Strategy Transfer Learning via Parametric Deviation Sets

Brian D. Ziebart BZIEBART@CS.CMU.EDU
Kevin Waugh WAUGH@CS.CMU.EDU
J. Andrew Bagnell DBAGNELL@RI.CMU.EDU
Geoffrey J. Gordon GGORDON@CS.CMU.EDU
Katia Sycara KATIA@CS.CMU.EDU
Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, USA 15213

Abstract

Accurately reasoning about agents' actions in strategic settings is a challenging artificial intelligence task. Many difficulties in learning to perform such reasoning arise due to the uncertainty in both agents' motives and the strategic games being played. In this paper, we address the problem of learning from observations of the agents' behavior in some games to predict play in different, but related, games. We introduce a deviation-based strategy prediction approach that, by also using game outcome features to describe different games in a common language, enables generalized strategy learning.

1. Introduction

Accurate predictions of the behavior of multiple agents in strategic settings are needed for many artificial intelligence applications, including opponent modeling, mechanism design, and behavior imitation tasks. There are two sources of information transfer to enable a predictor to perform this task: **probability** (similar behavior in similar situations) and **utility** (similar payoffs in similar situations). Past research on predicting game strategies has primarily focused on using probability-based transfer on games with known payoffs (Altman et al., 2006; McKelvey & Palfrey, 1995; Wright & Leyton-Brown, 2010) or learning utilities that rationalize strategies under assumed equilibria (Yang, 2009). New machine learning approaches using both sources of transfer are needed to enable prediction in more general settings.

A number of difficulties exist for the predictor in more general settings. First, the payoffs motivating observed behavior are generally unknown. Second, agents may appear to behave irrationally or inconsistently to the predictor due to limited game description availability for the learner. Third, the games of interest may have substantial structural differences (*e.g.*, different actions, number of players) and strategic differences (*e.g.*, desirabilities of outcomes) from one another.

To overcome these difficulties, we augment ideas from statistical machine learning for reasoning about uncertainty with game-theoretic concepts that explain strategic behavior. More specifically, our approach: (1) Recasts games into a feature-based perspective, expanding utility from a single number to **outcome features**—sets of variables that characterize an outcome in a language common across games; (2) Uses feature-based **deviation regret-matching** constraints to provide game-theoretic guarantees for predicted strategies relative to demonstrated strategies; and (3) Provides a strong predictive performance guarantee through its use of the **principle of maximum entropy**.

Combining these ideas, our approach, **maximum entropy deviation regret matching**, uses both sources of transfer—probability and utility—to enable strategy learning across all games sharing a common set of outcome features. We provide a set of experiments that demonstrate the approach's abilities to address the difficulties of this prediction task.

2. Preliminaries

We begin with a review of game theory concepts that our work builds upon and a discussion of related work on learning game strategies.

2.1. Games, strategies, deviations, and regrets

The canonical game setting, the **normal-form game**, is a single-step game that is defined by the payoffs received by each agent resulting from their joint behavior.

Definition 1. *A **normal-form game**, $G = (N, \mathcal{A}, Payoff)$, is comprised of: a set of players, N; a set of joint actions, $\mathcal{A}$; and a tuple of payoffs, $Payoff_i(a) : \mathcal{A} \mapsto \mathbb{R}$,*

This is a preliminary draft. Copyright 2011 by the author(s)/owner(s).

specifying the utility of game outcome $a \in \mathcal{A}$ to player $i \in N$. Each player controls one component of the action $a = (a_1, a_2, \ldots)$ with $a_i \in \mathcal{A}_i$.

Other players' actions, excluding player i's, are conveniently denoted as $a_{-i} \in \mathcal{A}_{-i}$ and payoffs are then also denoted $\text{Payoff}_i(a_i, a_{-i})$. A **strategy** is a probability distribution over the joint-actions, $P(A)$, with specific action probabilities expressed as $P(a)$ or $P(a_i, a_{-i})$. Strategies can be **independent**, and thus factor as $P(a) = \prod_{i \in N} P(a_i)$, or **coordinated**. Conceptually, coordinated behavior could arise from an external signaling mechanism, e.g., a traffic light; however, no external moderator is needed so long as the players have access to a public communications channel (Dodis et al., 2000).

A **deviation policy**, $\phi(a'_i|a_i) : \mathcal{A}_i \mapsto \Delta_{\mathcal{A}_i}$, is an important conceptual tool in game theory that defines a probabilistic mapping over A describing alternate behavior for player i when prescribed action a_i. It is useful for assessing rationality and defining equilibrium solution concepts:

Definition 2. *A strategy, $P(A)$, is said to have **no-regret** with respect to **deviation policy set** Φ if employing any deviation policy provides no expected gain, i.e., $\forall i, \phi_j \in \Phi$ $\textit{Regret}_i^{\phi_j}(P(A)) \leq 0$ where (denoting probabilistic expectations as: $\mathbb{E}_{P(A)}[f(a)] = \sum_{a \in \mathcal{A}} P(a) f(a)$):*

$$\textit{Regret}_i^{\phi_j}(P(A)) \triangleq \tag{1}$$
$$\mathbb{E}_{P(A)}\left[\sum_{a'_i \in \mathcal{A}_i} \phi_j(a'_i|a_i)\ \textit{Payoff}_i(a'_i, a_{-i}) - \textit{Payoff}_i(a)\right].$$

By considering the set of all deviation policies that switch from one action to another, the **correlated equilibrium** solution concept (Aumann, 1974) is obtained. By restricting the set of deviation policies, other equilibrium concepts are obtained as well: e.g., deviations that switch unconditionally to a fixed action lead to a minimax optimal Nash equilibrium in a zero-sum game (Nash, 1951).

