LECTURE OUTLINE

• Additional topics in approximate DP
• Stochastic shortest path problems
• Average cost problems
• Nonlinear versions of the projected equation
• Extension of $Q$-learning for optimal stopping
• Basis function adaptation
• Gradient-based approximation in policy space
For fixed policy $\mu$ to be evaluated, the solution of Bellman’s equation $J = TJ$ is approximated by the solution of

$$\Phi_r = \Pi T(\Phi_r)$$

whose solution is in turn obtained using a simulation-based method such as LSPE($\lambda$), LSTD($\lambda$), or TD($\lambda$).

These ideas apply to other (linear) Bellman equations, e.g., for SSP and average cost.

**Key Issue:** Construct framework where $\Pi T$ [or at least $\Pi T(\lambda)$] is a contraction.
STOCHASTIC SHORTEST PATHS

• Introduce approximation subspace

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

and for a given proper policy, Bellman’s equation and its projected version

\[ J = TJ = g + PJ, \quad \Phi r = \Pi T(\Phi r) \]

Also its \( \lambda \)-version

\[ \Phi r = \Pi T^{(\lambda)}(\Phi r), \quad T^{(\lambda)} = (1 - \lambda) \sum_{t=0}^{\infty} \lambda^t T^{t+1} \]

• **Question:** What should be the norm of projection?

• **Speculation based on discounted case:** It should be a weighted Euclidean norm with weight vector \( \xi = (\xi_1, \ldots, \xi_n) \), where \( \xi_i \) should be some type of long-term occupancy probability of state \( i \) (which can be generated by simulation).

• But what does “long-term occupancy probability of a state” mean in the SSP context?

• How do we generate infinite length trajectories given that termination occurs with prob. 1?
SIMULATION TRAJECTORIES FOR SSP

• We envision simulation of trajectories up to termination, followed by restart at state $i$ with some fixed probabilities $q_0(i) > 0$.

• Then the “long-term occupancy probability of a state” of $i$ is proportional to

\[
q(i) = \sum_{t=0}^{\infty} q_t(i), \quad i = 1, \ldots, n,
\]

where

\[
q_t(i) = P(i_t = i), \quad i = 1, \ldots, n, \ t = 0, 1, \ldots
\]

• We use the projection norm

\[
\|J\|_q = \sqrt{\sum_{i=1}^{n} q(i)(J(i))^2}
\]

[Note that $0 < q(i) < \infty$, but $q$ is not a prob. distribution. ]

• We can show that $\Pi T^{(\lambda)}$ is a contraction with respect to $\| \cdot \|_\xi$ (see the next slide).

• LSTD($\lambda$), LSPE($\lambda$), and TD($\lambda$) are possible.
CONTRACTION PROPERTY FOR SSP

• We have $q = \sum_{t=0}^{\infty} q_t$ so

$$q'P = \sum_{t=0}^{\infty} q'_t P = \sum_{t=1}^{\infty} q'_t = q' - q'_0$$

or

$$\sum_{i=1}^{n} q(i)p_{ij} = q(j) - q_0(j), \quad \forall \ j$$

• To verify that $\Pi_T$ is a contraction, we show that there exists $\beta < 1$ such that $\|Pz\|_q^2 \leq \beta \|z\|_q^2$ for all $z \in \mathbb{R}^n$.

• For all $z \in \mathbb{R}^n$, we have

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

$$= \sum_{j=1}^{n} z_j^2 \sum_{i=1}^{n} q(i)p_{ij} = \sum_{j=1}^{n} (q(j) - q_0(j)) z_j^2$$

$$= \|z\|_q^2 - \|z\|_{q_0}^2 \leq \beta \|z\|_q^2$$

where

$$\beta = 1 - \min_{j} \frac{q_0(j)}{q(j)}$$
AVERAGE COST PROBLEMS

- Consider a single policy to be evaluated, with single recurrent class, no transient states, and steady-state probability vector $\xi = (\xi_1, \ldots, \xi_n)$.
- The average cost, denoted by $\eta$, is independent of the initial state

$$\eta = \lim_{N \to \infty} \frac{1}{N} E \left\{ \sum_{k=0}^{N-1} g(x_k, x_{k+1}) \mid x_0 = i \right\}, \quad \forall \ i$$

