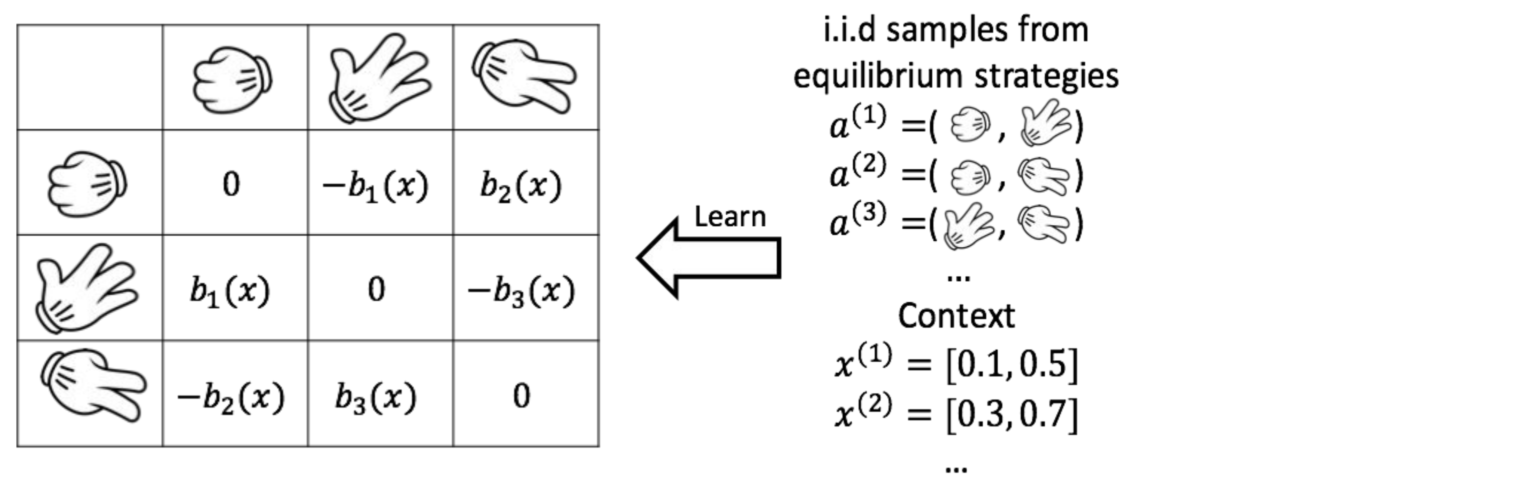
Large Scale Learning of Agent Rationality in Two-Player Zero-Sum Games

Chun Kai Ling[1], Fei Fang[2], J. Zico Kolter[1,3]
Department of Computer Science[1], *Institute for Software Research*[2], *Carnegie Mellon University*
Bosch Center for Artificial Intelligence[3]
chunkail@cs.cmu.edu, feifang@cmu.edu, zkolter@cs.cmu.edu

