Modeling Interaction via the Principle of Maximum Causal Entropy

Brian D. Ziebart Carnegie Mellon University
J. Andrew Bagnell Carnegie Mellon University
Anind K. Dey Carnegie Mellon University

Abstract

The principle of maximum entropy provides a powerful framework for statistical models of joint, conditional, and marginal 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 causally conditioned probabilities that can appropriately model the availability and influence of sequentially revealed side information. Using this principle, we derive Maximum Causal Entropy Influence Diagrams, a new probabilistic graphical framework for modeling decision making in settings with latent information, sequential interaction, and feedback. We describe the theoretical advantages of this model and demonstrate its applicability for statistically framing inverse optimal control and decision prediction tasks.

1. Introduction

The principle of maximum entropy (Jaynes, 1957) serves a foundational role in the theory and practice of constructing statistical models, with applicability to statistical mechanics (Jaynes, 1957), natural language processing, econometrics, and ecology (Dudík & Schapire, 2006). Conditional extensions of the principle that consider *side information*, and specifically Conditional Random Fields (Lafferty et al., 2001), have been applied with remarkable success in recognition, segmentation, and classification tasks, and are a preferred tool in natural language processing, machine vision, and activity recognition.

This work extends the maximum entropy approach to conditional probability distributions in settings characterized by *interaction with stochastic processes where side information from those processes is dynamic*, i.e., revealed over time. Importantly, in these settings, future side information is latent during earlier points of interaction. Consequentially, the value of side information should have no *causal* influence in our statistical models of interaction until after it is revealed, though the distribution from which it is drawn can be influential. More formally this means that if a future side information variable were secretly fixed to some value by *intervention* (Pearl, 2000) rather than sampled according to its conditional probability distribution, the distribution over all earlier variables would be unaffected by this change.

Conditional maximum entropy approaches are ill-suited for this setting as they assume all side information is available a priori. Building on the recent advance of the Marko-Massey theory of directed information (Massey, 1990), we propose the use of the *causally conditioned entropy* (Kramer, 1998) as this measure matches the information availability of our setting and has found applicability in the analysis of communication channels with feedback (Kramer, 1998), decentralized control (Tatikonda & Mitter, 2004), sequential investment and online compression with side information (Permuter et al., 2008).

We present the *principle of maximum causal entropy* (MaxCausalEnt), which prescribes a probability distribution by maximizing the entropy of a sequence of variables causally conditioned on sequentially revealed side information. This contribution extends the maximum entropy framework for statistical modeling to processes with information revelation, feedback, and interaction. We adopt the Influence Diagram, a *prescriptive* decision-making framework, as a convenient representation of the probability distribution of side information, its dynamic availability, and its relationships to other variables. Applying the MaxCausalEnt principle yields the MaxCausalEnt Influence Diagram, a novel *predictive* framework for estimating decision probabilities. We demonstrate that our framework provides a fully probabilistic approach to problems of inverse stochastic control, multiple agent behavior prediction in dynamic games, and modeling interaction with partially observable systems.

Preprint. To appear in the International Conference on Machine Learning (ICML).

2. Maximum Causal Entropy

Motivated by the task of modeling behavior with elements of sequential interaction, we introduce the principle of maximum causal entropy and describe its core theoretical properties.

2.1. Preliminaries

When faced with an ill-posed problem, the principle of maximum entropy (Jaynes, 1957) prescribes the use of "the least committed" probability distribution that is consistent with known problem constraints. This criterion is formally measured by Shannon's information entropy, $E_X[-\log P(X)]$, and many of the fundamental building blocks of statistics, including Gaussian and Markov random field distributions, maximize this entropy subject to moment constraints.

In the presence of *side information*, $\mathbf{X}$, that we do not desire to model, the standard prescription is to maximize the conditional entropy, $E_{\mathbf{Y},\mathbf{X}}[-\log P(\mathbf{Y}|\mathbf{X})]$, yielding, for example, the Conditional Random Field (CRF) (Lafferty et al., 2001). Though our intention is to similarly model conditional probability distributions, CRFs assume a knowledge of future side information, $X_{t+1:T}$, for each Y_t that does not match settings with dynamically revealed information. Attempts to marginalize over the joint distribution using a CRF are possible:

$$P(Y_t|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1}) \propto \sum_{\mathbf{X}_{t+1:T}, \mathbf{Y}_{t+1:T}} e^{\theta^\top F(\mathbf{X},\mathbf{Y})} P(\mathbf{X}_{t+1:T}|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1}). \quad (1)$$

However, we argue that entropy-based approaches that do not address the causal influence of side information are inadequate for interactive settings.

2.2. Directed Information and Causal Entropy

The *causally conditioned probability* (Kramer, 1998) from the Marko-Massey theory of directed information (Massey, 1990) is a natural extension of the conditional probability, $P(\mathbf{Y}|\mathbf{X})$, to the situation where each Y_t is conditioned on only a portion of the $\mathbf{X}$ variables, $\mathbf{X}_{1:t}$, rather than the entirety, $\mathbf{X}_{1:T}$. Following the previously developed notation (Kramer, 1998), the probability of $\mathbf{Y}$ *causally conditioned* on $\mathbf{X}$ is

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

The subtle, but significant difference from conditional probability, $P(\mathbf{Y}|\mathbf{X}) = \prod_{t=1}^{T} P(Y_t|\mathbf{X}_{1:T}, \mathbf{Y}_{1:t-1})$, serves as the underlying basis for our approach.

