6-2011

Computational Rationalization: The Inverse Equilibrium Problem

Kevin Waugh
Carnegie Mellon University

Brian D. Ziebart
Carnegie Mellon University

J. Andrew Bagnell
Carnegie Mellon University

Published In
Proceedings of the 28 th International Conference on Machine Learning.

Computational Rationalization: The Inverse Equilibrium Problem

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

Abstract

Modeling the purposeful behavior of imperfect agents from a small number of observations is a challenging task. When restricted to the single-agent decision-theoretic setting, inverse optimal control techniques assume that observed behavior is an approximately optimal solution to an unknown decision problem. These techniques learn a utility function that explains the example behavior and can then be used to accurately predict or imitate future behavior in similar observed or unobserved situations.

In this work, we consider similar tasks in competitive and cooperative multi-agent domains. Here, unlike single-agent settings, a player cannot myopically maximize its reward — it must speculate on how the other agents may act to influence the game's outcome. Employing the game-theoretic notion of regret and the principle of maximum entropy, we introduce a technique for predicting and generalizing behavior, as well as recovering a reward function in these domains.

1. Introduction

Predicting the actions of others in complex and strategic settings is an important facet of intelligence that guides our interactions—from walking in crowds to negotiating multi-party deals. Recovering such behavior from merely a few observations is an important and challenging machine learning task.

While mature computational frameworks for decision-making have been developed to **prescribe** the behavior that an agent *should* perform, such frameworks are often ill-suited for **predicting** the behavior that an agent *will* perform. Foremost, the standard assumption of decision-making frameworks that a criteria for preferring actions (*e.g.*, costs, motivations and goals) is known *a priori* often does not hold. Moreover, real behavior is typically not consistently optimal or completely rational; it may be influenced by factors that are difficult to model or subject to various types of error when executed. Meanwhile, the standard tools of statistical machine learning (*e.g.*, classification and regression) may be equally poorly matched to modeling purposeful behavior; an agent's goals often succinctly, but implicitly, encode a strategy that would require tremendous amounts of data to learn.

A natural approach to mitigate the complexity of recovering a full strategy for an agent is to consider identifying a compactly expressed utility function that *rationalizes* observed behavior: that is, identify rewards for which the demonstrated behavior is optimal and then leverage these rewards for future prediction. Unfortunately, the problem is fundamentally ill-posed: in general, many reward functions can make behavior seem rational, and in fact, the trivial, everywhere 0 reward function makes **all** behavior appear rational (Ng & Russell, 2000). Further, after removing such trivial reward functions, there may be **no** reward function for which the demonstrated behavior is optimal as agents may be imperfect and the real world they operate in may be only approximately represented.

In the single-agent decision-theoretic setting, inverse optimal control methods have been used to bridge this gap between the prescriptive frameworks and predictive applications (Abbeel & Ng, 2004; Ratliff et al., 2006; Ziebart et al., 2008a; 2010). Successful applications include learning and prediction tasks in personalized vehicle route planning (Ziebart et al., 2008a), robotic crowd navigation (Henry et al., 2010), quadruped foot placement and grasp selection (Ratliff et al., 2009). A reward function is learned by these techniques that both explains demonstrated behavior

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

and approximates the optimality criteria of prescriptive decision-theoretic frameworks.

As these methods only capture a single reward function and do not reason about competitive or cooperative motives, inverse optimal control proves inadequate for modeling the strategic interactions of multiple agents. In this paper, we consider the game-theoretic concept of regret as a necessary stand-in for the optimality criteria of the single-agent work. As with the inverse optimal control problem, the result is fundamentally ill-posed. We address this by requiring that for any utility function linear in known features, our learned model must have no more regret than that of the observed behavior. We demonstrate that this requirement can be re-cast as a set of equivalent convex constraints that we denote the *inverse correlated equilibrium* (ICE) polytope.

As we are interested in the effective prediction of behavior, we will use a maximum entropy criteria to select behavior from this polytope. We demonstrate that optimizing this criteria leads to mini-max optimal prediction of behavior subject to approximate rationality. We consider the dual of this problem and note that it generalizes the traditional log-linear maximum entropy family of problems (Della Pietra et al., 2002). We provide a simple and computationally efficient gradient-based optimization strategy for this family and show that only a small number of observations are required for accurate prediction and transfer of behavior. We conclude by considering a matrix routing game and compare the ICE approach to a variety of natural alternatives.

Before we formalize imitation learning in matrix games, motivate our assumptions and describe and analyze our approach, we will review the game-theoretic notions of regret and the correlated equilibrium.

2. Game Theory Background

Matrix games are the canonical tool of game theorists for representing strategic interactions ranging from illustrative toy problems, such as the "Prisoner's Dilemma" and the "Battle of the Sexes" games, to important negotiations, collaborations, and auctions. In this work, we employ a class of games with payoffs or utilities that are linear functions of features defined over the outcome space.

Definition 1. *A **linearly parameterized normal-form game**, or **matrix game**, $\Gamma = (N, \mathcal{A}, F)$, is composed of: a finite set of **players**, N; a set of **joint-actions** or **outcomes**, $\mathcal{A} = \times_{i \in N} A_i$, consisting of a finite set of **actions** for each player, A_i; a set of **outcome features**, $F = \{\theta_a^i \in \mathbb{R}^K\}$ for each outcome that induce a **parameterized utility function**, $u_i(a|w) = {\theta_a^i}^{\mathrm{T}} w$ – the reward for player i achieving outcome a w.r.t. **utility weights** w.*

For notational convenience, we let a_{-i} denote the vector a excluding component i and let $\mathcal{A}_{-i} = \times_{j \neq i, j \in N} A_i$ be the set of such vectors.

In contrast to standard normal-form games where the utility functions for game outcomes are known, in this work we assume that "true" utility weights, w^*, which govern observed behavior, are unknown. This allows us to model real-world scenarios where a cardinal utility is not available or is subject to personal taste.

We model the players with a distribution $\sigma \in \Delta_{\mathcal{A}}$ over the game's joint-actions. Coordination between players can exist, thus, this distribution need not factor into independent strategies for each player. Conceptually, a signaling mechanism, such as a traffic light, can be thought to sample a joint-action from σ and communicate to each player a_i, its portion of the joint-action. Each player can then consider deviating from a_i using a **modification function**, $f_i : A_i \mapsto A_i$ (Blum & Mansour, 2007).

