Beyond Equilibrium: Predicting Human Behavior in Normal-Form Games

James Wright & Kevin Leyton-Brown
University of British Columbia

July 15, 2010

Context

- Game theory: Mathematical study of behavior in idealized strategic multiagent settings.
  - Idealized agents, not human agents.
- Behavioral game theory: Aims to extend game theory to modelling human agents.
  - There are a wide range of BGT models in the literature.
  - BGT focuses on explaining behavior rather than predicting it.
  - Not much work compares different models' predictive power.

Game theory: Normal form game

In a normal form game:

- Each agent simultaneously chooses an action from a finite action set.
- Each combination of actions yields a known utility to each agent.
- The agents may choose actions either deterministically or stochastically.

Nash equilibrium

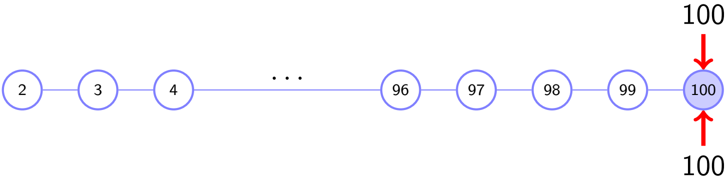
- In a Nash equilibrium, each agent best responds to the others.
- An agent best responds to other agents' actions by choosing a strategy that maximizes utility, conditional on the other agents' strategies.

$$BR_i(s_{-i}) = \arg\max_{s_i} u_i(s_i, s_{-i})$$

Example: Traveler's Dilemma

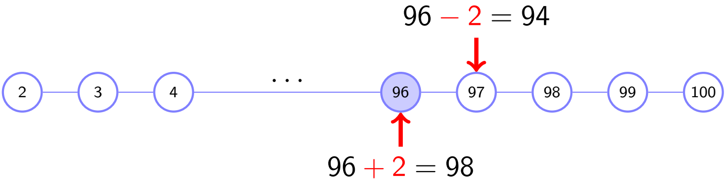
- Two players pick a number (2-100) simultaneously.

Example: Traveler's Dilemma



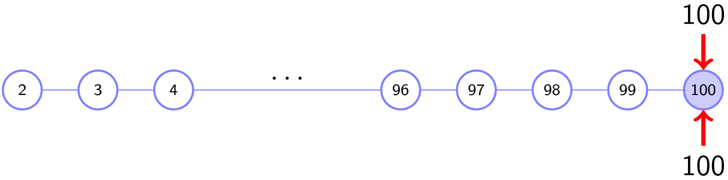
- Two players pick a number (2-100) simultaneously.
- If they pick the same number, that is their payoff.

Example: Traveler's Dilemma



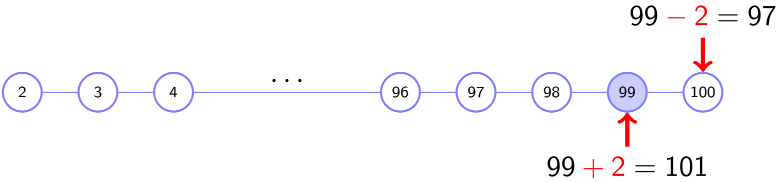
- Two players pick a number (2-100) simultaneously.
- If they pick the same number, that is their payoff.
- If the pick different numbers:
  - Lower player gets lower number, plus bonus of 2.
  - Higher player gets lower number, minus penalty of 2.

Example: Traveler's Dilemma



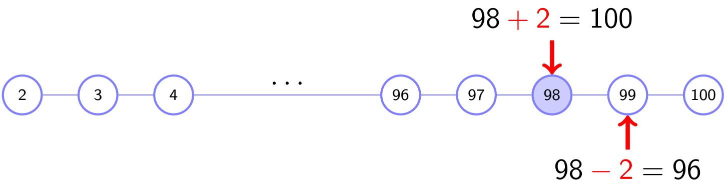
- Two players pick a number (2-100) simultaneously.
- If they pick the same number, that is their payoff.
- If the pick different numbers:
  - Lower player gets lower number, plus bonus of 2.
  - Higher player gets lower number, minus penalty of 2.
- Traveler's Dilemma has a unique Nash equilibrium.

