Predictive Inverse Optimal Control for Linear-Quadratic-Gaussian Systems

Xiangli Chen
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607
xchen40@uic.edu

Brian D. Ziebart
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607
bziebart@uic.edu

Abstract

Predictive inverse optimal control is a powerful approach for estimating the control policy of an agent from observed control demonstrations. Its usefulness has been established in a number of large-scale sequential decision settings characterized by complete state observability. However, many real decisions are made in situations where the state is not fully known to the agent making decisions. Though extensions of predictive inverse optimal control to partially observable Markov decision processes have been developed, their applicability has been limited by the complexities of inference in those representations. In this work, we extend predictive inverse optimal control to the linear-quadratic-Gaussian control setting. We establish close connections between optimal control laws for this setting and the probabilistic predictions under our approach. We demonstrate the effectiveness and benefit in estimating control policies that are influenced by partial observability on both synthetic and real datasets.

1 Introduction

Predicting sequences of behavior is an important task for many artificial intelligence applications. It is of key importance for human-robot interaction and human-computer interaction systems. For example, robots that more efficiently and safely navigate around people and user interfaces that can autonomously adapt to improve a user's task efficiency [3] each requires accurate behavior predictions. Perfect accuracy is an unrealistic objective for this task given the large number of behavior sequences that are possible. Instead, statistical methods are needed to characterize the inherent uncertainties of this prediction task and guide appropriate decision making in artificial intelligence applications.

Two main approaches for constructing predictive models for this behavior prediction task are: (1) direct policy estimation [22] learns a mapping from contextual situations to actions; and (2) inverse optimal control (IOC) methods (also known as inverse reinforcement learning and inverse planning) view behavior under a sequential decision process and estimate a reward/cost function that rationalizes demonstrated behavior [23, 19, 30, 2]. Often, a learned reward function from the latter approach generalizes well to other portions of the decision process's state space and even to other decision processes in which the same reward function is applicable. Policy estimation is not nearly as adaptive to contextual changes in the decision/control setting. However, inverse optimal control requires planning, decision, and control problems to be repeatedly solved[1]. This can be computationally expensive. As a result, inverse optimal control methods that have been beneficially employed to estimate decision policies for behavior prediction tasks have been restricted to settings with low dimensional state-action spaces and/or full state observability [31, 11, 29, 16] to make the optimal control problem tractable.

Unfortunately, the assumptions of a low-dimensional state-control space and full state observability often

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38. Copyright 2015 by the authors.

[1] Inverse optimal control methods that purportedly address the IOC problem by estimating value or cost-to-go functions without solving optimal control problems [9] more closely resemble (the dual optimization problem) of direct policy estimation methods.

do not match reality for many important prediction tasks. Often a human actor has only a partial knowledge of the "state of the world" and takes actions that are delayed responses to noisy observations of the actual world state. For example, a person may walk through an environment with occlusions and therefore have uncertainty about the locations of obstacles. Similarly, user interfaces may change in ways that users do not anticipate leading to observed behavior sequences affected by human response times. Extensions of inverse optimal control techniques address either high-dimensionality or partial observability, but not both. IOC methods for high-dimensional data assume a linear-quadratic control setting [29, 15], for which optimal control is tractable even for very high dimensional state-control spaces. IOC methods for partially-observable decision process representations [28, 7] have been limited to partially-observable Markov decision processes with small state-action spaces.

In this paper, we extend IOC methods to settings with both high-dimensional state-control spaces and partial observability. We specifically investigate the discrete-time linear-quadratic-Gaussian (LQG) control setting. This is a special sub-class of partially-observable decision processes for which optimal control is efficient even for large state and control dimensions. We formulate the inverse LQG problem from robust estimation first principles to obtain a predictive distribution over state-control sequences. Like the optimal control solution, which combines a Kalman filter [12] with a linear quadratic regulator [14], our approach finds a similar separation between state estimation and control policy estimation to provide probabilistic predictive distributions. We demonstrate the benefits of incorporating partial observability for predictive inverse optimal control in a synthetic control prediction task and for modeling mouse cursor pointing motions.

2 Background & Related Work

2.1 Linear quadratic Gaussian control

Linear-quadratic-Gaussian (LQG) control problems seek the optimal control policy of partially observed linear systems. Apart from its initial value, which is Gaussian distributed (1), the unobserved state of the system, denoted as value $\vec{x}_t$ (or random variable $\vec{X}_t$) at time t, evolves as a noisy linear function of the previous state $\vec{x}_{t-1}$ and control $\vec{u}_{t-1}$ (2). The state itself is not directly observed by the controller; instead, observation variables $\vec{z}_t$ (or as random variables, $\vec{Z}_t$) that are noisy linear functions of the state are observed (3). The state dynamic and observation noise are each conditional Gaussian distributions with a mean defined by linear relationships of the $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ matrices and covariance matrices Σ_{d_1}, Σ_d and Σ_o characterizing the noise:

$$\vec{\mathbf{X}}_1 \sim N(\vec{\mu}, \Sigma_{d_1}); \tag{1}$$

$$\vec{\mathbf{X}}_{t+1}|\vec{x}_t, \vec{u}_t \sim N(\mathbf{A}\vec{x}_t + \mathbf{B}\vec{u}_t, \Sigma_d); \tag{2}$$

$$\vec{\mathbf{Z}}_t|\vec{x}_t \sim N(\mathbf{C}\vec{x}_t, \Sigma_o). \tag{3}$$

The independence properties of the LQG control setting are illustrated by the Bayesian network in Figure 1.

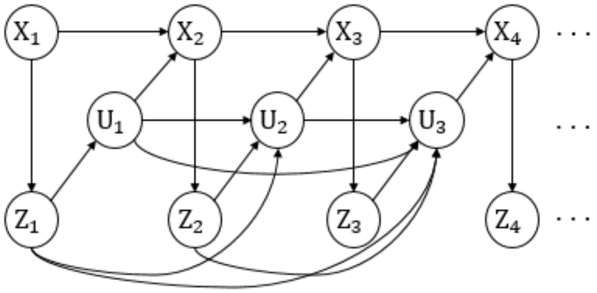

Figure 1: A probabilistic graphical model for the partially-observable control setting.