The **switch modification function**, for instance,

$$\text{switch}_i^{x \to y}(a_i) = \begin{cases} y & \text{if } a_i = x \\ a_i & \text{otherwise} \end{cases} \tag{1}$$

substitutes action y for recommendation x.

Instantaneous regret measures how much a player would benefit from a particular modification function when the coordination device draws joint-action a,

$$\begin{aligned} \text{regret}_i(a|f_i, w) &= u_i(f_i(a_i), a_{-i}|w) - u_i(a|w) && (2) \\ &= \left[\theta^i_{f_i(a_i), a_{-i}} - \theta^i_{a_i, a_{-i}}\right]^{\mathrm{T}} w && (3) \\ &= r_{i,a}^{f_i \mathrm{T}} w. && (4) \end{aligned}$$

Players do not have knowledge of the complete joint-action; thus, each must reason about the **expected regret** with respect to a modification function,

$$\begin{aligned} \sigma^{\mathrm{T}} R_i^{f_i} w &= \mathbb{E}_{a \sim \sigma}\left[\text{regret}_i(a|f_i, w)\right] && (5) \\ &= \sum_{a \in \mathcal{A}} \sigma_a r_{i,a}^{f_i \mathrm{T}} w. && (6) \end{aligned}$$

It is helpful to consider regret with respect to a class of modification functions. Two classes are particularly important for our discussion. First, **internal regret** corresponds to the set of modification functions where a single action is replaced by a new action, $\Phi_i^{\text{int}} =$

$\{\text{switch}_i^{x\to y}(\cdot) : \forall x, y \in A_i\}$. Second, **swap regret** corresponds to the set of all modification functions, $\Phi_i^{\text{swap}} = \{f_i\}$. We denote $\Phi = \cup_{i\in N} \Phi_i$.

The **expected regret with respect to Φ** and outcome distribution σ,

$$R^{\Phi}(\sigma, w) = \max_{f_i \in \Phi} \mathbb{E}_{a\sim\sigma} \left[\text{regret}_i(a|f_i, w)\right], \qquad (7)$$

is important for understanding the incentive to deviate from, and hence the stability of, the specified behavior. The most general modification class, Φ^{swap}, leads to the notion of **ε-correlated equilibrium** (Osborne & Rubinstein, 1994), in which σ satisfies $R^{\Phi^{\text{swap}}}(\sigma, w^*) \leq \varepsilon$. Thus, regret can be thought of as a substitute for utility when assessing the optimality of behavior in multi-agent settings.

3. Imitation Learning in Matrix Games

We are now equipped with the tools necessary to introduce our approach for imitation learning in multi-agent settings. As input, we observe a sequence of outcomes, $\{a_m\}_{m=1}^M$, sampled from σ, the **true behavior**. We denote the empirical distribution of this sequence, $\tilde{\sigma}$, the **demonstrated behavior**. We aim to learn a **predictive behavior** distribution, $\hat{\sigma}$ from these demonstrations. Moreover, we would like our learning procedure to extract the motives and intent for the behavior so that we may imitate the players in similarly structured, but unobserved games.

Imitation appears hard barring further assumptions. In particular, if the agents are unmotivated or their intentions are not coerced by the observed game, there is little hope of recovering principled behavior in a new game. Thus, we require some form of rationality.

3.1. Rationality Assumptions

We say that agents are *rational* under their true preferences when they are indifferent between $\hat{\sigma}$ and their true behavior if and only if $R^{\Phi}(\hat{\sigma}, w^*) \leq R^{\Phi}(\sigma, w^*)$.

As agents' true preferences w^* are unknown to the observer, we must consider an encompassing assumption that requires any behavior that we estimate to satisfy this property for all possible utility weights, or

$$\forall w \in \mathbb{R}^K,\ R^{\Phi}(\hat{\sigma}, w) \leq R^{\Phi}(\sigma, w). \qquad (8)$$

Any behavior achieving this restriction, *strong rationality*, is also rational, and, by virtue of the contrapositive, we see that unless we have additional information regarding the agents' true preferences, we must assume this strong assumption or we risk violating rationality.

Lemma 1. *If strong rationality does not hold for alternative behavior $\hat{\sigma}$ then there exist agent utilities such that they would prefer σ to $\hat{\sigma}$.*

By restricting our attention to behavior that satisfies strong rationality, at worst, agents acting according to unknown true preference w^* will be indifferent between our predictive distribution and their true behavior.

3.2. Inverse Correlated Equilibria

Unfortunately, a direct translation of the strong rationality requirement into constraints on the distribution $\hat{\sigma}$ leads to a non-convex optimization problem as it involves products of varying utility vectors and the behavior to be estimated. Fortunately, however, we can provide an equivalent concise convex description of the constraints on $\hat{\sigma}$ that ensures any feasible distribution satisfies strong rationality. We denote this set of equivalent constraints as the *Inverse Correlated Equilibria* (ICE) polytope:

Definition 2 (ICE Polytope)**.**

$$\hat{\sigma}^{\mathrm{T}} R_i^{f_i} = \sum_{f_j \in \Phi} \eta_{f_j}^{f_i} \tilde{\sigma}^{\mathrm{T}} R_j^{f_j}, \forall f_i \in \Phi \qquad (9)$$
$$\eta^{f_i} \in \Delta_{\Phi}, \forall f_i \in \Phi; \qquad \hat{\sigma} \in \Delta_{\mathcal{A}}.$$

Theorem 1. *A distribution, $\hat{\sigma}$, satisfies the constraints above for some η if and only if it satisfies strong rationality. That is, $\forall w \in \mathbb{R}^K$, $R^{\Phi}(\hat{\sigma}, w) \leq R^{\Phi}(\tilde{\sigma}, w)$ if and only if $\forall f_i \in \Phi, \exists \eta^{f_i} \in \Delta_{\Phi}$ such that $\hat{\sigma}^{\mathrm{T}} R_i^{f_i} = \sum_{f_j \in \Phi} \eta_{f_j}^{f_i} \tilde{\sigma}^{\mathrm{T}} R_j^{f_j}$.*

The proof of Theorem 1 is provided in the Appendix (Waugh et al., 2011).

We note that this polytope, perhaps unsurprisingly, is similar to the polytope of correlated equilibrium itself, but here is defined in terms of the behavior we observe instead of the (unknown) reward function. Given any observed behavior σ, the constraints are feasible as the demonstrated behavior satisfies them; our goal is to choose from these behaviors without estimating a full joint-action distribution. While the ICE polytope establishes a basic requirement for estimating rational behavior, there are generally infinitely many distributions consistent with its constraints.