Example: Traveler's Dilemma



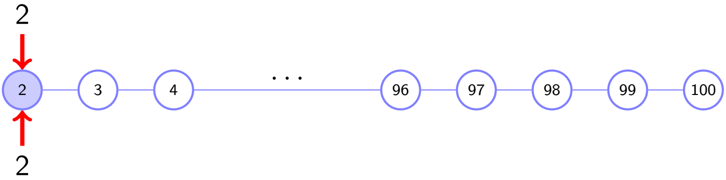
- Two players pick a number (2-100) simultaneously.
- If they pick the same number, that is their payoff.
- If the pick different numbers:
  - Lower player gets lower number, plus bonus of 2.
  - Higher player gets lower number, minus penalty of 2.
- Traveler's Dilemma has a unique Nash equilibrium.

Example: Traveler's Dilemma



- Two players pick a number (2-100) simultaneously.
- If they pick the same number, that is their payoff.
- If the pick different numbers:
  - Lower player gets lower number, plus bonus of 2.
  - Higher player gets lower number, minus penalty of 2.
- Traveler's Dilemma has a unique Nash equilibrium.

Example: Traveler's Dilemma



- Two players pick a number (2-100) simultaneously.
- If they pick the same number, that is their payoff.
- If the pick different numbers:
  - Lower player gets lower number, plus bonus of 2.
  - Higher player gets lower number, minus penalty of 2.
- Traveler's Dilemma has a unique Nash equilibrium.

Nash equilibrium and human subjects

- Nash equilibrium often makes counterintuitive predictions.
  - In Traveler's Dilemma: The vast majority of human players choose 97–100.
- Modifications to a game that don't change Nash equilibrium predictions at all can cause large changes in how human subjects play the game [Goeree & Holt 2001].
  - In Traveler's Dilemma: When the penalty is large, people play much closer to Nash equilibrium.
  - But the size of the penalty does not effect equilibrium.

Nash equilibrium and human subjects

- Nash equilibrium often makes counterintuitive predictions.
  - In Traveler's Dilemma: The vast majority of human players choose 97–100.
- Modifications to a game that don't change Nash equilibrium predictions at all can cause large changes in how human subjects play the game [Goeree & Holt 2001].
  - In Traveler's Dilemma: When the penalty is large, people play much closer to Nash equilibrium.
  - But the size of the penalty does not effect equilibrium.
- Clearly Nash equilibrium is not the whole story.
- Behavioral game theory proposes a number of models to better explain human behavior.

Behavioral game theory models

Themes:

1. Quantal response: Agents best-respond with high probability rather than deterministically best responding.
2. Iterative strategic reasoning: Agents can only perform limited steps of strategic "look-ahead".

One model is based on quantal response, two models are based on iterative strategic reasoning, and one model incorporates both.

BGT model: Quantal response equilibrium (QRE)

QRE model [McKelvey & Palfrey 1995]

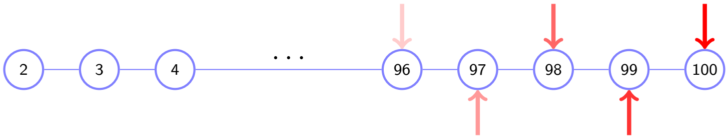
- Agents quantally best respond to each other.

$$QBR_i(s_{-i})(a_i) = \frac{e^{\lambda u_i(a_i, s_{-i})}}{\sum_{a'_i \in A_i} e^{\lambda u_i(a'_i, s_{-i})}}$$

- Precision parameter $\lambda \in [0, \infty)$ indicates how sensitive agents are to utility differences.
  - $\lambda = 0$ means agents choose actions uniformly at random.
  - As $\lambda \to \infty$, QRE approaches Nash equilibrium.

