alternatives such as sampling optimization (Shubert algorithm) [18]. Therefore, it cannot be applied to the problem studied in this paper.

Model predictive control (MPC) is an approach widely used to control dynamical systems with input and output constraints while ensuring the optimality of the system performance with respect to a given cost function [19]. The control input in MPC is typically calculated at each time-step by applying the first control in a sequence obtained from solving an optimal control problem over a finite or infinite horizon. The optimal problem is reformulated at each time step based on the available measurements. Traditionally, a full model of the system is required to solve the MPC problem and all of the control inputs are calculated centrally. However, in large-scale interconnected systems, such as power systems [20], [21], water distribution systems [21], [22], transport systems, manufacturing systems, biological systems, and irrigation systems [23], the assumption of knowing the whole model precisely is not realistic. The scenario where the models are learned during control is an active research area [24], [25]. This paper is another step in the direction where concepts from information and statistical learning theories are combined with those from control theory.

Contributions

The main contributions of this paper include:

- Formulation of dual control as a multi-objective problem, which quantifies the knowledge gain objective explicitly.
- Adaptation of concepts such as mutual information and relative entropy from information theory to quantify knowledge gain within a given active learning framework.
- Application of the framework to dual control of Markov Decision Processes as well as dual control of nonlinear systems.
- Deriving specific lower bounds on knowledge gain within the context of Gaussian process-based learning.

The next section presents the system model and problem formulation, which is independent of the learning framework. Section III provides definitions of and discusses relevant information and knowledge measures. Section IV presents an application of the approach to dual control of Markov Decision Processes where Dirichlet prior. The subsequent Section V applies the framework to dual control of nonlinear systems using GP Regression based on mutual information and relative entropy. Section VI illustrates the results with multiple numerical examples. The paper ends with the concluding remarks in Section VII.

II. System Model and Problem Formulation

In *dual control*, the controller aims to concurrently learn and control an unknown system. The first objective of the controller is to learn the unknown system in an efficient way. The second and concurrent objective is to control the system (optimally) based on the best available estimation. These two objectives are clearly in conflict. The control action that gives more information on system dynamics may not necessarily be the one optimally controlling the system and vice versa. It is the duality of these objectives, that motivates the "dual control" name of the problem.

Consider the discrete-time dynamical system

$$x_i(n+1) = f_i(x(n), u(n)), \;\; i = 1, 2, \ldots, d, \tag{1}$$

where $n = 1, 2, \ldots,$ denotes the time step. The state vector $x \in \mathcal{X}$ denotes that x belongs to the set $\mathcal{X} \subseteq \mathbb{R}^d$, the vector $u \in \mathcal{U}$ denotes the control action chosen from the finite set of available actions $\mathcal{U}$, and $f_i : \mathcal{X} \times \mathcal{U} \to \mathcal{X}$ is the mapping from the cross-product of state and control spaces to the the state space.

For notational convenience, define $z := [x, u]^T \in \mathcal{Z} := \mathcal{X} \times \mathcal{U}$, where $[\cdot]^T$ denotes the transpose operator. Then, the dynamical system is given by

$$x_i(n+1) = f_i(z(n)), \;\; i = 1, 2, \ldots, d.$$

It is assumed here that the dynamical system f is *not known except from past observations*. Therefore, it is estimated from available past data within a learning framework. Furthermore, the controller does not have the time or resources for collecting extensive information to identify the system. Therefore, methods such as extremum seeking or iterative learning cannot be applied to the problem.

Let the set

$$\mathcal{D}(n) = \{(z(0), x(1)), (z(1), x(2)), \ldots, (z(n-1), x(n))\} \tag{2}$$

denote the D data points obtained from past observations of the dynamical system.[1] For notational convenience, both $(z(n), x(n+1))$ and $z(n)$ will represent an observation for the rest of the paper. Using this data set a consistent estimate $\hat{f}$ of the original system f is obtained.

A finite-horizon control problem based on the estimated system $\hat{f}$ and initial state $x(0)$ is defined as follows:

$$\min_{u(0:N-1)} J(x(0), u(0:N-1), \hat{x}(1:N)) \tag{3}$$
$$= \min_{u(0:N-1)} \sum_{k=1}^{N} \beta^k J_k(\hat{x}(k-1), u(k-1), \hat{x}(k)),$$
$$\text{such that } \hat{x}(n+1) = \hat{f}(\hat{x}(n), u(n)), \;\; n = 0, \ldots, N-1,$$

[1] In some cases the observations, $y(n+1) = f(z(n)) + \mathcal{N}(0, \sigma)$, may be distorted by a Gaussian noise $\mathcal{N}(0, \sigma)$ with fixed variance σ.

where $\hat{x}(0) = x(0)$, $J_k \geq 0$ is the finite stage cost at time k, and $0 < \beta \leq 1$ is the future discount factor introduced to account for modeling uncertainty. Notice that this formulates a receding-horizon control problem where the system dynamics constituting the constraints are to be learned iteratively by choosing appropriate control actions.

In order to make this learning process more precise, an explicit knowledge objective is introduced

$$\max_{u(n)} \mathcal{I}_n(\hat{x}(n+1); x(n), u(n), \mathcal{D}(n)), \tag{4}$$

where the scalar $\mathcal{I}_n \geq 0$ denotes the amount of expected knowledge gain at time n for a given state $x(n)$, data set $\mathcal{D}(n)$, and chosen control action $u(n)$.

The presented approach to dual control repeatedly solves a finite horizon multi-objective optimization problem combining the objectives (3) and (4) subject to system dynamics as constraints. Since the system dynamics are not known, the control actions have to be chosen to satisfy not only (3) but also the knowledge acquisition goal (4) which is explicitly calculated using information theoretic metrics. Hence, the control and learning objectives are concurrently addressed in a principled way. The resulting framework is applied to dual control of MDPs using Dirichlet Processes is presented in Section IV and of nonlinear systems using GPR in Section V, respectively. Algorithms 1 and 2 in those sections illustrate two specific instances of this approach.

III. Information Measures for Dual Control

Concepts from the field of *information theory* provide a natural foundation for explicitly quantifying the knowledge objective $\mathcal{I}$ in the dual control problem defined (4). Multiple measures that are closely related to each other such as entropy, mutual information, relative entropy, and Fisher information can be used for this purpose. However, their application to the dual control problem still requires careful evaluation as it will be illustrated in the subsequent chapters.

Knowledge Gain as Uncertainty Reduction

One possible way of quantifying the (estimated) knowledge gain $\mathcal{I}$ is defining it as a reduction in (estimated) uncertainty.

Definition III.1 (Knowledge Gain as Uncertainty Reduction)**.** *Given the set of observations $\mathcal{D}$ (2) of the dynamical system (1), the knowledge gain $\mathcal{I}$ at the point $\bar{z} := [\bar{x}, \bar{u}]^T$ is defined as*

$$\mathcal{I}(\bar{z}; \mathcal{D}) := H_{before}(\mathcal{D}) - H_{after}(\mathcal{D} \cup (\bar{z})), \tag{5}$$

where H_{before} is the entropy before the observation, $(\bar{z}, f(\bar{z}))$, and H_{after} is the one after it.

Thus, entropy is used here as a measure of uncertainty. Note that, this is related to the well-known mutual information between two random variables X and Y, $I(X;Y) = H(X) - H(X|Y)$, which quantifies reduction in the uncertainty of X due to knowledge of Y [26].

Remark III.2. In this paper, a consistent learning framework is assumed where the estimates improve with each new piece of knowledge acquired similar to a consistent estimator where estimates converge to the true value as number of data points increase to infinity.

Depending on the type of entropy measure used the knowledge gain $\mathcal{I}$ in (5) can be calculated in different ways. In order to simplify the discussion, consider the single variable case without any loss of generality. Let $H(X)$ denote the entropy of a discrete random variable X, and $h(X)$ the differential entropy of its continuous counterpart [26] (see Appendix for the definitions).

Differential entropy is closely related to entropy, yet there are important differences between the two. While the range of entropy $H(X)$ is $[0, \infty)$, differential entropy is a mapping to $(-\infty, \infty)$. This may lead to subtraction of two quantities potentially diverging to $-\infty$ when differential entropy is used to calculate $\mathcal{I}$ in (5) even under the consistent learner assumption in Remark III.2. Therefore, differential entropy is not a good measure to use in computing knowledge gain as reduction in uncertainty. Fortunately, the problem can be remedied simply by using entropy power [27] or alternatively exponential entropy [28] which are closely related to each other.

Definition III.3 (Entropy power)**.** *The entropy power [27] of a random vector X with differential entropy $h(X)$ is*

$$h_e(X) := \frac{1}{2\pi e} e^{\frac{2}{d}h(X)},$$

where d is the dimension of the random vector.

Since the range of exponential entropy $[0, \infty)$ is the same as of entropy, it can be used as a measure of uncertainty of a continuous variable, and hence to define knowledge gain based on entropy power:

$$\mathcal{I}_e := h_{e,before} - h_{e,after} = \frac{1}{2\pi e}\left[e^{\frac{2}{d}h_{before}} - e^{\frac{2}{d}h_{after}}\right]. \tag{6}$$

Relative Entropy

An alternative method for quantifying knowledge gain is to use relative entropy or Kullback-Leibler (K-L) divergence, which measures the difference between two probability distributions [26].

