The Principle of Maximum Causal Entropy for Estimating Interacting Processes

Brian D. Ziebart, J. Andrew Bagnell, and Anind K. Dey

***Abstract*—The principle of maximum entropy provides a powerful framework for estimating joint, conditional, and marginal probability distributions. However, there are many important distributions with elements of interaction and feedback where its applicability has not been established. This work presents the principle of maximum causal entropy—an approach based on directed information theory for estimating an unknown process based on its interactions with a known process. We demonstrate the breadth of the approach using two applications: a predictive solution for inverse optimal control in decision processes and computing equilibrium strategies in sequential games.**

***Index Terms*—Maximum entropy, statistical estimation, causal entropy, directed information, inverse optimal control, inverse reinforcement learning, correlated equilibrium.**

I. Introduction

THE principle of maximum entropy [21] serves a foundational role in the theory and practice of constructing statistical models [55], with applicability to statistical mechanics [21], [22], natural language processing [3], [47], [36], [43], econometrics [15], finance [10], [7], ecology [12], and other fields [26]. It provides robust prediction guarantees by prescribing the probability distribution estimate that only *commits* as far as required to satisfy existing knowledge about an unknown distribution, and is otherwise as uncertain as possible [57], [17]. Conditional extensions of the principle that consider a sequence of *provided information* (*i.e.*, additional variables that are *not* predicted, but are related to random variables that *are* predicted), and conditional random fields [31] specifically, have been applied with remarkable success in recognition, segmentation, and classification tasks. They are a preferred tool in natural language processing, [31], [53], computer vision [30], [48], and activity recognition [32], [59] applications.

In this work, we extend the maximum entropy approach to estimating probability distributions in settings characterized by *interaction with a known process*. For example, consider the task of estimating an agent's interactions with a stochastic environment. The agent may know how each of its available actions in each of its possible states will probabilistically transition to future states, but, due to stochasticity, it does not know what value each future state will take until after selecting the sequence of actions temporally preceding it. Existing maximum entropy approaches either assume that all of the values generated by the known process are available *a priori* (maximum conditional entropy and conditional random field [31] models) or treat both the known and unknown processes with the same degree of ignorance (maximum joint entropy models). *Interaction with a known process* lies in between these two extremes, requiring a new technique to construct appropriate probability distribution estimates.

Building on the recent advance of the Marko-Massey theory of directed information [34], [35], we present the *principle of maximum causal entropy*. It prescribes a probability distribution by maximizing the entropy of a sequence of random variables conditioned on the information available from the known process at each time step. This contribution extends the maximum entropy framework for statistical estimation to interacting processes. We motivate and apply this approach on decision prediction tasks that are characterized by actions that stochastically influence a system's sequentially revealed state. The principle of maximum causal entropy unifies recent inverse optimal control approaches from computer science [41], [1], [65], [13], [4] with structural estimation methods from econometrics [52], providing predictive guarantees for the former, and a more generalizable formulation to the latter. We demonstrate the approach's applicability using examples from inverse optimal control and multi-player dynamic games.

Though we emphasize the connection to decision making and sequential games in this work, it is important to note that the principle of maximum causal entropy is not specific to decision making domains. It is a general approach that is applicable to any setting where sequential data is generated from two interacting processes—one known and one unknown. Further, maximum causal entropy is compatible with existing conditional and joint entropy maximization techniques.

II. The Process Estimation Task

This work addresses the problem of estimating an unknown process that is interacting with a known process. Formally, the unknown process is a probability distribution over a sequence of random variables $\mathbf{Y}_{1:T} = (Y_1, Y_2, \ldots, Y_T)$ that take on values $\mathbf{y}_{1:T} = (y_1, y_2, \ldots, y_T)$ from sets $\boldsymbol{\mathcal{Y}}_{1:T} = \mathcal{Y}_1 \times \mathcal{Y}_2 \times \ldots \times \mathcal{Y}_T$ (the **predicted sequence**), given a different sequence $\mathbf{X}_{1:T}$ of symbols from sets $\boldsymbol{\mathcal{X}}_{1:T}$ (the **provided**

Manuscript received August 17, 2011; revised October 9, 2012; accepted November 26, 2012. The material in this paper was presented in part at the International Conference on Machine Learning, Haifa, Israel, June 21-24, 2010 [62], and at the International Conference on Autonomous Agents and Multiagent Systems, Taipei, Taiwan, May 2-6, 2011 [63].

B. D. Ziebart is with the Department of Computer Science, University of Illinois at Chicago, Chicago, IL, 60607 USA (e-mail: bziebart@uic.edu); J. A. Bagnell and A. K. Dey are with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213 USA (e-mail: dbagnell@ri.cmu.edu, anind@cs.cmu.edu).

sequence) generated from the known process. We refer the reader to Appendix A for a detailed description of the notation employed throughout this paper.

A. Motivating Example

We consider a simple motivating example with two probabilistic sender-receivers communicating over a noiseless, bi-directional binary channel. Each receives a one-bit message from the channel and then sends a one-bit message. We refer to the mapping from each sender-receiver's history to its new message as a **response function**. The response function of one sender-receiver is known and the response function of the other is unknown. However, some properties of their joint interactions are known: over communication sequences of infinite length, the number of "0" bits and "1" bits sent over the channel are equal, and the frequency of five consecutive "1" bits over the channel is less than 3%.

Given this setting, how should the unknown sender-receiver response function be estimated? There are two important characteristics of the task to consider. First, the known properties of the channel's usage often do not fully constrain the unknown elements; many different schemes for the unknown encoder may realize those properties. Second, those known properties are defined in terms of the interaction between the known and unknown sender-receivers rather than isolated properties only of the unknown sender-receiver.

B. Conditional Probability Distribution Estimation

The typical process estimation approach uses available observations from the unknown process to fit a parametric conditional probability distribution. For example, a conditional multinomial distribution, logistic function, or Gaussian distribution could be employed using a stationarity assumption to estimate a time-invariant process, $P(Y_t|X_t, Y_{t-1})$. By employing observations from the interacting known and unknown processes, this approach allows the unknown process to be estimated independently of any knowledge of the provided process (beyond the observations).

As seen by our motivating example (Section II-A), it is quite natural to consider characteristics of interacting processes that are defined over the joint distribution of message sequences. Unfortunately, these types of known characteristics cannot be appropriately leveraged when independently estimating conditional probability distributions. For example, the property of having long-term parity in the communication channel bits explicitly depends on both the known process and the unknown process. To overcome these limitations, formulations that consider the interaction of the known and unknown processes—rather than treating them separately—are needed.

C. Joint Probability Decompositions

By the chain rule, any joint probability distribution can be represented as a product of conditional probabilities. The canonical factorization for a joint distribution of two sequences of random variables is:

$$P(\mathbf{X}_{1:T}, \mathbf{Y}_{1:T}) = \prod_{t=1}^{T} P\left(X_t|\mathbf{X}_{1:t-1}\right) \prod_{t=1}^{T} P\left(Y_t|\mathbf{Y}_{1:t-1}, \mathbf{X}_{1:T}\right),$$

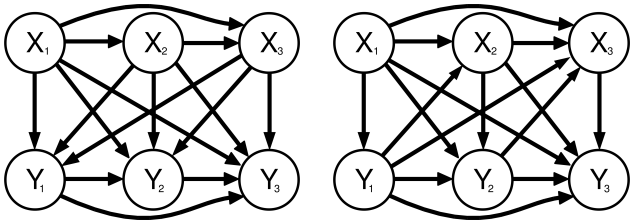

Fig. 1. Probabilistic graphical model representations [27] for: the canonical decomposition of the joint probability distribution where the three time step sequence of $\mathbf{Y}$ variables is conditioned on the sequence of $\mathbf{X}$ variables (**left**); and the decomposition of the three time step joint probability distribution into two interacting processes (**right**). The two differ in the direction of three edges connecting $\mathbf{X}_{2:3}$ variables with $\mathbf{Y}_{1:2}$ variables.

in which the probability distribution for given information, $P(\mathbf{X}_{1:T})$, is first formed, and then the probability of the sequence of predicted random variables $\mathbf{Y}_{1:T}$ is multiplied in using conditional probability distributions that condition on all of the provided variables to form the joint distribution (Figure 1, left). Unfortunately, this decomposition for the $\mathbf{X}_{1:T}$ random variables does not coincide with a known process, since the $\mathbf{X}_{1:T}$ random variables should also depend on previous $\mathbf{Y}_{1:T}$ random variables—the other sender-receiver's messages in our earlier example. Nor does the distribution for Y_t correspond to a known process; its distribution under this decomposition violates the properties of temporal processes—the conditional probabilities $P(Y_t|\mathbf{Y}_{1:t-1}, \mathbf{X}_{1:T})$ depend on future variables, $\mathbf{X}_{t+1:T}$, as also indicated by the anti-temporal edges in the left of Figure 1 (*e.g.*, from X_3 to Y_1).

An alternative application of the chain rule to the joint sequence distribution factors the provided and the predicted variables as two interacting temporal processes:

$$P(\mathbf{X}_{1:T}, \mathbf{Y}_{1:T}) = \overbrace{\prod_{t=1}^{T} P\left(X_t|\mathbf{X}_{1:t-1}, \mathbf{Y}_{1:t-1}\right)}^{\text{provided process}} \times \underbrace{\prod_{t=1}^{T} P\left(Y_t|\mathbf{Y}_{1:t-1}, \mathbf{X}_{1:t}\right)}_{\text{unknown process}}. \quad (1)$$

This decomposition coincides with settings having feedback, such as our simple communications example or sequential decision making processes. In decision settings, it is natural to have a model of the state dynamics process, $P(X_t|\mathbf{X}_{1:t-1}, \mathbf{Y}_{1:t-1})$, in which the next state depends on the previous controls, $\mathbf{Y}_{1:t-1}$, and states, $\mathbf{X}_{1:t-1}$. This feedback cycle is shown by the directed paths from, *e.g.*, Y_1 to X_2 to Y_2 on the right of Figure 1 and has a causal interpretation: future state variables are unknown when past controls are selected and, thus, their values have no direct influence on preceding control variables.

We make use of this decomposition throughout the remainder of this paper when we are estimating the latter process of the joint distribution (*i.e.*, the controller's decision process), $\prod_{t=1}^{T} P\left(Y_t|\mathbf{Y}_{1:t-1}, \mathbf{X}_{1:t}\right)$, when the former process (*i.e.*, the state-transition dynamics process), $\prod_{t=1}^{T} P\left(X_t|\mathbf{X}_{1:t-1}, \mathbf{Y}_{1:t-1}\right)$, is known.

III. The Principle of Maximum Causal Entropy

Motivated by the task of estimating a process based on its interactions with another known process without relying on local estimations, we introduce the principle of maximum causal entropy in this section.

A. Directed Information Theory

The Marko-Massey theory of directed information [34], [35] has investigated the components of the interacting temporal decomposition of the joint distribution (Equation 1). The **causally conditioned probability** [28],

$$P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) \triangleq \prod_{t=1}^{T} P(Y_t|\mathbf{Y}_{1:t-1}, \mathbf{X}_{1:t}), \tag{2}$$

reflects the causal restriction that future provided variables (*e.g.*, $\mathbf{X}_{\tau+1:T}$) do not influence earlier predicted variables (*e.g.*, $\mathbf{Y}_{1:\tau}$). In contrast to the **conditional probability**,

$$P(\mathbf{Y}_{1:T}|\mathbf{X}_{1:T}) \triangleq \prod_{t=1}^{T} P(Y_t|\mathbf{Y}_{1:t-1}, \mathbf{X}_{1:T}), \tag{3}$$

each Y_t variable of the causally conditioned probability (Equation 2) is only conditioned on previous variables, $\mathbf{X}_{1:t}$, rather than also being conditioned on future variables, $\mathbf{X}_{1:T}$. This subtle, but significant, difference from the conditional probability (Equation 3) serves as the basis for our approach. Multiplicatively combining with the complementary causally conditioned distribution,

$$P(\mathbf{X}_{1:T}||\mathbf{Y}_{1:T-1}) \triangleq \prod_{t=1}^{T} P(X_t|\mathbf{X}_{1:t-1}, \mathbf{Y}_{1:t-1}), \tag{4}$$

yields the joint probability distribution, $P(\mathbf{X}_{1:T}, \mathbf{Y}_{1:T}) = P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})\ P(\mathbf{X}_{1:T}||\mathbf{Y}_{1:T-1})$, following the same decomposition shown in Equation 1.

The uncertainty or "non-committedness" of a probability distribution is measured by the Shannon entropy [54]. The **conditional entropy**, $H(\mathbf{Y}_{1:T}|\mathbf{X}_{1:T}) = \mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[-\log P(\mathbf{y}_{1:T}|\mathbf{x}_{1:T})]$, measures this uncertainty when provided information, $\mathbf{x}_{1:T}$, is available up front. The analogous notion of **causal entropy** (Definition 1) for directed information theory measures the uncertainty present in the causally conditioned distribution of the $\mathbf{Y}_{1:T}$ variable sequence given the preceding partial $\mathbf{X}_{1:T}$ variable sequence.

Definition 1: The **causal entropy** [28], [46] of $\mathbf{Y}_{1:T}$ given $\mathbf{X}_{1:T}$ is:

$$\begin{aligned} H(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) &\triangleq \mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[-\log P(\mathbf{y}_{1:T}||\mathbf{x}_{1:T})] \\ &= \sum_{t=1}^{T} H(Y_t|\mathbf{Y}_{1:t-1}, \mathbf{X}_{1:t}). \end{aligned} \tag{5}$$

It can be interpreted as the expected number of *bits* (when $\log_2$ is employed) needed to minimally encode samples from the sequence $\mathbf{Y}_{1:T}$, iteratively over $t \in \{1, \ldots, T\}$, given the previous $\mathbf{Y}_{1:t-1}$ variables and sequentially revealed input, $\mathbf{X}_{1:t}$, up to that point in time, and excluding unrevealed future provided variables $\mathbf{X}_{t+1:T}$. It thus measures the compressibility of information in a feedback channel [28].

Causal entropy can be incorporated with other entropy measures using its conditional entropy decomposition from Definition 1—for instance, with joint variables and with traditional conditioning,

$$\begin{aligned} &H(\mathbf{W}_{1:T}, \mathbf{Y}_{1:T}||\mathbf{X}_{1:T}|\mathbf{Z}_{1:T}) \\ &\triangleq \sum_{t=1}^{T} H(W_t, Y_t|\mathbf{W}_{1:t-1}, \mathbf{Y}_{1:t-1}, \mathbf{X}_{1:t}, \mathbf{Z}_{1:T}), \end{aligned}$$

which we discuss in more detail in Appendix A.

Causal entropy has previously been applied in the analysis of communication channels with feedback [28], decentralized control [56], inferring causal relationships [49], [50], sequential investment and online compression with provided information [46]. This work contributes the notion of causal entropy for estimating probability distributions.

Definition 2: The **causal cross entropy** or **causal log-loss** of $\mathbf{Y}_{1:T}$ given $\mathbf{X}_{1:T}$ for causal distribution $\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$ under joint distribution $P(\mathbf{Y}_{1:T}, \mathbf{X}_{1:T})$ is:

$$\mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[-\log \hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})]. \tag{6}$$

The **causal log likelihood** of data distributed according to $P(\mathbf{Y}_{1:T}, \mathbf{X}_{1:T})$ is the negative of the causal log-loss (Definition 2). The causal log-loss measures the compressibility of a feedback channel when using a causally conditioned probability distribution estimate, $\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$, rather than the true causally conditioned distribution. It is a natural measure for evaluating a causally conditioned probability estimator and can be directly related to the expected growth rate of gambling on outcome sequence $\mathbf{Y}_{1:T}$ under uniform odds [9], [46].

