Scalable Opponent Strategy Estimation in Imperfect-Information Games

Tri-An D. Nguyen
Rutgers University
New Brunswick, New Jersey, USA

Garrett Seo
University of Michigan
Ann Arbor, Michigan, USA

Xintong Wang
Rutgers University
New Brunswick, New Jersey, USA

Abstract

Repeated economic interactions on the Web let agents learn how others behave, but private information makes opponent strategies hard to infer. We study how to estimate a fixed opponent's strategy from partial observations of repeated interactions. Existing library-based methods concentrate on one candidate strategy and work well only if it closely matches the opponent, while library-free methods with convergence guarantees incur costs exponential in the game horizon. Strategy Library Mixture Estimation (SLME) instead treats library strategies, such as negotiation tactics or styles of play, as building blocks that the opponent may switch between across encounters, and fits their maximum-likelihood mixture. Observation-Induced Strategy Reconstruction (OISR) computes a maximum-a-posteriori (MAP) estimate over the full strategy space from a compact, observation-induced representation, without constructing the full game tree. We prove that both methods converge to the global optimum of their objectives, OISR under suitably constructed priors. Under sufficient exploration, OISR recovers the opponent up to observational equivalence, and SLME does so whenever its library contains an observationally equivalent mixture; in these scenarios, best responses to the recovered models have vanishing regret. In Leduc poker and Deal or No Deal, with as many observed plays as the opponent has information sets, SLME reaches normalized regret below 0.01 against opponents that can be represented by its library, lower than library-free methods reach with 100 times more plays. With the same amount of observed data, OISR enumerates on average only 20% of the total number of opponent sequences. Against randomly sampled opponents, where no suitable library is known, OISR achieves average normalized regret 0.49 in online play in Leduc, versus 0.54 for the next best baseline and 1.6–3.9 for the other methods.

CCS Concepts

• **Computing methodologies** → **Multi-agent systems**; • **Theory of computation** → *Multi-agent learning*.

Keywords

opponent modeling, imperfect-information games, automated negotiation

1 Introduction

The growing use of algorithms acting on behalf of users, such as automated negotiators and bidding agents, enables repeated economic interactions on the Web at scale and creates opportunities to learn from accumulated experience. In negotiation, for example, an opponent's previous offers and rejections can inform which offers to make, how far to concede, and when to accept. Opponent modeling uses such past interactions to predict others' responses and inform future decisions [2]. When the opponent's strategy remains fixed across encounters, these observations support learning its behavior even as private valuations or cards change.

Learning from these observations faces two challenges. Hidden information obscures decision contexts: an offer may be rejected because of the opponent's private valuation or because of its negotiation strategy. At the same time, the number of information sets can grow exponentially with the interaction horizon. Existing library-based methods treat each candidate in a finite strategy library as a hypothesis about the opponent and concentrate on a single candidate, even when a mixture of candidates represents the opponent exactly [27, 13, 11], so they work well only if one candidate closely matches the opponent. Library-free methods with convergence guarantees, such as full-game MAP estimation (FMAP) [11], instead represent the opponent's entire strategy, whose size can grow exponentially with the interaction horizon.

We address these limitations with two methods for estimating a fixed opponent's strategy from repeated play in finite, two-player, general-sum games with imperfect information and perfect recall. Our first method, Strategy Library Mixture Estimation (SLME), treats library strategies as building blocks rather than competing hypotheses. It models an opponent that may follow a different library strategy, such as an established negotiation tactic or style of play, in each encounter, and estimates the mixture weights that maximize the observed-data likelihood. This preserves the structure supplied by the library while allowing behavior beyond its individual members.

Our second method, Observation-Induced Strategy Reconstruction (OISR), estimates the opponent's full strategy without a library. It solves an exact, compact maximum-a-posteriori (MAP) problem built from the hidden histories compatible with past observations: it retains the opponent's information sets along these histories, with every legal action at each, and fixes the remaining action distributions at the modes of a fixed prior. This reduction preserves the full MAP optimum without constructing the full game tree or representing the opponent's entire strategy during fitting.

We prove that both methods converge to the global optimum of their objectives, OISR under positive coherent priors that we construct, with explicit bounds on the remaining optimization error. As observed plays accumulate, provided the learner explores sufficiently and the per-play optimization error vanishes, the fitted models recover the opponent up to *observational equivalence*, even when the learner updates its strategy between plays: they converge to strategies that no learner can distinguish from the opponent's, so best-responding to them becomes as good as best-responding to the opponent. For SLME, recovery requires a library mixture observationally equivalent to the opponent; OISR, which fits over the full strategy space, needs no such condition.

We evaluate both methods in Leduc poker and two sizes of Deal or No Deal against library members, library mixtures, and randomly sampled opponents, on fixed datasets and in online play, where the learner refits its model and best-responds after each play. We report normalized regret: the utility the learner loses by best-responding to the fitted model instead of the opponent, relative to the loss from best-responding to a uniform opponent model. The two methods suit different opponents. Against opponents that can be represented by a mixture of its library strategies, SLME reaches normalized regret below 0.01 once the number of observed plays equals the opponent's number of information sets, lower than library-free methods achieve even with 100 times as many plays. With the same amount of observed data, OISR explicitly enumerates on average only 20% of the opponent sequences that FMAP constructs. Against randomly sampled opponents, which model settings where no suitable library is known, the library-based methods do no better than a best response to a uniform opponent model, whereas OISR keeps improving, reaching normalized regret between 0.021 and 0.059 with 100 observed plays per opponent information set. In online play in Leduc against these opponents, OISR achieves average normalized regret 0.49, comparable to 0.54 for FMAP and well below 1.6–3.9 for the other methods.

1.1 Related Work

Bayesian and library-based estimation. Bayesian approaches infer a posterior over opponent strategies and compute responses to it [27, 13]. Related models learn opponent-specific predictions from poker observations with unrevealed cards [26], or select actions using posterior beliefs over candidate behavioral types [1]. Finite-library implementations of Bayesian best response (BBR) treat each candidate as a hypothesis about the opponent's fixed strategy across all observed plays [27]. Their posterior weights can concentrate on a single candidate even when a mixture of library strategies represents the opponent exactly [11]. SLME treats the library weights as mixture-model parameters and estimates them by maximizing the observed-data likelihood, directly targeting the best-fitting library mixture while preserving the candidates' behavioral structure.

Direct strategy estimation. Southey et al. derive maximum-a-posteriori (MAP) estimates of opponent strategy parameters in Kuhn poker [28]. Mealing and Shapiro combine expectation-maximization (EM) inference of hidden information with sequence prediction and simulated play to model changing opponents in simplified poker [25]. Ganzfried's full MAP method (FMAP) maximizes a concave log posterior over the full opponent sequence-form representation using projected gradient steps [11]. OISR instead solves the full MAP problem through an observation-induced compact representation.

Equilibrium computation and opponent exploitation. Methods for approximating Nash equilibria in two-player zero-sum games include counterfactual regret minimization and its sampled and neural variants [32, 22, 5], as well as neural fictitious self-play [15]. Opponent exploitation adapts play using explicit opponent models [12] or estimated utilities of precomputed responses [4]. Robust response methods balance exploitation against exploitability [20, 19], while Opponent Exploitation Search (OX-Search) controls exploitability relative to a baseline [14]. Our work addresses the complementary task of estimating opponent strategies from partial observations for subsequent response computation.

Automated negotiation. Opponent models support offer generation, concession, and acceptance decisions [18, 3]. Bayesian approaches infer private preferences [16, 6], while Gaussian process models predict the utility of future opponent offers [30]. We estimate behavioral strategies without requiring knowledge of opponent utilities and evaluate our methods in variants of negotiation games based on Deal or No Deal [24].

2 Problem Formulation

Game formulation. We consider finite, two-player, general-sum extensive-form games with imperfect information and perfect recall. Let i denote the learner and $-i$ the opponent. Private information, such as a player's hand in poker or valuation in negotiation, is represented by chance moves whose outcomes are privately observed.

A history h is a sequence of player actions and chance outcomes from the start of a play. Let H denote the set of histories and $Z \subseteq H$ the set of terminal histories. At each nonterminal history h, $P(h) \in \{i, -i, \text{chance}\}$ identifies who acts next, and $A(h)$ is the set of available actions. At terminal history z, the learner receives utility $u_i(z) \in \mathbb{R}$. The learner knows the game rules (i.e., information structure and chance probabilities) and its own utility function; knowledge of the opponent's utility is not required. We write $p_{\text{chance}}(h)$ for the product of the probabilities of all chance outcomes along history h, with value one if there are no chance moves.

Let o_i be a function mapping each history h to the information available to player i. Under perfect recall, $o_i(h)$ contains the learner's complete record of observations and its own actions along h; at a terminal history z, it also includes the learner's utility $u_i(z)$. Let $\mathcal{I}_i$ denote the collection of player i's information sets, which partitions the histories at which player i acts. Two histories h and h' belong to the same information set I exactly when $o_i(h) = o_i(h')$. All histories in I have the same set of legal actions, denoted $A(I)$.

Strategies and realization weights. A *behavioral strategy* σ_i specifies an action distribution at each information set $I \in \mathcal{I}_i$ [21]. We write $\sigma_i(a \mid I)$ for the probability of choosing action $a \in A(I)$. Let Σ_i denote the set of behavioral strategies for the learner. For each history h, let $s_i(h)$ record player i's decisions along history h as an ordered sequence of pairs (I, a), where I is the information set and a is the action chosen there. Let S_i be the set of all such sequences, including the empty sequence $\emptyset$, used when player i has not yet acted. Under perfect recall, all histories in an information set I share the same sequence of the player's previous information sets and chosen actions. We call this the preceding sequence of I and denote it by $\rho(I)$. For any sequence $s \in S_i$, define its realization weight under σ_i as

$$y_{i,s}(\sigma_i) := \prod_{(I,a)\in s} \sigma_i(a \mid I), \qquad y_{i,\emptyset}(\sigma_i) = 1,$$

i.e., the product of player i's own action probabilities along s. We use $y_i(\sigma_i) = (y_{i,s}(\sigma_i))_{s\in S_i}$ to denote the vector of these weights. For $I \in \mathcal{I}_i$ and $a \in A(I)$, write Ia for the sequence $\rho(I)$ followed by (I, a). Since the action probabilities at I sum to one, realization

weights satisfy the *flow constraints*

$$\sum_{a\in A(I)} y_{i,Ia}(\sigma_i) = y_{i,\rho(I)}(\sigma_i) \qquad \text{for every } I \in \mathcal{I}_i.$$

Conversely, every nonnegative vector y with $y_\emptyset = 1$ that satisfies these constraints is the weight vector of any strategy that plays a at I with probability $y_{Ia}/y_{\rho(I)}$ wherever $y_{\rho(I)} > 0$. The set Y_i of realization-weight vectors is therefore a polytope; this representation is known as the *sequence form* [29]. We define Σ_{-i}, $s_{-i}(h)$, $\mathcal{I}_{-i}$, S_{-i}, $y_{-i}(\sigma_{-i})$, and Y_{-i} analogously for the opponent.

