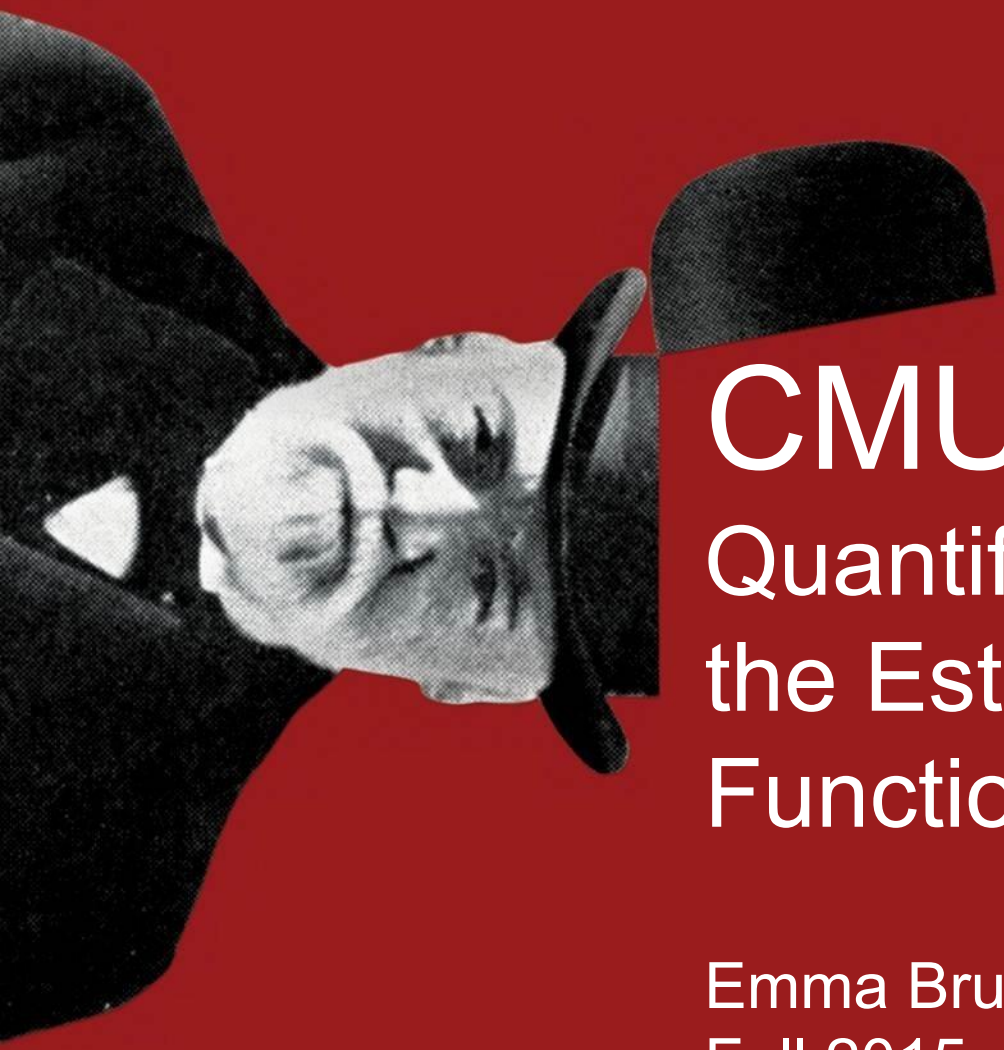
