LECTURE 23

LECTURE OUTLINE

• Additional topics in ADP
• Stochastic shortest path problems
• Average cost problems
• Generalizations
• Basis function adaptation
• Gradient-based approximation in policy space
• An overview
• **Policy Evaluation**: Bellman’s equation $J = TJ$ is approximated the projected equation

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

which can be solved by a simulation-based methods, e.g., LSPE($\lambda$), LSTD($\lambda$), or TD($\lambda$). Aggregation is another approach - simpler in some ways.

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

- **Important Issue**: Construct simulation framework where $\Pi T$ [or $\Pi T(\lambda)$] 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? How to implement it by simulation?

- 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 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 $\| \cdot \|_q$ (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

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

• 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 \( e = Pe \)).

\( \Pi F(\lambda) \) is a contraction w. r. 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 approx. 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 can generalize from DP to the case where $A$ is arbitrary, subject only to

$$I - \Pi A : \text{ invertible}$$

Also can deal with case where $I - \Pi A$ is (nearly) singular (iterative methods, see the text).

• Benefits of generalization:
  
  – Unification/higher perspective for projected equation (and aggregation) methods in approximate DP
  
  – An extension to a broad new area of applications, based on an approx. DP perspective

• 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
• **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 \(p_{ij}\) (connection to importance sampling)

• 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 $Q$)

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

Column Sampling According to $|A|$
(May Use Markov Chain $P$ $\sim |A|$)

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

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

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

• We solve the linear equation

\[
\sum_{k=0}^{t} \phi(i_k) \left( \phi(i_k) - \frac{a_{i_k j_k}}{p_{i_k j_k}} \phi(j_k) \right)' r_t = \sum_{k=0}^{t} \phi(i_k) b_{i_k}
\]

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

• Issues of singularity or near-singularity of \( I - \Pi A \) may be important; see the text.

• 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 text.
• 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 $t$.

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

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

$$\approx \sum_{i=1}^{n} \xi_i \sum_{j=1}^{n} a_{ij} \phi(i)\phi(j)'$$
BASIS FUNCTION ADAPTATION I

- An important issue in ADP is how to select basis functions.
- A possible approach is to introduce basis functions parametrized by a vector $\theta$, and optimize over $\theta$, i.e., solve a problem of the form
  \[
  \min_{\theta \in \Theta} F(\tilde{J}(\theta))
  \]
  where $\tilde{J}(\theta)$ approximates a cost vector $J$ on the subspace spanned by the basis functions.
- One 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.
- Another example is
  \[
  F(\tilde{J}(\theta)) = \| \tilde{J}(\theta) - T(\tilde{J}(\theta)) \|^2,
  \]
  where $\tilde{J}(\theta)$ is the solution of a projected equation.
Some optimization 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 references in the text.
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 differential cost vector.

• Consider minimizing $\eta(r)$ over $r$. Other than random search, we can try to solve the problem by a policy gradient method:

$$r_{k+1} = r_k - \gamma_k \nabla \eta(r_k)$$

• Approximate calculation of $\nabla \eta(r_k)$: If $\Delta \eta$, $\Delta g$, $\Delta P$ are the changes in $\eta$, $g$, $P$ 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 $r$. 
 APPROXIMATION IN POLICY SPACE II

• **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 \( \xi' \),

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

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!

• Much research on this subject, see the text.
6.231 DYNAMIC PROGRAMMING

OVERVIEW-EPILOGUE

- Finite horizon problems
  - Deterministic vs Stochastic
  - Perfect vs Imperfect State Info
- Infinite horizon problems
  - Stochastic shortest path problems
  - Discounted problems
  - Average cost problems
FINITE HORIZON PROBLEMS - ANALYSIS

• Perfect state info
  – A general formulation - Basic problem, DP algorithm
  – A few nice problems admit analytical solution

• Imperfect state info
  – Reduction to perfect state info - Sufficient statistics
  – Very few nice problems admit analytical solution
  – Finite-state problems admit reformulation as perfect state info problems whose states are prob. distributions (the belief vectors)
FINITE HORIZON PROBS - EXACT COMP. SOL.

- Deterministic finite-state problems
  - Equivalent to shortest path
  - A wealth of fast algorithms
  - Hard combinatorial problems are a special case (but # of states grows exponentially)

- Stochastic perfect state info problems
  - The DP algorithm is the only choice
  - Curse of dimensionality is big bottleneck

- Imperfect state info problems
  - Forget it!
  - Only small examples admit an exact computational solution
FINITE HORIZON PROBS - APPROX. SOL.

- Many techniques (and combinations thereof) to choose from

- Simplification approaches
  - Certainty equivalence
  - Problem simplification
  - Rolling horizon
  - Aggregation - Coarse grid discretization

- Limited lookahead combined with:
  - Rollout
  - MPC (an important special case)
  - Feature-based cost function approximation

- Approximation in policy space
  - Gradient methods
  - Random search
INFINITE HORIZON PROBLEMS - ANALYSIS

- A more extensive theory
- Bellman’s equation
- Optimality conditions
- Contraction mappings
- A few nice problems admit analytical solution
- Idiosynchracies of problems with no underlying contraction
- Idiosynchracies of average cost problems
- Elegant analysis
INF. HORIZON PROBS - EXACT COMP. SOL.

- Value iteration
  - Variations (Gauss-Seidel, asynchronous, etc)
- Policy iteration
  - Variations (asynchronous, based on value iteration, optimistic, etc)
- Linear programming
- Elegant algorithmic analysis
- Curse of dimensionality is major bottleneck
INFINITE HORIZON PROBS - ADP

- Approximation in value space (over a subspace of basis functions)

- Approximate policy evaluation
  - Direct methods (fitted VI)
  - Indirect methods (projected equation methods, complex implementation issues)
  - Aggregation methods (simpler implementation/many basis functions tradeoff)

- Q-Learning (model-free, simulation-based)
  - Exact Q-factor computation
  - Approximate Q-factor computation (fitted VI)
  - Aggregation-based Q-learning
  - Projected equation methods for opt. stopping

- Approximate LP

- Rollout

- Approximation in policy space
  - Gradient methods
  - Random search