Lecture Outline

- Discounted problems - Approximation on subspace \( \{ \Phi r \mid r \in \mathbb{R}^s \} \)
- Approximate (fitted) VI
- Approximate PI
- The projected equation
- Contraction properties - Error bounds
- Matrix form of the projected equation
- Simulation-based implementation
- LSTD and LSPE methods
REVIEW: APPROXIMATION IN VALUE SPACE

• Finite-spaces discounted problems: Defined by mappings $T_\mu$ and $T$ ($TJ = \min_\mu T_\mu J$).

• Exact methods:
  - VI: $J_{k+1} = TJ_k$
  - PI: $J_{\mu k} = T_\mu J_{\mu k}$, $T_{\mu k+1} J_{\mu k} = TJ_{\mu k}$
  - LP: $\min J c\trans J$ subject to $J \leq TJ$

• Approximate versions: Plug-in subspace approximation with $\Phi r$ in place of $J$
  - VI: $\Phi_{r_k} \approx TJ_{r_k}$
  - PI: $\Phi r_k \approx T_\mu \Phi r_k$, $T_{\mu k+1} \Phi r_k = T \Phi r_k$
  - LP: $\min_r c\trans \Phi r$ subject to $\Phi r \leq T \Phi r$

• Approx. onto subspace $S = \{\Phi r \mid r \in \mathbb{R}^s\}$ is often done by projection with respect to some (weighted) Euclidean norm.

• Another possibility is aggregation. Here:
  - The rows of $\Phi$ are probability distributions
  - $\Phi r \approx J_\mu$ or $\Phi r \approx J^*$, with $r$ the solution of an “aggregate Bellman equation” $r = DT_\mu(\Phi r)$ or $r = DT(\Phi r)$, where the rows of $D$ are probability distributions
APPROXIMATE (FITTED) VI

- Approximates sequentially $J_k(i) = (T^k J_0)(i)$, $k = 1, 2, \ldots$, with $\tilde{J}_k(i; r_k)$
- The starting function $J_0$ is given (e.g., $J_0 \equiv 0$)
- **Approximate (Fitted) Value Iteration:** A sequential “fit” to produce $\tilde{J}_{k+1}$ from $\tilde{J}_k$, i.e., $\tilde{J}_{k+1} \approx T \tilde{J}_k$ or (for a single policy $\mu$) $\tilde{J}_{k+1} \approx T_\mu \tilde{J}_k$

- After a large enough number $N$ of steps, $\tilde{J}_N(i; r_N)$ is used as approximation to $J^*(i)$
- Possibly use (approximate) projection $\Pi$ with respect to some projection norm,
  
  $\tilde{J}_{k+1} \approx \Pi T \tilde{J}_k$
WEIGHTED EUCLIDEAN PROJECTIONS

• Consider a weighted Euclidean norm

\[ \|J\|_\xi = \sqrt{\sum_{i=1}^{n} \xi_i (J(i))^2}, \]

where \( \xi = (\xi_1, \ldots, \xi_n) \) is a positive distribution (\( \xi_i > 0 \) for all \( i \)).

• Let \( \Pi \) denote the projection operation onto

\[ \mathcal{S} = \{ \Phi r \mid r \in \mathbb{R}^s \} \]

with respect to this norm, i.e., for any \( J \in \mathbb{R}^n \),

\[ \Pi J = \Phi r^* \]

where

\[ r^* = \arg \min_{r \in \mathbb{R}^s} \| \Phi r - J \|^2_\xi \]

• Recall that weighted Euclidean projection can be implemented by simulation and least squares, i.e., sampling \( J(i) \) according to \( \xi \) and solving

\[ \min_{r \in \mathbb{R}^s} \sum_{t=1}^{k} (\phi(i_t)'r - J(i_t))^2 \]
Fitted VI - Naive Implementation

- Select/sample a “small” subset $I_k$ of representative states

- For each $i \in I_k$, given $\tilde{J}_k$, compute

$$
(T\tilde{J}_k)(i) = \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u)(g(i,u,j) + \alpha \tilde{J}_k(j;r))
$$

- “Fit” the function $\tilde{J}_{k+1}(i;r_{k+1})$ to the “small” set of values $(T\tilde{J}_k)(i)$, $i \in I_k$ (for example use some form of approximate projection)

- “Model-free” implementation by simulation

- **Error Bound:** If the fit is uniformly accurate within $\delta > 0$, i.e.,

$$
\max_i |\tilde{J}_{k+1}(i) - T\tilde{J}_k(i)| \leq \delta,
$$

then

$$
\limsup_{k \to \infty} \max_{i=1,\ldots,n} (\tilde{J}_k(i,r_k) - J^*(i)) \leq \frac{\delta}{1 - \alpha}
$$

