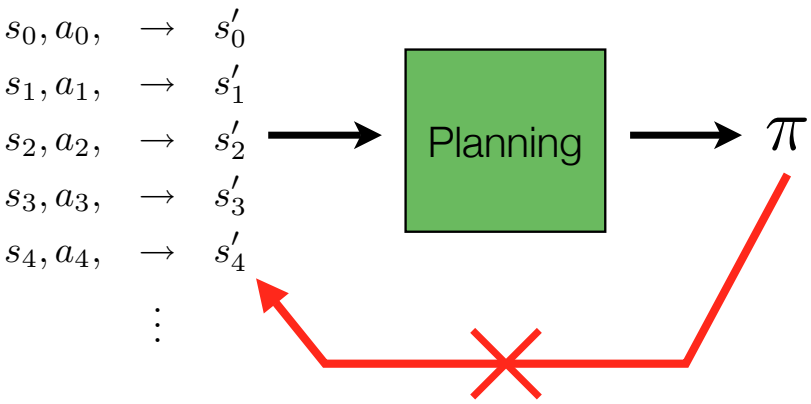
Planning with an Untrustworthy Model

Michael Bowling
University of Alberta

If you had a model, what should you do?

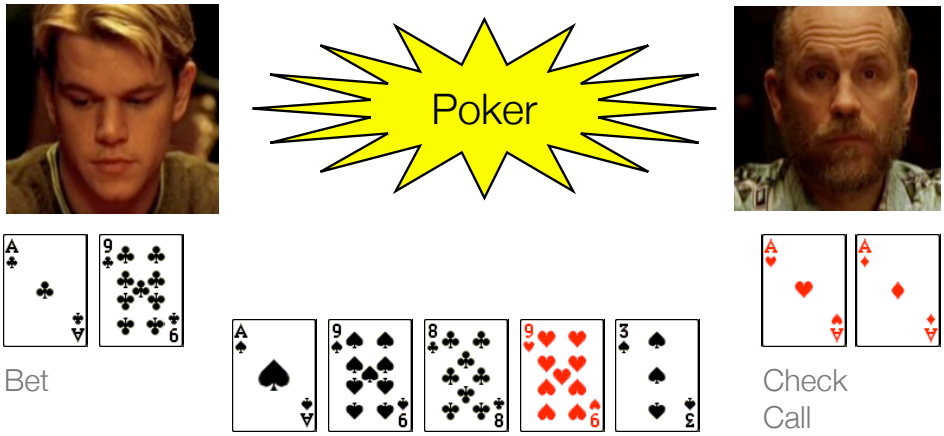
- But, what if...
  - You have to perform well in the world, not with respect to the model
  - The model was constructed from very little experience
  - The data-gathering policy was not deliberately exploring
  - The world could change

The Setup



"Mistake Already Made"

Who am I kidding?