2.2. Related work

Unfortunately, strategies employed by people often are not consistent with Nash or correlated equilibrium solution concepts for the designed payoff functions of games. Much research in strategy prediction has investigated relaxations of the rationality constraints of those equilibria to improve predictive performance. For example, the **quantal response equilibrium** (McKelvey & Palfrey, 1995) introduces a learned parameter that controls the degree of rationality of strategies with convergence to Nash equilibrium at one extreme. Many other techniques have been developed for predicting strategies in games with known payoffs (Altman et al., 2006; Cooper & Kagel, 2003; Halpern & Pass, 2009; Wright & Leyton-Brown, 2010).

Other strategy prediction techniques discard the game-theoretic notions of payoff, equilibrium, and regret entirely, and instead directly estimate the strategy from demonstrated behavior. For instance, case-based reasoning techniques have been employed to transfer observed computer-game strategies to new scenarios (Sharma et al., 2007). Direct policy estimation approaches have the advantage of being applicable in games where the payoffs are unknown or difficult to specify. However, generalization to games with structural differences, such as additional players with novel motives or different sets of actions, is difficult without a model capable of deeply learning the underlying motives of observed behavior.

In contrast to these techniques, our approach predicts strategies in different games by learning underlying deviation-based payoff functions that best explain observed behavior. Thus, it is capable of good generalization across games with structural and strategic differences without having the benefit of known payoffs. We take inspiration from recent research in imitation learning that learns a reward function for Markov decision processes that best explains sequences of behavior (Ziebart et al., 2008; Abbeel & Ng, 2004; Ziebart et al., 2010). These reward functions are learned in terms of state-action features. We extend this perspective to normal-form games by learning deviation-specific payoff functions in terms of game outcome features, capturing strategic properties of observed behavior.

Recent research taking the feature-based perspective for behavior transfer in games relates the regret of a source game (s) and a target game (t) with the same outcome feature set (Waugh et al., 2011). Specifically, this work requires that a prediction $P_t(A)$ have low regret for weights where $P_s(A)$ has low regret, or

$$\forall w, \max_{i, \phi_j \in \Phi_t} \text{Regret}_{i,w}^{\phi_j}(P_t(A)) \leq \kappa \max_{i, \phi_j \in \Phi_s} \text{Regret}_{i,w}^{\phi_j}(P_s(A)). \tag{2}$$

This approach does not benefit from knowledge of how actions relate across two games and requires a convex optimization for each prediction. In contrast, the approach introduced in this paper explicitly explains observed behavior in terms of specific deviations to enable generalization across different games and uses learned parameters corresponding to deviations from a separate training phase without additional optimization for predictions.

Elements of our approach and its motives have shaped other previously developed techniques that relate to strategic behavior, but with different purposes. Bayesian games (Harsanyi, 1967) provide some payoff uncertainty by associating a (well-specified) hidden type with each player that determines the player's payoff for each outcome, but this uncertainty extends to the players within the game rather than being restricted to the machine learner, as in our setting. The principle of maximum entropy has been employed by the maximum entropy correlated equilibria (Ortiz et al., 2007) to obtain a unique strategy in normal-form games with known payoffs.

3. Strategy Learning and Prediction

In this section, we introduce our maximum entropy regret matching (MaxEnt DRM) framework. We employ a feature-based perspective that generalizes across games by recasting regret in terms of outcome features. Then, we combine feature-based performance constraints, based on regret, with the principle of maximum entropy to obtain strategy predictions that inherit performance guarantees.

3.1. Viewing games via outcome features

We consider games where outcomes are described by vectors of outcome features, and the game's payoffs, which characterize the agents' desires, are determined by weights on these features.

Definition 3. *The payoffs of an **outcome-parameterized normal-form game**, $G = (N, \mathcal{A}, F, W)$, are characterized by: vectors of **outcome features**, $F : N \times \mathcal{A} \to \mathbb{R}^K$ (denoted as $\boldsymbol{f}_{i,a}$ or $\boldsymbol{f}_{i,a_i,a_{-i}}$), for each player $i \in N$ and joint action $a \in \mathcal{A}$; and a vector of payoff **outcome weights** $\boldsymbol{w}_i \in W$, associated with each player $i \in N$, such that: $Payoff_{i,\boldsymbol{w}}(a) = \boldsymbol{w}_i^T \boldsymbol{f}_{i,a}$.*

Outcome-parameterized normal-form games can be transformed into standard normal-form games (Definition 1) when features F and weights $\mathbf{w}$ are known. Similarly, regrets (and equilibrium concepts) can be extended to the feature-based payoff setting and are denoted as: $\text{Regret}^{\phi}_{i,\mathbf{w}}(P)$. However, we argue that the payoffs guiding strategies are difficult for an observing learner to precisely know. Indeed, in human subject experiments where we might expect that experimenter knows the exact payoffs, additional motives, such as maintaining a certain "image" in negotiation-like games (*e.g.*, that of a tough negotiator or altruist), reciprocating perceived selfless actions, and avoiding inequality have been shown to alter behavior in ways that would not be expected if the given payoffs were the sole factor underlying decisions (Gintis, 2009).

