Recall: Discounted infinite horizon

- Markov decision process (MDP) \((S, A, P, \gamma, g)\)
- \(\gamma\): discount factor
- Policy \(\pi = (\mu_0, \mu_1, \ldots), \mu_k : S \rightarrow A\)
- Value of a policy \(\pi\):
  \[
  J^\pi(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t g(x(t), u(t))|x_0 = x, \pi\right]
  \]
- Goal: find \(\pi^* \in \arg\min_{\pi \in \Pi} J^\pi\)

Today

- Recap + continuation of value iteration with function approximation
- Performance boosts
- Speed-ups
- Intermezzo: Extremely crude outline of (part of) the reinforcement learning field [as it might assist when reading some of the references]

Great references:
- Tsitsiklis and Van Roy, 1996, “Feature based methods for large scale dynamic programming”
- Bertsekas and Tsitsiklis, “Neuro-dynamic programming,” Chap. 6

DP/VI with function approximation

Pick some \(S' \subseteq S\) typically the idea is that \(|S'| << |S|\).
Iterate for \(i = 0, 1, 2, \ldots\):
- back-ups/\(s \in S': J^{i+1}(s) := \min_{u \in A} g(s, u) + \gamma \sum_{s'} P(s'|s, u) J^i(s')\)
- projection: find some \(\theta^{i+1}\) such that \(\forall s \in S': \hat{J}^{i+1}(s) = (\Pi_{SS'}(s)) J^{i+1}(s) \approx J^{i+1}(s)\)

Projection enables generalization to \(s \in S \setminus S'\), which in turn enables the Bellman back-ups in the next iteration.
\(\theta\) parameterizes the class of functions used for approximation of the cost-to-go function

Example --- piecewise linear
Recall: VI with function approximation need not converge!

\[ P(\varepsilon x_1, u) = 1; P(\varepsilon x_2, u) = 0 \]
\[ g(x_1, u) = 0; g(x_2, u) = 0; \]

Function approximator: \([1 \ 2] * \theta\)

VI w/ least squares function approximation diverges for \(\gamma > 5/6\) [see last lecture for details]

\[ \text{Contraction} \]

- **Fact.** The Bellman operator, \(T\), is a \(\gamma\)-contraction w.r.t. the infinity norm, i.e.,
  \[ \forall j_1, j_2 : \|T_{j_1} - T_{j_2}\|_\infty \leq \gamma \|j_1 - j_2\|_\infty \]

- **Theorem.** The Bellman operator has a unique fixed point \(J^* = TJ^*\) and for all \(J\) we have that \(T^kJ\) converges to \(J^*\) for \(k\) going to infinity.
  
  **Note:**
  \[ \|T^{(k)} J - J^*\|_\infty = \frac{\|T^{(k)} J - T^{(k-1)} J^*\|_\infty}{\gamma^k} \leq \gamma \|J - J^*\|_\infty \]

  I.e., with every back-up, the infinity norm distance to \(J^*\) decreases.

\[ \text{Guarantees for fixed point} \]

**Theorem.** Let \(J^*\) be the optimal value function for a finite MDP with discount factor \(\gamma\). Let the projection operator \(P\) be a non-expansion w.r.t. the infinity norm and let \(\bar{J}\) be any fixed point of \(P\). Suppose \(\|\bar{J} - J^*\|_\infty \leq \epsilon\). Then \(P\) converges to a value function \(J\) such that:

\[ \|J - J^*\|_\infty \leq \frac{2\epsilon}{1 - \gamma} \]

\[ \text{Proof} \]

\[ \text{Can we generally verify goodness of some estimate J despite not having access to J*} \]

**Fact.** Assume there are some \(J\) for which we have that \(\|J - TJ\|_\infty \leq \epsilon\). Then we have that

\[ \|J - PJ\|_\infty \leq \frac{\epsilon}{1 - \gamma} \]

**Proof by induction:**

Base case: We have \(\|J^{(1)} - TJ^{(0)}\|_\infty \leq \epsilon\).