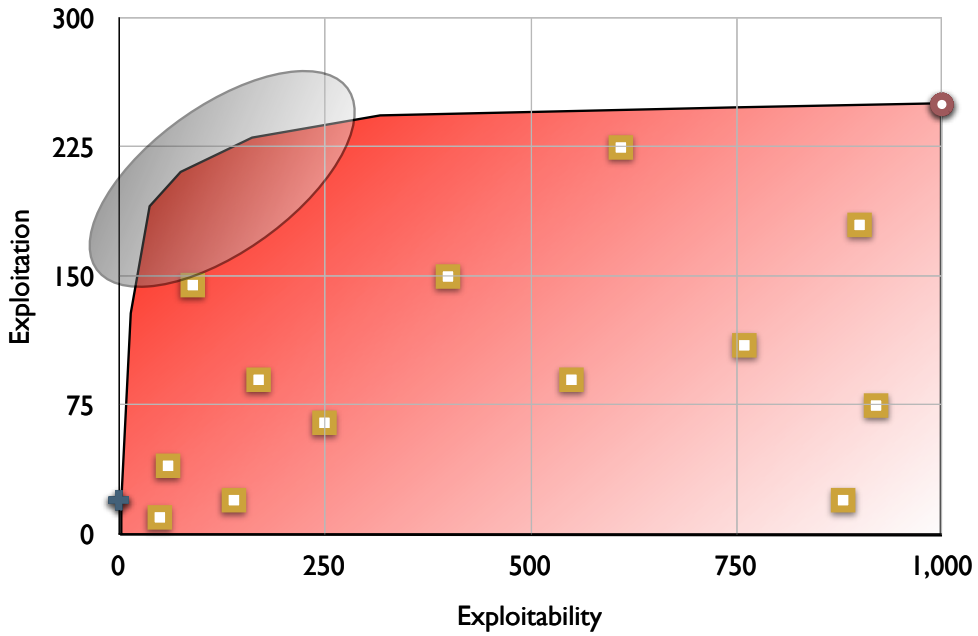


Opponent Modelling in Poker

- Given data from an opponent playing poker, we can **construct a model** and find an **optimal response to the model**
- But...
  - Need to perform well against the opponent not the model
  - Likely to have very little data
  - Data didn't involve deliberate exploration of the opponent's strategy
  - Opponent's strategy can change

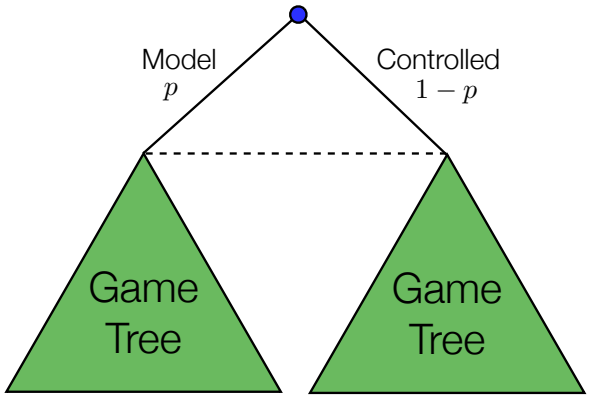
Restricted Nash Response

Johanson, Zinkevich & Bowling (NIPS, 2007)



Restricted Nash Response

Johanson, Zinkevich & Bowling (NIPS, 2007)



Opponent Modelling in Poker

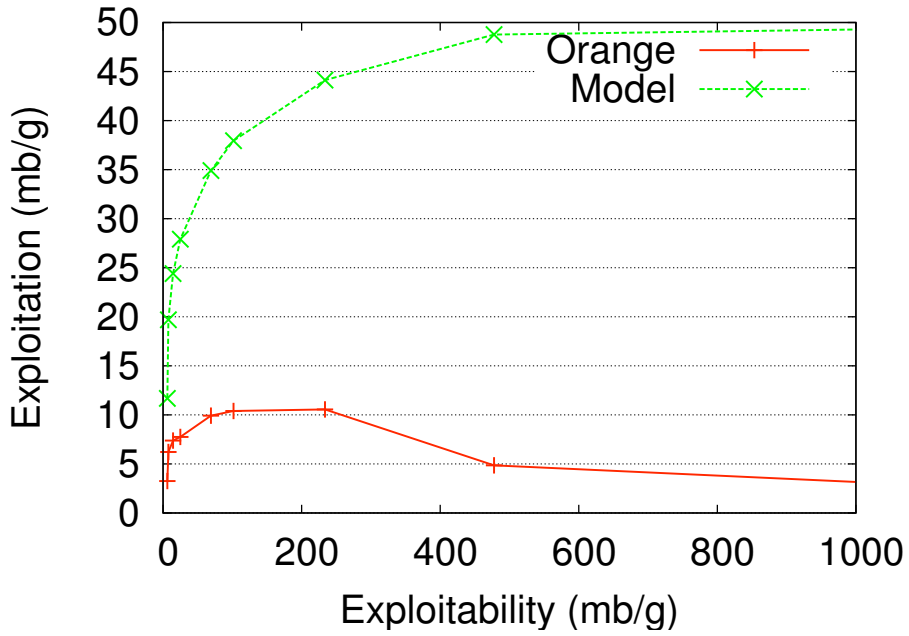
- Given data from an opponent playing poker, we can **construct a model** and find an **optimal response to the model**

- But...