Feature incompleteness is another crucial prospect to consider; for example, additional unknown outcome features may exist that form the true payoff function governing observed behavior, or the predictor's features may be a noisy approximation of the game's true outcome features. Statistical machine learning techniques are specifically designed to address the uncertainty of such incomplete modeling assumptions. However, the assumptions needed to make these statistical machine learning techniques practical are often difficult to combine with game-theoretic reasoning. The aim of our approach, therefore, is to enable the combination of statistical and game-theoretic reasoning, particularly in the transfer learning setting with incomplete features.

When features characterize game outcomes across many different games, cross-game learning can be beneficial even when the actions of those games have no direct correspondence. To provide an entirely feature-based view of observed behavior that enables transfer, we redefine deviations in terms of features as well, so that deviations in different games will relate.

Definition 4. *A **feature-based deviation policy** is a deviation policy $\phi(a'_i|a_i) = \varphi(F, a_i, a'_i)$ defined in terms of the game outcome features F.*

Thus, each feature-based deviation policies may be employed for any game that is characterized by the same set of features. We aim to employ these policies to enable strategy learning across games.

3.2. Feature-based regret matching

To generalize well from a small number of examples, salient properties of observed behavior that compactly represent the distribution are necessary to enable generalization and transfer across game settings. For single-agent, decision-theoretic behavior, these properties are typically associated with the reward of a decision process (Abbeel & Ng, 2004). For the strategic settings, we advocate regret-based properties.

A natural requirement for a strategy prediction $\hat{P}(A)$ is that it match regret (Equation 1) with the demonstrated strategy $\tilde{P}(A)$ evaluated on any feature-based deviation policy ϕ_j from some set Φ:

$$(\forall i \in N, \phi_j \in \Phi)\ \text{Regret}^{\phi_j}_{i,\mathbf{w}^*}(\hat{P}(A)) = \text{Regret}^{\phi_j}_{i,\mathbf{w}^*}(\tilde{P}(A)). \tag{3}$$

However, even if we assume that some true weights $\mathbf{w}^*$ parametrize the payoffs, those weights are unknown to us. To address this difficulty, our approach guarantees this property for all possible weights $\mathbf{w}$. This is accomplished by the set of constraints in Theorem 1.

Theorem 1 (Deviation regret matching[1]). *Regret matching (based on Equation 3),*

$$(\forall \boldsymbol{w})\ Regret^{\phi_j}_{i,\boldsymbol{w}}(\hat{P}(A)) = Regret^{\phi_j}_{i,\boldsymbol{w}}(\tilde{P}(A)) \tag{4}$$

is guaranteed for all choices of feature weights $\boldsymbol{w}$ and any deviation policy, $\phi_j \in \Phi$, if and only if the expected feature differences that result from employing the deviation policy matches under both strategy distributions:

$$\mathbb{E}_{\hat{P}(A)}\left[\sum_{a'_i \in A_i} \phi_j(a'_i|a_i)\, \boldsymbol{f}_i(a'_i, a_{-i}) - \boldsymbol{f}_i(a)\right]$$
$$= \mathbb{E}_{\tilde{P}(A)}\left[\sum_{a'_i \in A_i} \phi_j(a'_i|a_i)\, \boldsymbol{f}_i(a'_i, a_{-i}) - \boldsymbol{f}_i(a)\right]. \tag{5}$$

[1]The proofs of Theorems 1, 3, 4, and 5 are provided in the supplemental appendix.

These constraints are quite different from equilibrium constraints (Definition 2), which require that the behavior cannot benefit from any deviation policy $\phi \in \Phi$. However, they do ensure that a demonstrated equilibrium $\tilde{P}(A)$ is maintained by $\hat{P}(A)$ in source games.

Corollary 1. *If $\hat{P}(A)$ satisfies the constraints of Theorem 1 relative to distribution $\tilde{P}(A)$ then $\hat{P}(A)$ and $\tilde{P}(A)$ have exactly the same set of* ***equilibrium outcome weights****:*

$$(\forall \boldsymbol{w})\,(\forall \phi_j \in \Phi,\ \textit{Regret}^{\phi_j}_{i,\boldsymbol{w}}(\hat{P}(A)) \leq 0 \leftrightarrow \forall \phi_j \in \Phi,\ \textit{Regret}^{\phi_j}_{i,\boldsymbol{w}}(\tilde{P}(A)) \leq 0).$$

That is, there are no weights $\boldsymbol{w}$ *where either $\tilde{P}(A)$ or $\hat{P}(A)$ is in equilibrium and the other is not.*

Importantly, though, since the features and deviation policies considered may only approximate the strategic considerations of the game being played, weights providing an equilibrium are not *assumed*. If demonstrated strategy $\tilde{P}(A)$ is not in equilibrium for any choice of weights $\mathbf{w}$, these constraints assure $\hat{P}(A)$ matches the same demonstrated regrets.