Causal entropy (Kramer, 1998; Permuter et al., 2008),

$$H(\mathbf{Y}^T||\mathbf{X}^T) \triangleq E_{\mathbf{Y},\mathbf{X}}[-\log P(\mathbf{Y}^T||\mathbf{X}^T)] = \sum_{t=1}^{T} H(Y_t|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1}), \quad (3)$$

measures the uncertainty present in the causally conditioned distribution. It is easy to verify that it upper bounds the conditional entropy; intuitively this reflects the fact that conditioning on information from the future (i.e., acausally) can only decrease uncertainty. Using this notation, any joint distribution can be expressed as $P(\mathbf{Y}, \mathbf{X}) = P(\mathbf{Y}^T||\mathbf{X}^T)P(\mathbf{X}^T||\mathbf{Y}^{T-1})$. Our approach estimates $P(\mathbf{Y}^T||\mathbf{X}^T)$ based on a provided (explicitly or implicitly) distribution of side information $P(\mathbf{X}^T||\mathbf{Y}^{T-1}) = \prod_t P(X_t|\mathbf{X}_{1:t-1}, \mathbf{Y}_{1:t-1})$.

2.3. Maximum Causal Entropy Optimization

With the causal entropy (Equation 3) as our objective function, we now pose and solve the maximum causal entropy optimization problem. We constrain our distribution to match expected *feature functions*, $\mathcal{F}(\mathbf{X}, \mathbf{Y})$ with empirical expectations of those same functions, $\tilde{E}_{\mathbf{X},\mathbf{Y}}[\mathcal{F}(\mathbf{X}, \mathbf{Y})]$, yielding the following optimization problem:

$$\underset{\{P(Y_t|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1})\}}{\operatorname{argmax}} H(\mathbf{Y}^T||\mathbf{X}^T) \quad (4)$$
$$\text{such that: } E_{\mathbf{X},\mathbf{Y}}[\mathcal{F}(\mathbf{X},\mathbf{Y})] = \tilde{E}_{\mathbf{X},\mathbf{Y}}[\mathcal{F}(\mathbf{X},\mathbf{Y})]$$
$$\text{and } \forall_{\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1}} \sum_{Y_t} P(Y_t|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1}) = 1.$$

Theorem 1. *The distribution satisfying the maximum causal entropy constrained optimization (Equation 4) has a form defined recursively as:*

$$P_\theta(Y_t|\boldsymbol{X}_{1:t}, \boldsymbol{Y}_{1:t-1}) = \frac{Z_{Y_t|\boldsymbol{X}_{1:t}, \boldsymbol{Y}_{1:t-1}, \theta}}{Z_{\boldsymbol{X}_{1:t}, \boldsymbol{Y}_{1:t-1}, \theta}} \quad (5)$$

$$\log Z_{\boldsymbol{X}_{1:t}, \boldsymbol{Y}_{1:t-1}, \theta} = \log \sum_{Y_t} Z_{Y_t|\boldsymbol{X}_{1:t}, \boldsymbol{Y}_{1:t-1}, \theta}$$
$$= \underset{Y_t}{\text{softmax}} \Big(\sum_{X_{t+1}} P(X_{t+1}|\boldsymbol{X}_{1:t}, \boldsymbol{Y}_{1:t}) \log Z_{\boldsymbol{X}_{1:t+1}, \boldsymbol{Y}_{1:t}, \theta} \Big)$$

$$Z_{Y_t|\boldsymbol{X}_{1:t}, \boldsymbol{Y}_{1:t-1}, \theta} = e^{\sum_{X_{t+1}} P(X_{t+1}|\boldsymbol{X}_{1:t}, \boldsymbol{Y}_{1:t}) \log Z_{\boldsymbol{X}_{1:t+1}, \boldsymbol{Y}_{1:t}, \theta}}$$

$$\log Z_{\boldsymbol{X}_{1:T}, \boldsymbol{Y}_{1:T-1}, \theta} = \theta^\top \mathcal{F}(\boldsymbol{X}, \boldsymbol{Y}),$$

where $\text{softmax}_x f(x) \triangleq \log \sum_x e^{f(x)}$.

Proof (sketch).[1] We note that the (negated) primal objective function (Equation 4) is convex in the variables $P(\mathbf{Y}||\mathbf{X})$ and subject to linear constraints on feature function expectation matching, valid probability distributions, and non-causal influence of future

[1] Complete proofs and additional algorithm and experimental details are available in the supplement to this paper.

side information. Differentiating the Lagrangian of the causal maximum entropy optimization (Equation 4), and equating to zero, we obtain the general form:

$$P_\theta(Y_t|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t-1}) \propto \exp\Big\{\theta^\top E_{\mathbf{X},\mathbf{Y}}[\mathcal{F}(\mathbf{X},\mathbf{Y})|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t}] - \sum_{\tau>t} E_{\mathbf{X},\mathbf{Y}}[\log P_\theta(Y_\tau|\mathbf{X}_{1:\tau}, \mathbf{Y}_{1:\tau-1})|\mathbf{X}_{1:t}, \mathbf{Y}_{1:t}]\Big\}. \quad (6)$$

Substituting the more operational recurrence of Equation 5 into Equation 6 verifies the theorem. □

We note that Theorem 1 relies on strong duality to identify the form of this probability distribution; the sharp version of Slater's condition (Boyd & Vandenberghe, 2004) using the existence of a feasible point in the relative interior ensures this but requires that (1) prescribed feature counts are achievable, and (2) the distribution has full support. The first naturally follows if both model and empirical expectations are taken with respect to the provided model of side information, $P(\mathbf{X}^T||\mathbf{Y}^{T-1})$. For technical simplicity in this work, we will further assume full support for the modeled distribution, although relatively simple modifications (e.g., constraints hold within a small deviation ϵ) ensure the correctness of this form in all cases.

