What game are we playing? End-to-end Learning in Normal and Extensive Form 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

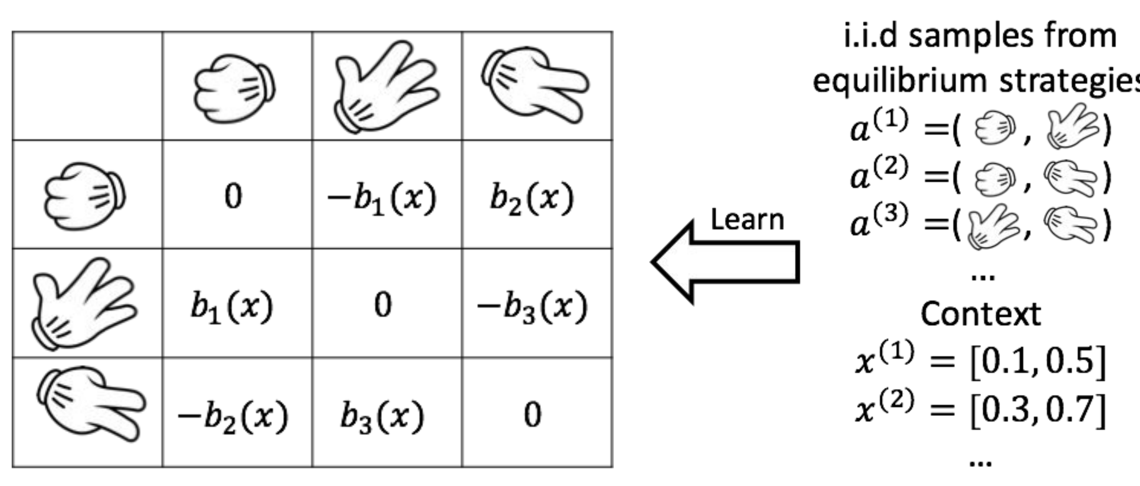
Carnegie Mellon University

1. Motivation

- Our objective is to learn underlying utilities of agents in zero-sum games by only observing player actions.
- 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.
- Learning the underlying utilities allows us to better understand the problem, as opposed to directly predicting strategies from context.

2. Setting

- Given a *context* x, we predict a matrix $\mathrm{P}(\mathrm{x})$, adapting to novel situations.
- Prior work either ignores context, or are restricted to special structural properties (e.g., symmetry in Vorobeychik, 2007).