Observations and estimation objective. Our goal is to estimate the unknown opponent strategy, represented as a behavioral strategy σ^*_{-i},[1] from n observed plays of a fixed game. Following prior work on opponent modeling [27, 13, 11], we assume that the opponent's strategy remains fixed across plays. In play t, let $z^t \in Z$ be the resulting terminal history, and σ_i^t the learner's behavioral strategy, which may depend on observations from earlier plays. The learner observes $o_i(z^t)$, which may be compatible with several terminal histories:

$$Z^t := \{z \in Z : o_i(z) = o_i(z^t)\}.$$

As the observation record includes the learner's past observations and actions, all histories $z \in Z^t$ have the same learner sequence $s_i(z) = s_i(z^t)$. For a candidate opponent strategy σ_{-i}, the probability of observing $o_i(z^t)$, conditional on the learner strategy in play t, is therefore

$$\begin{aligned} q^t(\sigma_{-i}) &:= \sum_{z\in Z^t} y_{i,s_i(z)}(\sigma_i^t)\, p_{\text{chance}}(z)\, y_{-i,s_{-i}(z)}(\sigma_{-i}) \\ &= y_{i,s_i(z^t)}(\sigma_i^t)\, f_{o_i(z^t)}(\sigma_{-i}), \end{aligned}$$

where, for each $o \in U := o_i(Z)$, the *opponent factor* $f_o(\sigma_{-i}) := \sum_{z\in Z:\, o_i(z)=o} p_{\text{chance}}(z)\, y_{-i,s_{-i}(z)}(\sigma_{-i})$ is the probability of o with the learner's realization weight divided out. We call two opponent strategies *observationally equivalent* if they have the same f_o for every $o \in U$. By the factorization, observationally equivalent strategies induce the same distribution of learner observations under every learner strategy.

Let $u_i(\sigma_i, \sigma_{-i})$ denote the learner's expected utility when it plays σ_i and the opponent plays σ_{-i}, and $\text{BR}(\sigma_{-i})$ a best response to σ_{-i}, a learner strategy that maximizes $u_i(\cdot, \sigma_{-i})$. The *regret* of a model $\widehat{\sigma}_{-i}$ is the expected utility the learner loses against the true opponent by best-responding to the model instead of to the true opponent:

$$R(\widehat{\sigma}_{-i}) := u_i\big(\text{BR}(\sigma^*_{-i}), \sigma^*_{-i}\big) - u_i\big(\text{BR}(\widehat{\sigma}_{-i}), \sigma^*_{-i}\big). \tag{1}$$

Since the learner's observation includes its utility, observationally equivalent opponents give every learner strategy the same expected utility, so every model observationally equivalent to the true opponent has zero regret. Thus, as the number of plays grows, we seek a model whose values of f_o approach those of the true opponent for every $o \in U$.

Our models use the likelihood of the recorded observations. To compute these likelihoods efficiently, we group plays whose likelihoods are identical. For each play t and compatible terminal history $z \in Z^t$, define the known coefficient

$$c_z^t := y_{i,s_i(z)}(\sigma_i^t)\, p_{\text{chance}}(z),$$

which does not depend on the candidate opponent strategy, so that $q^t(\sigma_{-i}) = \sum_{z\in Z^t} c_z^t\, y_{-i,s_{-i}(z)}(\sigma_{-i})$. We group plays into the same *observation type* when they yield the same learner observation and have the same collection of known coefficients. These plays have identical likelihoods under every candidate opponent strategy. Let V denote the recorded observation types and n_j the number of plays of type $v_j \in V$. Plays of type v_j have the same compatible histories Z_j and known coefficients $c_{j,z}$, so their common likelihood under candidate σ_{-i} is $q_j(\sigma_{-i}) = \sum_{z\in Z_j} c_{j,z}\, y_{-i,s_{-i}(z)}(\sigma_{-i})$. The observed-data log likelihood is

$$\ell_n(\sigma_{-i}) := \sum_{t=1}^{n} \log q^t(\sigma_{-i}) = \sum_{j=1}^{|V|} n_j \log q_j(\sigma_{-i}). \tag{2}$$

3 Strategy Library Mixture Estimation

We first consider estimation based on a fixed library of candidate opponent strategies. For example, such a library can incorporate domain knowledge through representative behavioral patterns, encoding concession patterns in negotiation [10] or bluffing frequencies in poker [27]. `SLME` directly estimates mixture weights by maximizing the observed-data likelihood, allowing multiple library members to jointly explain the observations.

Mixture model. Let $\{\sigma^m_{-i}\}_{m=1}^M$ be the library of opponent behavioral strategies. We model the opponent as independently selecting a library member at the beginning of each play and following it throughout that play. The selection probabilities $\pi \in \Delta_M$ remain fixed across plays, where Δ_M is the probability simplex over the M members and π_m is the probability of selecting library member m.

For observation type v_j, let $q_j^m := q_j(\sigma^m_{-i})$ denote its likelihood under member m, as defined in Section 2. Its likelihood under the mixture is

$$q_j(\pi) = \sum_{m=1}^{M} \pi_m q_j^m.$$

We estimate the mixture weights by maximizing the observed-data log likelihood:

$$\widehat{\pi} \in \arg\max_{\pi\in\Delta_M} \ell_n(\pi), \qquad \ell_n(\pi) = \sum_{j=1}^{|V|} n_j \log q_j(\pi). \tag{3}$$

The objective is concave in π because each $q_j(\pi)$ is linear in π.

Fitting and convergence. We fit the mixture weights using the EM algorithm, treating the selected library member in each play as a latent variable [7]. At iteration k, the E-step computes the responsibility

$$\gamma_j^{m,(k)} = \frac{\pi_m^{(k)} q_j^m}{\sum_{r=1}^M \pi_r^{(k)} q_j^r},$$

the conditional probability that member m generated an observation of type v_j. The M-step averages these responsibilities over the recorded plays:

$$\pi_m^{(k+1)} = \frac{1}{n} \sum_{j=1}^{|V|} n_j \gamma_j^{m,(k)}.$$

[1] By Kuhn's theorem [21], under perfect recall, every mixed strategy has a behavioral strategy that induces the same distribution over terminal histories against every opponent strategy. Thus, representing the opponent's strategy in behavioral form entails no loss of generality in terms of game outcomes.

Each pass uses the entire accumulated dataset. We have the following guarantee as a direct application of Zhao's convergence analysis [31].

Lemma 1 (SLME convergence). *Suppose $n \geq 1$ and every observation type has positive likelihood under at least one library member. Starting from $\pi_m^{(0)} = 1/M$ for all m, the exact EM iterates converge to a maximum-likelihood solution $\widehat{\pi}$ and satisfy*

$$0 \leq \frac{\ell_n(\widehat{\pi}) - \ell_n(\pi^{(k)})}{n} \leq \frac{\log M}{k+1}. \tag{4}$$

Proof. Let $f(\pi) := \ell_n(\pi)/n$. The EM update satisfies $\pi_m^{(k+1)} = \pi_m^{(k)} \nabla_m f(\pi^{(k)})$, so it coincides with the multiplicative gradient method. The objective has Zhao's required form: the coefficients n_j/n are positive and sum to one, and each likelihood vector $(q_j^1, \ldots, q_j^M)$ is nonnegative and nonzero. His convergence and rate guarantees therefore apply. □

Lemma 1 guarantees convergence to the maximum log likelihood over library mixtures for observations from a fixed set of plays.

Computation. We cache the log likelihood $\log q_j^m$ of each observation type v_j under each library strategy σ_{-i}^m. For each observation type, we enumerate its compatible terminal histories once and sum their probabilities under each library strategy. The likelihood table requires $O(|V|M)$ numerical entries, with $O(|V|)$ additional entries for the counts n_j. After preparation, each EM iteration on a fixed dataset of plays costs $O(|V|M)$ scalar operations and requires no further history enumeration or evaluation of library strategies.

For online use, a new play either increments the count of a recorded type or adds a new type whose likelihoods under each library strategy are computed once; each refit then runs EM on all accumulated types at $O(|V|M)$ per iteration, whereas finite-library BBR updates its posterior in $O(M)$ operations per play [27]. Appendix A analyzes the full time and space complexity of SLME.

4 Observation-Induced Strategy Reconstruction

When a strategy library is unavailable or cannot adequately represent the opponent's behavior, OISR estimates the opponent's behavioral strategy directly, using a compact representation constructed from the recorded observations. It optimizes action distributions at opponent information sets appearing in compatible terminal histories and fixes the remaining distributions at their prior modes. This reduction preserves the full MAP optimum for the chosen prior while avoiding explicit representation of information sets that do not affect the observed-data likelihood.