Theorem 2. *The gradient of the dual with respect to θ is $\big(\tilde{E}_{\boldsymbol{X},\boldsymbol{Y}}[\mathcal{F}(\boldsymbol{X},\boldsymbol{Y})] - E_{\boldsymbol{X},\boldsymbol{Y}}[\mathcal{F}(\boldsymbol{X},\boldsymbol{Y})]\big)$, which is the difference between the expected feature vector under the probabilistic model and the empirical feature vector given the complete policy, $\{P(Y_t|\boldsymbol{X}_{1:t}, \boldsymbol{Y}_{1:t-1})\}$.*

In many instances, the statistics of interest $\big(\tilde{E}_{\mathbf{X},\mathbf{Y}}[\mathcal{F}(\mathbf{X},\mathbf{Y})]\big)$ are only known approximately as they are obtained from small sample sets. We note that this uncertainty can be rigorously addressed by extending the duality analysis of Dudík & Schapire (2006), leading to parameter regularization that may be naturally adopted in the causal setting as well.

Theorem 3. *The maximum causal entropy distribution minimizes the worst case prediction log-loss,*

$$\inf_{P(\boldsymbol{Y}||\boldsymbol{X})} \sup_{\tilde{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T)} \sum_{\boldsymbol{Y},\boldsymbol{X}} \tilde{P}(\boldsymbol{Y},\boldsymbol{X}) \log P(\boldsymbol{Y}^T||\boldsymbol{X}^T),$$

given that $\tilde{P}(\boldsymbol{Y},\boldsymbol{X}) = \tilde{P}(\boldsymbol{Y}^T||\boldsymbol{X}^T)P(\boldsymbol{X}^T||\boldsymbol{Y}^{T-1})$ and feature expectations $E_{\tilde{P}(\boldsymbol{X},\boldsymbol{Y})}[\mathcal{F}(\boldsymbol{X},\boldsymbol{Y})]$ when $\boldsymbol{X}$ is sequentially revealed from a known distribution.

Theorem 3 follows naturally from Grünwald & Dawid (2003) and extends their "robust Bayes" results to the interactive setting as one justification for the maximum causal entropy approach. The theorem can be understood by viewing maximum causal entropy as a *maximin* game where nature chooses a distribution to maximize a predictor's perplexity while the predictor tries to minimize it. By duality, the *minimax* view of the theorem is equivalent. This strong result is not shared when maximizing alternate entropy measures (e.g., conditional or joint entropy) and marginalizing out future side information (as in Equation 1).

3. MaxCausalEnt Influence Diagrams

We now apply MaxCausalEnt to the Influence Diagram (Howard & Matheson, 1984), a graphical framework that subsumes Bayesian Networks, augmenting their capabilities to reason about latent variables with *decisions* and *utilities* so that inference includes optimal decision making. MaxCausalEnt expands the applicability of Influence Diagrams from the *prescription* of optimal decisions to the *prediction* of decision-making and the recovery of explanatory utility weights from observed decision sequences. It also generalizes the probability distribution derived in Theorem 1 to settings with partially observable side information, and provides a convenient graphical representation for the MaxCausalEnt variables and their relationships.

3.1. Variables, Dependencies, and Features

A Maximum Causal Entropy Influence Diagram (MaxCausalEnt ID) is structurally characterized by square decision nodes ($\mathbf{Y}$), circular uncertainty nodes ($\mathbf{X}$), diamond value nodes ($\mathbf{V}$), and directed edges. The role of an edge depends on the type of node to which it is a parent. The parents of a decision node, $par(Y_i)$, are known when the decision is made. An uncertainty node's parents, $par(X_i)$, specify its conditional probability distribution, $P(X_i|par(X_i))$. The parents of a value node, $par(V_i)$, in a MaxCausalEnt ID indicate the form of *feature functions* of the value node $\mathcal{F}_{V_i} : par(V_i) \to \Re^k$. We restrict our consideration in this work to decision settings with perfect recall[2].

3.2. Causal Decision Entropy Maximization

We define the *causal decision probability* of a MaxCausalEnt ID as $P(\mathbf{Y}||par(\mathbf{Y})) \triangleq \prod_t P(Y_t|par(Y_t))$ and the *causal decision entropy* as $H(\mathbf{Y}||par(\mathbf{Y})) \triangleq E_{\mathbf{X},\mathbf{Y}}[-\log P(\mathbf{Y}||par(\mathbf{Y}))]$. We maximize this entropy while matching the additive combination of expected and empirical feature functions, $E_{\mathbf{X},\mathbf{Y}}[\sum_V \mathcal{F}(V)] = \tilde{E}_{\mathbf{X},\mathbf{Y}}[\sum_V \mathcal{F}(V)]$. The key distinction between this setting and the setting discussed in Section 2 is that some variables in $\mathbf{X}$ may never be directly observed in the MaxCausalEnt ID.

3.3. Inference and Learning Algorithms

Algorithm 1 illustrates the procedure for inferring de-

[2]Variables observed during previous decisions are either observed in future decisions or irrelevant (i.e., value node descendants of future decisions are conditionally independent from that variable given other observed variables).

