Modeling Interaction via the Principle of Maximum Causal Entropy

Brian D. Ziebart BZIEBART@CS.CMU.EDU
J. Andrew Bagnell DBAGNELL@RI.CMU.EDU
Anind K. Dey ANIND@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

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 models for sequential data with revealed information, interaction, and feedback, and demonstrate their 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, natural language processing, econometrics, and ecology (Dudík & Schapire, 2006). Conditional extensions of the principle that consider a sequence of *side information* (i.e., additional variables that are *not* predicted, but are related to variables that *are* predicted), 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, 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 is dynamic*, i.e., revealed over time. For example, consider an agent interacting with a stochastic environment. The agent may have a model for the distribution of future states given its current state and possible actions, but, due to stochasticity, it does not know what value a future state will take until after selecting the sequence of actions temporally preceding it. Thus, future states have no *causal influence* over earlier actions. 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 present the *principle of maximum causal entropy* (MaxCausalEnt). It prescribes a probability distribution by maximizing the entropy of a sequence of variables conditioned on the side information available at each time step. This contribution extends the maximum entropy framework for statistical modeling to processes with information revelation, feedback, and interaction. We motivate and apply this approach on decision prediction tasks, where actions stochastically influence revealed side information (the state of the world) with examples from inverse optimal control, multi-player dynamic games, and interaction with partially observable systems.

Though we focus on the connection to decision making and control in this work, it is important to note that the principle of maximum causal entropy is not specific to those domains. It is a general approach that is applicable to any sequential data where future side information's assumed lack of causal influence over earlier variables is reasonable.

2. Maximum Causal Entropy

Motivated by the task of modeling decisions with elements of sequential interaction, we introduce the principle of maximum causal entropy, describe its core theoretical properties, and provide efficient algorithms for inference and learning.

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010. Copyright 2010 by the author(s)/owner(s).

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, $E_Y[-\log P(Y)]$, 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, $\mathbf{X}_{t+1:T}$, for each Y_t that does not match settings with dynamically revealed information. Despite not being originally formulated for such uses, marginalizing over the CRF's joint distribution is possible:

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

in what we refer to as a *latent CRF* model. However, we argue that entropy-based approaches like this that do not address the causal influence of side information are inadequate for interactive settings.

In the context of this paper, we focus on modeling the sequential actions of an agent interacting with a stochastic environment. Thus, we replace the predicted variables, $\mathbf{Y}$, and side information, $\mathbf{X}$, with sequences of action variables, $\mathbf{A}$, and state variables, $\mathbf{S}$.

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, $P(\mathbf{A}|\mathbf{S})$, to the situation where each A_t is conditioned on only a portion of the $\mathbf{S}$ variables, $\mathbf{S}_{1:t}$, rather than the entirety, $\mathbf{S}_{1:T}$. Following the previously developed notation (Kramer, 1998), the probability of $\mathbf{A}$ *causally conditioned* on $\mathbf{S}$ is

$$P(\mathbf{A}^T||\mathbf{S}^T) \triangleq \prod_{t=1}^{T} P(A_t|\mathbf{S}_{1:t}, \mathbf{A}_{1:t-1}). \tag{2}$$

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

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

$$\begin{aligned} H(\mathbf{A}^T||\mathbf{S}^T) &\triangleq E_{\mathbf{A},\mathbf{S}}[-\log P(\mathbf{A}^T||\mathbf{S}^T)] \qquad (3) \\ &= \sum_{t=1}^{T} H(A_t|\mathbf{S}_{1:t}, \mathbf{A}_{1:t-1}), \end{aligned}$$

measures the uncertainty present in the causally conditioned distribution. It upper bounds the conditional entropy; intuitively this reflects the fact that additionally conditioning on information from the future (i.e., acausally) only decreases uncertainty. The causal entropy has previously 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).