- Bellman’s equation is $J = FJ$ with

$$FJ = g - \eta e + PJ$$

where $e$ is the unit vector $e = (1, \ldots, 1)$.
- The projected equation and its $\lambda$-version are

$$\Phi_r = \Pi F(\Phi_r), \quad \Phi_r = \Pi F^{(\lambda)}(\Phi_r)$$

- A problem here is that $F$ is not a contraction with respect to any norm (since $e = Pe$).
- However, $\Pi F^{(\lambda)}$ turns out to be a contraction with respect to $\| \cdot \|_\xi$ assuming that $e$ does not belong to $S$ and $\lambda > 0$ [the case $\lambda = 0$ is exceptional, but can be handled - see the text].
- LSTD($\lambda$), LSPE($\lambda$), and TD($\lambda$) are possible.
GENERALIZATION/UNIFICATION

- Consider approximate solution of \( x = T(x) \), where

\[
T(x) = Ax + b, \quad A \text{ is } n \times n, \quad b \in \mathbb{R}^n
\]

by solving the projected equation \( y = \Pi T(y) \), where \( \Pi \) is projection on a subspace of basis functions (with respect to some Euclidean norm).

- We will generalize from DP to the case where \( A \) is arbitrary, subject only to

\[
I - \Pi A : \text{invertible}
\]

- Benefits of generalization:
  - Unification/higher perspective for TD methods in approximate DP
  - An extension to a broad new area of applications, where a DP perspective may be helpful

- Challenge: Dealing with less structure
  - Lack of contraction
  - Absence of a Markov chain
GENERALIZED PROJECTED EQUATION

- Let $\Pi$ be projection with respect to

$$\|x\|_\xi = \sqrt{\sum_{i=1}^{n} \xi_i x_i^2}$$

where $\xi \in \mathbb{R}^n$ is a probability distribution with positive components.

- If $r^*$ is the solution of the projected equation, we have $\Phi r^* = \Pi(A\Phi r^* + b)$ or

$$r^* = \arg \min_{r \in \mathbb{R}^s} \sum_{i=1}^{n} \xi_i \left( \phi(i)' r - \sum_{j=1}^{n} a_{ij} \phi(j)' r^* - b_i \right)^2$$

where $\phi(i)'$ denotes the $i$th row of the matrix $\Phi$.

- Optimality condition/equivalent form:

$$\sum_{i=1}^{n} \xi_i \phi(i) \left( \phi(i) - \sum_{j=1}^{n} a_{ij} \phi(j) \right)' r^* = \sum_{i=1}^{n} \xi_i \phi(i) b_i$$

- The two expected values can be approximated by simulation.
**SIMULATION MECHANISM**

Row Sampling According to $\xi$

- **Row sampling**: Generate sequence $\{i_0, i_1, \ldots\}$ according to $\xi$, i.e., relative frequency of each row $i$ is $\xi_i$

- **Column sampling**: Generate $\{(i_0, j_0), (i_1, j_1), \ldots\}$ according to some transition probability matrix $P$ with
  
  $$p_{ij} > 0 \quad \text{if} \quad a_{ij} \neq 0,$$

  i.e., for each $i$, the relative frequency of $(i, j)$ is $p_{ij}$

- **Row sampling may be done using a Markov chain with transition matrix $Q$ (unrelated to $P$)**

- **Row sampling may also be done without a Markov chain - just sample rows according to some known distribution $\xi$ (e.g., a uniform)**
ROW AND COLUMN SAMPLING

Row Sampling According to $\xi$
(May Use Markov Chain $Q$)

Column Sampling
According to Markov Chain $P \sim |A|$

- **Row sampling** $\sim$ State Sequence Generation in DP. Affects:
  - The projection norm.
  - Whether $\Pi A$ is a contraction.

- **Column sampling** $\sim$ Transition Sequence Generation in DP.
  - Can be totally unrelated to row sampling. Affects the sampling/simulation error.
  - “Matching” $P$ with $|A|$ is beneficial (has an effect like in importance sampling).