Algorithm 1 MaxCausalEnt ID Inference Procedure

1: **for all** $V \in \mathbf{V}$ **do**
2: Associate V with $\text{argmax}_{Y_{\text{index}} \in \text{ancest}(V)}$ index
3: **end for**
4: **for** $i = |\mathbf{Y}|$ to 1 **do**
5: **for all** values $(Y_i', par(Y_i)')$ **do**
6: $Z_i(Y_i'|par(Y_i)') \leftarrow 0$
7: **for all** V_j associated with Y_i **do**
8: **for all** values $par(V_j)'$ **do**
9: Compute $P(par(V_j)'|par(Y_i)', Y_i')$
10: **end for**
11: $Z_i(Y_i'|par(Y_i)') \leftarrow Z_i(Y_i'|par(Y_i)') + E_{par(V_j)}[\theta^\top \mathcal{F}_{V_j}(par(V_j))|Y_i', par(Y_i)']$
12: **end for**
13: **for all** values $par(Y_{i+1})'$ **do**
14: Compute $P(par(Y_{i+1})'|Y_i', par(Y_i)')$
15: **end for**
16: $Z_i(Y_i'|par(Y_i)') \leftarrow Z_i(Y_i'|par(Y_i)') + E_{par(Y_{i+1})'}[Z_{i+1}(par(Y_{i+1})')|Y_i', par(Y_i)']$
17: **end for**
18: **for all** $par(Y_i)'$ **do**
19: $Z_i(par(Y_i)') \leftarrow \text{softmax}_{Y_i} Z_i(Y_i|par(Y_i)')$
20: **end for**
21: **end for**

Algorithm 2 MaxCausalEnt ID Gradient Calculation

1: Compute $\tilde{E}[\mathcal{F}] \leftarrow \frac{1}{T}\sum_t E_{\mathbf{X},\mathbf{Y}}[\sum_V \mathcal{F}(V)|\tilde{\mathbf{x}}_t, \tilde{\mathbf{y}}_t]$
2: Compute $Z(par(Y)), Z(Y|par(Y))$ via Algorithm 1 for Parameters θ
3: **for all** decision nodes Y_i **do**
4: Replace Y_i with an uncertainty node with probabilities $P(y_i|par(Y_i)) = e^{Z(y_i|par(Y_i)) - Z(par(Y_i))}$
5: **end for**
6: $E[\mathcal{F}] \leftarrow \mathbf{O}^k$
7: **for all** V **do**
8: **for all** values par(V)' **do**
9: Compute $P(par(V)')$ using $\{P(y_i|par(Y_i))\}$
10: **end for**
11: $E[\mathcal{F}] \leftarrow E[\mathcal{F}] + E[F(V)]$
12: **end for**
13: $\nabla_\theta \log P(\mathbf{y}||par(\mathbf{y})) \leftarrow \tilde{E}[\mathcal{F}] - E[\mathcal{F}]$

cision probabilities in the MaxCausalEnt ID based on Theorem 1. We assume as a subroutine an algorithm for calculating the marginal probabilities of variables conditioned on a set of fixed evidence variables in a Bayesian Network (e.g., variable elimination or belief propagation). Expectations over the unobserved uncertainty nodes that are ancestors of value variables are employed in line 11, replacing the exact evaluations, $\theta^\top \mathcal{F}(\mathbf{X}, \mathbf{Y})$, of Equation 4. This added expectation preserves the convexity of the optimization since the feature matching constraint remains linear in the causally conditioned probabilities. Standard gradient-based optimization techniques can be employed using the gradient calculated in Algorithm 2.

4. Applications

We now present a series of applications with increasing complexity of interaction: (1) control with stochastic dynamics; (2) multiple agent interaction; and (3) interaction with a partially observable system.

4.1. Inverse Optimal Stochastic Control

Optimal control frameworks, such as the Markov Decision Process (MDP) and the Linear-Quadratic Regulator (LQR), provide rich representations of interactions with stochastic systems. Inverse optimal control (IOC) (Kalman, 1964; Boyd et al., 1994) is the problem of recovering a cost function that makes a particular controller or policy (near)-optimal. Recent work has demonstrated that IOC is a powerful technique for modeling the decision-making behavior of intelligent agents in problems as diverse as robotics (Ratliff et al., 2009), personal navigation (Ziebart et al., 2008), and cognitive science (Ullman et al., 2009). Many recent IOC approaches (Abbeel & Ng, 2004; Ziebart et al., 2008) consider cost functions linear in a set of features, and attempt to find behaviors that induce the same feature counts as the policy to be mimicked ($E[\sum_t \mathcal{F}_{S_t}] = \tilde{E}[\sum_t \mathcal{F}_{S_t}]$); by linearity such behaviors must achieve the same expected cost or value. Unfortunately, matching feature counts is fundamentally ill-posed – usually no truly optimal policy will achieve those feature counts, but many stochastic policies (and policy mixtures) may satisfy this constraint.

Ziebart et al. (2008) resolve this ambiguity by using the classical maximum entropy criteria to select a single distribution from all the distributions over decisions that match feature counts. They provide an exact solution for deterministic MDPs and propose an approximate solution for the Inverse Optimal Stochastic Control (IOSC) problem for MDPs with stochastic dynamics that is equivalent to the conditional entropy model with latent state variables (Equation 1).