Using this notation, any joint distribution can be expressed as $P(\mathbf{A}, \mathbf{S}) = P(\mathbf{A}^T||\mathbf{S}^T)P(\mathbf{S}^T||\mathbf{A}^{T-1})$. Our approach estimates a *policy*—the factors, $P(A_t|\mathbf{S}_{1:t}, \mathbf{A}_{1:t-1})$, of $P(\mathbf{A}^T||\mathbf{S}^T)$—based on a provided (explicitly or implicitly) distribution of side information $P(\mathbf{S}^T||\mathbf{A}^{T-1}) = \prod_t P(S_t|\mathbf{A}_{1:t-1}, \mathbf{S}_{1:t-1})$.

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*, $\mathcal{F}(\mathbf{S}, \mathbf{A})$ with empirical expectations of those same functions, $\tilde{E}_{\mathbf{S},\mathbf{A}}[\mathcal{F}(\mathbf{S}, \mathbf{A})]$, yielding the following optimization problem:

$$\begin{aligned} &\underset{\{P(A_t|\mathbf{S}_{1:t},\mathbf{A}_{1:t-1})\}}{\text{argmax}} \; H(\mathbf{A}^T||\mathbf{S}^T) \qquad (4) \\ &\text{such that: } E_{\mathbf{S},\mathbf{A}}[\mathcal{F}(\mathbf{S}, \mathbf{A})] = \tilde{E}_{\mathbf{S},\mathbf{A}}[\mathcal{F}(\mathbf{S}, \mathbf{A})] \\ &\quad \text{and } \forall_{\mathbf{S}_{1:t},\mathbf{A}_{1:t-1}} \sum_{A_t} P(A_t|\mathbf{S}_{1:t}, \mathbf{A}_{1:t-1}) = 1, \\ &\text{and given: } P(\mathbf{S}^T||\mathbf{A}^{T-1}). \end{aligned}$$

We assume for explanatory simplicity that feature functions factor as: $\mathcal{F}(\mathbf{S}, \mathbf{A}) = \sum_t F(S_t, A_t)$, and that state dynamics are first-order Markovian, $P(\mathbf{S}^T||\mathbf{A}^{T-1}) = \prod_t P(S_t|A_{t-1}, S_{t-1})$.

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

$$\begin{aligned} P_\theta(A_t|S_t) &= \frac{Z_{A_t|S_t,\theta}}{Z_{S_t,\theta}} \qquad (5) \\ \log Z_{A_t|S_t,\theta} &= \theta^\top F(S_t, A_t) + \sum_{S_{t+1}} P(S_{t+1}|S_t, A_t) \log Z_{S_{t+1},\theta} \\ \log Z_{S_t,\theta} &= \log \sum_{A_t} Z_{A_t|S_t,\theta} = \underset{A_t}{\text{softmax}} \log Z_{A_t|S_t,\theta} \end{aligned}$$

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

Proof (sketch). The (negated) primal objective function (Equation 4) is convex in the variables $P(\mathbf{A}||\mathbf{S})$ and subject to linear constraints on feature function expectation matching, valid probability distributions, and the non-causal influence of future side information. Differentiating the Lagrangian of the maximum causal entropy optimization (Equation 4), and equating to zero, we obtain the general form:

$$P_\theta(A_t|S_t) \propto \exp\Big\{\theta^\top E_{\mathbf{S},\mathbf{A}}[\mathcal{F}(\mathbf{S},\mathbf{A})|S_t,A_t] - \sum_{\tau>t} E_{\mathbf{S},\mathbf{A}}[\log P_\theta(A_\tau|S_\tau)|S_t,A_t]\Big\}. \quad (6)$$

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 constraints 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, $P(\mathbf{S}^T||\mathbf{A}^{T-1})$. 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* $\big(\tilde{E}_{\boldsymbol{S},\boldsymbol{A}}[\mathcal{F}(\boldsymbol{S},\boldsymbol{A})] - E_{\boldsymbol{S},\boldsymbol{A}}[\mathcal{F}(\boldsymbol{S},\boldsymbol{A})]\big)$*, which is the difference between the empirical feature vector given the complete policy,* $\{P(A_t|S_t)\}$*, and the expected feature vector under the probabilistic model.*