3.3. Principle of Maximum Entropy

As we are interested in the problem of statistical prediction of strategic behavior, we must find a mechanism to resolve the ambiguity remaining after accounting for the rationality constraints. The **principle of maximum entropy** provides a principled method for

choosing such a distribution. This choice leads to not only statistical guarantees on the resulting predictions, but to efficient optimization.

The Shannon **entropy** of a distribution $\hat{\sigma}$ is defined as $H(\hat{\sigma}) = -\sum_{x \in \mathcal{X}} \hat{\sigma}_x \log \hat{\sigma}_x$. The **principle of maximum entropy** advocates choosing the distribution with maximum entropy subject to known (linear) constraints (Jaynes, 1957):

$$\sigma_{\text{MaxEnt}} = \operatorname*{argmax}_{\hat{\sigma} \in \Delta_{\mathcal{X}}} H(\hat{\sigma}), \quad \text{subject to:} \qquad (10)$$
$$g(\hat{\sigma}) = 0 \text{ and } h(\hat{\sigma}) \leq 0.$$

The resulting log-linear family of distributions (*e.g.*, logistic regression, Markov random fields, conditional random fields) are widely used within statistical machine learning. For our problem, the constraints are precisely that the distribution is in the ICE polytope, ensuring that whatever distribution is learned has no more regret than the demonstrated behavior.

Importantly, the maximum entropy distribution subject to our constraints enjoys the following guarantee:

Lemma 2. *The maximum entropy ICE distribution minimizes over all strongly rational distributions the worst-case log-loss , $-\sum_{a \in \mathcal{A}} \sigma_a \log \hat{\sigma}_a$, when σ is chosen adversarially and subject to strong rationality.*

The proof of Lemma 2 follows immediately from the result of Grünwald and Dawid (2003).

In the context of multi-agent behavior, the principle of maximum entropy has been employed to obtain correlated equilibria with predictive guarantees in normal-form games when the utilities are known *a priori* (Ortiz et al., 2007). We will now leverage its power with our rationality assumption to select predictive distributions in games where the utilities are unknown.

3.4. Prediction of Behavior

Let us first consider prediction of the demonstrated behavior using the principle of maximum entropy and our strong rationality condition. After, we will extend to behavior transfer and analyze the error introduced as a by-product of sampling $\tilde{\sigma}$ from σ.

The mathematical program that maximizes the entropy of $\hat{\sigma}$ under strong rationality with respect to $\tilde{\sigma}$,

$$\operatorname*{argmax}_{\hat{\sigma}, \eta} H(\hat{\sigma}), \quad \text{subject to:} \qquad (11)$$
$$\hat{\sigma}^{\mathrm{T}} R_i^{f_i} = \sum_{f_j \in \Phi} \eta_{f_j}^{f_i} \tilde{\sigma}^{\mathrm{T}} R_j^{f_j}, \forall f_i \in \Phi$$
$$\eta^{f_i} \in \Delta_{\Phi}, \forall f_i \in \Phi; \qquad \hat{\sigma} \in \Delta_{\mathcal{A}},$$

is convex with linear constraints, feasible, and bounded. That is, it is simple and can be efficient solved in this form. Before presenting our preferred dual optimization procedure, however, let us describe an approach for behavior transfer that further illustrates the advantages of this approach over directly estimating σ.

3.5. Transfer of Behavior

A principal justification of inverse optimal control techniques that attempt to identify behavior in terms of utility functions is the ability to consider what behavior might result if the underlying decision problem were changed while the interpretation of features into utilities remain the same (Ng & Russell, 2000; Ratliff et al., 2006). This enables prediction of agent behavior in a no-regret or agnostic sense in problems such as a robot encountering novel terrain (Silver et al., 2010) as well as route recommendation for drivers traveling to unseen destinations (Ziebart et al., 2008b).

Econometricians are interested in similar situations, but for much different reasons. Typically, they aim to validate a model of market behavior from observations of product sales. In these models, the firms assume a fixed pricing policy given known demand. The econometrician uses this fixed policy along with product features and sales data to estimate or bound both the consumers' utility functions as well as unknown production parameters, like markup and production cost (Berry et al., 1995; Nevo, 2001; Yang, 2009). In this line of work, the observed behavior is considered accurate to start with; it is not suitable for settings with limited observations.

Until now, we have considered the problem of identifying behavior in a single game. We note, however, that our approach enables behavior *transfer* to games equipped with the same features. We denote this unobserved game as $\bar{\Gamma}$. As with prediction, to develop a technique for behavior transfer we assume a link between regret and the agents' preferences across the known space of possible preferences. Furthermore, we assume a relation between the regrets in both games.

Property 1 (Transfer Rationality)**.** *For some constant $\kappa > 0$,*

$$\forall w, \ \bar{R}^{\bar{\Phi}}(\bar{\sigma}, w) \leq \kappa R^{\Phi}(\sigma, w). \qquad (12)$$

Roughly, we assume that under preferences with low regret in the original game, the behavior in the unobserved game should also have low regret. By enforcing this property, if the agents are performing well with respect to their true preferences, then the transferred behavior will also be of high quality.

As we are not privileged to know κ and this property is not guaranteed to hold, we introduce a slack variable to allow for violations of the strong rationality constraints to guaranteeing feasibility. Intuitively, the *transfer-ICE polytope* we now optimize over requires that for any linear reward function and for every player, the predicted behavior in a new game must have no more regret than demonstrated behavior does in the observed game using the same parametric form of reward function. The corresponding mathematical program is:

$$\max_{\hat{\sigma},\eta,\nu} H(\hat{\sigma}) - C\nu, \quad \text{subject to:} \tag{13}$$
$$\hat{\sigma}^{\mathrm{T}} \bar{R}_i^{f_i} - \sum_{f_j \in \Phi} \eta_{f_j}^{f_i} \tilde{\sigma}^{\mathrm{T}} R_j^{f_j} \leq \nu, \forall f_i \in \bar{\Phi}$$
$$\sum_{f_j \in \Phi} \eta_{f_j}^{f_i} \tilde{\sigma}^{\mathrm{T}} R_j^{f_j} - \hat{\sigma}^{\mathrm{T}} \bar{R}_i^{f_i} \leq \nu, \forall f_i \in \bar{\Phi}$$
$$\eta^{f_i} \in \Delta_{\Phi}, \forall f_i \in \bar{\Phi}; \;\; \hat{\sigma} \in \Delta_{\mathcal{A}}; \;\; \nu \geq 0.$$

In the above formulation, $C > 0$ is a slack penalty parameter, which allows us to choose the trade-off between obeying the rationality constraints and maximizing the entropy. Additionally, we have omitted κ above by considering it intrinsic to R.