BGT models: Level-k models

- Each agent has one of 3 levels: Level-0, level-1, or level-2.
- Level-0 agents choose uniformly at random.
- Level-1 agents believe that all opponents are level-0.
- Level-2 agents believe that all opponents are level-1.
- Two variants considered:
  - Lk
  - Quantal level-k (QLk)



BGT model: Lk

Lk model [Costa-Gomes et al. 2001]

- Each level-k agent makes a "mistake" with probability ϵ_k, or best responds to level-$(k-1)$ opponent with probability $1-\epsilon_k$.
- Level-k agents aren't aware that level-$(k-1)$ agents will make "mistakes".

$$
\begin{aligned}
IBR_{i,0} &= A_i, \\
IBR_{i,k} &= BR_i(IBR_{-i,k-1}), \\
\pi_{i,0}^{Lk}(a_i) &= |A_i|^{-1}, \\
\pi_{i,k}^{Lk}(a_i) &= \begin{cases} (1-\epsilon_k)/|IBR_{i,k}| & \text{if } a_i \in IBR_{i,k}, \\ \epsilon_k/(|A_i| - |IBR_{i,k}|) & \text{otherwise.} \end{cases}
\end{aligned}
$$

BGT model: QLk

QLk model [Stahl & Wilson 1994]

- Each agent quantally responds to next-lower level.
- Each QLk agent level has its own precision (λ_k), and its own beliefs about lower-level agents' precisions ($\mu_{k,\ell}$).

$$\pi_{i,0}^{QLk}(a_i) = |A_i|^{-1},$$
$$\pi_{i,1}^{QLk} = QBR_i(\pi_{-i,0}^{QLk} \mid \lambda_1),$$
$$\pi_{i,2}^{QLk} = QBR_i(\gamma \mid \lambda_2).$$

BGT model: Cognitive hierarchy

- Each agent has a non-negative level.
- An agent of level m best responds to the truncated, true distribution of levels from 0 to $m-1$.
- Poisson-CH [Camerer et al. 2004]: Levels are assumed to have a Poisson distribution.

$$\pi_{i,0}^{PCH}(a_i) = |A_i|^{-1},$$

$$\pi_{i,m}^{PCH}(a_i) = \begin{cases} |TBR_{i,m}|^{-1} & \text{if } a_i \in TBR_{i,m}, \\ 0 & \text{otherwise.} \end{cases}$$

$$TBR_{i,m} = BR_i\left(\sum_{\ell=0}^{m-1} F(\ell)\pi_{-i,\ell}^{PCH}\right)$$

Prediction using Nash equilibrium

- We would like to compare BGT models' prediction performance to Nash equilibrium.
- Unmodified Nash equilibrium is not suitable for predictions:
    1. Games often have multiple Nash equilibria.
    2. A Nash equilibrium will often assign probability 0 to some actions.

Prediction using Nash equilibrium

- We would like to compare BGT models' prediction performance to Nash equilibrium.
- Unmodified Nash equilibrium is not suitable for predictions:
  1. Games often have multiple Nash equilibria.
  2. A Nash equilibrium will often assign probability 0 to some actions.
- We constructed two different Nash-based models to deal with multiple equilibria:
  - UNEE: Take the average of all Nash equilibria.
  - NNEE: Predict using the post-hoc "best" Nash equilibrium.
- Both models avoid probability 0 predictions via a tunable error probability.

Experimental setup: Overview

What do we need to compare predictive models?

What do we need to compare predictive models?

1. Evaluation criteria
    - Metric to measure performance
    - Statistical test to evaluate significance

Experimental setup: Overview

What do we need to compare predictive models?