In many instances, the statistics of interest $\big(\tilde{E}_{\mathbf{S},\mathbf{A}}[\mathcal{F}(\mathbf{S},\mathbf{A})]\big)$ for the gradient (Theorem 2) 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, i.e.,*

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

given $\tilde{P}(\boldsymbol{A},\boldsymbol{S}) = \tilde{P}(\boldsymbol{A}^T||\boldsymbol{S}^T)P(\boldsymbol{S}^T||\boldsymbol{A}^{T-1})$ *and feature expectations* $E_{\tilde{P}(\boldsymbol{S},\boldsymbol{A})}[\mathcal{F}(\boldsymbol{S},\boldsymbol{A})]$ *when* $\boldsymbol{S}$ *is sequentially revealed from a known distribution and actions are sequentially predicted using only previously revealed variables.*

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).

2.4. Inference and Learning Algorithms

The procedure for inferring decision probabilities in the MaxCausalEnt model based on Theorem 1 is illustrated by Algorithm 1.

Algorithm 1 MaxCausalEnt Inference Procedure

```
for $t = T$ to 1 do
    if $t = T$ then
        $\forall_{A_t,S_t} \log Z_{A_t|S_t,\theta} \leftarrow \theta^\top F(A_t,S_t)$
    else
        $\forall_{A_t,S_t} \log Z_{A_t|S_t,\theta} \leftarrow \theta^\top F(A_t,S_t) + E_{S_{t+1}}[\log Z_{S_{t+1},\theta}|S_t,A_t]$
    end if
    $\forall_{S_t} \log Z_{S_t,\theta} \leftarrow \operatorname{softmax}_{A_t} \log Z_{A_t|S_t}$
    $\forall_{A_t,S_t} P(A_t|S_t) \leftarrow \frac{Z_{A_t|S_t,\theta}}{Z_{S_t,\theta}}$
end for
```

Using the resulting action distributions and empirical feature functions, $\tilde{E}(\mathcal{F})$, the gradient is obtained by employing Algorithm 2.

Algorithm 2 MaxCausalEnt Gradient Calculation

```
for $t = 1$ to $T$ do
    if $t = 1$ then
        $\forall_{S_t,A_t} D_{S_t,A_t} \leftarrow P(S_t)P(A_t|S_t)$
    else
        $\forall_{S_t,A_t} D_{S_t,A_t} \leftarrow \sum_{S_{t-1},A_{t-1}} D_{S_{t-1},A_{t-1}} P(A_t|S_{t-1},A_{t-1})P(A_t|S_t)$
    end if
    $E[\mathcal{F}] \leftarrow E[\mathcal{F}] + \sum_{S_t,A_t} D_{S_t,A_t} F(A_t,S_t)$
end for
$\nabla_\theta \log P(\tilde{\mathbf{A}}||\tilde{\mathbf{S}}) \leftarrow \tilde{E}[\mathcal{F}] - E[\mathcal{F}]$
```

As a consequence of convexity, standard gradient-based optimization techniques converge to the maximum likelihood estimate of feature weights, $\hat{\theta}$.

2.5. Graphical Representation

We extend the influence diagram graphical framework (Howard & Matheson, 1984) as a convenient representation for the MaxCausalEnt variables and their relationships. The structural elements of the repre-

Table 1. Graphical representation structural elements.

Type	**Symbol**	**Parent relationship**
Decision	A	Specifies observed variables when A is selected
Uncertainty	S	Specifies conditional probability, $P(S\|\text{par}(S))$
Utility	U	Specifies feature functions, $\theta^\top F(\text{par}(U)) \to \Re$

sentation are outlined in Table 1. We constrain the graph to possess perfect recall[1]. Every graphical representation can then be reduced to the earlier causal entropy form (Equation 4) by marginalizing over each decision's non-parent uncertainty nodes to obtain side information transition dynamics and expected feature functions. Whereas in traditional influence diagrams, the parent-dependent values for decisions nodes that provide the highest expected utility are inferred, in the MaxCausalEnt setting, utilities should be interpreted as potentials for inferring a probability distribution over decisions.