Given the state and observation dynamics, the LQG optimal control problem is to obtain the control policy $\pi : \vec{\boldsymbol{\mathcal{U}}}_{1:t-1} \times \vec{\boldsymbol{\mathcal{Z}}}_{1:t} \rightarrow \vec{\mathcal{U}}_t$, that minimizes an expected cost that is quadratically defined in terms of a cost matrix $\mathbf{M}$:

$$\mathbb{E}_{f(\vec{\mathbf{Z}}_{1:T+1}, \vec{\mathbf{X}}_{1:T+1}, \vec{\mathbf{U}}_{1:T})} \left[\sum_{t=1}^{T+1} \vec{X}_t^T \mathbf{M} \vec{X}_t \right]. \tag{4}$$

The optimal control law is obtained by separating the problem into a state estimation task and a linear-quadratic-regulation optimal control problem for the estimated state. State estimation is accomplished using a Kalman filter [12]. Due to the linear characteristics of the problem, only the mean of the state estimate is needed. For the estimate of the state's mean conditioned on previous and current observations and previous controls, $\hat{x}_t(+)$, the optimal control policy is recursively defined [25] as:

$$\vec{u}_t = -\mathbf{L}_t \hat{x}_t(+),$$
$$\hat{x}_{t+1} = \mathbf{A}\hat{x}_t + \mathbf{B}u_t + \mathbf{K}_t(z_t - \mathbf{C}(\mathbf{A}\hat{x}_t + \mathbf{B}u_t)), \hat{x}_1 = \mathbb{E}[x_1],$$

where $\mathbf{K_t}$ is the Kalman gain:

$$\mathbf{K}_t = \mathbf{H}_t \mathbf{C}^T (\mathbf{C}\mathbf{H}_t\mathbf{C}^T + \Sigma_o)^{-1},$$

and $\mathbf{H}_t$ is determined by the following matrix Riccati difference equation that runs forward in time,

$$\mathbf{H}_{t+1} = \mathbf{A}(\mathbf{H}_t - \mathbf{H}_t\mathbf{C}^T(\mathbf{C}\mathbf{H}_t\mathbf{C}^T + \Sigma_o)^{-1}\mathbf{C}\mathbf{H}_t)\mathbf{A}^T + \Sigma_x,$$
$$\mathbf{H}_1 = \mathbb{E}[X_1 X_1^T] = \Sigma_x + \mu\mu^T$$

The feedback gain matrix:

$$\mathbf{L}_t = (\mathbf{B}^T\mathbf{F}_{t+1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{F}_{t+1}\mathbf{A},$$

where $\mathbf{F}_t$ is determined by the following matrix Riccati difference equation that runs backward in time,

$$\mathbf{F}_t = \begin{cases} \mathbf{M} + \mathbf{A}^T\mathbf{F}_{t+1}\mathbf{A} & t \le T \\ \quad -\mathbf{A}^T\mathbf{F}_{t+1}\mathbf{B}(\mathbf{B}^T\mathbf{F}_{t+1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{F}_{t+1}\mathbf{A} & \\ \mathbf{M} & t = T+1. \end{cases} \tag{5}$$

Linear-quadratic regulator (LQR) can be viewed as the full-observability special case of LQG with $\mathbf{C} = \mathbf{I}$ and $\Sigma_o = \mathbf{0}$ where $\mathbf{I}$ is the identity matrix so that the observation variable $\vec{z}_t$ is equivalent to the unobserved state $\vec{x}_t$.

2.2 Inverse optimal control

In contrast with the optimal control problem of obtaining a policy that minimizes some cumulative expected cost, inverse optimal control [20, 1] takes (samples from) a policy and tries to obtain a cost function for which observed behavior is optimal, ideally. Early approaches often assumed a linear functional form [5, 24, 1] for the cost function in terms of *features* $\mathbf{f}$, $\text{cost}(x_t) = \theta^T\mathbf{f}(x_t)$, and optimality for some choice of weights θ. In practice, observed behavior is not consistently optimal for any linear cost function [1] for this family of functions, and the optimality assumption breaks down. Even though the linear weights of the function are unknown and behavior can be arbitrarily sub-optimal, any policy that has the same expected feature statistics, $\mathbb{E}[\sum_t \mathbf{f}(x_t)]$, as the demonstrated feature statistic expectation is guaranteed to have the same expected cost [1]. Mixtures of optimal policies can instead be employed to guarantee the same expected costs as (sub-optimal) demonstrated behavior [1].

We distinguish IOC approaches that match the expected cost of demonstrated behavior with those that attempt to provide predictions for behavior. The principle of maximum entropy has previously been employed for this purpose. Maximum entropy IOC [30] provides a stochastic control policy that robustly minimizes the predictive logloss for policies that, in expectation, generate certain expected features. In contrast, mixtures of optimal policies [1] can produce infinite logloss when they provide no support for demonstrated policies. We build upon the maximum entropy IOC approach in this work.

2.3 Directed information theory

We view the LQG setting using concepts and measures from directed information theory [17, 18, 13, 26, 21]. The joint distribution of states, observations, and controls is factored into two causally conditioned probability distributions [13],

$$f(\vec{\mathbf{x}}_{1:T}, \vec{\mathbf{z}}_{1:T}, \vec{\mathbf{u}}_{1:T}) = f(\vec{\mathbf{u}}_{1:T}||\vec{\mathbf{x}}_{1:T}, \vec{\mathbf{z}}_{1:T})\; f(\vec{\mathbf{x}}_{1:T}, \vec{\mathbf{z}}_{1:T}||\vec{\mathbf{u}}_{1:T-1}), \tag{6}$$

$$\text{where: } f(\vec{\mathbf{u}}_{1:T}||\vec{\mathbf{x}}_{1:T}, \vec{\mathbf{z}}_{1:T}) \triangleq \prod_{t=1}^{T} f(\vec{u}_t|\vec{\mathbf{u}}_{1:t-1}, \vec{\mathbf{x}}_{1:t}, \vec{\mathbf{z}}_{1:t}) \tag{7}$$

$$\text{and } f(\vec{\mathbf{x}}_{1:T}, \vec{\mathbf{z}}_{1:T}||\vec{\mathbf{u}}_{1:T-1}) \triangleq \prod_{t=1}^{T} f(\vec{x}_t, \vec{z}_t|\vec{\mathbf{x}}_{1:t-1}, \vec{\mathbf{u}}_{1:t-1}, \vec{\mathbf{z}}_{1:t-1}). \tag{8}$$

The causal entropy [13] of the control policy,

$$H(\vec{\mathbf{U}}_{1:T}||\vec{\mathbf{X}}_{1:T}, \vec{\mathbf{Z}}_{1:T}) \triangleq \mathbb{E}\left[-\log_2 f(\vec{\mathbf{U}}_{1:T}||\vec{\mathbf{X}}_{1:T}, \vec{\mathbf{Z}}_{1:T})\right] = \mathbb{E}\left[\sum_{t=1}^{T} H(\vec{U}_t|\vec{\mathbf{X}}_{1:t}, \vec{\mathbf{Z}}_{1:t}, \vec{\mathbf{U}}_{1:t-1})\right], \tag{9}$$

is a measure of the uncertainty of the causally conditioned probability distribution (8). It can be interpreted as the amount of information or "surprise" (in bits when using base-2) present in expectation for a control sequence $\vec{\mathbf{u}}_{1:T}$ sampled from the joint state, observation, control distribution (6), given only previous observation and control variables.