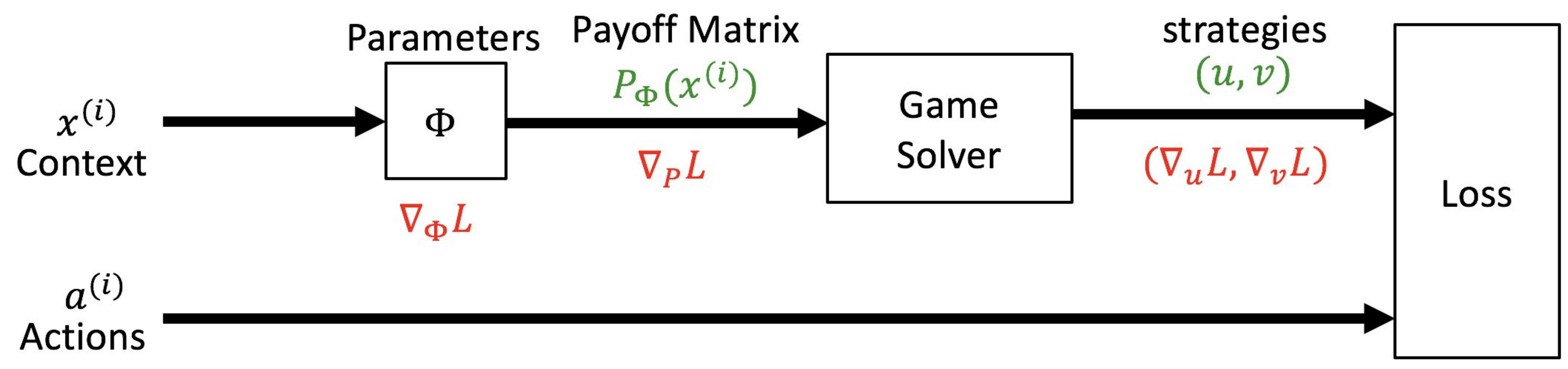
3. Contributions

- We assume that players act according to the logit Quantal Response Equilibrium (QRE, McKelvey, 1993).
- We propose a differentiable game solver to find the QRE.
- We derive gradients for 'differentiating through' game solutions, allowing for training to be done end-to-end using stochastic gradient descent to minimize log-loss.
- Our method scales up to larger extensive form games by exploiting the sequence form representation.
- Successfully learned payoffs for a range of synthetic data.

Algorithm 1: Learning parameters Φ using SGD

Input: training data $\{(x^{(i)}, a^{(i)})\}$, learning rate η, Φ_{init}
for *ep in* $\{0, \ldots, ep_{max}\}$ **do**
  Sample $(x^{(i)}, a^{(i)})$ from training data;
  Forward pass: Compute $P_\Phi(x^{(i)})$, QRE (u, v) and loss $L(a^{(i)}, u, v)$;
  Backward pass: Compute gradients $\nabla_u L, \nabla_v L, \nabla_P L, \nabla_\Phi L$;
  Update parameters: $\Phi \leftarrow \Phi - \eta \nabla_\Phi L$;
end

A typical use-case of our differentiable module



4. Normal Form Games

- Solution for QRE in zero-sum games is unique, smooth, and equivalent to a min-max problem with entropy regularization.

$$\min_{u \in \mathbb{R}^n} \max_{v \in \mathbb{R}^m} \quad u^T P v - H(v) + H(u) \qquad \text{subject to} \quad 1^T u = 1, \quad 1^T v = 1$$

- Convex-concave problem: efficient solution with Newton's method

$$Q = \begin{bmatrix} \mathrm{diag}(\frac{1}{u}) & P & 1 & 0 \\ P^T & -\mathrm{diag}(\frac{1}{v}) & 0 & 1 \\ 1^T & 0 & 0 & 0 \\ 0 & 1^T & 0 & 0 \end{bmatrix} \qquad Q \begin{bmatrix} \Delta u \\ \Delta v \\ \Delta \mu \\ \Delta \nu \end{bmatrix} = - \begin{bmatrix} Pv + \log u + 1 + \mu 1 \\ P^T u - \log v - 1 + \nu 1 \\ 1^T u - 1 \\ 1^v - 1 \end{bmatrix}$$

Hessian of Lagrangian — Newton Step

- Implicit differentiation (Dontchev & Rockafellar, 2009) yields gradients for backpropagation expressed by Jacobian of KKT conditions (Amos & Kolter, 2017)

$$[y_u \;\; y_v \;\; y_\mu \;\; y_\nu]^T = Q^{-1} [-\nabla_u L \;\; -\nabla_v L \;\; 0 \;\; 0]^T$$
$$\nabla_P L = y_u v^T + u y_v^T$$

5. Extensive Form Games

- We apply sequence form representation (Von Stengel, 1996) for computational efficiency.
- Dilated entropy regularization: entropy of behavioral strategy weighted by probabilities (in isolation of chance and other players).

$$\min_u \max_v u^T P v + \sum_{i \in \mathcal{I}_u} \sum_{a \in \mathcal{A}_i} u_a \log \frac{u_a}{u_{p_i}} - \sum_{i \in \mathcal{I}_v} \sum_{a \in \mathcal{A}_i} v_a \log \frac{v_a}{v_{p_i}} \qquad Eu - e = 0, \quad Fv - f = 0.$$