3. 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.

3.1. Inverse Optimal Stochastic Control

Optimal control frameworks, such as Markov decision processes (MDPs) and linear-quadratic regulators (LQRs), provide rich representations of interactions with stochastic systems. Inverse optimal control (IOC) is the problem of recovering a cost function that makes a particular controller or policy (nearly) optimal (Kalman, 1964; Boyd et al., 1994). 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 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 ($E[\sum_t \mathbf{f}_{S_t}] = \tilde{E}[\sum_t \mathbf{f}_{S_t}]$); by linearity such behaviors achieve the same expected value. For settings with vectors of continuous states and actions, matching quadratic moments, e.g., $E[\sum_t \mathbf{s}_t \mathbf{s}_t^\top] = \tilde{E}[\sum_t \mathbf{s}_t, \mathbf{s}_t^\top]$, guarantees equivalent performance under quadratic cost functions, e.g., $\sum_t s_t^\top \mathbf{Q} s^t$ for unknown $\mathbf{Q}$.

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) will satisfy the constraint. Ziebart et al. (2008) resolve this ambiguity by using the classical maximum entropy criteria to select a single policy from all the distributions over decisions that match feature counts. However, for inverse optimal stochastic control (IOSC)—characterized by stochastic state dynamics—their proposed approach is to marginalize over future state variables (Equation 1). For IOSC, the maximum causal entropy approach provides prediction guarantees (Theorem 3) and a softened interpretation of optimal decision theory, while the latent CRF approach provides neither.

3.1.1. MDP and LQR Formulations

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 framework to model this setting, as shown in Figure 1.

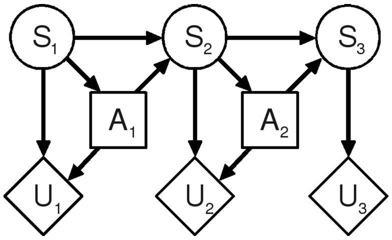

Figure 1. The graphical representation for MaxCausalEnt inverse optimal control. For MDPs: $U_t(S_t, A_t) = \theta^\top \mathbf{f}_{S_t}$, and for LQRs: $U_t(\mathbf{s}_t, \mathbf{a}_t) = \mathbf{s}_t^\top \mathbf{Q} \mathbf{s}_t + \mathbf{a}_t^\top \mathbf{R} \mathbf{a}_t$.

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

$$\begin{aligned} Q_\theta^{\text{soft}}(A_t, S_t) &= E_{S_{t+1}}[V_\theta^{\text{soft}}(S_{t+1})|S_t, A_t] \qquad (7) \\ V_\theta^{\text{soft}}(S_t) &= \underset{A_t}{\text{softmax}}\, Q_\theta^{\text{soft}}(A_t, S_t) + \theta^\top \mathbf{f}_{S_t}, \end{aligned}$$

and for the continuous, quadratic-reward setting to

$$\begin{aligned} Q_\theta^{\text{soft}}(\mathbf{a}_t, \mathbf{s}_t) &= E_{\mathbf{s}_{t+1}}[V_\theta^{\text{soft}}(\mathbf{s}_{t+1})|\mathbf{s}_t, \mathbf{a}_t] + \mathbf{a}_t^\top \mathbf{R} \mathbf{a}_t \qquad (8) \\ V_\theta^{\text{soft}}(\mathbf{s}_t) &= \underset{\mathbf{a}_t}{\text{softmax}}\, Q_\theta^{\text{soft}}(\mathbf{a}_t, \mathbf{s}_t) + \mathbf{s}_t^\top \mathbf{Q} \mathbf{s}_t. \end{aligned}$$

Note that by replacing the *softmax*[2] function with the *maximum*, this algorithm becomes equivalent to the

[1] Variables observed during previous decisions are either observed in future decisions or irrelevant (i.e., d-separated from future value nodes by observed variables).

