Maximum Causal Entropy Correlated Equilibria for Markov Games

Brian D. Ziebart, J. Andrew Bagnell, Anind K. Dey
Carnegie Mellon University
bziebart@cs.cmu.edu, dbagnell@ri.cmu.edu, anind@cs.cmu.edu

ABSTRACT

Motivated by a machine learning perspective—that game-theoretic equilibria constraints should serve as guidelines for predicting agents' strategies, we introduce maximum causal entropy correlated equilibria (MCECE), a novel solution concept for general-sum Markov games. In line with this perspective, a MCECE strategy profile is a uniquely-defined joint probability distribution over actions for each game state that minimizes the worst-case prediction of agents' actions under log-loss. Equivalently, it maximizes the worst-case growth rate for gambling on the sequences of agents' joint actions under uniform odds. We present a convex optimization technique for obtaining MCECE strategy profiles that resembles value iteration in finite-horizon games. We assess the predictive benefits of our approach by predicting the strategies generated by previously proposed correlated equilibria solution concepts, and compare against those previous approaches on that same prediction task.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Theory, Algorithms

Keywords

Game theory, correlated equilibria, maximum entropy

1. INTRODUCTION

Agents often need to predict the future behavior of other agents [9] to appropriately choose their own actions. Equilibria solution concepts, such as Nash equilibria [22], and the more general correlated equilibria (CE) [1], which allow agents to coordinate their actions, are important constructs for multi-agent games that provide certain individual or group performance guarantees based on assumed rationality. Agents playing many decentralized, adaptive strategies (such as no-regret learning) will converge to CE [23, 10, 14, 11], but the particular set of convergence CE will vary depending on the strategies employed. From an applied machine learning perspective, existing equilibria concepts are often not useful for prediction. First, they generally do not fully specify a unique strategy profile, making strategy prediction under-specified without additional assumptions. Second, they are typically not designed to provide any predictive performance guarantees.

We introduce the maximum causal entropy correlated equilibria (MCECE) solution concept to enable equilibria-based prediction for general-sum Markov games. It extends maximum entropy correlated equilibria (MaxEntCE) for normal-form games [24] to the dynamic game setting by specifying the unique CE strategy profile with the *fewest* additional assumptions. This property is useful for three main purposes. First, for prescriptive settings, the resulting MCECE strategy profile best conceals the underlying motives of agents in a manner we specify in Section 3. This can often be an important consideration when revealing too much information can lead to future exploitation. Second, for predictive purposes, the MCECE strategy profile minimizes the worst-case log-loss when predicting the actions of agents assumed to act according to an unknown CE strategy. Thus, it is theoretically justified for predicting the actions of agents assumed to be jointly behaving rationally. Third, for gambling on the sequence of agents' actions, the MCECE strategy profile maximizes the worst-case expected investment growth rate under uniform odds.

We present in Section 4 an efficient algorithm for obtaining MCECE based on convex optimization that ultimately reduces to a dynamic programming algorithm over time steps of finite games. In contrast with our predictive approach, previously developed CE solution concepts impose very strong assumptions on agents' preference over possible CE strategy profiles to provide unique payoffs [20, 15, 12]. In Section 5, we evaluate the predictive benefits of the MCECE and other strategy profiles at predicting the strategies of one another.

2. BACKGROUND

We first review concepts in game theory and information theory to properly situate the contributions of this paper.

2.1 Games and Equilibria

The canonical set of games studied within game theory are one-shot games with matrices of payoffs.

Definition 1. *A* ***normal-form game****, is defined by a set of agents N, a set of joint agent actions A, and a utility vector $U : A \mapsto \mathbb{R}^N$, specifying the* payoffs *for each agent $i \in N$ for joint action $a \in A$. Each agent controls a portion $a_i \in A_i$ of the joint action $a = \times_{i \in N} a_i$.*

Cite as: Maximum Causal Entropy Correlated Equilibria for Markov Games, Brian D. Ziebart, J. Andrew Bagnell and Anind K. Dey, *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, Tumer, Yolum, Sonenberg and Stone (eds.), May, 2–6, 2011, Taipei, Taiwan, pp. XXX-XXX.

In a normal-form game (*Definition* 1), each agent ($i \in N$) simultaneously selects an action ($a_i \in A_i$) and receives a numerical **payoff**, $U_{a,i} \in \mathbb{R}$, based on the combination of actions, $a \in A$.

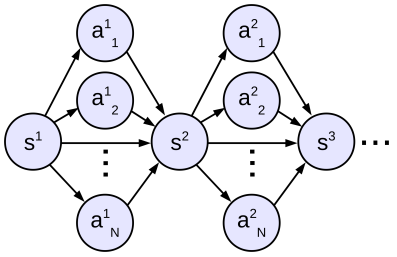

Figure 1: The sequence of states and (Markovian) actions of a Markov game.

Markov games (*Definition* 2) generalize normal-form games to sequential settings. In a Markov game, the joint actions of N agents at time t, denoted a^t, stochastically lead to a next state as shown in Figure 1.

Definition 2. *A* ***Markov game*** *is defined by a set of states (S) representing the joint states of N agents, a set of actions (A), a probabilistic state transition function, $T : S \times A \mapsto \Delta_S$, and a utility function, $Utility_i : N \times S \times A \mapsto \mathbb{R}$.*

Agents choose **strategy profiles**, $\pi \in \Delta_A$, specifying next actions for each situation that are either **mixed** (*i.e.*, stochastic) or **pure** (*i.e.*, deterministic); and either **correlated** (*i.e.*, joint functions) or **independent** based on a (discounted, $0 < \gamma \leq 1$) cumulative expected utility:

$$\begin{aligned} &\text{ExpUtil}_i^{\pi}(a^t, s^t) \qquad (1)\\ &\triangleq \mathbb{E}_{S^{t+1:T}, A^{t+1:T}}\left[\sum_{\tau \geq t} \gamma^{\tau} \text{Utility}_i(s^{\tau}, a^{\tau}) \middle| a^t, s^t, \pi\right], \end{aligned}$$

where we denote the variables being marginalized over in the expectation using subscript. We assume in Equation 1 and throughout this paper that the strategy profile is mixed and **Markovian**[1], meaning it depends only on the current state and time step.

