Inverse Correlated Equilibrium for Matrix Games

Kevin Waugh
waugh@cs.cmu.edu
Department of Computer Science
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA USA 15213

Brian D. Ziebart
bziebart@cs.cmu.edu
The Robotics Institute
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA USA 15213

J. Andrew Bagnell
dbagnell@ri.cmu.edu
The Robotics Institute
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA USA 15213

Abstract

Modeling the joint behavior of multiple imperfect agents from a small number of observations is a difficult, but important task. In the single-agent, decision-theoretic setting, inverse optimal control has been successfully employed. It views observed behavior as an approximately optimal solution to an unknown decision problem, and learns the decision problem's parameters that best explains the observed behavior. In this work, we introduce the Inverse Correlated Equilibrium problem, the multi-agent extension of inverse optimal control to normal-form games. We describe two approaches for this problem and discuss how they enable prediction of behavior without knowledge of the game's reward function. The first approach solves a convex optimization problem, but, unfortunately, often restricts the ability to generalize more than we desire. The second approach is more computationally intensive, but exhibits better worst-case performance guarantees.

1 Introduction

Though frameworks for optimal or rational decision making are useful for prescribing behavior that an agent should perform, they are often ill-suited for predicting actual behavior. This is because real behavior is typically not consistently optimal or rational; it may be influenced by factors that are difficult to model or subject to various types of error when executed. Recent research in imitation learning, and specifically inverse optimal control, has bridged the gap between prescriptive and predictive applications of decision frameworks in the single-agent setting [1, 8, 11, 12]. Successful applications include learning and prediction tasks in personalized vehicle route planning [12], robotic crowd navigation [4], quadruped foot placement and grasp selection [9]. A reward function is learned by those techniques that best explains demonstrated behavior and approximates the optimality criteria of prescriptive decision-theoretic frameworks.

Motivated by those successes, we extend inverse optimal control to multi-agent settings in this work. Game-theoretic concepts such as rationality and regret measures replace the utility and optimality measures of decision theory in this setting, and help assess how decision are made. Our approach learns analogs to these from limited (and potentially irrational and suboptimal) observational data. Importantly, the probabilistic reasoning that results from our formulation allows an appropriate treatment of the imperfections inherent in real data. In the following sections, we will review game the-

ory and the principle of maximum entropy. After, we present two approaches that combine these two concepts to provide strong predictive performance guarantees using the rationality constructs of game theory. We conclude with some brief experimental results and a discussion of future work.

2 Background

2.1 Game Theory

Matrix games are the basic building block used by Game Theory to study the strategic interactions between multiple agents in an environment. Most well known matrix games, including the "Prisoners Dilemma" game and the "Battle of the Sexes" game, are merely toy problems with interesting strategic properties, but many interesting problems, such as negotiations among parties with conflicting interests or collaboration within teams, can be represented as matrix games.

Definition 1. *A **normal-form game** [10], or **matrix game**, is a tuple $\Gamma = (N, \mathcal{A}, u)$ where*

- *N is a finite set of **players**,*
- *$\mathcal{A} = \times_{i \in N} A_i$ is the set of **joint-actions**, where A_i is player i's finite set of **actions**, and*
- *u contains a $u_i : \mathcal{A} \mapsto \mathbb{R}$, the **utility function** for player i, for every $i \in N$.*

For notational convenience, we let $\mathcal{A}_{-i} = \times_{j \neq i, j \in N} A_j$ represent the set of actions for players other than i, and define $a_{-i} \in \mathcal{A}_{-i}$ to be the vector a excluding component i. We let $A_{\max} = \max_{i \in N} |A_i|$.

In this work, we are concerned with the multi-agent setting where players' utilities are parametrized. A vector of features is associated with each joint combination of actions.

Definition 2. *Let the tuple $\Xi = (N, \mathcal{A}, F)$ denote a normal-form game with a linear utility function. That is, for every joint-action, each player has a feature vector $f_i(a) \in [-1, 1]^K$ and $u_i(a|w^*) = \langle f_i(a), w^* \rangle$ for some (unknown to us) vector of feature weights $w^* \in \mathbb{R}^K$.*

To allow for agents to cooperate, as not all games of interest are necessarily competitive in a strong sense, we will consider a solution concept that incorporates a coordination device. In particular, prior to play, the environment will draw a joint-action x from a distribution σ and communicate to each player x_i, their portion of the joint-action. Then, each player will simultaneously choose an action $a_i = g_i(x_i)$. After, each player is assigned reward by their utility function depending on the joint-action a. We call the function g_i a modification function [2].