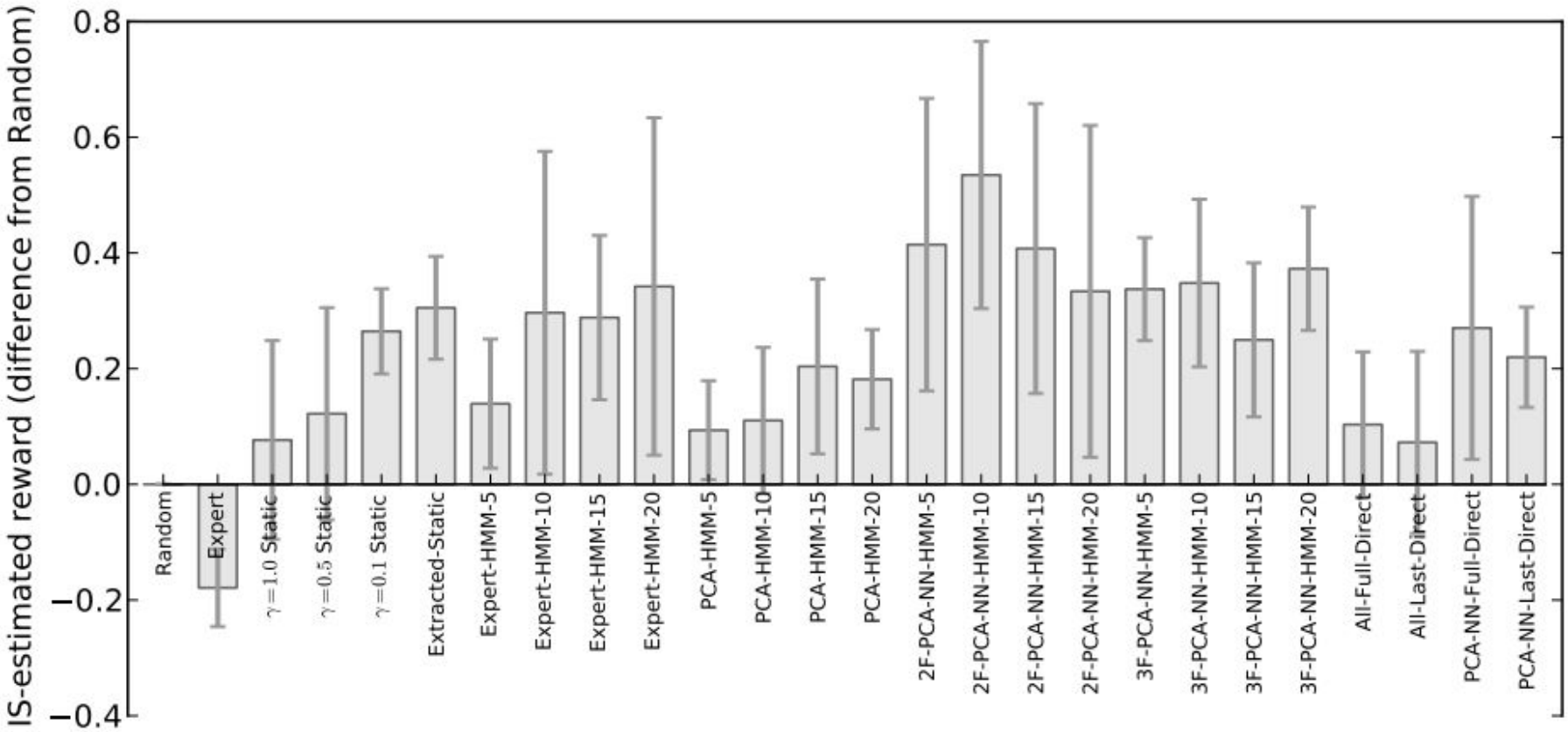