Due to the specific independence properties of partial observability in the LQG setting (shown in Figure 1), the control policy reduces to:

$$f(\vec{\mathbf{u}}_{1:T}||\vec{\mathbf{z}}_{1:T}) \triangleq \prod_{t=1}^{T} f(\vec{u}_t|\vec{\mathbf{u}}_{1:t-1}, \vec{\mathbf{z}}_{1:t}), \tag{10}$$

and the causal entropy of the control policy reduces to $H(\vec{\mathbf{U}}_{1:T}||\vec{\mathbf{Z}}_{1:T})$. However, as we shall see, representing the control distribution in its more general form and constraining it to possess the required independence properties is crucial for our approach.

3 Inverse Linear Quadratic Gaussian Control

We employ a robust estimation approach for learning the control policy in a way that generalizes to different settings (§3.1). We show that this approach can be posed as a convex optimization problem (§3.2) leading to a maximum causal entropy problem (§3.3). The dual solution decomposes into a state estimation component and a (softened) optimal control component, enabling efficient inference (§3.4).

3.1 Robust policy estimation

We consider a set of policies denoted by Ξ that are similar to observed sequences of states, observations, and controls (defined precisely in §3.3). We follow the robust estimation formulation [27, 10] of maximum entropy inverse optimal control [28] to select the single policy with the best worst-case predictive guarantees from this set. This can be viewed as a two-player game in which the policy estimate, $\hat{f}$, that minimizes loss is first chosen, and then an evaluation policy, $f \in \Xi$, is adversarially chosen that maximizes the loss subject to matching known/observed properties of the actual policy:

$$\min_{\{\hat{f}(\vec{\mathbf{u}}_{1:T}||\vec{\mathbf{z}}_{1:T},\vec{\mathbf{x}}_{1:T})\}} \max_{\substack{\{f(\vec{\mathbf{u}}_{1:T}||\vec{\mathbf{z}}_{1:T},\vec{\mathbf{x}}_{1:T})\} \\ \in\,\Xi}} \text{Loss}(\hat{f}, f) \quad (11)$$

$$\geq \max_{\substack{\{f(\vec{\mathbf{u}}_{1:T}||\vec{\mathbf{z}}_{1:T},\vec{\mathbf{x}}_{1:T})\} \\ \in\,\Xi}} \min_{\{\hat{f}(\vec{\mathbf{u}}_{1:T}||\vec{\mathbf{z}}_{1:T},\vec{\mathbf{x}}_{1:T})\}} \text{Loss}(\hat{f}, f) \quad (12)$$

$$= \max_{\{f(\vec{\mathbf{u}}_{1:T}||\vec{\mathbf{z}}_{1:T},\vec{\mathbf{x}}_{1:T})\}\in\Xi} \text{Loss}(f, f). \quad (13)$$

In general, weak Lagrangian duality holds and the dual optimization problem (12) provides a lower-bound on the primal optimization problem (11). The **causal log-loss**,

$$\text{Loss}(\hat{f}, f) = E_f[-\log \hat{f}(\vec{\mathbf{U}}_{1:T}||\vec{\mathbf{Z}}_{1:T}\vec{\mathbf{X}}_{1:T})], \quad (14)$$

measures the amount of "surprise" (in bits when $\log_2$ is used) when control sequences sampled from f are observed while control sequences from $\hat{f}$ are expected.

When it is employed as the loss function, the dual optimization problem reduces to maximizing the causal entropy (13): $\text{Loss}(f, f) = H(\vec{\mathbf{U}}_{1:T}||\vec{\mathbf{Z}}_{1:T}, \vec{\mathbf{X}}_{1:T})$.

3.2 A convex definition of the LQG policy set

We seek to strengthen our analysis of the dual solution so that primal-dual equality holds (12). This strong duality requires the set of policies (10) to be convex [6], which is not obviously the case. We introduce the **partial observability causal simplex** (Definition 1), which extends the causal simplex [28] to the partial-observability setting. It is defined by affine constraints that ensure that members of the set factor according to (10). This is accomplished by preventing unobserved variables ($\vec{\mathbf{x}}_{1:T}$) and not-yet-revealed variables ($\vec{\mathbf{z}}_{t+1:T}$) from influencing a control variable's distribution (y_t).

Definition 1. *The* **partial observability causal simplex** *for* $f(\vec{\mathbf{u}}_{1:T}||\vec{\mathbf{z}}_{1:T}, \vec{\mathbf{x}}_{1:T})$ *denoted by* Δ*, is defined by the following set of constraints:*

$$\forall \vec{\mathbf{u}}_{1:T} \in \vec{\mathcal{U}}_{1:T}, \vec{\mathbf{x}}_{1:T}, \mathbf{x}'_{1:T} \in \vec{\mathcal{X}}_{1:T}, \vec{\mathbf{z}}_{1:T}, \mathbf{z}'_{1:T} \in \vec{\mathcal{Z}}_{1:T},$$

$$f(\vec{\mathbf{u}}_{1:T}||\vec{\mathbf{z}}_{1:T}, \vec{\mathbf{x}}_{1:T}) \geq 0, \quad (15)$$

$$\int_{\vec{\mathbf{u}}'_{1:T}\in\vec{\mathcal{U}}_{1:T}} f(\vec{\mathbf{u}}'_{1:T}||\vec{\mathbf{z}}_{1:T}, \vec{\mathbf{x}}_{1:T})\, d\vec{\mathbf{u}}'_{1:T} = 1, \quad (16)$$

$$f(\vec{\mathbf{u}}_{1:T}||\vec{\mathbf{z}}_{1:T}, \vec{\mathbf{x}}_{1:T}) = f(\vec{\mathbf{u}}_{1:T}||\vec{\mathbf{z}}_{1:T}, \vec{\mathbf{x}}'_{1:T}). \quad (17)$$