Definition III.4 (Relative Entropy)**.** *The K-L divergence between distributions p and q on $\mathcal{Y}$ is defined as*

$$D(p \parallel q) := \sum_{y \in \mathcal{Y}} p(y) \ln\left(\frac{p(y)}{q(y)}\right). \tag{7}$$

The $K - L$ divergence is not a symmetric quantity, $D(p \parallel q) \neq D(q \parallel p)$, and hence, is not a true distance metric.

Fisher Information

Let $f(X; \theta)$ be the parameterized pdf of the random variable X. The *score* is a random variable

$$score := \frac{\partial}{\partial \theta} \ln(f(X; \theta). \tag{8}$$

Definition III.5 (Fisher information)**.** *Fisher information [26] is the variance of the score defined in (8),*

$$F(\theta) := E_\theta \left[\frac{\partial}{\partial \theta} \ln(f(x;\theta) \right]^2,$$

where E_θ denotes the conditional expectation given θ.

Relationships between Information Measures

Mutual information and relative entropy are related to each other through the equation $I(X;Y) = D(p_{X,Y}||p_X p_Y)$, where p_X and p_Y are the pmfs of the random variables X and Y.

The well-known Cramer-Rao inequality [26] relates the variance $\mathrm{var}(\mathrm{T})$ of any unbiased estimator $T(X)$ of the parameter θ to Fisher information:

$$F(\theta) \geq \frac{1}{\mathrm{var}(\mathrm{T})}. \tag{9}$$

The de Bruijn's identity connects differential entropy and Fisher information as follows:

$$\frac{\partial}{\partial t} h(X + \sqrt{t}Z)\Big|_{t=0} = 0.5 F(X),$$

where Z is a zero mean unit variance Gaussian random variable and

$$F(X) = \int_{-\infty}^{+\infty} f(x) \left[\frac{\partial}{\partial x} \ln f(x) \right]^2 dx.$$

While the (differential) entropy is related to the volume V of its typical (or "support") set $\mathcal{A}$, Fisher information is related to the surface area of the typical set [29]

$$S(X) = \frac{\partial}{\partial t} h_e(X + \sqrt{t}Z)\Big|_{t=0} = 0.5 F(X) h_e(X),$$

or

$$\frac{\partial}{\partial t} \log\left(h_e(X + \sqrt{t}Z) \right)\Big|_{t=0} = 0.5 F(X).$$

Remark III.6. **Fisher information can be seen as a directional derivative of entropy and the knowledge gain $\mathcal{I}$ in Definition III.1 can be interpreted as a subgradient of entropy.** Therefore, the two are closely related as they both provide a direction for entropy, and hence uncertainty, reduction. Furthermore, minimizing variance (of the estimator) maximizes knowledge gain which follows from the Cramer-Rao bound (9) as well as the analysis in Section V-B.

The information metrics studied in this section are used for dual control of MDPs in Section IV and of nonlinear systems in Section V under different learning schemes. The choice of which specific information measure to use is heavily problem-dependent and constrained by computational limits. However, the analysis in this section provides insights to facilitate such decisions in various problems.

IV. Dual Control of Markov Decision Processes

Markov decision processes provide a mathematical machinery to describe systems whose evolution is partly due to the actions taken by a decision maker and partly random. There are many problems that can be formulated with this language and there is a rich history of applying MDPs to different problem areas. For example, in [30] a scenario is studied where at each time step a decision should be made on how much water to be used for electricity generation when alternative methods of generation exist. In [31] the problem of commodity acquisition in face of changing prices and expected demand pattern is studied. The problem of how much investment should be made by an insurance company in the presence of random claims, expenses, and stockbrokers cut-offs is considered in [32]. For more examples on practical applications of MDPs the reader may refer to [33].

To be able to use MDP to model a system a few assumptions should be made. First, the number of states that the system operates in should be finite. Second, the number of actions that the decision maker can take at each state are finite. Third, the Markov property should hold, i.e. the effects of an action taken in a state depend only on that state and not on the prior history. Moreover, it is assumed that a reward (penalty) is associated with each transition to a new state after applying an action and the goal would be to maximize (minimize) the sum of all rewards (penalties) over time.

Under the aforementioned assumptions the required variables are defined. Let $\mathcal{S} = \{s_1, \ldots, s_n\}$ be a finite set of states and $\mathcal{A} = \{a_1, \ldots, a_m\}$ be a finite set of actions. Let Π_i be the transition matrix when action a_i is applied where $\Pi_i(j,l)$, the jl-th entry of Π_i, is the probability of transition from state s_j to s_l under action a_i. Moreover, let $R(s,a)$ be the (expected) reward gained by applying action a at state s. Consider the following finite-horizon cost function

$$J(k_0, \pi(k_0)) = \sum_{k=k_0}^{k_0+N-1} R(s(k), a(k)). \tag{10}$$

where $\pi(k_0) = \{a(k_0), \ldots, a(k_0+N-1)\}$. The decision problem would be to find the sequence of actions $\pi(k_0)$ such that $J(k_0, \pi(k_0))$ is minimized. This problem is well-known and a vast body of research work has addressed this and other similar problems, see [34], [35] for a thorough treatment of such problems. However, the majority of those works make assumptions on a having perfect knowledge of the the problem, e.g. knowing $\Pi_1, \ldots, \Pi_m$ precisely. Note that Π_i is the true and unknown transition matrix when action a_i is applied.

In this section, the problem is addressed when the transition matrices $\Pi_1, \ldots, \Pi_m$ are unknown and need to be estimated while the expected reward based on the known estimates is to be maximized at each step. Hence, it is a special case of the general formulation in Section II.

A. Learning the Transition Matrices

Classical maximum likelihood (ML) or Maximum A Posteriori (MAP) schemes can be used to estimate the transition matrices of an MDP [36], [37]. To take advantage of the conjugacy property of Dirichlet distribution with the multinomial

distribution [38], a prior Dirichlet distribution is assumed on each row of the transition matrices. This facilitates recursively estimating the transition matrices via counting the transitions between each pair of states under an action to obtain ML or MAP estimates of the transition matrices, the reader may refer to [39], [40] for more information.

First, denote the estimate of Π_i as P_i. The mean-variance estimator of [41] is used to estimate the transition matrices $\Pi_1, \ldots, \Pi_m$. While other methods can be also used in this context, the same principles introduced here to maximize the knowledge gain will be applicable. To this end assume each row j of each estimate P_i denoted by $P_i(j,:)$ is a Dirichlet random variable with hyper-parameters $\alpha_i = [\alpha_i(j,1), \ldots, \alpha_i(j,n)]$ where $\alpha_i(j,1), \ldots, \alpha_i(j,n)$ are initially set to be positive real numbers. Additionally, assume that at time k an action $a(k) = a_i$ is applied and let $s(k) = s_j$. After the application of action a_i the estimate of the true transition matrix Π_i, P_i, will be updated as follows

$$\begin{aligned} P_i(j,l) &:= \alpha_i(j,l)/\bar{\alpha}_i(j) \\ \Sigma_{ll} &:= \frac{\alpha_i(j,l)(\bar{\alpha}_i(j) - \alpha_i(j,l))}{\bar{\alpha}_i(j)^2(\bar{\alpha}_i(j)+1)} \end{aligned} \tag{11}$$

where

$$\alpha_i(j,l) := \begin{cases} \alpha_i(j,l)+1 & \overline{T} = 1 \\ \alpha_i(j,l) & \overline{T} = 0 \end{cases} \tag{12}$$

with

$$\overline{T} = \left(s(k_0) = s_j \ \ \& \ \ s(k_0+1) = s_l \ \ \& \ \ a(k_0) = a_i\right),$$

and

$$\bar{\alpha}_i(j) := \sum_{l=1}^{n} \alpha_i(j,l). \tag{13}$$

We note that $\bar{\alpha}_i(j) > 0$ due to the fact that each $\alpha_i(j,l)$, $l = 1, \ldots, n$ remains positive before and after applying (12). Specifically, the Boolean variable $\overline{T}$ is true when under action a_i at time k a transition from state s_j to state s_l occurs.

The main problem is how to choose $a(k)$ in order to improve the performance of the estimator described above. Different strategies based on the information measures introduced in Section III are investigated next.

B. Quantifying Information

Knowledge gain as uncertainty reduction and Fisher information defined in Section III are used next to quantify knowledge acquired in the MDP setting for learning under a Dirichlet prior. As the first step, the entropy of a Dirichlet variable is given. If x is a Dirichlet random variable with hyper-parameters $\alpha = [\alpha_1, \ldots, \alpha_n]$, then the information entropy of x is

$$h(x) = \log B(\alpha) + (\bar{\alpha} - n)\psi(\bar{\alpha}) - \sum_{\ell=1}^{n} (\alpha_\ell - 1)\psi(\alpha_\ell) \tag{14}$$

where $\bar{\alpha} = \sum_{\ell=1}^{n} \alpha_\ell$, $\psi(\cdot)$ is the digamma function of its argument, and

$$B(\alpha) = \frac{\prod_{\ell=1}^{n} \Gamma(\alpha_\ell)}{\Gamma(\bar{\alpha})} \tag{15}$$

with $\Gamma(\cdot)$ being the gamma function of its argument.