CMU 15-889e

Quantifying the Quality of the Estimated Value Function or Policy

Emma Brunskill

Fall 2015



from Manel, Liu, Levine, Brunskill & Popovic AAMAS 2013

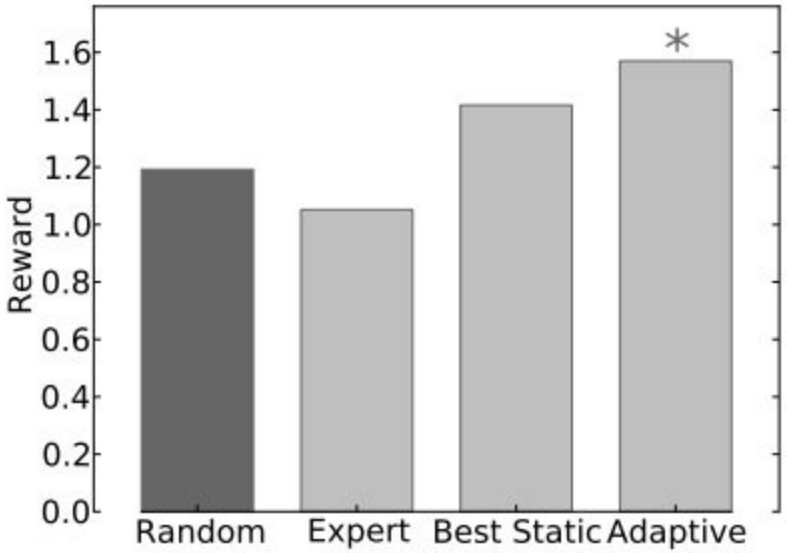
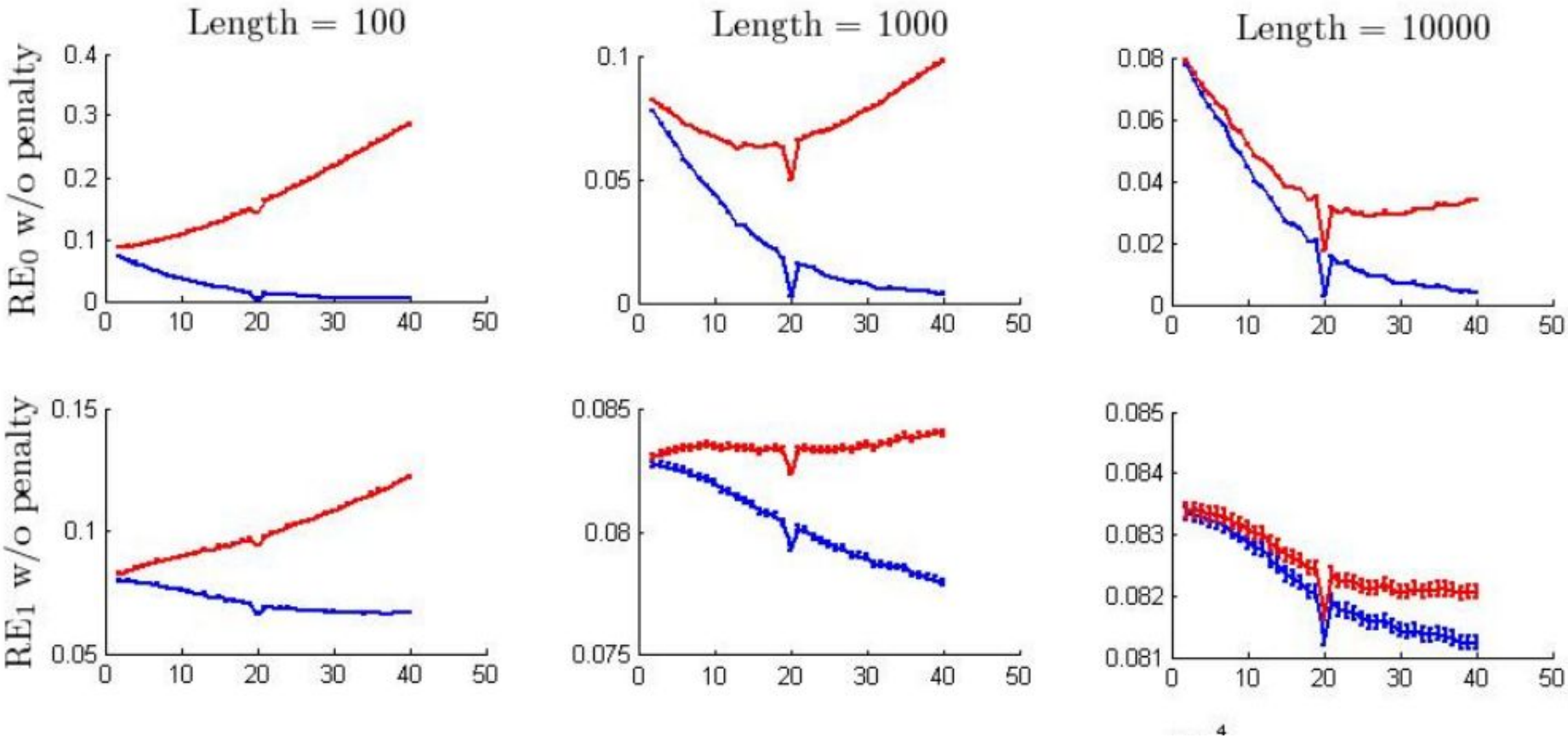
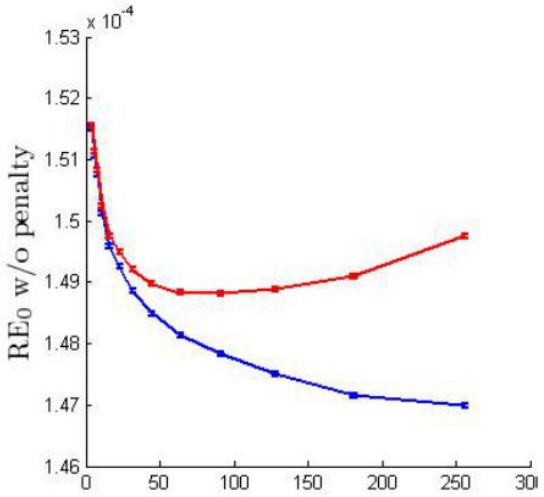