The feature-matching concept is easily extended to discrete-time, continuous-state, continuous-action settings where quadratic properties of actions ($\mathbf{a}$) and states ($\mathbf{s}$) can be matched in expectation: $E[\sum_t \mathbf{a}_t \mathbf{a}_t^\top] = \tilde{E}[\sum_t \mathbf{a}_t \mathbf{a}_t^\top]$ and $E[\sum_t \mathbf{s}_t \mathbf{s}_t^\top] = \tilde{E}[\sum_t \mathbf{s}_t \mathbf{s}_t^\top]$. Cost functions are then linear in these features, e.g., $\sum_t Tr[\mathbf{s}\,\mathbf{s}^\top \mathbf{Q}]$.

4.1.1. MaxCausalEnt ID Formulation

In the IOSC problem, side information (states) and decisions (actions) are inter-dependent with the distribution of side information provided by the known

dynamics, $P(\mathbf{S}^T||\mathbf{A}^{T-1}) = \prod_t P(S_t|S_{t-1}, A_{t-1})$. We employ the MaxCausalEnt ID framework, as shown in Figure 1, to solve this problem exactly.

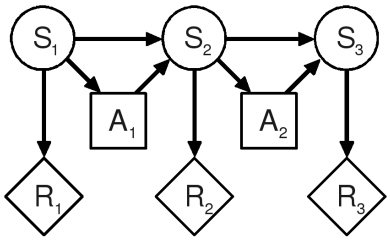

Figure 1. The MaxCausalEnt ID inverse optimal control representation for MDPs with state-based features.

Using the action-based cost-to-go (Q) and state-based value (V) notation, the inference procedure for MDP MaxCausalEnt IOC reduces to

$$Q_\theta^{\text{soft}}(a_t, s_t) = \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) V_\theta^{\text{soft}}(s_{t+1}) \qquad (7)$$
$$V_\theta^{\text{soft}}(s_t) = \underset{a_t}{\text{softmax}}\, Q_\theta^{\text{soft}}(a_t, s_t) + \theta^\top \mathcal{F}_{s_t},$$

and for the continuous, quadratic-reward setting to

$$Q_\theta^{\text{soft}}(\mathbf{a}_t, \mathbf{s}_t) = \int_{\mathbf{s}_{t+1}} P(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) V_\theta^{\text{soft}}(\mathbf{s}_{t+1}) d\mathbf{s}_{t+1} + \mathbf{a}_t^\top \mathbf{R} \mathbf{a}_t$$
$$V_\theta^{\text{soft}}(\mathbf{s}_t) = \underset{\mathbf{a}_t}{\text{softmax}}\, Q_\theta^{\text{soft}}(\mathbf{a}_t, \mathbf{s}_t) + \mathbf{s}_t^\top \mathbf{Q} \mathbf{s}_t. \qquad (8)$$

Note that by replacing the *softmax*[3] function with the *maximum*, this algorithm becomes equivalent to the (stochastic) value iteration algorithm (Bellman, 1957) for finding the optimal control policy. The softened version yields a stochastic policy, $\pi_\theta(a|s) \propto e^{Q_\theta^{\text{soft}}(a,s)}$.

For the special case where dynamics are linear functions with Gaussian noise, many continuous optimal control problems permit closed-form solutions. The same is true of inference for Inverse MaxCausalEnt LQR. Assuming the dynamics are $\mathbf{s}_{t+1} \sim N(\mathbf{A}\mathbf{s}_t + \mathbf{B}\mathbf{a}_t, \Sigma)$, Equation 8 reduces to:

$$Q_\theta^{\text{soft}}(\mathbf{a}_t, \mathbf{s}_t) = \begin{bmatrix} \mathbf{a}_t \\ \mathbf{s}_t \end{bmatrix}^\top \begin{bmatrix} \mathbf{B}^\top \mathbf{D} \mathbf{B} + \mathbf{R} & \mathbf{A}^\top \mathbf{D} \mathbf{B} \\ \mathbf{B}^\top \mathbf{D} \mathbf{A} & \mathbf{A}^\top \mathbf{D} \mathbf{A} \end{bmatrix} \begin{bmatrix} \mathbf{a}_t \\ \mathbf{s}_t \end{bmatrix}$$
$$V_\theta^{\text{soft}}(\mathbf{s}_t) = \mathbf{s}_t^\top (\mathbf{C}_{s,s} + Q - \mathbf{C}_{a,s}^\top \mathbf{C}_{a,a}^{-1} \mathbf{C}_{a,s}) \mathbf{s}_t + \text{const},$$

where $\mathbf{C}$ and $\mathbf{D}$ are recursively computed as: $\mathbf{C}_{a,a} = \mathbf{B}^\top \mathbf{D} \mathbf{B}; \mathbf{C}_{s,a} = \mathbf{C}_{a,s}^\top = \mathbf{B}^\top \mathbf{D} \mathbf{A}; \mathbf{C}_{s,s} = \mathbf{A}^\top \mathbf{D} \mathbf{A}$; and $\mathbf{D} = \mathbf{C}_{s,s} + \mathbf{Q} - \mathbf{C}^\top \mathbf{C}_{a,a}^{-1} \mathbf{C}_{a,s}$.