B. Causally Conditioned Probability Distributions via Affine Constraints

Unfortunately, the definition of causally conditioned probabilities as products of conditional probabilities (Equation 2) is not well-suited for optimization procedures—it is a non-linear function of the unknown conditional probabilities, $\{P(y_t|\mathbf{y}_{1:t-1}, \mathbf{x}_{1:t})\}_{t\in\{1,\ldots,T\}, \mathbf{x}_{1:t}\in\boldsymbol{\mathcal{X}}_{1:t}, \mathbf{y}_{1:t}\in\boldsymbol{\mathcal{Y}}_{1:t}}$. In this section, we introduce an affinely constrained definition of causally conditioned probabilities that supports convex optimization, and show that it is equivalent to the previous definition.

Definition 3: The class of **causally conditioned probability distributions**, denoted Ξ, is defined by the following **causal polytope** of affine constraints for any $P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) \in \Xi$[1]:

$$\forall \mathbf{x}_{1:T} \in \boldsymbol{\mathcal{X}}_{1:T}, \mathbf{y}_{1:T} \in \boldsymbol{\mathcal{Y}}_{1:T},\ P(\mathbf{y}_{1:T}||\mathbf{x}_{1:T}) \geq 0; \tag{7}$$

$$\forall \mathbf{x}_{1:T} \in \boldsymbol{\mathcal{X}}_{1:T}, \sum_{\mathbf{y}_{1:T}\in\boldsymbol{\mathcal{Y}}_{1:T}} P(\mathbf{y}_{1:T}||\mathbf{x}_{1:T}) = 1; \text{ and} \tag{8}$$

$$\forall \tau \in \{1, \ldots, T\},\ \mathbf{y}_{1:T} \in \boldsymbol{\mathcal{Y}}_{1:T}, \mathbf{x}_{1:T} \in \boldsymbol{\mathcal{X}}_{1:T}, \mathbf{x}'_{1:T} \in \boldsymbol{\mathcal{X}}_{1:T}$$
$$\text{such that: } \mathbf{x}_{1:\tau} = \mathbf{x}'_{1:\tau},$$
$$\sum_{\mathbf{y}_{\tau+1:T}\in\boldsymbol{\mathcal{Y}}_{\tau+1:T}} (P(\mathbf{y}_{1:T}||\mathbf{x}_{1:T}) - P(\mathbf{y}_{1:T}||\mathbf{x}'_{1:T})) = 0. \tag{9}$$

[1]Though we present the form for discrete-valued random variables, a similar set of constraints defines the causally conditioned probability distribution over continuous-valued variables.

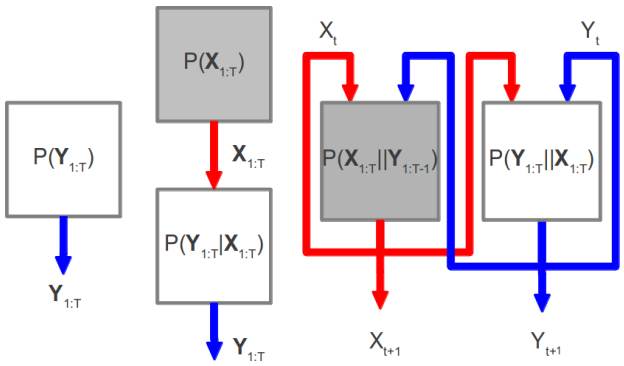

Fig. 2. Three probability distribution estimation tasks of the unknown white components (given the known gray components): joint distribution estimation of $P(\mathbf{Y}_{1:T})$ (**left**); conditional distribution estimation of $P(\mathbf{Y}_{1:T}|\mathbf{X}_{1:T})$ (**center**); and causally conditioned distribution estimation of $P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$ given a known interacting process $P(\mathbf{X}_{1:T}||\mathbf{Y}_{1:T-1})$ (**right**), which can be estimated by maximizing the joint entropy, the conditional entropy, and the causal entropy, respectively.

The final set of constraints (Equation 9) ensures **causal independence**—that past conditioned variables $\mathbf{y}_{1:\tau}$ are not functions of future conditioning variables $\mathbf{x}_{\tau+1:T}$. Without it, a conditional probability distribution would be defined.

Theorem 1: Using the definition of causally conditioned probabilities in terms of affine constraints (Definition 3), interdependent causally conditioned probability distributions together form valid joint probability distributions[2]:

$$\Big(\forall \mathbf{x}_{1:T} \in \boldsymbol{\mathcal{X}}_{1:T}, \mathbf{y}_{1:T} \in \boldsymbol{\mathcal{Y}}_{1:T},$$
$$P(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}) = P(\mathbf{y}_{1:T}||\mathbf{x}_{1:T})\ P(\mathbf{x}_{1:T}||\mathbf{y}_{1:T-1})\Big)$$
$$\Rightarrow \Big(\forall \mathbf{x}_{1:T} \in \boldsymbol{\mathcal{X}}_{1:T}, \mathbf{y}_{1:T} \in \boldsymbol{\mathcal{Y}}_{1:T},\ P(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}) \geq 0 \quad (10)$$
$$\text{and} \sum_{\substack{\mathbf{x}_{1:T} \in \boldsymbol{\mathcal{X}}_{1:T}, \\ \mathbf{y}_{1:T} \in \boldsymbol{\mathcal{Y}}_{1:T}}} P(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}) = 1\Big). \quad (11)$$

Corollary 1: The causally conditioned probability distributions defined according to affine constraints (Definition 3) are equivalent to the causally conditioned probability distributions defined by the decomposition into a product of conditional probabilities: $P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) = \prod_{t=1}^{T} P(Y_t|\mathbf{Y}_{1:t-1}, \mathbf{X}_{1:t})$.

This formulation of the causally conditioned probability distribution using the causal polytope (Definition 3) enables the efficient optimization for the principle of maximum causal entropy. Throughout the remainder of this paper, whenever the variables of an optimization correspond to a causally conditioned probability distribution, they should be interpreted to reside within the causal polytope of Definition 3.

C. Maximum Causal Entropy

The principle of maximum entropy [21] prescribes the probability distribution estimator that is the "least committed" (or most uncertain) apart from matching known properties of the distribution being estimated. This is realized by maximizing the Shannon entropy [54] subject to constraints. Many of the fundamental building block distributions of statistics (*e.g.*, Gaussians), though often derived by other means, can be obtained by this approach using moment-matching constraints. In fact, there is a general duality between maximum entropy (or conditional entropy) estimation problems and maximum likelihood (or conditional likelihood) estimation of exponential family probability distributions [23]. For example, maximizing the conditional entropy, $H_{\hat{P}}(\mathbf{Y}_{1:T}|\mathbf{X}_{1:T}) = \mathbb{E}_{\hat{P}(\mathbf{Y}_{1:T}, \mathbf{X}_{1:T})}[-\log \hat{P}(\mathbf{y}_{1:T}|\mathbf{x}_{1:T})]$, given constraints on cliques of variables, $\mathcal{C}_i \subseteq \{1, \ldots, T\}$,

$$\forall i \in \{1, \ldots, K\}, \mathbb{E}_{\hat{P}(\mathbf{Y}_{1:T}, \mathbf{X}_{1:T})}[f_i(\mathbf{y}_{\mathcal{C}_i}, \mathbf{x}_{\mathcal{C}_i})] = \mathbb{E}_{P(\mathbf{Y}_{1:T}, \mathbf{X}_{1:T})}[f_i(\mathbf{y}_{\mathcal{C}_i}, \mathbf{x}_{\mathcal{C}_i})], \quad (12)$$

where $\mathbf{x}_{\mathcal{C}_i}$ is a subset of $\mathbf{x}_{1:T}$ and $P(\mathbf{Y}_{1:T}, \mathbf{X}_{1:T})$ denotes the true distribution being estimated, yields **conditional random fields** [31] (Figure 2, center),

$$\hat{P}(\mathbf{y}_{1:T}|\mathbf{x}_{1:T}) \propto e^{\sum_{i=1}^{K} \theta_i f_i(\mathbf{y}_{\mathcal{C}_i}, \mathbf{x}_{\mathcal{C}_i})}, \quad (13)$$

a state-of-the-art statistical estimation technique.

We extend the principle of maximum entropy to estimate (with estimator $\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$) an *unknown* causally conditioned probability distribution, $P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$, that interacts with a *known* causally conditioned probability distribution, $P(\mathbf{X}_{1:T}||\mathbf{Y}_{1:T-1})$, as shown on the right of Figure 2. Together, these probability distributions satisfy a set of constraints defined in terms of the joint distribution $\hat{P}(\mathbf{Y}_{1:T}, \mathbf{X}_{1:T}) = \hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})P(\mathbf{X}_{1:T}||\mathbf{Y}_{1:T-1})$ (Definition 4) and the unknown distribution can be obtained as the result of a convex optimization problem (Theorem 2).

Definition 4: The **principle of maximum causal entropy** prescribes the causally conditioned entropy-maximizing probability distribution estimator, $\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$, from the causal polytope Ξ (Definition 3):

$$\underset{\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) \in \Xi}{\operatorname{argmax}} H_{\hat{P}}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) \quad (14)$$
$$\text{such that: } \mathbf{g}\left(\hat{P}(\mathbf{Y}_{1:T}, \mathbf{X}_{1:T})\right) = \mathbf{0}, \text{ and}$$
$$\mathbf{h}\left(\hat{P}(\mathbf{Y}_{1:T}, \mathbf{X}_{1:T})\right) \geq \mathbf{0},$$

for affine functions[3] $\mathbf{g} : \Delta_{\mathcal{X}_{1:T}, \mathcal{Y}_{1:T}} \to \mathbb{R}^M$ and $\mathbf{h} : \Delta_{\mathcal{X}_{1:T}, \mathcal{Y}_{1:T}} \to \mathbb{R}^N$.

More specifically, the two affine constraints can always be written as expectations of feature functions, $\boldsymbol{\mathcal{F}}_g : \boldsymbol{\mathcal{Y}}_{1:T} \times \boldsymbol{\mathcal{X}}_{1:T} \to \mathbb{R}^M$ and $\boldsymbol{\mathcal{F}}_h : \boldsymbol{\mathcal{Y}}_{1:T} \times \boldsymbol{\mathcal{X}}_{1:T} \to \mathbb{R}^N$ ($\mathbf{c}_g \in \mathbb{R}^M$, $\mathbf{c}_h \in \mathbb{R}^N$):

$$\mathbf{g}\left(\hat{P}(\mathbf{Y}_{1:T}, \mathbf{X}_{1:T})\right) = \mathbb{E}_{\hat{P}(\mathbf{Y}_{1:T}, \mathbf{X}_{1:T})}[\boldsymbol{\mathcal{F}}_g(\mathbf{y}_{1:T}, \mathbf{x}_{1:T})] + \mathbf{c}_g \quad (15)$$

$$\mathbf{h}\left(\hat{P}(\mathbf{Y}_{1:T}, \mathbf{X}_{1:T})\right) = \mathbb{E}_{\hat{P}(\mathbf{Y}_{1:T}, \mathbf{X}_{1:T})}[\boldsymbol{\mathcal{F}}_h(\mathbf{y}_{1:T}, \mathbf{x}_{1:T})] + \mathbf{c}_h, \quad (16)$$

which we will make use of later in this work. We note that these constraints are also affine in the unknown

[2]The proofs of each theorem and corollary are presented in Appendix B.

[3]More generally, $\mathbf{h}(P(\mathbf{Y}_{1:T}, \mathbf{X}_{1:T}))$ can be a convex function and the same strong Lagrangian duality developed in this work applies, subject to appropriate primal feasibility requirements. An example of this is in estimation techniques for rationalizing observed game play [60].

$\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$ variables, but not in the multiplicative factors of those variables, *e.g.*, $\hat{P}(Y_t|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1})$.

Theorem 2: The maximum causal entropy optimization problem (Equation 14) is a convex optimization problem.

D. Lagrangian Duality

The primal problem of the maximum causal entropy optimization (Equation 14) is a potentially high-dimensional one in the space of probability distributions. The Lagrangian dual may be much more compact when the feature function dimensionality, $M+N$, is smaller than the causally conditioned probability distribution's dimensionality, $|\boldsymbol{\mathcal{X}}_{1:T}||\boldsymbol{\mathcal{Y}}_{1:T}|$. As we shall show, its optimal solution is also an optimal solution for the primal problem.

Theorem 3: The Lagrangian dual optimization problem of the primal maximum causal entropy problem (Definition 4) is:

$$\min_{\boldsymbol{\lambda},\boldsymbol{\gamma}\geq\mathbf{0}} \sum_{x_1\in\mathcal{X}_1} P(x_1)\log Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(x_1) + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{c}_g + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{c}_h \tag{17}$$

$$\text{where:}\quad Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1}) = \sum_{y_t\in\mathcal{Y}} Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(y_t|\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})$$

$$\text{and}\quad Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(y_t|\mathbf{x}_{1:t},\mathbf{y}_{1:t-1}) = \begin{cases} e^{\sum_{x_{t+1}\in\mathcal{X}} P(x_{t+1}|\mathbf{x}_{1:t},\mathbf{y}_{1:t})\log Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{x}_{1:t+1},\mathbf{y}_{1:t})} & t<T \\ e^{\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\mathcal{F}}_g(\mathbf{x}_{1:T},\mathbf{y}_{1:T})+\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\mathcal{F}}_h(\mathbf{x}_{1:T},\mathbf{y}_{1:T})} & t=T. \end{cases}$$

The solution to this problem (*i.e.*, the estimated probability distribution), can be expressed recursively as:

$$\hat{P}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(y_t|\mathbf{x}_{1:t},\mathbf{y}_{1:t-1}) = \frac{Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(y_t|\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})}{Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})}. \tag{18}$$

Due to convexity (Theorem 2), the optimal values obtained by the dual and primal optimization problems are equivalent under mild technical considerations. (Theorem 4).

Theorem 4: Strong Lagrangian duality [6], *i.e.*, no gap between the primal optimization problem (Equation 14) and the dual optimization problem (Equation 17), holds for the maximum causal entropy estimation task when a feasible solution to the primal optimization (Equation 14) exists on the relative interior, *i.e.*, $P(\mathbf{y}_{1:T}||\mathbf{x}_{1:T}) > 0$ $(\forall \mathbf{y}_{1:T}\in\boldsymbol{\mathcal{Y}}_{1:T}, \mathbf{x}_{1:T}\in\boldsymbol{\mathcal{X}}_{1:T})$.