Induction: We also have for any \(i > 1\):

\[ \|T^i J^{(0)} - J^i\|_\infty \leq \frac{\epsilon + \gamma \|T^{i-1} J^{(0)} - T^{i-1} J^{(0)} - w^{(i-1)}\|_\infty}{\gamma^i} \]

\[ \leq \epsilon + \gamma \left( \epsilon (1 + \gamma + \gamma^2 + \ldots + \gamma^{i-2}) \right) \]

\[ \leq \epsilon + \frac{\epsilon}{1 - \gamma} \]

\[ \text{What if the projection fails to be a non-expansion} \]

**Assume \(\bar{J}\) only introduces a little bit of noise, i.e.,**

\[ \forall i \leq \infty : \|P^i J^{(0)} - \bar{J}^{(0)}\|_\infty \leq \epsilon \]

Or, more generally, we have a noisy sequence of back-ups:

\[ J^{(i+1)} \leftarrow TJ^{(i)} + w^{(i)} \] with the noise \(w^{(i)}\) satisfying

\[ \|w^{(i)}\|_\infty \leq \epsilon \]

**Fact.** \(\|P^i J^{(0)} - \bar{J}^{(0)}\|_\infty \leq \epsilon (1 + \gamma + \gamma^2 + \ldots + \gamma^{i-1})\) and as a consequence

\[ \|J^{(0)}\|_\infty \leq \frac{\epsilon}{1 - \gamma} \]
Improving performance with a given value function

1. Multi-stage lookahead aka Receding/Moving horizon
   - Rather than using greedy policy \( \mu \) w.r.t. approximate value function, with
     \[ \mu(s) = \arg \min_a g(s, a) + \gamma \sum_{s'} P(s'|s, a) J(s') \]
   - Two-stage lookahead:
     - At time \( t \) perform back-ups for all \( s' \) which are successor states of \( s \).
     - Then use these backed up values to perform the back-up for \( s \).
   - N stage lookahead: similarly, perform back-ups to N-stages of successor states of \( s \), backward in time
   - Can’t guarantee N-stage lookahead provides better performance [Can guarantee tighter infinity norm bound on attained value function estimates by N-stage lookahead.]
   - Example application areas in which it has improved performance chess, backgammon

[See also Bertsekas and Tsitsiklis, 6.1.2]

2. Roll-out policies
   - Given a policy \( \pi \), choose the current action \( u \) by evaluating the cost incurred by taking action \( u \) followed by executing the policy \( \pi \) from then onwards
   - Guaranteed to perform better than the baseline policy on top of which it builds (thanks to general guarantees of policy iteration algorithm)
   - Baseline policy could be obtained with any method
   - Practicalities
     - Todo --- fill in

[See also Bertsekas and Tsitsiklis, 6.1.3]
**Speed-ups**
- Parallelization
  - VI lends itself to parallelization
- Multi-grid, Coarse-to-fine grid, Variable resolution grid
- Prioritized sweeping
- Richardson extrapolation
- Kuhn triangulation

**Prioritized sweeping**
- Dynamic programming (DP) / Value iteration (VI):
  For $i=0, 1, \ldots$
  $$J^{(i+1)}(s) \leftarrow \min_{u \in A} g(s, u) + \gamma \sum_{s'} P(s'|s, u) J^{(i)}(s')$$
- Prioritized sweeping idea: focus updates on states for which the update is expected to be most significant
- Place states into priority queue and perform updates accordingly
  - For every Bellman update: compute the difference $J^{(i+1)} - J^{(i)}$ and update the priority of the states $s'$ from which one could transition into $s$ based upon the above difference and the transition probability of transitioning into $s$
  - For details: See Moore and Atkeson, 1993, "Prioritized sweeping: RL with less data and less real time"

**Richardson extrapolation**
- Generic method to improve the rate of convergence of a sequence
- Assume $h$ is the grid-size parameter in a discretization scheme
- Assume we can approximate $J^{(h)}(x)$ as follows:
  $$J^{(h)}(x) = J(x) + J_1(x)h + o(h)$$
- Similarly:
  $$J^{(h/2)}(x) = J(x) + J_1(x)h/2 + o(h)$$
- Then we can get rid of the order $h$ error term by using the following approximation which combines both:
  $$2J^{(h/2)}(x) - J^{(h)}(x) = J(x) + o(h)$$

**Kuhn triangulation**
- Allows efficient computation of the vertices participating in a point’s barycentric coordinate system and of the convex interpolation weights (aka the barycentric coordinates)
- See Munos and Moore, 2001 for further details.