4.1.2. Inverse Helicopter Control

We demonstrate the MaxCausalEnt approach to inverse stochastic optimal control on the problem of building a controller for a learned helicopter model (Abbeel et al., 2007) with linearized stochastic dynamics. Most existing approaches to IOSC (Ratliff et al., 2006; Abbeel & Ng, 2004) have both practical and theoretical difficulties in the presence of imperfect demonstrated behavior, leading to unstable controllers due to large changes in cost weights (Abbeel et al., 2007) or poor predictive accuracy (Ratliff et al., 2006). To test the robustness of our approach, we generated five 100 time-step sub-optimal training trajectories by noisily sampling actions from an optimal LQR controlled designed for hovering using the linearized stochastic simulator of Abbeel et al. (2007).

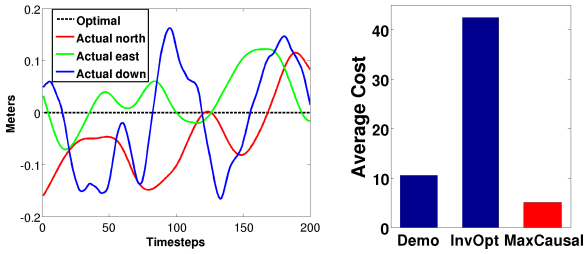

Figure 2. Left: An example sub-optimal helicopter trajectory attempting to hover around the origin point. Right: The average cost under the original cost function of: (1) demonstrated trajectories; (2) the optimal controller using the inverse optimal control model; and (3) the optimal controller using the maximum causal entropy model.

We contrast performance between maximum margin planning (Ratliff et al., 2006) (labeled IOSC in Figure 2) and MaxCausalEnt trained using demonstrated trajectories. Performance was evaluated by generating trajectories from the optimal controller of each model and measuring their cost under the *true* cost function used to generate the original sub-optimal demonstration trajectories. The InvOpt model performs poorly because there is no optimal trajectory *for any cost function* that matches demonstrated features. On the other hand, the function learned by MaxCausalEnt IOC not only induces the same feature counts– and hence equal cost on the unknown cost function– under the learned probabilistic policy, but because of the quadratic cost function its learned controller's optimal policy is always *at least as good* as the demonstrated behavior on the original, unknown cost function. Figure 2 demonstrates this; the resulting learned optimal policy outperforms the demonstrated behavior on the original, unknown cost function. In this sense, MaxCausalEnt provides a rigorous approach to learning a cost function for such stochastic optimal control problems: it is both predictive and can guarantee good performance of the learned controller.

[3]The continuous version of the softened maximum is defined as: $\text{softmax}_{\mathbf{x}} f(\mathbf{x}) \triangleq \log \int_{\mathbf{x}} e^{f(\mathbf{x})}\, d\mathbf{x}$.

4.2. Inverse Dynamic Games

Modeling the interactions of multiple agents is an important task for uncovering the motives of negotiating parties, planning a robot's movement in a crowded environment, and assessing the perceived roles of interacting agents (Ullman et al., 2009). While game and decision theory can prescribe the optimal action policy when the utilities of agents are known, often these utilities are not known and only observed behavior is available. We investigate the setting where the interleaved sequences of actions from R agents, $\mathbf{A}$, and state sequences, $\mathbf{S}$, generated from stochastic dynamics ($P(S_{t+1}|S_t, A_t)$), are observed. The learning task is to recover each player's reward function, $\mathbf{w}_j^\top \mathbf{f}(\mathbf{A}, \mathbf{S})$.

4.2.1. MaxCausalEnt ID Formulation

In the multi-agent setting, side information (the sequence of states) is governed by other agents' policies, $\pi_k(A|S)$. Given those other agents' policies, the conditional distribution of side information (states) is known and the j-th agent's maximum causal entropy policy is obtained with the following optimization:

$$\underset{\pi_j(A|S)}{\text{argmax}}\ H(\mathbf{A}||\mathbf{S}) \tag{9}$$
$$E[\sum_t \mathbf{f}_j(S_t)|\forall_k \pi_k(A|S)] = \tilde{E}[\sum_t \mathbf{f}_j(S_t)].$$

For many learning setting, only state-action traces are available and not full policies. Unfortunately when trying to jointly learn the behavior of multiple agents, the distribution of side information is no longer known (it depends on the other agents' unknown policies), and the MaxCausalEnt approach cannot be applied in a straight-forward manner. We instead employ a cyclic coordinate descent approach, which iteratively maximizes the causal entropy of one agent's actions while matching expected feature counts under the other agents' learned policies with the agent's "empirical" feature counts (i.e., expected feature counts of the agents demonstrated actions under the other agents' learned policies), and settle for local optima.

4.2.2. Pursuit-Evasion Modeling

We consider a generalization of the pursuit-evasion multi-agent setting (Parsons, 1976) with three agents operating in a four-by-four grid world. Each agent has a mobility, $m_i \in [0, 1]$, which corresponds to the probability of success when attempting to move in one of the four cardinal directions, and a utility for being co-located with each of the other agents. In the prediction task, we are provided with the mobilities of each agent and a time sequence of their actions and locations, and hope to recover the underlying co-location utility matrix.

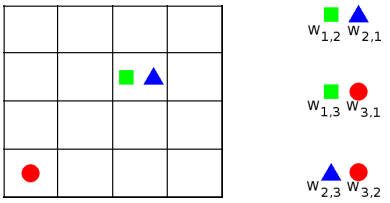

Figure 3. The pursuit-evasion grid with three agents and their co-location utilities. Agent X has a mobility of m_X and a utility of $w_{X,Y}$ when co-located with agent Y.