Given demonstrated strategies from multiple games in some set $\mathcal{G}$, the previous guarantees can be easily extended to be in terms of the average of regrets over all the games: $\frac{1}{|\mathcal{G}|}\sum_{G\in\mathcal{G}} \text{Regret}^{\phi_j(G)}_{i,w}(\tilde{P}_G(A), G)$. For brevity, we do not make this explicit in our notation.

Many potential estimated strategies $\hat{P}(A)$ satisfy the constraints of Equation 5 (*e.g.*, trivially, $\tilde{P}(A)$ does). However, most do not generalize well. A secondary criterion is needed to select a strategy distribution with predictive qualities in a way that leads to strategy transfer across games.

3.3. Maximizing entropy

The **principle of maximum entropy** (Jaynes, 1957) advocates estimating a probability distribution by selecting the distribution with the fewest additional assumptions possible beyond matching specified properties of empirical samples. The notion of additional assumptions is quantified using Shannon's **information entropy** (Shannon, 1948), $H(P(X)) \triangleq -\sum_{x\in\mathcal{X}} P(x)\log_2 P(x)$.

For the deviation regret-matching strategy setting (Theorem 1), the following optimization is prescribed by the principle of maximum entropy.

Definition 5. *A* ***maximum entropy deviation regret-matching strategy*** *(MaxEnt DRM) for a set of deviation policies, Φ, is a probability distribution further defined by:*

$$\hat{P}_{ME}(A) \triangleq \underset{\hat{P}(A)}{\text{argmax}}\, H(\hat{P}(A)) \textit{ such that:} \quad (6)$$

$$(\forall \phi_j \in \Phi)\ \mathbb{E}_{\hat{P}(A)}\left[\sum_{a'_i\in A_i} \phi_j(a'_i|a_i)\,\boldsymbol{f}_i(a'_i, a_{-i}) - \boldsymbol{f}_i(a)\right] = \mathbb{E}_{\tilde{P}(A)}\left[\sum_{a'_i\in A_i} \phi_j(a'_i|a_i)\,\boldsymbol{f}_i(a'_i, a_{-i}) - \boldsymbol{f}_i(a)\right].$$

The maximum entropy approach provides a useful worst-case predictive guarantee.

Theorem 2 ((Grünwald & Dawid, 2003))**.** *A maximum entropy distribution, $\hat{P}_{ME}$ (Definition 5), minimizes the worst-case predictive log-loss, $E_{\tilde{P}(A)}[-\log \hat{P}(a)]$, when predicting an unknown empirical distribution $\tilde{P}(A)$ constrained so that distribution estimate $\hat{P}(A)$ matches specified properties of $\tilde{P}(A)$.*

This general result is the underlying motivation for many state-of-the-art probabilistic techniques, such as logistic regression, Markov random fields, and conditional random fields (Lafferty et al., 2001).

3.4. MaxEnt DRM distribution

It is insightful to express the MaxEnt DRM strategy $\hat{P}_{ME}$ in terms of the Lagrange multipliers, θ, for the constraints of Definition 5. These multipliers can be employed for strategy transfer in any game with the same types of features and deviation policies.

Theorem 3. *The maximum entropy deviation regret-matching strategy (Definition 5) has the following parametric form:*

$$P(a) \propto e^{\sum_{\phi_{j,(i)}\in\Phi} \theta_j^T \left(\sum_{a'_i\in A_i} \phi_j(a'_i|a_i)\,\boldsymbol{f}_i(a'_i, a_{-i}) - \boldsymbol{f}_i(a)\right)} \quad (7)$$

where model parameters $\{\theta_j\}$ are parameters causing the distribution to satisfy the constraints of Definition 5.

Similar to the outcome-parameterized weights $\mathbf{w}$ (Definition 3), the MaxEnt DRM model's parameters θ can be loosely interpreted as specifying deviation-specific feature-based payoffs. Under the assumption that the deviation policies across games correspond to behavior with similar outcome features with similar desirability, the model provides good generalization across different games.

3.5. Generalized deviation policies

Many types of feature-based deviation policies (Definition 4) exist. A key aspect to note is that the deviation policies need not be accurate reflections of how an agent would actually choose to deviate. Indeed, deviation policies to actions that an agent would avoid are of significant importance for learning strategy estimates that similarly avoid those particular actions. We discuss a few methods for constructing deviation policies in this section.

When commonalities exist in the actions of all source and target games, these can be encoded as features and used for deviations. For example, a "walk away" action is typically available in negotiations that prevents any deal from being reached. A simple and natural deviation policy to consider is the "always walk away" policy, which can be realized using features by adding a single indicator function to game outcomes for each player's "walk away" action.

For coordinated behavior, conditional deviation policies are also useful to consider. In our experiments, we employ a general Gibbs measure to generate deviation functions[2] of the form:

$$\phi_j(a_i'|a_i) \propto e^{\mathbb{E}_{\bar{P}_j(A_{-i}|a_i)}[\alpha^{\mathrm{T}}\mathbf{f}_{a_i',a_{-i}}]} \tag{8}$$

where $\bar{P}(a_{-i}|a_i)$ is some assumed distribution (*e.g.*, uniform).

3.6. Learning algorithms

The coordinated formulation of our approach provides advantageous convexity properties (Theorem 4) that enable efficient learning through well-studied gradient-based optimization procedures.