The *expected knowledge gain in terms of entropy power* of applying an action a_i at state $s(k)$ is defined as

$$\mathcal{I}_e(s(k), a_i) = \sum_{\ell=1}^{n} P_i(j,\ell) \left[h_e\Big(P_i(j,:)\Big) - h_e\left(\widehat{P}_i^{\ell}(j,:)\right)\right] \tag{16}$$

where $s_j = s(k)$,

$$\begin{aligned} \widehat{P}_i^{\ell}(j,l) &= \widehat{\alpha}_i^{\ell}(j,l)/\bar{\alpha}_i^{\ell}(j) \\ \widehat{\Sigma}_{ll}^{\ell} &= \frac{\widehat{\alpha}_i^{\ell}(j,l)(\bar{\alpha}_i^{\ell}(j) - \widehat{\alpha}_i^{\ell}(j,l))}{\bar{\alpha}_i^{\ell}(j)^2(\bar{\alpha}_i^{\ell}(j)+1)} \end{aligned} \tag{17}$$

$$\widehat{\alpha}_i^{\ell}(j,l) = \begin{cases} \alpha_i(j,l)+1 & l = \ell \\ \alpha_i(j,l) & l \neq \ell \end{cases}, \tag{18}$$

and $\bar{\alpha}_i^{\ell}(j) = \bar{\alpha}_i(j) + 1$ for $\ell = 1, \ldots, n$.

Thus, at each time k, the best action $a(k)$ with respect to entropy power is given by

$$a(k) = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \ \mathcal{I}_e(s(k), a).$$

Next, consider the Fisher Information measure as an alternative. As above consider x to be a Dirichlet random variable with hyper-parameters $\alpha = [\alpha_1, \ldots, \alpha_n]$. The ij-th entry of the Fisher information matrix $F(x)$, $F_{ij}(x)$, in light of Definition III.5 is

$$F_{ij}(x) = \begin{cases} \psi'(\alpha_j) - \psi'(\alpha_0), & i = j \\ -\psi'(\alpha_0) & i \neq j \end{cases} \tag{19}$$

where $\psi'(\cdot)$ is the trigamma function, i.e. it is the first derivative of the digamma function.

The *expected knowledge gain in terms of Fisher information* of applying an action a_i at state $s(k)$ is defined as

$$\mathcal{I}^F(s(k), a_i) = \sum_{\ell=1}^{n} P_i(j,\ell) \left[\text{tr}\Big(F(P_i(j,:))\Big) - \text{tr}\left(F(\widehat{P}_i^{\ell}(j,:))\right)\right] \tag{20}$$

where tr$(\cdot)$ is the trace of its argument, $s_j = s(k)$, $\widehat{P}_i^{\ell}(j,l)$ is given by (17), $\widehat{\alpha}_i^{\ell}(j,l)$ is described by (18), and $\bar{\alpha}_i^{\ell}(j) = \bar{\alpha}_i(j) + 1$ for $\ell = 1, \ldots, n$.

C. Dual Control of MDP

Consider now a Markov decision problem where the transition matrices are unknown, which corresponds to dual control of a MDP. Taking inspiration from model predictive control strategies, define at time k_0

$$\overline{J}(k_0, \pi(k), P_i) = w_{\mathcal{D}} J(k_0, \pi(k)) + w_{\mathcal{I}} \mathcal{I}(s(k_0), a(k_0)). \tag{21}$$

where $w_{\mathcal{D}}$ and $w_{\mathcal{I}}$ are some nonnegative weighting scalars, $\mathcal{I}(s(k_0), a(k_0))$ is the expected knowledge gain after applying action $a(k_0)$ at time k_0 and state $s(k_0)$ and can be either defined by (16) or (20). We rewrite (21) as

$$\overline{J}(k_0, \pi(k_0), P_i) = \sum_{k=k_0}^{k_0+N} \overline{R}(s(k), a(k)) \tag{22}$$

with

$$\overline{R}(s(k_0), a(k_0)) = w_{\mathcal{D}} R_{a(k_0)}(s(k_0), a(k_0)) + w_{\mathcal{I}} \mathcal{I}(s(k_0), a(k_0)),$$
$$\overline{R}(s(k), a(k)) = w_{\mathcal{D}} R(s(k), a(k)),\ k = k_0 + 1 : k_0 + N - 1.$$

Now the control strategy is to first apply the first action in the sequence $\pi^\star(k_0) = \{a^\star(k_0), \ldots, a^\star(k_0 + N - 1)\}$, $a^\star(k_0)$, at time k_0 where

$$\pi^\star(k_0) = \underset{\pi(k_0)}{\operatorname{argmax}}\, \overline{J}(k_0, \pi(k_0), P_i). \tag{23}$$

Secondly, update the estimates $P_1, \ldots, P_m$ by observing the outcome of the action via (11)-(13), form $\overline{J}(k_0 + 1, \pi(k_0 + 1), P_i)$ using the new estimates and repeat the procedure. This dual control strategy for Markov Decision processes with unknown transition matrices is outlined in Algorithm 1.

Algorithm 1 Dual Control of Markov Decision Processes with Unknown Transition Matrices.

Input: $\mathcal{S}$, $\mathcal{A}$, $w_{\mathcal{D}}$, $w_{\mathcal{I}}$, $R(s,a)$, $\mathcal{I}(s,a)$
  initialize $P_1, \ldots, P_m$
  $s := s(0)$
  $\pi^\star(0) := \underset{\pi(0)}{\operatorname{argmax}}\, \overline{J}(0, \pi(0), P_i)$
  $\{\pi^\star(k) = \{a^\star(k), \ldots, a^\star(k + N - 1)\}\}$
  apply $a := a^\star(0)$
  for $k_0 \in \{1, 2, \ldots\}$ **do**
    $s := s(k_0)$
    update $P_1, \ldots, P_m$ via (11)-(13)
    $\pi^\star(k_0) := \underset{\pi(k_0)}{\operatorname{argmax}}\, \overline{J}(k_0, \pi(k_0), P_i)$
    apply $a := a^\star(k_0)$
  end for

V. Dual Control of Nonlinear Systems

In the dual control problem formulated, information is needed to learn the dynamical system and the learning framework provides the basis for explicitly quantifying and actively obtaining the next piece of information.

A. Learning the System Dynamics using GP

A short overview of Gaussian Process (GP) regression is presented next for completeness [2], [42]. Deriving the estimate $\hat{f}$ using the set of available data points, $\mathcal{D}$, is known as the *regression* problem in the pattern recognition (machine learning) literature. This task falls under the umbrella of supervised learning since $\mathcal{D}$ is used here as a labeled learning set. The learning process involves selection of an a-priori "model" which allows the learned function $\hat{f}$ to be expressed in terms of a set of parameters and specific basis functions. At the same time an error measure between the original function $\mathbf{f}$ and $\hat{f}$ is minimized using the learning data set. GP regression provides a "non-parametric" version of this basic idea.

A GP is formally defined as a collection of random variables, any finite number of which have a joint Gaussian distribution. In general, it is completely specified by its mean function $m(z)$ which is assumed to be zero here for simplicity, and covariance function

$$c_i : (\mathcal{Z}, \mathcal{Z}) \to \mathbb{R},\ \ c_i(z, \tilde{z}) := E[f_i(z) f_i(\tilde{z})],$$

for $i = 1, \ldots, d$ and $\forall z, \tilde{z} \in \mathcal{Z}$. Hence, the GP is characterized in this special case entirely by its covariance function $c(z, \tilde{z})$. Given a set of data $\mathcal{D}$ and assuming fixed Gaussian observation noise, the covariance matrix is defined as the sum of a *kernel matrix* Q and noise variance σ:

$$C_{ij} := Q_{ij} + \sigma,\ \forall i, j = 1, \ldots, \mathrm{card}(\mathcal{D}) \tag{24}$$

where $\mathrm{card}(\mathcal{D})$ is the cardinality of the data set $\mathcal{D}$. While it is possible to choose here any (positive definite) kernel function $q(z, \tilde{z}) : (\mathcal{Z}, \mathcal{Z}) \to \mathbb{R}$, one classical choice is the Gaussian kernel,

$$q(z, \tilde{z}) = \exp\left[-\frac{1}{2} \|z - \tilde{z}\|^2\right], \tag{25}$$

which leads to the kernel matrix $Q_{ij} = q(z_i, z_j)$, where $z_i,\ z_j \in \mathcal{D}$. Note that GP makes use of the well-known *kernel trick* here by representing an infinite dimensional continuous function using a (finite) set of continuous basis functions and associated vector of real parameters in accordance with the *representer theorem* [10].