Sub-gradient-based optimization[4] with adaptive learning rate $\eta_i \in \mathbb{R}^+$ can be employed to obtain optimal parameters $(\boldsymbol{\lambda}^*,\boldsymbol{\gamma}^*) = \lim_{i\to\infty}(\boldsymbol{\lambda}_{(i)},\boldsymbol{\gamma}_{(i)})$ using parameter updates:

$$\begin{aligned} \boldsymbol{\lambda}_{(i+1)} &\leftarrow \boldsymbol{\lambda}_{(i)} + \eta_i\left(\mathbb{E}_{\hat{P}(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[\boldsymbol{\mathcal{F}}_g(\mathbf{y}_{1:T},\mathbf{x}_{1:T})] + \mathbf{c}_g\right) \\ \boldsymbol{\gamma}_{(i+1)} &\leftarrow \max\{\mathbf{0}, \\ &\qquad \boldsymbol{\gamma}_{(i)} + \eta_i\left(\mathbb{E}_{\hat{P}(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[\boldsymbol{\mathcal{F}}_h(\mathbf{y}_{1:T},\mathbf{x}_{1:T})] + \mathbf{c}_h\right)\}, \end{aligned} \tag{19}$$

with expectations calculated according to the Lagrangian dual's form of the probability distribution (Equation 18).

[4]Other convex optimization techniques (*e.g.*, gradient ascent, Newton's method, interior-point methods) with guarantees of convergence to a global optima are also applicable for different sets of constraints.

E. Maximum Causal Likelihood

The equivalency of maximum entropy estimation and maximum likelihood estimation in exponential random families [23] extends to the causally conditioned setting.

Theorem 5: Subject to moment-matching constraints, *i.e.*, $\mathbb{E}_{\hat{P}(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[\boldsymbol{\mathcal{F}}_g(\mathbf{x}_{1:T})] = \mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[\boldsymbol{\mathcal{F}}_g(\mathbf{x}_{1:T})]$ (via $\mathbf{c}_g = -\mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[\boldsymbol{\mathcal{F}}_g(\mathbf{x}_{1:T})]$ in Equation 15), using statistics from the true distribution within Definition 4 and no inequality constraints, maximizing the causal entropy is equivalent to maximizing the (log) causal likelihood (Definition 2) of the true data distribution,

$$\max_{\boldsymbol{\lambda}} \mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[\log \hat{P}_{\boldsymbol{\lambda}}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})]. \tag{20}$$

Often, moment statistics are estimated from a limited number of samples, $\mathbb{E}_{\tilde{P}(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[\boldsymbol{\mathcal{F}}_g(\mathbf{x}_{1:T},\mathbf{y}_{1:T})]$, where the sample distribution $\tilde{P}(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})$ is obtained from n samples of the joint distribution. This causes probabilistically bounded approximation error (Theorem 6) as opposed to using the true joint distribution as in Theorem 5.

Theorem 6: If $\bar{\mathbf{f}}_g$ are sample means of the statistic $\boldsymbol{\mathcal{F}}_g(\mathbf{x}_{1:T},\mathbf{y}_{1:T}) \in [\mathbf{f}_{\min},\mathbf{f}_{\max}]$ (*i.e.*, under distribution $\tilde{P}(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})$), then the difference between sample mean and expectation is bounded as:

$$\begin{aligned} &P\left(\left|\left|\bar{\mathbf{f}}_g - \mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[\boldsymbol{\mathcal{F}}_g(\mathbf{y}_{1:T},\mathbf{x}_{1:T})]\right|\right|_\infty \geq \epsilon\right) \\ &\qquad \leq 2K\exp\left(-\frac{2n\epsilon^2}{||\mathbf{f}_{g,\max}-\mathbf{f}_{g,\min}||_\infty^2}\right). \end{aligned}$$

The constraints of the maximum causal entropy primal optimization problem (Equation 14) can be relaxed to allow a small amount of slack to address this approximation error. This leads to regularized maximum causal likelihood estimation in the dual optimization problem [12], which is a common statistical estimation technique to avoid overfitting to a small sample data set.

F. Robust Performance Guarantees

Though the principle of maximum entropy is often justified with the philosophical argument that no additional assumptions should be made except known constraints [21], it can instead be derived as a robust estimation procedure for minimizing the predictive log-loss [57], [17]. We present the principle of maximum causal entropy's derivation as a robust causally conditioned probability estimator in this section.

We consider the setting in which the joint probability distribution, $P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})$, that is obtained by combining an unknown causally conditioned probability distribution, $P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$, with a known causally conditioned probability distribution, $P(\mathbf{X}_{1:T}||\mathbf{Y}_{1:T-1})$, is known to satisfy a set of convex constraints (*e.g.*, those from Equation 14). We would like to construct an estimator $\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$ that minimizes the causal log-loss (Definition 2) evaluated according to the joint distribution $P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})$. However, since the joint distribution is only partially known, our estimator can be made more robust (*i.e.*, better worst-case performance) compared to estimators that make unwarranted assumptions, by instead treating unknown factors of the joint distribution as being

chosen adversarially (*i.e.*, to maximize the estimator's log-loss).

This setting can be viewed as a two-step game in which the first player chooses an estimate for each possible contingency of the sequence, $\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$, and then the second player, with knowledge of this choice, chooses the distribution $P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$ from a restricted subset of causally conditioned probability distributions that satisfy known constraints. The resulting **adversarial causal log-loss minimization** task is:

$$\inf_{\substack{\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) \\ \in \Xi}} \sup_{\substack{P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) \\ \in \Xi}} \mathbb{E}_{P(\mathbf{Y},\mathbf{X})}[-\log \hat{P}(\mathbf{y}_{1:T}||\mathbf{x}_{1:T})] \quad (21)$$
$$\text{such that: } \mathbf{g}(P(\mathbf{Y}_{1:T}, \mathbf{X}_{1:T})) = \mathbf{0} \text{ and}$$
$$\mathbf{h}(P(\mathbf{Y}_{1:T}, \mathbf{X}_{1:T})) \geq \mathbf{0}.$$

Theorem 7: Adversarial log-loss minimization (Equation 21) is equivalent to maximizing the causal entropy subject to the same sets of constraints (Equation 14) under the assumptions of Theorem 4.

G. Generalization and Special Cases

Entropy measures are implicitly relative to a uniform probability distribution. They can be generalized using the **relative causal entropy** or **causal Kullback-Leibler divergence** [29],

$$\mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}\left[\log \frac{P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})}{P_0(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})}\right], \quad (22)$$

given baseline causally conditioned probability distribution $P_0(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$. This is generally necessary to ensure invariance to variable transformations when estimating continuous probability distributions [55]. In the causal setting, this yields log causally conditioned probability distributions that are in proportion to $P_0(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$. While it may be attractive to incorporate background knowledge by using a "non-uniform" relative probability distribution, the equivalent model can be learned (possibly from a larger set of potential baseline distributions $\{P_0(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})\}$) by incorporating constraint: $\mathbb{E}_{\tilde{P}(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[\log P_0(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})] = \mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[\log P_0(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})]$ and using a standard "uniform" relative probability distribution.

When $P(\mathbf{X}_{1:T}||\mathbf{Y}_{1:T-1})$ is a deterministic probability distribution (*i.e.*, $\forall t \in \{1,\dots,T\}, \mathbf{x}_{1:t} \in \boldsymbol{\mathcal{X}}_{1:t}, \mathbf{y}_{1:t} \in \boldsymbol{\mathcal{Y}}_{1:t}, P(\mathbf{x}_{1:t}||\mathbf{y}_{1:t-1}) \in \{0,1\}$, in contrast to $[0,1]$), there is no uncertainty in the future provided information given the sequence of previously occurring predicted variables. In this case, the causal entropy reduces to the conditional entropy, where the conditioning information can be thought of as the variables specifying the next conditioned variable given a history, $\mathcal{T}_{\mathbf{x}_{1:t-1},\mathbf{y}_{1:t-1}} \in \mathcal{X}_t$. This special case has been investigated for modeling decision making [65] and applied to predicting the driving routes and destinations of drivers on road networks [66], and the movements of pedestrians for more intelligent robotic path planning [67], [19].

IV. Inverse Optimal Control

Stochastic decision problems closely match the causal assumption of our approach. Typically, prescribing the optimal action to employ given a cost or reward measure has been of central focus for decision theorists. However, understanding the inverse problem—the recovery of a reward function that motivates observed behavior in sequential decision settings—is also important for a number of behavior forecasting applications. Though our formulation of the maximum causal entropy estimation approach does not rely on a control-based perspective, we show in this section that it provides a general probabilistic solution to the inverse optimal control problem.

A. Background

Inverse optimal control (also known as inverse reinforcement learning) [25], [5], [41] describes the problem of recovering an agent's reward function, given a controller or policy, when the remainder of the decision process is known. We consider the discrete decision process formulation where the rewards motivating behavior are linearly parameterized [41], [1][5].

Definition 5: A **parametric-reward Markov decision problem** (θ-MDP) is defined as a tuple, $\mathcal{M}_{\theta\text{-MDP}} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \boldsymbol{\mathcal{F}}, \boldsymbol{\theta})$, comprising: a set of states, $\mathcal{S}$; a set of actions, $\mathcal{A}$; state transition dynamics, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$, probabilistically mapping state-action pairs, $s_t \in \mathcal{S}$ and $a_t \in \mathcal{A}$ to next state $s_{t+1} \in \mathcal{S}$ according to $\mathcal{T}(S_{t+1}|S_t, A_t)$; a mapping, $\boldsymbol{\mathcal{F}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^K$ of state-action pairs to feature vectors, denoted $\mathbf{f}(s,a)$; and a weight vector, $\boldsymbol{\theta} \in \mathbb{R}^K$, compactly parameterizing the rewards.

The reward received for selecting action $a \in \mathcal{A}$ when in state $s \in \mathcal{S}$ is $R_{\boldsymbol{\theta}}(s,a) = \boldsymbol{\theta}^{\mathrm{T}}\mathbf{f}(s,a)$, and the total reward[6] received over time is: $\mathbb{E}_{P(\mathbf{S}_{1:T},\mathbf{A}_{1:T})}[\sum_{t=1}^{T} R_{\boldsymbol{\theta}}(s_t,a_t)]$. The MDP is **solved** for a specific vector of reward weights by finding the **policy**, $\pi : \mathcal{S} \to \mathcal{A}$ (more generally stochastic, $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$), prescribing an action to take in each state, that maximizes the total reward.

Inverse optimal control techniques that assume behavior is optimal for some choice of reward weights [41], [8] are often ill-posed [41]—many reward weights, including degeneracies (*e.g.*, the *all zeros* reward vector), will make observed behavior optimal—and, when observed behavior is noisy and inherently sub-optimal, degenerate solutions will often be the only reward parameters that can make observe behavior optimal.

Abbeel & Ng [1] resolve some of these difficulties by recovering a (mixture of) $\boldsymbol{\theta}$-MDP solution(s) guaranteeing the same reward (in expectation) as the demonstrated trajectories for any choice of parameter $\boldsymbol{\theta}$. This reduces to matching the optimal controller's expected **feature counts**, $\boldsymbol{\mathcal{F}}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = \sum_{t=1}^{T} \mathbf{f}(s_t, a_t)$ with those of the demonstrated trajectories:

$$\mathbb{E}_{P(\mathbf{S}_{1:T},\mathbf{A}_{1:T})}\left[\boldsymbol{\mathcal{F}}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})\right] = \mathbb{E}_{\tilde{P}(\mathbf{S}_{1:T},\mathbf{A}_{1:T})}\left[\boldsymbol{\mathcal{F}}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})\right],$$

where $P(\mathbf{S}_{1:T}, \mathbf{A}_{1:T}) = \mathcal{T}(\mathbf{S}_{1:T}||\mathbf{A}_{1:T-1})\ \pi(\mathbf{A}_{1:T}||\mathbf{S}_{1:T})$. Unfortunately, when sub-optimal behavior is demonstrated

[5]The continuous state and control setting has been investigated [62] and applied to predicting computer cursor pointing targets from partial motion trajectories [64].

[6]We consider finite horizons, T, in this work, but infinite horizons can also be considered by requiring the decision process to terminate with some probability after each time step, *i.e.*, a discount factor, or that some states are absorbing to make the total reward received finite.

(due to the agent's imperfection or inevitable approximations of the model), the approach can assign zero probability to demonstrated training behavior [65]. Other inverse optimal control approaches can avoid this issue [51], [40], but lack reward-equivalency performance guarantees.

Structural estimation [52] takes a latent-variable perspective to the problem. In addition to assuming that the reward function of a state-action pair is a linear function of known features, $\mathbf{f}(s_t, a_t)$, unobservable influences also contribute to the reward. These are incorporated as exogenous "shock" error terms, $\epsilon(s_t, a_t)$: $\hat{\pi}(s_t) = \text{argmax}_{a_t \in \mathcal{A}_t} Q(s_t, a_t) + \epsilon(s_t, a_t)$. Only certain error distributions [37] admit closed-form solutions, which match the maximum causal entropy's prescribed distribution in the discrete choice setting [52]. Unfortunately, establishing appropriate error term distributions for the influences of latent variables in other decision settings is difficult. As we show in this work, the maximum causal entropy approach can be applied to decision estimation tasks and multi-agent strategic decision making without requiring explicit construction of latent variable influences.

B. Maximum Causal Entropy Inverse Optimal Control

We formulate the inverse optimal control problem as a maximum causal entropy estimation task. Despite the differences in formulation, a number of important connections to decision theory result. For control and decision-making domains, the predicted variables, $\mathbf{Y}_{1:T}$, correspond to an agent's sequence of employed actions, $\mathbf{A}_{1:T}$. The variables with known dynamics, $\mathbf{X}_{1:T}$, correspond to the agent's sequence of states, $\mathbf{S}_{1:T}$. We assume Markovian state dynamics, denoted $\mathcal{T}(\mathbf{S}_{1:T} || \mathbf{A}_{1:T-1}) \triangleq \prod_{t=1}^{T} \mathcal{T}(S_t | S_{t-1}, A_{t-1})$, that are either explicitly provided or estimated from data using a separate procedure. Since future states are only revealed *after* actions are selected, they should have no causal influence over earlier actions. This matches the causal assumptions of the maximum causal entropy model. We refer to the causally conditioned action distribution as a stochastic **policy**, $\pi(\mathbf{A}_{1:T} || \mathbf{S}_{1:T})$, with factors that are often Markovian, $\pi(A_t | S_t)$.

Definition 6: The **maximum causal entropy inverse optimal control** estimator is a special case of the general maximum causal entropy optimization (Equation 14) problem, $\text{argmax}_{\hat{\pi}(\mathbf{A}_{1:T} || \mathbf{S}_{1:T})} H_{\hat{\pi}}(\mathbf{A}_{1:T} || \mathbf{S}_{1:T})$, with linear equality constraints,

$$g_i(\hat{P}(\mathbf{S}_{1:T}, \mathbf{A}_{1:T})) = \mathbb{E}_{\hat{P}(\mathbf{S}_{1:T}, \mathbf{A}_{1:T})}[\mathcal{F}_i(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})] - \mathbb{E}_{\tilde{P}(\mathbf{S}_{1:T}, \mathbf{A}_{1:T})}[\mathcal{F}_i(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})]. \quad (23)$$

The maximum causal entropy distribution of Equation 18 simplifies greatly when feature functions linearly decompose over time steps, *i.e.*, $\mathcal{F}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = \sum_t \mathbf{f}(s_t, a_t)$.