1. Evaluation criteria
   - Metric to measure performance
   - Statistical test to evaluate significance
2. Experimental data
   - Training data to fit model parameters
   - Test data to evaluate models on

1. Evaluation criteria: Metric

- We score the performance of a model by the likelihood of the test data:

$$\mathbf{P}(\mathcal{D}_{test} \mid \mathcal{M}, \vec{\theta}^*).$$

1. Evaluation criteria: Metric

- We score the performance of a model by the likelihood of the test data:

$$\mathbf{P}(\mathcal{D}_{test} \mid \mathcal{M}, \vec{\theta}^{\,*}).$$

- The parameters are chosen to maximize the likelihood of the training data:

$$\vec{\theta}^{\,*} = \arg\max_{\vec{\theta}} \mathbf{P}(\mathcal{D}_{train} \mid \mathcal{M}, \vec{\theta}).$$

1. Evaluation criteria: Statistical test

- For lower-variance estimate of performance, we use 10-fold cross-validation.

1. Evaluation criteria: Statistical test

- For lower-variance estimate of performance, we use 10-fold cross-validation.
- Problem: Results may depend upon the particular partition into folds.

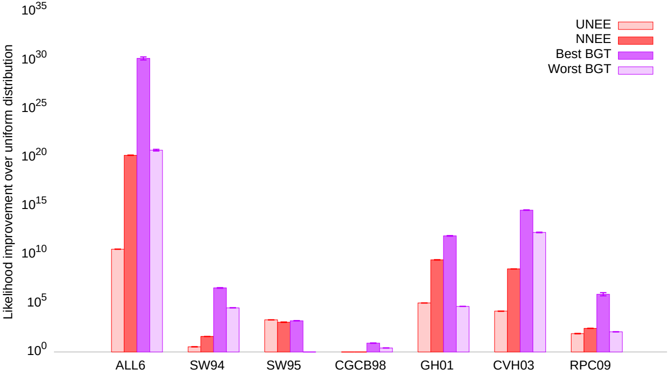
1. Evaluation criteria: Statistical test

- For lower-variance estimate of performance, we use 10-fold cross-validation.
- Problem: Results may depend upon the particular partition into folds.
- We average over multiple cross-validation runs.
- We can then compute 95% confidence interval by assuming a t-distribution of these averages [Witten & Frank 2000].

2. Experimental data

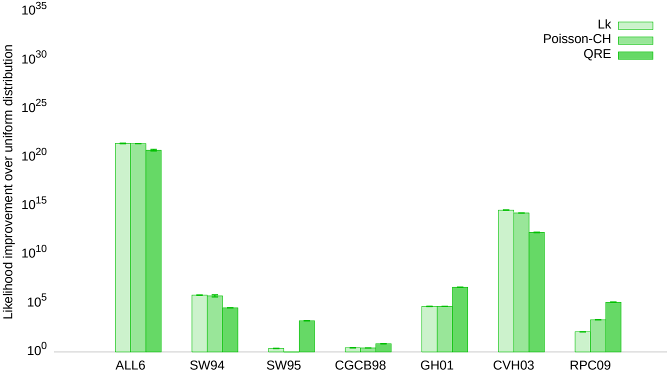
- Data from six experimental studies, plus a combined dataset:
  - `SW94`: 400 observations from [Stahl & Wilson 1994]
  - `SW95`: 576 observations from [Stahl & Wilson 1995]
  - `CGCB98`: 1296 observations from [Costa-Gomes et al. 1998]
  - `GH01`: 500 observations from [Goeree & Holt 2001]
  - `CVH03`: 2992 observations from [Cooper & Van Huyck 2003]
  - `RPC09`: 1210 observations from [Rogers et al. 2009]
  - `ALL6`: All 6974 observations
- Subjects played 2-player normal form games once each.
- Each action by an individual player is a single observation.