Given the data set $\mathcal{D}$, define the vector

$$k(\mathcal{D}, z) := [q(z_1, z), \ldots q(z_D, z)] \tag{26}$$

and scalar

$$\kappa := q(z, z) + \sigma = 1 + \sigma. \tag{27}$$

Then, the predictive distribution at a given point z, $p_{\hat{x}_i}(\mathcal{D}, z)$, is a Gaussian random variable, $\mathcal{N}(\hat{f}_i, v_i)$, with the mean $\hat{f}_i$ and variance v:

$$\hat{f}_i(\mathcal{D}, z) := k^T C^{-1} \bar{f}_i(\mathcal{D}) \text{ and } v(\mathcal{D}, z) := \kappa - k^T C^{-1} k, \tag{28}$$

where $\bar{f}_i(\mathcal{D}) = [f_i(z_0), f_i(z_2), \ldots, f_i(z_{D-1})]^T$. Note that the variance is independent of the individual state dimension. Equation (28) is a key result that defines GP regression. The mean function $\hat{f}_i(z)$ of the Gaussian distribution provides a prediction of the objective function $\mathbf{f}(z)$ in each dimension i. The variance function $v(z)$ indicates the uncertainty level of the predictions provided by $\hat{f}$.

B. Quantifying Information

Two specific definitions of information, mutual information and relative entropy (Kullback-Leibler distance), are presented next using the GPR learning framework.

Mutual Information Metric: One possible way of quantifying the estimated knowledge gain from choosing a control action $\bar{u} \in \mathcal{U}$ at any given state $\bar{x} \in \mathcal{X}$, is defining it as a reduction in estimated uncertainty at $\bar{z} := [\bar{x}, \bar{u}]^T$ using (5). Based on GP regression and given the observed data $\mathcal{D}$, each state dimension, $\hat{x}_i$, is estimated as a multivariate random variable, $\mathcal{N}(\hat{f}_i, v_i)$, with *predictive distribution*

$$p_{\hat{x}_i}(\mathcal{D}, z) = \frac{1}{\sqrt{2\pi v_i}} \exp\left(-\frac{1}{2} \frac{(\hat{x}_i - \hat{f}_i)^2}{v}\right), \tag{29}$$

where $\hat{f}_i$ and v are defined in (28), i.e. this is the conditional distribution of $\hat{f}_i(z)$ given data $\mathcal{D}$.

The differential entropy of the predictive Gaussian distribution (29) at any point $z \in \mathcal{Z}$ before observing the new data point, $\bar{z}$, is

$$H_b(\mathcal{D}, z) = \frac{d}{2}\ln(2\pi e) + \frac{d}{2}\ln\left(v(\mathcal{D}, z)\right),$$

and the one after observing it is

$$H_a(\mathcal{D} \cup \{\bar{z}\}, z) = \frac{d}{2}\ln(2\pi e) + \frac{d}{2}\ln\left(v(\mathcal{D} \cup \{\bar{z}\}, z)\right).$$

Here, the entropy value is the same and added over each d dimensions of the state space. Note again that the entropy is independent from the value of the system mapping $\mathbf{f}(\bar{z})$ [43].

The functions H_b and H_a quantify the uncertainty at a given point $z \in \mathcal{Z}$. The aggregate uncertainty values H_{before} and H_{after}, however, are computed over the whole space $\mathcal{Z}$. In order to simplify the analysis and notation, define a dense sampling of $\mathcal{Z}$, $\Theta := \{z_1, \ldots, z_T : z_i \in \mathcal{Z},\ z_i \notin \mathcal{D},\ \forall i\}$ [1]. Then, an proxy for H_{before} and H_{after} are defined as follows:

$$\hat{H}_{before}(\mathcal{D}) := \sum_{z \in \Theta} H_b(\mathcal{D}, z)$$

and

$$\hat{H}_{after}(\mathcal{D} \cup \{\bar{z}\}) := \sum_{z \in \Theta \setminus \bar{z}} H_a(\mathcal{D} \cup \{\bar{z}\}, z) + \frac{d}{2}\ln(\sigma),$$

where σ in (24) is the observation noise.

Let,

$$\hat{H}_{before}(\mathcal{D}) = H_b(\mathcal{D}, \bar{z}) + \sum_{z \in \Theta \setminus \bar{z}} H_b(\mathcal{D}, z).$$

The following lemma provides the basis for a useful approximation.

Lemma V.1. *In the problem setup considered, given the data set $\mathcal{D}$ and the next observation $\bar{z} \in \mathcal{Z},\ \bar{z} \notin \mathcal{D}$, define*

$$\rho(\bar{z}) := \sum_{z \in \Theta \setminus \bar{z}} H_b(\mathcal{D}, z) - H_a(\mathcal{D} \cup \{\bar{z}\}, z),$$

which is equivalent to

$$\rho(\bar{z}) = \frac{d}{2} \sum_{z \in \Theta \setminus \bar{z}} \ln\left(v(\mathcal{D}, z)\right) - \ln\left(v(\mathcal{D} \cup \{\bar{z}\}, z)\right).$$

Then, the following hold:

1) $\rho(\bar{z}) \geq 0$.
2) $\rho(\bar{z}) \to 0$ *as* $\mathrm{card}(\mathcal{D}) \to \infty$.

Proof. Define [42]:

$$C_{D+1}^{-1} = \begin{bmatrix} M & m \\ m^T & \mu \end{bmatrix}, \tag{30}$$

where

$$\begin{aligned} \bar{k} := Q(z, \bar{z}),\ \ \mu := v^{-1}, \\ k_{D+1} = [k_D,\ \bar{k}]^T,\ \ m := -\mu C_D^{-1} k, \\ M := C_D^{-1} + \frac{1}{\mu} m m^T,\ \ \varepsilon := \kappa - \bar{k}. \end{aligned} \tag{31}$$

It follows directly from (26) and (28) that

$$v(\mathcal{D} \cup \{\bar{z}\}, z) = \kappa - k_{D+1}^T C_{D+1}^{-1} k_{D+1}.$$

Straightforward algebraic manipulations lead to

$$v(\mathcal{D} \cup \{\bar{z}\}, z) = v \cdot \left(1 - \frac{(v - \varepsilon)^2}{v^2}\right).$$

The ratio of variances of the predictive distributions before and after observing $\bar{z}$ at any point z is then

$$R(\mathcal{D}, \bar{z}, z) := \frac{v(\mathcal{D}, z)}{v(\mathcal{D} \cup \{\bar{z}\}, z)} = \frac{v}{v \cdot \left(1 - \frac{(v - \varepsilon)^2}{v^2}\right)}. \tag{32}$$

Since $\varepsilon \geq 0$ (otherwise $v(\mathcal{D} \cup \{\bar{z}\}, z)$ would be negative), the ratio (32) has to be greater than one, $R(\mathcal{D}, \bar{z}, z) \geq 1$, which establishes the result in part 1.

As $\mathrm{card}(\mathcal{D}) \to \infty$ in part 2, both the variance v and variable ε converge to σ due to consistency property of GP regression [2]. Consequently, $R(\mathcal{D}, \bar{z}, z)$ convergences to one, and hence,

$$\rho(\bar{z}) = \frac{d}{2} \sum_{z \in \Theta \setminus \bar{z}} \ln\left(R(\mathcal{D}, \bar{z}, z)\right) \to 0,$$

which completes the proof. □

Using Lemma V.1, the expected knowledge gain that approximates (5) is

$$\begin{aligned} \hat{\mathcal{I}}(\bar{z}; \mathcal{D}) &= \frac{d}{2}\ln\left(v(\mathcal{D}, \bar{z})\right) + \rho(\bar{z}) + \frac{d}{2}\ln(2\pi e/\sigma). \\ &\Rightarrow \hat{\mathcal{I}}(\bar{z}; \mathcal{D}) \geq \frac{d}{2}\ln\left(v(\mathcal{D}, \bar{z})\right) + \frac{d}{2}\ln(2\pi e/\sigma). \end{aligned} \tag{33}$$

This inequality is independent of the sampling Θ which leads to the following result:

Proposition V.2. *Given $\mathcal{D}$ and GP model with predictive distribution (29), a lower bound on the estimated knowledge gain from observation, $(\bar{z})$, is*

$$\mathcal{I}(\bar{z}; \mathcal{D}) \geq \frac{d}{2}\ln\left(v(\mathcal{D}, \bar{z})\right) + \frac{d}{2}\ln(2\pi e/\sigma),$$

where $v(\mathcal{D}, \bar{z})$ is defined in (28). Moreover, this bound becomes tighter as $\mathrm{card}(\mathcal{D}) \to \infty$.

Remark V.3. It is interesting to note that the knowledge gain is independent of the system dynamics, f, i.e. only the observation point $\bar{z}$ plays a role in the definition; see e.g. (5) and (33).

Corollary V.4. *Given $\mathcal{D}$ and GP model with predictive distribution (29), a lower bound on the estimated knowledge gain based on entropy power from observation, $(\bar{z})$, is*

$$\mathcal{I}_e(\bar{z}; \mathcal{D}) \geq v(\mathcal{D}, \bar{z}),$$

where $v(\mathcal{D}, \bar{z})$ is defined in (28). Moreover, this bound becomes tighter as $\mathrm{card}(\mathcal{D}) \to \infty$.

Proof. The proof follows directly from the definition of knowledge gain based on entropy power (6) and Lemma V.1. □

Relative Entropy (K-L) Metric: If p and q are both Gaussian distributions, then the K-L divergence (7) between them is

$$D(p \parallel q) = \frac{1}{2}\left[\ln\left(\frac{v_q}{v_p}\right) + \frac{v_p}{v_q} + \frac{(m_p - m_q)^2}{v_q} - 1\right], \quad (34)$$

where $m_p,\ m_q$ and $v_p,\ v_q$ are the respective means and variances.