Objective and prior. We place fixed, independent Dirichlet priors on the opponent's action distributions at its information sets. For each $I \in \mathcal{I}_{-i}$ and $a \in A(I)$, let $\lambda(I,a) \geq 0$ be the pseudocount, so that the Dirichlet shape parameter is $\alpha(I,a) = 1 + \lambda(I,a)$. We call the prior *positive* if every pseudocount is positive. Adding the log-prior terms to the observed-data log likelihood in Eq. (2) gives the MAP objective, up to an additive constant independent of the strategy:

$$J_n(\sigma_{-i}) := \ell_n(\sigma_{-i}) + \sum_{\substack{I \in \mathcal{I}_{-i} \\ a \in A(I)}} \lambda(I,a) \log \sigma_{-i}(a \mid I). \tag{5}$$

Compact representation. To represent the opponent's unknown action distributions, let $\widetilde{\mathcal{I}}_{-i}$ be the set of all opponent information sets appearing in at least one compatible terminal history $z \in Z_j$ for some recorded type $v_j \in V$. We call these information sets *retained* and assume that $\widetilde{\mathcal{I}}_{-i} \neq \emptyset$; otherwise the observed plays contain no opponent decisions to fit.

Assume a positive prior, and write $\lambda(I) := \sum_{a \in A(I)} \lambda(I,a)$. At a non-retained information set, the action distribution does not affect the observed-data likelihood and appears only in its own log-prior term, maximized at the prior mode, $\sigma_{-i}(a \mid I) = \lambda(I,a)/\lambda(I)$. OISR fixes these distributions at their prior modes, making their total log-prior contribution a constant K and leaving only the retained action distributions to optimize:

$$J_n(\sigma_{-i}) = \ell_n(\sigma_{-i}) + \sum_{\substack{I \in \widetilde{\mathcal{I}}_{-i} \\ a \in A(I)}} \lambda(I,a) \log \sigma_{-i}(a \mid I) + K, \tag{6}$$

where $K := \sum_{I \in \mathcal{I}_{-i} \setminus \widetilde{\mathcal{I}}_{-i}} \sum_{a \in A(I)} \lambda(I,a) \log\big(\lambda(I,a)/\lambda(I)\big)$. Optimizing the retained distributions and completing the rest at their prior modes therefore solves the full MAP problem exactly.

Let $\widetilde{S}_{-i} := \{\emptyset\} \cup \{Ia : I \in \widetilde{\mathcal{I}}_{-i},\ a \in A(I)\}$ be the retained sequences, and let $\widetilde{Y}_{-i}$ be the set of restrictions of realization-weight vectors in Y_{-i} to these sequences. For $y_{-i} \in \widetilde{Y}_{-i}$ with positive entries, the retained action probabilities are $\sigma_{-i}(a \mid I) = y_{-i,Ia}/y_{-i,\rho(I)}$. We write $J_n(y_{-i})$ for the objective of this strategy, completed with the prior modes at non-retained information sets.

Fitting and monotonicity. Grouping each type's compatible terminal histories by the opponent's sequence gives the sparse coefficient matrix C:

$$C_{j,s} := \sum_{\substack{z \in Z_j \\ s_{-i}(z) = s}} c_{j,z}, \qquad q_j(y_{-i}) = \sum_{s \in \widetilde{S}_{-i}} C_{j,s} y_{-i,s}. \tag{7}$$

We call s an *endpoint* if $C_{j,s} > 0$ for some j: it records the opponent's complete sequence of decisions along a compatible terminal history.

OISR fits the retained action distributions by EM, treating the opponent's sequence in each recorded play as unobserved. Let $\sigma_{-i}^{(k)}$ be the opponent model after k EM iterations, and let $y_{-i}^{(k)}$ be its realization weights on $\widetilde{S}_{-i}$. For a current model assigning positive probability to every observed type, the E-step computes

$$\gamma_{j,s}^{(k)} = \frac{C_{j,s} y_{-i,s}^{(k)}}{q_j(y_{-i}^{(k)})},$$

the conditional probability that a play of type v_j has opponent sequence s. With n_j recorded plays of type v_j, the expected number of times the opponent chose a at I is

$$N^{(k)}(I,a) = \sum_j n_j \sum_{\substack{s \in \widetilde{S}_{-i} \\ (I,a) \in s}} \gamma_{j,s}^{(k)}.$$

The M-step maximizes the expected complete-data log posterior separately at each retained information set. Writing $N^{(k)}(I) := \sum_{a \in A(I)} N^{(k)}(I,a)$, the resulting update is

$$\sigma_{-i}^{(k+1)}(a \mid I) = \frac{N^{(k)}(I,a) + \lambda(I,a)}{N^{(k)}(I) + \lambda(I)}. \tag{8}$$

Thus, each update normalizes the expected action counts augmented by the prior pseudocounts. Non-retained distributions remain fixed at their prior modes. Appendix B.3.1 shows that

$$J_n(\sigma_{-i}^{(k+1)}) - J_n(\sigma_{-i}^{(k)}) \geq \sum_{\substack{I \in \widetilde{\mathcal{I}}_{-i} \\ a \in A(I)}} \left(N^{(k)}(I,a) + \lambda(I,a)\right) \log \frac{\sigma_{-i}^{(k+1)}(a \mid I)}{\sigma_{-i}^{(k)}(a \mid I)} \geq 0, \tag{9}$$

so each EM pass does not decrease the objective.

Optimization guarantees. The EM updates above apply to any positive Dirichlet pseudocounts. For *global convergence*, we impose a sufficient condition ensuring that the log-prior contribution is concave in realization weights. We call the prior *coherent* if:

$$\lambda(I,a) \geq \sum_{\substack{I' \in \mathcal{I}_{-i} \\ \rho(I') = Ia}} \lambda(I') \qquad \text{for every } I \in \mathcal{I}_{-i},\ a \in A(I). \tag{10}$$

Here I' ranges over all possible next opponent information sets after action a at I. Intuitively, coherence constrains the relative strengths of the local priors: every play reaching one of these next opponent information sets must first choose a at I, so their total pseudocount cannot exceed $\lambda(I,a)$. Appendix B.1 gives the derivation and a construction of a prior with positive, coherent pseudocounts.

Coherence ensures J_n is concave in $\widetilde{Y}_{-i}$ with positive entries (Appendix B.1), so every local maximum is a global maximum. Appendix B.3.2 proves the following convergence guarantee for exact EM:

Theorem 1 (MAP optimization). *Suppose $n \geq 1$, the prior is coherent with positive pseudocounts, and every row of C has a positive entry. Then exact EM, initialized with positive action probabilities at every retained information set and the prior modes elsewhere, satisfies*

$$\lim_{k \to \infty} J_n(\sigma_{-i}^{(k)}) = J_n^*, \quad \text{where } J_n^* := \max_{\sigma_{-i} \in \Sigma_{-i}} J_n(\sigma_{-i})$$

This theorem guarantees convergence to the global MAP objective value for observations from a fixed set of plays. We now construct OISR's initialization, which satisfies the condition of Theorem 1 and yields an explicit convergence rate. Let

$$x \in \arg\max_{y_{-i} \in \widetilde{Y}_{-i}} \min_{\text{endpoint } s} y_{-i,s}, \qquad \Gamma := \frac{1}{\min_{\text{endpoint } s} x_s}.$$

Let y_{-i}^{def} be the prior-mode strategy's realization weights restricted to $\widetilde{S}_{-i}$, and let $\tau := \sum_{I \in \widetilde{\mathcal{I}}_{-i} : \rho(I) = \emptyset} \lambda(I)$ be the total pseudocount at the opponent's first retained decisions. OISR initializes the realization weights on $\widetilde{S}_{-i}$ to

$$y_{-i}^{(0)} := \frac{nx + \tau y_{-i}^{\text{def}}}{n + \tau}. \tag{11}$$

Since $\widetilde{\mathcal{I}}_{-i} \neq \emptyset$, $\tau > 0$, so this mixture is feasible and has positive entries, even if x does not. Appendix B.2 computes x and Γ in linear time over the retained sequences. Appendix B.3.3 proves the following bound.

Lemma 2 (Convergence rate). *Under the assumptions of Theorem 1, exact EM initialized by Eq.* (11) *satisfies, for every $k \geq 1$,*

$$0 \leq J_n^* - J_n(\sigma_{-i}^{(k)}) \leq \frac{B_0}{k},$$

where $B_0 := n \log \Gamma + n \log(1 + \tau/n) + \tau \log(1 + n/\tau)$.

Computation. For each distinct observation type, we enumerate its compatible terminal histories and aggregate their known coefficients $c_{j,z}$ by opponent sequence to construct $C_{j,s}$ (Eq. (7)). Let $E := \{(j,s) : C_{j,s} > 0\}$. As an additional speedup, OISR merges the actions that no compatible history takes at each retained information set into a single escape action carrying their total pseudocount; these actions affect J_n only through their prior terms, so the merge leaves the MAP optimum and EM's guarantees unchanged (Appendix B.4). Using sparse likelihood calculations and backward and forward passes over the sequences $\bar{S} \subseteq \widetilde{S}_{-i}$ along compatible terminal histories, each EM iteration then costs $O(|V| + |E| + |\bar{S}|)$ scalar operations, without further history enumeration (Appendix B.4). Appendix B.4 accounts separately for preparation costs and additional storage.

Online, a new play increments its type's count or adds a new type, whose coefficients may expand the retained representation; previously computed coefficients are reused, and each refit runs EM on all accumulated types.

5 Statistical Guarantees

We study opponent recovery as the number of observed plays n grows, for a fixed opponent strategy σ_{-i}^*. We assume sufficient coverage by the learner strategies generating the observations: there exists $\theta > 0$ such that, almost surely, for sufficiently large n,

$$\frac{1}{n} \sum_{t=1}^{n} y_{i,s}(\sigma_i^t) \geq \theta \qquad \text{for every } s \in S_i. \tag{12}$$

A fixed learner strategy assigning positive probability to every legal action satisfies this condition.

Observational recovery. The following result applies to both estimators. It requires the fitted model's log likelihood to fall below that of the true opponent by at most a sublinear amount.

Theorem 2 (Observational recovery). *Under Eq.* (12)*, let $\widehat{\sigma}_{-i,n}$ be a model fitted from the first n observed plays. If, almost surely,*

$$\ell_n(\widehat{\sigma}_{-i,n}) \geq \ell_n(\sigma_{-i}^*) - o(n), \tag{13}$$

then, almost surely, $f_o(\widehat{\sigma}_{-i,n}) \longrightarrow f_o(\sigma_{-i}^*)$, for every $o \in U$.