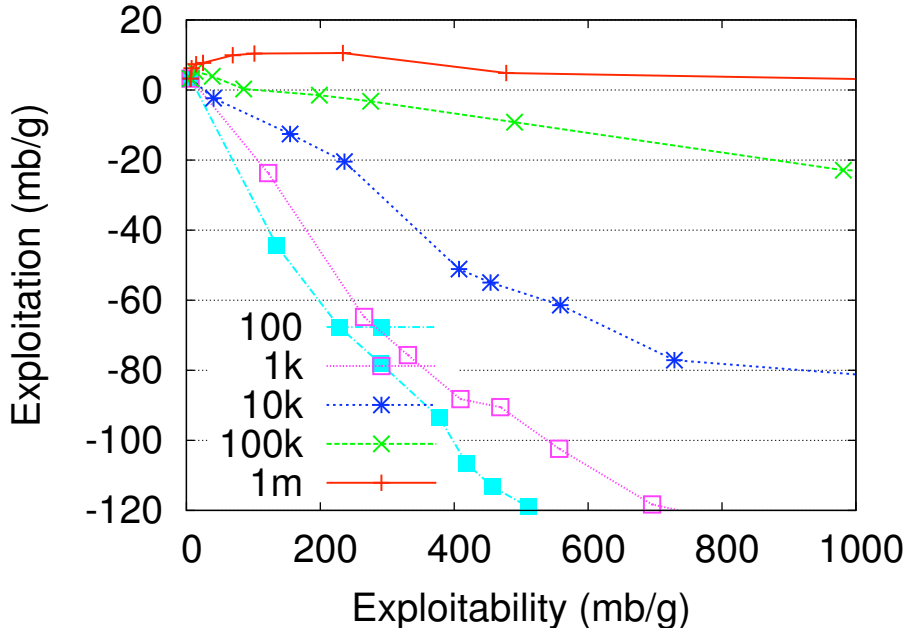
  - Need to perform well against the opponent not the model
  - Likely to have very little data
  - Data didn't involve deliberate exploration of the opponent's strategy