- *Theorem*: solution with dilated entropy regularization is realization equivalent to QRE of the game in reduced normal form.
- Solutions to min-max problem are obtained using Newton's method.

$$\Xi(u)_{ab} = \begin{cases} -\frac{1+J_a}{u_a}, a = b \\ \frac{1}{u_b}, p_{\rho_a} = b \\ \frac{1}{u_a}, p_{\rho_b} = a \end{cases} \quad \Xi(v)_{a'b'} = \begin{cases} -\frac{1+J_{a'}}{v_{a'}}, a' = b' \\ \frac{1}{v_{b'}}, p_{\rho_{a'}} = b' \\ \frac{1}{v_{a'}}, p_{\rho_{b'}} = a' \end{cases} \quad Q = \begin{bmatrix} -\Xi(u) & P & E^T & 0 \\ P^T & \Xi(v) & 0 & F^T \\ E & 0 & 0 & 0 \\ 0 & F & 0 & 0 \end{bmatrix} \quad Q \begin{bmatrix} \Delta u \\ \Delta v \\ \Delta \mu \\ \Delta \nu \end{bmatrix} = -g(u, v, \mu, \nu)$$

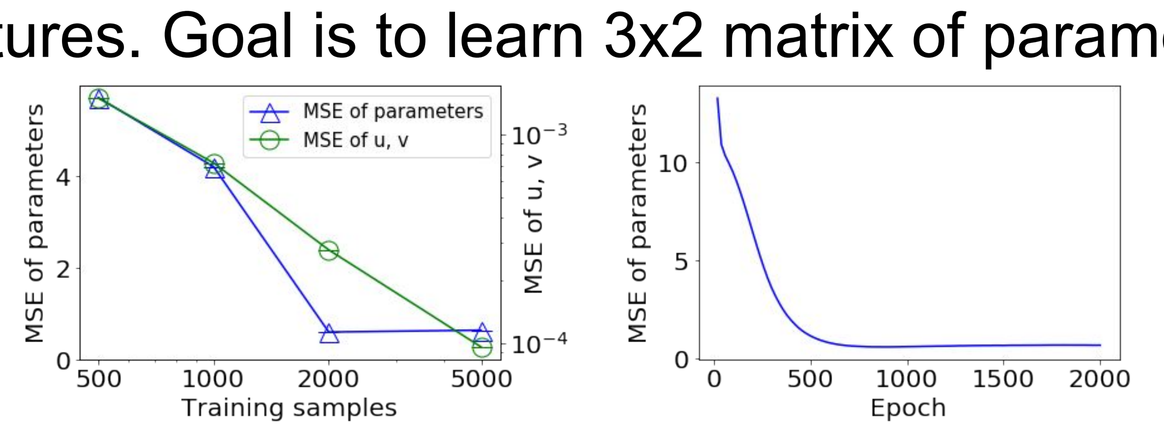
Hessian of Lagrangian — Newton Step (full details in paper)

- Gradient expressions are identical to the normal form case (sec. 4).

6. Experiments

A. Featurized Rock-Paper-Scissors

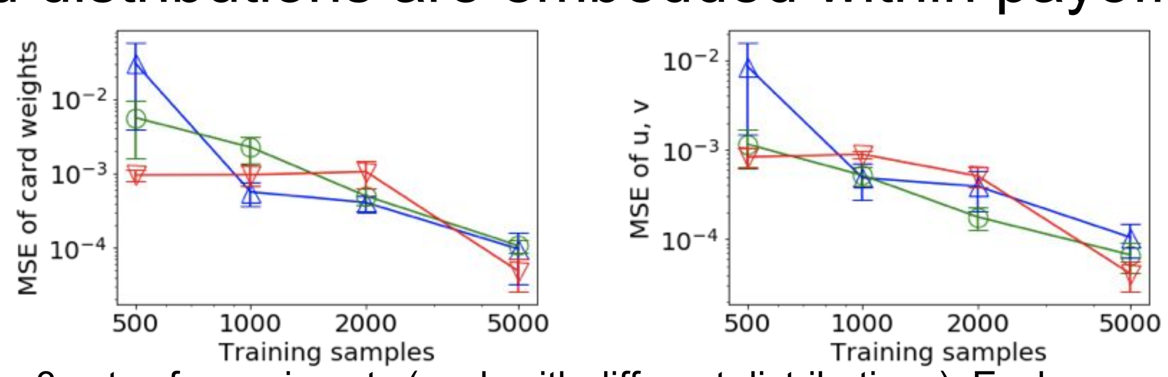
- Payoffs for each combination is a linear combination of 2 features. Goal is to learn 3x2 matrix of parameters.