Definition 3. *A **modification function** for player i is a function $g_i : A_i \mapsto A_i$.*

In a sense, we can measure the quality of the coordinating distribution by how much a player can benefit by deviating from the recommended action. This measure is known as swap regret and leads to the solution concept known as a correlated equilibrium.

Definition 4. *The **instantaneous regret** [2] experienced by player i for choosing action a_i when all other players play a_{-i} with respect to modification function g_i is,*

$$\text{regret}_i(a, g_i|w) = u_i(g_i(a_i), a_{-i}|w) - u_i(a|w). \tag{1}$$

Definition 5. *The switch modification function, $\text{switch}_i^{x \to y} : A_i \mapsto A_i$, is defined as,*

$$\text{switch}_i^{x \to y}(a_i) = \begin{cases} y & \text{if } a_i = x \\ a_i & \text{otherwise} \end{cases}. \tag{2}$$

Definition 6. *The **expected regret for switching from action x to action y** for player i when joint-actions are drawn from σ is*

$$\text{r}_i^{\textit{switch}}(x \to y|\sigma, w) = \mathbb{E}_{a \sim \sigma} \left[\text{regret}_i(a, \text{switch}_i^{x \to y} |w) \right]. \tag{3}$$

Definition 7. *The **expected internal regret on action x** for player i under distribution σ is*

$$\text{r}_i^{\textit{internal}}(x|\sigma, w) = \max_{y \in A_i} \mathbb{E}_{a \sim \sigma} \left[\text{regret}_i(a, \text{switch}_i^{x \to y} |w) \right]. \tag{4}$$

Definition 8. *The **expected swap regret** [2] for player i when joint-actions are drawn from σ is*

$$\mathrm{R}_i^{\textit{swap}}(\sigma|w) = \sum_{x \in A_i} \mathrm{r}_i^{\textit{internal}}(x|\sigma, w). \tag{5}$$

*A distribution σ over $\mathcal{A}$ is an **ε-correlated equilibrium** [2] if for all $i \in N$,*

$$\mathrm{R}_i^{\textit{swap}}(\sigma|w) \leq \varepsilon. \tag{6}$$

In words, a distribution σ is a correlated equilibrium if no player can benefit by deviating from the recommended action given that all other players play according to the recommendation. A distribution is an ε-correlated equilibrium if no player can benefit more than ε by deviating, again assuming the others play according to their recommendation.

For readers familiar with the Nash equilibrium solution concept, we note that the correlated equilibrium is an extension of this concept. That is, a Nash equilibrium is a correlated equilibrium where the recommended actions are chosen independently. Additionally, a persuasive argument for the validity of the correlated equilibrium can be made as one can be reached by simple evolutionary dynamics, whereas no such dynamics are known for the Nash equilibrium [2].

2.2 The Principle of Maximum Entropy

Information theory provides tools for evaluating the predictive power of a distribution. One key quantity is Shannon's information entropy, which measures the uncertainty or information content.

Definition 9. *The **entropy** of a distribution σ over a finite set X is*

$$H(\sigma) = -\sum_{x \in X} \sigma(x) \log \sigma(x). \tag{7}$$

The **principle of maximum entropy** advocates choosing the distribution with maximum entropy subject to known constraints [5].

Definition 10. *The **maximum entropy distribution** is defined as:*

$$\sigma_{\textit{MaxEnt}} \triangleq \operatorname*{argmax}_{\sigma} H(\sigma) \quad \textit{subject to: } g(\sigma) = 0 \textit{ and } h(\sigma) \leq 0. \tag{8}$$

The constraint functions, $g : \sigma \mapsto \mathbb{R}^{K_1}$ and $h : \sigma \mapsto \mathbb{R}^{K_2}$, are typically chosen to capture the important or most salient characteristics of a distribution. When those functions are affine and convex, respectively, in the elements of the probability distribution, σ, finding this distribution is a convex optimization problem. This distribution has important guarantees for prediction.