Theorem 4. *The log-likelihood function of the MaxEnt DRM parameters θ (Equation 7) is convex.*

Due to this convexity, the MaxEnt DRM model's parameters (i.e., the log-likelihood of Equation 7) can be optimized using gradient-based techniques.

Theorem 5. *The gradient for the MaxEnt DRM model's log-likelihood is:* $\nabla_\theta \log L(\theta|\tilde{P}) =$

$$\left\{ \mathbb{E}_{\tilde{P}(A)} \left[\sum_{a_i' \in A_i} \phi_j(a_i'|a_i) \boldsymbol{f}_i(a_i', a_{-i}) - \boldsymbol{f}_i(a) \right] - \mathbb{E}_{\hat{P}(A)} \left[\sum_{a_i' \in A_i} \phi_j(a_i'|a_i) \boldsymbol{f}_i(a_i', a_{-i}) - \boldsymbol{f}_i(a) \right] \right\}$$

for the vector of parameters θ_j associated with deviation policy $\phi_j(a_i'|a_i)$.

The corresponding gradient for demonstrated strategies from multiple games can be obtained by simply adding together the sample-size-weighted gradients of each game.

Gradient-based parameter updates, $\theta_{t+1} \leftarrow \theta_t - \gamma_t \nabla_\theta \log L(\theta_t|\tilde{P})$, with an adaptive learning rate, γ_t, are guaranteed to converge to the global optima as a consequence of Theorem 4. When the space of actions (and parameters) is reasonably sized, the expectations can be computed efficiently in a straight-forward manner according to Equation 7. Simulation-based approaches, e.g., Markov chain Monte Carlo, can be employed to approximate those expectations in large or infinite action spaces.

[2]Stochastic deviation policies do not provide more representational power than deterministic deviation policies. However, it can be convenient to specify randomized deviation policies to relate different games.

4. Experiments

We now perform strategy prediction experiments to investigate the benefits of the MaxEnt DRM approach when learning to predict strategic behavior.

4.1. The treasure hunt coordination game

Treasure hunt games are a class of simultaneous two-player game specified as $G_{TH} = (\{\mathbf{t}_s\}, \{\mathbf{w}_i\}, T_{\text{heavy}})$ in which each player i chooses a treasure-hunting site, $s \in S$ that contains treasures $\mathbf{t}_s$ that provide player-dependent utility $\mathbf{w}_i^{\mathrm{T}}\mathbf{t}_s$ if collected. The game is strategic because players split the treasures when they choose the same site, but are able to collect heavy types of treasures from set T_{heavy} that require both players to move. The game is fully known to each player[3]. To extensively investigate the predictive capabilities of our approach, we employ a large amount of synthetic strategy data rather than a small amount of human subject strategy data.

Table 1. Payoff features for the two players in the coordinated treasure hunting site selection problem with two sites. The payoff for each player i is obtained from the dot product of the table vectors with $(\mathbf{w}_i^\top : \mathbf{1}_{t_i \notin T_{\text{heavy}}}; \mathbf{w}_i^\top : \mathbf{1}_{t_i \in T_{\text{heavy}}})$, where ":" represents the element-wise Frobenius product with the indicator function.

	Site 1	Site 2
Site 1	$\begin{pmatrix} \mathbf{t}_1/2 \\ \mathbf{t}_1/2 \end{pmatrix}, \begin{pmatrix} \mathbf{t}_1/2 \\ \mathbf{t}_1/2 \end{pmatrix}$	$\begin{pmatrix} \mathbf{t}_1 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{t}_2 \\ 0 \end{pmatrix}$
Site 2	$\begin{pmatrix} \mathbf{t}_2 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{t}_1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} \mathbf{t}_2/2 \\ \mathbf{t}_2/2 \end{pmatrix}, \begin{pmatrix} \mathbf{t}_2/2 \\ \mathbf{t}_2/2 \end{pmatrix}$

We generate strategy distributions $P(A)$ using the (nearly) convergent solution of a subgradient-based correlated equilibrium solver. This is motivated by results showing simple adaptive strategies (such as no-regret learning) converge to some subset of correlated equilibria (Gordon et al., 2008). Depending on the players' weights on different treasures, many different interesting equilibria can result with strategies known from classical games, such as Chicken, the El Farol Bar Problem (Arthur, 1994), and Battle of the Sexes. Cooperation—by mutually choosing to treasure hunt together or alone and employing a randomized mechanism for fairly resolving preferences for sites—is common. However, adversarial behavior also arises when joining an-

[3]In our experiments, the treasures present in each site s, (*i.e.*, elements of $\mathbf{t}_s$) are drawn i.i.d. from Bernoulli(0.5) as is whether a particular type of treasure t_i is "heavy" $P(t_i \in T_{\text{heavy}}) = 0.5$. Players' payoffs for different treasures are drawn i.i.d. from Uniform$[0, 1]$. Except where noted, we consider the treasure hunt game with 5 sites and 8 types of treasures.

other player at a particular site would be unilaterally advantageous.