Thus, the fitted models asymptotically predict the distribution of learner observations correctly under every learner strategy. Appendix C gives the proof.

Application to SLME. Suppose the library contains an observationally equivalent mixture: there exists $\pi^* \in \Delta_M$ such that

$$\sum_{m=1}^{M} \pi_m^* f_o(\sigma_{-i}^m) = f_o(\sigma_{-i}^*) \qquad \text{for every } o \in U.$$

Then $\ell_n(\pi^*) = \ell_n(\sigma_{-i}^*)$. With uniform initialization and k_n exact EM iterations, Lemma 1 gives

$$\ell_n(\widehat{\pi}_n) \geq \ell_n(\sigma_{-i}^*) - \frac{n \log M}{k_n + 1}.$$

Thus, $k_n \to \infty$ ensures Eq. (13), and the fitted mixtures recover the opponent up to observational equivalence. This conclusion requires an observationally equivalent library mixture; it does not require the true opponent to be a library member.

Application to OISR. For any fixed pseudocounts $\lambda(I, a) \geq 0$, a vanishing MAP objective gap per play,

$$\frac{J_n^* - J_n(\widehat{\sigma}_{-i,n})}{n} \longrightarrow 0 \qquad \text{almost surely,}$$

also implies Eq. (13). To see this, write $y^* := y_{-i}(\sigma_{-i}^*)$ and choose a fixed realization-weight vector y^{pos} with positive entries. For $n \geq 2$, the feasible vector $y_n = (1 - 1/n)y^* + y^{\text{pos}}/n$ has log likelihood at least $\ell_n(\sigma_{-i}^*) - 2$. Its action probabilities are at least y_{Ia}^{pos}/n, so its log-prior contribution is bounded below by $-O(\log n)$. Hence

$$J_n^* \geq \ell_n(\sigma_{-i}^*) - O(\log n).$$

Since $J_n(\sigma) \leq \ell_n(\sigma)$, an $o(n)$ MAP objective gap implies the required likelihood bound.

Statistical recovery requires a vanishing MAP objective gap per play, without requiring coherence. For OISR with a fixed positive coherent prior and the initialization in Eq. (11), Lemma 2 bounds the MAP gap per play by $B_0/(nk_n)$ after k_n exact EM iterations. For a fixed finite game and prior, Γ and τ are uniformly bounded, and therefore so is B_0/n (Appendix B.3.3). Choosing $k_n \to \infty$ ensures observational recovery.

Strategy recovery and best-response regret. Observational recovery also yields the following.

Corollary 1. *Under the assumptions of Theorem 2, almost surely,*

$$R(\widehat{\sigma}_{-i,n}) \longrightarrow 0.$$

If every strategy observationally equivalent to σ_{-i}^* has the same realization weights as σ_{-i}^*, then, almost surely,

$$y_{-i}(\widehat{\sigma}_{-i,n}) \longrightarrow y_{-i}(\sigma_{-i}^*).$$

Proof. Because the learner's terminal observation includes its utility, write $u_i(o)$ for the utility determined by o. Bounding payoff prediction errors uniformly over learner strategies gives

$$R(\widehat{\sigma}_{-i,n}) \leq 2 \sum_{o \in U} |u_i(o)| \left| f_o(\widehat{\sigma}_{-i,n}) - f_o(\sigma_{-i}^*) \right| \longrightarrow 0.$$

For strategy recovery, compactness of Y_{-i} and continuity of each f_o imply that every accumulation point of the fitted realization weights is observationally equivalent to the true opponent. The identifiability assumption makes every such point equal to $y_{-i}(\sigma_{-i}^*)$, proving convergence. □

6 Empirical Evaluation

We compare SLME and OISR with three baselines in Leduc poker and two sizes of Deal or No Deal. We assess sample efficiency and performance against library members, library mixtures, and randomly sampled opponents using fixed observation datasets (Section 6.2), followed by online play in Leduc (Section 6.3). Appendix D reports observation-distribution error and sensitivity to OISR's prior strength. Our source code is available at https://anonymous.4open.science/r/opponent-modeling-811E/. We used AI coding assistants (Claude Code and OpenAI Codex) to help implement and test the experiment code, run the experiments, compute statistics, generate figures, and check the proofs.

6.1 Experimental Setup

Games, baselines, and strategy libraries. We evaluate both learner roles, denoting the player who acts first by P1 and the second by P2. All games are implemented in OpenSpiel [23]. Leduc poker has 468 information sets per player. Players observe their own card, the public card, and all betting actions; the opponent's card is revealed only if neither player folds and the play ends in a showdown. We use two Deal or No Deal variants [24] with up to three alternating proposals. Each player's private valuation is drawn independently and uniformly from three types, each favoring a different item category. Utilities are additive over allocated items, with zero payoff if no agreement is reached. The small game contains one item per category and has 195 and 1,560 information sets for P1 and P2, respectively; the medium game contains two items per category and has 2,190 and 59,130 information sets. Players observe all proposals and responses, their own valuation, and their own payoff, but never the opponent's valuation.

We compare against Bayesian best response (BBR) [27], a finite-library baseline, and two library-free baselines: the EM estimation component of Mealing and Shapiro (MS-EM) [25] and Ganzfried's full MAP method (FMAP) [11]. For FMAP, we use a Dirichlet prior with $\alpha = 2$ and a realization-weight floor of 10^{-6}. Given the high memory and runtime costs of our FMAP implementation, we evaluate it only in Leduc and small Deal or No Deal.[2]

For each game, SLME and BBR share a fixed library of heuristic strategies that give positive probability to all actions. The Leduc library contains 48 strategies that vary in calling, betting with strong hands, bluffing, and responses to earlier betting. The Deal or No Deal library contains 60 strategies from four families: concession, responses to opponent concessions [10], proposals of similarly valued item combinations [9], and fixed utility targets. Within these families, we vary target values, concession rates, and randomness in action selection.

Ground-truth opponent strategies. We consider three classes, defined by their relationship to the strategy library. A *library member* uses one of the library's strategies. A *library mixture* samples uniformly from varying subsets of candidates (up to 32 in Leduc and 24 for Deal or No Deal) before each play and follows the selected strategy throughout that play. These two classes are *representable* by the library. We verified that each library mixture is observationally distinct from every library member by computing the KL divergence between their observation distributions under a uniformly playing learner. We also use *randomly sampled opponents* to evaluate settings where no useful library is known. We generate each opponent by randomly choosing its action probabilities at every information set. Each opponent's strategy remains fixed throughout its interactions with the learner.

We repeat each setting across 10 independent *worlds*, each specifying an opponent from each class and random seeds for generating play. All methods face the same opponents within each world, and the strategy libraries remain fixed across worlds.

[2] Our FMAP implementation uses Gurobi for projections. In nine preliminary runs on medium Deal or No Deal with 592 observations, our implementation did not complete any fits after more than 68 hours per run.

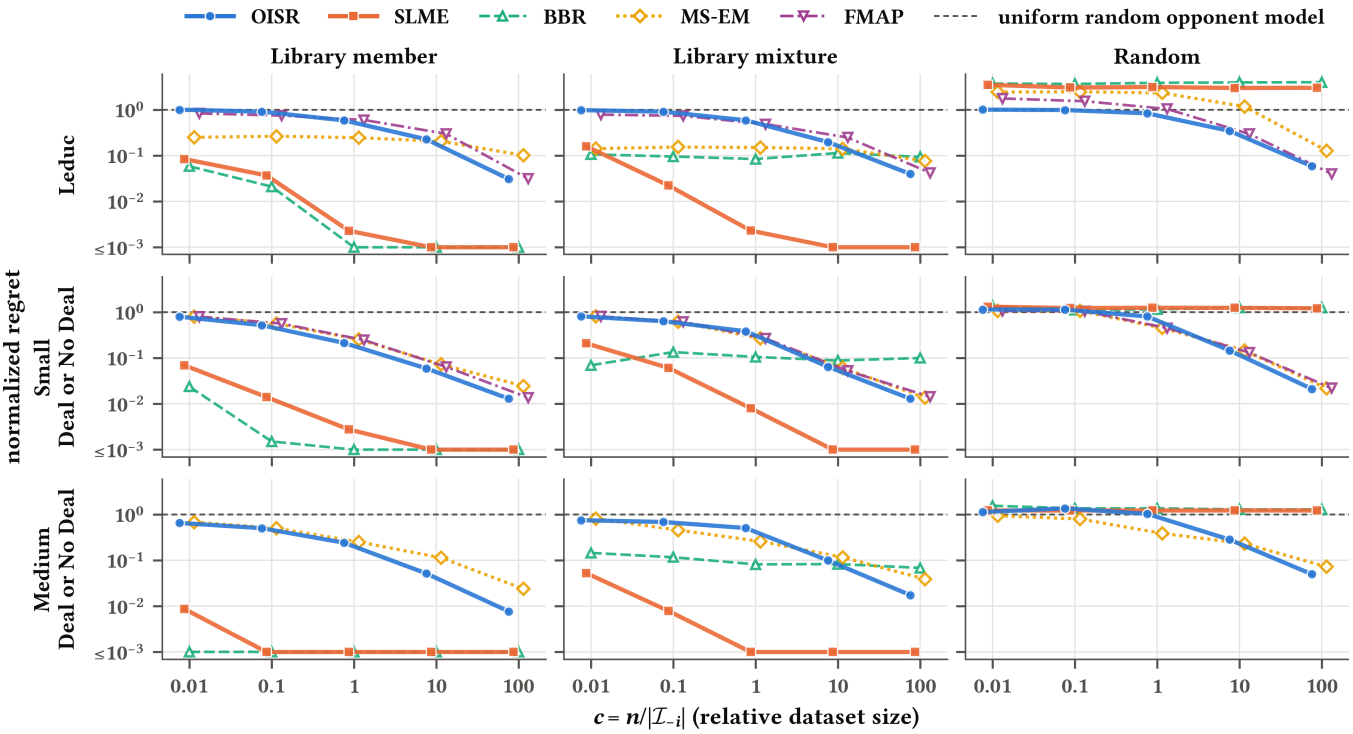

Figure 1: Normalized regret of offline fitted models (Eq. (14)). The pointwise 95% CIs have median width 0.082 normalized-regret units (max 1.71).