We observe that this program is almost identical to the behavior prediction program introduced above. We have simply made substitutions of the regret matrices and modification sets in the appropriate places. That is, if $\bar{\Gamma} = \Gamma$, we recover prediction with a slack.

Given $\hat{\sigma}$ and ν, we can bound the violation of the strong rationality constraint for any utility vector.

Lemma 3. *If $\hat{\sigma}$ violates the strong rationality constraints in the slack formulation by ν then for all w*

$$R^{\Phi}(\hat{\sigma}, w) \leq R^{\Phi}(\tilde{\sigma}, w) + \nu \left|\left|w\right|\right|_1. \tag{14}$$

One could choose to institute multiple slack variables, say one for each $f_i \in \bar{\Phi}$, instead of a single slack across all modification functions. Our choice is motivated by the interpretation of the dual multipliers presented in the next section. There, we will also address selection of an appropriate value for C.

4. Duality and Efficient Optimization

In this section, we will derive, interpret and describe a procedure for optimizing the dual program for solving the MaxEnt ICE problem. We will see that the dual multipliers can be interpreted as utility vectors and that optimization in the dual has computational advantages. We begin by presenting the dual of the transfer program.

Algorithm 1 Dual MaxEnt ICE

Input: $T, \gamma, C > 0$, $R, \bar{R}, \Phi$ and $\bar{\Phi}$
$\forall f_i \in \bar{\Phi}, \; \alpha^{f_i}, \beta^{f_i} \leftarrow 1/(|\bar{\Phi}|K + 1)$
for t from 1 to T **do**
  /* compute the gradient */
  $\forall a \in \bar{\mathcal{A}}, \; z_a \leftarrow \exp\left(-\sum_{f_i \in \bar{\Phi}} \bar{r}_{i,a}^{f_i \mathrm{T}} (\alpha^{f_i} - \beta^{f_i})\right)$
  $Z \leftarrow \sum_{a \in \bar{\mathcal{A}}} z_a$
  for $f_i \in \bar{\Phi}$ **do**
    $f_j^* \leftarrow \mathrm{argmax}_{f_j \in \Phi} \tilde{\sigma}^{\mathrm{T}} R_j^{f_j} (\alpha^{f_i} - \beta^{f_i})$
    $g^{f_i} \leftarrow \tilde{\sigma}^{\mathrm{T}} R_{j^*}^{f_j^*} - \sum_{a \in \bar{\mathcal{A}}} z_a \bar{r}_{i,a}^{f_i} / Z$
  end for
  /* descend and project */
  $\gamma_t \leftarrow \gamma/\sqrt{t}$
  $\rho \leftarrow 1 + \sum_{f_i,k} \alpha_k^{f_i} \exp(-\gamma_t g_k^{f_i}) + \beta_k^{f_i} \exp(\gamma_t g_k^{f_i})$
  $\forall f_i \in \bar{\Phi}, k \in K, \; \alpha_k^{f_i} \leftarrow C\alpha_k^{f_i} \exp(-\gamma_t g_k^{f_i})/\rho$
  $\forall f_i \in \bar{\Phi}, k \in K, \; \beta_k^{f_i} \leftarrow C\beta_k^{f_i} \exp(\gamma_t g_k^{f_i})/\rho$
end for
return (α, β)

$$\min_{\alpha,\beta,\xi} \sum_{f_i \in \bar{\Phi}} \max_{f_j \in \Phi} \left[\tilde{\sigma}^{\mathrm{T}} R_j^{f_j} (\alpha^{f_i} - \beta^{f_i})\right] + \log Z(\alpha, \beta)$$

$$\text{subject to: } \xi + \sum_{f_i \in \bar{\Phi}} \sum_{k=1}^{K} \alpha_k^{f_i} + \beta_k^{f_i} = C, \; \alpha, \beta, \xi \geq 0.$$

where $Z(\alpha, \beta)$ is the partition function,

$$Z(\alpha, \beta) = \sum_{a \in \bar{\mathcal{A}}} \exp\left(-\sum_{f_i \in \bar{\Phi}} \bar{r}_{i,a}^{f_i \mathrm{T}} (\alpha^{f_i} - \beta^{f_i})\right).$$

Removing the equality constraint is equivalent to disallowing any slack. We derive the dual in the appendix (Waugh et al., 2011).

For $C > 0$, the dual's feasible set has non-empty interior and is bounded. Therefore, by Slater's condition, strong duality holds – there is no duality gap. In particular, we can use a dual solution to recover $\hat{\sigma}$.

Lemma 4. *Given a dual solution, (α, β), we can recover the primal solution, $\hat{\sigma}$. Specifically,*

$$\hat{\sigma}_a = \exp\left(-\sum_{f_i \in \bar{\Phi}} \bar{r}_{i,a}^{f_i \mathrm{T}} (\alpha^{f_i} - \beta^{f_i})\right) / Z(\alpha, \beta). \tag{15}$$

Intuitively, the probability of predicting an outcome is small if that outcome has high regret.

In general, the dual multipliers are utility vectors associated with each modification function in $\bar{\Phi}$. Under

the slack formulation, there is a natural interpretation of these variables as a single utility vector. Given a dual solution, (α, β) with slack penalty C, we choose

$$\lambda^{f_i} = \alpha^{f_i} - \beta^{f_i}, \tag{16}$$

$$\pi^{f_i} = \frac{1}{C} \sum_{k=1}^{K} \alpha_k^{f_i} + \beta_k^{f_i}, \text{and} \tag{17}$$

$$\hat{w} = \sum_{f_i \in \bar{\Phi}} \pi^{f_i} \lambda^{f_i}. \tag{18}$$

That is, we can associate with each modification function a probability, π^{f_i}, and a utility vector, λ^{f_i}. Thus, a natural estimate for $\hat{w}$ is the expected utility vector. Note, $\sum_{f_i \in \bar{\Phi}} \pi^{f_i}$ need not sum to one. The remaining mass, ξ, is assigned to the zero utility vector.

The above observation implies that introducing a slack variable coincides with bounding the L_1 norm of the utility vectors under consideration by C. This insight suggests that we choose $C \geq ||w^*||_1$, if possible, as smaller values of C will exclude w^* from the feasible set. If a bound on the L_1 norm is not available, we may solve the prediction problem on the observed game without slack and use $||\hat{w}||_1$ as a proxy.