4.2. Strategy prediction metrics and models

In the strategy prediction task, we consider sets of treasure hunt games that differ only in the treasures present at sites $\{\mathbf{t}_s\}$. The predictor can only "see" the treasure types $\{\mathbf{t}_s\}$ that are in set T_{obs} and has neither knowledge of which types are heavy T_{heavy} nor player utilities $\{\mathbf{w}_i\}$. This is reflective of the strategy prediction setting in general, where ulterior motives beyond those of a game's design are known to influence behavior. Given strategy samples $\tilde{P}(a)$ drawn from $P(a)$ from some games in this set, the predictor must estimate the strategy $\hat{P}(a)$ for withheld games. The partial knowledge of the game and the non-uniqueness of correlated equilibria make this a challenging prediction task. We employ the Kullback-Leibler (KL) divergence, $D_{KL}(P(A)||\hat{P}(A))$, to assess the predictive performance of $\hat{P}(A)$. It measures the additional uncertainty implied by using distribution $\hat{P}(A)$ in place of distribution $P(A)$.

We compare the maximum entropy regret-matching approach against four other models: A **Nearest Neighbor** model, where the strategy corresponding to the "closest" source game, $\min_{G \in \mathcal{G}_{\text{source}}} \sum_s ||\mathbf{t}_s^G - \mathbf{t}_s^t||_1$, is predicted; an **Average Strategy** model where the average of source strategies $\frac{1}{|\mathcal{G}|} \sum_{G \in \mathcal{G}} P_G(A)$ is predicted; a joint action logistic regression (**Joint Logistic**) model that assumes behavior is centrally (but suboptimally) chosen according to distribution $P(a) \propto \exp\{\sum_{i \in N} \theta_i^{\mathrm{T}} \mathbf{f}_{a,i}\}$ with learned utility weights θ; and an opponent-oblivious logistic regression (**Individual Logistic**) model that ignores the strategic behavior of the other players by assuming it is uniformly distributed over all actions, yielding distribution $P(a_i) \propto \exp\{\theta_i^{\mathrm{T}} \sum_{a_{-i} \in A_{-i}} \mathbf{f}_a\}$. We employ a weakly-weighted Dirichlet prior (pseudo-count) of 0.02 over the multinomial distributions of actions for the Nearest Neighbor and Average Strategy approaches. The log-likelihoods for the logistic models are convex functions of θ and are optimized using standard gradient-based techniques.

4.3. Strategy prediction comparisons

We begin our experiments with idealized settings and then introduce more challenging assumptions that characterize the difficulties of the strategy transfer task. We randomly draw source games and 100 target games from this class of games and report the average KL-divergence between $\hat{P}(A)$ and $P(A)$ using 10 repeated experiments with different randomly drawn player outcome weight vectors. Thus, each quantity reported in our experimental results is based on 1,000 samples.

In our first set of experiments, we assume ideal observability—of features and source strategies. We adjust the number of source games to investigate the relation between the amount of available training data and the predictive performance for our baseline and MaxEnt DRM models with two deviation policy set sizes, $|\Phi| = \{2, 20\}$.

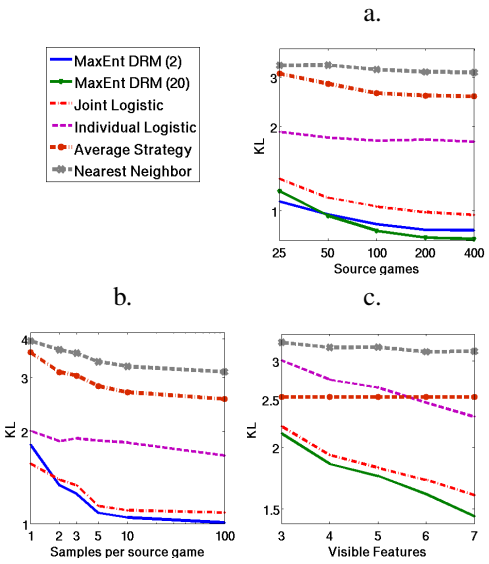

Figure 1. Predictive performance on: (a) withheld data in the ideal observability setting, $\tilde{P} = P$, $T_{\text{obs}} = T$ for varying numbers of source games; (b) the perfect feature observation setting ($T_{\text{obs}} = T$) and a varying number of per-game observed strategies ($\tilde{P} \neq P$); and (c) the imperfect feature observation setting ($T_{\text{obs}} \subset T, \tilde{P} = P$).

As shown by the experiments (Figure 1a), baseline approaches operating directly on strategies (Nearest Neighbor and Average Strategy) perform very poorly due to the relatively large number of action combinations in the treasure hunt game. Additionally, small differences in the the allocations of treasures lead to significant differences in strategies, making the distance-based approach inferior to the simpler strategy averaging approach when provided with limited amounts of training data. Of the two logistic approaches, the joint logistic approach provides lower loss than the individual logistic approach. This is reflective of the fact that many of the strategies generated for the treasure hunt game possess a large degree of coordination. While providing better performance than the direct strategy-based approaches, and attesting to the benefit of utility-related parameter learning, neither the joint nor the individual logistic regression approach outperforms the MaxEnt DRM approach with either a small (2) or large (20) number of deviation bases. Thus, the strategic considerations of the MaxEnt DRM approach are validated on this dataset.