Using (29), define the predictive distribution $p_b(\mathcal{D}, z)$ before observing $\bar{z}$ and $p_a(\mathcal{D} \cup \{\bar{z}\}, z)$ after the observation. A new problem that was not encountered before with the entropy-based knowledge metric arises next. When deciding which $\bar{z}$ to choose, it is not possible to know beforehand the value $\mathbf{f}(\bar{z})$, which is now explicitly a part of this knowledge metric. This complication can be addressed in two different ways: (a) by ignoring the means of p_a and p_b or (b) by replacing $\mathbf{f}(\bar{z})$ with $\hat{f}(\bar{z})$ using the existing data $\mathcal{D}$. For convenience, the former is called for the rest of the paper as *zero-mean K-L knowledge* and the latter as *approximate K-L knowledge*.

Following similar steps as before, the **zero-mean K-L knowledge** metric over Θ is defined as

$$\hat{\mathcal{I}}_{KLz}(\bar{z}; \mathcal{D}) = \sum_{z \in \Theta} D(p_a(\mathcal{D} \cup \{\bar{z}\}, z) \parallel p_b(\mathcal{D}, z)) \quad (35)$$

$$= \sum_{z \in \Theta} \frac{d}{2}\left[\ln\left(\frac{v_{p_b}}{v_{p_a}}\right) + \frac{v_{p_a}}{v_{p_b}} - 1\right], \quad (36)$$

where $v_{p_a} = v(\mathcal{D} \cup \{\bar{z}\}, z)$ and $v_{p_b} = v(\mathcal{D}, \bar{z})$. Note that, $v(\mathcal{D} \cup \{\bar{z}\}, \bar{z}) = \sigma$ in accordance with (24).

Define next the quantity

$$\eta(\mathcal{D}, \bar{z}) := \sum_{z \in \Theta \setminus \bar{z}} \frac{d}{2}\left[\ln\left(\frac{v_{p_b}}{v_{p_a}}\right) + \frac{v_{p_a}}{v_{p_b}} - 1\right], \quad (37)$$

$$= \sum_{z \in \Theta \setminus \bar{z}} \frac{d}{2}\left[\ln(R) + \frac{1}{R} - 1\right],$$

where R is given in (32).

Proposition V.5. *Given $\mathcal{D}$ and GP model with predictive distribution (29), a lower bound on the estimated zero-mean K-L knowledge gain from observation, $(\bar{z})$, as defined in (35) is*

$$\mathcal{I}_{KLz}(\bar{z}; \mathcal{D}) \geq \frac{d}{2}\left[\ln\left(v(\mathcal{D}, \bar{z})\right) + \frac{\sigma}{v(\mathcal{D}, \bar{z})}\right] + \frac{d}{2}\left(\log(1/\sigma) - 1\right),$$

where $v(\mathcal{D}, \bar{z})$ is defined in (28). Moreover, this bound becomes tighter as $\mathrm{card}(\mathcal{D}) \to \infty$.

Proof. The inequality follows from

$$\mathcal{I}_{KLz}(\bar{z}; \mathcal{D}) = \frac{d}{2}\ln\left(v(\mathcal{D}, \bar{z})\right) + \eta(\mathcal{D}, \bar{z}) + \frac{d}{2}\left(\log(1/\sigma) - 1\right),$$

where $\eta(\mathcal{D}, \bar{z})$ is defined in (37). Since $R \geq 1$ from (32), η is always non-negative, $\eta \geq 0$, and the inequality holds. Furthermore, following an argument similar to the one in the proof of Lemma V.1, η converges to zero as $\mathrm{card}(\mathcal{D}) \to \infty$, which completes the proof. □

The **approximate K-L knowledge** extends the definition in (35) as follows:

$$\hat{\mathcal{I}}_{KLa}(\bar{z}; \mathcal{D}) = \sum_{z \in \Theta} D(p_a(\mathcal{D} \cup \{\bar{z}\}, z) \parallel p_b(\mathcal{D}, z))$$
$$= \sum_{z \in \Theta}\left(\frac{d}{2}\left[\ln\left(\frac{v_{p_b}}{v_{p_a}}\right) + \frac{v_{p_a}}{v_{p_b}} - 1\right] + \sum_{i=1}^{d} \frac{(m_{a,i} - m_{b,i})^2}{2v_{p_b}}\right), \quad (38)$$

where $v_{p_a} = v(\mathcal{D} \cup \{\bar{z}\}, z)$, $v_{p_b} = v(\mathcal{D}, \bar{z})$, $m_{a,i} = \hat{f}_i(\mathcal{D} \cup \{\bar{z}\}, z)$, and $m_{b,i} = \hat{f}_i(\mathcal{D}, z)$ as in (28).

Define next the quantity ζ similar to η. Straightforward algebraic manipulations lead to:

$$\zeta(\mathcal{D}, \bar{z}) := \sum_{z \in \Theta \setminus \bar{z}} D(p_a(\mathcal{D} \cup \{\bar{z}\}, z) \parallel p_b(\mathcal{D}, z)),$$
$$= \sum_{z \in \Theta \setminus \bar{z}}\left(\frac{d}{2}\left[\ln(R) + \frac{1}{R} - 1\right] + \sum_{i=1}^{d} \frac{(v - \varepsilon)^2}{2v^3}(\hat{f}_i(\bar{z}) - \hat{f}_i(z))^2\right), \quad (39)$$

where R is given in (32), ε in (31), and $v = v(\mathcal{D}, \bar{z})$. Since $\mathbf{f}(\bar{z})$ is naturally not available when deciding on which $\bar{z}$ to choose, it is replaced by $\hat{f}_i(\bar{z})$ above $\forall i$. Since $\zeta \geq 0$ and $\zeta \to 0$ as $\mathrm{card}(\mathcal{D}) \to \infty$, the following counterpart of Proposition V.5 immediately follows:

Proposition V.6. *Given $\mathcal{D}$ and GP model with predictive distribution (29), a lower bound on the estimated approximate K-L knowledge gain as defined in (38) from observation, $(\bar{z})$, is*

$$\mathcal{I}_{KLa}(\bar{z}; \mathcal{D}) \geq \frac{d}{2}\left[\ln\left(v(\mathcal{D}, \bar{z})\right) + \frac{\sigma}{v(\mathcal{D}, \bar{z})}\right] + \frac{d}{2}\left(\log(1/\sigma) - 1\right),$$

where $v(\mathcal{D}, \bar{z})$ is defined in (28). Moreover, this bound becomes tighter as $\mathrm{card}(\mathcal{D}) \to \infty$.

Proof. Similar to that of Proposition V.5. □

It is interesting to note that the lower bounds in both Proposition V.5 and V.6 are the same.

C. Dual Control

Using the results from the previous section, the dual control problem is formulated as

$$\min_{u(0:N-1)} J(x(0), u(0:N-1), \hat{x}(1:N)) \quad (40)$$
$$= \min_{u(0:N-1)} \sum_{k=1}^{N} \beta^k \left[w_{\mathcal{C}}(k) J_k(\hat{x}(k-1), u(k-1), \hat{x}(k)) - w_{\mathcal{I}}(k)\mathcal{I}_k(\hat{x}(k-1), u(k-1), \hat{x}(k), \mathcal{D}(k))\right],$$
such that $\hat{x}(n+1) = \hat{f}(\hat{x}(n), u(n)),\ \ n = 0, \ldots, N-1,$

where $\hat{x}(0) = x(0)$, $\mathcal{I}_k \geq 0$ is the expected knowledge gain at time k. The weighting factors $w_{\mathcal{C}}(k)$ and $w_{\mathcal{I}}(k)$ are nonnegative values and capture the balance between control and knowledge gain dual objectives. Note that, unlike the stage cost J_k, the expected knowledge gain $\mathcal{I}_k$ is time-varying due to $\mathcal{D}(k)$. In order to remedy computational complications this creates, a special case can be considered where $w_{\mathcal{I}}(1) > 0$ and $w_{\mathcal{I}}(k) = 0,\ \ k = 2, \ldots, N$. As in the case of MPC,

the dual control problem (40) is solved repeatedly at each time step to choose the best control action u from a finite set despite calculating the future controls over the horizon. The dual control strategy for nonlinear systems using GPR is described in Algorithm 2.

The choice of weights $w_{\mathcal{C}}$ and $w_{\mathcal{I}}$ poses an interesting research question, especially if they are time-varying. There are multiple options and interpretations. If $w_{\mathcal{C}} = 0$ and $w_{\mathcal{I}} > 0$ until a certain time instance and changed to $w_{\mathcal{C}} > 0$ and $w_{\mathcal{I}} = 0$ afterwards, this special case corresponds to first learning the system and then controlling it based on the knowledge obtained, which is similar to conventional system identification. An alternative is starting from the same initial condition but gradually decreasing $w_{\mathcal{I}}$ while increasing $w_{\mathcal{C}}$. This is then similar to "cooling" in simulated annealing [44]. Yet another and more dynamic solution is inspired by how window lengths are set in TCP-IP. At each step the measured value of the system state is compared to the estimation obtained from the GPR algorithm and an estimation error is calculated. If the estimation error is less than a pre-determined threshold, then the value of $w_{\mathcal{I}}$ is decreased. However, if the estimation error is larger than the threshold the value of $w_{\mathcal{I}}$ is set to its original value. This dynamic scheme can be interpreted as active learning based on need.