C. Inference as Softened Optimal Control

Surprisingly, though formulated from information theory, the maximum causal entropy probability distribution is a generalization of optimal control laws governing decision theory. By replacing the log partition functions, $\log Z_{\boldsymbol{\theta}}(\mathbf{x}_{1:t}, \mathbf{y}_{1:t-1})$ and $\log Z_{\boldsymbol{\theta}}(y_t | \mathbf{x}_{1:t}, \mathbf{y}_{1:t-1})$ with analogs to state-action values, $Q^{\text{soft}}_{\boldsymbol{\theta}}(a_t, s_t)$, and state values, $V^{\text{soft}}_{\boldsymbol{\theta}}(s_t)$, which we will call value potentials, the connection to the Bellman equation [2] is established by Corollary 2.

Corollary 2 (of Theorem 3): The maximum causal entropy distribution constrained to match feature functions (Definition 6) that decompose linearly over time, *i.e.*, $\boldsymbol{\mathcal{F}}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = \sum_{t=1}^{T} \mathbf{f}(s_t, a_t)$, and with Markovian dynamics, can be re-expressed as:

$$Q^{\text{soft}}_{\boldsymbol{\theta},t}(a_t, s_t) = \mathbb{E}_{\mathcal{T}(S_{t+1} | s_t, a_t)}[V^{\text{soft}}_{\boldsymbol{\theta},t+1}(s_{t+1}) | s_t, a_t] + \boldsymbol{\theta}^{\mathrm{T}} \mathbf{f}(s_t, a_t) \quad (24)$$

$$V^{\text{soft}}_{\boldsymbol{\theta},t}(s_t) = \underset{a_t \in \mathcal{A}}{\text{softmax}}\, Q^{\text{soft}}_{\boldsymbol{\theta},t}(a_t, s_t), \quad (25)$$

where $\text{softmax}_{x \in \mathcal{X}} f(x) \triangleq \log \sum_{x \in \mathcal{X}} e^{f(x)}$ provides a smooth (*i.e.*, differentiable) interpolation of the maximum of different functions.

The gap between an action's value potential and the state's value potential, $Q^{\text{soft}}_{\boldsymbol{\theta},t}(s, a) - V^{\text{soft}}_{\boldsymbol{\theta},t}(s)$, determines that action's probability within the maximum causal entropy inverse optimal control model: $\hat{\pi}_{\boldsymbol{\theta}}(a | s) = e^{Q^{\text{soft}}_{\boldsymbol{\theta},t}(s,a) - V^{\text{soft}}_{\boldsymbol{\theta},t}(s)}$. When the gaps of multiple actions approach equality, the probabilities of those actions become uniform under the distribution. In the opposite limit, when the gap between one action and all others grows large, the softmax operation behaves like the maximum function and the stochastic maximum causal entropy policy approaches determinism, converging to the optimal policy of the Bellman equation [2], which only differs in the use of the max/softmax function in Equation 25.

V. Maximum Entropy Correlated Equilibria for Markov Games

The second setting we consider is the rational behavior of multiple players in sequential games. In this setting, the utilities governing players' decisions are known and obtaining equilibrium strategies for players is of interest.

A. Games and Equilibria

We consider sequential games with perfect information (*i.e.*, each player knows the complete state of the game). Markov games (*Definition* 7) formalize this setting, with each player choosing an action at each point in time based on the known state of the game, and players receiving some utility based on the combination of actions in each state. The canonical set of games studied within game theory—**one-shot** or **normal-form** games—are a special case of Markov game with only one time step of joint actions.

Definition 7: A **Markov game** is defined by a set of states ($\mathcal{S}$) representing the joint states of N agents (from set $\mathcal{N}$), a set of joint actions ($\boldsymbol{\mathcal{A}}_{1:N}$), a probabilistic state transition function, $\mathcal{T}(\mathbf{S}_{1:T} || \mathbf{A}_{1:T-1}) = \prod_{t=1}^{T} \mathcal{T}(S_t | S_{t-1}, A_{t-1})$ and a utility function, $\text{Utility}_i(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = \sum_{t=1}^{T} \mathbf{u}_i(s_t, a_t) \in \mathbb{R}$, specifying player i's utility for a sequence of states and actions.

Players choose **strategy profiles**, $\pi(\mathbf{A}_{1:T,1:N} || \mathbf{S}_{1:T})$, specifying (a distribution of) next actions for each situation. We consider the most general set of strategy profiles: **mixed** (*i.e.*, stochastic) and **correlated** (*i.e.*, joint functions in which

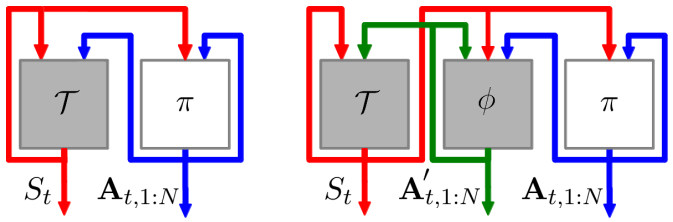

Fig. 3. The interacting strategy profile and game dynamics defining the (expected) utility (**left**); and the interaction between strategy profile, deviation policy, and game dynamics defining (expected) deviation utility (**right**) for Markov games.

players' actions can be interdependent) based on a cumulative expected utility. The rationality of a strategy profile is based on the relative utility each player obtains by deviating from the strategy profile in different ways. A **deviation policy**, $\phi(\mathbf{A'}_{1:T,1:N}||\mathbf{A}_{1:T,1:N}, \mathbf{S}_{1:T})$ specifies changes to the strategy profile. The utilities obtained under the strategy profile modified by the deviation policy are based on the interactions of causally conditioned probability distributions (as shown in Figure 3):

$$\text{ExpUtil}_{\pi,i} \triangleq \sum_{\substack{\mathbf{s}_{1:T} \in \boldsymbol{\mathcal{S}}_{1:T}, \\ \mathbf{a}_{1:T,1:N} \in \boldsymbol{\mathcal{A}}_{1:T,1:N}}} \pi(\mathbf{a}_{1:T,1:N}||\mathbf{s}_{1:T})\ \mathcal{T}(\mathbf{s}_{1:T}||\mathbf{a}_{1:T,1:N}) \times \text{Utility}_i(\mathbf{s}_{1:T}, \mathbf{a}_{1:T,1:N}) \tag{26}$$

$$\text{DevUtil}_{\pi,\phi,i} \triangleq \sum_{\substack{\mathbf{s}_{1:T} \in \boldsymbol{\mathcal{S}}_{1:T}, \\ \mathbf{a}_{1:T,1:N} \in \boldsymbol{\mathcal{A}}_{1:T,1:N}, \\ \mathbf{a}'_{1:T,1:N} \in \boldsymbol{\mathcal{A}}_{1:T,1:N},}} \pi(\mathbf{a}_{1:T,1:N}||\mathbf{s}_{1:T})\ \mathcal{T}(\mathbf{s}_{1:T}||\mathbf{a}'_{1:T,1:N}) \times \phi(\mathbf{a}'_{1:T,1:N}||\mathbf{a}_{1:T,1:N}, \mathbf{s}_{1:T})\ \text{Utility}_i(\mathbf{s}_{1:T}, \mathbf{a}'_{1:T,1:N}). \tag{27}$$

Often, the set of deviation policies corresponding to **switch functions**, Φ_{switch}, are considered, which allow one player to switch from a provided action, $a_{t,i}$, to an alternate action, $a_{t,i}{}'$, (the remaining action mapping does not switch actions, *i.e.*, $a'_{t,i} = a_{t,i}$).

Definition 8: A **correlated equilibrium** (CE) for a Markov game is a mixed joint strategy profile, $\pi^{CE}(\mathbf{A}_{1:T,1:N}||\mathbf{S}_{1:T})$, where no expected gain is obtained for any agent by employing a switch deviation policy. This is guaranteed with the following set of constraints:

$$\forall \phi_j \in \Phi_{\text{switch}},\ \text{Regret}_{\pi,\phi_j,i} \leq 0. \tag{28}$$

where $\text{Regret}_{\pi,\phi,i} \triangleq \text{DevUtil}_{\pi,\phi,i} - \text{ExpUtil}_{\pi,i}$ and ϕ_j is a switch for player i's action.

Correlated equilibria (*Definition* 8) generalize **Nash equilibria** [39], which further require agents' actions in each state to be independent, *i.e.*, $\pi(\mathbf{A}_{1:T,1:N}||\mathbf{S}_{1:T}) = \prod_{i=1}^{N} \pi(\mathbf{A}_{1:T,i}||\mathbf{S}_{1:T})$. Agents in a CE can coordinate their actions to obtain higher expected utilities. Conceptually, each agent is provided an action, $a_{t,i}$, and knows the conditional distribution of other agents' actions, $P(\mathbf{a}_{t,-i}|a_{t,i})$. To be in correlated equilibrium requires that no agent has an incentive to switch from action $a_{t,i}$ to a deviation action, $a'_{t,i}$, given that knowledge. Traffic lights are a canonical example of a **signaling device** designed to produce CE strategies. Given other agents' prescribed strategies (go on green), an agent will have incentive (equivalently, non-positive deviation regret) to obey a prescribed action (stop on red) rather than deviating (go on red). However, this coordination mechanism is not required as long as the players have access to a public communications channel [11]. Past research has shown that many decentralized, adaptive strategies will converge to some subset of strategies within the set of CE [42], [14], [18], [16], and not necessarily to the more restrictive Nash equilibrium.

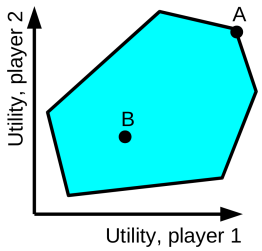

Fig. 4. A correlated equilibria polytope with: (A) an equilibrium maximizing social welfare, $\sum_{i \in \mathcal{N}} \text{Utility}_i(\mathbf{a}_{1:N})$, and (B) a maximum entropy correlated equilibrium.

The deviation regret constraints (Equation 28) define an N-dimensional convex polytope of CE solutions in the space of agents' joint utility payoffs (Figure 4). Exactly representing this polytope is generally intractable for Markov games, because the number of corners of the polytope grows exponentially with the game's time horizon. Efficient approximation approaches have been employed [38], [33], but tractable applicability has been limited to small games [33]. For the far more modest goal of finding an arbitrary CE in a range of compact games, algorithms that are polynomial in the number of agents have been developed [45], [24] and extended to sequential games [20].

The **maximum entropy correlated equilibria** (MaxEntCE) solution concept for normal-form games [44] selects the unique joint strategy profile that maximizes the joint entropy of players' actions subject to linear deviation regret inequality constraints (Equation 28). This approach provides the predictive guarantees of maximum entropy [17] in the single time step (normal-form) multi-agent game setting.

TABLE I
THE GAME OF CHICKEN AND FOUR STRATEGY PROFILES THAT ARE IN CORRELATED EQUILIBRIUM.

	Stay	Swerve
Stay	0,0	4,1
Swerve	1,4	3,3

CE 1

0	1
0	0

CE 2

0	0
1	0

CE 3

0	$\frac{1}{3}$
$\frac{1}{3}$	$\frac{1}{3}$

CE 4

$\frac{1}{4}$	$\frac{1}{4}$
$\frac{1}{4}$	$\frac{1}{4}$

Consider the game of Chicken (where each agent hopes the other will *Swerve*) and the correlated equilibria that define its utility polytope in TABLE I. *CE 4* is the maximum entropy correlated equilibrium and its predictive guarantee is apparent: all other CE have infinite log-loss when evaluated under the distribution of at least one other CE; the MaxEntCE is the only CE that assigns positive probability to the {Stay, Stay}

action combination. We extend these predictive guarantees to the Markov game setting in this work.

B. Maximum Causal Entropy Correlated Equilibria

We extend the maximum entropy correlated equilibrium approach to sequential games by posing it as a maximum causal entropy problem. The **causally conditioned entropy** measure (Equation 5) for this multi-player game setting is:

$$H(\mathbf{A}_{1:T,1:N}||\mathbf{S}_{1:T}) \triangleq \sum_{t=1}^{T} H(\mathbf{A}_{t,1:N}|\mathbf{A}_{1:t-1,1:N}, \mathbf{S}_{1:t}). \quad (29)$$

For the possible sequences of states and actions in a Markov game, it corresponds to the uncertainty associated with only the actions of the players in such sequences.

Definition 9: A **maximum causal entropy correlated equilibrium** (MCECE) solution maximizes the causal entropy (Equation 29), while satisfying the correlated equilibrium constraints (Equation 28) given the game dynamics $\mathcal{T}(S_{t+1}|S_t, A_{t,1:N})$. The regret constraints can be expressed in the maximum causal entropy framework as:

$$\begin{aligned}\mathcal{F}_{h,j}(\mathbf{a}_{1:T,1:N}, \mathbf{s}_{1:T}) &= -\text{Utility}_i(\mathbf{a}_{1:T,1:N}, \mathbf{s}_{1:T}) \quad (30)\\ &+ \frac{\mathcal{T}(\mathbf{s}_{1:T}||\mathbf{a}'_{1:T,1:N})\ \phi_j(\mathbf{a}'_{1:T,1:N}||\mathbf{a}_{1:T,1:N}, \mathbf{s}_{1:T})}{\mathcal{T}(\mathbf{s}_{1:T}||\mathbf{a}_{1:T,1:N})} \times \\ &\text{Utility}_i(\mathbf{a}'_{1:T,1:N}, \mathbf{s}_{1:T}) - \epsilon,\end{aligned}$$

for deviation policies, $\phi_j(\mathbf{a}'_{1:T,1:N}||\mathbf{a}_{1:T,1:N}, \mathbf{s}_{1:T})$, corresponding to each *switch* function for each player $i \in \mathcal{N}$.

By including a small amount of slack, $\epsilon \geq 0$, to provide primal feasibility, **sub-game equilibria** are realized by the Lagrangian dual solution, meaning that even in states where the probability of being reached converges towards 0, under the strategy profile and state dynamics, the strategy profile satisfies equilibria constraints (Equation 28) in all sub-games starting from those states.

Theorem 8: Subject to the feasibility requirements of Theorem 4, the MCECE strategy profile, $\pi_{\lambda}^{\text{MCECE}}(a_t|s_t)$, has the recursive form (with $\boldsymbol{\lambda} \geq \mathbf{0}$):

$$\pi_{\boldsymbol{\lambda}}(\mathbf{a}_{t,1:N}|s_t) \propto \exp\Bigg\{ H(\mathbf{a}_{t+1:T,1:N}||\mathbf{s}_{t+1:T}|\mathbf{a}_{t,1:N}, \mathbf{s}_t) - \sum_{i\in\mathcal{N}, a_{t,i}'\in\mathcal{A}_i} \lambda_{t,i,s_t,a_{t,i},a_{t,i}'} \text{Regret}_{\pi,i}(\mathbf{a}_{t,1:N}, s_t, a_{t,i}') \Bigg\}.$$

where $\text{Regret}_{\pi,i}(\mathbf{a}_{t,1:N}, s_t, a'_{t,i})$ denotes the regret of a particular switch function from $a_{t,i}$ to $a'_{t,i}$ given the other players' actions $\mathbf{a}_{t,-i}$.

Thus, by employing the principle of maximum causal entropy, we have expanded the maximum entropy correlated equilibria solution concept [44] to the Markov game setting.

VI. Conclusion

In this work, we introduced the principle of maximum causal entropy for estimating probability distributions where elements of interaction and feedback exist. We demonstrated its applicability and effectiveness on two very different statistical estimation tasks—discrete control and strategic Markovian games—to illustrate its generality.