We generate data for this setting using the following procedure. First, for each agent, mobilities ($U[0.2, 1.0]$) and co-location utilities ($U[-1.0, 1.0]$) are sampled. Next, a potentially sub-optimal policy for each agent is generated by solving the optimal sequential decision problem (maximizing expected utility) for ten different limited time horizons, and mixing these policies by switching between them uniformly at random. Lastly, the agents take random initial locations and from the stochastic policy and state dynamics, a trajectory of 40 time-steps is sampled. Five training trajectories and one testing trajectory are generated for six different parameter samples. Despite its simplicity, this setting produces suprising rich behavior. For example, a first evader may help its pursuer corner a more desirable second evader so that the first evader will be spared. Due to the sensitivity of the optimal policy to time horizon and symmetries in the state space (broken at random), the resulting policies are often stochastic.

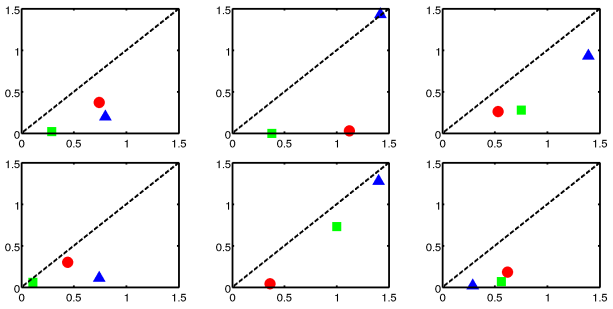

Figure 4. The average per-action perplexities of the latent CRF model and the MaxCausalEnt model plotted against each other for three agents from six different pursuit-evasion settings. The MaxCausalEnt model outperforms the latent CRF model in the region below the dotted line.

A comparison between the latent CRF model (Equation 1) trained to maximize data likelihood and the maximum causal entropy model is shown in Figure 4 using perplexity, $\frac{1}{T}\sum_{a_t,s_t} \log P(a_t|s_t)$, as the evaluation metric. In addition to this empirical demonstration of the benefit of the MaxCausalEnt approach on

this problem, a conceptual understanding of the distinction between the two approaches can be obtained by realizing that the latent CRF approach is equivalent to employing $Q(a_t, s_t) = \text{softmax}_{s_{t+1}}(V(s_{t+1}) + \log P(s_{t+1}|s_t, a_t))$ within Equation 7. This has a disconcerting interpretation that the agent somehow *chooses* the next state by "paying" an extra $\log P(s_{t+1}|s_t, a_t)$ penalty to ignore the actual provided stochasticity of the problem dynamics.

4.3. Inverse Diagnostics

Many important problems can be framed as interaction with a partially observable stochastic system. In medical diagnosis, for example, tests are conducted, the results of which may lead to additional tests to narrow down probable conditions or diseases and to prescribe treatments, which are adjusted based on patient response. Motivated by the objective of learning good diagnosis policies from experts, we investigate the Inverse Diagnostics problem of modeling interaction with partially observed systems.

4.3.1. MaxCausalEnt ID Formulation

In this setting, the partially observed set of variables (related by a Bayesian Network in this application) serves as side information. Inference over the latent variables from this set is required to infer decision probabilities. Additionally, decisions can influence the variables, causally changing their values, and the implications of these interventions must also be assessed. Vectors of features are associated with observing or manipulating each variable and we employ our MaxCausalEnt ID model with these features as value nodes as shown in Figure 5[4].

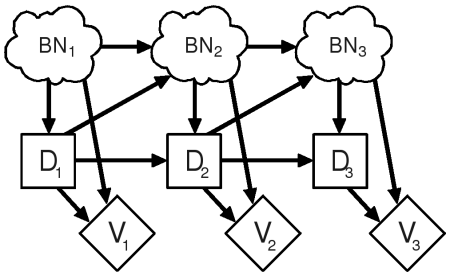

Figure 5. The MaxCausalEnt ID representation of the diagnostic problem.

4.3.2. Fault Diagnosis Experiments

We apply our inverse diagnostics approach to the vehicle fault detection Bayesian Network (Heckerman et al., 1994) shown in Figure 6. Apart from the relationship between *Battery Age* and *Battery* (exponentially increasing probability of failure with battery age), the remaining conditional probability distributions are deterministic-or's (i.e., failure in any parent causes a failure in the child).

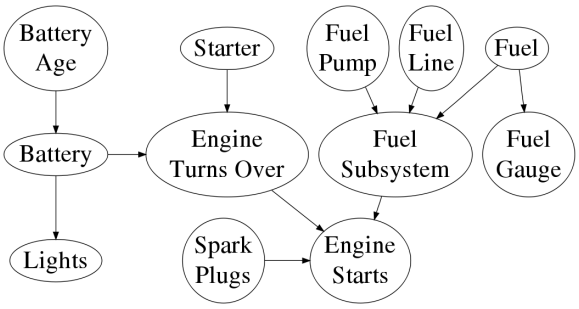

Figure 6. The vehicle fault detection Bayesian Network.

Each variable in the network can be tested, revealing whether it is operational (or the battery's age), and the *Battery, Fuel, Fuel Line, Fuel Pump, Spark Plugs*, and *Starter* can each be replaced (making it and potentially its descendants operational). Replacements and tests are both characterized *action features*: a cost to the vehicle owner, a profit for the mechanic, and a time requirement. Ideally a sequence of tests and replacements would minimize the expected cost to the vehicle owner, but an over-booked mechanic might instead choose to minimize the total repair time so that other vehicles can be serviced, and a less ethical mechanic might seek to optimize personal profit.