Algorithm 2 Dual Control of Nonlinear Systems using GPR

Input: J, $\mathcal{U}$, $\mathcal{X}$, w_c, w_I, β.
  initialize $\hat{x}(0) = x(0)$, $\mathcal{D}$, select $u(0)$.
  for $k \in \{1, \dots\}$ **do**
    observe $x(k)$.
    extend data set $\mathcal{D} = \mathcal{D} \cup ([x(k-1), u(k-1)], x(k))$.
    update the system estimate $\hat{f}$ using GPR (29).
    choose $u(k)$ solving (40); apply as next control action.
  end for

VI. Numerical Examples

The applicability of the framework developed is illustrated with three different problems. Firstly, the methodology presented in Section IV is applied to an MDP where the transition matrices are not known. Secondly, the algorithm of Section V is used to address the problem of controlling an unknown nonlinear system with linear inputs. Finally, the framework is applied to dual control of an unknown nonlinear system with nonlinear control inputs. The `Python` and `Matlab` scripts to generate the results in this section can be found at [45].

A. Dual Control of a Markov Decision Process

This section presents two specific scenarios to illustrate the application of the presented dual control framework to Markov Decision Processes (MDPs) as discussed in Section IV. The first scenario investigates the problem of estimating the transition matrices of an MDP with $n = 5$ states and $m = 10$ actions. The error between the entries of the estimated transition matrices and their real values are compared for two cases. In the first case at each step k an action is picked randomly; in the second case at each step the action that maximizes the knowledge gain is chosen and applied to the system. The average of the estimation error $\sum_{i=1}^{m} \sum_{j=1,l=1}^{j=n,l=n} |\Pi_i(j,l) - P_i(j,l)|$, after running the experiment for 100 times for different MDPs is depicted in Fig. 1. The estimation error of the proposed method is consistently less than the error where the actions are picked randomly. This exhibits the performance gain obtained by implementing the learning strategy based on applying $a(k) = \underset{a \in \mathcal{A}}{\text{argmax}}\ \mathcal{I}_e(s(k), a)$ at each step k.

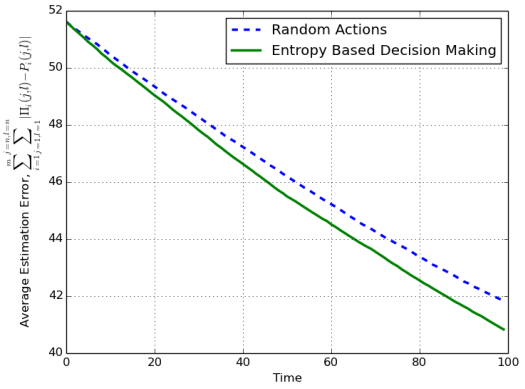

Fig. 1. The average estimation error after repeating the experiment for 100 times as actions are applied randomly and in a way that maximizes the knowledge gain.

The second scenario addresses an MDP problem where the transition matrices are not known. The number of states is $n = 5$ and there are $m = 5$ actions. The finite horizon length is $N = 10$. The weight of the knowledge gain in (21), $w_{\mathcal{I}} = 10$ and $\mathcal{I}(s(k_0), a(k_0)) = \mathcal{I}_e(s(k_0), a(k_0))$ given by (16). The Algorithm 1 is applied to this problem. The result is depicted in Fig. 2 is compared with the case where $w_{\mathcal{I}} = 0$ and the true transition matrices are used. The experiment is run 100 times. It is observed that the rewards obtained from applying the proposed method in this paper converges to the one achieved via implementing the optimal strategy with exact information about the transition matrices associated with each action after nearly 200 steps.

B. Dual Control of a Logistic Map with Linear Input

The framework developed in Section V is applied to dual control of a logistic map with linear input. The logistic map is controlled with additive actions while being identified using the GP method described in Algorithm 2:

$$x(n+1) = r\, x(n)\,(1 - x(n)) + u(n).$$

The controller knows here that the control is linear (additive), and utilizes this extra knowledge in identifying the system which simplifies the problem significantly. The system description (input-output relationship) from the perspective of the controller is:

$$y(n+1) = \hat{h}(y(n)) + u(n).$$

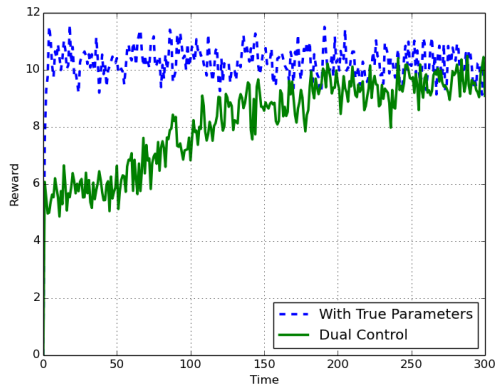

Fig. 2. The reward obtained by applying Algorithm 1 compared to the gain where a receding horizon problem is applied to the system with known transition matrices.

The control actions are taken from the finite set

$$\mathcal{U} = \{u_i \in [-1, 1] : u_{i+1} = u_i + 0.02,\ i = 1, \ldots, 101\}.$$

The kernel variance is 0.5 and the weights in the objective function (40) are chosen as $w_{\mathcal{I}} = w_{\mathcal{C}} = 1$. The goal is stabilize the system at $x^* = 0.8$, which constitutes the constant reference signal. The starting point is $x_0 = 0.1$. The control actions and state estimation errors over time (in each step based on arrived data points) for $r = 3.5$ and the corresponding trajectory of the logistic map are depicted in Figures 3 and 4. Note that, in this case the logistic map acts only as a nonlinear system with a limit cycle rather than behaving chaotically. It is observed that approximately the first 10 steps are used by the algorithm to explore or learn the system after which the trajectory approaches to the target. The Fig. 5 shows the estimated function versus the original mapping for $u = 0$ as well as one standard deviation from estimated value. It can be seen that the variance is minimum, i.e. the estimate is best, around the target value.

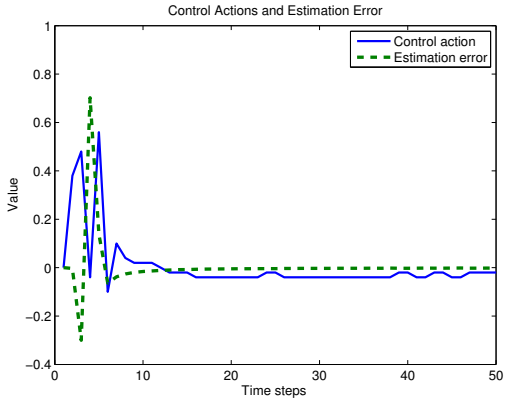

Fig. 3. The control actions and state estimation errors for logistic map with $r = 3.5$ and linear control.

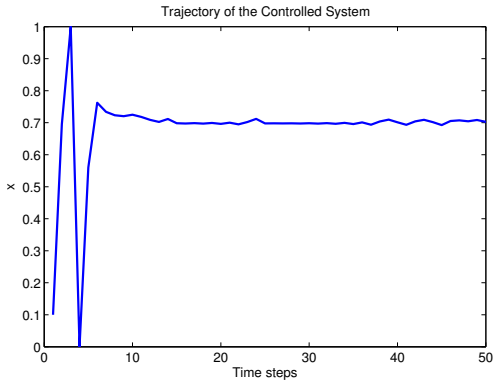

Fig. 4. The controlled trajectory of the logistic map for $r = 3.5$ and linear control.

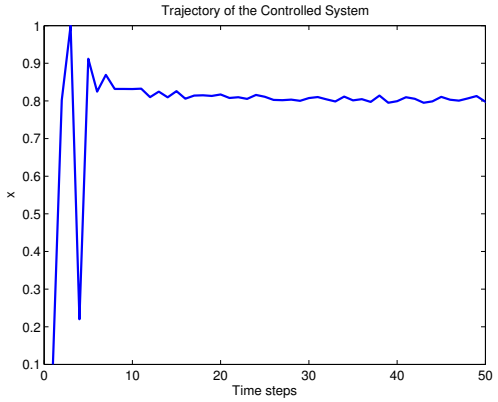

Fig. 5. The logistic map and its estimate along with one standard deviation for $u = 0$ and $r = 3.5$ after 100 iterations (data points).

C. Dual Control of a Cart with Inverted Pendulum

The framework presented in Section V is next applied to the problem of controlling the position of a cart with an inverted pendulum governed by the set of equations [46], [47]:

$$x_1(n+1) = x_1(n) + T\, x_2(n), \tag{41}$$

$$\begin{aligned} x_2(n+1) = \quad & x_2(n) + \frac{T}{M + m\sin^2(x_3(n))} \left[\mathbf{u}(n)\right. \\ & + mLx_4^2(n)\sin(x_3(n)) - bx_2(n) \\ & \left. - mg\cos(x_3(n))\sin(x_3(n))\right], \end{aligned}$$

$$x_3(n+1) = x_3(n) + T\, x_4(n), \tag{42}$$

$$\begin{aligned} x_4(n+1) = \quad & x_4(n) + \frac{T}{L\left(M + m\sin^2(x_3(n))\right)} \\ & \left[-\mathbf{u}(n)\cos(x_3(n)) + (M+m)g\sin(x_3(n))\right. \\ & + bx_2(n)\cos(x_3(n)) \\ & \left. - mLx_4^2(n)\cos(x_3(n))\sin(x_3(n))\right], \end{aligned} \tag{43}$$

$$y(n) = x_1(n), \tag{44}$$

where $T = 0.05$ is the sampling period, $y = x_1$ is the position of the cart, $x_2 \approx dx/dt$ is the cart velocity $x_3 = \theta$ is the inverted pendulum angle, $x_4 \approx d\theta/dt$ is the angular velocity.