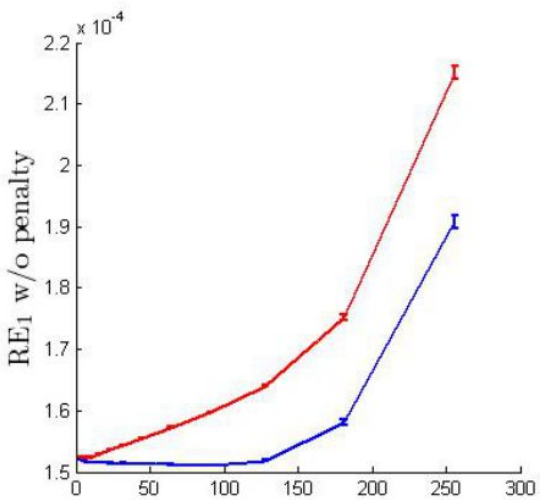
Figure 5: Experimental results on N=2000 players distributed equally among the four conditions. Adaptive received the highest average reward, and was the only policy that differed significantly from Random (p=0.002), as indicated by the asterisk.

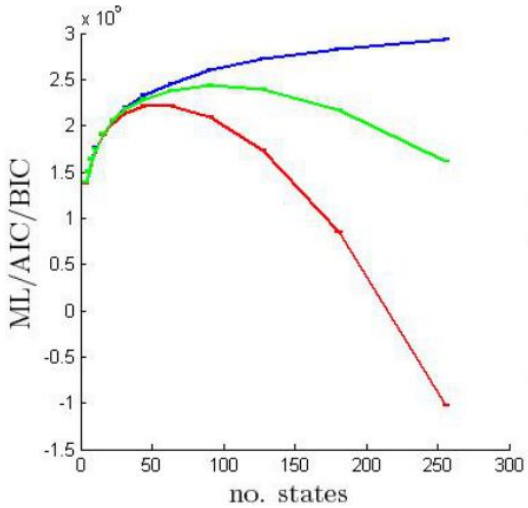
from Manel, Liu, Levine, Brunskill & Popovic AAMAS 2013

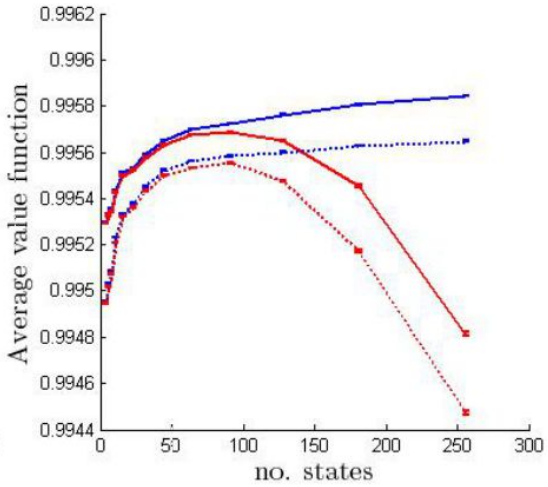


from Hallak, Di-Castro, Mannor
KDD 2013









from Hallak, Di-Castro, Mannor
KDD 2013

Quantifying the Quality of the Estimated V/Q or Policy: 3 Important Lines of Work

1. Bound estimation error
2. Direct unbiased estimate of future performance
3. Selecting among classes of models/ approximation classes / policies
    - Problem with model prediction error
    - Problem with value function
    - Options? IS, reward prediction errror...