- Able to learn parameters and accurately predict player strategies even in novel contexts.

B. One Card Poker

- Variant with 4 cards and nonuniform card distributions.
- Learn players' perceived card distributions from actions of player (these may not be true distributions).
- Card distributions are embedded within payoff matrix.



3 sets of experiments (each with different distributions). Each experiment is run 5 times. Error bars denote standard errors

- Results show that learning of attributes other than actual payoffs is possible (e.g. strategies of chance player).
- Able to learn when payoffs are nonlinear in parameters.

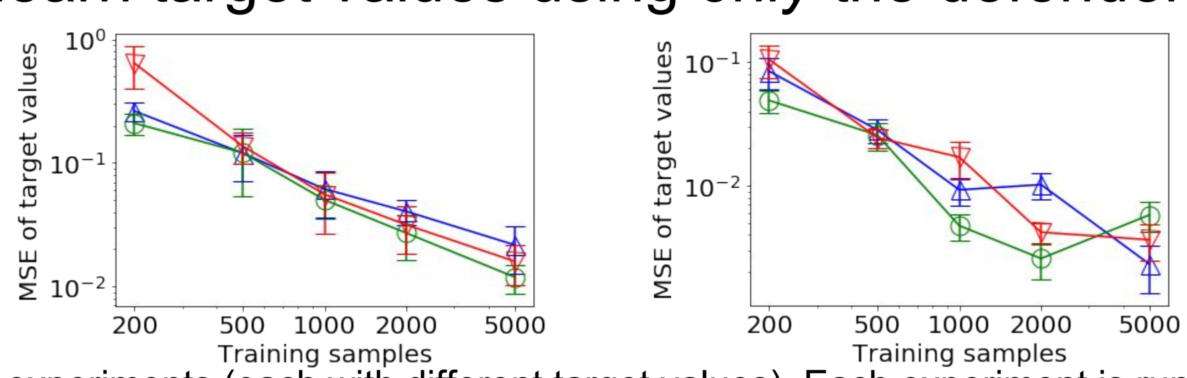
C. Resource Allocation Security Game

- There are N distinct targets with differing values. K defensive resources are split between each target. Each resource stops an attack with independently with probability 0.5.

	$(0,3)$	$(1,2)$	$(2,1)$	$(3,0)$
Target 1	$-R_1$	$-R_1/2$	$-R_1/4$	$-R_1/8$
Target 2	$-R_2/8$	$-R_2/4$	$-R_2/2$	$-R_2$

Example payoffs with N=2 and K=3

- Diminishing returns implies defender should spread his resources.
- The game proceeds in T iterations. After each iteration, the attacker is informed if the attack was successful and is allowed to alter his strategy. The defender is not allowed to reallocate his resources.
- It is unlikely that one obtains data from both attacker and defender.
- Our goal is learn target values using *only* the defender's actions.



3 sets of experiments (each with different target values). Each experiment is run 10 times. Error bars denote standard errors. Results are for N=2, K=5, T=1 (right) and T=2 (left).

7. Conclusion

- Issues regarding nonidentifiability occur with overparameterization.
- Future work include faster solvers for larger extensive form games, extension to non zero-sum games and application to real datasets and other domains (e.g. RL).