Our second set of experiments relaxes the ideal observability assumptions. We use a source size of 50 games and vary the amount of observed sequences per game and then the

number of features observed. The predictive performance (Figure 1b-c) degrades in these settings, as expected. With extremely little amounts of data (one strategy sample per observed game), the joint logistic model outperforms the MaxEnt DRM model, but this relationship quickly reverses as the amount of data increases. In the visibility-varying experiment, the MaxEnt DRM continues to perform best for each amount of visible features.

Table 2. Transfer learning performance on qualitatively different target game.

	MaxEnt DRM	Joint Logistic	Ind. Logistic
KL	1.31	1.70	2.47

Our third and final experiment (Table 2) evaluates the predictive performance of each approach on a very different target game setting with 8 treasure sites (instead of 5) and a different per-site treasure probability, drawn from Bernoulli(0.25). A source dataset of 50 games was employed. Note that direct strategy-based learning approaches (Nearest Neighbor and Average Strategy) cannot be applied due to the differences in source and target strategy spaces. We find an even larger disparity between the MaxEnt DRM approach and the joint logistic model in this scenario, further demonstrating its value for generalized strategy prediction.

5. Future Work

A number of future directions are encouraged by the principled combination of game theory with machine learning in this paper. More sophisticated deviation policy selection is likely to prove fruitful, making the addition of search or sparsifying regularization techniques one interesting direction for future investigation. Extension of this approach beyond normal-form games to sequential games and extensive form games is another important direction. Finally, we plan to establish empirical validation of this approach on real-world data—particularly on observed behavior where game formulations do not yet exist.

Acknowledgements

This research has been funded by ARO MURI award number W911NF0810301 and the Office of Naval Research through the Distributed Reasoning in Reduced Information Spaces MURI.

References

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

Altman, A., Bercovici-Boden, A., and Tennenholtz, M. Learning in one-shot strategic form games. *Machine Learning: ECML 2006*, pp. 6–17, 2006.

Arthur, W.B. Inductive reasoning and bounded rationality. *The American economic review*, 84(2):406–411, 1994.

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

Cooper, D. and Kagel, J.H. Lessons learned: generalizing learning across games. *The American Economic Review*, 93(2):202–207, 2003. ISSN 0002-8282.

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

Gintis, H. *The bounds of reason: game theory and the unification of the behavioral sciences*. Princeton University Press, 2009. ISBN 0691140529.

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

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.

Halpern, J.Y. and Pass, R. Iterated regret minimization: A new solution concept. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-09)*, pp. 153–158, 2009.

Harsanyi, J.C. Games with incomplete information played by "Bayesian" players, i-iii. part i. the basic model. *Management science*, 14(3):159–182, 1967. ISSN 0025-1909.

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

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

McKelvey, R.D. and Palfrey, T.R. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1): 6–38, 1995. ISSN 0899-8256.

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

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

Shannon, C. E. A mathematical theory of communication. *Bell system technical journal*, 27, 1948.

Sharma, M., Holmes, M., Santamaria, J., Irani, A., Isbell, C., and Ram, A. Transfer learning in real-time strategy games using hybrid CBR/RL. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, 2007.

Waugh, Kevin, Ziebart, Brian D., and Bagnell, J. Andrew. Computational rationalization: The inverse equilibrium problem. In *Proc. of the International Conference on Machine Learning*, 2011.

Wright, J.R. and Leyton-Brown, K. Beyond equilibrium: Predicting human behavior in normal-form games. In *Twenty-Fourth Conference on Artificial Intelligence (AAAI-10)*, 2010.

Yang, Z. Correlated equilibrium and the estimation of discrete games of complete information. Working paper, http://www.econ.vt.edu/faculty/2008vitas_research/joeyang_research.htm, 2009.

Ziebart, B. D., Bagnell, J. A., and Dey, A. K. Modeling interaction via the principle of maximum causal entropy. In *Proc. ICML*, pp. 1255–1262, 2010.

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

A. Proofs of Theorems

Proof of Theorem 1. First, note that:

$$\begin{aligned}
&\mathbf{w}^{\mathrm{T}}\mathbb{E}_{P(A)}\left[\sum_{a_i'\in A_i}\phi_j(a_i'|a_i)\,\mathbf{f}_i(a_i',a_{-i})-\mathbf{f}_i(a)\right]\\
&=\mathbb{E}_{P(A)}\left[\sum_{a_i'\in A_i}\phi_j(a_i'|a_i)\,\mathbf{w}^{\mathrm{T}}\mathbf{f}_i(a_i',a_{-i})-\mathbf{w}^{\mathrm{T}}\mathbf{f}_i(a)\right]\\
&=\text{Regret}_{i,\mathbf{w}}^{\phi_j}(P(A)).
\end{aligned}$$

Using this, we can easily show Equation 3 if Equation 5. Multiplying both sides of Equation 5 by any $\mathbf{w}^{\mathrm{T}}$, we have:

$$\text{Regret}_{i,\mathbf{w}}^{\phi_j}(\hat{P}(A))=\text{Regret}_{i,\mathbf{w}}^{\phi_j}(\tilde{P}(A)).$$