[2] The continuous version of the softened maximum is

(stochastic) value iteration algorithm (Bellman, 1957) for finding the optimal control policy. The relative magnitudes of the action values in the MaxCausalEnt model control the amount of stochasticity in the resulting action 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, the quadratic MaxCausalEnt model permits a closed-form solution and, given dynamics $\mathbf{s}_{t+1} \sim N(\mathbf{A}\mathbf{s}_t + \mathbf{B}\mathbf{a}_t, \Sigma)$, Equation 8 reduces to:

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

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 $\mathbf{D} = \mathbf{C}_{s,s} + \mathbf{Q} - \mathbf{C}^\top\mathbf{C}_{a,a}^{-1}\mathbf{C}_{a,s}$.

3.1.2. Inverse Helicopter Control

We demonstrate the MaxCausalEnt approach to IOSC on the problem of building a controller for a helicopter with linearized stochastic dynamics. 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 (Figure 2) by noisily sampling actions from an optimal LQR controlled designed for hovering using the linearized stochastic helicopter simulator of Abbeel et al. (2007).

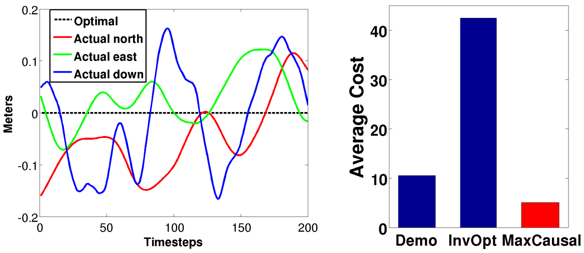

Figure 2. Left: An example sub-optimal helicopter trajectory attempting to hover around the origin point. Right: The average cost under the original cost function of: (1) demonstrated trajectories; (2) the optimal controller using the inverse optimal control model; and (3) the optimal controller using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

We contrast between the policies obtained from the maximum margin planning (Ratliff et al., 2006) (labeled **InvOpt** in Figure 2) and MaxCausalEnt models trained using demonstrated trajectories. Using the *true* cost function, we measure the cost of trajectories sampled from the optimal policy under the cost function learned by each model. The InvOpt model performs poorly because there is no optimal trajectory *for any cost function* that matches demonstrated features. In contrast, by design the MaxCausalEnt model is guaranteed to match the performance of demonstrated behavior (**Demo**) in expectation even if that behavior is sub-optimal. Additionally, because of the quadratic cost function, the optimal controller using the MaxCausalEnt cost function is always *at least as good* as the demonstrated behavior on the original, unknown cost function, and often better, as shown in Figure 2. 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.

3.2. 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 equilibria or optimal policies when the utilities of agents are known, often the utilities are *not* known and only observed behavior is available for modeling tasks. We investigate recovering the agents' utilities from those observations.

3.2.1. Markov Game Formulation

We consider a Markov game where agents act in sequence, taking turns after observing the preceding agents' actions and the resulting stochastic outcome sampled from known state dynamics, $P(S_{t+1}|A_t, S_t)$. Agents are assumed to act based on a utility function that is linear in features associated with each state, $w_i^\top \mathbf{f}_S$ and to know the other agents' utilities and policies.

Learning a single agent's MaxCausalEnt policy, π_i, given the others' policies, π_{-i}, reduces to an IOSC problem where the entropy of the agent's actions given state is maximized while matching state features:

$$\underset{\pi_i(A|S)}{\text{argmax}}\ H(\mathbf{A}^{(i)}||\mathbf{S}^{(i)}) \tag{9}$$
$$\text{such that: } E[\sum_t \mathbf{f}_i(S_t)] = \tilde{E}[\sum_t \mathbf{f}_i(S_t)]$$
$$\text{and given: } \pi_{-i}(A|S) \text{ and } P(\mathbf{S}||\mathbf{A}).$$