$$\forall \tau \in \{0, \ldots, T\} \textit{ such that } \vec{\mathbf{z}}_{1:\tau} = \vec{\mathbf{z}}'_{1:\tau}, \vec{\mathbf{x}}_{1:\tau} = \vec{\mathbf{x}}'_{1:\tau},$$

$$\int_{\vec{\mathbf{u}}_{\tau+1:T}\in\vec{\mathcal{U}}_{\tau+1:T}} f(\vec{\mathbf{u}}_{1:T}||\vec{\mathbf{z}}_{1:T}, \vec{\mathbf{x}}_{1:T})\, d\vec{\mathbf{u}}_{\tau+1:T} \quad (18)$$

$$= \int_{\vec{\mathbf{u}}_{\tau+1:T}\in\vec{\mathcal{U}}_{\tau+1:T}} f(\vec{\mathbf{u}}_{1:T}||\vec{\mathbf{z}}'_{1:T}, \vec{\mathbf{x}}'_{1:T})\, d\vec{\mathbf{u}}_{\tau+1:T}.$$

The non-negativity constraints (15) and normalization constraints (16) ensure a valid probability distribution. The next set of constraints (17) enforces partial observability—the controls do not depend on the hidden state. The final set of constraints (Equation 18) ensures that only previous $\vec{x}$ and $\vec{z}$ variables influence controls $\vec{u}$ (causal conditioning). Because all of the equalities and inequalities are affine, the partial observability causal simplex is a convex set.

3.3 Maximum causal entropy estimation

Redefining the domain of the estimated policy $\hat{f}(\mathbf{u}_{1:T}||\mathbf{z}_{1:T})$ using the partial observability causal simplex (Definition 1) enables strong duality. The dual of the robust policy estimation formulation (Section 3.1), reduces to maximizing the causal entropy (9) as a selection measure from the set of policies (Δ) matching quadratic state expectation constraints (Definition 2).

Definition 2. *The* **maximum causal entropy inverse LQG policy** *is obtained from:*

$$\operatorname*{argmax}_{\{f(\vec{\mathbf{u}}_{1:T}||\vec{\mathbf{z}}_{1:T},\vec{\mathbf{x}}_{1:T})\}\in\Delta} H(\vec{\mathbf{U}}_{1:T}||\vec{\mathbf{Z}}_{1:T}, \vec{\mathbf{X}}_{1:T}) \quad (19)$$

$$\textit{such that: } \mathbb{E}\left[\sum_{t=1}^{T+1} \vec{\mathbf{X}}_t \vec{\mathbf{X}}_t^T\right] = \tilde{\mathbb{E}}\left[\sum_{t=1}^{T+1} \vec{\mathbf{X}}_t \vec{\mathbf{X}}_t^T\right], \quad (20)$$

where Δ *is the partial observability causal simplex of Definition 1,* $\mathbb{E}[\cdot]$ *is the expectation under the estimated policy, and* $\tilde{\mathbb{E}}[\cdot]$ *is the empirical expectation from observed behavior sequence data.*

This choice of constraints is motivated by inverse optimal control (Section 2.2). They ensure that the stochastic control policy matches the performance of observed behavior on unknown state-based quadratic cost functions[2] (Corollary 1).

[2] We employ state-based functions for notational simplicity. Control-based functions could also be explicitly added with an additional constraint or implicitly by incorporating an "action memory" into the state vector.

Corollary 1 ([1]). *For any unknown quadratic cost function, parameterized by matrix* $\mathbf{M}$*, matching expected* feature counts *guarantees equivalent performance on the unknown cost function:*

$$\forall \mathbf{M} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|},$$
$$\mathbb{E}\left[\sum_{t=1}^{T+1} \vec{\mathbf{X}}_t \vec{\mathbf{X}}_t^T\right] = \tilde{\mathbb{E}}\left[\sum_{t=1}^{T+1} \vec{\mathbf{X}}_t \vec{\mathbf{X}}_t^T\right]$$
$$\implies \mathbb{E}\left[\sum_{t=1}^{T+1} \vec{\mathbf{X}}_t^T \mathbf{M} \vec{\mathbf{X}}_t\right] = \tilde{\mathbb{E}}\left[\sum_{t=1}^{T+1} \vec{\mathbf{X}}_t^T \mathbf{M} \vec{\mathbf{X}}_t\right],$$

Many different mixture distributions over deterministic policies can satisfy this constraint [1]. Thus, the causal entropy (19) can be viewed as a tie-breaking criterion that resolves the ill-posedness of inverse optimal and provides strong robust prediction guarantees.

3.4 Predictive inverse LQG distribution

The Lagrangian dual provides a value-equivalent solution to the primal constrained optimization problem (19)[3], while leading to a more compact representations of the policy.

Theorem 1. *The solution to the partially-observable maximum causal entropy problem (Definition 2) takes the following recursive form where M is the Lagrangian multiplier matrix:*

$$f(\vec{u}_t|\vec{\mathbf{u}}_{1:t-1}, \vec{\mathbf{z}}_{1:t}) = e^{Q(\vec{\mathbf{u}}_{1:t}, \vec{\mathbf{z}}_{1:t}) - V(\vec{\mathbf{u}}_{1:t-1}, \vec{\mathbf{z}}_{1:t})} \tag{21}$$

where:

$$Q(\vec{\mathbf{u}}_{1:t}, \vec{\mathbf{z}}_{1:t}) = \begin{cases} \mathbb{E}[\vec{\mathbf{X}}_{T+1}^T \mathbf{M} \vec{\mathbf{X}}_{T+1} | \vec{\mathbf{u}}_{1:T}, \vec{\mathbf{z}}_{1:T}] & t = T; \\ \mathbb{E}[\vec{\mathbf{X}}_{t+1}^T \mathbf{M} \vec{\mathbf{X}}_{t+1} & \\ \quad + V(\vec{\mathbf{U}}_{1:t}, \vec{\mathbf{Z}}_{1:t+1}) | \vec{\mathbf{u}}_{1:t}, \vec{\mathbf{z}}_{1:t}] & t < T \end{cases} \tag{22}$$

$$\begin{aligned} V(\vec{\mathbf{u}}_{1:t-1}, \vec{\mathbf{z}}_{1:t}) &= \operatorname*{softmax}_{\vec{u}_t} Q(\vec{\mathbf{u}}_{1:t}, \vec{\mathbf{z}}_{1:t}) \\ &\triangleq \log \int_{\vec{u}_t} e^{Q(\vec{\mathbf{u}}_{1:t}, \vec{\mathbf{z}}_{1:t})} d\vec{u}_t. \end{aligned} \tag{23}$$

The probability distribution can be interpreted as a softened relaxation of the Bellman optimal policy criterion [4] where the softmax function replaces the max function: $\text{softmax}_x f(x) \triangleq \log \int_x e^{f(x)}$. It serves as a smooth interpolator of the maximum function.