- Independent row and column sampling allows exploration at will! Resolves the exploration problem that is critical in approximate policy iteration.
LSTD-LIKE METHOD

- Optimality condition/equivalent form of projected equation

\[ \sum_{i=1}^{n} \xi_i \phi(i) \left( \phi(i) - \sum_{j=1}^{n} a_{ij} \phi(j) \right) \quad r^* = \sum_{i=1}^{n} \xi_i \phi(i) b_i \]

- The two expected values are approximated by row and column sampling (batch 0 → t).

- We solve the linear equation

\[ \sum_{k=0}^{t} \phi(i_k) \left( \phi(i_k) - \frac{a_{i_k,j_k}}{p_{i_k,j_k}} \phi(j_k) \right) \quad r_t = \sum_{k=0}^{t} \phi(i_k) b_{i_k} \]

- We have \( r_t \rightarrow r^* \), regardless of \( \Pi A \) being a contraction (by law of large numbers; see next slide).

- An LSPE-like method is also possible, but requires that \( \Pi A \) is a contraction.

- Under the assumption \( \sum_{j=1}^{n} |a_{ij}| \leq 1 \) for all \( i \), there are conditions that guarantee contraction of \( \Pi A \); see the paper by Bertsekas and Yu, “Projected Equation Methods for Approximate Solution of Large Linear Systems,” 2009.
JUSTIFICATION W/ LAW OF LARGE NUMBERS

- We will match terms in the exact optimality condition and the simulation-based version.

- Let $\hat{\xi}_i^t$ be the relative frequency of $i$ in row sampling up to time $t$.

- We have

$$\frac{1}{t+1} \sum_{k=0}^{t} \phi(i_k)\phi(i_k)' = \sum_{i=1}^{n} \hat{\xi}_i^t \phi(i)\phi(i)' \approx \sum_{i=1}^{n} \xi_i \phi(i)\phi(i)'.$$

$$\frac{1}{t+1} \sum_{k=0}^{t} \phi(i_k)b_{i_k} = \sum_{i=1}^{n} \hat{\xi}_i^t \phi(i)b_i \approx \sum_{i=1}^{n} \xi_i \phi(i)b_i$$

- Let $\hat{p}_{ij}^t$ be the relative frequency of $(i, j)$ in column sampling up to time $t$.

$$\frac{1}{t+1} \sum_{k=0}^{t} a_{i_kj_k} \phi(i_k)\phi(j_k)'$$

$$= \sum_{i=1}^{n} \hat{\xi}_i^t \sum_{j=1}^{n} \hat{p}_{ij}^t \frac{a_{ij}}{p_{ij}} \phi(i)\phi(j)'$$

$$\approx \sum_{i=1}^{n} \xi_i \sum_{j=1}^{n} a_{ij} \phi(i)\phi(j)'.$$
Consider a system of the form
\[ x = T(x) = Af(x) + b, \]
where \( f : \mathbb{R}^n \to \mathbb{R}^n \) is a mapping with scalar components of the form \( f(x) = (f_1(x_1), \ldots, f_n(x_n)) \).

Assume that each \( f_i : \mathbb{R} \to \mathbb{R} \) is nonexpansive:
\[ |f_i(x_i) - f_i(\bar{x}_i)| \leq |x_i - \bar{x}_i|, \quad \forall i, x_i, \bar{x}_i \in \mathbb{R} \]
This guarantees that \( T \) is a contraction with respect to any weighted Euclidean norm \( \| \cdot \|_\xi \) whenever \( A \) is a contraction with respect to that norm.

Algorithms similar to LSPE [approximating \( \Phi r_{k+1} = \Pi T(\Phi r_k) \)] are then possible.

Special case: In the optimal stopping problem of Section 6.4, \( x \) is the \( Q \)-factor corresponding to the continuation action, \( \alpha \in (0, 1) \) is a discount factor, \( f_i(x_i) = \min\{c_i, x_i\} \), and \( A = \alpha P \), where \( P \) is the transition matrix for continuing.

If \( \sum_{j=1}^n p_{ij} < 1 \) for some state \( \bar{i} \), and \( 0 \leq P \leq Q \), where \( Q \) is an irreducible transition matrix, then \( \Pi((1-\gamma)I + \gamma T) \) is a contraction with respect to \( \| \cdot \|_\xi \) for all \( \gamma \in (0, 1) \), even with \( \alpha = 1. \)
BASIS FUNCTION ADAPTATION I

• An important issue in ADP is how to select basis functions.