The dual formulation of our program has important inherent computational advantages. First, it is a optimization over a simple set that is particularly well-suited for gradient-based optimization, a trait not shared by the primal program. Second, the number of dual variables, $2|\Phi|K$, is typically much fewer than the number of primal variables, $|\mathcal{A}| + 2|\Phi|^2$. Though the work per iteration is still a function of $|\mathcal{A}|$ (to compute the partition function), these two advantages together let us scale to larger problems than if we consider optimizing the primal objective. Computing the expectations necessary to descend the dual gradient can leverage recent advances in the structured, compact game representations: in particular, any graphical game with low-treewidth or finite horizon Markov game (Kakade et al., 2003) enables these computations to be performed in time that scales only polynomially in the number of decision makers or time-steps.

Algorithm 1 employs exponentiated gradient descent (Kivinen & Warmuth, 1995) to find an optimal dual solution. The step size parameter, γ, is commonly taken to be $\sqrt{2 \log |\bar{\Phi}|K}/\Delta$, with Δ being the largest value in any $R_i^{f_i}$. With this step size, if the optimization is run for $T \geq 2\Delta^2 \log\left(|\bar{\Phi}|K\right)/\epsilon^2$ iterations then the dual solution will be within ϵ of optimal. Alternatively, one can exactly measure the duality gap on each iteration and halt when the desired accuracy is achieved. This is often preferred as the lower bound on the number of iterations is conservative in practice.

5. Sample Complexity

In practice, we do not have full access to the agents' true behavior – if we did, prediction would be straightforward and not require our estimation technique. Instead, we can only approximate it through finite observation of play. In real applications there are costs associated with gathering these observations and, thus, there are inherent limitations on the quality of this approximation. In this section, we will analyze the sensitivity of our approach to these types of errors.

First, although $|\mathcal{A}|$ is exponential in the number of players, our technique only accesses $\tilde{\sigma}$ through products of the form $\tilde{\sigma} R_j^{f_j}$. That is, we need only approximate these products accurately, not the distribution $\tilde{\sigma}$. As a result, we can bound the approximation error in terms of $|\Phi|$ and K.

Theorem 2. *With probability at least $1 - \delta$, for any w, by observing $M \geq \frac{2}{\epsilon^2} \log \frac{2|\Phi|K}{\delta}$ outcomes we have $R^{\Phi}(\tilde{\sigma}, w) \leq R^{\Phi}(\sigma, w) + \epsilon\Delta \, ||w||_1$.*

The proof is an application of Hoeffding's inequality and is provided in the Appendix (Waugh et al., 2011). As an immediate corollary, considering only the true, but unknown, reward function w^*:

Corollary 1. *With probability at least $1 - \delta$, by sampling according to the above rule, $R^{\Phi}(\hat{\sigma}, w^*) \leq R^{\Phi}(\sigma, w^*) + (\epsilon\Delta + \nu) \, ||w^*||_1$ for $\hat{\sigma}$ with slack ν.*

That is, so long as we assume bounded utility, with high probability we need only logarithmic many samples in terms of $|\Phi|$ and K to closely approximate $\sigma R_j^{f_j}$ and avoid a large violation of our rationality condition.

We note that choosing $\Phi = \Phi^{\text{int}}$ is particularly appealing, as $|\Phi^{\text{int}}| \leq |N|A^2$, compared to $|\Phi^{\text{swap}}| \leq |N|A!$. As internal regret closely approximates swap regret, we do not lose much of the strategic complexity by choosing the more limited set, but we require both fewer observations and fewer computational resources.

6. Experimental Results

To evaluate our approach experimentally, we designed a simple routing game shown in Figure 1. Seven drivers in this game choose how to travel home during rush hour after a long day at the office. The different road segments have varying capacities, visualized by the line thickness in the figure, that make some of them more or less susceptible to congestion or to traffic accidents. Upon arrival home, each driver records the total time and distance they traveled, the gas that they used, and the amount of time they spent stopped at intersections or in traffic jams – their utility features.

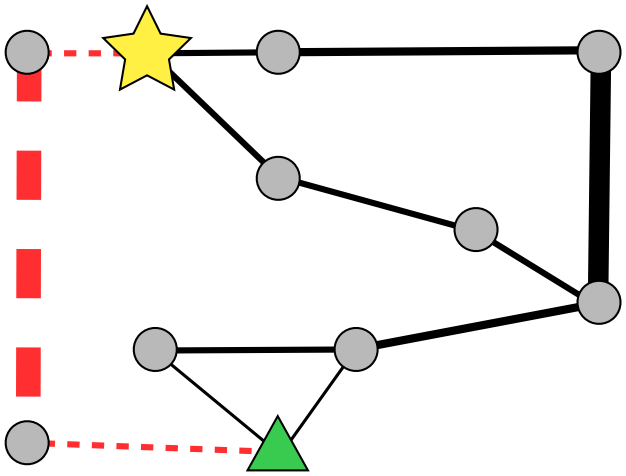

Figure 1. A visualization of the routing game.

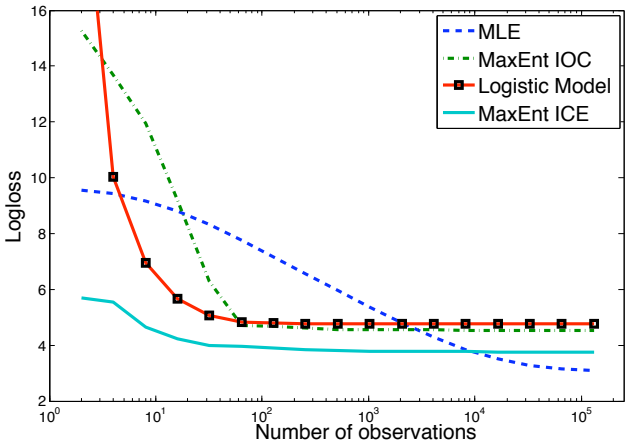

Figure 2. Prediction error (log-loss) as a function of number of observations.

In this game, each of the drivers chooses from four possible routes (solid lines in Figure 1), yielding over 16,000 possible outcomes. We obtained an ε-social welfare maximizing correlated equilibrium for those drivers where the drivers preferred mainly to minimize their travel time, but were also slightly concerned with gas usage. The demonstrated behavior $\tilde{\sigma}$ was sampled from this true behavior distribution σ.