Unfortunately, in the LQG setting, the value functions of Theorem 1 are still unwieldy since they depend on the entire history of actions and observations. As in optimal LQG control [14], a more practical algorithm is obtained by separating state estimation from the policy distribution. Assuming a Gaussian *belief* of the current state $\vec{\mathbf{X}}_t | b_t(\vec{\mathbf{u}}_{1:t-1}, \vec{\mathbf{z}}_{1:t}) \sim N(\vec{\mu}_{b_t}, \Sigma_{b_t})$ that is based on the entire history, the policy can be recursively obtained according to Theorem 2.

Theorem 2. *Given a belief state which summarizes the history* $\vec{\mathbf{u}}_{1:t-1}, \vec{\mathbf{z}}_{1:t}$ *up to time step t (i.e.,* $\vec{\mathbf{X}}_t | b_t \sim N(\vec{\mu}_{b_t}, \Sigma_{b_t})$*), the recurrence values* (22),(23) *are Markovian quadratic functions of the form:*

$$Q(\vec{u}_t, \vec{\mu}_{b_t}) = \begin{bmatrix} \vec{u}_t \\ \vec{\mu}_{b_t} \end{bmatrix}^T \mathbf{W}_t \begin{bmatrix} \vec{u}_t \\ \vec{\mu}_{b_t} \end{bmatrix} \tag{24}$$

$$V(\vec{z}_t, \vec{u}_{t-1}, \vec{\mu}_{b_{t-1}}) = \begin{bmatrix} \vec{z}_t \\ \vec{u}_{t-1} \\ \vec{\mu}_{b_{t-1}} \end{bmatrix}^T \mathbf{D}_t \begin{bmatrix} \vec{z}_t \\ \vec{u}_{t-1} \\ \vec{\mu}_{b_{t-1}} \end{bmatrix} \tag{25}$$

$$\mathbf{W}_t = \begin{cases} \begin{bmatrix} \mathbf{B} & \mathbf{A} \end{bmatrix}^T \mathbf{M} \begin{bmatrix} \mathbf{B} & \mathbf{A} \end{bmatrix} & t = T \\ \begin{bmatrix} \mathbf{B} & \mathbf{A} \end{bmatrix}^T \mathbf{M} \begin{bmatrix} \mathbf{B} & \mathbf{A} \end{bmatrix} & t < T \\ \quad + \mathbf{D}_{t+1(U\mu,Z)} \begin{bmatrix} \mathbf{CB} & \mathbf{CA} \end{bmatrix} & \\ \quad + \begin{bmatrix} \mathbf{CB} & \mathbf{CA} \end{bmatrix}^T \mathbf{D}_{t+1(Z,U\mu)}, & \\ \quad + \begin{bmatrix} \mathbf{CB} & \mathbf{CA} \end{bmatrix}^T \mathbf{D}_{t+1(Z,Z)} \begin{bmatrix} \mathbf{CB} & \mathbf{CA} \end{bmatrix} & \\ \quad + \mathbf{D}_{t+1(U\mu,U\mu)} & \end{cases} \tag{26}$$

$$\mathbf{D}_t = \mathbf{P}_t^T (\mathbf{W}_{t(\mu,\mu)} - \mathbf{W}_{t(U,\mu)}^T \mathbf{W}_{t(U,U)}^{-1} \mathbf{W}_{t(U,\mu)}) \mathbf{P}_t \tag{27}$$

where

$$\mathbf{P}_t = \begin{bmatrix} \mathbf{E}_{t+1} & \mathbf{B} - \mathbf{E}_{t+1}\mathbf{CB} & \mathbf{A} - \mathbf{E}_{t+1}\mathbf{CA} \end{bmatrix}$$

$$\mathbf{E}_{t+1} = (\Sigma_d + \mathbf{A}\Sigma_{b_t}^T \mathbf{A}^T)^T \mathbf{C}^T (\Sigma_o + \mathbf{C}(\Sigma_d + \mathbf{A}\Sigma_{b_t}^T \mathbf{A}^T)^T \mathbf{C}^T)^{-1}$$

The probabilistic control policy for a belief state with mean $\vec{\mu}_{b_t}$ is then:

$$\vec{\mathbf{U}}_t | \vec{\mu}_{b_t} \sim N\left(-\mathbf{W}_{t(U,U)}^{-1} \mathbf{W}_{t(U,\mu)} \vec{\mu}_{b_t}, -\frac{1}{2}\mathbf{W}_{t(U,U)}^{-1}\right) \tag{28}$$

Theorem 3 establishes the connection to optimal control: the mean/mode of the control distribution is the optimal control and, in fact, the (stochastic) maximum causal entropy probabilistic control policy can be directly obtained from the optimal control solution.

Theorem 3. *The terms of the stochastic control policy* (28) *are related to the LQG optimal control laws as:*

$$\mathbf{W}_{t(U,U)} = \mathbf{B}^T \mathbf{F}_{t+1} \mathbf{B}; \qquad \mathbf{W}_{t(U,\mu)} = \mathbf{B}^T \mathbf{F}_{t+1}, \tag{29}$$

where $\mathbf{F}_{t+1}$ *is defined by the optimal control law*(5)*, and the Lagrangian multiplier matrix M in*(26) *is given as the cost matrix in*(5)*.*

Thus, existing methods for solving LQG optimal control problems can be used to recover the stochastic control policy given the cost matrix $\mathbf{M}$.

[3] Strong duality is subject to mild feasibility requirements on feature matching.

3.5 Model Fitting

According to the previously developed theory of maximum causal entropy [28] the gradient of the Lagrangian dual form with respect to the Lagrangian multipliers matrix $\mathbf{M}$ is:

$$\mathbb{E}_f\left[\sum_{t=1}^{T+1}\vec{\mathbf{X}}_t\vec{\mathbf{X}}_t^T\right]-\tilde{\mathbb{E}}_f\left[\sum_{t=1}^{T+1}\vec{\mathbf{X}}_t\vec{\mathbf{X}}_t^T\right] \qquad (30)$$

This comes from the constraint of the convex optimization problem (20).