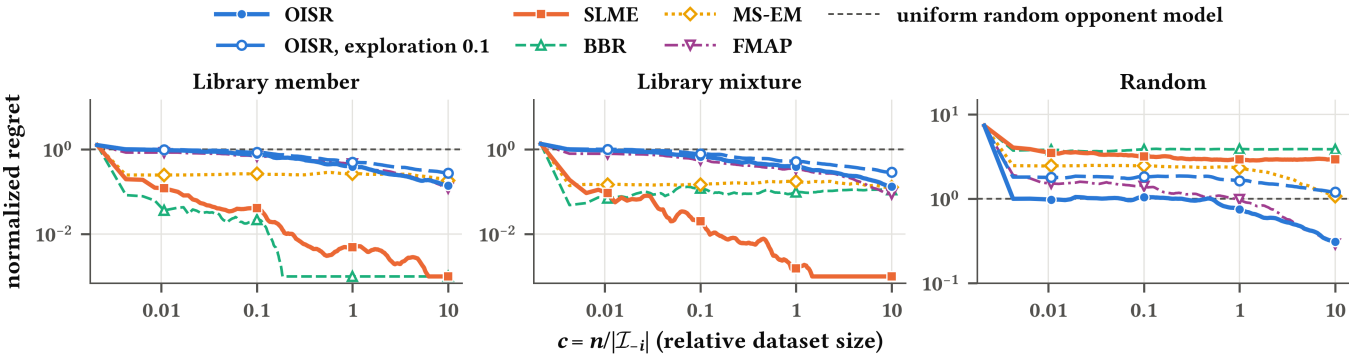

Figure 2: Normalized regret per play during online play in Leduc (Eq. (14)). Curves are smoothed; markers indicate the offline dataset sizes in Figure 1. The pointwise 95% CIs have median width 0.094 normalized-regret units (max 3.09).

Evaluation metrics. We report best-response regret $R(\widehat{\sigma}_{-i,n})$ (Eq. (1)) of a model fitted from n observed plays. We compute best responses using OpenSpiel [23] and evaluate their expected utilities against the ground-truth opponent by full game-tree traversal. Appendix D.1 additionally evaluates prediction accuracy using the KL divergence.

To aid interpretation across payoff scales, we normalize by the regret of a best response to a uniform opponent model $\sigma_{-i}^{\text{unif}}$, which assigns equal probability to every legal action at each information set. For each setting, we divide the average regret across the 10 worlds and both learner roles, indexed by w, by the corresponding average reference regret:

$$\bar{R} = \frac{\sum_w R_w(\widehat{\sigma}_{-i,n})}{\sum_w R_w(\sigma_{-i}^{\text{unif}})}. \tag{14}$$

A value of zero indicates optimal responses, one matches the reference, and values above one indicate worse performance. We aggregate regrets before normalization to avoid unstable ratios in individual cases where a best response to the uniform opponent model already has near-zero regret.

Across all experiments we compute pointwise 95% bias-corrected and accelerated (BCa) bootstrap confidence intervals [8] using

100,000 resamples of the ten independent worlds, keeping both learner roles together.

We use the *relative dataset size* $c = n/|\mathcal{I}_{-i}|$, where n is the number of observed plays and $|\mathcal{I}_{-i}|$ the opponent's number of information sets. Offline datasets target $c \in \{0.01, 0.1, 1, 10, 100\}$, rounding n up to an integer; smaller datasets are prefixes of larger ones.

6.2 Offline Opponent Modeling

We examine how the quality of responses to fitted opponent models changes with the amount of observed play n. For each world and learner role, all methods receive the first n observations from the same sequence of plays. Fig. 1 compares normalized regret (Eq. (14)) across the three opponent classes and the dataset sizes.

OISR uses the coherent prior of Appendix B.1.1 with strength $B_{\text{prior}} = 0.1$. For SLME, OISR, and FMAP, we bound the remaining optimization gap using the maximum feasible improvement predicted by the objective's linearization [17]. Fitting stops when this bound per play falls below $10^{-6}/\sqrt{n}$ or reaches 100,000 optimization iterations; we evaluate the final iterate.[3] BBR computes posterior weights directly, while MS-EM processes each play once.

Library members and mixtures. A suitable library enables effective responses from fewer observations, and SLME retains this advantage when the opponent combines multiple library strategies (Fig. 1, left two columns). Across all three games, SLME achieves normalized regret below 0.07 at $c = 0.1$ and below 0.01 at $c = 1$ against both library members and library mixtures. Where evaluated, OISR and FMAP first achieve normalized regret below 0.07 at $c = 10$ or 100, using roughly 100–1000 times as many observations as SLME at $c = 0.1$; MS-EM remains above this threshold in Leduc at all tested dataset sizes. BBR achieves lower regret with fewer observations against individual library members, but its regret against library mixtures remains high, between 0.07 and 0.15. These results support using SLME when observations are limited and domain knowledge supplies a compact library of canonical strategies whose mixtures capture plausible opponent behavior.

Randomly sampled opponents. Against the randomly sampled opponents, SLME and BBR perform similarly or worse than the uniform-model reference at all dataset sizes. OISR keeps improving, achieving regret between 0.021 and 0.059 across the three games at $c = 100$. Using its compact representation, it achieves regret comparable to FMAP: 0.059 versus 0.040 in Leduc and 0.021 versus 0.022 in small Deal or No Deal. MS-EM performs better on some smaller Deal or No Deal datasets, but OISR has lower regret in all three games on the largest tested datasets.

OISR substantially reduces explicit sequence enumeration, especially when observations are limited. Across the three games, it enumerates on average 0.9%, 20%, and 82% of the opponent sequences represented by FMAP at $c = 0.01$, 1, and 100, respectively. The reduction is largest in medium Deal or No Deal: with P2 as the opponent, these fractions are just 0.13% at $c = 0.01$ and 6.12% at $c = 1$. In a separate timing comparison of 100 Leduc fits using one thread and a 600-second limit per fit, OISR meets the optimization tolerance in every fit, with median times ranging from 0.09 seconds at the smallest dataset size to 217 seconds at the largest. FMAP reaches the time limit in 47 fits.

[3]The optimization tolerance is met in all 900 OISR fits, 734 of 900 SLME fits, and 269 of 600 FMAP fits.

6.3 Online Opponent Modeling: Leduc Poker

We next evaluate opponent modeling during online play, where each method collects observations under its learner's current strategy; this follows prior work settings [11, 27]. The learner plays 4,680 hands of Leduc against a fixed opponent, starting with uniform play. After each hand, the learner updates its opponent model and computes a best response for the next hand. We also evaluate an OISR exploration variant that mixes uniform play into the best response with probability 0.1 at each learner decision. Fig. 2 reports the normalized regret of the learner's strategy used in each hand, pooled over 10 worlds and both learner roles.

For SLME, OISR, and FMAP, each update uses the previous fit for initialization and runs for at most 100 optimization iterations, stopping earlier when the per-play optimization-gap bound falls below $10^{-3}/\sqrt{n}$ after n observed plays. If an update fails, the model from the previous play is used.

The online results in Fig. 2 reinforce the value of a suitable strategy library when observations are limited. Against both library members and library mixtures, SLME achieves normalized regret below 0.01 within 110 observed plays and averages at most 0.003 over the 4,680 plays.

Against randomly sampled opponents, direct strategy estimation is more effective than relying on the fixed library. Using its compact representation, OISR achieves average normalized regret 0.49, comparable to 0.54 for full-game FMAP and substantially below 1.6–3.9 for MS-EM and the library-based methods. Uniform exploration with probability 0.1 satisfies Eq. (12) but raises OISR's average normalized regret against randomly sampled opponents from 0.49 to 1.4, so over the tested horizon it provides no payoff benefit, despite ensuring coverage.

7 Conclusion and Discussion

We develop two complementary methods for opponent modeling from partial observations in imperfect-information games. SLME fits mixtures of library strategies and is the method of choice when the opponent's strategy space decomposes into library components, whereas OISR reconstructs the full strategy from a compact representation when no suitable library is known. Under sufficient exploration, OISR recovers the opponent up to observational equivalence, and SLME does so whenever its library contains an observationally equivalent mixture.

Our analysis assumes a finite, two-player game and a fixed opponent; with several unknown opponents, the likelihood becomes multilinear in their strategies and the objective is no longer concave. Future work includes modeling adaptive opponents, exploration that balances information gathering and payoff, and automatic library construction.

References

[1] Stefano V. Albrecht and Subramanian Ramamoorthy. 2013. A game-theoretic model and best-response learning method for ad hoc coordination in multiagent systems. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1155–1156.

[2] Stefano V. Albrecht and Peter Stone. 2018. Autonomous agents modelling other agents: a comprehensive survey and open problems. *Artificial Intelligence*, 258, 66–95. doi:10.1016/j.artint.2018.01.002.

[3] Tim Baarslag, Mark J. C. Hendrikx, Koen V. Hindriks, and Catholijn M. Jonker. 2016. Learning about the opponent in automated bilateral negotiation: a comprehensive survey of opponent modeling techniques. *Autonomous Agents and Multi-Agent Systems*, 30, 5, 849–898. doi:10.1007/s10458-015-9309-1.

[4] Nolan Bard, Michael Johanson, Neil Burch, and Michael Bowling. 2013. Online implicit agent modelling. In *Proceedings of the Twelfth International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 255–262.

[5] Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. 2019. Deep counterfactual regret minimization. In *Proceedings of the 36th International Conference on Machine Learning* (Proceedings of Machine Learning Research). Vol. 97. PMLR, 793–802.

[6] Shengbo Chang and Katsuhide Fujita. 2024. COMB: scalable concession-driven opponent models using Bayesian learning for preference learning in bilateral multi-issue automated negotiation. *Group Decision and Negotiation*, 33, 5, 1143–1190. doi:10.1007/s10726-024-09889-7.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39, 1, 1–22. doi:10.1111/j.2517-6161.1977.tb01600.x.