First, we evaluate the differences between the true behavior distribution σ and the predicted behavior distribution $\hat{\sigma}$ trained from observed behavior sampled from $\tilde{\sigma}$. In Figure 2 we compare the prediction accuracy when varying the number of observations using log-loss, $-\sum_{a \in \mathcal{A}} \sigma_a \log \hat{\sigma}_a$. The baseline algorithms we compare against are: a maximum likelihood estimate of the distribution over the joint-actions with a uniform prior, an exponential family distribution parameterized by the outcome's utilities trained with logistic regression, and a maximum entropy inverse optimal control approach (Ziebart et al., 2008a) trained individually for each player.

In Figure 2, we see that MaxEnt ICE predicts behavior with higher accuracy than all other algorithms when the number of observations is limited. In particular, it achieves close to its best performance with as few at 16 observations. The maximum likelihood estimator eventually overtakes it, as expected since it will ultimately converge to σ, but only after 10,000 observations, or about as many observations as there are outcomes in the game. This experiment demonstrates that learning underlying utility functions to estimate observed behavior can be much more data-efficient for small sample sizes, and additionally, that the regret-based assumptions of MaxEnt ICE are both reasonable and beneficial in our strategic routing game setting.

Table 1. Transfer error (log-loss) on unobserved games.

Problem	Logistic Model	MaxEnt Ice
Add Highway	4.177	3.093
Add Driver	4.060	3.477
Gas Shortage	3.498	3.137
Congestion	3.345	2.965

Next, we evaluate behavior transfer from this routing game to four similar games, the results of which are displayed in Table 1. The first game, *Add Highway*, adds the dashed route to the game. That is, we model the city building a new highway. The second game, *Add Driver*, adds another driver to the game. The third game, *Gas Shortage*, keeps the structure of the game the same, but changes the reward function to make gas mileage more important to the drivers. The final game, *Congestion*, adds construction to the major roadway, delaying the drivers.

These transfer experiments even more directly demonstrate the benefits of learning utility weights rather than directly learning the joint-action distribution; direct strategy-learning approaches are incapable of being applied to general transfer setting. Thus, we only compare against the Logistic Model. We see from Table 1 that MaxEnt ICE outperforms the Logistic Model in all of our tests. For reference, in these new games, the uniform strategy has a loss of approximately 6.8 in all games, and the true behavior has a loss of approximately 2.7.

7. Conclusion

In this paper, we extended inverse optimal control to multi-agent settings by combining the principle of maximum entropy with the game-theoretic notion of regret. We observed that our formulation has a particularly appealing dual program, which led to a simple gradient-based optimization procedure. Perhaps the most appealing quality of our technique is its theoretical and practical sample complexity. In our experiments, MaxEnt ICE performed exceptionally well after only 0.1% of the game had been observed.

Acknowledgments

This work is supported by the ONR MURI grant N00014-09-1-1052 and by the National Sciences and Engineering Research Council of Canada (NSERC).

References

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

Berry, S., Levinsohn, J., and Pakes, A. Automobile prices in market equilibrium. *Econometrica*, 63(4): 841–90, July 1995.

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

Della Pietra, S., Della Pietra, V., and Lafferty, J. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19 (4):380–393, 2002. ISSN 0162-8828.

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.

Henry, P., Vollmer, C., Ferris, B., and Fox, D. Learning to navigate through crowded environments. In *Proceedings of Robotics and Automation*, 2010.

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

Kakade, S., Kearns, M., Langford, J., and Ortiz, L. Correlated equilibria in graphical games. In *Proceedings of Electronic Commerce*, pp. 42–47, 2003.

Kivinen, J. and Warmuth, M. K. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132:1–63, 1995.

Nevo, A. Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342, March 2001.

Ng, A. and Russell, S. Algorithms for inverse reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2000.

Ortiz, L. E., Shapire, R. E., and Kakade, S. M. Maximum entropy correlated equilibrium. In *Proceedings of Artificial Intelligence and Statistics*, pp. 347–354, 2007.

Osborne, M.J. and Rubinstein, A. *A course in game theory*. The MIT press, 1994. ISBN 0262650401.

Ratliff, N., Bagnell, J. A., and Zinkevich, M. Maximum margin planning. In *Proceedings of the International Conference on Machine Learning*, 2006.

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

Silver, D., Bagnell, J. A., and Stentz, A. Learning from demonstration for autonomous navigation in complex unstructured terrain. *International Journal of Robotics Research*, 29(1):1565 – 1592, October 2010.

Waugh, K., Ziebart, B., and Bagnell, J. A. Computational rationalization: The inverse equilibrium problem. *arXiv*, abs/1103.5254, 2011.

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., Maas, A., Bagnell, J. A., and Dey, A. Maximum entropy inverse reinforcement learning. In *Proceeding of the AAAI Conference on Artificial Intelligence*, 2008a.

Ziebart, B. D., Maas, A., Dey, A. K., and Bagnell, J. A. Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior. In *Proceedings of the International Conference on Ubiquitous Computing*, 2008b.

Ziebart, B. D., Bagnell, J. A., and Dey, A. K. Modeling interaction via the principle of maximum causal entropy. In *Proceedings of the International Conference on Machine Learning*, 2010.

Appendix

Rationality Properties and Primal Programs

The proof of Theorem 1 relies upon the following technical lemmas.

Lemma 5.

$$b^{\mathrm{T}}w \leq \max_{a_i \in A} {a_i}^{\mathrm{T}}w \Leftrightarrow \exists \lambda \in \Delta_A \ s.t.\ b^{\mathrm{T}}w \leq \lambda^{\mathrm{T}}Aw.$$

Proof of Lemma 5. Given $b^{\mathrm{T}}w \leq \max_{a_i \in A} {a_i}^{\mathrm{T}}w$, choose

$$\lambda_i = \begin{cases} 1 & \text{if } a_i = \operatorname{argmax}_{a_i \in A} {a_i}^{\mathrm{T}}w \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

Thus, $b^{\mathrm{T}}w \leq \max_{a_i \in A} {a_i}^{\mathrm{T}}w = \lambda^{\mathrm{T}}Aw$.
Given $\exists \lambda \in \Delta_A$ s.t. $b^{\mathrm{T}}w \leq \lambda^{\mathrm{T}}Aw$,

$$\begin{aligned} & b^{\mathrm{T}}w \leq \lambda^{\mathrm{T}}Aw && (20)\\ \leq & \sum_{a_j \in A} \lambda_{a_j} \max_{a_i \in A} {a_i}^{\mathrm{T}}w && (21)\\ = & \left[\max_{a_i \in A} {a_i}^{\mathrm{T}}w\right] \sum_{a_j \in A} \lambda_{a_j} && (22)\\ = & \max_{a_i \in A} {a_i}^{\mathrm{T}}w && (23) \end{aligned}$$