We can compute the expectation of quadratic state moments over the distribution of state-control-observation trajectories provided by our estimated policy also via conditioning on the mean and variance of the belief state:

$$\begin{aligned}
&\mathbb{E}_f\left[\vec{\mathbf{X}}_t\vec{\mathbf{X}}_t^T\right]\\
&=\mathbb{E}_{f(\vec{\mathbf{u}}_{1:t-1},\vec{\mathbf{z}}_{1:t})}\left[\mathbb{E}_{f(\vec{\mathbf{x}}_t|\vec{\mathbf{u}}_{1:t-1},\vec{\mathbf{z}}_{1:t})}\left[\vec{\mathbf{X}}_t\vec{\mathbf{X}}_t|\vec{\mathbf{U}}_{1:t-1},\vec{\mathbf{Z}}_{1:t}\right]\right]\\
&=\mathbb{E}\left[\Sigma_{b_t}+\vec{\mu}_{b_t}\vec{\mu}_{b_t}^T\right]\\
&=\Sigma_{b_t}+\mathbf{Var}\left[\vec{\mu}_{b_t}\right]+\mathbb{E}\left[\vec{\mu}_{b_t}\right]\mathbb{E}\left[\vec{\mu}_{b_t}\right]^T .
\end{aligned}$$

The mean and variance of the belief state $\vec{\mu}_{b_t},\Sigma_{b_t}$ is recursively computed according to a Kalman filter [12].

4 Experimental Validation

We evaluate the performance of our inverse LQG approach on controlled data (Sec. 4.1) and real mouse cursor movement data (Sec. 4.2) to investigate its benefits in comparison to full observability models of behavior by average empirical causal log-loss over test data. We call empirical casual log-loss as **trajectory log-loss**. Assume we have N trajectories $\{\vec{\mathbf{x}}_{1:T_{n+1}},\vec{\mathbf{z}}_{1:T_n},\vec{\mathbf{u}}_{1:T_n}\}_{n=1}^N$ for test, and f is our probabilistic control policy learned from training data, then the average trajectory log-loss is:

$$\begin{aligned}
&-\frac{1}{N}\sum_{n=1}^{N}\log f(\vec{\mathbf{u}}_{1:T_n}||\vec{\mathbf{z}}_{1:T_n})=\\
&=-\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n}\log f(\vec{\mathbf{u}}_t|\vec{\mathbf{u}}_{1:t-1},\vec{\mathbf{z}}_{1:t})\\
&=-\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n}\log f(\vec{\mathbf{u}}_t|\vec{\mu}_{b_t}).
\end{aligned}$$

4.1 Controlled demonstrations of the benefits of partial observability

In our first set of experiments, we investigate the benefits of incorporating partial-observability into predictive inverse optimal control. This provides some insights into whether it is sufficient to simply ignore partial observability and use inverse optimal control (IOC) models that assume full observability. We vary the state and observation noise of a LQG control problem and measure the average empirical trajectory log-loss compared to treating the problem as a fully-observed linear-quadratic regulator (LQR)control process.

We collect data via an optimal LQG controller [14] applied to a spring-mass system:

$$\vec{\mathbf{X}}_{t+1}=\mathbf{A}\vec{\mathbf{X}}_t+\mathbf{B}\vec{\mathbf{U}}_t+\varepsilon_s \qquad \vec{\mathbf{Z}}_t=\mathbf{C}\vec{\mathbf{X}}_t+\varepsilon_o$$

$$\vec{\mathbf{X}}_1\sim N(\vec{0},\Sigma_{d_1}) \qquad \varepsilon_x\sim N(\vec{0},\Sigma_d) \qquad \varepsilon_o\sim N(\vec{0},\Sigma_o).$$

$$\mathbf{A}=\begin{bmatrix}0&1\\-1&0\end{bmatrix} \qquad \mathbf{B}=\begin{bmatrix}0\\1\end{bmatrix} \qquad \mathbf{C}=\begin{bmatrix}1&0\end{bmatrix}$$

The controller minimizes the following expected quadratic cost function:

$$J=\mathbb{E}\left[\sum_{t=1}^{T+1}\vec{\mathbf{X}}_t^T\mathbf{Q}\vec{\mathbf{X}}_t\right],$$

where we set (using $\mathbf{I}$ as the identity matrix):

$$\mathbf{Q}=\mathbf{I}_{2\times2} \quad \Sigma_{d1}=\Sigma_d=\sigma_d*\mathbf{I}_{2\times2} \quad \Sigma_o=\sigma_o*\mathbf{I}_{1\times1}.$$

From the setting of the observation dynamic $\mathbf{C}$, only the first row of the two row state $\vec{\mathbf{X}}_t$ is observed which provides partial-observability scnario.

For each experiment where we vary the noise of the system, we generate 2000 state-observation-control trajectories with length $T=30$ by applying the optimal LQG controller. We use the first 1000 trajectories as the training data and remaining 1000 trajectories as testing data. To simulate the LQR model via LQG setting, we set $\mathbf{C}=\mathbf{I}_{2\times2}$ and let $\vec{\mathbf{Z}}_t=\vec{\mathbf{X}}_t$. We note that the average trajectory log-loss can be negative, as in these experiments, because it is taken over a continuous distribution.

Figure 2, above, shows that LQG has significantly better performance than LQR when the observation noise σ_o increases. This is because the controller is basing its controls on noisy observations that are increasingly different from the true state. Also, it shows that the log loss decreases as the observation noise increases. This is because as the observation noise increases, the controllers system state estimates are less certain. Paying large costs for controls becomes less worthwhile, and controls from a smaller range (closer to 0) are instead produced. These lower variance controls are easier to predict and have small log loss. As shown in Figure 2, below, when the state noise σ_d increases, it dominates the test performance of both LQG and LQR. This is because as the state noise increases, state estimation becomes increasingly error-prone for both LQG and LQR.