Appendix A
Notational Conventions

We lowercase **values of variables** (*e.g.*, x_t, y_t), capitalize **random variables** (*e.g.*, X_t, Y_t), embolden sequence **multivariates** (*e.g.*, $\mathbf{x}_{1:T}$, $\mathbf{Y}_{1:t}$), and denote **sets** with calligraphy (*e.g.*, $\mathcal{X}_t$, $\boldsymbol{\mathcal{Y}}_{1:T}$), where temporal ranges, *e.g.*, $(1, 2, \ldots, T)$, are compactly represented as, *e.g.*, $1\!:\!T$. We generally employ T as the index of the last variable of the sequence (multivariate) and other time indexes, *e.g.*, t and τ, as indexes to other temporal positions in the sequence.

A **probability distribution** over random variables, *e.g.*, $P(\mathbf{Y}_{1:T})$, which is a member of the **probability simplex** $\Delta_{\boldsymbol{\mathcal{Y}}_{1:T}}$, implies the probability for each specific value, *e.g.*, $P(\mathbf{y}_{1:T}) \triangleq P(\mathbf{Y}_{1:T} = \mathbf{y}_{1:T})$. We denote estimated probability distributions as $\hat{P}(X_1)$ and sample probability distributions as $\tilde{P}(X_1)$. **Expectations** over random variables make the distribution of the random variables explicit, *e.g.*, $\mathbb{E}_{P(\mathbf{X}_{1:t})}[f(\mathbf{x}_{1:t})|x_1] = \sum_{\mathbf{x}_{2:t}\in\boldsymbol{\mathcal{X}}_{2:t}} P(\mathbf{x}_{2:t}|x_1) f(\mathbf{x}_{1:t})$.

When the distribution defining the **entropy** is unclear (*i.e.*, not P), we denote with subscript the defining distribution, *e.g.*, $H_{\hat{P}}(X) = \mathbb{E}_{\hat{P}(X)}[-\log \hat{P}(x)]$. An **entropy** can be conditioned on specific values, *e.g.*, $H(\mathbf{Y}_{2:T}|y_1) = \mathbb{E}_{P(\mathbf{Y}_{1:T})}[-\log P(\mathbf{Y}_{2:T})|y_1]$ or, in the causal entropy case, $H(\mathbf{Y}_{t+1:T}||\mathbf{X}_{t+1:T}|\mathbf{y}_{1:t}, \mathbf{x}_{1:t}) = \mathbb{E}_{P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})}[-\log P(\mathbf{Y}_{t+1:T}||\mathbf{X}_{t+1:T})|\mathbf{y}_{1:t}, \mathbf{x}_{1:t}]$.

Appendix B
Proofs of the Theorems

Proof of Theorem 1: Equation 10 is trivially implied by the non-negativity constraints on both causally conditioned probabilities (Equation 7).

To show that the second constraint (Equation 11) is implied, we must first introduce additional notation. We let $[\mathbf{x}_{1:\tau}; \mathbf{x}'_{\tau+1:T}]$ denote a partial replacement sequence for $\mathbf{x}_{1:T}$ in which $\mathbf{x}_{\tau+1:T}$ have been replaced with a different sequence of symbols, $\mathbf{x}'_{\tau+1:T}$. The proof procedure operates by "unzipping" the joint distribution:

$$\begin{aligned}&\sum_{\substack{\mathbf{x}_{1:T}\in\boldsymbol{\mathcal{X}}_{1:T}\\ \mathbf{y}_{1:T}\in\boldsymbol{\mathcal{Y}}_{1:T}}} P(\mathbf{y}_{1:T}||\mathbf{x}_{1:T})\ P(\mathbf{x}_{1:T}||\mathbf{y}_{1:T-1})\\ &\overset{(a)}{=} \sum_{\substack{\mathbf{x}_{1:T}\in\boldsymbol{\mathcal{X}}_{1:T}\\ \mathbf{y}_{1:T-1}\in\boldsymbol{\mathcal{Y}}_{1:T-1}}} \left(\sum_{y_T\in\mathcal{Y}_T} P(\mathbf{y}_{1:T}||\mathbf{x}_{1:T})\right)\ P(\mathbf{x}_{1:T}||\mathbf{y}_{1:T-1})\\ &\overset{(b)}{=} \sum_{\substack{\mathbf{x}_{1:T}\in\boldsymbol{\mathcal{X}}_{1:T}\\ \mathbf{y}_{1:T-1}\in\boldsymbol{\mathcal{Y}}_{1:T-1}}} \left(\sum_{y_T\in\mathcal{Y}_T} P(\mathbf{y}_{1:T}||[\mathbf{x}_{1:T-1}; x'_T])\right) P(\mathbf{x}_{1:T}||\mathbf{y}_{1:T-1})\\ &\overset{(c)}{=} \sum_{\substack{\mathbf{x}_{1:T-1}\in\boldsymbol{\mathcal{X}}_{1:T-1}\\ \mathbf{y}_{1:T-1}\in\boldsymbol{\mathcal{Y}}_{1:T-1}}} \left(\sum_{y_T\in\mathcal{Y}_T} P(\mathbf{y}_{1:T}||[\mathbf{x}_{1:T-1}; x'_T])\right)\\ &\qquad \left(\sum_{x_T\in\mathcal{X}_T} P(\mathbf{x}_{1:T}||\mathbf{y}_{1:T-1})\right)\end{aligned}$$

$$\overset{(d)}{=} \sum_{\substack{\mathbf{x}_{1:T-1} \in \boldsymbol{\mathcal{X}}_{1:T-1} \\ \mathbf{y}_{1:T-1} \in \boldsymbol{\mathcal{Y}}_{1:T-1}}} \left(\sum_{y_T \in \mathcal{Y}_T} P(\mathbf{y}_{1:T} || [\mathbf{x}_{1:T-1}; x'_T]) \right) \times \left(\sum_{x_T \in \mathcal{X}_T} P(\mathbf{x}_{1:T} || [\mathbf{y}_{1:T-2}; y'_{T-1}]) \right)$$

$$\overset{(e)}{=} \sum_{\substack{\mathbf{x}_{1:\tau} \in \boldsymbol{\mathcal{X}}_{1:\tau} \\ \mathbf{y}_{1:\tau} \in \boldsymbol{\mathcal{Y}}_{1:\tau}}} \left(\sum_{\mathbf{y}_{\tau+1:T} \in \boldsymbol{\mathcal{Y}}_{\tau+1:T}} P(\mathbf{y}_{1:T} || [\mathbf{x}_{1:\tau}; \mathbf{x}'_{\tau+1:T}]) \right) \times \left(\sum_{\mathbf{x}_{\tau+1:T} \in \boldsymbol{\mathcal{X}}_{\tau+1:T}} P(\mathbf{x}_{1:T} || [\mathbf{y}_{1:\tau}; \mathbf{y}'_{\tau+1:T-1}]) \right)$$

$$\overset{(f)}{=} \left(\sum_{\mathbf{y}_{1:T} \in \boldsymbol{\mathcal{Y}}_{1:T}} P(\mathbf{y}_{1:T} || [\mathbf{x}'_{1:T}]) \right) \left(\sum_{\mathbf{x}_{1:T} \in \boldsymbol{\mathcal{X}}_{1:T}} P(\mathbf{x}_{1:T} || [\mathbf{y}'_{1:T-1}]) \right)$$

$$\overset{(g)}{=} 1.$$

The variable(s) that appear in only one of the causally conditioned distributions are separately marginalized over (a) and that independent marginal replaced with a replacement sequence via the property of Equation 9 (b). Due to this replacement, one additional variable then only appears in the other causally conditioned probability distribution and the procedure can alternate with separate marginalization (c) and replacement (d). This operation can be repeated to the τ^{th} elements of the sequences (as shown) (e) and then for the entire sequence (f). Lastly, using the normalization property of the two causally conditioned probability distributions (Equation 8) concludes the proof (g). ∎

Proof of Corollary 1: Using the definition of the conditional probability in terms of the marginalized joint probability (Theorem 1), and following the same partial replacement notation and "unzipping" procedure of the proof of Theorem 1, we have:

$$P(y_\tau | \mathbf{y}_{1:\tau-1}, \mathbf{x}_{1:\tau}) = \frac{\sum_{\substack{\mathbf{y}_{\tau+1:T} \in \boldsymbol{\mathcal{Y}}_{\tau+1:T}, \\ \mathbf{x}_{\tau+1:T} \in \boldsymbol{\mathcal{X}}_{\tau+1:T}}} P(\mathbf{y}_{1:T}, \mathbf{x}_{1:T})}{\sum_{\substack{\mathbf{y}_{\tau:T} \in \boldsymbol{\mathcal{Y}}_{\tau:T}, \\ \mathbf{x}_{\tau+1:T} \in \boldsymbol{\mathcal{X}}_{\tau+1:T}}} P(\mathbf{y}_{1:T}, \mathbf{x}_{1:T})}$$

$$= \frac{\sum_{\substack{\mathbf{y}_{\tau+1:T} \in \boldsymbol{\mathcal{Y}}_{\tau+1:T}, \\ \mathbf{x}_{\tau+1:T} \in \boldsymbol{\mathcal{X}}_{\tau+1:T}}} P(\mathbf{y}_{1:T} || \mathbf{x}_{1:T}) P(\mathbf{x}_{1:T} || \mathbf{y}_{1:T-1})}{\sum_{\substack{\mathbf{y}_{\tau:T} \in \boldsymbol{\mathcal{Y}}_{\tau:T}, \\ \mathbf{x}_{\tau+1:T} \in \boldsymbol{\mathcal{X}}_{\tau+1:T}}} P(\mathbf{y}_{1:T} || \mathbf{x}_{1:T}) P(\mathbf{x}_{1:T} || \mathbf{y}_{1:T-1})}$$

$$= \frac{\left(\sum_{\mathbf{y}_{\tau+1:T} \in \boldsymbol{\mathcal{Y}}_{\tau+1:T}} P(\mathbf{y}_{1:T} || [\mathbf{x}_{1:\tau}; \mathbf{x}'_{\tau+1:T}]) \right)}{\left(\sum_{\mathbf{y}_{\tau:T} \in \boldsymbol{\mathcal{Y}}_{\tau:T}} P(\mathbf{y}_{1:T} || [\mathbf{x}_{1:\tau}; \mathbf{x}'_{\tau+1:T}]) \right)}.$$

$$\text{Thus, } \prod_{\tau=1}^{T} P(y_\tau | \mathbf{y}_{1:\tau-1}, \mathbf{x}_{1:\tau}) = \prod_{\tau=1}^{T} \frac{\left(\sum_{\mathbf{y}_{\tau+1:T} \in \boldsymbol{\mathcal{Y}}_{\tau+1:T}} P(\mathbf{y}_{1:T} || [\mathbf{x}_{1:\tau}; \mathbf{x}'_{\tau+1:T}]) \right)}{\left(\sum_{\mathbf{y}_{\tau:T} \in \boldsymbol{\mathcal{Y}}_{\tau:T}} P(\mathbf{y}_{1:T} || [\mathbf{x}_{1:\tau}; \mathbf{x}'_{\tau+1:T}]) \right)} = \frac{P(\mathbf{y}_{1:T} || \mathbf{x}_{1:T})}{\left(\sum_{\mathbf{y}_{1:T} \in \boldsymbol{\mathcal{Y}}_{1:T}} P(\mathbf{y}_{1:T} || [\mathbf{x}'_{1:T}]) \right)} = P(\mathbf{y}_{1:T} || \mathbf{x}_{1:T}).$$

Similarly, $P(\mathbf{x}_{1:T} || \mathbf{y}_{1:T-1}) = \prod_{t=1}^{T} P(x_t | \mathbf{x}_{1:t-1}, \mathbf{y}_{1:t-1})$ following an analogous argument for the causally conditioned $\mathbf{x}_{1:T}$ variables.

To complete the proof, we must show that the *product of conditional probabilities* definition (Equation 2) of the causally conditioned probability satisfies the causal polytope definition (Definition 3). The non-negativity constraint (Equation 7) is satisfied by the non-negativity of conditional probability distributions. The remaining two constraints are satisfied as a consequence of noting that:

$$\sum_{\mathbf{y}_{\tau:T} \in \boldsymbol{\mathcal{Y}}_{\tau:T}} \prod_{t=\tau}^{T} P(y_t | \mathbf{y}_{1:t-1}, \mathbf{x}_{1:t}) = \underbrace{\sum_{y_\tau \in \mathcal{Y}_\tau} P(y_\tau | \mathbf{y}_{1:\tau-1}, \mathbf{x}_{1:\tau}) \cdots \sum_{y_T \in \mathcal{Y}_T} P(y_T | \mathbf{y}_{1:T-1}, \mathbf{x}_{1:T})}_{T-\tau+1 \text{ summations}} = 1,$$

because each conditional probability distribution (starting from the right-most) normalizes to 1. Thus, Equation 8 is satisfied for $\tau = 1$ and Equation 9 is satisfied as

$$\sum_{\mathbf{y}_{\tau+1:T} \in \boldsymbol{\mathcal{Y}}_{\tau+1:T}} P(\mathbf{y}_{1:T} || \mathbf{x}_{1:T}) - P(\mathbf{y}_{1:T} || \mathbf{x}'_{1:T})$$

$$= \left(\prod_{t=1}^{\tau} P(y_t | \mathbf{y}_{1:t-1}, \mathbf{x}_{1:t}) \right) \left(\sum_{\substack{\mathbf{y}_{\tau+1:T} \\ \in \boldsymbol{\mathcal{Y}}_{\tau+1:T}}} \prod_{t=\tau+1}^{T} P(y_t | \mathbf{y}_{1:t-1}, \mathbf{x}_{1:t}) - \sum_{\mathbf{y}_{\tau+1:T} \in \boldsymbol{\mathcal{Y}}_{\tau+1:T}} \prod_{t=\tau+1}^{T} P(y_t | \mathbf{y}_{1:t-1}, \mathbf{x}'_{1:t}) \right) = 0,$$

completing the proof. ∎

Proof of Theorem 2: The negative causally conditioned entropy, $-H_{\hat{P}}(\mathbf{Y_{1:T}} || \mathbf{X_{1:T}})$, is a conic combination of $-\hat{P}(\mathbf{y}_{1:T} || \mathbf{x}_{1:T}) \log \hat{P}(\mathbf{y}_{1:T} || \mathbf{x}_{1:T})$ terms, which are each convex for $\hat{P}(\mathbf{y}_{1:T} || \mathbf{x}_{1:T}) \geq 0$. The optimization constraints based on the joint probability terms are all affine in the unknown causally conditioned probability terms. The intersection with the causal polytope (Definition 3) is also convex. Thus, the overall optimization is a convex optimization problem. ∎

We now prove two lemmas that are needed for the proof of Theorem 3.

Lemma 1: The Lagrangian dual optimization problem's solution is the probability distribution recursively defined according to Equation 18.