The parameter values are: $b = 12.98$, $M = 1.378$, $L = 0.325$, $g = 9.8$, and $m = 0.051$.

The control objective is to fix the position of the cart to $y^* = 0.5$, i.e. $J_n(\hat{x}(n)) = |\hat{x}_1(n) - 0.5|$ at time step n and the variance v in (28) is used to approximate the knowledge gain objective $\mathcal{I}_n$ in (5). The cart is controlled adopting a one-step look-ahead strategy using control actions $u \in \{-10, -9, \ldots, 9, 10\}$. The weights in the objective function (40) are chosen as $w_{\mathcal{I}} = 1$ and $w_{\mathcal{C}} = 20$. To simplify the problem, it is assumed that the controller knows (41) and (44), but has to learn the main system dynamics (42)-(43).

The actual and the estimated position of the cart as well as its velocity are depicted in Fig. 6 and the selected control input is shown in Fig. 7. The performance is rather satisfactory considering that the trajectory is within 10% distance of the target within 30 steps. It is important to note that the controller does not know the model (42)-(43) at all here, and learns (a representation of) the system while controlling it at the same time.

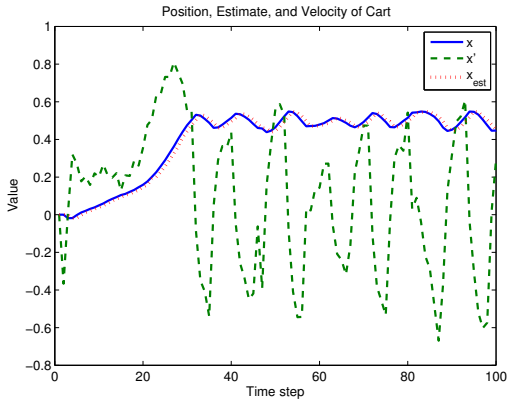

Fig. 6. The actual and estimated positions and velocity of the cart with inverted pendulum.

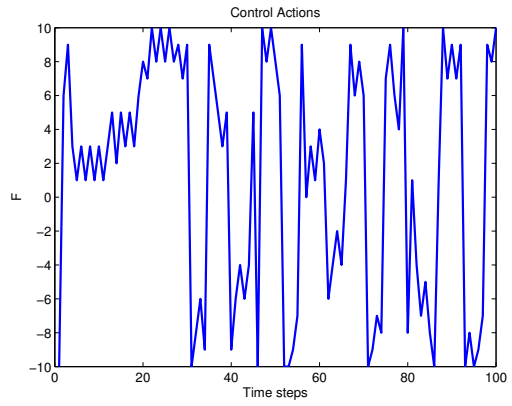

Fig. 7. The magnitude of the control input, u.

The performance of the dual controller is further evaluated by comparing it to that of the one-step look ahead controller with full information. The full information controller knows the entire system dynamics (41)-(44). The resulting position and velocity of the cart are depicted in Fig. 8 and the selected control input under full information is shown in Fig. 9. It is worth noting that the difference between the dual control and full information cases is rather small considering that the dual controller has to learn the dynamics (42)-(43) while concurrently trying to achieve the chosen position objective.

While the one-step look ahead strategy (under full or limited information) may not result in the best controller for addressing the given "cart with an inverted pendulum" problem, that is not the main point of the example here. The example system and the control strategy are merely used as a way to illustrate how learning and information can be integrated to (dual) control decisions when system dynamics are not available to the controller.

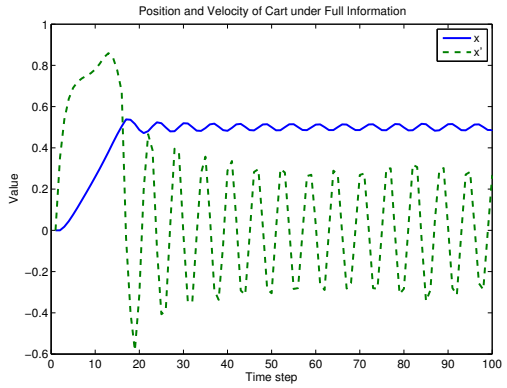

Fig. 8. The actual position and velocity of the cart with inverted pendulum under full information control.

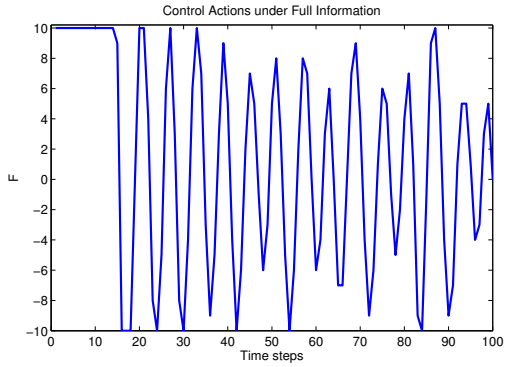

Fig. 9. The magnitude of the control input, u, under full information.

VII. Conclusion

A novel multi-objective optimization approach to dual control is presented where the active learning objective is explicitly quantified using measures from information theory. Specifically, entropy-based uncertainty reduction, Fisher information, and relative entropy are used to quantify the knowledge gain, which is balanced against a standard finite

horizon control objective. The framework is applied to Markov Decision Processes and discrete-time nonlinear systems via a Dirichlet prior and Gaussian Process Regression as respective learning methods. Thus, the broad applicability and usefulness of the presented approach is demonstrated in diverse problem settings. In addition, the links between uncertainty variance in learning and various information-theoretic measures are investigated.

Although dual control is a well-known problem, the presented approach opens a new direction for exploring decision making under limited information [1], [48], [49] by integrating methods from information and statistical learning theories to control theory. Future research questions include a study of system stability under dual control and optimal (dynamic) quantization of control actions, and the analysis of the framework properties from control and learning perspectives.

ACKNOWLEDGEMENT

The authors would like to thank Michael Cantoni and Girish Nair for their insightful comments and suggestions.

APPENDIX

Definition A.1 (Entropy of a discrete random variable)**.** *The entropy of a discrete random variable X is*

$$H(X) := -\sum_x p(x)\log(p(x)),$$

where $p(x)$ (or $p_X(x)$) denotes the probability mass function (pmf) of X [26].

Definition A.2 (Entropy of a continuous random variable)**.** *The differential entropy [26] of a continuous random variable X with a probability density function (pdf) $f(x)$ is*

$$h(X) := -\int_x f(x)\log(f(x))dx.$$

REFERENCES

[1] T. Alpcan, "A framework for optimization under limited information," *Journal of Global Optimization*, vol. 55, no. 3, pp. 681–706, March 2013.

[2] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.

[3] B. Wittenmark, "Adaptive Dual Control Methods: An Overview," in *Proc. of 5th IFAC symposium on Adaptive Systems in Control and Signal Processing*, Budapest, Hungary, June 1995, pp. 67–72.

[4] A. A. Feldbaum, "Dual control theory. I-IV," Automation and Remote Control, vol. 21, pp. 874-880, 1960, and vol. 22, pp. 1033-1039, 1-12, 109-11, 1961.

[5] B. Wittenmark, "Adaptive dual control," in *Control Systems, Robotics and Automation, Encyclopedia of Life Support Systems (EOLSS), Developed under the auspices of the UNESCO*. Oxford, UK: Eolss Publishers, Jan. 2002.

[6] J. G. Pierce, "A New Look at the Relation Between Information Theory and Search Theory," Office of Naval Research, Arlington, VA, USA, Tech. Rep., June 1978. [Online]. Available: http://handle.dtic.mil/100.2/ADA063845

[7] E. T. Jaynes, "Entropy and Search-Theory," in *Maximum-Entropy and Bayesian Methods in Inverse Problems*, ser. Fundamental Theories of Physics, vol 14, C. R. Smith and J. W. T. Grandy, Eds. Netherlands: Springer, 1985, p. 443. [Online]. Available: http://bayes.wustl.edu/etj/articles/search.pdf

[8] Q. Zhu and J. Oommen, "On the optimal search problem: the case when the target distribution is unknown," in *Proc. of XVII Intl. Conf. of Chilean Computer Science Society*, Valparaiso, Chile, November 1997, pp. 268–277.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[10] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[11] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge University Press, 2003. [Online]. Available: http://www.inference.phy.cam.ac.uk/mackay/itila

[12] M. Tipping, "Bayesian Inference: An Introduction to Principles and Practice in Machine Learning," in *Advanced Lectures on Machine Learning*, ser. Lecture Notes in Computer Science, O. Bousquet, U. von Luxburg, and G. Rätsch, Eds. Berlin, Germany: Springer Berlin Heidelberg, 2004, vol. 3176, pp. 41–62. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-28650-9_3

[13] B. Settles, "Active Learning Literature Survey," University of Wisconsin–Madison, Madison, Wisconsin, USA, Computer Sciences Technical Report 1648, January 2010.