Midterm Review: Topics So Far

- Background on RL
- Batch RL in large domains techniques
  - FVI
  - LSPI
  - Approx model-based learning, MCTS approaches
- Choosing/Creating features for use in batch RL
  - L1/L2
  - OMP
  - BatchIFDD+
- Evaluating quality of result from batch RL
  - Bound estimation error for a model
  - Direct unbiased estimate of future performance (using importance sampling)
  - Selecting among models

Learning Objectives for Class so Far

- What's a learning objective?
- This is my expectation of what you should be able to do after completing this part of the class

.

Learning Objectives for Class so Far

- Define MDP (standard definition and to take a problem and define as a MDP)
- Execute value iteration and policy iteration on tabular MDP to compute a policy

Learning Objectives for Class so Far

- Define MDP (standard definition and to take a problem and define as a MDP)
- Execute value iteration and policy iteration on tabular MDP to compute a policy
- Implement an algorithm that can approximately compute a policy given batch data from a large domain. Explain what is the objective it is trying to minimize/maximize. Understand convergence guarantees.

Learning Objectives for Class so Far

- Define MDP (standard definition and to take a problem and define as a MDP)
- Execute value iteration and policy iteration on tabular MDP to compute a policy
- Implement an algorithm that can approximately compute a policy given batch data from a large domain. Explain what is the objective it is trying to minimize/maximize. Understand convergence guarantees.
- Implement a feature selection method for use in an policy evaluation method.
- Compare the benefits and drawbacks of >=2 feature selection algorithms

Learning Objectives for Class so Far

- Define MDP (standard definition and to take a problem and define as a MDP)
- Execute value iteration and policy iteration on tabular MDP to compute a policy
- Implement an algorithm that can approximately compute a policy given batch data from a large domain. Explain what is the objective it is trying to minimize/maximize. Understand convergence guarantees.
- Implement a feature selection method for use in an policy evaluation method.
- Compare the benefits and drawbacks of >=2 feature selection algorithms
- Describe, calculate & bound different forms of error that can contribute to mismatches in the expected and actual performance of a policy. Understand the benefits of making some of these errors

Learning Objectives for Class so Far

- Define MDP (standard definition and to take a problem and define as a MDP)
- Execute value iteration and policy iteration on tabular MDP to compute a policy
- Implement an algorithm that can approximately compute a policy given batch data from a large domain. Explain what is the objective it is trying to minimize/maximize. Understand convergence guarantees.
- Implement a feature selection method for use in an policy evaluation method.
- Compare the benefits and drawbacks of >=2 feature selection algorithms
- Describe, calculate & bound different forms of error that can contribute to mismatches in the expected and actual performance of a policy. Understand the benefits of making some of these errors
- Describe and characterize properties of at least two methods for evaluating the quality of a policy computed using a batch RL algorithm. Be able to provide example applications where you would prefer one or the other, and support your argument.

Midterm

- Single-sided sheet of notes, no laptops, no calculators
- Other tips: look back at summary slides or “key points” from past lectures
- Prior midterms which contains some standard RL material (if unsure, just send a note to Piazza)
- https://www.cs.cmu.edu/~ggordon/780-fall07/fall06/docs/15780f06-midterm-sol.pdf
- https://www.cs.cmu.edu/~ggordon/780-fall07/docs/15780f07-midterm-sol.pdf
- http://www.cs.cmu.edu/~guestrin/Class/10701-S07/ [ML but has some stuff on RL]
- http://www.cs.cmu.edu/afs/cs/academic/class/15780-s13/www/lec/gradai.midterm.2013.solutions.pdf
- We will also post a few extra pointers and/or exercises for practice

Projects!

- 1 page proposal due October 23 by 5pm.
- Send to me and Christoph
- Encourage you to work in pairs, but if a few people want to work solo (or in groups of 3), that may be fine-- come talk to us.
- Proposal should cover
  - Team name
  - Short description of project
  - Plan for accomplishing
- We will meet with all the teams after reading the proposals
- We welcome your ideas! Tying it to your research is highly encouraged
- Also a list of potential project ideas will be posted to piazza