Lemma 1 (from [3])**.** *The maximum entropy distribution minimizes the worst-cast log-loss, $-\sum_{x \in \mathcal{X}} \tilde{\sigma}(x) \log \sigma(x)$, when nature adversarially chooses $\tilde{\sigma}(x)$ subject to provided constraints.*

As log-loss is a common criteria for evaluating machine learning predictions, the principle of maximum entropy serves as an underlying justification for many existing machine learning techniques (*e.g.*, logisitic regression, Markov random fields, conditional random fields). In the context of multi-agent behavior, it has been employed to obtain correlated equilibria with predictive guarantees in normal-form games when the utilities are known [7]. We employ the principle of maximum entropy in this work to provide predictive guarantees in settings when players' utilities are unknown.

3 The Inverse Correlated Equilibrium Problem

Imitation learning aims to accurately predict future behavior from examples of past decisions. That is, we cannot make assumptions as to what goal each agent is attempting to achieve, or even if they are acting in an optimal fashion to achieve this goal. Formally, we model our interactions as normal-form games with unknown utility functions. As a substitute for the utility function, we assume features for each joint-action are available assume further that the true reward function is a linear function of those features. It is arguably much easier to model real world situations under

this framework than it is to construct an accurate reward function. For example, often it is the case that traveling a distance or spending time to complete a task will result from a particular decision, both of which can be easily measured though exactly how these quantities translate into "utility" is unclear and may depend on the agents' internal preferences.

Problem (Inverse Correlated Equilibrium)**.** *Given a normal-form game with unknown utility function, $\Xi = (N, \mathcal{A}, F)$ and a sequence of observations, $\{a^{(j)} \in \mathcal{A}\}$ for $j = \{1, 2, \ldots, m\}$, we denote the empirical distribution of $\{a^{(j)}\}$ as $\tilde{\sigma}$, the* ***demonstrated behavior****. From $\tilde{\sigma}$, we wish to produce a correlating distribution σ that can accurately predict the future behavior of the agents in the system.*

Property 1. *We say that σ preserves the quality of $\tilde{\sigma}$ if for any $w \in \mathbb{R}^K$ and for all players $i \in N$,*

$$\mathrm{R}_i^{\textit{swap}}(\sigma|w) \leq \mathrm{R}_i^{\textit{swap}}(\tilde{\sigma}|w). \tag{9}$$

That is, for any reward function, σ is no further from being a correlated equilibrium than $\tilde{\sigma}$.

3.1 Feature Regret Matching

Our first imitation learning approach employs constraints so that the distribution's expected switch regret matches that of the demonstrated joint-action distribution, $\tilde{\sigma}$, for all possible utility functions:

$$\forall_{w \in \mathbb{R}^K} \left\{ \forall_{i \in N, x, y \in A_i} \mathrm{r}_i^{\text{switch}}(x \to y|\sigma, w) = \mathrm{r}_i^{\text{switch}}(x \to y|\tilde{\sigma}, w) \right\}. \tag{10}$$

Note that matching the expected switch regrets is stronger than preserving the quality of $\tilde{\sigma}$.

Due to the linear definition of the game utility functions, matching the expected switch regret of the demonstrated behavior is satisfied by distributions that match expected feature differences with the empirical distribution, as shown by the constraints in the following definition.

Definition 11. *The maximum entropy feature regret matching distribution is*

$$\underset{\sigma \in \Delta_{\mathcal{A}}}{\operatorname{argmax}} H(\sigma) \tag{11}$$

$$\forall_{i \in N, x, y \in A_i, k} \sum_{a_{-i} \in \mathcal{A}_{-i}} \sigma(x, a_{-i}) \left[f_i^k(y, a_{-i}) - f_i^k(x, a_{-i}) \right] = \sum_{a_{-i} \in \mathcal{A}_{-i}} \tilde{\sigma}(x, a_{-i}) \left[f_i^k(y, a_{-i}) - f_i^k(x, a_{-i}) \right].$$

Here, we use superscript k to denote the kth entry in the feature vector.

Lemma 2. *The maximum entropy feature regret matching distribution preserves the quality of $\tilde{\sigma}$.*

This distribution provides minimal worst-case log-loss guarantees of all the distributions that match feature regrets with the demonstrated distribution's feature regrets. However, note that for N players and K features there are $O(NA_{\max}^2 K)$ constraints, which each correspond to a free parameter in the Lagrangian of the optimization. A demonstrated action distribution has $O(A_{\max}^N)$ values. Thus, for small numbers of players there are more free parameters than data points, and generalization is difficult—the learned distribution will exactly match the demonstrated distribution.