To obtain strategy profiles, it is useful to consider the amount of utility gained by switching from a provided action, a_i^t, to an alternate action, $a_i^{t'}$, called a **deviation action**, when: all agents' actions, a^t, are known (Equation 2); or when other agents' actions, denoted $a_{-i}^t \in A_{-i}^t$, are unknown and averaged over according to the strategy profile, π (Equation 3):

$$\begin{aligned} &\text{ExpDevGain}_i^{\pi}(a^t, s^t, a_i^{t'}) \qquad (2)\\ &\quad \triangleq \text{ExpUtil}_i^{\pi}(\{a_{-i}^t, a_i^{t'}\}, s^t) - \text{ExpUtil}_i^{\pi}(a^t, s^t)\\ &\text{ExpRegret}_i^{\pi}(a_i^t, a_i^{t'}, s^t) \qquad (3)\\ &\quad \triangleq \mathbb{E}_{A_{-i}^t}\left[\text{ExpDevGain}_i^{\pi}(a^t, s^t, a_i^{t'}) \middle| a_i^t, s^t\right]. \end{aligned}$$

Definition 3. *A* ***correlated equilibrium*** *(CE) for a Markov game is a mixed joint strategy profile, π^{CE}, where no expected gain is obtained for any agent by substituting an action, $a_i^{t'}$ that deviates from the strategy. This is guaranteed with the following set of constraints:*

$$\forall_{t \in T, i \in N, s^t \in S, a_i^t \in S, a_i^{t'} \in S} \; ExpRegret_i^{\pi^{CE}}(a_i^t, a_i^{t'}, s^t) \leq 0. \quad (4)$$

CE (*Definition* 3) generalize **Nash equilibria** [22], which further require agents' actions in each state to be independent. Agents in a CE can coordinate their actions to obtain higher expected utilities. Conceptually, each agent is provided an action, a_i^t, and knows the conditional distribution of other agents' actions, $P(a_{-i}^t | a_i^t)$. To be in correlated equilibrium requires that no agent has an incentive to switch from action a_i^t to a deviation action, $a_i^{t'}$. 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 its prescribed action (stop on red) rather than deviating (go on red). this coordination mechanism is not required as long as players have access to a public communications channel [6]. Past research has shown that many decentralized, adaptive strategies will converge to a CE [23, 10, 14, 11], and not necessarily to more restrictive equilibria, such as the Nash equilibrium.

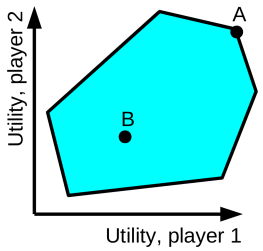

Figure 2: A CE polytope with a CE-Q equilibria (point A) maximizing average utility and a MaxEntCE (point B).

The deviation regret constraints (Equation 4) define an N-dimensional convex polytope of CE solutions in the space of agents' joint utility payoffs (Figure 2). 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 [21, 5], but tractable applicability has been limited to small games (15 or fewer joint actions combinations per state) [5]. 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 [25, 18] and extended to sequential games [16].

Our objective is different; we desire neither the (approximate) entire convex polytope of CE strategy profiles nor an arbitrary CE strategy profile. Rather, we desire a single CE strategy profile with certain properties that are useful for predictive purposes. This can be approached using optimization techniques. For a single-shot (*i.e.*, **normal-form**) game, there are $O(N|A_i|^2)$ regret constraints that are linear in a total of $O(|A|)$ strategy variables, $\{\pi(a)\}_{a \in A}$, and CE solutions can be efficiently obtained by solving a linear

[1]Markovian strategy profiles are a consequence of the MCECE formulation and commonly assumed in other solution concept formulations.

program or a convex program:

$$\max_{\pi} f_0(\pi(A)) \text{ such that:} \quad (5)$$
$$\forall_{i,a_i,a_i'} \sum_{a_{-i}} \pi(a)\big(\text{Utility}\left(\{a_{-i}, a_i'\}\right)\text{–Utility}(a)\big) \leq 0,$$
$$\forall_a \pi(a) \geq 0, \text{ and } \sum_a \pi(a) = 1.$$

depending on the objective function, f_0.

A **correlated-Q equilibria** (CE-Q) [12] employs a linear or convex function of strategy probabilities for the selection metric objective of Equation 5 to obtain utility-unique strategy profiles[2]. A number of objectives have been proposed:

- **Utilitarian** (uCE-Q) maximizes the sum of agents' utilities, $\sum_{i=1}^{N} \mathbb{E}[\text{Utility}_i(a)|\pi]$;
- **Dictatorial** (dCE-Q) maximizes a specific agent's utility, $\mathbb{E}[\text{Utility}_i(a)|\pi]$;
- **Republican** (rCE-Q) maximizes the highest agent's utility, $\max_i \mathbb{E}[\text{Utility}_i(a)|\pi]$; and
- **Egalitarian** (eCE-Q) maximizes lowest agent's utility, $\min_i \mathbb{E}[\text{Utility}_i(a)|\pi]$.

More generally, strategies that penalize one or more agents are also possible. For example, grim-trigger strategies have been recognized as viable sub-game strategies that disincentivize an agent's undesirable actions. Two punishment-based selection criteria that we consider in this work are:

- **Disciplinarian** (xCE-Q) minimizes a specific agent's utility, $\mathbb{E}[\text{Utility}_i(a)|\pi]$; and
- **Inegalitarian** (iCE-Q) maximizes utility differences between two (groups of) agents, $\mathbb{E}[\text{Utility}_i(a) - \text{Utility}_j(a)|\pi]$.

The strong assumptions about agents' preferences constrain CE-Q solutions to cover corners of the CE polytope (Figure 2).

2.2 Entropy, Prediction, and Gambling

Information theory provides powerful tools for constructing predictive probability distributions. One of its basic measures is **Shannon's information entropy**, $H(P) \triangleq -\sum_{x \in X} P(x) \log_2 P(x)$, which measures the uncertainty of distribution P. Information theory has many connections to problems in gambling. For example, the entropy of distribution P, and the exponential rate at which a gambler who knows P can expect his investment to grow are related by Theorem 4.