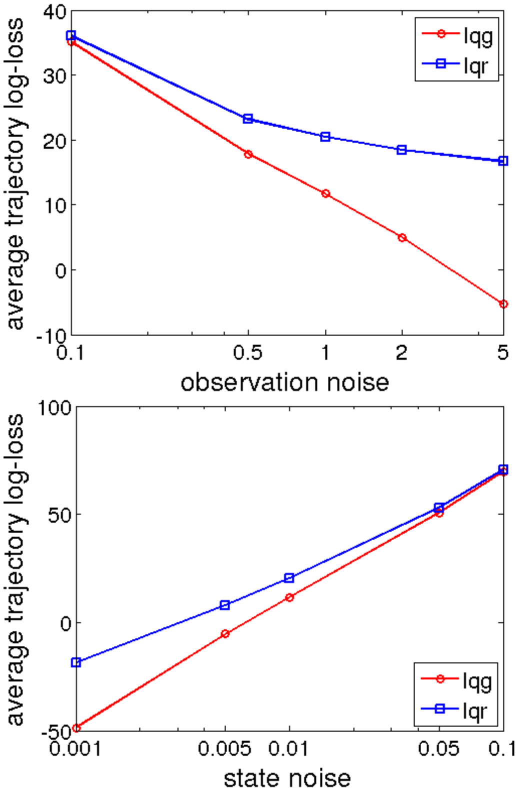

Figure 2: *Above:* Withheld average trajectory log-loss as the observation noise, σ_o, increases (with fixed state transition dynamic noise $\sigma_d = 0.01$). *Below:* Withheld average trajectory log-loss as the state transition dynamics noise, σ_d, increases (with fixed observation noise $\sigma_o = 1.0$).

4.2 Estimating mouse cursor pointing trajectories

Modeling mouse cursor pointing motions is an important machine learning problem for human-computer interaction tasks. A number of interventional techniques have been developed to facilitate pointing target acquisition (e.g., adjusting the control-display ratio dynamics, enlarging targets, etc.) [3]. However, better predictions of intended target are required for these intervention techniques to be successfully employed in the wild [29].

We use data captured from 20 non-motor-impaired computer users performing computer cursor pointing tasks to assess the benefits of the LQG approach versus previous LQR models that have been employed for this task [29]. Users are presented with a sequence of circular clicking targets to select and their mouse cursor data is collected at 100Hz. We specifically investigate whether incorporating a response delay using our LQG framework provides better predictions than the LQR approach, which assumes an instantaneous response to the changes in mouse cursor position. Our assumption is instead that due to the imprecise human abilities for fine-grained control, cursor navigation is essentially an open loop control problem and that incorporating feedback delay will produce better policy estimates. Some evidence of this is demonstrated by the cursor pointing trajectories in Figure 3.

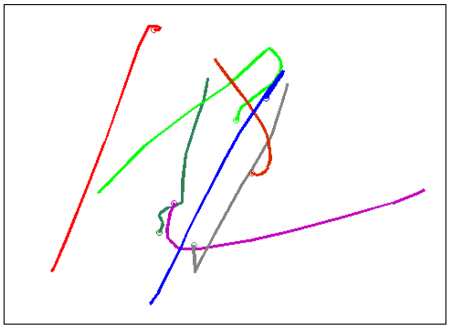

Figure 3: Example mouse cursor trajectories terminating at small circle positions exhibiting characteristics of delayed feedback.

We follow the previous work's control formulation [29]. The instantaneous state

$$\vec{\mathbf{x}}_t \triangleq [x_t \; y_t \; \dot{x}_t \; \dot{y}_t \; \ddot{x}_t \; \ddot{y}_t \;]^\top$$

is represented by the relative position, velocity, and acceleration vectors towards the target and orthogonal to the target at discrete points in time. These dynamics (e.g., velocities and accelerations) are defined according to difference equations,

$$\begin{pmatrix} \dot{x}_t \\ \dot{y}_t \end{pmatrix} = \begin{pmatrix} x_t - x_{t-1} \\ y_t - y_{t-1} \end{pmatrix} \tag{31}$$

$$\begin{pmatrix} \ddot{x}_t \\ \ddot{y}_t \end{pmatrix} = \begin{pmatrix} \dot{x}_t - \dot{x}_{t-1} \\ \dot{y}_t - \dot{y}_{t-1} \end{pmatrix}, \tag{32}$$

and can easily be expressed as a linear dynamics model with the control vector $\vec{u}_t$ representing the change in position. Under this dynamics model, mouse pointing motion data follows a linear relationship (with optional zero-meaned Gaussian noise, ϵ):

$$\vec{\mathbf{x}}_{t+1} = \mathbf{A}\vec{\mathbf{x}}_t + \mathbf{B}\vec{\mathbf{u}}_t + \epsilon.$$

where

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \end{bmatrix} \qquad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Though the cursor positions are located at discrete pixels and the controls are the discrete differences of these pixel locations, using a discrete model for estimating the control policy is not feasible. Specifically, since the dimensionality of the state space is six, any reasonably fine-grained discretization of each dimension (position, velocity, acceleration) will lead to an intractably large discrete decision process that is farther exacerbated by partial observability (as a partially-observed Markov decision process).

We consider a control system with a delayed observation of t_0 time steps. This is formally represented in the LQG model by augmenting the LQG state with the previous t_0 states and having the observation dynamics only reveal the state from t_0 time steps ago. For example, a delay one model has the following dynamics matrices:

$$\mathbf{A}' = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \qquad \mathbf{B}' = \begin{bmatrix} \mathbf{B} \\ \mathbf{0} \end{bmatrix} \qquad \mathbf{C}' = \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix}.$$

We additionally compare our prediction approach against a direct policy estimation method: k^{th}-order Markov models of different orders $k = \{1, 2, 3, 4\}$. For this continuous state-action setting, estimating the Markov model reduces to a linear regression problem of the form:

$$\hat{\vec{s}}_t = [\vec{s}_{t-1}\ \vec{s}_{t-2}\ \ldots\ \vec{s}_{t-k}]\vec{\alpha} + \epsilon, \tag{33}$$

with zero-mean Gaussian noise $\epsilon \sim N(0, \sigma^2)$. The state at each time is the x and y position of the mouse cursor. Regression parameters $\vec{\alpha}$ are estimated by minimizing the sum of squared errors, as is standard in ordinary linear regression. Control estimates $\hat{\vec{u}}_t$ are simply the difference between the next state estimate, $\hat{\vec{s}}_t$ and the previous state, $\vec{s}_{t-1}$, with the distribution determined by the Gaussian model.

Of the $4,949$ mouse cursor trajectories, we randomly select $3,000$ as training data and use the remaining $1,949$ for evaluation. In Figure 4, we evaluate different choices of delayed feedback, t_0. We have no observation noise. Note that the model is equivalent to the LQR setting when $t_0 = 0$. As shown in this figure, the LQG setting with $t_0 = 3$ delay has the best performance. The Markov models of 3^{rd} and 4^{th} order outperform the LQR model, but are not more predictive than the LQG model with delay of 1, 2, 3, or 4. (Note that 1^{st} order Markov model is significantly worse and does not appear in the figure.) The advantage of the LQG model over the noise-less LQR inverse optimal control model and the direct policy estimation of the Markov model shows that modeling mouse pointing motions as an LQG problem is advantageous compared to the previous LQR model, which assumes instantaneous responses.