Proof: We begin by obtaining the form of the probability distribution in the Lagrangian dual optimization problem. Note that since the domain of the objective (the causal entropy) is only on the non-negative causally conditioned probability terms, $\hat{P}(\mathbf{y}_{1:T} || \mathbf{x}_{1:T})$, thus the non-negativity constraints from the causal polytope are superfluous, and we will suppress them. Differentiating the Lagrangian of the maximum causal entropy optimization (Equation 14), where the causal probability constraints are replaced with the locally normalizing

constraints (which are equivalent by Corollary 1),

$$\begin{aligned}\Lambda(\hat{P},\boldsymbol{\lambda},\boldsymbol{\gamma}) &= H_{\hat{P}}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) \\ &+ \lambda^{\mathrm{T}}\mathbf{g}(\hat{P}(\mathbf{X}_{1:T},\mathbf{Y}_{1:T})) + \gamma^{\mathrm{T}}\mathbf{h}(\hat{P}(\mathbf{X}_{1:T},\mathbf{Y}_{1:T})) \\ &- \sum_{\substack{t=1:T,\\ \mathbf{x}_{1:t}\in\boldsymbol{\mathcal{X}}_{1:t},\\ \mathbf{y}_{1:t-1}\in\boldsymbol{\mathcal{Y}}_{1:t-1}}} C(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})\left(1-\sum_{y_t\in\mathcal{Y}}\hat{P}(y_t|\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})\right),\end{aligned} \tag{31}$$

we have (with $\boldsymbol{\gamma}\geq\mathbf{0}$):

$$\begin{aligned}\nabla_{\{\hat{P}(y_t|\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})\}}\Lambda(\hat{P},\boldsymbol{\lambda},\boldsymbol{\gamma}) &= C(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1}) \\ &+ \hat{P}(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})\Big(-\log\hat{P}(y_t|\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})-1 \\ &+ H_{\hat{P}}(\mathbf{Y}_{t+1:T}||\mathbf{X}_{t+1:T}|\mathbf{x}_{1:t},\mathbf{y}_{1:t}) \\ &+ \boldsymbol{\lambda}^{\mathrm{T}}\mathbb{E}_{\hat{P}(\mathbf{X}_{1:T},\mathbf{Y}_{1:T})}[\boldsymbol{\mathcal{F}}_g(\mathbf{x}_{1:T},\mathbf{y}_{1:T})|\mathbf{x}_{1:t},\mathbf{y}_{1:t}] \\ &+ \boldsymbol{\gamma}^{\mathrm{T}}\mathbb{E}_{\hat{P}(\mathbf{X}_{1:T},\mathbf{Y}_{1:T})}[\boldsymbol{\mathcal{F}}_h(\mathbf{x}_{1:T},\mathbf{y}_{1:T})|\mathbf{x}_{1:t},\mathbf{y}_{1:t}]\Big).\end{aligned} \tag{32}$$

Equating the gradient to $\mathbf{0}$ and solving for $\hat{P}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(y_t|\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})$ yields:

$$\begin{aligned}\hat{P}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(y_t|\mathbf{x}_{1:t},\mathbf{y}_{1:t-1}) \propto \exp\Big\{ & H_{\hat{P}}(\mathbf{Y}_{t+1:T}||\mathbf{X}_{t+1:T}|\mathbf{x}_{1:t},\mathbf{y}_{1:t}) \\ & + \boldsymbol{\lambda}^{\mathrm{T}}\mathbb{E}_{\hat{P}(\mathbf{X}_{1:T},\mathbf{Y}_{1:T})}[\boldsymbol{\mathcal{F}}_g(\mathbf{x}_{1:T},\mathbf{y}_{1:T})|\mathbf{x}_{1:t},\mathbf{y}_{1:t}] \\ & + \boldsymbol{\gamma}^{\mathrm{T}}\mathbb{E}_{\hat{P}(\mathbf{X}_{1:T},\mathbf{Y}_{1:T})}[\boldsymbol{\mathcal{F}}_h(\mathbf{x}_{1:T},\mathbf{y}_{1:T})|\mathbf{x}_{1:t},\mathbf{y}_{1:t}]\Big\}.\end{aligned}$$

Starting with the recursive relationship constraining the causally conditioned probability distribution (Equation 32), we go further to prove the operational recurrence of the theorem (Equation 18). We begin by factoring out the $\hat{P}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})$ multiplier. We prove the lemma by substituting our recursive definitions (Equation 17 and Equation 18) to show that they are solutions to the recurrence.

$$\begin{aligned}&\frac{C(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})}{\hat{P}(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})}-1 \\ &-\left(\sum_{\tau=t}^{T}\mathbb{E}_{\hat{P}(\mathbf{X}_{1:T},\mathbf{Y}_{1:T})}\Big[\log\hat{P}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(y_\tau|\mathbf{x}_{1:\tau},\mathbf{y}_{1:\tau-1})\Big|\mathbf{x}_{1:t},\mathbf{y}_{1:t}\Big]\right) \\ &+\mathbb{E}_{\hat{P}(\mathbf{X}_{1:T},\mathbf{Y}_{1:T})}\Big[\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\mathcal{F}}_g(\mathbf{x}_{1:T},\mathbf{y}_{1:T})+\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\mathcal{F}}_h(\mathbf{x}_{1:T},\mathbf{y}_{1:T})\Big|\mathbf{x}_{1:t},\mathbf{y}_{1:t}\Big] \\ =&\frac{C(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})}{\hat{P}(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})}-1-\sum_{\tau=t}^{T-1}\mathbb{E}_{\hat{P}(\mathbf{X}_{1:T},\mathbf{Y}_{1:T})}\Big[\\ &\sum_{x_{\tau+1}\in\mathcal{X}}P(x_{\tau+1}|\mathbf{x}_{1:\tau},\mathbf{y}_{1:\tau})\log Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{x}_{1:\tau+1},\mathbf{y}_{1:\tau}) \\ &-\log Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{x}_{1:\tau},\mathbf{y}_{1:\tau-1})\Big|\mathbf{x}_{1:t},\mathbf{y}_{1:t}\Big] \\ &-\mathbb{E}_{\hat{P}(\mathbf{X}_{1:T},\mathbf{Y}_{1:T})}[\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\mathcal{F}}_g(\mathbf{x}_{1:T},\mathbf{y}_{1:T})+\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\mathcal{F}}_h(\mathbf{x}_{1:T},\mathbf{y}_{1:T}) \\ &-\log Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{x}_{1:T},\mathbf{y}_{1:T-1})|\mathbf{x}_{1:t},\mathbf{y}_{1:t}] \\ &+\mathbb{E}_{\hat{P}(\mathbf{X}_{1:T},\mathbf{Y}_{1:T})}\Big[\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\mathcal{F}}_g(\mathbf{x}_{1:T},\mathbf{y}_{1:T}) \\ &+\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\mathcal{F}}_h(\mathbf{x}_{1:T},\mathbf{y}_{1:T})\Big|\mathbf{x}_{1:t},\mathbf{y}_{1:t}\Big]\end{aligned} \tag{33}$$

$$\begin{aligned}=&\frac{C(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})}{\hat{P}(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})}-1 \\ &-\sum_{\tau=t}^{T}\mathbb{E}_{\hat{P}(\mathbf{X}_{1:T},\mathbf{Y}_{1:T})}\Big[\log Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{x}_{1:\tau+1},\mathbf{y}_{1:\tau})- \\ &\quad\log Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{x}_{1:\tau},\mathbf{y}_{1:\tau-1})\Big|\mathbf{x}_{1:t},\mathbf{y}_{1:t}\Big] \\ &-\mathbb{E}_{\hat{P}(\mathbf{X}_{1:T},\mathbf{Y}_{1:T})}[\log Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{x}_{1:T},\mathbf{y}_{1:T-1}|\mathbf{x}_{1:t},\mathbf{y}_{1:t})] \\ =&\frac{C(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})}{\hat{P}(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})}-1-\log Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})=0.\end{aligned}$$

Thus, setting $C(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1}) = \hat{P}(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1}) + \log Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})\hat{P}(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})$, which is only a function of $\mathbf{x}_{1:t}$ and $\mathbf{y}_{1:t-1}$ (and, importantly, not y_t), proves the distribution form. ■

Lemma 2: Under the Lagrangian dual's form of the probability distribution (Lemma 1), $\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$, and another distribution $P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T}) = P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})P(\mathbf{X}_{1:T}||\mathbf{Y}_{1:T-1})$, the conditioned causal log-loss (Definition 2) has the following relationship:

$$\begin{aligned}&E_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[-\log\hat{P}(\mathbf{Y}_{\tau+1:T}||\mathbf{X}_{\tau+1:T}|\mathbf{y}_{1:\tau},\mathbf{x}_{1:\tau})] = \\ &\quad\sum_{x_{\tau+1}\in\mathcal{X}_{\tau+1}}P(x_{\tau+1}|\mathbf{x}_{1:\tau},\mathbf{y}_{1:\tau})\log Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{x}_{1:\tau+1},\mathbf{y}_{1:\tau}) \\ &\quad-\mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\mathcal{F}}_g(\mathbf{x}_{1:T},\mathbf{y}_{1:T}) \\ &\qquad+\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\mathcal{F}}_h(\mathbf{x}_{1:T},\mathbf{y}_{1:T})|\mathbf{x}_{1:\tau},\mathbf{y}_{1:\tau}].\end{aligned}$$

Proof: Using the recursive form under the dual (Equation 17 and Equation 18) obtained in Lemma 1, we have:

$$\begin{aligned}&\mathbb{E}_{P(\mathbf{X}_{1:T},\mathbf{Y}_{1:T})}\left[\sum_{t=\tau+1}^{T}-\log\hat{P}(y_t|\mathbf{x}_{1:t},\mathbf{y}_{t-1})\Big|\mathbf{x}_{1:\tau},\mathbf{y}_{1:\tau}\right] \\ =&\,\mathbb{E}_{P(\mathbf{X}_{1:T},\mathbf{Y}_{1:T})}\Bigg[-\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\mathcal{F}}_g(\mathbf{x}_{1:T},\mathbf{y}_{1:T})-\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\mathcal{F}}_h(\mathbf{x}_{1:T},\mathbf{y}_{1:T}) \\ &-\sum_{t=\tau+1}^{T-1}\sum_{x_{t+1}\in\mathcal{X}_t}P(x_{t+1}|\mathbf{x}_{1:t},\mathbf{y}_{1:t})\log Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{x}_{1:t+1},\mathbf{y}_{1:t}) \\ &+\sum_{t=\tau+1}^{T}\log Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{x}_{1:t},\mathbf{y}_{1:t-1})\Big|\mathbf{x}_{1:\tau},\mathbf{y}_{1:\tau}\Bigg] \\ =&-\mathbb{E}_{P(\mathbf{X}_{1:T},\mathbf{Y}_{1:T})}\Bigg[\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\mathcal{F}}_g(\mathbf{x}_{1:T},\mathbf{y}_{1:T}) \\ &+\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\mathcal{F}}_h(\mathbf{x}_{1:T},\mathbf{y}_{1:T})\Big|\mathbf{x}_{1:\tau},\mathbf{y}_{1:\tau}\Bigg] \\ &+\sum_{x_{\tau+1}\in\mathcal{X}_{\tau+1}}P(x_{\tau+1}|\mathbf{x}_{1:\tau},\mathbf{y}_{1:\tau})\log Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{x}_{1:\tau+1},\mathbf{y}_{1:\tau}),\end{aligned}$$

which proves the lemma. ■

Proof of Theorem 3: Plugging the dual optimization problem's optimal solution form (Equation 18) into the La-

grangian (Equation 31), we have:

$$\begin{aligned}
&\inf_{\boldsymbol{\lambda},\boldsymbol{\gamma}\geq\mathbf{0}} \sup_{\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})} \Lambda(\hat{P},\boldsymbol{\lambda},\boldsymbol{\gamma}) \\
&= \inf_{\boldsymbol{\lambda},\boldsymbol{\gamma}\geq\mathbf{0}} \Lambda(\hat{P}_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}),\boldsymbol{\lambda},\boldsymbol{\gamma}) \\
&= \inf_{\boldsymbol{\lambda},\boldsymbol{\gamma}\geq\mathbf{0}} H_{\hat{P}}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) + \mathbb{E}_{\hat{P}}\Big[\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}(\hat{P}(\mathbf{X}_{1:T},\mathbf{Y}_{1:T})) \\
&\qquad + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{h}(\hat{P}(\mathbf{X}_{1:T},\mathbf{Y}_{1:T}))\Big].
\end{aligned}$$

Substituting in the result of Lemma 2 for $H_{\hat{P}}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$ under the special case that $P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) = \hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$ proves the dual optimization form of the theorem. The form of the distribution is provided by Lemma 1. ∎

Proof of Theorem 4: The primal optimization problem is convex (Theorem 2), thus by *Slater's condition* for affine inequality constraints [6], as long as there is a feasible solution satisfying the constraint set in the primal optimization problem on the relative interior, then strong Lagrangian duality holds—there is no duality gap between the primal optimization problem and the dual optimization problem:

$$\begin{aligned}
&\sup_{\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})} \inf_{\boldsymbol{\lambda},\boldsymbol{\gamma}\geq\mathbf{0}} \Lambda(\hat{P},\boldsymbol{\lambda},\boldsymbol{\gamma}) \\
&= \inf_{\boldsymbol{\lambda},\boldsymbol{\gamma}\geq\mathbf{0}} \sup_{\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})} \Lambda(\hat{P},\boldsymbol{\lambda},\boldsymbol{\gamma}).
\end{aligned}$$

For our problem, this requires:

$$\begin{aligned}
&\exists P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) \in \Xi \text{ such that:} \\
&\forall \mathbf{y}_{1:T} \in \boldsymbol{\mathcal{Y}}_{1:T}, \forall \mathbf{x}_{1:T} \in \boldsymbol{\mathcal{X}}_{1:T},\ P(\mathbf{y}_{1:T}||\mathbf{x}_{1:T}) > 0, \qquad (34) \\
&\mathbf{g}(P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})) = \mathbf{0},\ \text{and}\ \ \mathbf{h}(P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})) \geq \mathbf{0}.
\end{aligned}$$

Note that when the only primal feasible solution violates strict positivity (Equation 34), non-finite dual parameters would be required. This can be alleviated by allowing small slack in the equality and inequality constraints,

$$\begin{aligned}
&\forall i \in \{1,\ldots,M\},\ |\mathbf{g}_i(P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T}))| \leq \epsilon \\
&\forall j \in \{1,\ldots,N\},\ \mathbf{h}_j(P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})) \geq -\epsilon,
\end{aligned}$$

leading to Lagrangian multiplier regularization in the dual optimization problem [12]. We also discuss adding slack in Section III-E to deal with finite sample approximations. ∎

Proof of Theorem 5: Writing the Lagrangian dual (Equation 17) for these constraints and then relying on Lemma 2, we have:

$$\begin{aligned}
&\min_{\boldsymbol{\lambda}} \sum_{x_1\in\mathcal{X}_1} P(x_1)\log Z_{\boldsymbol{\lambda}}(x_1) \\
&\qquad - \boldsymbol{\lambda}^{\mathrm{T}}\mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[\boldsymbol{\mathcal{F}}_g(\mathbf{y}_{1:T},\mathbf{x}_{1:T})] \\
&= \min_{\boldsymbol{\lambda}}\ -\mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[\log \hat{P}_{\boldsymbol{\lambda}}(\mathbf{y}_{1:T}||\mathbf{x}_{1:T})] \\
&= \max_{\boldsymbol{\lambda}}\ \mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[\log \hat{P}_{\boldsymbol{\lambda}}(\mathbf{y}_{1:T}||\mathbf{x}_{1:T})],
\end{aligned}$$

completing the proof. ∎

Proof of Theorem 6: Letting each sample's k^{th} moment statistic be bounded within $[f_{g,k}^{\min}, f_{g,k}^{\max}]$, by Hoeffding's inequality, we have:

$$\begin{aligned}
&P\left(\left|\bar{f}_{g,k} - \mathbb{E}_{\tilde{P}(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}\left[\boldsymbol{\mathcal{F}}_{g,k}(\mathbf{y}_{1:T},\mathbf{x}_{1:T})\right]\right| \geq \epsilon\right) \\
&\leq 2\exp\left(-\frac{2n\epsilon^2}{(f_{g,k}^{\max} - f_{g,k}^{\min})^2}\right),
\end{aligned}$$