The distribution of side information (i.e., the agent's next state, S_{t+N} given the agent's previous state and

defined as: $\text{softmax}_{\mathbf{x}} f(\mathbf{x}) \triangleq \log \int_{\mathbf{x}} e^{f(\mathbf{x})}\, d\mathbf{x}$.

action, S_t and A_t, is obtained by marginalizing over the other agents' states and actions:

$$P(S_{t+N}|A_t,S_t) = E_{\mathbf{S}_{t+1:t+N-1},\mathbf{A}_{t+1:t+N-1}}\Big[P(S_{t+N}|S_{t+N-1},A_{t+N-1})\Big|S_t,A_t\Big].$$

Difficulties arise, however, because the policies of other agents are not known in our setting. Instead, all of the agents' utilities and policies are learned from demonstrated trajectories. Maximizing the joint causal entropy of all agents' actions leads either to non-convexity (multiplicative functions of agent policies in the feature matching constraints) or an assumption that agents share a common utility function.

We settle for potentially non-optimal solution policies by employing a cyclic coordinate descent approach. It iteratively maximizes the causal entropy of agent i's actions subject to constraints:

$$\hat{\pi}_i \leftarrow \operatorname*{argmax}_{\pi_i} F_p(\pi_i|\hat{\pi}_{-i}),$$

where F_p is the Lagrangian primal of Equation 9. However, instead of matching observed feature counts, which may be infeasible given the estimate of $\hat{\pi}_{-i}$, the expectation of agent i's features given its empirical actions under the current estimate of agent policies, $\tilde{\mathcal{F}}(\mathbf{S},\mathbf{A}) = E_{\mathbf{A},\mathbf{S}}[\sum_{S\in\mathbf{S}} \mathbf{f}_S|\hat{\pi}]$, is matched.

3.2.2. 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 (Figure 3). Each agent has a mobility, $m_i \in [0.2, 1]$, which corresponds to the probability of success when attempting to move in one of the four cardinal directions, and a utility, $w_{i,j} \in [-1, 1]$, for being co-located with each of the other agents. Unlike traditional pursuer-evader, there is no "capture" event in this game; agents continue to act after being co-located. The mobilities of each agent and a time sequence of their actions and locations are provided and the task is to predict their future actions.

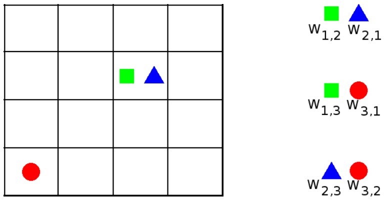

Figure 3. The pursuit-evasion grid with three agents and their co-location utilities. Agent X has a mobility of m_X and a utility of $w_{X,Y}$ when co-located with agent Y.

We generate data for this setting using the following procedure. First, mobilities and co-location utilities are sampled (uniformly from their domain). Next, optimal policies[3], $\pi_i^{t*}(A|S)$, for a range of time horizons $t \in \{T_0, ..., T_0 + \Delta T\}$, are computed with complete knowledge of other agents' utilities and future policies. A stochastic policy, $\tilde{\pi}_i(A|S)$, is obtained by first sampling a time horizon from $P(t) = \frac{1}{\Delta T}$, and then sampling an action from the optimal policy for that time horizon. Lastly, starting from random initial locations, trajectories of length 40 time steps (five for training and one for testing) are sampled from the policy and state dynamics for six different parameters. Despite its simplicity, this setting produces surprisingly rich behavior. For example, under certain optimal policies, a first evader will help its pursuer corner a more desirable second evader so that the first evader will be spared.

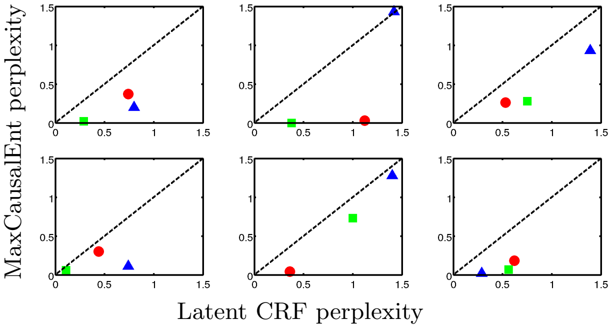

Figure 4. The average per-action perplexities of the latent CRF model and the MaxCausalEnt model plotted against each other for three agents from six different pursuit-evasion settings. The MaxCausalEnt model outperforms the latent CRF model in the region below the dotted line.