Carnegie Mellon University

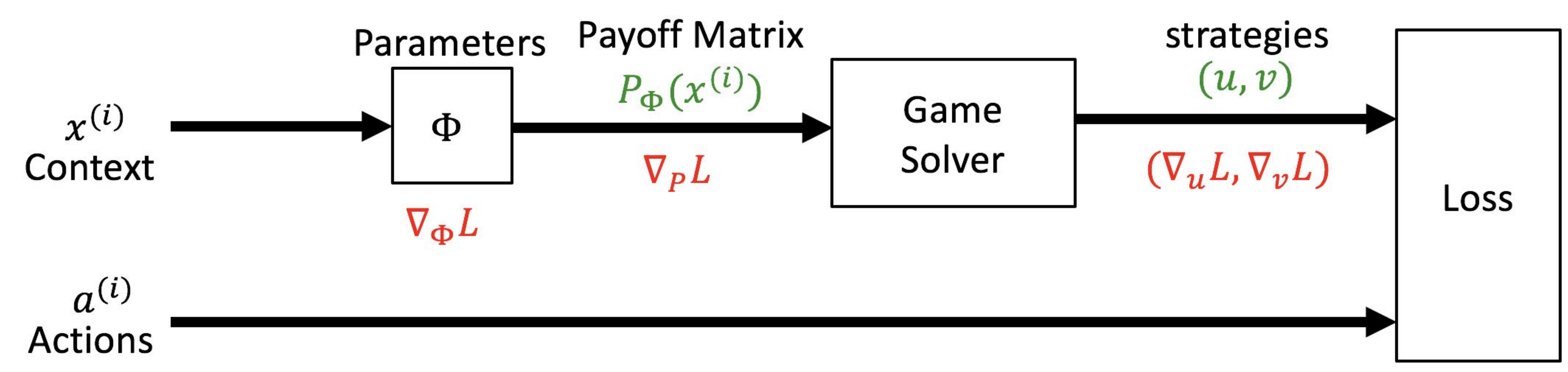
1. Introduction

- Game theory finds optimal strategies based on known payoffs. Our setting, sometimes known as inverse game theory (Kuleshov, Waugh et al, 2011) is the reverse.
- Our objective is to learn underlying game parameters by observing player actions *only*.
- Learning utilities allows us to better understand the problem, as opposed to directly predicting strategies.
- Example: Given a context (features) x, learn a payoff matrix $P(x)$ which has equilibrium distributions similar to the actions observed



2. End-to-end learning

- Based on recent work by Ling et. al. (2018)
- Assumes that players act according to the logit Quantal Response Equilibrium (QRE, McKelvey, 1993) of the reduced normal form.
- By differentiating `through' game solutions, training is performed end-to-end using SGD to minimize log-loss.
- Scales up to larger extensive form games by exploiting the sequence form representation (Von Stengel, 1996).



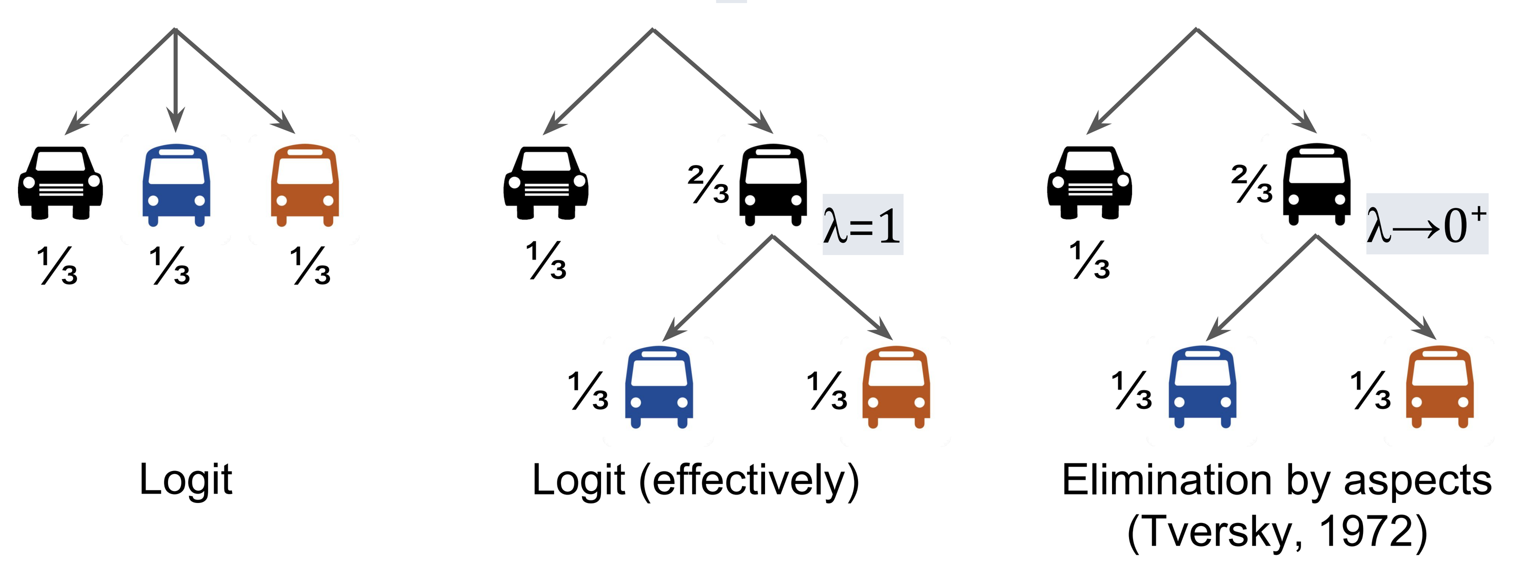
- Relies on the strong assumption that players play according to the QRE of the reduced normal form.
- Scales poorly in size of the game, requires solving a regularized min-max problem and a linear system for every training point.