[8] Bradley Efron. 1987. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82, 397, 171–185. doi:10.1080/01621459.1987.10478410.

[9] P. Faratin, C. Sierra, and N.R. Jennings. 2002. Using similarity criteria to make issue trade-offs in automated negotiations. *Artificial Intelligence*, 142, 2, 205–237. doi:10.1016/S0004-3702(02)00290-4.

[10] Peyman Faratin, Carles Sierra, and Nick R. Jennings. 1998. Negotiation decision functions for autonomous agents. *Robotics and Autonomous Systems*, 24, 3, 159–182. doi:10.1016/S0921-8890(98)00029-3.

[11] Sam Ganzfried. 2027. Consistent opponent modeling in imperfect-information games. In *Game Theory and AI for Security*. Springer Nature Switzerland, Cham, 3–21.

[12] Sam Ganzfried and Tuomas Sandholm. 2011. Game theory-based opponent modeling in large imperfect-information games. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 533–540.

[13] Sam Ganzfried and Qingyun Sun. 2018. Bayesian opponent exploitation in imperfect-information games. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, 1–8. doi:10.1109/CIG.2018.8490452.

[14] Zhenxing Ge, Zheng Xu, Tianyu Ding, Linjian Meng, Bo An, Wenbin Li, and Yang Gao. 2024. Safe and robust subgame exploitation in imperfect information games. In *Proceedings of the 41st International Conference on Machine Learning* (Proceedings of Machine Learning Research). Vol. 235. PMLR, 15255–15270.

[15] Johannes Heinrich and David Silver. 2016. Deep reinforcement learning from self-play in imperfect-information games. arXiv: 1603.01121.

[16] Koen Hindriks and Dmytro Tykhonov. 2008. Opponent modelling in automated multi-issue negotiation using Bayesian learning. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, 331–338.

[17] Martin Jaggi. 2013. Revisiting Frank-Wolfe: projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning* (Proceedings of Machine Learning Research) number 1. Vol. 28. PMLR, 427–435.

[18] N. R. Jennings, P. Faratin, A. R. Lomuscio, S. Parsons, M. J. Wooldridge, and C. Sierra. 2001. Automated negotiation: prospects, methods and challenges. *Group Decision and Negotiation*, 10, 2, 199–215. doi:10.1023/A:1008746126376.

[19] Michael Johanson and Michael Bowling. 2009. Data biased robust counter strategies. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics* (Proceedings of Machine Learning Research). Vol. 5. PMLR, 264–271.

[20] Michael Johanson, Martin Zinkevich, and Michael Bowling. 2007. Computing robust counter-strategies. In *Advances in Neural Information Processing Systems*. Vol. 20. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2007/file/6e7b33fdea3adc80ebd648fffb665bb8-Paper.pdf.

[21] H. W. Kuhn. 1953. 11. Extensive games and the problem of information. In *Contributions to the Theory of Games, Volume II*. Harold William Kuhn and Albert William Tucker, (Eds.) Princeton University Press, Princeton, 193–216. doi:10.1515/9781400881970-012.

[22] Marc Lanctot, Kevin Waugh, Martin Zinkevich, and Michael Bowling. 2009. Monte Carlo sampling for regret minimization in extensive games. In *Advances in Neural Information Processing Systems*. Vol. 22. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2009/file/00411460f7c92d2124a67ea0f4cb5f85-Paper.pdf.

[23] Marc Lanctot et al. 2019. OpenSpiel: a framework for reinforcement learning in games. *CoRR*. arXiv: 1908.09453.

[24] Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or No Deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2443–2453. doi:10.18653/v1/D17-1259.

[25] Richard Mealing and Jonathan L. Shapiro. 2017. Opponent modeling by expectation–maximization and sequence prediction in simplified poker. *IEEE Transactions on Computational Intelligence and AI in Games*, 9, 1, 11–24. doi:10.1109/TCIAIG.2015.2491611.

[26] Marc Ponsen, Jan Ramon, Tom Croonenborghs, Kurt Driessens, and Karl Tuyls. 2008. Bayes-relational learning of opponent models from incomplete information in no-limit poker. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*. AAAI Press, 1485–1486.

[27] Finnegan Southey, Michael Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. 2005. Bayes' bluff: opponent modelling in poker. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*. AUAI Press, Arlington, Virginia, USA, 550–558.

[28] Finnegan Southey, Bret Hoehn, and Robert C. Holte. 2009. Effective short-term opponent exploitation in simplified poker. *Machine Learning*, 74, 2, 159–189. doi:10.1007/s10994-008-5091-5.

[29] Bernhard von Stengel. 1996. Efficient computation of behavior strategies. *Games and Economic Behavior*, 14, 2, 220–246. doi:10.1006/game.1996.0050.

[30] Colin R. Williams, Valentin Robu, Enrico H. Gerding, and Nicholas R. Jennings. 2011. Using Gaussian processes to optimise concession in complex negotiations against unknown opponents. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume One*. AAAI Press, 432–438.

[31] Renbo Zhao. 2023. Convergence rate analysis of the multiplicative gradient method for PET-type problems. *Operations Research Letters*, 51, 1, 26–32.

[32] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. 2007. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems*. Vol. 20. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2007/file/08d98638c6fcd194a4b1e6992063e944-Paper.pdf.

A SLME Computation

We count costs in calls to the game's rules, which never require the game tree. From a history, a call returns in unit time the acting player and its information set or, for a given action or chance outcome, its legality, successor, and observation by the learner, plus a chance outcome's probability; listing legal actions or outcomes, or computing a library strategy's distribution over them, takes linear time. Arithmetic, log, and exp cost one unit. Histories have at most $d_{\max}$ actions and chance outcomes, with at most $b_{\max}$ legal at each.

Summing likelihoods over hidden histories, such as an opponent's unrevealed cards, is standard [27, Section 3.3]; SLME and OISR do it without the game tree. Hashing each play's observation and the product of its learner action probabilities, which fix its known coefficients (Section 2), groups the plays into types in expected time linear in the plays' total record length N_R. For each type, a depth-first search follows the recorded play. It takes each event that the learner observed exactly as recorded, branches over all legal events only where the learner did not, extends only histories whose observations match the start of the recorded one, and keeps the terminal histories whose observations equal it. By perfect recall (Section 2), every prefix of a compatible history passes this test, so the search finds exactly Z_j, and the chance and learner probabilities along each give $c_{j,z}$. The search stores only its current path and the untried branches along it, so it never constructs the game tree. It takes $O(d_{\max} N_C)$ time, where N_C counts its maximal paths over all types, ending at a kept history or at a mismatch.

SLME accumulates each q_j^m, a sum over Z_j (Section 2), during the search in $O(Mb_{\max}d_{\max}N_H)$ time for the $N_H := \sum_j |Z_j|$ histories found, each with at most $d_{\max}$ opponent actions. An EM pass takes $O(|V|M)$ time (Section 3), and under Lemma 1's assumptions, at most $(\log M)/\varepsilon$ passes from uniform weights reach a per-play gap of $\varepsilon > 0$. SLME takes expected time $O(N_R + d_{\max}N_C + Mb_{\max}d_{\max}N_H + |V|M\log(M)/\varepsilon)$ and $O(N_R + b_{\max}d_{\max} + |V|M)$ space for the type keys, the search, and the table.

B OISR

Throughout, σ and y_s abbreviate σ_{-i} and $y_{-i,s}$. Except in Appendix B.1.1, we fix the data ($n \geq 1$) and Theorem 1's assumptions: a positive coherent prior and a positive entry in every row of C. A vector or strategy is *positive* if all entries or action probabilities are.

B.1 The Prior in Realization Weights

We analyze EM in realization weights, where the log likelihood and, by coherence, the summed log prior are concave. Substituting $\sigma(a \mid I) = y_{Ia}/y_{\rho(I)}$ into the retained prior terms of Eq. (6) and collecting each $\log y_s$, with $\log y_\emptyset = 0$, gives

$$J_n(y) - K = \ell_n(y) + \textstyle\sum_{s\in\widetilde{S}_{-i}} \kappa_s \log y_s, \tag{15}$$

where $\kappa_\emptyset := 0$ and $\kappa_{Ia} := \lambda(I,a) - \sum_{I'\in\widetilde{\mathcal{I}}_{-i}:\rho(I')=Ia} \lambda(I')$. Coherence (Eq. (10)) bounds $\lambda(I,a)$ below by the sum over all I' with $\rho(I') = Ia$, retained or not, so every $\kappa_s \geq 0$ and J_n is concave in y.

The prior acts like imaginary, fully observed plays, allowed to stop at any retained sequence s, with likelihood proportional to y_s and weight κ_s. A sequence *extends* each of its prefixes, itself included; the imaginary plays whose sequence extends Ia weigh

$$\textstyle\sum_{s \text{ extends } Ia} \kappa_s = \lambda(I,a), \tag{16}$$

by backward induction, as the plays continuing to a retained I' with $\rho(I') = Ia$ weigh $\sum_{a'} \lambda(I',a') = \lambda(I')$, which κ_{Ia} subtracts. Summing Eq. (16) over the opponent's first retained decisions, all imaginary plays weigh τ, and with the recorded plays, $W := n + \tau$.

B.1.1 A positive coherent prior. We set the pseudocounts by depth, with bounds that certify coherence without listing the information sets. The depth d_I of I is the number of decisions the opponent makes before reaching I, and D is the largest depth. Each information set of depth d gets the same total pseudocount $\lambda_d > 0$, split evenly: $\lambda(I,a) := \lambda_{d_I}/|A(I)|$, which is positive and needs only d_I and $|A(I)|$, not the game tree. Let A_d and F_d bound, at depth d, the number of actions and the number of information sets I' with $\rho(I') = Ia$ after each action a. For $d < D$, $\lambda(I,a) \geq \lambda_d/A_d$, and each of the at most F_d information sets after a has total pseudocount λ_{d+1}; at depth D, none follow. So $\lambda_d \geq A_dF_d\lambda_{d+1}$ for all $d < D$ makes the prior coherent. We set $\lambda_D := B_{\text{prior}}$, the prior strength, and, with the smallest bounds the rules allow, $\lambda_d := 1.1\,A_dF_d\lambda_{d+1}$ for $d < D$, or $\lambda_d := B_{\text{prior}} + 1.1\,A_dF_d\lambda_{d+1}$ in Leduc.