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 of predictive performance. MaxCausalEnt consistently outperforms the latent CRF model. The explanation for this is based on the observation that under the latent CRF model, the action-value of Equation 7 is instead: $Q(a_t, s_t) = \text{softmax}_{s_{t+1}}(V(s_{t+1}) + \log P(s_{t+1}|s_t, a_t))$. This has a disconcerting interpretation that the agent *chooses* the next state by "paying" an extra $\log P(s_{t+1}|s_t, a_t)$ penalty to ignore the true stochasticity of the state transition dynamics, resulting in an unrealistic probability distribution.

[3] "Ties" in action value lead to uniform distributions over the corresponding actions.

3.3. Inverse Diagnostics

Many important interaction tasks involve partial observability. In medical diagnosis, for example, sequences of symptoms, tests, and treatments are used to identify (and mediate) unknown illnesses. Motivated by the objective of learning diagnosis policies from experts, we investigate the inverse diagnostics problem of modeling interaction with partially observed systems.

3.3.1. Bayes Net Diagnosis Formulation

We consider a set of variables (distributed according to a known dynamic Bayesian network) that evolve over time based in part on employed actions, $\mathbf{A}_{1:T}$, as shown in Figure 5. Those actions are made with only partial knowledge of the Bayes net variables, as relayed through observation variables $\mathbf{O}_{1:T}$. Previous actions determine what information from the hidden variables is revealed in the next time step.

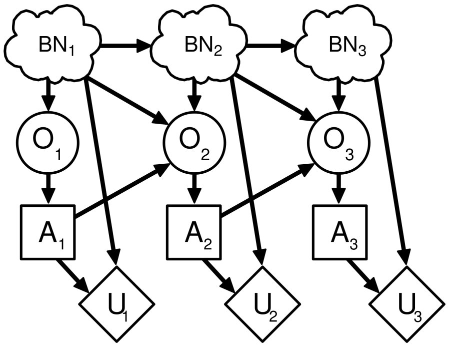

Figure 5. The MaxCausalEnt representation of the diagnostic problem with an abstract dynamic Bayesian network represented as BN nodes. Perfect recall edges from all past observations and actions to future actions are suppressed.

We assume that the utility for the state of the Bayes net and actions invoked is an unknown linear function of known feature vectors, $\theta^\top \mathbf{f}_{BN_t,A_t}$. We formulate the modeling problem as a maximum causal entropy optimization by maximizing the causally conditioned entropy of actions given observations, $H(\mathbf{A}^T||\mathbf{O}^T)$. We marginalize over the latent Bayes net variables to obtain side information dynamics, $P(O_{t+1}|O_t, S_t) = E_{BN_{1:t}}[P(O_{t+1}|\text{par}(O_{t+1}|\mathbf{O}_{1:t}, \mathbf{A}_{1:t}]$ and expected feature functions, $E[\mathbf{f}_t|\mathbf{O}_{1:t}, \mathbf{A}_{1:t}] = E_{BN_{1:t}}[F_{U_t}(\text{par}(U_t))|\mathbf{O}_{1:t}, \mathbf{A}_{1:t}]$ to estimate the full-history policy, $P(A_t|\mathbf{O}_{1:t}, \mathbf{A}_{1:t})$ using Algorithm 1 and Algorithm 2.

3.3.2. Vehicle Diagnosis Experiments

We apply our inverse diagnostics approach to the vehicle fault detection Bayesian network (Heckerman et al., 1994) shown in Figure 6 with fully specified conditional probability distributions. 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).