□

Lemma 6.

$$\forall w \in \mathbb{R}^K, b^{\mathrm{T}}w \leq \max_{i \in N} {a_i}^{\mathrm{T}}w \Leftrightarrow \exists \lambda \in \Delta_A \ s.t.\ b = \lambda^{\mathrm{T}}A.$$

Proof of Lemma 6.

$$\begin{aligned} & \forall w \in \mathbb{R}^K, b^{\mathrm{T}}w \leq \max_{a_i \in A} {a_i}^{\mathrm{T}}w && (24)\\ \Leftrightarrow & \forall w \in \mathbb{R}^K, \exists \lambda \in \Delta_A \text{ s.t. } b^{\mathrm{T}}w \leq \lambda^{\mathrm{T}}Aw && (25)\\ \Leftrightarrow & \forall w \in \mathbb{R}^K, \exists \lambda \in \Delta_A \text{ s.t. } \left[b - \lambda^{\mathrm{T}}A\right]^{\mathrm{T}}w \leq 0 && (26) \end{aligned}$$

$\Leftrightarrow$ the following linear program has optimal value 0

$$\begin{aligned} & \max_{w,t} b^{\mathrm{T}}w - t && (27)\\ & \text{subject to: } t \geq {a_i}^{\mathrm{T}}w, \forall a_i \in A. \end{aligned}$$

The following linear feasibility problem is the dual of the above program

$$\begin{aligned} & \min_{\lambda} 0 && (28)\\ & \text{subject to: } b = \lambda^{\mathrm{T}}A \\ & \qquad \lambda \in \Delta_A. \end{aligned}$$

By strong duality for linear programming, the primal has value 0 iff the dual is feasible, which is exactly when $\exists \lambda \in \Delta_A$ s.t. $b = \lambda^{\mathrm{T}}A$. □

Proof of Theorem 1.

$$\forall w \in \mathbb{R}^K,\ R^{\Phi}(\hat{\sigma}, w) \leq R^{\Phi}(\tilde{\sigma}, w) \tag{29}$$

$$\Leftrightarrow \forall w \in \mathbb{R}^K, \max_{f_i \in \Phi} \hat{\sigma}^{\mathrm{T}} R_i^{f_i} w \leq \max_{f_i \in \Phi} \tilde{\sigma}^{\mathrm{T}} R_i^{f_i} w \tag{30}$$

$$\Leftrightarrow \forall f_i \in \Phi, \forall w \in \mathbb{R}^K, \hat{\sigma}^{\mathrm{T}} R_i^{f_i} w \leq \max_{f_j \in \Phi} \tilde{\sigma}^{\mathrm{T}} R_j^{f_j} w \tag{31}$$

$$\Leftrightarrow \forall f_i \in \Phi, \exists \eta^{f_i} \in \Delta_{\Phi} \text{ s.t. } \hat{\sigma}^{\mathrm{T}} R_i^{f_i} = \sum_{f_j \in \Phi} \eta_{f_j}^{f_i} \tilde{\sigma}^{\mathrm{T}} R_j^{f_j} \tag{32}$$

The last step makes use of our second technical lemma. □

Derivation of the Dual Program

The Lagrange dual is

$$\min_{\alpha,\beta,\gamma,\delta,u,v,\xi} \max_{\hat{\sigma},\eta,\nu} -\sum_{a \in \bar{\mathcal{A}}} \hat{\sigma}_a \log \hat{\sigma}_a - C\nu - \sum_{f_i \in \bar{\Phi}} \left(\hat{\sigma} \bar{R}_i^{f_i} - \sum_{f_j \in \Phi} \eta_{f_j}^{f_i} \tilde{\sigma}^{\mathrm{T}} R_j^{f_j} - \nu \right) \alpha^{f_i} \tag{33}$$

$$- \sum_{f_i \in \bar{\Phi}} \left(\sum_{f_j \in \Phi} \eta_{f_j}^{f_i} \tilde{\sigma}^{\mathrm{T}} R_j^{f_j} - \hat{\sigma} \bar{R}_i^{f_i} - \nu \right) \beta^{f_i} \tag{34}$$

$$+ \sum_{f_i \in \bar{\Phi}} \left(1 - \sum_{f_j \in \Phi} \eta_{f_j}^{f_i} \right) \gamma^{f_i} + \left(1 - \sum_{a \in \bar{\mathcal{A}}} \hat{\sigma}_a \right) \delta \tag{35}$$

$$+ \sum_{f_i \in \bar{\Phi}} \sum_{f_j \in \Phi} \eta_{f_j}^{f_i} u_{f_j}^{f_i} + \sum_{a \in \mathcal{A}} \hat{\sigma}_a v_a + \nu\xi \tag{36}$$

$$\text{subject to: } \alpha, \beta, u, v, \xi \geq 0 \tag{37}$$

To solve the unconstrained inner optimization, we take derivatives w.r.t. σ, η and ν and set equal to 0:

$$\log \hat{\sigma}_a = -1 - \sum_{f_i \in \bar{\Phi}} \bar{r}_{i,a}^{f_i} (\alpha^{f_i} - \beta^{f_i}) - \delta + v_a = 0, \tag{38}$$

$$\tilde{\sigma}^{\mathrm{T}} R_j^{f_j} (\alpha^{f_i} - \beta^{f_i}) - \gamma^{f_i} + u_{f_j}^{f_i} = 0, \quad \forall f_i, \in \bar{\Phi}, f_j \in \Phi, \text{ and} \tag{39}$$

$$\xi - C + \sum_{f_i \in \bar{\Phi}} \alpha^{f_i} + \beta^{f_i} = 0. \tag{40}$$

Substituting into the Lagrangian, we get

$$\min_{\alpha,\beta,\gamma,\delta,u,v,\xi} \sum_{f_i \in \bar{\Phi}} \gamma^{f_i} + \delta + \exp(-1-\delta) \sum_{a \in \bar{\mathcal{A}}} \exp\left(-\sum_{f_i \in \bar{\Phi}} \bar{r}_{i,a}^{f_i} (\alpha^{f_i} - \beta^{f_i}) + v_a \right) \tag{41}$$

$$\text{subject to: } \tilde{\sigma}^{\mathrm{T}} R_j^{f_j} (\alpha^{f_i} - \beta^{f_i}) - \gamma^{f_i} + u_{f_j}^{f_i} = 0, \quad \forall f_i \in \bar{\Phi}, f_j \in \Phi \tag{42}$$