To generate a dataset of observations and replacements, a stochastic policy is obtained by adding Gaussian noise, $\epsilon_{s,a}$, to each action's future expected value, $Q^*(s, a)$, under the optimal policy for a fixed set of weights and selecting the highest noisy-valued action, $Q^*(s, a) + \epsilon_{s,a}$, to execute at each time-step. Different vehicle failure samples are generated from the Bayesian Network conditioned on the vehicle's engine failing to start, and the stochastic policy is sampled until the vehicle is operational.

We evaluate the prediction error rate and perplexity of our model in Figure 7. We compare against a Markov Model that ignores the underlying mechanisms for decision making and simply predicts behavior in proportion to the frequency it has previously been observed (with small pseudo-count priors). Our approach consistently outperforms the Markov Model even with an order of magnitude less training data. The classification error rate quickly reaches the limit implied by the inherent stochasticity of the data generation process.

[4] An objective function over the Bayesian Network variables can also be incorporated into a value node at each time-step and/or at the end of the sequence, as shown.

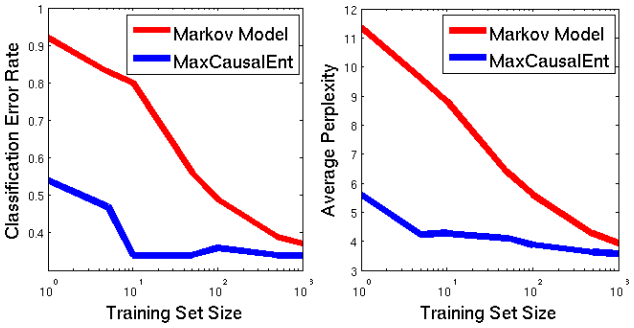

Figure 7. Error rate and perplexity of the MaxCausalEnt ID model and Markov Model for diagnosis action prediction as training set size (log-scale) increases.

5. Conclusion and Future Work

We have extended the principle of maximum entropy to settings with sequentially revealed information in this work. We demonstrated the applicability of the resulting principle of maximum causal entropy for estimating causally conditioned probability distributions in stochastic control, multi-agent interaction, and partially observable settings. In addition to further investigating modeling applications of maximum causal entropy, our future work will investigate its applicability on non-modeling tasks in dynamics settings. For instance, we note that the proposed principle provides a natural criteria for efficiently identifying a correlated equilibrium in dynamic Markov games, generalizing the approach to static games of Ortiz et al. (2007).

References

Abbeel, P., Coates, A., Quigley, M., & Ng, A. Y. (2007). An application of reinforcement learning to aerobatic helicopter flight. *NIPS* (pp. 1–8).

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. *Proc. ICML* (pp. 1–8).

Bellman, R. (1957). A Markovian decision process. *Journal of Mathematics and Mechanics*, *6*, 679–684.

Boyd, S., Ghaoui, L. E., Feron, E., & Balakrishnan, V. (1994). *Linear matrix inequalities in system and control theory*, vol. 15 of *Studies in Applied Mathematics*.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Dudík, M., & Schapire, R. E. (2006). Maximum entropy distribution estimation with generalized regularization. *Proc. COLT* (pp. 123–138).

Grünwald, P. D., & Dawid, A. P. (2003). Game theory, maximum entropy, minimum discrepancy, and robust bayesian decision theory. *Annals of Statistics*, *32*, 1367–1433.

Heckerman, D., Breese, J. S., & Rommelse, K. (1994). Troubleshooting under uncertainty. *Communications of the ACM* (pp. 121–130).

Howard, R. A., & Matheson, J. E. (1984). Influence diagrams. *Readings on the Principles and Applications of Decision Analysis* (pp. 721–762). Strategic Decisions Group.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, *106*, 620–630.

Kalman, R. (1964). When is a linear control system optimal? *Trans. ASME, J. Basic Engrg.*, *86*, 51–60.

Kramer, G. (1998). *Directed information for channels with feedback*. Doctoral dissertation, Swiss Federal Institute of Technology (ETH) Zurich.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. ICML* (pp. 282–289).

Massey, J. L. (1990). Causality, feedback and directed information. *Proc. IEEE International Symposium on Information Theory and Its Applications* (pp. 27–30).

Ortiz, L. E., Shapire, R. E., & Kakade, S. M. (2007). Maximum entropy correlated equilibrium. *AISTATS* (pp. 347–354).

Parsons, T. D. (1976). Pursuit-evasion in a graph. In *Theory and applications of graphs*, 426–441. Springer-Verlag.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Permuter, H. H., Kim, Y.-H., & Weissman, T. (2008). On directed information and gambling. *Proc. IEEE International Symposium on Information Theory* (pp. 1403–1407).

Ratliff, N., Bagnell, J. A., & Zinkevich, M. (2006). Maximum margin planning. *Proc. ICML* (pp. 729–736).

Ratliff, N. D., Silver, D., & Bagnell, J. A. (2009). Learning to search: Functional gradient techniques for imitation learning. *Auton. Robots*, *27*, 25–53.

Tatikonda, S., & Mitter, S. (2004). Control under communication constraints. *Automatic Control, IEEE Transactions on*, *49*, 1056–1068.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. (2009). Help or hinder: Bayesian models of social goal inference. *Proc. NIPS* (pp. 1874–1882).

Ziebart, B. D., Maas, A., Bagnell, J. A., & Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. *Proc. AAAI* (pp. 1433–1438).