Model comparisons: Nash equilibrium vs. BGT



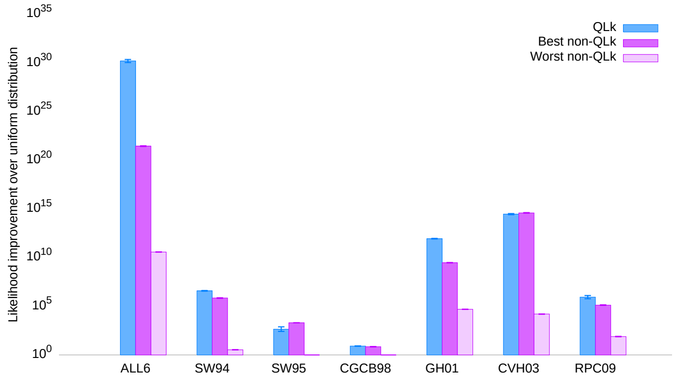
- UNEE worse than every BGT model (except `GH01` and `SW95`).
- Even NNEE worse than QLk and QRE in most datasets.
- BGT models typically predict human behavior better than Nash equilibrium-based models.

Model comparisons: Lk and CH vs. QRE



- Lk and Poisson-CH performance roughly similar.
- No ordering between Lk/Poisson-CH and QRE.
- Iterative models and quantal response appear to capture distinct phenomena.

Model comparisons: QLk



- We would expect a model with both iterative and quantal response components to perform best.
- That is the case: QLk is the best predictive model on almost every dataset.

Deeper analysis

1. Is the Poisson distribution helpful in cognitive hierarchy?
2. Are higher-level agents helpful in level-k?
3. Does payoff scaling matter in QRE?
4. Is heterogeneity necessary in QLk?

Deeper analysis

1. Is the Poisson distribution helpful in cognitive hierarchy?
2. Are higher-level agents helpful in level-k?
3. Does payoff scaling matter in QRE?
4. Is heterogeneity necessary in QLk?

Heterogeneity

In QLk, different agent levels:

- have different precisions (λ_k).
- have different beliefs about the relative proportions of other levels.
    - Level-k believes that 100% of the population is level-$(k-1)$.
- have different beliefs about the precisions of other levels ($\mu_{k,\ell}$).

$$\pi_{i,0}^{QLk}(a_i) = |A_i|^{-1},$$
$$\pi_{i,1}^{QLk} = QBR_i(\pi_{-i,0}^{QLk} \mid \lambda_1),$$
$$\pi_{i,2}^{QLk} = QBR_i(\gamma \mid \lambda_2).$$

Simplified hybrid model

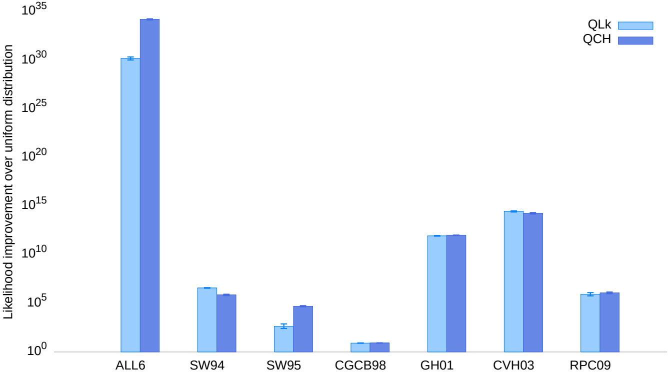
Question 4: Is heterogeneity necessary in QLk?

- Combine quantal response of QLk with truncated true beliefs of cognitive hierarchy
- In quantal cognitive hierarchy model (QCH), all agent levels:
  - respond quantally (as in QLk).
  - respond to truncated, true distribution of lower levels (as in cognitive hierarchy).
  - have the same precision λ.
  - are aware of the true precision of lower levels.

$$\pi_{i,0}^{QCH}(a_i) = |A_i|^{-1}$$

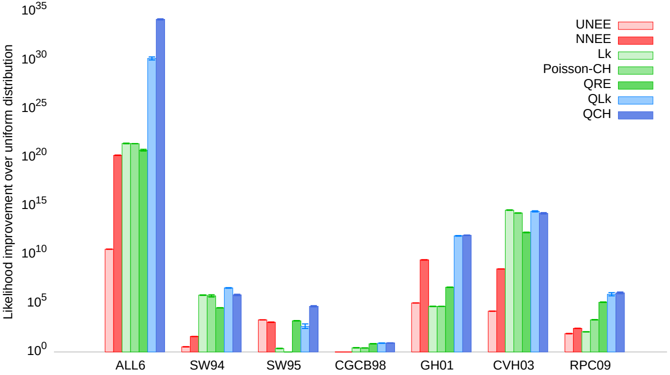
$$\pi_{i,m}^{QCH}(a_i) = QBR_i\left(\sum_{\ell=0}^{m-1} \alpha_\ell \pi_{j,\ell}^{QCH} \mid \lambda\right)$$