  - Opponent's strategy can change

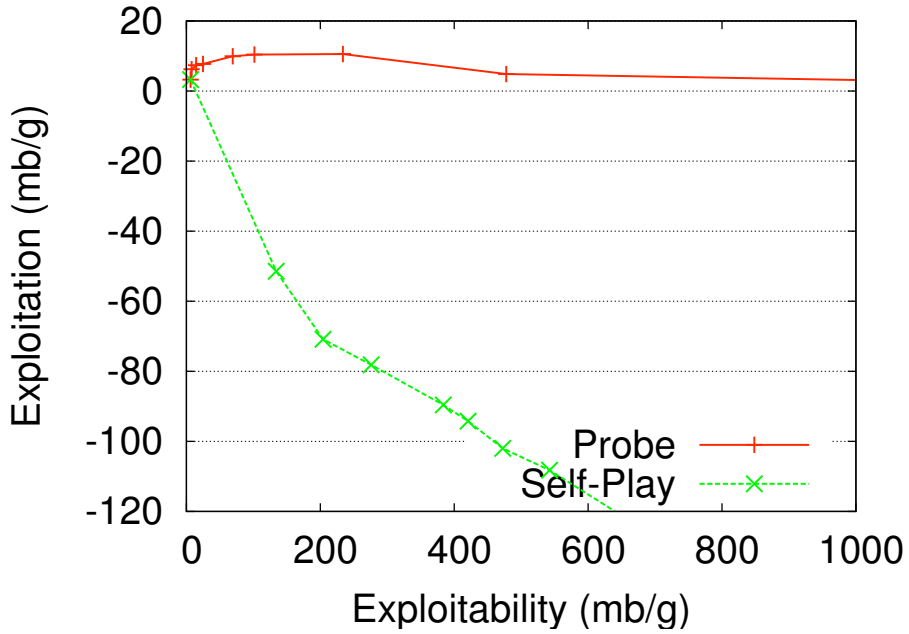
Model vs. Real Performance



Lots vs. Little Data

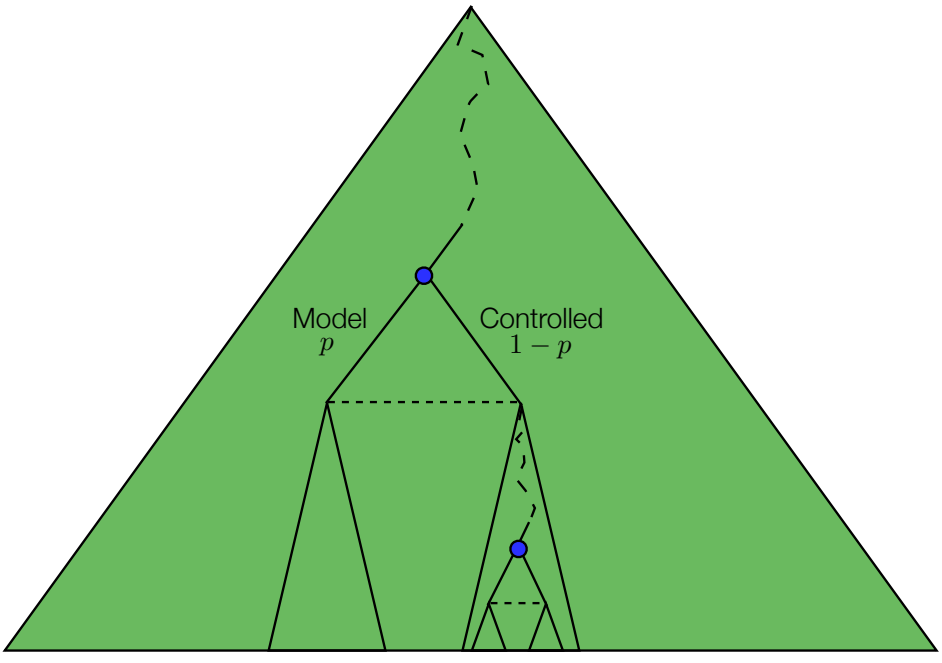


Good vs. Bad Exploration

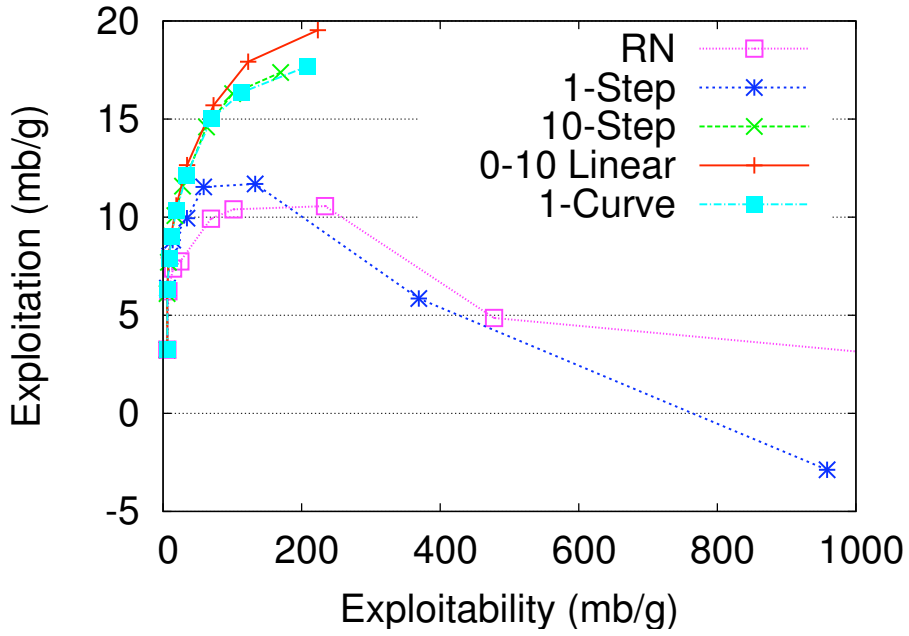


Data Biased Response

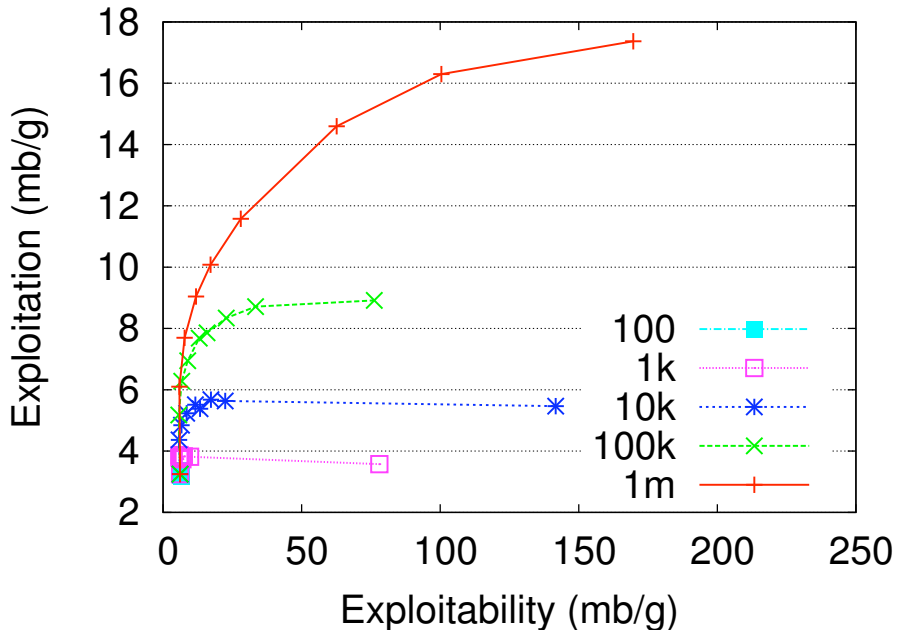
Johanson & Bowling (AISTATS, 2009)