By the union bound:

$$\begin{aligned}
&P\left(\bigcup_{k=1}^{K}\left|\bar{f}_{g,k} - \mathbb{E}_{\tilde{P}(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}\left[\boldsymbol{\mathcal{F}}_{g,k}(\mathbf{y}_{1:T},\mathbf{x}_{1:T})\right]\right| \geq \epsilon\right) \\
&\leq \sum_{k=1}^{K} P\left(\left|\bar{f}_{g,k} - \mathbb{E}_{\tilde{P}(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}\left[\boldsymbol{\mathcal{F}}_{g,k}(\mathbf{y}_{1:T},\mathbf{x}_{1:T})\right]\right| \geq \epsilon\right).
\end{aligned}$$

Combining these, and recognizing that:

$$\begin{aligned}
&P\left(\bigcup_{k=1}^{K}\left|\bar{f}_{g,k} - \mathbb{E}_{\tilde{P}(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}\left[\boldsymbol{\mathcal{F}}_{g,k}(\mathbf{y}_{1:T},\mathbf{x}_{1:T})\right]\right| \geq \epsilon\right) = \\
&P\left(\left|\left|\bar{f}_{g,k} - \mathbb{E}_{\tilde{P}(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}\left[\boldsymbol{\mathcal{F}}_{g,k}(\mathbf{y}_{1:T},\mathbf{x}_{1:T})\right]\right|\right|_{\infty} \geq \epsilon\right),
\end{aligned}$$

while letting $\mathbf{f}_{g,\max} = \max_k f_{g,k}^{\max}$ and $\mathbf{f}_{g,\min} = \min_k f_{g,k}^{\min}$, proves the theorem. ∎

We now prove an important saddle point existence lemma needed for Theorem 7.

Lemma 3: Under the restriction that $P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$ is from the set $\Gamma \subseteq \Xi$ of causally conditioned probability distributions satisfying provided equality and inequality constraints, $\mathbf{g}\left(P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})\right) = \mathbf{0}$ and $\mathbf{h}\left(P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})\right) \geq \mathbf{0}$ (Equation 14), and assuming strong Lagrangian duality holds (Theorem 4), the causal log-loss (Definition 2),

$$\begin{aligned}
&CLL(\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}), P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})) \\
&= \mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[-\log \hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})] \\
&= -\sum P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) P(\mathbf{X}_{1:T}||\mathbf{Y}_{1:T-1}) \log \hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}),
\end{aligned}$$

has a saddle point, $P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) = \hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) = P^*(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$: the maximum causal entropy distribution (Definition 4). In other words,

$$\begin{aligned}
&\sup_{P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})\in\Gamma} CLL(P^*(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}), P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})) \\
&\overset{(a)}{=} CLL(P^*(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}), P^*(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})) \\
&\overset{(b)}{=} \inf_{\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})\in\Xi} CLL(\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}), P^*(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}))
\end{aligned}$$

Proof: Equality (a): For any $P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) \in \Gamma$, as a

special case of Lemma 2,

$$\begin{aligned}
&CLL(P^*(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}), P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})) \\
&= \sum_{x_1 \in \mathcal{X}_1} P(x_1) \log Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(x_1) \\
&\qquad - \mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[\boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{\mathcal{F}}_g(\mathbf{x}_{1:T}, \mathbf{y}_{1:T})] \\
&\qquad - \mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{\mathcal{F}}_h(\mathbf{x}_{1:T}, \mathbf{y}_{1:T})] \\
&\leq \sum_{x_1 \in \mathcal{X}_1} P(x_1) \log Z_{\boldsymbol{\lambda},\boldsymbol{\gamma}}(x_1) + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{c}_g + \boldsymbol{\gamma}^{\mathrm{T}} \mathbf{c}_h \qquad (35) \\
&= CLL(P^*(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}), P^*(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})).
\end{aligned}$$

The inequality follows from the constraints on the set Γ: any $P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) \in \Gamma$ satisfies (with equality) $\mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[\boldsymbol{\mathcal{F}}_g(\mathbf{x}_{1:T}, \mathbf{y}_{1:T})] = -\mathbf{c}_g$ and (with inequality) $\mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[\boldsymbol{\mathcal{F}}_h(\mathbf{x}_{1:T}, \mathbf{y}_{1:T})] \geq -\mathbf{c}_h$. Note that Equation 35 is the dual optimization objective (Equation 17) and it reaches its optima at $P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) = P^*(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$.

Equality (b): For any $\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$,

$$\begin{aligned}
&CLL(\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}), P^*(Y_{1:T}||\mathbf{X}_{1:T}) \\
&= \mathbb{E}_{P^*(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[-\log \hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})] \\
&\geq \mathbb{E}_{P^*(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[-\log P^*(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})] \\
&= CLL(P^*(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}), P^*(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})).
\end{aligned}$$

The inequality follows from an "information bound" on the causal Kullback-Leibler divergence (Equation 22): $\mathbb{E}_{P^*(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[\log \frac{P^*(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})}{\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})}] \geq 0$, which is tight when $\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) = P^*(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$. ∎

Proof of Theorem 7: In what follows, we let Ξ denote the causal polytope defining causally conditioned probability distributions and let Γ denote the subset of Ξ that satisfies: $\mathbf{g}(P(\mathbf{Y}_{1:T}, \mathbf{X}_{1:T})) = 0$ and $\mathbf{h}(P(\mathbf{Y}_{1:T}, \mathbf{X}_{1:T})) \geq 0$.

$$\begin{aligned}
&\inf_{\substack{\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) \\ \in \Xi}} \sup_{\substack{P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) \\ \in \Gamma}} \mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[-\log \hat{P}(\mathbf{y}_{1:T}||\mathbf{x}_{1:T})] \\
&= \sup_{\substack{P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) \\ \in \Gamma}} \inf_{\substack{\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) \\ \in \Xi}} \mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[-\log \hat{P}(\mathbf{y}_{1:T}||\mathbf{x}_{1:T})] \\
&= \sup_{P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) \in \Gamma} \mathbb{E}_{P(\mathbf{Y}_{1:T},\mathbf{X}_{1:T})}[-\log P(\mathbf{y}_{1:T}||\mathbf{x}_{1:T})] \\
&= \sup_{P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) \in \Gamma} H(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}),
\end{aligned}$$

The first equality (minimax) follows from the existence of the saddle point established in Lemma 3. The second follows from the fact that setting the estimate to the adversarially chosen distribution $\hat{P}(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T}) = P(\mathbf{Y}_{1:T}||\mathbf{X}_{1:T})$ is then optimal for the minimization. Finally, the result follows from the definition of causal entropy (Definition 1). ∎

Proof of Corollary 2: Following the proof of Lemma 1, we substitute the softened maximum causal entropy recurrence (Equation 24 and Equation 25) into Equation 33 to verify it is a solution to the Lagrangian dual optimization problem.

$$\begin{aligned}
&-\sum_{\tau=t}^{T-1} \mathbb{E}_{\hat{P}(\mathbf{S}_{1:T},\mathbf{A}_{1:T})}\Big[\sum_{s_{\tau+1} \in \mathcal{S}} P(s_{\tau+1}|s_\tau, a_\tau) V_{\boldsymbol{\theta}}(s_{\tau+1}) \\
&\qquad + \boldsymbol{\theta}^{\mathrm{T}} \mathbf{f}(s_\tau, a_\tau) - V_{\boldsymbol{\theta}}(s_\tau) \Big| s_t, a_t \Big] \\
&\qquad - \mathbb{E}_{\hat{P}(\mathbf{S}_{1:T},\mathbf{A}_{1:T})}[\boldsymbol{\theta}^{\mathrm{T}} \mathbf{f}(s_T, a_T) - V_{\boldsymbol{\theta}}(s_T) | \mathbf{s}_{1:t}, \mathbf{a}_{1:t}] \\
&\qquad + \mathbb{E}_{\hat{P}(\mathbf{S}_{1:T},\mathbf{A}_{1:T})}\Big[\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{\mathcal{F}}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) \Big| \mathbf{s}_{1:t}, \mathbf{a}_{1:t}\Big] \\
&= \mathbb{E}_{\hat{P}(\mathbf{S}_{1:T},\mathbf{A}_{1:T})}\Big[\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{\mathcal{F}}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) \Big| \mathbf{s}_{1:t}, \mathbf{a}_{1:t}\Big] \\
&\qquad - \mathbb{E}_{\hat{P}(\mathbf{S}_{t:T},\mathbf{A}_{t:T})}\Big[\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{\mathcal{F}}(\mathbf{s}_{t:T}, \mathbf{a}_{t:T}) \Big| s_t, a_t\Big] + V_{\boldsymbol{\theta}}(s_t),
\end{aligned}$$

where $\hat{P}(\mathbf{S}_{1:T}, \mathbf{A}_{1:T}) = \mathcal{T}(\mathbf{S}_{1:T}||\mathbf{A}_{1:T-1})\ \hat{\pi}(\mathbf{A}_{1:T}||\mathbf{S}_{1:T})$. Thus, setting $C(\mathbf{s}_{1:t}, \mathbf{a}_{1:t-1})$ to the remaining terms, $\boldsymbol{\theta}^{\mathrm{T}} \sum_{\tau=1}^{t-1} \mathbf{f}(s_\tau, a_\tau) + V_\theta(s_t)$, completes the proof. ∎

Proof of Theorem 8: We first re-express the optimization problem in terms of each of the alternative actions corresponding to the set of switch functions.

$$\begin{aligned}
&\operatorname*{argmax}_{\pi(\mathbf{A}_{1:T,1:N}||\mathbf{S}_{1:N}) \in \Xi} H_\pi(\mathbf{A}_{1:T,1:N}||\mathbf{S}_{1:T}) \text{ such that: } \qquad (36) \\
&\forall t \in \{1, \ldots, T\}, i \in \mathcal{N}, a_{t,i} \in \mathcal{A}_i, {a_{t,i}}' \in \mathcal{A}_i, s_{1:t} \in \boldsymbol{\mathcal{S}}_{\mathbf{1:t}}, \\
&\qquad \mathbf{a}_{1:t-1,1:N} \in \boldsymbol{\mathcal{A}}_{1:t-1,1:N}, \\
&\qquad \text{Regret}_{\pi,i}(a_{t,i}, {a_{t,i}}', \mathbf{s}_{1:t}, \mathbf{a}_{1:t-1,1:N}) \leq 0,
\end{aligned}$$

where $\text{Regret}_{\pi,i}(a_{t,i}, {a_{t,i}}', \mathbf{s}_{1:t}, \mathbf{a}_{1:t-1,1:N})$ is the regret corresponding to the switch function for player i from action $a_{t,i}$ to $a'_{t,i}$ at time t given history $\mathbf{a}_{1:t-1,1:N}$ and $\mathbf{s}_{1:t}$.

We find the form of the probability distribution by finding the optimal solution of the Lagrangian dual optimization problem. We suppress the probabilistic positivity constraints and normalization constraints with the understanding that the resulting probability distribution must normalize to 1.

The Lagrangian for the optimization of Equation 36 when using entire history-dependent probability distributions and parameters is:

$$\begin{aligned}
\Lambda(\pi, \lambda) = H_\pi(\mathbf{a}_{1:T,1:N}||\mathbf{s}_{1:T}) - \\
\sum_{\substack{t \in \{1, \ldots, T\}, i \in \mathcal{N}, \\ a_{t,i} \in \mathcal{A}_i, {a_{t,i}}' \in \mathcal{A}_i, \\ \mathbf{s}_{1:t} \in \boldsymbol{\mathcal{S}}_{1:t}, \\ \mathbf{a}_{1:t-1,1:N} \in \boldsymbol{\mathcal{A}}_{1:t-1,1:N}}} \lambda_{t,i,a_{t,i},{a_{t,i}}',\mathbf{s}_{1:t},\mathbf{a}_{1:t-1,1:N}} \times \\
\text{Regret}_{\pi,i}(a_{t,i}, {a_{t,i}}', \mathbf{s}_{1:t}, \mathbf{a}_{1:t-1,1:N}).
\end{aligned}$$

Taking the partial derivative with respect to a history-dependent action probability for a particular state, we have:

$$\begin{aligned}
&\frac{\partial \Lambda(\pi, \lambda)}{\partial \pi(\mathbf{a}_{t,1:N}|\mathbf{s}_{1:t}, \mathbf{a}_{1:t-1,1:N})} \\
&= P(\mathbf{a}_{1:t}, \mathbf{s}_{1:t}) \Big(H_\pi(\mathbf{A}_{t:T}||\mathbf{S}_{t:T}|\mathbf{a}_{1:t}, \mathbf{s}_{1:t}) \\
&\quad - \sum_{i \in \mathcal{N}, {a_{t,i}}' \in \mathcal{A}_i} \lambda_{t,i,a_{t,i},{a_{t,i}}',\mathbf{s}_{1:t},\mathbf{a}_{1:t-1,1:N}} \times \\
&\qquad \text{Regret}_{\pi,i}({a_{t,i}}', \mathbf{s}_{1:t}, \mathbf{a}_{1:t,1:N}) \Big) \\
&= P(\mathbf{a}_{1:t}, \mathbf{s}_{1:t}) \Big(-\log \pi(\mathbf{a}_{t,1:N}|\mathbf{s}_{1:t}, \mathbf{a}_{1:t-1,1:N}) \qquad (37) \\
&\quad + H_\pi(\mathbf{A}_{t+1:T}||\mathbf{S}_{t+1:T}|\mathbf{s}_{1:t}, \mathbf{a}_{1:t})
\end{aligned}$$

$$- \sum_{i \in \mathcal{N}, \mathbf{a}_{t,i}{}' \in \mathcal{A}_i} \lambda_{t,i,a_{t,i},\mathbf{a}_{t,i}{}',\mathbf{s}_{1:t},\mathbf{a}_{1:t-1,1:N}} \times \text{Regret}_{\pi,i}(a_{t,i}{}', \mathbf{s}_{1:t}, \mathbf{a}_{1:t,1:N}) \Big),$$

where here the regret is conditioned on the other players' actions $\mathbf{a}_{t,-i}$. The form of the history-dependent distribution,

$$\pi(\mathbf{a}_{t,1:N}|\mathbf{s}_{1:t}, \mathbf{a}_{1:t-1,1:N}) \propto \exp\Big\{ H_\pi(\mathbf{S}_{t+1:T}||\mathbf{A}_{t+1:T}|\mathbf{a}_{1:t}, \mathbf{s}_{1:t}) - \sum_{i \in \mathcal{N}, a_{t,i}{}' \in \mathcal{A}_i} \lambda_{t,i,a_{t,i},a_{t,i}{}',\mathbf{s}_{1:t},\mathbf{a}_{1:t-1,1:N}} \times \text{Regret}_{\pi,i}(a_{t,i}{}', \mathbf{s}_{1:t}, \mathbf{a}_{1:t,1:N}) \Big\}, \quad (38)$$

is obtained by equating Equation 37 to zero and dividing off the (constant) probability term, $P(\mathbf{a}_{1:t}, \mathbf{s}_{1:t})$. ■

APPENDIX C
ALTERNATIVE ENTROPY MAXIMIZATION APPROACHES

Can the same process estimates obtained by maximizing the causal entropy instead be obtained by maximizing more familiar entropy measures? The connection to the Bellman equation [2] established in Section IV-C allows us to answer this question by illustrating and interpreting the differences when employing other entropy measures.