THEOREM 4 ([4]). *The **doubling rate**, which specifies the wealth growth rate, $O(2^{W(P,b)})$, for random outcomes distributed according to P with bets in proportion to b and payoff multipliers, o, such that $\forall_x o(x) \geq 1$ and $\sum_x o(x)^{-1} = 1$, is:*

$$W(P, b) = \sum_{x \in X} P(x) \log(b(x)o(x)).$$

It is maximized by $b(x)^* = P(x)$ for uniform odds and provides an optimal doubling rate, $W^*(P) = \log |X| - H(P)$.

[2]Unique strategy profiles are not guaranteed by the CE-Q solution concept—multiple actions can provide the same agent utility vector. We ignore this ambiguity and employ a single CE-Q from the possible set.

More generally, a gambler (or predictor) may not know the distribution P, but instead knows some constraints that P satisfies. For example, linear equality constraints, $g(P) = 0$, and inequality constraints, $h(P) \leq 0$, are common.

Definition 5. *The **principle of maximum entropy** [17] prescribes the **maximum entropy probability distribution** subject to equality and inequality constraints:*

$$\operatorname*{argmax}_P H(P) \text{ such that: } g(P) = 0 \text{ and } h(P) \leq 0.$$

The maximum entropy distribution (*Definition* 5) provides important predictive guarantees (Theorem 6).

THEOREM 6 ([13]). *The **maximum entropy distribution** minimizes the worst case predictive **log-loss**,*

$$\inf_{P(X)} \sup_{\tilde{P}(X)} -\sum_{x \in X} \tilde{P}(x) \log P(x),$$

subject to constraints $g(P) = 0$ and $h(P) \leq 0$.

Additionally, the gambling asset allocation that maximizes the worst-case growth-rate for this setting is:

$$b(X)^* = \operatorname*{argmax}_{b(X)} \min_{P(\mathbf{X})} W(\mathbf{X}) \quad (6)$$

subject to equality and inequality constraints.

COROLLARY 7 (THEOREM 6 *and* THEOREM 4). *The optimal gambling asset allocation, $b(X)^*$, is proportional to the maximum entropy distribution when the payoff multipliers are uniform.*

2.3 Maximum Entropy Correlated Equilibria

The **maximum entropy correlated equilibria** (MaxEntCE) solution concept for normal-form games [24] selects the unique joint strategy profile that satisfies the principle of maximum entropy (*Definition* 5) subject to linear deviation regret inequality constraints (Equation 4). This approach provides the predictive and gambling guarantees of maximum entropy (THEOREM 6 and COROLLARY 7) to the one-shot, normal-form multi-agent game setting.

Table 1: 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 1. We relate these strategy profiles to the more specific equilibria described in Section 2.1. *CE 1* and *CE 2* are both dictatorial, disciplinarian, and inegalitarian CE (for different agents) and republican CE (but ambiguous). *CE 3* is a utilitarian CE and an egalitarian CE. *CE 4* is the maximum entropy CE. Its predictive guarantee is apparent: all other CE have infinite log-loss for 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 setting in this work.

3. MAXIMUM CAUSAL ENTROPY CORRELATED EQUILIBRIA

Extension of the MaxEntCE solution concept [24] to the Markov game setting is not straight-forward. The first difficulty is that the deviation regret constraints of normal-form games (Equation 5), contain expectations over future actions (Equation 4) when extended to the Markov game setting. This creates non-linear constraints that are products of the unknown variables, making optimization difficult.

THEOREM 8. *A linear/convex program formulation of CE for Markov games is possible by considering as variables the entire sequence of joint agent actions for the sequence of revealed states, $\eta(A^{1:T}|S^{1:T})$, and employing appropriate inequality constraints (deviation regret guarantees) and equality constraints (forcing the strategy over sequences to factor into products of Markovian strategies) on marginal distributions using linear function of $\eta(A^{1:T}|S^{1:T})$ variables.*

Naïvely formulating the Markov game CE strategy profiles into a linear/convex program is possible (THEOREM 8), but the number of constraints and variables grow exponentially with the time horizon.

The second difficulty is that there are many entropy measures that could be applied as objective functions. For example, the **conditional entropy** and **joint entropy** are natural entropy measures to consider. However, neither appropriately extends the predictive and gambling guarantees of the maximum entropy approach to the sequential Markov game setting. They either assume the availability of future outcome information (violating the problem setting), or are not risk-neutral to the stochasticity of the Markov game's transition dynamics.

We instead advocate the less common **causally conditioned entropy** measure [19],

$$H(\mathbf{A}^T||\mathbf{S}^T) \triangleq \sum_t H(A^t|A^{1:t-1}, S^{1:t}). \tag{7}$$

For the possible sequences of states and actions through a Markov game, it corresponds to the uncertainty associated with only the actions in such sequences. It is based on the **causally conditioned probability distribution**, $P(\mathbf{A}^T||\mathbf{S}^T) \triangleq \prod_t P(A^t|A^{1:t-1}, S^{1:t})$, which conditions each set of correlated actions only on actions and states that have been revealed at that point in time and not on future states, as in the conditional probability distribution $P(\mathbf{A}|\mathbf{S}) = \prod_t P(A^t|A^{1:t-1}, S^{1:t}, S^{t+1:T})$.