3a. Nested Logit (prior work)

- Logit model: action probabilities based on softmax

$$u_a^* = \frac{\exp(U_a/\lambda)}{\sum_{a' \in \mathcal{A}} \exp(U_{a'}/\lambda)} \quad \text{or} \quad u_a^* = \arg\max_{u \in \Delta_u} u^T U - \lambda H(u)$$

- λ corresponds to the level of (ir)rationality.
- Using logits to model player actions may lead to pathologies (Debreu 1960).
- *Nested Logits* generalize logits by grouping actions into nests, with different λ for each nest.



3b. Nested Logit QRE

- Obtained by extending nested logits to the multiplayer setting, or alternatively, adding varying levels of rationality at each infoset.
- Natural extension of dilated entropy regularization (Kroer et. al., 2018)

$$\min_u \max_v u^T P v + \sum_{h \in \mathcal{I}_u} \lambda_h \sum_{a \in \mathcal{A}_h} u_a \log \frac{u_a}{u_{p_h}} - \sum_{h \in \mathcal{I}_v} \lambda_h \sum_{a \in \mathcal{A}_h} v_a \log \frac{v_a}{v_{p_h}}$$

$$\text{subject to} \quad Eu = e, \quad Fv = f.$$

u, v: sequence form probs , p_h: action preceding infoset h.

- From a machine learning standpoint: additional λ parameters to be learned. Gradients of loss wrt λ are similar to Ling et. al., which is derived using the implicit function theorem.

$$\nabla_P L = y_u v^T + u y_v^T \qquad \nabla_{\lambda_h} L = \kappa_h^T y_u \quad \text{where}$$

$$(\kappa_h)_a = \begin{cases} 1 + \log(u_a / u_{p_{\rho_a}}), & \rho_a = h \\ -1, & h \in C_a \end{cases}$$

$$\Xi(u)_{ab} = \begin{cases} \frac{\lambda_{\rho_a} + \sum_{h' \in C_a} \lambda_{h'}}{u_a}, & a = b \\ -\frac{\lambda_{\rho_a}}{u_b}, & p_{\rho_a} = b \\ -\frac{\lambda_{\rho_b}}{u_a}, & p_{\rho_b} = a \end{cases} \qquad \begin{bmatrix} y_u \\ y_v \\ y_\mu \\ y_\nu \end{bmatrix} = \begin{bmatrix} -\Xi(u) & P & E & 0 \\ P^T & \Xi(v) & 0 & F \\ E^T & 0 & 0 & 0 \\ 0 & F^T & 0 & 0 \end{bmatrix}^{-1} \begin{bmatrix} -\nabla_u L \\ -\nabla_v L \\ 0 \\ 0 \end{bmatrix}$$

and quantities involving v are defined analogously to u.

4. Efficient Forward and backward passes

- Utilize first order methods (FOM) of Chambolle and Pock (2015) to solve convex-concave problems of the form

$$\min_{Ex=x_0} \max_{Fy=y_0} x^T P y + \mathcal{E}(x) - \mathcal{F}(y)$$

- Requires fast computation of "best response" subproblems

$$\arg\min_{Ex=x_0} x^T c_x + \mathcal{E}(x) \quad \text{and} \quad \arg\min_{Fy=y_0} y^T c_y + \mathcal{F}(y)$$

- Forward pass: subproblem may be solved easily by a single tree traversal or sparse matrix-vector multiplication (Kroer et. al, Hoda et. al., 2010)
- Backwards pass involves solving a large linear system to obtain y. (See 3b).
- Surprisingly, this may be recast to another convex-concave problem by observing that the linear system is precisely the KKT conditions of

$$\min_x \max_y \quad x^T P y + \frac{1}{2} x^T \Xi(u) x - \frac{1}{2} y^T \Xi(v) y + \nabla_u L^T x + \nabla_v L^T y$$