Next, we show not Equation 3 if not Equation 5 to complete the proof. Assuming not Equation 5:

$$\begin{aligned}
&\mathbb{E}_{\hat{P}(A)}\left[\sum_{a_i'\in A_i}\phi(a_i'|a_i)\,\mathbf{f}_i(a_i',a_{-i})-\mathbf{f}_i(a)\right] \qquad (9)\\
&-\mathbb{E}_{\tilde{P}(A)}\left[\sum_{a_i'\in A_i}\phi(a_i'|a_i)\,\mathbf{f}_i(a_i',a_{-i})-\mathbf{f}_i(a)\right]\neq\mathbf{0}.
\end{aligned}$$

That is, at least one entry, k, of this vector must be non-zero. Let $\mathbf{w}'=\mathbf{e}_k$, where $\mathbf{e}_k$ is the vector where the k^{th} entry is 1 and the others are 0. Multiplying Equation 9 by $\mathbf{w}'^{\mathrm{T}}$, we have:

$$\begin{aligned}
&\text{Regret}_{i,\mathbf{w}'}^{\phi_j}(\hat{P}(A))-\text{Regret}_{i,\mathbf{w}'}^{\phi_j}(\tilde{P}(A))\neq 0\\
&\Rightarrow\exists\mathbf{w}:\text{Regret}_{i,\mathbf{w}}^{\phi_j}(\hat{P}(A))\neq\text{Regret}_{i,\mathbf{w}}^{\phi_j}(\tilde{P}(A)).
\end{aligned}$$

□

Proof of Theorem 3. First, we note that $\tilde{P}(A)$ is a feasible solution to the constraints of the convex optimization in Definition 5. We restrict our consideration to mixed strategies, which reside on the relative interior of the convex constraint set. Thus, by Slater's condition, strong duality holds and no duality gap exists.

Ignoring the probabilistic non-negativity constraint (which will directly follow), the Lagrangian has the form:

$$\begin{aligned}
\Lambda(\theta,Z,P(A))=&H(P(A))+Z\left(1-\sum P(A)\right)+\\
&\mathbb{E}_{P(A)}\left[\sum_j\theta_j^{\mathrm{T}}\left(\sum_{a_i'}\phi_j(a_i'|a_i)\,\mathbf{f}_{a_i',a_{-i}}-\mathbf{f}_a\right)\right].
\end{aligned}$$

Differentiating with respect to $P(a)$ and pushing the additive constant term into Z we obtain:

$$-\log P(a)-Z+\sum_j\theta_j^{\mathrm{T}}\left(\sum_{a_i'}\phi_j(a_i'|a_i)\,\mathbf{f}_{a_i',a_{-i}}-\mathbf{f}_a\right).$$

After equating to zero, we find the distribution form to be:

$$P(a)=\frac{e^{\sum_j\theta_j^{\mathrm{T}}\left(\sum_{a_i'}\phi_j(a_i'|a_i)\,\mathbf{f}_{a_i',a_{-i}}-\mathbf{f}_a\right)}}{Z}$$

□

Proof of Theorem 4. This fact follows directly from convex duality, discussed in the proof of Theorem 3. □

Proof of Theorem 5. The log probability of a single demonstrated action $\tilde{a}$ is:

$$\begin{aligned}
\log\hat{P}(\tilde{a})=&\sum_{\phi_{j,(i)}\in\Phi}\theta_j^{\mathrm{T}}\left(\sum_{a_i'\in A_i}\phi_j(a_i'|\tilde{a}_i)\,\mathbf{f}_i(\tilde{a}_i',\tilde{a}_{-i})-\mathbf{f}_i(\tilde{a})\right)\\
&-\log\sum_a e^{\sum_{\phi_{j,(i)}\in\Phi}\theta_j^{\mathrm{T}}\left(\sum_{a_i'\in A_i}\phi_j(a_i'|a_i)\,\mathbf{f}_i(a_i',a_{-i})-\mathbf{f}_i(a)\right)}.
\end{aligned}$$

Differentiating with respect to θ_j, we have:

$$\begin{aligned}
\frac{\partial\log\hat{P}(\tilde{a})}{\partial\theta_j}=&\left(\sum_{a_i'\in A_i}\phi_j(a_i'|\tilde{a}_i)\,\mathbf{f}_i(a_i',\tilde{a}_{-i})-\mathbf{f}_i(\tilde{a})\right)\\
&-\sum_{a\in A}\hat{P}(a)\left(\sum_{a_i'\in A_i}\phi_j(a_i'|a_i)\,\mathbf{f}_i(a_i',a_{-i})-\mathbf{f}_i(a)\right)
\end{aligned}$$

Thus, because $\log L(\theta|\tilde{P}(A))=-\sum_{a\in\mathcal{A}}\tilde{P}(a)\log\hat{P}(a)$:

$$\begin{aligned}
\frac{\partial\log L(\theta|\tilde{P}(A))}{\partial\theta_j}=&\mathbb{E}_{\hat{P}(A)}\left[\sum_{a_i'\in A_i}\phi_j(a_i'|a_i)\,\mathbf{f}_i(a_i',a_{-i})-\mathbf{f}_i(a)\right]\\
&-\mathbb{E}_{\tilde{P}(A)}\left[\sum_{a_i'\in A_i}\phi_j(a_i'|a_i)\,\mathbf{f}_i(a_i',a_{-i})-\mathbf{f}_i(a)\right]
\end{aligned}$$

□