Maximizing the conditional entropy of actions given states, $H_{\tilde{\pi}}(\mathbf{A}_{1:T}|\mathbf{S}_{1:T})$, provides a distribution of the form: $\hat{\pi}_{\boldsymbol{\theta}}(\mathbf{a}_{1:T}|\mathbf{s}_{1:T}) \propto \exp\{\sum_{t=1}^{T} \boldsymbol{\theta}^{\mathrm{T}} \mathbf{f}(s_t, a_t)\}$. As future states are latent, a common approach [65], [58] is to marginalize over the future latent states and actions, yielding a recursive expression for the conditional probability, $\hat{\pi}_{\boldsymbol{\theta}}(a_t|s_t) = e^{Q^{\text{cond}}_{\boldsymbol{\theta},t}(a_t,s_t) - V^{\text{cond}}_{\boldsymbol{\theta},t}(s_t)}$:

$$Q^{\text{cond}}_{\boldsymbol{\theta},t}(a_t, s_t) = \boldsymbol{\theta}^{\mathrm{T}} \mathbf{f}(s_t, a_t) + \underset{s_{t+1} \in \mathcal{S}_{t+1}}{\text{softmax}} \Big\{ \log \mathcal{T}(s_{t+1}|s_t, a_t) + V^{\text{cond}}_{\boldsymbol{\theta},t+1}(s_{t+1}) \Big\}$$

$$V^{\text{cond}}_{\boldsymbol{\theta},t}(s_t) = \underset{a_t \in \mathcal{A}}{\text{softmax}}\, Q^{\text{cond}}_{\boldsymbol{\theta},t}(a_t, s_t).$$

It can be interpreted as allowing the (softmax) selection of the next state s_{t+1} with the best state value potential with a penalty of $\log \mathcal{T}(s_{t+1}|s_t, a_t)$ incurred for realizing the desired state dynamics transition. In contrast, under the maximum causal entropy distribution and the Bellman equation, the expectation over the next state is taken according to the dynamics model.

Maximizing the joint entropy $H_{\tilde{\pi}}(\mathbf{A}_{1:T}, \mathbf{S}_{1:T})$ subject to constraints enforcing the dynamics distribution yields the following recursive definition of the conditional probability $\hat{\pi}_{\boldsymbol{\theta}}(a_t|s_t) = e^{Q^{\text{joint}}_{\boldsymbol{\theta},t}(a_t,s_t) - V^{\text{joint}}_{\boldsymbol{\theta},t}(s_t)}$:

$$Q^{\text{joint}}_{\boldsymbol{\theta},t}(a_t, s_t) = \mathbb{E}_{\mathcal{T}(s_{t+1}|s_t,a_t)}[V^{\text{joint}}_{\boldsymbol{\theta},t+1}(s_{t+1})|s_t, a_t] + \boldsymbol{\theta}^{\mathrm{T}} \mathbf{f}(s_t, a_t) + H_{\mathcal{T}}(S_{t+1}|s_t, a_t)$$

$$V^{\text{joint}}_{\boldsymbol{\theta},t}(s_t) = \underset{a_t \in \mathcal{A}}{\text{softmax}}\, Q^{\text{joint}}_{\boldsymbol{\theta},t}(a_t, s_t).$$

In contrast to the maximum causal entropy distribution (and the Bellman equation), more probability mass is assigned to actions leading towards portions of the state space where the dynamics are more stochastic. We refer the reader to our previous work [61] for an illustrative example of these differences.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Richard King Mellon Foundation, the Quality of Life Technology Center, and the Office of Naval Research Reasoning in Reduced Information Spaces project MURI for support of this research. We thank: Martial Hebert, Nathan Ratliff, and Andrew Maas for collaborations on projects that helped to drive this line of research; Geoff Gordon and Miro Dudík for useful discussions; and our reviewers for their valuable comments and suggestions.

REFERENCES

[1] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proc. International Conference on Machine Learning*, 2004, pp. 1–8.
[2] R. Bellman, "A Markovian decision process," *Journal of Mathematics and Mechanics*, vol. 6, pp. 679–684, 1957.
[3] A. Berger, V. Pietra, and S. Pietra, "A maximum entropy approach to natural language processing," *Computational linguistics*, vol. 22, no. 1, pp. 39–71, 1996.
[4] A. Boularias, J. Kober, and J. Peters, "Relative entropy inverse reinforcement learning," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2011, pp. 182–189.
[5] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, "Linear matrix inequalities in system and control theory," *SIAM*, vol. 15, 1994.
[6] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, March 2004.
[7] P. Buchen and M. Kelly, "The maximum entropy distribution of an asset inferred from option prices," *Journal of Financial and Quantitative Analysis*, vol. 31, no. 01, pp. 143–159, 1996.
[8] U. Chajewska, D. Koller, and D. Ormoneit, "Learning an agent's utility function by observing behavior," in *In Proc. of the International Conference on Machine Learning*, 2001, pp. 35–42.
[9] T. Cover and J. Thomas, *Elements of information theory*. John Wiley and sons, 2006.
[10] J. Cozzolino and M. Zahner, "The maximum-entropy distribution of the future market price of a stock," *Operations Research*, vol. 21, no. 6, pp. 1200–1211, 1973.
[11] Y. Dodis, S. Halevi, and T. Rabin, "A cryptographic solution to a game theoretic problem," in *Advances in Cryptology*. Springer, 2000, pp. 112–130.
[12] M. Dudík, S. J. Phillips, and R. E. Schapire, "Maximum entropy density estimation with generalized regularization and an application to species distribution modeling," *J. Mach. Learn. Res.*, vol. 8, pp. 1217–1260, 2007.
[13] K. Dvijotham and E. Todorov, "Inverse Optimal Control with Linearly-solvable MDPs," in *Proc. International Conference on Machine Learning*, 2010, pp. 335–342.
[14] D. Foster and R. Vohra, "Calibrated Learning and Correlated Equilibrium," *Games and Economic Behavior*, vol. 21, no. 1-2, pp. 40–55, 1997.
[15] A. Golan, G. Judge, and D. Miller, *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. Wiley, 1996.
[16] G. Gordon, A. Greenwald, and C. Marks, "No-regret learning in convex games," in *Proc. International Conference on Machine Learning*. ACM, 2008, pp. 360–367.
[17] P. D. Grünwald and A. P. Dawid, "Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory," *Annals of Statistics*, vol. 32, pp. 1367–1433, 2004.
[18] S. Hart and A. Mas-Colell, "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, vol. 68, no. 5, pp. 1127–1150, 2000.
[19] P. Henry, C. Vollmer, B. Ferris, and D. Fox, "Learning to Navigate Through Crowded Environments," in *Proc. International Conference on Robotics and Automation*, 2010, pp. 981–986.
[20] W. Huang and B. von Stengel, "Computing an extensive-form correlated equilibrium in polynomial time," *Internet and Network Economics*, pp. 506–513, 2008.
[21] E. T. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 106, pp. 620–630, 1957.
[22] ——, "Information theory and statistical mechanics, II," *Physical review*, vol. 108, no. 2, pp. 171–190, 1957.

[23] ——, "On the rationale of maximum-entropy methods," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 939–952, 1982.
[24] S. Kakade, M. Kearns, J. Langford, and L. Ortiz, "Correlated equilibria in graphical games," in *Proceedings of the 4th ACM Conference on Electronic Commerce*. ACM, 2003, pp. 42–47.
[25] R. Kalman, "When is a linear control system optimal?" *Trans. ASME, J. Basic Engrg.*, vol. 86, pp. 51–60, 1964.
[26] J. Kapur, *Maximum-entropy models in science and engineering*. John Wiley & Sons, 1989.
[27] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009.
[28] G. Kramer, "Directed information for channels with feedback," Ph.D. dissertation, Swiss Federal Institute of Technology (ETH) Zurich, 1998.
[29] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 49–86, 1951.
[30] S. Kumar and M. Hebert, "Discriminative random fields," *Int. J. Comput. Vision*, vol. 68, no. 2, pp. 179–201, 2006.
[31] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. International Conference on Machine Learning*, 2001, pp. 282–289.
[32] L. Liao, D. Fox, and H. Kautz, "Extracting places and activities from GPS traces using hierarchical conditional random fields," *Int. J. Rob. Res.*, vol. 26, no. 1, pp. 119–134, 2007.
[33] L. Mac Dermed and C. L. Isbell, "Solving Stochastic Games," in *Proc. Neural Information Processing Systems*, 2009, pp. 1186–1194.
[34] H. Marko, "The bidirectional communication theory – a generalization of information theory," in *IEEE Transactions on Communications*, 1973, pp. 1345–1351.
[35] J. L. Massey, "Causality, feedback and directed information," in *Proc. IEEE International Symposium on Information Theory and Its Applications*, 1990, pp. 27–30.
[36] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy Markov models for information extraction and segmentation," in *Proc. International Conference on Machine Learning*, 2000, pp. 591–598.
[37] D. McFadden, "Conditional logit analysis of qualitative choice behavior," *Frontiers in Econometrics*, pp. 105–142, 1974.
[38] C. Murray and G. Gordon, "Multi-robot negotiation: approximating the set of subgame perfect equilibria in general-sum stochastic games," in *Proc. Neural Information Processing Systems*, 2007, pp. 1001–1008.
[39] J. Nash, "Non-cooperative games," *Annals of mathematics*, vol. 54, no. 2, pp. 286–295, 1951.
[40] G. Neu and C. Szepesvári, "Apprenticeship learning using inverse reinforcement learning and gradient methods," in *Proc. UAI*, 2007, pp. 295–302.
[41] A. Y. Ng and S. Russell, "Algorithms for inverse reinforcement learning," in *Proc. International Conference on Machine Learning*, 2000, pp. 663–670.
[42] Y. Nyarko, "Bayesian learning leads to correlated equilibria in normal form games," *Economic Theory*, vol. 4, no. 6, pp. 821–841, 1994.
[43] F. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 295–302.
[44] L. E. Ortiz, R. E. Shapire, and S. M. Kakade, "Maximum entropy correlated equilibrium," in *Proc. International Conference on Artificial Intelligence and Statistics*, 2007, pp. 347–354.
[45] C. Papadimitriou and T. Roughgarden, "Computing equilibria in multiplayer games," in *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2005, pp. 82–91.
[46] H. H. Permuter, Y.-H. Kim, and T. Weissman, "On directed information and gambling," in *Proc. IEEE International Symposium on Information Theory*, 2008, pp. 1403–1407.
[47] V. D. Pietra, V. D. Pietra, and J. Lafferty, "Inducing features of random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 380–393, 1997.
[48] A. Quattoni, M. Collins, and T. Darrell, "Conditional random fields for object recognition," in *In Neural Information Processing Systems*, 2004.
[49] C. Quinn, T. Coleman, N. Kiyavash, and N. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *Journal of computational neuroscience*, vol. 30, no. 1, pp. 17–44, 2011.
[50] M. Raginsky, "Directed information and Pearl's causal calculus," in *Annual Allerton Conference on Communication, Control, and Computing*, 2011, pp. 958–965.
[51] D. Ramachandran and E. Amir, "Bayesian inverse reinforcement learning," in *Proc. IJCAI*, 2007, pp. 2586–2591.
[52] J. Rust, "Maximum likelihood estimation of discrete control processes," *SIAM Journal on Control and Optimization*, vol. 26, pp. 1006–1024, 1988.
[53] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 2003, pp. 134–141.
[54] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, 1948.
[55] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Transactions on Information Theory*, vol. 26, no. 1, pp. 26–37, 1980.
[56] S. Tatikonda and S. Mitter, "Control under communication constraints," *Automatic Control, IEEE Transactions on*, vol. 49, no. 7, pp. 1056–1068, 2004.
[57] F. Topsøe, "Information theoretical optimization techniques," *Kybernetika*, vol. 15, no. 1, pp. 8–27, 1979.
[58] M. Toussaint, "Robot trajectory optimization using approximate inference," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 1049–1056.
[59] D. L. Vail, M. M. Veloso, and J. D. Lafferty, "Conditional random fields for activity recognition," in *Proc. International Conference on Autonomous Systems and Multiagent Systems*, 2007, pp. 1–8.
[60] K. Waugh, B. D. Ziebart, and J. A. Bagnell, "Computational rationalization: The inverse equilibrium problem," in *Proc. of the International Conference on Machine Learning*, 2011, pp. 1169–1176.
[61] B. D. Ziebart, "Modeling purposeful adaptive behavior with the principle of maximum causal entropy," Ph.D. dissertation, Carnegie Mellon University, 2010.
[62] B. D. Ziebart, J. A. Bagnell, and A. K. Dey, "Modeling interaction via the principle of maximum causal entropy," in *Proc. International Conference on Machine Learning*, 2010, pp. 1255–1262.
[63] ——, "Maximum causal entropy correlated equilibria for Markov games," in *International Conference on Autonomous Agents and Multiagent Systems*, 2011, pp. 207–214.
[64] B. D. Ziebart, A. K. Dey, and J. A. Bagnell, "Probabilistic pointing target prediction via inverse optimal control," in *Proceedings of the ACM International Conference on Intelligent User Interfaces*, 2012, pp. 1–10.
[65] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning," in *Proc. AAAI Conference on Artificial Intelligence*, 2008, pp. 1433–1438.
[66] B. D. Ziebart, A. Maas, A. K. Dey, and J. A. Bagnell, "Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior," in *Proc. International Conference on Ubiquitous Computing*, 2008, pp. 322–331.
[67] B. D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. Srinivasa, "Planning-based prediction for pedestrians," in *Proc. Intelligent Robots and Systems*, 2009, pp. 3931–3936.

Brian D. Ziebart is an Assistant Professor in the Department of Computer Science at the University of Illinois at Chicago. He received his PhD in Machine Learning from Carnegie Mellon University in 2010, where he was also a postdoctoral fellow. His research interests include machine learning, decision theory, game theory, robotics, and assistive technologies.

J. Andrew (Drew) Bagnell is an Associate Professor in the Robotics Institute and Machine Learning Departments at Carnegie Mellon University. He received his PhD from Carnegie Mellon in 2004. Bagnell's research focuses on the intersection of machine learning with computer vision, optimal control, and robotics. His interests in machine learning range from algorithmic and theoretical development to delivering fielded learning-based systems.

Anind K. Dey is an Associate Professor in the Human-Computer Interaction (HCI) Institute at Carnegie Mellon University. He received his Ph.D. in Computer Science from Georgia Tech in 2000, and was a Senior Researcher for Intel Research Berkeley and an Adjunct Assistant Professor at UC Berkeley from 2001 to 2004. His main research focus lies at the intersection of HCI and ubiquitous computing.