Definition 9. *A **maximum causal entropy correlated equilibrium** (MCECE) solution maximizes the causal entropy while being constrained to have no action deviation regrets[3]:*

$$\pi^{MCECE} \triangleq \underset{\pi}{\operatorname{argmax}}\, H(\boldsymbol{A}^T||\boldsymbol{S}^T) \tag{8}$$
$$= \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}_{A^{1:T},S^{1:T}}\left[\sum_{t=1}^{T} -\log P(a^t|s^t)\right]$$
$$\text{such that: } \forall_{t,\, i,\, a_i^t,\, a_i^{t\prime},\, s^t}\ \ ExpRegret_i^{\pi}(a_i^t,\ a_i^{t\prime},\ s^t) \le 0,$$
$$\forall_{t,\, s^t,\, a^t}\ P(a^t|s^t) \ge 0,\ \ \forall_{t,\, s^t}\ \sum_{a^t} P(a^t|s^t) = 1,$$
$$\pi \text{ factors as: } P(\boldsymbol{A}^T||\boldsymbol{S}^T) \text{ and given: } P(S^{t+1}|S^t, A^t).$$

*We further constrain the strategy profile to have **sub-game equilibria**, meaning that even in states that are unreachable under the strategy profile and state dynamics, the strategy profile is constrained to satisfy Equation 8 in all sub-games starting from those states.*

Maximizing the causal entropy (Equation 8) has been previously employed to match characteristics of demonstrated behavior in decision settings using equality constraints [26]. MCECE represents the first inequality-constrained application of the principle of maximum causal entropy

Based on the view of conditional entropy as a measure of predictability [4], the MCECE solution concept offers two important predictive guarantees:

THEOREM 10 (*extension of* [24]). *Given an MCECE strategy profile, no agent may decrease the predictability of her action sequence without creating deviation regret for herself.*

THEOREM 11 (*extension of* [13]). *The MCECE solution strategy profile, π^{MCECE} minimizes the worst-case log prediction loss for the sequences of joint actions, i.e.,*

$$\inf_{P(\boldsymbol{A}^T||\boldsymbol{S}^T)} \sup_{\tilde{P}(\boldsymbol{A}^T||\boldsymbol{S}^T)} -\sum_{\boldsymbol{a}\in\boldsymbol{A},\boldsymbol{s}\in\boldsymbol{S}} \tilde{P}(\boldsymbol{a},\boldsymbol{s}) \log P(\boldsymbol{a}^T||\boldsymbol{s}^T), \tag{9}$$

of all the CE satisfying deviation regret constraints, where $\tilde{P}(\boldsymbol{A}^T||\boldsymbol{S}^T)$ is the (worst possible for prediction) empirical CE strategy and the joint, $\tilde{P}(\boldsymbol{A},\boldsymbol{S})$, is the distribution of states and actions under that strategy profile and the known state transition dynamics.

The second result (THEOREM 11) is particularly relevant to our machine learning perspective, because it justifies the MCECE strategy profile as a robust predictive model of agents' actions when they jointly behave rationally.

4. CORRELATED EQUILIBRIA FINDING

We turn our attention from the theoretical properties of the MCECE solution concept to developing an algorithm that obtains the MCECE for a fixed-horizon Markov game more efficiently than the naïve convex optimization that follows the formulation of Theorem 8. Despite the non-compact formulation of the näive MCECE convex program, the strategy profile can be expressed compactly.

LEMMA 12. *The MCECE strategy profile for a Markov game is also Markovian.*

Proof (sketch). Intuitively, independence maximizes entropy. Since the utility structure and game dynamics are history-independent, nothing prevents the MCECE from also being history-independent. □

THEOREM 13. *The MCECE strategy profile, $\pi_\lambda^{MCECE}(a^t|s^t)$, has the following recursive form (with $\lambda \ge 0$):*

$$\pi_\lambda(a^t|s^t) \propto e^{-\left(\sum_{i,a_i^{t\prime}} \lambda_{i,s^t,a_i^t,a_i^{t\prime}}\ ExpDevGain_i^{\pi}\left(a^t,s^t,a_i^{t\prime}\right)\right)} \tag{10}$$
$$+ExpEnt\left(a^t,s^t\right),$$

[3]Markovian policies are a consequence of the MCECE formulation. See Lemma 13.

where $ExpEnt(a^t, s^t) \triangleq$
$\mathbb{E}_{P(A^{t+1},S^{t+1})}\left[ExpEnt(a^{t+1}, s^{t+1}) + H(a^{t+1}|s^{t+1})|a^t, s^t\right]$.

We obtain optimal Lagrange variables, $\{\lambda^*_{i,s,a_i,a_i'}\} \geq 0$, by optimizing the Lagrange dual,

$$L_D(\lambda) = \mathbb{E}_{A^{1:T},S^{1:T}}\Big[-\log P_\lambda(a^{1:T}||s^{1:T})\Big] - \sum_{t,i,s^t,a_i^t,a_i^{t'}} \lambda_{t,i,s^t,a_i^t,a_i^{t'}} \text{ExpRegret}_i^{\pi_\lambda}(a_i^t, a_i^{t'}, s^t), \quad (11)$$

using gradient-based optimization and the dual's gradient,

$$\nabla_\lambda L_D(\lambda) = \left\{ \sum_{a_{-i}^t} P(a^t|s^t)\Big(\text{ExpUtil}_i^{\pi_\lambda}(\{a_{-i}^t, a_i^{t'}\}, s^t) - \text{ExpUtil}_i^{\pi_\lambda}(a^t, s^t)\Big)\right\}, \quad (12)$$

as shown in Algorithm 1.

Algorithm 1 Find MCECE equilibria for finite horizon

1: $\lambda^{(1)} = \{\lambda^{(1)}_{t,i,s^t,a_i^t,a_i^{t'}}\} \leftarrow$ (arbitrary) positive initial values.
2: $x \leftarrow 1$
3: **while** not converged **do**
4: Compute $\pi_\lambda^{(x)} = \{\pi_\lambda(a^t|s^t)\}$ from $\lambda^{(x)}$ using a subroutine.
5: Compute $L_D(\lambda^{(x)})$ and $\nabla_\lambda L_D(\lambda^{(x)})$ directly via Equation 11 and Equation 12 using $\pi_\lambda^{(x)}$
6: Update $\lambda^{(x+1)}$ from $\{\lambda,\ L_D,\ \nabla_\lambda L_D\}^{(1:x)}$ using gradient-based optimization update rules
7: $x \leftarrow x + 1$
8: **end while**

Remark 14. *For time-varying policies, future strategy probabilities (and dual parameters) are independent of earlier strategy and dual parameters given the state. As a result, the "parallel" updates for dual parameters across time (Algorithm 1) can be sequentially ordered for improved efficiency.*

Following Remark 14, Algorithm 1 can be re-expressed as a sequential dynamic programming algorithm (Algorithm 2) resembling value iteration [2] that iteratively computes both future expected utilities and expected entropies. It also suggests the parallel updating of the dual parameters (or the primal policy) as a general approach for overcoming the limitations of value iteration for finding stationary CE strategy profiles in general-sum Markov games. However, full discussion is beyond this paper's scope.