- But there is a potential serious problem!
AN EXAMPLE OF FAILURE

• Consider two-state discounted MDP with states 1 and 2, and a single policy.
  – Deterministic transitions: 1 → 2 and 2 → 2
  – Transition costs ≡ 0, so $J^*(1) = J^*(2) = 0$.
• Consider (exact) fitted VI scheme that approximates cost functions within $S = \{(r, 2r) \mid r \in \mathbb{R}\}$ with a weighted least squares fit; here $\Phi = (1, 2)'$
• Given $\tilde{J}_k = (r_k, 2r_k)$, we find $\tilde{J}_{k+1} = (r_{k+1}, 2r_{k+1})$, where $\tilde{J}_{k+1} = \Pi_\xi(T\tilde{J}_k)$, with weights $\xi = (\xi_1, \xi_2)$:
  $$ r_{k+1} = \arg\min_r \left[ \xi_1 (r - (T\tilde{J}_k)(1))^2 + \xi_2 (2r - (T\tilde{J}_k)(2))^2 \right] $$
• With straightforward calculation
  $$ r_{k+1} = \alpha \beta r_k, \quad \text{where } \beta = 2(\xi_1 + 2\xi_2)/(\xi_1 + 4\xi_2) > 1 $$
• So if $\alpha > 1/\beta$ (e.g., $\xi_1 = \xi_2 = 1$), the sequence $\{r_k\}$ diverges and so does $\{\tilde{J}_k\}$.
• Difficulty is that $T$ is a contraction, but $\Pi_\xi T$ (= least squares fit composed with $T$) is not.
NORM MISMATCH PROBLEM

• For fitted VI to converge, we need $\Pi_\xi T$ to be a contraction; $T$ being a contraction is not enough.

We need a $\xi$ such that $T$ is a contraction w. r. to the weighted Euclidean norm $\| \cdot \|_\xi$.

Then $\Pi_\xi T$ is a contraction w. r. to $\| \cdot \|_\xi$.

We will come back to this issue, and show how to choose $\xi$ so that $\Pi_\xi T_\mu$ is a contraction for a given $\mu$. 

Fitted Value Iteration with Projection

Subspace $S = \{ \Phi r \mid r \in \mathbb{R}^s \}$
Approximate Policy Iteration (APPI)

1. **Guess Initial Policy**
2. **Evaluate Approximate Cost**
   \( \tilde{J}_\mu(r) = \Phi r \) Using Simulation
3. **Generate “Improved” Policy** \( \overline{\mu} \)

**Evaluation of typical \( \mu \):** Linear cost function approximation \( \tilde{J}_\mu(r) = \Phi r \), where \( \Phi \) is full rank \( n \times s \) matrix with columns the basis functions, and \( i \)th row denoted \( \phi(i)' \).

**Policy “improvement”** to generate \( \overline{\mu} \):

\[
\overline{\mu}(i) = \arg \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u) \left( g(i, u, j) + \alpha \phi(j)' r \right)
\]

**Error Bound** (same as approximate VI): If

\[
\max_i |\tilde{J}_{\mu_k}(i, r_k) - J_{\mu_k}(i)| \leq \delta, \quad k = 0, 1, \ldots
\]

the sequence \( \{\mu^k\} \) satisfies

\[
\limsup_{k \to \infty} \max_i \left( J_{\mu_k}(i) - J^*(i) \right) \leq \frac{2\alpha \delta}{(1 - \alpha)^2}
\]
APPROXIMATE POLICY EVALUATION

• Consider approximate evaluation of $J_\mu$, the cost of the current policy $\mu$ by using simulation.
  
  – **Direct policy evaluation** - generate cost samples by simulation, and optimization by least squares
  
  – **Indirect policy evaluation** - solving the projected equation $\Phi r = \Pi T_\mu (\Phi r)$ where $\Pi$ is projection w/ respect to a suitable weighted Euclidean norm

• Recall that projection can be implemented by simulation and least squares
Given the current policy $\mu$:

- We solve the projected Bellman’s equation

$$\Phi r = \Pi T_\mu(\Phi r)$$

- We approximate the solution $J_\mu$ of Bellman’s equation

$$J = T_\mu J$$

with the projected equation solution $\tilde{J}_\mu(r)$
KEY QUESTIONS AND RESULTS

• Does the projected equation have a solution?

• Under what conditions is the mapping $\Pi T_\mu$ a contraction, so $\Pi T_\mu$ has unique fixed point?