3.2 Internal Regret Matching

To provide better generalization from demonstrated joint action distributions, and thus allowing generalization in two-player scenarios, a less constrained optimization is needed. Fortunately, the above constraints are tighter than is necessary. That is, there exists distributions σ that preserve the quality of $\tilde{\sigma}$, but that do not match the expected switch regrets. To achieve this goal, we will construct a set of linear inequalities guaranteeing that, under any utility function, the internal regret on every action for every player cannot exceed that of the demonstrated behavior.

Definition 12. *Choose $\mathcal{K}_i(x, y'|F, \tilde{\sigma})$ to be the corner points of the polytope defined by the constraints*

$$\sum_{a_{-i} \in \mathcal{A}_{-i}} \tilde{\sigma}(x, a_{-i}) \left[\langle f_i(y, a_{-i}), w \rangle - \langle f_i(y', a_{-i}), w \rangle \right] \geq 0, \forall y \in A_i \tag{12}$$

$$-1 \leq w_i \leq 1 \tag{13}$$

Algorithm 1 MaxEntICE-ExpGrad

Let $\sigma^{(1)} \leftarrow \text{Uniform}(\mathcal{A})$
Let $w^{(1)} \leftarrow \mathbf{1}$
Let $\eta \leftarrow \sqrt{N \log(A_{\max})/T}/K$
For $k = 1, 2, \ldots, T$
  $w^{(k)} \leftarrow \max_{|w|\leq 1} \max_{i \in N} \mathrm{R}_i^{\text{internal}}(\sigma|w) - \max_{i \in N} \mathrm{R}_i^{\text{internal}}(\tilde{\sigma}|w)$
  $i^{(k)}, x^{(k)}, y^{(k)} \leftarrow \text{argmax}_{i \in N, x, y \in A_i} \sum_{a_{-i} \in A_i} \sigma(x, a_{-i}) \left[\left\langle f_i(y, a_{-i}), w^{(k)} \right\rangle - \left\langle f_i(x, a_{-i}), w^{(k)} \right\rangle \right]$
  $\partial \sigma^{(k)}_{x, a_{-i^{(k)}}} \leftarrow \begin{cases} \left\langle f_{i^{(k)}}(y^{(k)}, a_{-i^{(k)}}), w \right\rangle & \text{if } x = y^{(k)} \\ \left\langle -f_{i^{(k)}}(x^{(k)}, a_{-i^{(k)}}), w \right\rangle & \text{if } x = x^{(k)} \\ 0 & \text{otherwise} \end{cases}$
  $w_a^{(k+1)} \leftarrow w_a^{(k)} \cdot \exp(-\eta \cdot \partial \sigma_a^{(k)})$
  $\sigma^{(k+1)} \leftarrow w^{(k+1)} / \left|\left| w^{(k+1)} \right|\right|_1$

Player i would be best off to deviate from x to y' under any utility function in $\mathcal{K}_i(x, y'|F, \tilde{\sigma})$. Furthermore, the utility functions in $\mathcal{K}_i(x, y'|F, \tilde{\sigma})$ *c*over all of the utility functions where this is the case. That is, any utility function where player i would prefer to switch from x to y' is a positive combination of the utility functions in this set. As a consequence, we only need to check that a candidate σ have no more internal regret than $\tilde{\sigma}$ for a finite number of possible utility functions.

Lemma 3. *If, for all players $i \in N$, all actions $x, y' \in A_i$ and all $w \in \mathcal{K}_i(x, y'|F, \tilde{\sigma})$, σ satisfies*

$$\max_{y \in A_i} \sum_{a_{-i} \in \mathcal{A}_{-i}} \sigma(x, a_{-i}) \left[\langle f_i(y, a_{-i}), w \rangle - \langle f_i(x, a_{-i}), w \rangle \right] \leq \sum_{a_{-i} \in \mathcal{A}_{-i}} \tilde{\sigma}(x, a_{-i}) \left[\langle f_i(y', a_{-i}), w \rangle - \langle f_i(x, a_{-i}), w \rangle \right], \tag{14}$$

it has no more expected internal regret than $\tilde{\sigma}$ and, thus, preserves its quality.

As in the feature regret matching case, we can use convex optimization techniques to choose the maximum entropy distribution that satisfies these constraints as to minimize the log-loss in the worst-case. Unfortunately, the number of constraints above, though finite, can be exponential in K. Luckily, most do not matter when it comes to determining the maximum entropy distribution. That is, using constraint generation is an appealing alternative to enumerating the corner points of $\mathcal{K}_i(x, y'|F, \tilde{\sigma})$. Compared with feature regret matching, this approach more easily generalizes from smaller amount of data since fewer measures must be estimated and fewer constraints satisfied. However, it necessarily does not match as many of the characteristics of demonstrated behavior.