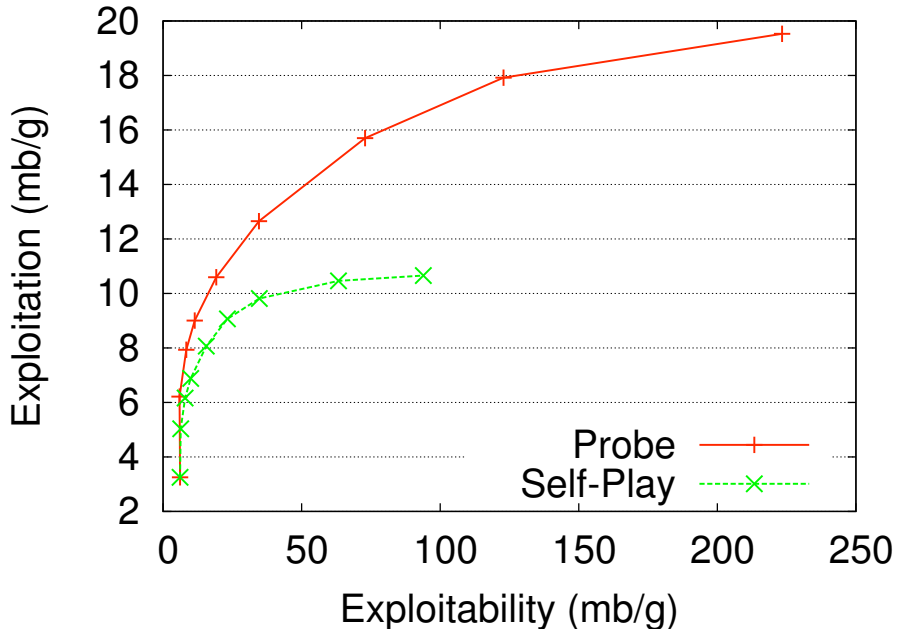
Model vs. Real Performance (Revisited)



Lots vs. Little Data (Revisited)



Good vs. Bad Exploration (Revisited)



It Really Works

- Benchmarks
  - “Always Fold” loses 750 (mb/g)
  - “Pro” players typically win 50
- Amateur players lose ~110 to our “best” programs
- DBR using 1 million hands of amateur data beats amateurs by ~350!

A Bajan Reinterpretation

- A “Proper” Bajan:

$$\underset{\pi}{\operatorname{argmax}} \int_{m \in M} \overbrace{\Pr(m|\mathcal{Z})}^{\text{Likelihood}} \overbrace{f(m)}^{\text{Prior}} \operatorname{Value}_m(\pi)$$

- Robustness to the Prior:

$$\underset{\pi}{\operatorname{argmax}} \overbrace{\min_{f \in F}}^{\text{Robust}} \int_{m \in M} \Pr(m|\mathcal{Z})\, f(m)\, \operatorname{Value}_m(\pi)$$

  - If F contains all possible priors... maximizes worst-case performance
  - If F contains a single prior... Bayes-optimal
  - If F is a particular form of the prior (independent Dirichlets with fixed strength)... data-biased response

Can we use this in RL?

- Ummm... sure
- Discrete state MDPs?
  - Reduces to a 2-player zero-sum stochastic game (“easy”)
  - Not implemented yet
  - How does this relate to Xu and Mannor’s RP Tradeoff?
- MDPs with approximation?
  - Talk to me and help me figure it out

Questions?