Other partial-observability mechanisms are likely to also influence the cursor pointing motions and improve the prediction of common overshooting and correcting motions. For example, a noisy observation model is more appropriate than the delayed perfect observation model we employ. However, our experiments provide a solid first step in predicting pointing motion control sequences using the LQG framework.

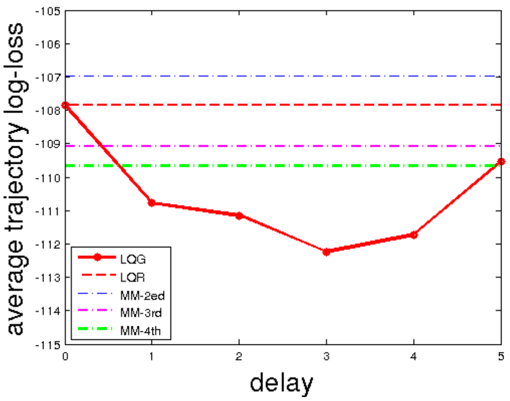

Figure 4: Average trajectory log-loss of: the LQG model with various amounts of delay, t_0; the LQR model; Markov models of order 2,3,4.

5 Discussion and Future Work

In this paper, we extended maximum entropy inverse optimal control to the LQG control setting. We established a separation property that allows inference in the resulting model to be performed efficiently. Despite the formulation of our approach being distinct from optimal LQG control, we found close connections between the two methods, including the ability to use an LQG solver as an integral part of the inverse LQG inference procedure. We demonstrated the advantages of the LQG representation for predictive inverse optimal control both on a synthetic dataset and on real mouse cursor data.

Of significant future interest are general methods for approximating non-linear control problems with observed state-observation-action trajectories placed within the LQG framework and using inverse optimal control to construct predictive models. These linearizing approximations have been well-studied in the control literature [14], but it remains to be seen whether reasonable learning can occur when an entire trajectory distribution must be approximated rather than the deterministic optimal controller.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No.1227495.

References

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proc. International Conference on Machine Learning*, pages 1–8, 2004.

[2] Monica Babes, Vukosi Marivate, Kaushik Subramanian, and Michael L Littman. Apprenticeship learning about multiple intentions. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 897–904, 2011.

[3] Ravin Balakrishnan. "Beating" Fitts law: virtual enhancements for pointing facilitation. *International Journal of Human-Computer Studies*, 61(6):857–874, 2004.

[4] R. Bellman. A Markovian decision process. *Journal of Mathematics and Mechanics*, 6:679–684, 1957.

[5] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. Linear matrix inequalities in system and control theory. *SIAM*, 15, 1994.

[6] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004.

[7] Jaedeug Choi and Kee-Eung Kim. Inverse reinforcement learning in partially observable environments. *The Journal of Machine Learning Research*, 12:691–730, 2011.

[8] T.M. Cover and J.A. Thomas. *Elements of information theory*. John Wiley and sons, 2006.

[9] Krishnamurthy Dvijotham and Emanuel Todorov. Inverse Optimal Control with Linearly-solvable MDPs. In *Proc. International Conference on Machine Learning*, pages 335–342, 2010.

[10] P. D. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, 32:1367–1433, 2004.

[11] P. Henry, C. Vollmer, B. Ferris, and D. Fox. Learning to Navigate Through Crowded Environments. In *Proc. International Conference on Robotics and Automation*, pages 981–986, 2010.

[12] R.E. Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

[13] G. Kramer. Capacity results for the discrete memoryless network. *Proc. IEEE Transactions on Information Theory*, 49(1):4–21, Jan 2003.

[14] Huibert Kwakernaak and Raphael Sivan. *Linear optimal control systems*, volume 1. Wiley-Interscience New York, 1972.

[15] Sergey Levine and Vladlen Koltun. Continuous inverse optimal control with locally optimal examples. In *International Conference on Machine Learning (ICML 2012)*, 2012.

[16] Sergey Levine and Vladlen Koltun. Variational policy search via trajectory optimization. In *Advances in Neural Information Processing Systems*, pages 207–215, 2013.

[17] Hans Marko. The bidirectional communication theory – a generalization of information theory. In *IEEE Transactions on Communications*, pages 1345–1351, 1973.

[18] James L. Massey. Causality, feedback and directed information. In *Proc. IEEE International Symposium on Information Theory and Its Applications*, pages 27–30, 1990.

[19] Gergely Neu and Csaba Szepesvári. Apprenticeship learning using inverse reinforcement learning and gradient methods. In *Proc. UAI*, pages 295–302, 2007.

[20] Andrew Y. Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *Proc. International Conference on Machine Learning*, pages 663–670, 2000.

[21] Haim H. Permuter, Young-Han Kim, and Tsachy Weissman. On directed information and gambling. In *Proc. IEEE International Symposium on Information Theory*, pages 1403–1407, 2008.

[22] D. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *Advances in Neural Information Processing Systems 1*, 1989.

[23] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *Proc. IJCAI*, pages 2586–2591, 2007.

[24] N. Ratliff, J. A. Bagnell, and M. Zinkevich. Maximum margin planning. In *Proc. ICML*, pages 729–736, 2006.

[25] Robert F Stengel. *Optimal control and estimation*. Courier Dover Publications, 2012.

[26] S. Tatikonda and S. Mitter. Control under communication constraints. *Automatic Control, IEEE Transactions on*, 49(7):1056–1068, 2004.

[27] F. Topsøe. Information theoretical optimization techniques. *Kybernetika*, 15(1):8–27, 1979.

[28] Brian D. Ziebart, J. Andrew Bagnell, and Anind K. Dey. Modeling interaction via the principle of maximum causal entropy. In *Proc. International Conference on Machine Learning*, pages 1255–1262, 2010.

[29] Brian D. Ziebart, Anind K. Dey, and J. Andrew Bagnell. Probabilistic pointing target prediction via inverse optimal control. In *Proceedings of the ACM International Conference on Intelligent User Interfaces*, pages 1–10, 2012.

[30] Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proc. AAAI Conference on Artificial Intelligence*, pages 1433–1438, 2008.

[31] Brian D. Ziebart, Andrew Maas, Anind K. Dey, and J. Andrew Bagnell. Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior. In *Proc. International Conference on Ubiquitous Computing*, pages 322–331, 2008.