• Assumption: The Markov chain corresponding to $\mu$ has a single recurrent class and no transient states, with steady-state prob. vector $\xi$, so that

$$\xi_j = \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} P(i_k = j \mid i_0 = i) > 0$$

Note that $\xi_j$ is the long-term frequency of state $j$.

• Proposition: (Norm Matching Property) Assume that the projection $\Pi$ is with respect to $\| \cdot \|_\xi$, where $\xi = (\xi_1, \ldots, \xi_n)$ is the steady-state probability vector. Then:

  (a) $\Pi T_\mu$ is contraction of modulus $\alpha$ with respect to $\| \cdot \|_\xi$.

  (b) The unique fixed point $\Phi_{r^*}$ of $\Pi T_\mu$ satisfies

$$\| J_\mu - \Phi_{r^*} \|_\xi \leq \frac{1}{\sqrt{1 - \alpha^2}} \| J_\mu - \Pi J_\mu \|_\xi$$
PRELIMINARIES: PROJECTION PROPERTIES

- Important property of the projection $\Pi$ on $S$ with weighted Euclidean norm $\| \cdot \|_\xi$. For all $J \in \mathbb{R}^n$, $\Phi r \in S$, the Pythagorean Theorem holds:

$\| J - \Phi r \|_\xi^2 = \| J - \Pi J \|_\xi^2 + \| \Pi J - \Phi r \|_\xi^2$

- The Pythagorean Theorem implies that the projection is nonexpansive, i.e.,

$\| \Pi J - \Pi \bar{J} \|_\xi \leq \| J - \bar{J} \|_\xi$, for all $J, \bar{J} \in \mathbb{R}^n$.

To see this, note that

$\| \Pi (J - \bar{J}) \|_\xi^2 \leq \| \Pi (J - \bar{J}) \|_\xi^2 + \| (I - \Pi) (J - \bar{J}) \|_\xi^2$

$= \| J - \bar{J} \|_\xi^2$
PROOF OF CONTRACTION PROPERTY

• Lemma: If $P$ is the transition matrix of $\mu$,
  $$\|Pz\|_\xi \leq \|z\|_\xi, \quad z \in \mathbb{R}^n,$$
where $\xi$ is the steady-state prob. vector.

Proof: For all $z \in \mathbb{R}^n$

$$\|Pz\|_\xi^2 = \sum_{i=1}^{n} \xi_i \left( \sum_{j=1}^{n} p_{ij}z_j \right)^2 \leq \sum_{i=1}^{n} \xi_i \sum_{j=1}^{n} p_{ij}z_j^2$$

$$= \sum_{j=1}^{n} \sum_{i=1}^{n} \xi_i p_{ij}z_j^2 = \sum_{j=1}^{n} \xi_j z_j^2 = \|z\|_\xi^2.$$

The inequality follows from the convexity of the quadratic function, and the next to last equality
follows from the defining property $\sum_{i=1}^{n} \xi_i p_{ij} = \xi_j$

• Using the lemma, the nonexpansiveness of $\Pi$,
and the definition $T_{\mu}J = g + \alpha PJ$, we have

$$\|\Pi T_{\mu}J - \Pi T_{\mu}\tilde{J}\|_\xi \leq \|T_{\mu}J - T_{\mu}\tilde{J}\|_\xi = \alpha \|P(J - \tilde{J})\|_\xi \leq \alpha \|J - \tilde{J}\|_\xi$$

for all $J, \tilde{J} \in \mathbb{R}^n$. Hence $\Pi T_{\mu}$ is a contraction of modulus $\alpha$. 
PROOF OF ERROR BOUND

- Let $\Phi r^*$ be the fixed point of $\Pi T$. We have

\[
\| J_\mu - \Phi r^* \|_\xi \leq \frac{1}{\sqrt{1 - \alpha^2}} \| J_\mu - \Pi J_\mu \|_\xi.
\]

Proof: We have

\[
\| J_\mu - \Phi r^* \|_\xi^2 = \| J_\mu - \Pi J_\mu \|_\xi^2 + \| \Pi J_\mu - \Phi r^* \|_\xi^2 \\
= \| J_\mu - \Pi J_\mu \|_\xi^2 + \| \Pi T J_\mu - \Pi T (\Phi r^*) \|_\xi^2 \\
\leq \| J_\mu - \Pi J_\mu \|_\xi^2 + \alpha^2 \| J_\mu - \Phi r^* \|_\xi^2,
\]

where

- The first equality uses the Pythagorean Theorem
- The second equality holds because $J_\mu$ is the fixed point of $T$ and $\Phi r^*$ is the fixed point of $\Pi T$
- The inequality uses the contraction property of $\Pi T$.