B.2 Initialization

The endpoints carry the recorded plays, and the rate of Lemma 2 depends, through Γ, on how small their initial weights are, so the initialization of Eq. (11) mixes in feasible weights x whose smallest endpoint weight, $1/\Gamma$, is as large as possible. Feasible weights, those in $\widetilde{Y}_{-i}$, are the nonnegative y with $y_\emptyset = 1$ that satisfy the flow constraints at the retained information sets, by the converse in Section 2, as the prefixes of retained sequences are retained.

Each retained I with $\rho(I) = s$ receives the full weight y_s and splits it among its actions. So for every endpoint extending s to get weight at least one, s needs weight at least

$$r_s := \max\Big\{\mathbf{1}\{s \text{ is an endpoint}\},\ \max_{I\in\widetilde{\mathcal{I}}_{-i}:\rho(I)=s} \textstyle\sum_{a\in A(I)} r_{Ia}\Big\}, \tag{17}$$

where an empty maximum is zero. A backward pass over the retained sequences computes r in linear time. An endpoint exists, as the rows of C are nonzero, and Eq. (17) makes r at least one there and at its prefixes, so $r_\emptyset \geq 1$. If a feasible y has smallest endpoint weight m, backward induction with the flow constraints gives $y_s \geq m\,r_s$, so $m \leq 1/r_\emptyset$. A forward pass attains this bound in linear time: from $x_\emptyset := 1$, at each retained I, it gives each action a the weight $r_{Ia}/r_\emptyset$ and adds the surplus $x_{\rho(I)} - \sum_a r_{Ia}/r_\emptyset$, nonnegative as, inductively, $x_{\rho(I)} \geq r_{\rho(I)}/r_\emptyset$, to one action. So x is feasible with every endpoint weight at least $1/r_\emptyset$, and $\Gamma = r_\emptyset$. Induction on Eq. (17) also bounds r_s by the number of endpoints extending s, so $\Gamma \leq |S_{-i}|$ for every dataset, the uniform bound used in Section 5.

As $\widetilde{\mathcal{I}}_{-i} \neq \emptyset$, the opponent's first decision on some compatible history is retained, so $\tau > 0$ by positivity. So $y^{(0)} = (nx + \tau y^{\text{def}})/W$, a mixture of feasible weights, is feasible and positive, with $y^{(0)} \geq (\tau/W)\,y^{\text{def}}$ and, at endpoints, $y_s^{(0)} \geq nx_s/W \geq n/(W\Gamma)$.

B.3 Proofs of Theorem 1 and Lemma 2

B.3.1 Monotonicity and an optimal fixed point. Jensen's inequality and the exact change in the retained log prior give the first inequality of Eq. (9), EM's ascent bound with the prior added [7, Section 2 and Theorem 1]. By the update (8), its right side equals $\sum_I (N^{(k)}(I) + \lambda(I))\, D_{\mathrm{KL}}\big(\sigma^{(k+1)}(\cdot \mid I) \,\|\, \sigma^{(k)}(\cdot \mid I)\big)$, which is positive unless the pass leaves the model unchanged. Positive pseudocounts suffice, without coherence, and keep every iterate positive.

On the compact set of behavioral strategies, J_n is upper semicontinuous and finite exactly at the positive strategies, so it attains J_n^* at a positive strategy σ^{opt}, with weights y^{opt} and expected counts N^{opt}, which plays the prior modes at the non-retained information sets (Section 4). By Eq. (9), a pass from σ^{opt} cannot change it without exceeding J_n^*, so $\sigma^{\text{opt}}(a \mid I) = (N^{\text{opt}}(I,a) + \lambda(I,a))/(N^{\text{opt}}(I) + \lambda(I))$.

B.3.2 Global convergence. To compare $\sigma^{(k)}$ with σ^{opt}, let

$$\mathcal{D}_k := \textstyle\sum_{I\in\widetilde{\mathcal{I}}_{-i}} \big(N^{\text{opt}}(I) + \lambda(I)\big)\, D_{\mathrm{KL}}\big(\sigma^{\text{opt}}(\cdot \mid I) \,\|\, \sigma^{(k)}(\cdot \mid I)\big),$$

which weights each retained I by the expected weight of real and imaginary plays reaching it under σ^{opt} (Eq. (16)); it is nonnegative, and finite as $\sigma^{(k)}$ is positive. Let $\Delta_k := J_n^* - J_n(\sigma^{(k)})$, and let $\beta(s) := \sum_j n_j\gamma_{j,s} + \kappa_s$ be the expected weight of real and imaginary plays whose complete opponent sequence is s, which totals $W = n + \tau$; superscripts mark the model, and $0\log 0 = 0$. Then

$$\begin{aligned} \mathcal{D}_{k+1} &\leq \textstyle\sum_s \beta^{\text{opt}}(s) \log\big(\beta^{\text{opt}}(s)/\beta^{(k)}(s)\big) \\ &\leq \textstyle\sum_j n_j D_{\mathrm{KL}}\big(\gamma_{j,\cdot}^{\text{opt}} \,\|\, \gamma_{j,\cdot}^{(k)}\big) = \mathcal{D}_k - \Delta_k. \end{aligned} \tag{18}$$

For the equality, expanding each $\log(y_s^{\text{opt}}/y_s^{(k)})$ into action log-ratios and applying Eq. (16) and the fixed point gives $\sum_s \beta^{\text{opt}}(s)\log(y_s^{\text{opt}}/y_s^{(k)}) = \mathcal{D}_k$. By Eq. (15), Δ_k is this sum with each type's part $n_j \sum_s \gamma_{j,s}^{\text{opt}} \log(y_s^{\text{opt}}/y_s^{(k)})$ replaced by $n_j \log(q_j(y^{\text{opt}})/q_j(y^{(k)}))$; as $\gamma_{j,\cdot}^{\text{opt}}$ sums to one and $y_s/q_j(y) =$

$\gamma_{j,s}/C_{j,s}$ where $C_{j,s} > 0$, the difference is the n_j-weighted divergence between each type's two posteriors over hidden sequences.

The second inequality follows by the log-sum inequality at each s, merging $n_j\gamma_{j,s}$ and the shared $\kappa_s \geq 0$ (coherence). For the first, read β/W as the distribution of a play's complete opponent sequence, generated one decision at a time: by Eq. (16) with the update (8) and the fixed point, respectively, a play that reaches a retained I continues with a with probability $\sigma^{(k+1)}(a \mid I)$ under $\beta^{(k)}$ and $\sigma^{\mathrm{opt}}(a \mid I)$ under β^{opt}, so the chain rule for relative entropy splits the first inequality's right side into $\mathcal{D}_{k+1}$ and nonnegative terms for stopping and for which retained information set comes next.

Summing Eq. (18) over the first k passes gives $\sum_{t<k} \Delta_t \leq \mathcal{D}_0$; as the gaps never increase (Eq. (9)), $\Delta_k \leq \mathcal{D}_0/k$ for $k \geq 1$, which tends to zero and proves Theorem 1.

B.3.3 The rate. Since $\Delta_k \leq \mathcal{D}_0/k$, Lemma 2 follows once $\mathcal{D}_0 \leq B_0$. As in the proof of Eq. (18), $\mathcal{D}_0 = \sum_s \beta^{\mathrm{opt}}(s) \log(y_s^{\mathrm{opt}}/y_s^{(0)})$. The real plays have total weight n and lie on endpoints, where $y_s^{\mathrm{opt}} \leq 1$ and $y_s^{(0)} \geq n/(W\Gamma)$ (Appendix B.2), so they contribute at most $n\log(W\Gamma/n)$. The part τy^{def} of Eq. (11) gives $y^{(0)} \geq (\tau/W)\, y^{\mathrm{def}}$, and the prior modes maximize $\sum_s \kappa_s \log y_s$, which by Eq. (16) equals the retained log prior $\sum_{I,a} \lambda(I,a) \log \sigma(a \mid I)$, so the imaginary plays, of total weight τ, contribute at most $\tau \log(W/\tau)$. With $W = n + \tau$, the two bounds sum to B_0. As $\log(1+u) \leq u$, $B_0/n \leq \log\Gamma + \tau/n + 1$, bounded for a fixed game and prior, as Section 5 uses: $\Gamma \leq |S_{-i}|$ (Appendix B.2), and τ is at most the prior's total at the opponent's first decisions.

B.4 Computation

Escape actions. To make each pass cost $O(|V| + |E| + |\bar{S}|)$, OISR represents only the sequences $\bar{S}$ along compatible terminal histories (Section 4), including the empty sequence, plus at most one escape weight per retained information set. At a retained I, the actions a with $Ia \notin \bar{S}$, which no compatible history takes, enter J_n only through their prior terms $\lambda(I,a)\log y_{Ia}$, and the flow constraint at I only through their total weight: such an Ia is no endpoint and precedes no retained information set, so $\kappa_{Ia} = \lambda(I,a)$ in Eq. (15). Where such actions exist, OISR merges them into one escape action e with their total pseudocount $\lambda(I,e)$. By Gibbs' inequality, splitting the escape weight in proportion to their pseudocounts maximizes their prior terms, making J_n the merged objective plus a constant; so splitting a merged fit gives a full MAP fit with the same objective gap. The merged model keeps all that Appendix B.3 uses of the prior (positivity, $\kappa \geq 0$, Eq. (16), and $\sum_s \kappa_s = \tau$), with $\kappa_{Ie} = \lambda(I,e)$. Merging preserves feasible endpoint weights and the totals $\lambda(I)$, hence Γ, τ, and B_0, so Theorem 1 and Lemma 2 hold for the merged model, hence for the split fits.