$$\xi + \sum_{f_i \in \bar{\Phi}} \alpha^{f_i} + \beta^{f_i} = C, \tag{43}$$

$$\alpha, \beta, u, v, \xi \geq 0. \tag{44}$$

We note that u are slack variables, and that, by inspection, $v = 0$ at optimality. Thus, an equivalent program is

$$\min_{\alpha,\beta,\gamma,\delta,\xi} \sum_{f_i \in \bar{\Phi}} \gamma^{f_i} + \delta + \exp(-1-\delta) \sum_{a \in \bar{\mathcal{A}}} \exp\left(- \sum_{f_i \in \bar{\Phi}} \bar{r}_{i,a}^{f_i} (\alpha^{f_i} - \beta^{f_i}) \right) \tag{45}$$

$$\text{subject to: } \tilde{\sigma}^{\mathrm{T}} R_j^{f_j} \lambda^{f_i} \leq \gamma^{f_i}, \quad \forall f_i \in \bar{\Phi}, f_j \in \Phi \tag{46}$$

$$\xi + \sum_{f_i \in \bar{\Phi}} \alpha^{f_i} + \beta^{f_i} \leq C, \tag{47}$$

$$\alpha, \beta, \xi \geq 0. \tag{48}$$

We eliminate δ by setting its partial derivative to 0, solving for δ

$$\delta = \log \left(\sum_{a \in \bar{\mathcal{A}}} \exp \left(- \sum_{f_i \in \bar{\Phi}} \bar{r}_{i,a}^{f_i} (\alpha^{f_i} - \beta^{f_i}) \right) \right) - 1 \tag{49}$$

and substituting back into the objective

$$\min_{\alpha,\beta,\gamma,\xi} \sum_{f_i \in \bar{\Phi}} \gamma^{f_i} + \log \left(\sum_{a \in \bar{\mathcal{A}}} \exp \left(- \sum_{f_i \in \bar{\Phi}} \bar{r}_{i,a}^{f_i} (\alpha^{f_i} - \beta^{f_i}) \right) \right) - 1 \tag{50}$$

$$\text{subject to: } \tilde{\sigma}^{\mathrm{T}} R_j^{f_j} (\alpha^{f_i} - \beta^{f_i}) \leq \gamma^{f_i}, \quad \forall f_i \in \bar{\Phi}, f_j \in \Phi \tag{51}$$

$$\xi + \sum_{f_i \in \bar{\Phi}} \alpha^{f_i} + \beta^{f_i} \leq C, \tag{52}$$

$$\alpha, \beta, \xi \geq 0. \tag{53}$$

By inspection, at optimality, $\gamma^{f_i} = \max_{f_j \in \Phi} \tilde{\sigma}^{\mathrm{T}} R_j^{f_j} (\alpha^{f_i} - \beta^{f_i})$. Thus an equivalent program is

$$\min_{\alpha,\beta,\xi} \sum_{f_i \in \bar{\Phi}} \left[\max_{f_j \in \Phi} \tilde{\sigma}^{\mathrm{T}} R_j^{f_j} (\alpha^{f_i} - \beta^{f_i}) \right] + \log \left(\sum_{a \in \bar{\mathcal{A}}} \exp \left(- \sum_{f_i \in \bar{\Phi}} \bar{r}_{i,a}^{f_i} (\alpha^{f_i} - \beta^{f_i}) \right) \right) - 1 \tag{54}$$

$$\xi + \sum_{f_i \in \bar{\Phi}} \alpha^{f_i} + \beta^{f_i} = C, \tag{55}$$

$$\alpha, \beta, \xi \geq 0. \tag{56}$$

Proof of Lemma 4. In the derivation of the dual program, we observed that at optimality

$$\log \hat{\sigma}_a = -1 - \sum_{f_i \in \bar{\Phi}} \bar{r}_{i,a}^{f_i} (\alpha^{f_i} - \beta^{f_i}) - \delta + v_a = 0. \tag{57}$$

Noting $v = 0$ and substituting for the optimal δ, we get

$$\log \hat{\sigma}_a = - \sum_{f_i \in \bar{\Phi}} \bar{r}_{i,a}^{f_i} (\alpha^{f_i} - \beta^{f_i}) - \log \left(\sum_{a' \in \bar{\mathcal{A}}} \exp \left(- \sum_{f_j \in \bar{\Phi}} \bar{r}_{j,a'}^{f_j} (\alpha^{f_j} - \beta^{f_j}) \right) \right). \tag{58}$$

All that remains is to exponentiate both sides. □

Sample Complexity

Proof of Theorem 2.

$$P\left(\max_{f_i\in\Phi,k\in K}|\tilde{\sigma}R_i^{f_i}-\sigma R_i^{f_i}|_k\geq\epsilon\Delta M\right)\leq P\left(\bigcup_{f_i\in\Phi,k\in K}|\tilde{\sigma}R_i^{f_i}-\sigma R_i^{f_i}|_k\geq\epsilon\Delta M\right) \tag{59}$$

$$\leq\sum_{f_i\in\Phi,k\in K}P\left(|\tilde{\sigma}R_i^{f_i}-\sigma R_i^{f_i}|_k\geq\epsilon\Delta M\right) \tag{60}$$

$$\leq\sum_{f_i\in\Phi,k\in K}2\exp\left(\frac{-\epsilon^2M}{2}\right) \tag{61}$$

$$=2|\Phi|K\exp\left(\frac{-\epsilon^2M}{2}\right) \tag{62}$$

$$\leq\delta \tag{63}$$

We use the union bound in step 2, and Hoeffding's inequality in step 3. Solving for M, we get our result

$$M\geq\frac{2}{\epsilon^2}\log\frac{2|\Phi|K}{\delta}. \tag{64}$$

□

Proof of Corollary 1. We have $\forall w, R^\Phi(\hat{\sigma},w)\leq R^\Phi(\tilde{\sigma},w)+\nu\,||w||_1$, where ν depends on the choice of the slack's penalty. Thus, we have $R^\Phi(\hat{\sigma},w^*)\leq R^\Phi(\tilde{\sigma},w^*)+\nu\,||w||_1\leq R^\Phi(\sigma,w^*)+(\epsilon\Delta+\nu)\,||w^*||_1$ with probability at least $1-\delta$, so long as M is as large as Theorem 2 deems. We can make ν as small as we like by increasing the slack penalty. □