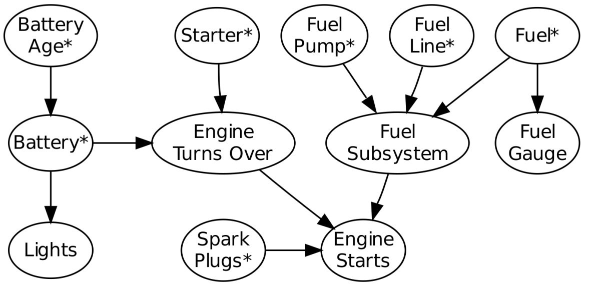

Figure 6. The vehicle fault detection Bayesian network with replaceable variables denoted with an asterisk. All variables are binary (working or not) except Battery Age.

A mechanic can either test a component of the vehicle (revealing its status) or repair a component (making it and potentially its descendants operational). Replacements and tests are both characterized by three action features: (1) a cost to the vehicle owner; (2) a profit for the mechanic; and (3) a time requirement. Ideally the sequence of mechanic's actions would minimize the expected cost to the vehicle owner, but an over-booked mechanic might instead choose to minimize the total repair time, 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 feature weights and the highest noisy-valued action, $Q^*(s, a) + \epsilon_{s,a}$, is selected 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.

In Figure 7, we evaluate the prediction error rate and perplexity of our model and 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). The MaxCausalEnt 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 stochasticity of the data generation process.

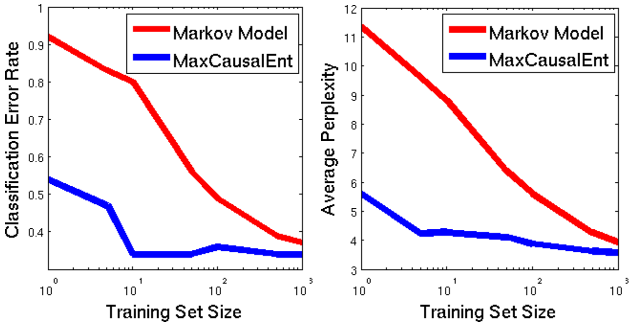

Figure 7. Error rate and perplexity of the MaxCausalEnt model and Markov model for diagnosis action prediction as training set size (log-scale) increases.

4. Conclusion and Future Work

We 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 learning policies in stochastic control, multi-agent interaction, and partially observable settings. In addition to further investigating modeling applications, our future work will investigate the applicability of MaxCausalEnt 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 normal-form games of Ortiz et al. (2007).

Acknowledgments

The authors gratefully acknowledge the Richard King Mellon Foundation, the Quality of Life Technology Center, and the Office of Naval Research MURI N00014-09-1-1052 for support of this research. We also thank our reviewers for many useful suggestions.

References

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

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

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

Boyd, S. and Vandenberghe, L. *Convex Optimization.* Cambridge University Press, 2004.

Boyd, S., Ghaoui, L. El, Feron, E., and Balakrishnan, V. *Linear Matrix Inequalities in System and Control Theory*, volume 15 of *Studies in Applied Mathematics.* 1994.

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

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

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

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

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

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

Kramer, G. *Directed Information for Channels with Feedback.* PhD thesis, Swiss Federal Institute of Technology (ETH) Zurich, 1998.

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

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

Ortiz, L. E., Shapire, R. E., and Kakade, S. M. Maximum entropy correlated equilibria. In *AISTATS*, pp. 347–354, 2007.

Parsons, T. D. Pursuit-evasion in a graph. In *Theory and Applications of Graphs*, pp. 426–441. Springer-Verlag, 1976.

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

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

Ratliff, N., Silver, D., and Bagnell, J. A. Learning to search: Functional gradient techniques for imitation learning. *Auton. Robots*, 27(1):25–53, 2009.

Tatikonda, S. and Mitter, S. Control under communication constraints. *Automatic Control, IEEE Transactions on*, 49(7):1056–1068, July 2004. ISSN 0018-9286. doi: 10.1109/TAC.2004.831187.

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

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