Costs. Preparation groups the plays and runs the search of Appendix A, in expected time $O(N_R + d_{\max}N_C)$. For each history $z \in Z_j$ found, OISR adds $c_{j,z}$ to $C_{j,s_{-i}(z)}$ and enters the opponent's decisions along z into hash tables of the retained information sets and of $\bar{S}$, in expected time $O(d_{\max})$, plus $O(b_{\max})$ to set the pseudocounts of each new information set. Each retained I is followed by some $Ia \in \bar{S}$, and each history found has at most $d_{\max}$ opponent decisions, so $|\widetilde{I}_{-i}| < |\bar{S}| \leq 1 + d_{\max}N_H$, and the merged model has fewer than $2|\bar{S}|$ weights. Preparation, including the $O(|\bar{S}|)$ computation of x and $y^{(0)}$, thus takes expected time $O(N_R + d_{\max}N_C + b_{\max}d_{\max}N_H)$. A pass computes the q_j and $\gamma_{j,s}$ over the nonzeros of C in $O(|V| + |E|)$ time; a backward pass over $\bar{S}$ accumulates the expected number of recorded plays whose sequence extends each s, which at $s = Ia$ is $N(I,a)$, zero for escape actions; and a forward pass applies the update (8) and recomputes the weights, in $O(|\bar{S}|)$ time. Beyond $O(|V| + |E| + |\bar{S}|)$ numbers for C, the counts, and the model, OISR uses $O(N_R + b_{\max}d_{\max})$ space for type keys and the search.

C Proof of Theorem 2

Let r_n be the empirical distribution of observation records in U. For an opponent strategy σ, define the average observation distribution

$$p_n^{\sigma}(o) := \left(\frac{1}{n}\sum_{t=1}^{n} y_{i,s_i(o)}(\sigma_i^t)\right) f_o(\sigma),$$

with $s_i(o)$ the learner's sequence recorded in o. Write $p_n^* := p_n^{\sigma^*_{-i}}$ and $\hat{p}_n := p_n^{\hat{\sigma}_{-i,n}}$. The strong law for bounded martingale differences gives $\|r_n - p_n^*\|_1 \to 0$ almost surely. The coverage condition bounds $p_n^*(o)$ away from zero whenever $f_o(\sigma^*_{-i}) > 0$, and no other observation can occur. Consequently, $D_{\mathrm{KL}}(r_n \| p_n^*) \to 0$ almost surely.

The known learner factors cancel from likelihood ratios, giving

$$D_{\mathrm{KL}}(r_n \| \hat{p}_n) = D_{\mathrm{KL}}(r_n \| p_n^*) + \frac{\ell_n(\sigma^*_{-i}) - \ell_n(\hat{\sigma}_{-i,n})}{n}.$$

Equation (13) therefore implies $D_{\mathrm{KL}}(r_n \| \hat{p}_n) \to 0$. Pinsker's inequality and the triangle inequality yield $\|\hat{p}_n - p_n^*\|_1 \to 0$. Dividing each coordinate difference by the corresponding average learner realization weight, which is at least θ, proves the claim.

D Experiments

D.1 KL Divergence from the True Opponent

We compare opponent strategies by the KL divergence between the distributions they induce on the observation of a learner that plays uniformly at random,

$$D_{\mathrm{KL}}\left(q(\sigma_{-i}) \,\|\, q(\sigma'_{-i})\right) = \sum_j q_j(\sigma_{-i}) \log \frac{q_j(\sigma_{-i})}{q_j(\sigma'_{-i})}, \tag{19}$$

where j ranges over this learner's possible observations, each its own type (Section 2). Observationally equivalent strategies induce the same distribution, so a positive divergence certifies that two strategies are not observationally equivalent.

We measure how well a fitted model $\widehat{\sigma}_{-i,n}$ predicts what the learner observes by the divergence of Eq. (19) from the true opponent strategy, $D_{\mathrm{KL}}(q(\sigma^*_{-i}) \,\|\, q(\widehat{\sigma}_{-i,n}))$, which is zero exactly when the model is observationally equivalent to the true opponent. For SLME and BBR, $q_j(\widehat{\sigma}_{-i,n})$ is $q_j(\pi)$ under the fitted weights or posterior.

In Fig. D.1, as with regret (Fig. 1), the library methods level off against randomly sampled opponents while OISR keeps improving at large data sizes, and at $c = 100$, MS-EM's divergence is 1.7 to 13 times OISR's across all games and opponents. Unlike regret, SLME's divergence keeps falling against library members and library mixtures after its responses are nearly optimal, by a factor of 55 to 94 from $c = 1$ to $c = 100$; against randomly sampled opponents in Deal or No Deal, OISR's divergence first rises, so that MS-EM's is lower at intermediate data sizes; and at $c = 100$, FMAP's is lower

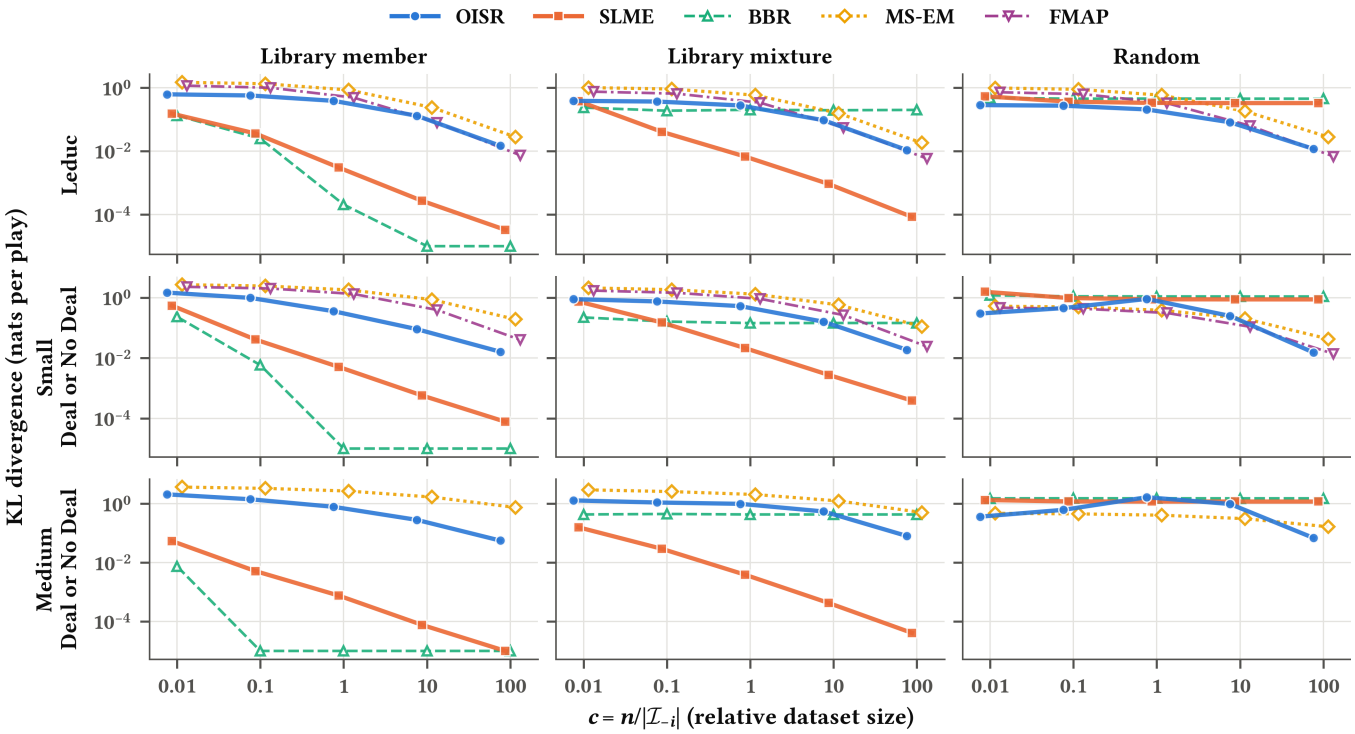

Figure D.1: KL divergence between the observation distributions of the true opponent strategy and each offline model, averaged over 10 worlds and both learner roles. The pointwise 95% CIs have median width 0.045 nats (max 1.24).

than OISR's in Leduc, but higher in small Deal or No Deal against library members and mixtures (0.040 and 0.024 against 0.016 and 0.019).

D.2 Prior Strength

We vary B_{prior}, which scales OISR's coherent prior: larger values more strongly favor uniform action probabilities (Appendix B.1.1). We test $\{10^{-6}, 10^{-5}, \ldots, 10^{3}\}$ in Leduc using the offline dataset sizes and stopping criterion (Section 6.2), pooling all 10 worlds at each of the three opponent classes; 5 of the 3,000 fits reach the iteration limit.

Fig. D.2 shows that stronger regularization limits the gains from additional observations. With five observed plays ($c = 0.01$), all tested strengths give normalized regret near the uniform-model reference (0.96–1.00). The default $B_{\text{prior}} = 0.1$ is within 5% of the lowest pooled regret among tested strengths for $c \leq 1$, but at $c = 100$ its regret is 0.039, compared with approximately 0.020 for $B_{\text{prior}} \leq 10^{-3}$. A smaller fixed strength $B_{\text{prior}} = 0.01$ is within 8% of the lowest pooled regret at every tested dataset size.

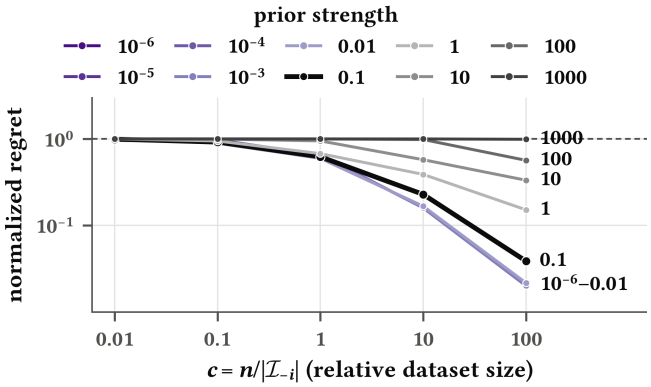

Figure D.2: OISR's normalized regret in Leduc by prior strength, pooled over the three opponent classes and both learner roles across the ten worlds (Eq. (14)). The pointwise 95% CIs have median width 0.039 normalized-regret units (max 0.183).