• A possible approach is to introduce basis functions that are parametrized by a vector $\theta$, and optimize over $\theta$, i.e., solve the problem

$$\min_{\theta \in \Theta} F(\tilde{J}(\theta))$$

where $\tilde{J}(\theta)$ is the solution of the projected equation.

• One example is

$$F(\tilde{J}(\theta)) = \| \tilde{J}(\theta) - T(\tilde{J}(\theta)) \|^2$$

• Another example is

$$F(\tilde{J}(\theta)) = \sum_{i \in I} | J(i) - \tilde{J}(\theta)(i) |^2,$$

where $I$ is a subset of states, and $J(i)$, $i \in I$, are the costs of the policy at these states calculated directly by simulation.
Some algorithm may be used to minimize $F(\tilde{J}(\theta))$ over $\theta$.

A challenge here is that the algorithm should use low-dimensional calculations.

One possibility is to use a form of random search (the cross-entropy method); see the paper by Menache, Mannor, and Shimkin (Annals of Oper. Res., Vol. 134, 2005)

Another possibility is to use a gradient method. For this it is necessary to estimate the partial derivatives of $\tilde{J}(\theta)$ with respect to the components of $\theta$.

It turns out that by differentiating the projected equation, these partial derivatives can be calculated using low-dimensional operations. See the paper by Menache, Mannor, and Shimkin, and the paper by Yu and Bertsekas (2009).
APPROXIMATION IN POLICY SPACE I

- Consider an average cost problem, where the problem data are parametrized by a vector $r$, i.e., a cost vector $g(r)$, transition probability matrix $P(r)$. Let $\eta(r)$ be the (scalar) average cost per stage, satisfying Bellman’s equation

$$\eta(r)e + h(r) = g(r) + P(r)h(r)$$

where $h(r)$ is the corresponding differential cost vector.

- Consider minimizing $\eta(r)$ over $r$ (here the data dependence on control is encoded in the parametrization). We can try to solve the problem by non-linear programming/gradient descent methods.

- **Important fact:** If $\Delta \eta$ is the change in $\eta$ due to a small change $\Delta r$ from a given $r$, we have

$$\Delta \eta = \xi'(\Delta g + \Delta Ph),$$

where $\xi$ is the steady-state probability distribution/vector corresponding to $P(r)$, and all the quantities above are evaluated at $r$:

$$\Delta \eta = \eta(r + \Delta r) - \eta(r),$$

$$\Delta g = g(r + \Delta r) - g(r), \quad \Delta P = P(r + \Delta r) - P(r)$$
Proof of the gradient formula: We have, by “differentiating” Bellman’s equation,

\[ \Delta \eta(r) \cdot e + \Delta h(r) = \Delta g(r) + \Delta P(r)h(r) + P(r)\Delta h(r) \]

By left-multiplying with \( \xi' \),

\[ \xi' \Delta \eta(r) \cdot e + \xi' \Delta h(r) = \xi' \left( \Delta g(r) + \Delta P(r)h(r) \right) + \xi' P(r)\Delta h(r) \]

Since \( \xi' \Delta \eta(r) \cdot e = \Delta \eta(r) \) and \( \xi' = \xi' P(r) \), this equation simplifies to

\[ \Delta \eta = \xi'(\Delta g + \Delta Ph) \]

Since we don’t know \( \xi \), we cannot implement a gradient-like method for minimizing \( \eta(r) \). An alternative is to use “sampled gradients”, i.e., generate a simulation trajectory \((i_0, i_1, \ldots)\), and change \( r \) once in a while, in the direction of a simulation-based estimate of \( \xi'(\Delta g + \Delta Ph) \).

Important Fact: \( \Delta \eta \) can be viewed as an expected value!

There is much research on this subject, see e.g., the work of Marbach and Tsitsiklis, and Konda and Tsitsiklis, and the refs given there.
6.231 Dynamic Programming and Stochastic Control
Fall 2011

For information about citing these materials or our Terms of Use, visit: http://ocw.mit.edu/terms.