Using interior-point methods, an $\epsilon-$optimal MCECE strategy profile is obtained in $O(|S|^{\frac{T}{2}}|A|^{\frac{T}{2}}\log\frac{1}{\epsilon})$ time using the naïve formulation of Theorem 8. Using Algorithm 2, this is reduced to $O(|S|T|A|^{\frac{1}{2}}\log\frac{|S|T}{2\epsilon})$ time.

We employ existing sub-gradient optimization methods [3] for the convex objective and linear inequality constraints to obtain the strategy profiles for the interior optimization (Line 4) of Algorithm 2. This provides looser runtime bounds than interior-point optimization methods, but is simpler to implement and still practical for the purposes of this paper.

Algorithm 2 Value iteration approach for obtaining MCECE

1: $\forall_{i,a,s}\ \text{ExpUtil}_i(a, s) \leftarrow \text{Utility}_i(a, s)$
2: $\forall_{a,s}\ \text{ExpEnt}(a, s) \leftarrow 0$
3: **for** $t = T$ to 1 **do**
4: For each state, s^t, obtain $\{\pi_\lambda(a^t|s^t)\}$ using ExpUtil and ExpEnt values in the following optimization:

$$\underset{\pi(a^t|s^t)}{\text{argmax}}\ H(a^t|s^t) + \mathbb{E}\left[\text{ExpEnt}(a, s)|s^t, \pi(a^t|s^t)\right]$$

$$\text{such that: } \sum_{a_{-i}\in A_{-i}} P(a^t|s^t)\Big(\text{ExpRegret}(\{a_{-i}^t, a_i^t\}, s^t) - \text{ExpRegret}(a^t, s^t)\Big) \leq 0$$

$$\forall_{a^t} P(a^t|s^t) \geq 0 \text{ and } \sum_{a^t\in A^t} P(a^t|s^t) = 1.$$

5: $\forall_{i\in N,a\in A,s\in S}\ \text{ExpUtil}_i'(a, s) \leftarrow \gamma\sum_{a^t\in A,s^t\in S}\pi(a^t|s)P(s^t|s, a)\ \text{ExpUtil}_i(a^t, s^t)$
6: $\forall_{s\in S,a\in A}\ \text{ExpEnt}'(a, s) \leftarrow \gamma\sum_{a^t,s^t}\pi(a^t|s^t)P(s^t|s, a)\big(\text{ExpEnt}(a', s') + H(a'|s')\big)$
7: $\forall_{i\in N,a\in A,s\in S}\ \text{ExpUtil}_i(a, s) \leftarrow \text{ExpUtil}_i'(a, s) + \text{Utility}_i(a, s)$
8: $\forall_{a\in A,s\in S}\ \text{ExpEnt}(a, s) \leftarrow \text{ExpEnt}'(a, s)$
9: **end for**

5. EXPERIMENTAL EVALUATION

In this section, we demonstrate that the theoretical robust predictive guarantees of the MCECE are realized in practice. Following Zinkevich et al. [27], we generate random Markov games for evaluation. We compute different strategy profiles for each generated game using existing CE solution concepts and evaluate how well they predict one another.

5.1 Setup

We generate random stochastic Markov games according to the following procedure. For each of $|S|$ states in the Markov game, each agent has $|A_i|$ actions from which to choose, and there are $|A|$ joint actions total. The state transition dynamics, $P(S_{t+1}|S_t, A_t)$, depend on the combination of agents' actions (and state) and are drawn uniformly from the simplex of probabilities. The utility obtained by each agent in each state, $\text{Utility}_i(s)$, is drawn uniformly from $\{0, 0.1, 0.2, ..., 0.9\}$. A discount factor of $\gamma = .75$ is incorporated in each game. It's important to note that we did not optimize these random game parameters to obtain desired results; we expect the results for the games we evaluate to extend to a wide range of games—random or otherwise.

We generate time-varying strategy profiles for MCECE using Algorithm 2 and for the CE-Q variants using projected sub-gradient optimization. The CE-Q strategies we evaluate are a subset of those described in Section 2.1. iC-EQ maximizes the positive margin of agent 1's utility over agent 2's utility. We repeat this process for 100 random games for each choice of game parameters and investigate the properties of the resulting CE strategy profiles.

As shown in Figure 3, the uncertainty of action sequences increases linearly with the size of the action set, as one might expect. Previous experiments have primarily considered two-player Markov games [12, 27]. The larger number of players we consider in this paper greatly increases the game complexity since the game description grows exponen-

Table 2: Predictive bake-off evaluation of the first action in a ten timestep horizon using 100 random Markov games. Log-loss and non-support measures (equivalent to total gambling loss) are evaluated.

Seven equilibria strategy profiles evaluated on random Markov games with **three** agents, two states, and two actions/agent.

	MCECE	uCE-Q	d_1CE-Q	d_2CE-Q	x_1CE-Q	x_2CE-Q	iCE-Q	Average
MCECE	———	1.951 0.0%	1.951 0.0%	1.967 0.0%	**2.010** 0.0%	**1.992** 0.0%	1.974 0.0%	**1.974** 0.0%
uCE-Q	3.377 22.3%	———	2.039 4.5%	**1.888** 5.7%	2.647 21.2%	3.072 20.8%	2.444 13.8%	2.578 14.7%
d_1CE-Q	3.442 18.9%	**1.866** 3.5%	———	2.511 5.6%	3.328 18.8%	2.321 17.6%	**1.798** 9.8%	2.544 12.4%
d_2CE-Q	3.462 17.0%	1.872 1.7%	2.536 3.3%	———	2.576 15.6%	3.489 17.2%	3.060 12.1%	2.833 11.2%
x_1CE-Q	2.897 0.2%	2.472 0.0%	2.798 0.0%	2.450 0.0%	———	2.375 0.0%	2.764 0.0%	2.626 0.0%
x_2CE-Q	**2.877** 0.5%	2.605 0.0%	2.251 0.0%	2.905 0.0%	2.373 0.2%	———	2.116 0.0%	2.521 0.1%
iCE-Q	3.378 5.3%	2.279 1.9%	**1.902** 0.0%	2.989 2.5%	3.116 5.1%	2.276 4.2%	———	2.657 3.2%

Seven equilibria strategy profiles evaluated on random Markov games with **four** agents, two states, and two actions/agent.

