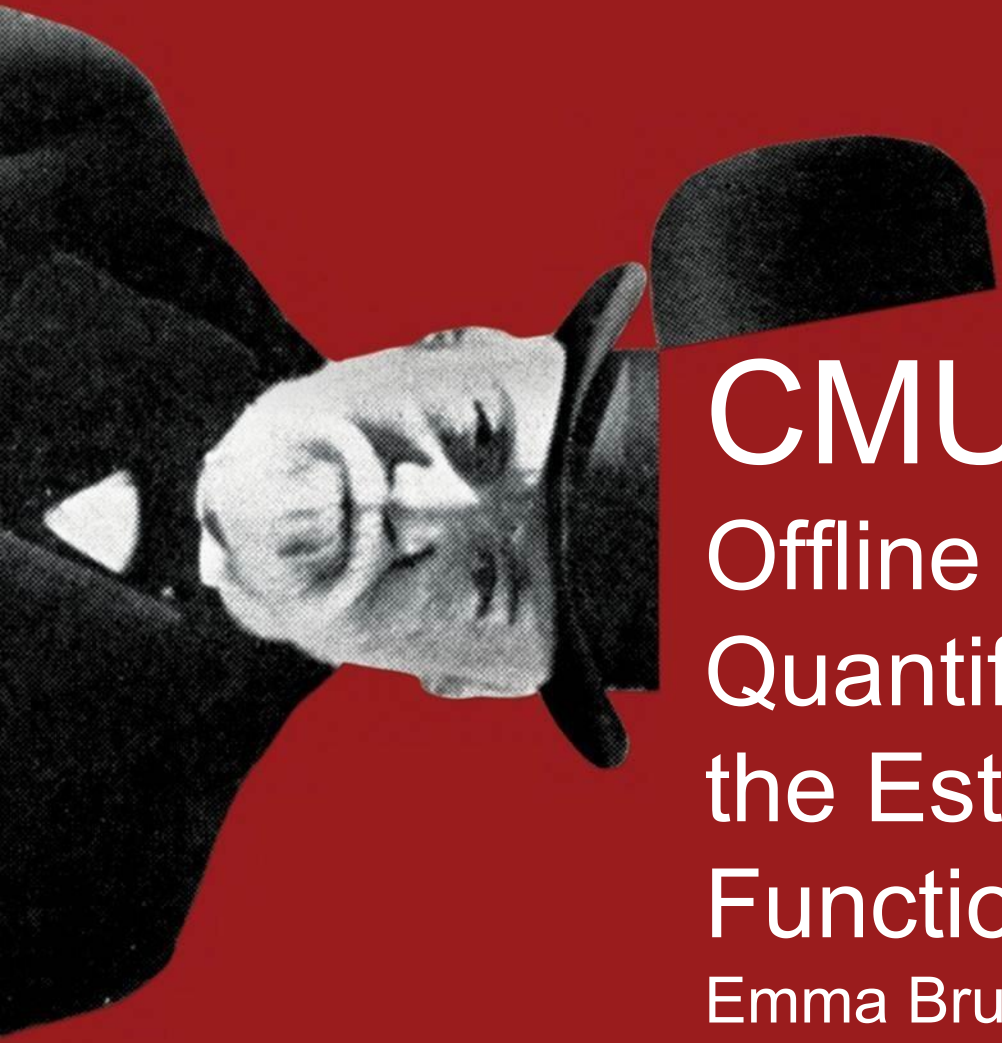


CMU 15-889e

Offline Batch RL: Quantifying the Quality of the Estimated Value Function or Policy

Emma Brunskill

Fall 2015

Setting: From Offline Data to Good Future Decisions

- Have a set of input data
  - Set of trajectories or (s,a,r,s’) tuples
- Want to determine a good policy for future use
- Challenges include:
  - State/action space may be large/infinite
  - Data sampled from one distribution and often want to estimate other policies that would induce different distributions

From Offline Data to Good Future Decisions, so far

- Feature-based batch approximate RL algorithms
  - Fitted value / q iteration
  - Least squares policy iteration
- Techniques for choosing features
  - Take a large pool of features and select a few using L1/L2/OMP methods
  - Start with a small set of features and generate additional features as needed (e.g. BatchIFDD+)

From Offline Data to Good Future Decisions, next

- Feature-based batch approximate RL algorithms
- Techniques for choosing features
- But how good is our solution?
  - **Evaluating the quality of the resulting V/Q/□**

Review from Intro Lecture: Desirable Properties for a RL Algorithm

Convergence ← Discussed for FVI/LSPI

Consistency

Small generalization error

Small estimation error

Small approximation error

High learning speed

Safety

Offline Batch RL:

Desired Properties of Algorithm & Output Policy

Convergence ← Discussed for FVI/LSPI

Consistency

Small generalization error

Small estimation error

Small approximation error

High learning speed

Safety

Quantifying the Quality of the Estimated V/Q or Policy

- Very important, interesting & still open research area!
- We will cover some of the key ideas

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

1. Bound estimation error
   - Focus is on error due to finite samples
   - Does not address approximation error
   - If know model class, or have an extremely expressive model class, estimation error may dominate generalization error

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

- No direct separation of estimation and approximation error
- Can be very high variance

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

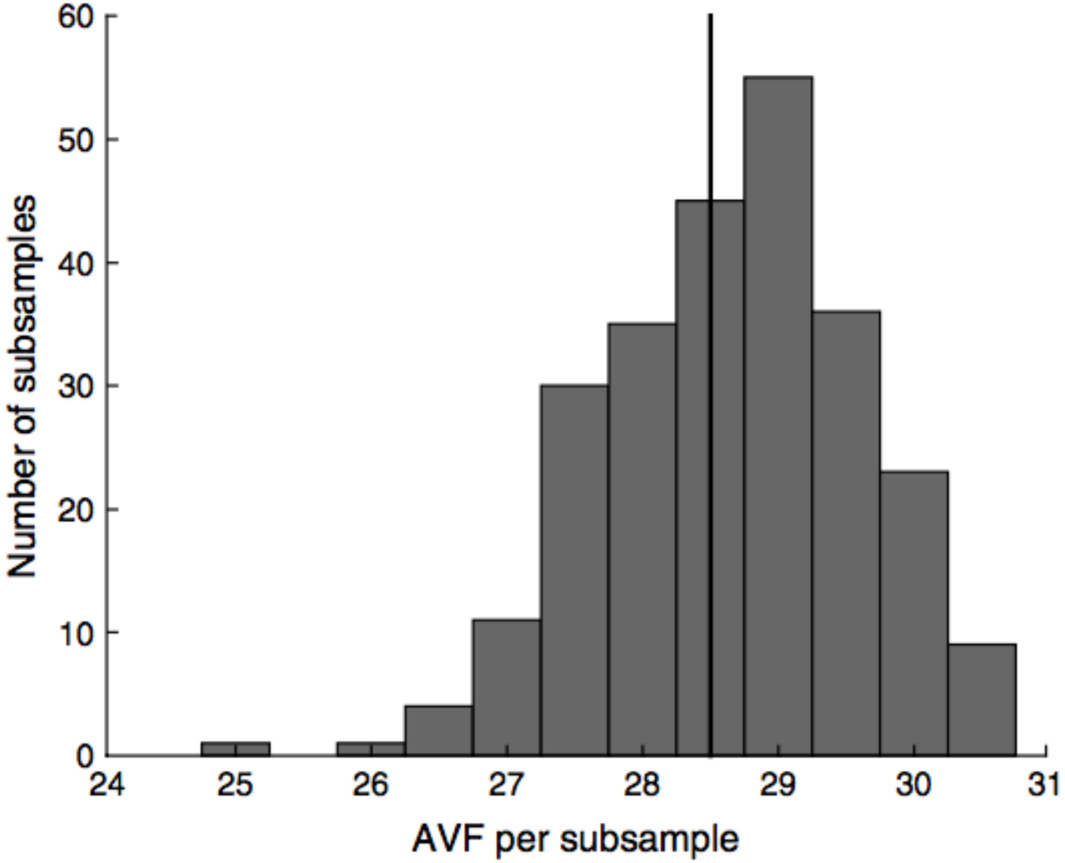
- Use generalization performance to pick
- May not know generalization performance of selected choice, just that it is best of set

Quantifying the Quality of the Estimated V/Q or Policy: Today

1. **Bound estimation error**
2. Direct unbiased estimate of future performance
3. Selecting among classes of models/ approximation classes / policies

Impact of Finite Sample / Estimation Error

Figure 1 **Mail Catalog Problem: A Histogram of the AVF of the Historical Policy for a Partition of the Customers to 250 Subsamples**



Note. The discount factor per period is $\alpha = 0.98$. The policy used is the historical (mixed) policy used by the firm, and the value function is weighted uniformly across states. The AVF obtained from the full data is \$28.54, and is plotted as a vertical line. The empirical standard deviation is \$0.97.

Mannor et al.: Bias and Variance Approximation in Value Function Estimates 312 Management Science. 2007

Bounding Estimation Error: Key Points

- Be able to reproduce simulation lemma proof
- Understand key idea about bias and variance for a single policy (how to calculate, what’s being approximated)
- Understand problem for control setting and one way to get around
- Know that these approaches ignore approximation error