QCH vs. QLk



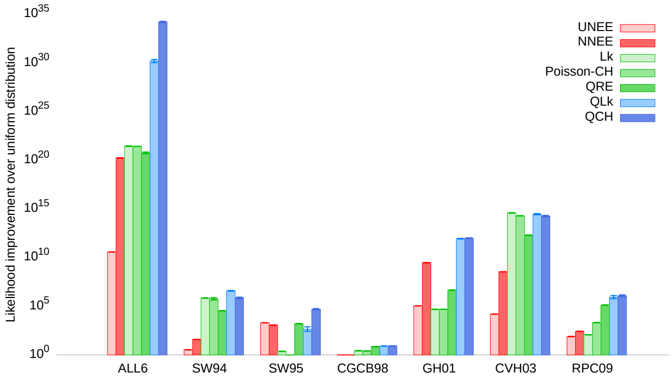
- QCH predicts somewhat better than QLk on most datasets, including the combined dataset.
- A less heterogeneous model has roughly the same predictive power as QLk.

Summary



- We compared predictive performance of four BGT models.
- BGT models typically predict human behavior better than Nash equilibrium-based models.
- Recommended specific models: QLk or QCH.

Thank you!



- We compared predictive performance of four BGT models.
- BGT models typically predict human behavior better than Nash equilibrium-based models.
- Recommended specific models: QLk or QCH.

Previous work

Paper	Nash	QLk	Lk	CH	QRE
[Stahl and Wilson, 1994]	t	t			
[McKelvey and Palfrey, 1995]	f				f
[Stahl and Wilson, 1995]	f	t			
[Costa-Gomes et al., 1998]	f		f		
[Haruvy et al., 1999]		t			
[Costa-Gomes et al., 2001]	f		f		
[Haruvy et al., 2001]		t			
[Morgan and Sefton, 2002]	f				p
[Weizsäcker, 2003]	t				t
[Camerer et al., 2004]	f			p	
[Costa-Gomes and Crawford, 2006]	f		f		
[Stahl and Haruvy, 2008]		t			
[Rey-Biel, 2009]	t		t		
[Georganas et al., 2010]	f		f		
[Hahn et al., 2010]				p	
[Camerer et al., 2001]				f	f
[Chong et al., 2005]	f			p	p
[Crawford and Iriberri, 2007]	p		p		p
[Costa-Gomes et al., 2009]	f		f	f	f
[Rogers et al., 2009]	f			f	f

A 'p' indicates that the study evaluated out-of-sample prediction performance for that model; a 't' indicates statistical tests of training sample performance; an 'f' indicates comparison of training sample fit only.
Only five studies compared more than one of the non-Nash models we considered.

Bibliography

Camerer, C., Ho, T., and Chong, J. (2001).
Behavioral game theory: Thinking, learning, and teaching.
Nobel Symposium on Behavioral and Experimental Economics.

Camerer, C., Ho, T., and Chong, J. (2004).
A cognitive hierarchy model of games.
QJE, 119(3):861–898.

Chong, J., Camerer, C., and Ho, T. (2005).
Cognitive hierarchy: A limited thinking theory in games.
Experimental Business Research, Vol. III: Marketing, accounting and cognitive perspectives, pages 203–228.

Costa-Gomes, M. and Crawford, V. (2006).
Cognition and behavior in two-person guessing games: An experimental study.
AER, 96(5):1737–1768.

Costa-Gomes, M., Crawford, V., and Broseta, B. (1998).