	MCECE	uCE-Q	d_1CE-Q	d_2CE-Q	x_1CE-Q	x_2CE-Q	iCE-Q	Average
MCECE	———	3.451 0.0%	**3.468** 0.0%	**3.476** 0.0%	**3.518** 0.0%	**3.509** 0.0%	3.475 0.0%	**3.483** 0.0%
uCE-Q	7.308 25.8%	———	3.955 9.4%	3.652 8.5%	6.495 21.8%	6.814 22.0%	5.591 17.9%	5.636 17.6%
d_1CE-Q	6.914 22.5%	3.831 5.8%	———	4.746 10.2%	7.566 21.2%	5.536 17.6%	3.895 11.2%	5.415 14.8%
d_2CE-Q	7.109 23.1%	**3.408** 6.4%	4.767 10.7%	———	5.643 18.3%	7.651 21.5%	6.976 19.7%	5.926 16.6%
x_1CE-Q	**4.603** 0.0%	4.300 0.0%	5.000 0.0%	3.849 0.0%	———	3.918 0.0%	4.372 0.0%	4.340 0.0%
x_2CE-Q	4.633 0.1%	4.401 0.0%	3.917 0.0%	4.986 0.0%	3.944 0.0%	———	**3.171** 0.0%	4.175 0.0%
iCE-Q	6.380 11.5%	5.070 5.0%	3.884 2.8%	6.501 8.0%	5.991 8.7%	4.311 5.5%	———	5.356 6.9%

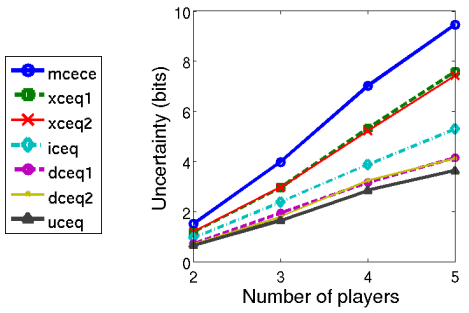

Figure 3: The average causal entropy measure of the 10 time step action sequences that are generated by different correlated equilibria solution concepts for random 2-state, 2-action Markov games.

tially with the number of players. In our experiments, the MCECE strategy profile is the most uncertain (by design) and many of the previously investigated CE-Q solutions are the most deterministic.

5.2 Evaluation Metrics

The predictive guarantees of the MCECE strategy profiles relate to the log-loss of predicting actions distributed according to P_{obs} with distribution P_{pred}:

$$-\sum_{a \in A} P_{\text{obs}}(a) \log P_{\text{pred}}(a).$$

In the context of this work, P_{obs} and P_{pred} are each probabilities in CE strategy profiles for Markov games. Unfortunately, many of the strategy profiles provide no support for some action combinations that are possible in other strategy profiles. In other words, they predict that some action combinations occur with 0% probability when they do in fact occur with positive probability. This corresponds to an infinite log-loss.

Instead of using the typical log-loss measures, which is often infinite except for the MCECE strategy profile, we instead employ two measures. The first is the log-loss on the action combinations that do have support[4]. The second is the percentage of action combinations that have no support. The latter can be interpreted as the degree of infiniteness that the log loss would have. Equivalently, under the gambling perspective, it can be interpreted as the percentage of instances all of a gambler's money would be lost.

5.3 Action Prediction Comparison

The results of this comparison across strategy profiles are shown in Table 2 (for three and four agents). We note that in some cases one C-EQ strategy profile may better predict another than the MCECE strategy profile when their objectives are closely aligned. For example, the x_2CE-Q predicts iCE-Q fairly accurately since both are punishing agent 2 to some degree. However, overall the MCECE solution profile provides a much more robust prediction of other strategy profiles (and full support) on average (right column).

We employ the relationships between log-loss and dou-

[4]To address ϵ approximation error, we employ a minimum P_{obs} threshold of 0.1% for assessing non-support and a maximum penalty threshold of 16.6 bits ($-\log_2 0.00001$) for small support—both to the benefit of CE-Q strategy profiles.

Table 3: Doubling rates of CE as gambling allocations in the three and four player game settings.

	Three players	Four players
MCECE	**1.026**	**0.517**
μCE-Q	0.422	-1.636
d_1CE-Q	0.456	-1.415
d_2CE-Q	0.167	-1.926
x_1CE-Q	0.374	-0.346
x_2CE-Q	0.479	-0.175
iCE-Q	0.343	-1.356

bling rate from Section 2.2 to illustrate the benefits of different CE for gambling in Table 3. For the three player experimental setting, all CE strategies are expected to have positive investment growth rate (under uniform, fair odds). However, many also have a probability of losing all money (Table 2). Betting according to the MCECE distribution provides the largest growth rate—with the expectation of doubling an investment after each bet. For the four player setting, the MCECE distribution is the only one with positive expected investment growth. Thus, the theoretical properties for prediction under log-loss and gambling under uniform odds provided by the MCECE are realized in practice.

6. CONCLUSIONS

This paper was motivated by a fundamental question: given that agents act rationally (*i.e.*, according to an unknown correlated equilibrium) within a known Markov game, and no other information is available, what predictions of agents' action sequences should be employed? We employed an extension of information theory and the principle of maximum entropy to develop a predictive solution concept that addresses this question. We demonstrated the robustness of its predictions across a wide range of existing value-based CE solution concepts. In many settings where decisions are made based on the action sequences of self-interested, communicating autonomous agents, this assumption is reasonable. We have shown in theory and in practice the predictive guarantees of the approach and connected its guarantees to the gambling setting. Generalizing this approach to extensive form games is of interest, however it introduces non-convexity. We could replace the joint entropy measure of past coordinate descent approaches with the causal entropy measure [7].