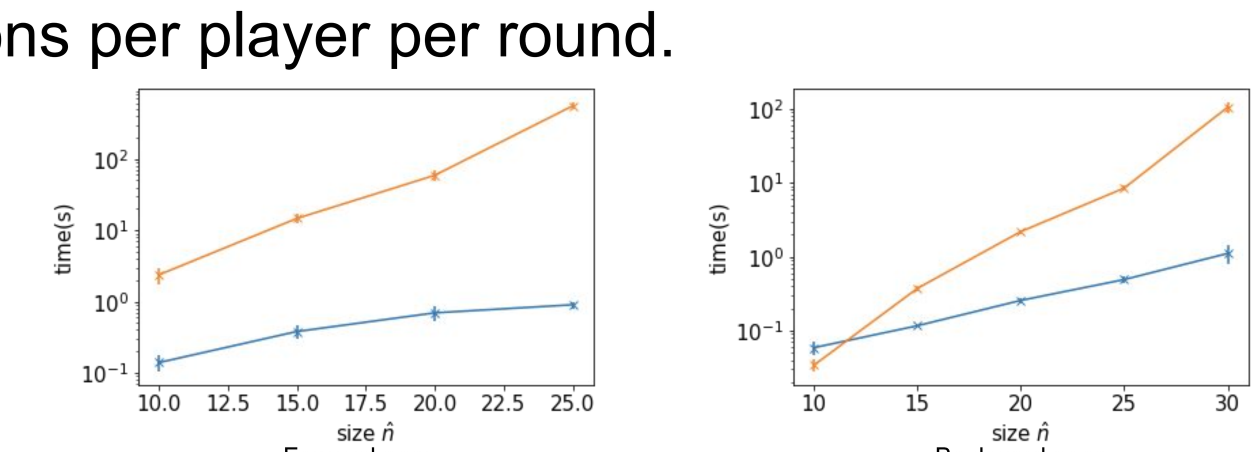
$$\text{subject to} \quad Ex = 0 \qquad Fy = 0.$$

- This may be solved again using the FOM of Chambolle and Pock, with subproblems that may be solved in linear time by exploiting tree structures of extensive form games

5. Experiments

5.1 Synthetic Payoff Matrices

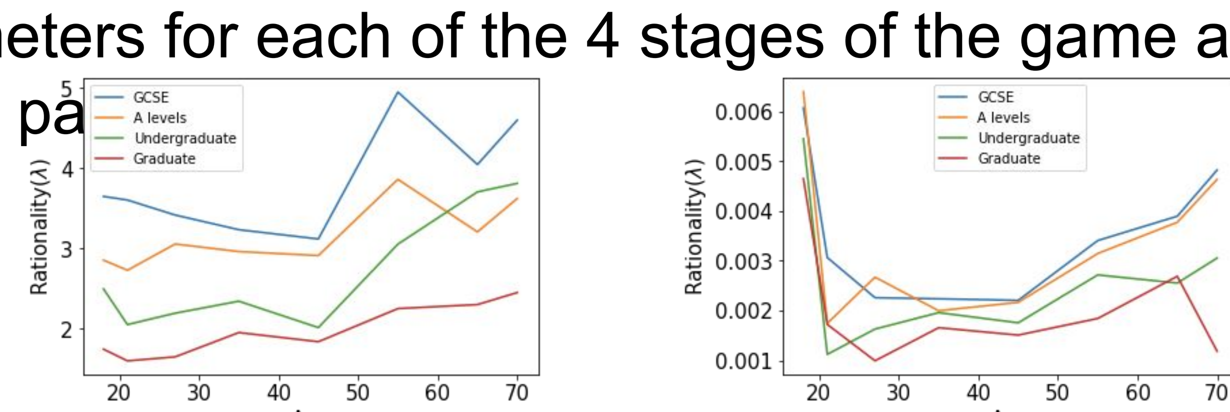
- Extensive form games with depth 2 random payoff matrices and $\hat{n}$ actions per player per round.



- Our method is orders of magnitude faster than our previous work

5.2 Information gathering dataset

- The information gathering game (Hunt et. al., 2016) is a one-player 4-stage game which requires players to trade-off between paying for exploration or taking a potentially risky action.
- We used age and education as features and learned λ-parameters for each of the 4 stages of the game as a function of these pa



- The better educated and middle aged demographics enjoy higher rationality (lower λ) corroborating with Hunt et. al..