RL: Lecture 19

Harvey Mudd College

April 8, 2020

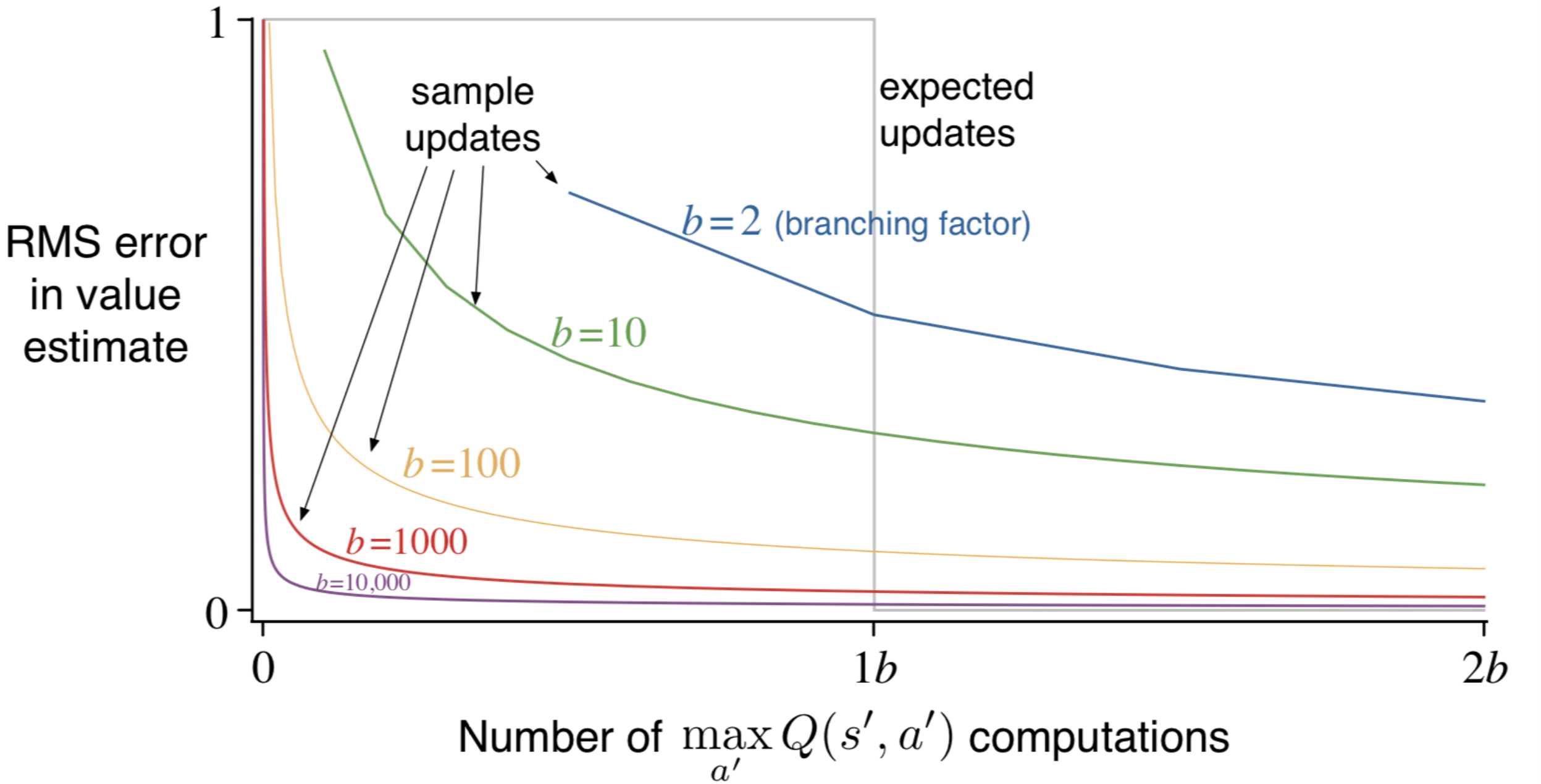
Neil Rhodes

Expected updates > sample updates?

Sample update (less compuation):
$Q(s,a) = Q(s,a) + \alpha[R + \gamma \max_{a'} Q(S', a') - Q(s,a)]$

Expected update (more accurate):
$Q(s,a) = \sum_{s',r} p(s', r|s,a)\alpha[r + \gamma \max_{a'} Q(s', a')]$

Expected updates > sample updates?



Trajectory Sampling

Sampling:

- Uniform

vs.

- According to on-policy distribution

When to plan

- In the background
- At decision time

What is a rollout algorithm?

1. Begin with current state
2. Simulate trajectories starting with that state (following a rollout policy)
3. Choose action with highest estimated value

Improves the rollout policy

Monte Carlo Tree Search (MCTS)

Repeat while time remaining starting with current state:
1. Selection: Select a leaf node in the expanded tree
2. Expansion: Expand a child of the leaf node
3. Simulation: Follow rollout-policy from expanded node to simulate complete episode
4. Backup: Backup action values to nodes in the tree

Selection: choose a leaf node in the MCTS tree

Traverse MCTS tree until reach a node with unexpanded children.

Use Upper Confidence Bound for Trees (UCT) to decide most promising child.

$q(v)$: Total simulation reward for node v
$n(v)$: Total number of visits (simulation backups) for node v

$$UCT(v) = \frac{q(v)}{n(v)} + c\sqrt{\frac{\log n(v.parent)}{n(v)}}$$

Expansion: expand selected node

If selected node is not terminal, choose an untried action and and create a new MCTS node for the state that generates

Simulation: rollout starting at expanded node

Monte Carlo simulation using rollout policy until terminal state is reached.

Record total reward.

Backup: backup action values to nodes in the MCTS tree

Update $n(v)$ and $q(v)$ for each node v in the MCTS tree from simulation node up to root.

Note: If two-player competitive game, adjust reward to reflect who made the move.

For example, if reward is +1 (player 1 won):

- For v reached from player 1 move, increment $q(v)$
- For v reached from player 2 move, decrement $q(v)$

Selecting a final action

Probably don't want exploration term in UCT

- Child with highest $\frac{q(v)}{n(v)}$, or
- Child of root with highest $N(v)$—it's the one that was explored the most so must have been most promising overall.