Our view in this paper has been agnostic, apart from assuming joint rationality. Often, past agent behavior may be known. Important future work extends this approach to when such additional information *is* available. This can be accomplished with the addition of behavior-matching equality constraints to the MCECE solution concept optimization. Additionally, since actual behavior may only be approximately jointly rational, relaxing the inequality constraints using dual regularization [8] is another important direction. Extending the maximum entropy approach to behavior that is guided by unknown utility functions remains as an important future problem.

Acknowledgments

The authors gratefully acknowledge the Richard K. 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 Geoff Gordon and Miro Dudík for useful discussions.

7. REFERENCES

[1] R. Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1):67–96, 1974.

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

[3] S. Boyd, L. Xiao, and A. Mutapcic. Subgradient methods. *Lecture notes of EE392o, Stanford University, Autumn Quarter*, 2003.

[4] T. Cover and J. Thomas. *Elements of information theory*. Wiley, 2006.

[5] L. M. Dermed and C. L. Isbell. Solving Stochastic Games. In *Proc. NIPS*, pages 1186–1194, 2009.

[6] Y. Dodis, S. Halevi, and T. Rabin. A cryptographic solution to a game theoretic problem. In *Advances in Cryptology*, pages 112–130. Springer, 2000.

[7] M. Dudík and G. Gordon. A sampling-based approach to computing equilibria in succinct extensive-form games. In *Proc. UAI*, pages 151–160, 2009.

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

[9] S. Ficici and A. Pfeffer. Modeling how humans reason about others with partial information. In *Proc. AAMAS*, pages 315–322, 2008.

[10] D. Foster and R. Vohra. Calibrated Learning and Correlated Equilibrium. *Games and Economic Behavior*, 21(1-2):40–55, 1997.

[11] G. Gordon, A. Greenwald, and C. Marks. No-regret learning in convex games. In *Proc. ICML*, pages 360–367. ACM, 2008.

[12] A. Greenwald and Hall. Correlated Q-learning. In *Proc. ICML*, pages 242–249, 2003.

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

[14] S. Hart and A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

[15] J. Hu and M. Wellman. Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Proc. ICML*, pages 242–250, 1998.

[16] W. Huang and B. von Stengel. Computing an extensive-form correlated equilibrium in polynomial time. *Internet and Network Economics*, pages 506–513, 2008.

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

[18] S. Kakade, M. Kearns, J. Langford, and L. Ortiz. Correlated equilibria in graphical games. In *Proc. Electronic Commerce*, pages 42–47. ACM, 2003.

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

[20] M. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proc. ICML*, pages 157–163, 1994.

[21] C. Murray and G. Gordon. Multi-robot negotiation: approximating the set of subgame perfect equilibria in general-sum stochastic games. In *Proc. NIPS*, pages 1001–1008, 2007.

[22] J. Nash. Non-cooperative games. *Annals of mathematics*, 54(2):286–295, 1951.

[23] Y. Nyarko. Bayesian learning leads to correlated equilibria in normal form games. *Economic Theory*, 4(6):821–841, 1994.

[24] L. E. Ortiz, R. E. Shapire, and S. M. Kakade. Maximum entropy correlated equilibria. In *Proc. AISTATS*, pages 347–354, 2007.

[25] C. Papadimitriou and T. Roughgarden. Computing equilibria in multi-player games. In *Proc. SODA*, pages 82–91, 2005.
[26] B. D. Ziebart, J. A. Bagnell, and A. K. Dey. Modeling interaction via the principle of maximum causal entropy. In *Proc. ICML*, pages 1255–1262, 2010.
[27] M. Zinkevich, A. Greenwald, and M. Littman. Cyclic equilibria in Markov games. In *Proc. NIPS*, volume 18, pages 1641–1648, 2006.

APPENDIX

A. PROOFS OF THEOREMS

THEOREM 8 *(Proof).* Consider optimizing over conditional strategy sequence variables, $\eta(a^{1:T}|s^{1:T})$, that represent the probability of an entire sequence of actions given the entire sequence of states. Action-state strategy probabilities, $\pi(a^t|s^{t:T})$, can be obtained by marginalizing over a linear function of conditional sequence variables:

$$\pi(a^t|s^{t:T}) = \sum_{s^{1:t-1}} \sum_{a^{1:t-1}} \sum_{a^{t+1:T}} P(a^{1:T}, s^{1:t-1}|s^{t:T}) \quad (13)$$

$$= \sum_{s^{1:t-1}} \sum_{a^{1:t-1}} \sum_{a^{t+1:T}} \eta(a^{1:T}|s^{1:T}) \prod_{\tau=1}^{t-1} P(s^\tau|a^{\tau-1}, s^{\tau-1}).$$

Crucially, to match the Markov game setting, the conditional distribution of actions at time step t should be equivalent regardless of future state variables, $s^{t+1:T}$, since those variables are not yet known in the Markov game:

$$\forall_{t,a^t,s^t,s^{t+1:T},\tilde{s}^{t+1:T}} \pi(a^t|s^t, s^{t+1:T}) = \pi(a^t|s^t, \tilde{s}^{t+1:T}). \quad (14)$$

The constraints (Equation 14) are linear of conditional strategy sequence variables via the steps of Equation 13.

The expected regret can similarly be expressed as a linear function of conditional strategy sequence variables:

$$\begin{aligned}&\text{ExpRegret}_i^\pi(a_i^t, a_i^{t\prime}, s^t)\\ &= \sum_{a_{-i}^t, a^{t+1:T}, s^{1:t-1}, s^{t+1:T}} \eta(a^{1:T}|s^{1:T}) \frac{\prod_{\tau=1}^T P(s^{\tau+1}|s^\tau, a^\tau)}{P(s^{t+1}|s^t, a^t)} \times \\ &\left(P(s^{t+1}|s^t, a^{t\prime}) \left(\sum_{\tau>t} \text{Util}(s^\tau, a^\tau) + \text{Util}(s^t, a^{t\prime}) \right) \right.\\ &\left. - P(s^{t+1}|s^t, a^t) \left(\sum_{\tau>t} \text{Util}(s^\tau, a^\tau) + \text{Util}(s^t, a^t) \right) \right)\end{aligned}$$