[14] D. J. C. MacKay, "Information-Based Objective Functions for Active Data Selection," *Neural Computation*, vol. 4, no. 4, pp. 590–604, 1992. [Online]. Available: http://www.mitpressjournals.org/doi/abs/10.1162/neco.1992.4.4.590

[15] S. Seo, M. Wallat, T. Graepel, and K. Obermayer, "Gaussian Process Regression: Active Data Selection and Test Point Rejection," in *Proc. of IEEE-INNS-ENNS Intl. Joint Conf. on Neural Networks IJCNN*, vol. 3, Como, Italy, July 2000, pp. 241–246.

[16] C. Freeman and Y. Tan, "Iterative Learning Control With Mixed Constraints for Point-to-Point Tracking," *IEEE Trans. on Control Systems Technology*, vol. 21, no. 3, pp. 604–616, May 2013.

[17] D. Nesic, Y. Tan, W. Moase, and C. Manzie, "A Unifying Approach to Extremum Seeking: Adaptive Schemes based on Estimation of Derivatives," in *Proc. of 49th IEEE Conf. on Decision and Control (CDC)*, Atlanta, Georgia, USA, December 2010, pp. 4625–4630.

[18] D. Nešić, T. Nguyen, Y. Tan, and C. Manzie, "A Non-gradient Approach to Global Extremum Seeking: An Adaptation of the Shubert Algorithm," *Automatica*, vol. 49, no. 3, pp. 809–815, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0005109812006036

[19] J. B. Rawlings, "Tutorial Overview of Model Predictive Control," *IEEE Control Systems*, vol. 20, no. 3, pp. 38–52, June 2000.

[20] S. Kouro, P. Cortes, R. Vargas, U. Ammann, and J. Rodriguez, "Model Predictive Control: A Simple and Powerful Method to Control Power Converters," *IEEE Trans. Industrial Electronics*, vol. 56, no. 6, pp. 1826–1838, June 2009.

[21] R. R. Negenborn, Z. Lukszo, and H. Hellendoorn, Eds., *Intelligent Infrastructures*, ser. Intelligent Systems, Control and Automation: Science and Engineering, Vol. 42. Netherlands: Springer Netherlands, 2010.

[22] L. Zhang, M. Pan, and S. Quan, "Model predictive control of water management in PEMFC," *Journal of Power Sources*, vol. 180, no. 1, pp. 322–329, 2008.

[23] M. Kearney, M. Cantoni, and P. Dower, "Model predictive control for systems with scheduled load and its application to automated irrigation channels," in *Proc. of IEEE Intl Conf on Networking, Sensing and Control (ICNSC)*, Delft, Netherlands, April 2011, pp. 186–191.

[24] G. Marafioti, "Enhanced Model Predictive Control: Dual Control Approach and State Estimation Issues," Ph.D. dissertation, Norwegian University of Science and Technology, 2010.

[25] J. Rathouský and V. Havlena, "MPC-based approximate dual controller by information matrix maximization," *International Journal of Adaptive Control and Signal Processing*, vol. 27, no. 11, pp. 974–999, 2013.

[26] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 1991.

[27] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, July 1948. [Online]. Available: http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf

[28] L. Campbell, "Exponential entropy as a measure of extent of a distribution," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 5, no. 3, pp. 217–225. [Online]. Available: http://dx.doi.org/10.1007/BF00533058

[29] M. H. M. Costa and T. M. Cover, "On the Similarity of the Entropy Power Inequality and the Brunn Minkowski Inequality," Stanford University, California, USA, Tech. Rep. 48, September 1993.

[30] J. D. Little, "The use of storage water in a hydroelectric system," *Operations Research*, vol. 3, no. 2, pp. 187–197, 1955.

[31] B. G. Kingsman, "Purchasing raw materials with uncertain fluctuating prices," *European Journal of Operational Research*, vol. 25, no. 3, pp. 358–372, 1986.
[32] D. J. White and J. M. Norman, "Control of cash reserves," *Journal of the Operational Research Society*, vol. 16, pp. 309–328, 1965.
[33] D. J. White, "A survey of applications of Markov decision processes," *Journal of the Operational Research Society*, vol. 44, no. 11, pp. 1073–1096, November 1993. [Online]. Available: http://www.jstor.org/stable/2583870
[34] W. B. Powell, *Introduction to Approximate Dynamic Programming*, 2nd ed. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., September 2011.
[35] M. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, ser. Wiley Series in Probability and Statistics. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2009, vol. 414.
[36] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1554–1563, 1966.
[37] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.
[38] R. K. Hankin, "A Generalization of the Dirichlet Distribution," *Journal of Statistical Software*, vol. 33, no. 11, pp. 1–18, 2010.
[39] L. F. Bertuccelli, "Robust decision-making with model uncertainty in aerospace systems," Ph.D. dissertation, Massachusetts Institute of Technology, 2008.
[40] T. Minka, "Estimating a Dirichlet distribution," 2012. [Online]. Available: http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirichlet.pdf
[41] L. F. Bertuccelli and J. P. How, "Estimation of non-stationary Markov chain transition models," in *Proc. of 47th IEEE Conference on Decision and Control (CDC)* , Cancun, Mexico, December 2008, pp. 55–60.
[42] D. J. C. MacKay, "Introduction to Gaussian Processes," in *Neural Networks and Machine Learning*, ser. NATO ASI Series, C. M. Bishop, Ed. Berlin, Heidelberg, Germany: Springer Verlag, 1998, pp. 133–166.
[43] N. Ahmed and D. Gokhale, "Entropy expressions and their estimators for multivariate distributions," *IEEE Trans. on Information Theory*, vol. 35, no. 3, pp. 688–692, May 1989.
[44] R. Rutenbar, "Simulated annealing algorithms: an overview," *IEEE Circuits and Devices Magazine*, vol. 5, no. 1, pp. 19–26, January 1989.
[45] I. Shames. (2013) Simulations scripts of "An Information-based Learning Framework for Dual Control". [Online]. Available: https://dl.dropboxusercontent.com/u/4527019/Simulations/Dual/Dual.zip
[46] D. Wang and J. Huang, "A Neural Network Based Method for Solving Discrete-Time Nonlinear Output Regulation Problem in Sampled-Data Systems," in *Advances in Neural Networks - ISNN 2004*, ser. Lecture Notes in Computer Science, F. Yin, J. Wang, and C. Guo, Eds. Springer Berlin / Heidelberg, 2004, vol. 3174, pp. 97–97, 10.1007/978-3-540-28648-6_9.
[47] ——, "A Neural Network-based Approximation Method for Discrete-time Nonlinear Servomechanism Problem," *IEEE Trans. on Neural Networks*, vol. 12, no. 3, pp. 591–597, May 2001.
[48] T. Alpcan, "A framework for optimization under limited information," in *5th International ICST Conference on Performance Evaluation Methodologies and Tools*, ser. VALUETOOLS '11. ICST, Brussels, Belgium, Belgium: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2011, pp. 234–243.
[49] ——, "A Risk-Based Approach to Optimisation under Limited Information," in *Proc. of the 20th Intl. Symp. on Mathematical Theory of Networks and Systems (MTNS)*, Melbourne, Australia, July 2012.

Tansu Alpcan received his B.S. degree in electrical engineering from Bogazici University, Istanbul, Turkey in 1998. He received his M.S. and Ph.D. degrees in electrical and computer engineering from University of Illinois at Urbana-Champaign in 2001 and 2006, respectively. His research involves applications of distributed decision making, game theory, optimization, and control to various security and resource allocation problems in networked and energy systems. He is recipient of multiple research and best paper awards from UIUC and IEEE. He has played a role in organization of several workshops and conferences such as IEEE Infocom, ICC, GameComm, and GameSec as TPC member, associate editor, co-chair, chair, and steering board member. He is the (co-)author of more than 100 journal and conference articles, two edited volumes, as well as the book "Network Security: A Decision and Game Theoretic Approach" published by the Cambridge University Press in 2011. He has worked as a senior research scientist in Deutsche Telekom Laboratories, Berlin, Germany, between 2006-2009, and as Assistant Professor in Technical University of Berlin from 2009 until 2011. He has been with the Department of Electrical and Electronic Engineering at the University of Melbourne as Senior Lecturer since October 2011 and as Associate Professor and Reader since January 2015.

Iman Shames Iman Shames is a Senior Lecturer and McKenzie fellow at the Department of Electrical and Electronic Engineering, the University of Melbourne. Previously, he was an ACCESS Postdoctoral Researcher at the ACCESS Linnaeus Centre, the KTH Royal Institute of Technology, Stockholm, Sweden. He received his B.Sc. degree in Electrical Engineering from Shiraz University, Iran in 2006, and the Ph.D. degree in engineering and computer science from the Australian National University, Canberra, Australia in 2011. His current research interests include, but are not limited to, optimisation theory, mathematical systems theory, sensor networks, and secure cyber-physical systems.