Next, we describe a method for solving the maximum entropy internal regret matching problem in time polynomial in the number of features. Consider the optimization problem:

$$\min_{\sigma \sim \mathcal{A}} \max_{w} \max_{i \in N} \mathrm{R}_i^{\text{internal}}(\sigma|w) - \max_{i \in N} \mathrm{R}_i^{\text{internal}}(\tilde{\sigma}|w) \tag{15}$$
$$\text{subject to: } ||w||_1 \leq 1$$

We observe that this optimization has an optimal value of 0 that is achieved by any distribution that matches the internal regret of $\tilde{\sigma}$. Furthermore, the objective of the outer minimization is convex in σ. The objective of the inner maximization, though not convex in w, can be solved efficiently using a combination of case analysis and linear programming. That is, we can compute gradients with respect to σ. As exponentiated gradient descent implicitly incorporates an entropy regularizer to the objective that it minimizes, we can efficiently approximate the maximum entropy internal regret matching distribution [6]. This approach is presented as Algorithm 1.

Lemma 4. *MaxEntICE-ExpGrad has a running time of $O(N^2 A_{\max}^{N+4} K \cdot \mathrm{LP}(K, N A_{\max}^2 + K)$ and is within ϵ of the optimal maximum entropy internal regret matching policy after $T = 4K^2 N \log(A)/\epsilon^2$ iterations, where $\mathrm{LP}(n, m)$ is the complexity of solving a linear program with n variables and m constraints.*

	Straight	Stop
Straight	-10,-10	1,0
Stop	0, 1	0,0

Figure 1: Payoff matrix for chicken

	Straight	Stop
Straight	0.037	0.333
Stop	0.333	0.297

Figure 2: σ_{MaxEnt} for chicken

4 Experimental Results

For our experiments, we used the well known game Chicken, displayed in Figure 1. Chicken is a two-player general sum game. Each player decides whether to drive straight, or to avoid the other player. If both players drive straight, they collide and both receive a penalty. If one player drives straight while the other remains stopped, the moving player gets a slight reward. This game models traffic at an intersection and a correlation device in this case could be a traffic signal.

As features, we used the actual game's reward. Thus, the true reward function is $w^* = 1$. The demonstrated behavior provided to the algorithm has one player driving straight 60% of the time, and the other the remaining 40% – a correlated equilibrium. The maximum entropy internal regret matching distribution is shown in Figure 2. Though this distribution has the players crashing about 4% of the time, it is also a correlated equilibrium.

5 Conclusion and Future Work

In this paper, we have extended inverse optimal control to the multi-agent setting by combining the game-theoretic concept of regret with the information-theoretic concept of maximum entropy. The resulting probability distribution over joint actions provides important regret-based guarantees and predictive guarantees with respect to demonstrated training data. The objective of our future work is to transfer the knowledge learned using this technique in one game setting to make predictions of behavior in other game settings or with other combinations of players. Extending this approach to sequential games and stochastic games also remains as important future work.

References

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-first International Conference on Machine Learning*, 2004.

[2] A. Blum and Y. Mansour. *Algorithmic Game Theory*, chapter Learning, Regret Minimization and Equilibria, pages 79–102. Cambridge University Press, 2007.

[3] 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.

[4] P. Henry, C. Vollmer, B. Ferris, and D. Fox. Learning to navigate through crowded environments. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 981–986. IEEE, 2010.

[5] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, May 1957.

[6] J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132, 1995.

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

[8] N. Ratliff, J. A. Bagnell, and M. Zinkevich. Maximum margin planning. In *Proceedings of The Twenty-Third International Conference on Machine Learning*, 2006.

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

[10] E. Tarados and V. V. Vazirani. *Algorithmic Game Theory*, chapter Basic Solution Concepts and Computational Issues, pages 3–28. Cambridge University Press, 2007.

[11] B. D. Ziebart, J. A. Bagnell, and A. K. Dey. Modeling interaction via the principle of maximum causal entropy. In *ICML*, 2010.

[12] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. Dey. Maximum entropy inverse reinforcement learning. In *Proceeding of The Twenty-Third AAAI Conference on Artificial Intelligence*, 2008.