All constraints are linear in conditional variables, so when $-f_0$ is a linear or convex function, the MCECE optimization (Equation 15) is a linear program or convex program.

$$\underset{\{\eta(a^{1:T}|s^{1:T})\}}{\text{argmax}} \; f_0(\{\eta(a^{1:T}|s^{1:T})\}) \quad (15)$$

$$\text{such that: } \forall_{t,i,s^t,a_i^t,a_i^{t\prime}} \text{ExpRegret}_i^\pi(a_i^t, a_i^{t\prime}, s^t) \le 0$$

$$\forall_{t,s^t,a^t} \pi(a^t|s^{t:T}) \ge 0, \qquad \forall_{t,s^t} \sum_{a^t} \pi(a^t|s^{t:T}) = 1$$

$$\forall_{t,a^t,s^t,s^{t+1:T},\tilde{s}^{t+1:T}} \pi(a^t|s^t, s^{t+1:T}) = \pi(a^t|s^t, \tilde{s}^{t+1:T}).$$

This formulation has $O(T|S|^T|A|)$ non-redundant constraints and a total of $O(|S|^T|A|^T)$ variables. □

THEOREM 10 *(Proof).* Ignoring all the deviation regret constraints in our notation, consider the decomposition of the causally conditioned entropy using the chain rule:

$$\begin{aligned}&\underset{\{\pi(a^t|s^t)\}}{\text{argmax}} \; H(A^T||S^T)\\ &= \underset{\{\pi(a^t|s^t)\}}{\text{argmax}} \left(H(A_i^T||S^T) + H(A_{-i}^T||S^T) \right)\\ &= \left\{ \pi^{\text{MCECE}}(a_{-i}^t|s_t) \right\} \cup \underset{\{\pi(a_i^t|s^t)\}}{\text{argmax}} \; H(A_i^T||S^T, A_{-i}^T).\end{aligned}$$

As shown, this is equivalent to a causally conditioned entropy maximization of agent i's strategy profile (with the suppressed deviation regret constraints) given the combined MCECE strategy profile of the other agents. By definition this is the least predictable strategy profile that agent i can employ (subject to any deviation regret constraints). □

THEOREM 11 *(Proof sketch).* As a special case of [13], the causal entropy can be expressed as:
$H(\tilde{P}(\mathbf{A}^T||\mathbf{S}^T)) = \inf_{P(\mathbf{A}^T||\mathbf{S}^T)} \mathbb{E}_{\tilde{P}(\mathbf{A},\mathbf{S})}[-\log P(\mathbf{A}^T||\mathbf{S}^T)]$.
Choosing $\tilde{P}(\mathbf{Y}^T||\mathbf{X}^T)$ that maximizes this is then:
$\sup_{\tilde{P}(\mathbf{A}^T||\mathbf{S}^T)} \inf_{P(\mathbf{A}^T||\mathbf{S}^T)} \mathbb{E}_{\tilde{P}(\mathbf{A},\mathbf{S})}[-\log P(\mathbf{A}^T||\mathbf{S}^T)]$, which is invariant to swapping the sup and inf operation order. □

THEOREM 13 *(Proof).* We find the form of the probability distribution by finding the optimal point of the Lagrangian.

$$\underset{\pi}{\text{argmax}} \, H(\mathbf{A}^T||\mathbf{S}^T) \text{ such that:} \quad (16)$$

$$\forall_{t,i,a_i^t,a_i^{t\prime},s^{1:t},a^{1:t-1}} \text{ExpRegret}_i^\pi(a_i^t, a_i^{t\prime}, s^{1:t}, a^{1:t-1}) \le 0$$

and probabilistic/causal constraints on π.

The Lagrangian for the optimization of Equation 16 when using entire history-dependent probability distributions and parameters is:

$$\begin{aligned}&\Lambda(\pi, \lambda) = H(a^{1:T}||s^{1:T})\\ &- \sum_{t,i,a_i^t,a_i^{t\prime},s^{1:t},a^{1:t-1}} \lambda_{t,i,a_i^t,a_i^{t\prime},s^{1:t},a^{1:t-1}} \text{ExpRegret}_i^\pi(a_i^t, a_i^{t\prime}, s^{1:t}, a^{1:t-1}) \quad (17)\end{aligned}$$

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

$$\begin{aligned}&\frac{\partial \Lambda(\pi, \lambda)}{\partial P(a^t|s^{1:t}, a^{1:t-1})} = -\log P(a^t|s^{1:t}, a^{1:t-1}) \quad (18)\\ &- \sum_{s^{t+1:T}, a^{t+1:T}} P(s^{t+1:T}, a^{t+1:T}) \log \prod_{\tau=t}^T P(a^\tau|s^{1:\tau}, a^{1:\tau-1})\\ &- \sum_{i,a_i^{t\prime}} \lambda_{t,i,a_i^t,a_i^{t\prime},s^{1:t},a^{1:t-1}} \text{ExpDevGain}_i^\pi(a_i^t, a_i^{t\prime}, s^{1:t}, a^{1:t-1}).\end{aligned}$$

Equating Equation 18 to zero provides the form of the history dependent distribution:

$$\begin{aligned}&P(a^t|s^{1:t}, a^{1:t-1}) \propto \exp\Big\{ \quad (19)\\ &\sum_{s^{t+1:T}, a^{t+1:T}} P(s^{t+1:T}, a^{t+1:T}) \log \prod_{\tau=t}^T P(a^\tau|s^{1:\tau}, a^{1:\tau-1})\\ &- \sum_{i,a_i^{t\prime}} \lambda_{t,i,a_i^t,a_i^{t\prime},s^{1:t},a^{1:t-1}} \text{ExpDevGain}_i^\pi(a_i^t, a_i^{t\prime}, s^{1:t}, a^{1:t-1}) \Big\}.\end{aligned}$$

Convex duality in this optimization relies on a feasible solution on the relative interior of the constraint set. This can be accomplished by adding an infinitesimally small amount of slack, ϵ, to the constraint set [24]. Following the argument that the MCECE is Markovian (Lemma 12), Equation 19 reduces to the Markovian form of the theorem. □