Q.E.D.
• The solution $\Phi r^*$ satisfies the orthogonality condition: The error

$$\Phi r^* - (g + \alpha P\Phi r^*)$$

is “orthogonal” to the subspace spanned by the columns of $\Phi$.

• This is written as

$$\Phi'\Xi(\Phi r^* - (g + \alpha P\Phi r^*)) = 0,$$

where $\Xi$ is the diagonal matrix with the steady-state probabilities $\xi_1, \ldots, \xi_n$ along the diagonal.

• Equivalently, $Cr^* = d$, where

$$C = \Phi'\Xi(I - \alpha P)\Phi, \quad d = \Phi'\Xi g$$

but computing $C$ and $d$ is HARD (high-dimensional inner products).
SOLUTION OF PROJECTED EQUATION

- Solve $Cr^* = d$ by matrix inversion: $r^* = C^{-1}d$
- Alternative: Projected Value Iteration (PVI)
  \[ \Phi r_{k+1} = \Pi T(\Phi r_k) = \Pi(g + \alpha P\Phi r_k) \]
  Converges to $r^*$ because $\Pi T$ is a contraction.

- PVI can be written as:
  \[ r_{k+1} = \arg \min_{r \in \mathbb{R}^s} \| \Phi r - (g + \alpha P\Phi r_k) \|_\xi^2 \]
  By setting to 0 the gradient with respect to $r$,
  \[ \Phi' \Xi (\Phi r_{k+1} - (g + \alpha P\Phi r_k)) = 0, \]
  which yields
  \[ r_{k+1} = r_k - (\Phi' \Xi \Phi)^{-1}(Cr_k - d) \]
SIMULATION-BASED IMPLEMENTATIONS

- **Key idea:** Calculate simulation-based approximations based on $k$ samples

\[ C_k \approx C, \quad d_k \approx d \]

- Approximate matrix inversion $r^* = C^{-1}d$ by

\[ \hat{r}_k = C_k^{-1}d_k \]

This is the **LSTD** (Least Squares Temporal Differences) method.

- **PVI method** $r_{k+1} = r_k - (\Phi'\Xi\Phi)^{-1}(Cr_k - d)$ is approximated by

\[ r_{k+1} = r_k - G_k(C_kr_k - d_k) \]

where

\[ G_k \approx (\Phi'\Xi\Phi)^{-1} \]

This is the **LSPE** (Least Squares Policy Evaluation) method.

- **Key fact:** $C_k$, $d_k$, and $G_k$ can be computed with low-dimensional linear algebra (of order $s$; the number of basis functions).
SIMULATION MECHANICS

- We generate an infinitely long trajectory \((i_0, i_1, \ldots)\) of the Markov chain, so states \(i\) and transitions \((i, j)\) appear with long-term frequencies \(\xi_i\) and \(p_{ij}\).

- After generating each transition \((i_t, i_{t+1})\), we compute the row \(\phi(i_t)'\) of \(\Phi\) and the cost component \(g(i_t, i_{t+1})\).

- We form

\[
d_k = \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t) g(i_t, i_{t+1}) \approx \sum_{i,j} \xi_i p_{ij} \phi(i) g(i, j) = \Phi' \Xi g = d
\]

\[
C_k = \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t) (\phi(i_t) - \alpha \phi(i_{t+1}))' \approx \Phi' \Xi (I-\alpha P) \Phi = C
\]

Also in the case of LSPE

\[
G_k = \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t) \phi(i_t)' \approx \Phi' \Xi \Phi
\]

- Convergence based on law of large numbers.

- \(C_k, d_k,\) and \(G_k\) can be formed incrementally. Also can be written using the formalism of temporal differences (this is just a matter of style).
OPTIMISTIC VERSIONS

- Instead of calculating nearly exact approximations \( C_k \approx C \) and \( d_k \approx d \), we do a less accurate approximation, based on few simulation samples.
- Evaluate (coarsely) current policy \( \mu \), then do a policy improvement.
- This often leads to faster computation (as optimistic methods often do).
- Very complex behavior (see the subsequent discussion on oscillations).
- The matrix inversion/LSTD method has serious problems due to large simulation noise (because of limited sampling) - particularly if the \( C \) matrix is ill-conditioned.
- LSPE tends to cope better because of its iterative nature (this is true of other iterative methods as well).
- A stepsize \( \gamma \in (0,1] \) in LSPE may be useful to damp the effect of simulation noise.

\[
r_{k+1} = r_k - \gamma G_k (C_k r_k - d_k)
\]