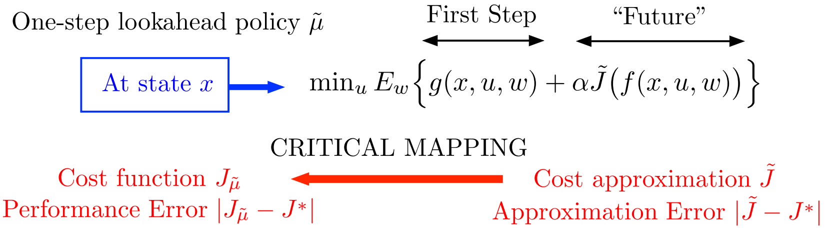
John Tsitsiklis Celebration Event
Panel Discussion on RL
October 7, 2023

Dimitri Bertsekas
Arizona State University

How do mainstream theory and RL practice connect?
I will argue NOT WELL

On-Line Approximation in Value Space (Model-Based)
System equation: $f(x,u,w)$, Cost per stage: $g(x,u,w)$, α-Discounted



One-step lookahead policy $\tilde{\mu}$

First Step

“Future”

At state x

$$\min_u E_w\Big\{g(x,u,w)+\alpha\tilde{J}\big(f(x,u,w)\big)\Big\}$$

CRITICAL MAPPING

Cost function $J_{\tilde{\mu}}$

Performance Error $|J_{\tilde{\mu}}-J^*|$

Cost approximation $\tilde{J}$

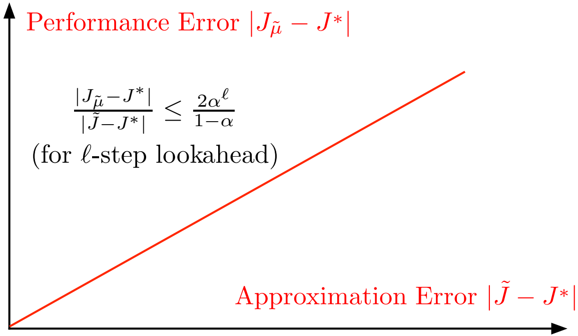
Approximation Error $|\tilde{J}-J^*|$

- Replace optimal cost J^* with an approximation $\tilde{J}$ in Bellman's equation
- Defines a lookahead policy $\tilde{\mu}$ with $\tilde{\mu}(x)$ being the minimizing u above

KEY QUESTIONS

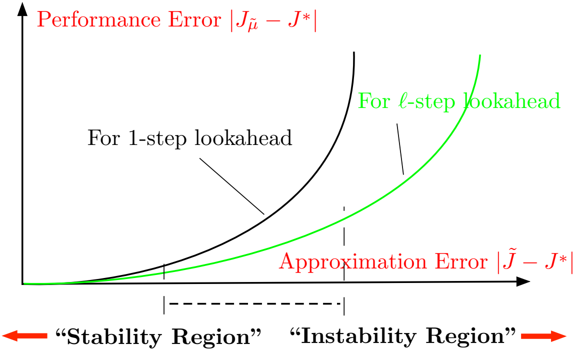
- What is the relation between $J_{\tilde{\mu}}$ and $\tilde{J}$?
- How does multistep lookahead affect this relation?

The Linear Error Bound Model: An Example of Bad Theory



- These bounds are well-known to be conservative
- ... but they are broadly thought to be “qualitatively" correct
- THE REALITY IS FAR DIFFERENT
- The bounds are not only unrealistic, they are misleading
- They misdirect theoretical research and confuse the practitioners

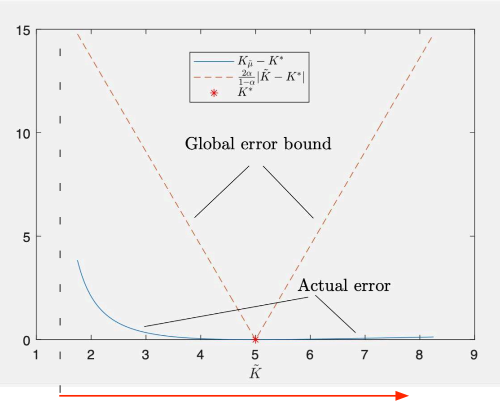
The Real Relation is Superlinear



A key fact: The critical mapping is a Newton Step for solving the Bellman equation (Newton/SOR for multistep lookahead)

- Far-reaching implications for both theory and practice
- Convergence threshold defined by the region of convergence of Newton's method
- Inside the two regions, better training/more data, improving confidence intervals has marginal effect
- There is a critical stability threshold (for undiscounted problems)

An α-Discounted Linear Quadratic Example



Region of stability

- One-step lookahead
- One-dimensional problem - unstable system - undiscounted
- $J^*(x) = K^* x^2, \quad \tilde{J}(x) = \tilde{K} x^2, \quad J_{\tilde{\mu}}(x) = K_{\tilde{\mu}} x^2$
- Details in my Lessons from AlphaZero book (2022)

A Computational Study (Laidlaw, Russell, Dragan, 2023)

Extensive tests using a dataset of 155 MDPs and “current" methods. Quotes:

- “There is a large gap between the current theory and practice of RL"
- “Deep RL works impressively in some environments and fails catastrophically in others"
- “Current theory does not quite have the ability to predict this"
- “We find that prior bounds do not correlate well with when deep RL succeeds vs. fails"

Among their empirical findings:

- An important mechanism to make methods “work" is to increase the lookahead, NOT do more sampling, explore better, etc, to improve $\tilde{J}$
- With long enough lookahead, an exactly optimal policy is obtained (a theoretical fact known since the 60s)