IE 3186: Approximate Dynamic Programming **Fall 2018**

Lecture 10: Aggregation, Feature-based VI, Natural PG

Lecturer: Daniel Jiang *Scribes: Mingyuan Xu, Tarik Bilgic*

Main Reference: Tsitsiklis, J.N. and Van Roy, B. *Feature-based methods for large scale dynamic programming.* Machine Learning, 22(1-3), pp. 59-94, 1996.

Paper Discussion: Amari, S.I.. *Natural gradient works efficiently in learning.* Neural Computation, 10(2), pp. 251-276, 1998.

Kakade, S.M. *A natural policy gradient.* Advances in Neural Information Processing Systems, 2002.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. *Trust region policy optimization.* International Conference on Machine Learning, 2015.

Peters, J., Schaal, S. *Natural actor-critic.* Neurocomputing, 71(7-9), pp. 1180-1190, 2008.

Previously we learned the following results:

- AVI has an error bound when per-iteration Bellman error controlled in $\|\cdot\|_\infty$.
- AVI may diverge.
- Using a special norm $\|\cdot\|_\xi$, we can design a projected AVI which converges.
- AVI for control works for the special case of optimal stopping by using $\|\cdot\|_\xi$.

10.1 State Aggregation

Let the state space S be partitioned into $\{S_j\}$ for $j = 1, 2, \cdots, m$. So $\cup_j S_j = S$ and $S_k \cap S_l = \emptyset$.

- Assume the partition is given (we are not considering adaptive partitioning).
- Try to learn one value W_j per partition S_j. The vector $W = (W_1, \ldots, W_m)$ becomes the parameter of the value function approximation: $\bar{J}(W)(i) = \sum_j W_j \mathbb{1}_{\{i \in S_j\}}$.
- With large enough m and well chosen sets, $\bar{J}(W)(i) \approx J^*(i)$.

10.1.1 Asynchronous VI algorithm for learning the weights

- Let $\Gamma_j \subseteq \{0, 1, 2, \cdots\}$ be the iterations at which S_j's value is updated. $|\Gamma_j| = \infty$
- Let $p^j(\cdot)$ be a distribution over S_j

The algorithm is as follows:

1. On iteration $n+1$, sample $X^{n+1} = (X_1^{n+1}, X_2^{n+1}, \cdots, X_m^{n+1})$ where $X_j^{n+1} \sim p^j(\cdot)$.
2. $W_j^{n+1}(j) = (1 - \alpha^{n+1}(j))W^n(j) + \alpha^{n+1}(j)[T\bar{J}(W^n)(X_j^{n+1})], \qquad n+1 \in \Gamma_j,$
   $W_j^{n+1}(j) = W_j^n(j), \qquad n+1 \notin \Gamma_j.$

This is for simplicity of notation; in a practical implementation, sample X_j^{n+1} only if j being updated.

Theorem 10.1. *Assume standard stepsize assumption conditions hold, then:*

(i): $W^n \to W^*$ *a.s. where* W^* *solves*

$$W^*(j) = \sum_{i \in S_i} p^j(i)(T\bar{J}(W^*))(i).$$

(ii): For each aggregate status $j = \{1, 2, \cdots, m\}$*, let* $e_j = \max_{k,\, l \in S_j} |J^*(k) - J^*(l)|$ *and* π^{W^*} *be the greedy policy w.r.t.* $\bar{J}(W^*)$*. Then, the value function approximation satisfies*

$$\|\bar{J}(W^*) - J^*\|_\infty \leq \|\mathbf{e}\|_\infty / (1 - \gamma).$$

(iii): $\|J^{W^*} - J^*\|_\infty \leq 2\gamma\|\mathbf{e}\|_\infty/(1-\gamma)^2$*, where* J^{W^*} *is the performance of policy* π^{W^*}*.*

(iv): There exsits an MDP for which (ii), (iii) are tight.

Main ideas:

Let $T' : \mathbb{R}^m \to \mathbb{R}^m$, $(T'W)(j) = \mathbf{E}_{p^j}[(T\bar{J}(W))(X_j)] = \sum_i p^j(i)((T\bar{J}(W)(i)).$

T' returns the average value of $T\bar{J}(W)$ over the partition j.

Then $W^{n+1}(j) = (1 - \alpha)W^n(j) + \alpha((T'W^n)(j) + \text{sampling noise}).$

Thus the original problem is converted to prove the contraction property of T'.

Define another function that takes full value functions $J \in \mathbb{R}^n$ to aggregated ones: $(\bar{J}^{-1}(J))(j) = \sum_{i \in S_j} p^j(i)J(i)$. The following holds, which is called *pseudo inverse property*:

$$(\bar{J}^{-1}(\bar{J}(W)))(j) = \sum_{i \in S_j} p^j(i)\bar{J}(W)(i) = W(j).$$

Also, note that $T' = J^{-1} \circ T \circ \bar{J}$.

The following are true:

1. T is a γ- contraction on $\|\cdot\|_\infty$.
2. $\|\bar{J}(W) - \bar{J}(W')\|_\infty \leq \|W - W'\|_\infty$. This is because

$$\begin{aligned}\|\bar{J}(W) - \bar{J}(W')|_\infty &= \max_i |(\bar{J}(W)(i) - \bar{J}(W')(i)| \\ &= \max_j |W(j) - W'(j)| = \|W - W'\|_\infty.\end{aligned}$$

3. $\|\bar{J}^{-1}(J) - \bar{J}^{-1}(J')\|_\infty \leq \|J - J'\|_\infty$

Proof. $$\begin{aligned}\|\bar{J}^{-1}(J) - \bar{J}^{-1}(J')\|_\infty &= \max_{j \in \{1,\ldots,m\}} \left| \sum_{i \in S_j} p^j(i)(J(i) - J'(i)) \right| \\ &\leq \max_j \max_{i \in S_j} |J(i) - J'(i)| \\ &= \max_{\text{all states } i} |J(i) - J'(i)| \leq \|J - J'\|_\infty.\end{aligned}$$

4. $\|T'W - T'W'\|_\infty \leq \gamma \|W - W'\|_\infty$ can be proved by 1, 2, and 3.

Proof. We now prove the theorem.

(i) Apply our standard stochastic approximation/SGD results with T' to show convergence to a fixed point W^*.

(ii) Using a constant to approximate J^* in S_j, minimum error is $\|\mathbf{e}\|_\infty/2$. Therefore,

$$\min_W \|\bar{J}(W) - J^*\|_\infty = \|\mathbf{e}\|_\infty/2.$$

Let $\hat{W}$ be a vector that achieves the minimum: $\|\bar{J}(\hat{W}) - \bar{J}^*\|_\infty = \|\mathbf{e}\|_\infty/2 =: \epsilon$. First, a preliminary inequality:

$$\begin{aligned}\|W^* - \hat{W}\|_\infty &\leq \|W^* - T'\hat{W}\|_\infty + \|T'\hat{W} - \hat{W}\|_\infty \\ &\leq \gamma\|W^* - \hat{W}\|_\infty + \|\hat{J}^{-1}T\bar{J}\hat{W} - \bar{J}^{-1}\bar{J}\hat{W}\|_\infty \\ &\leq \gamma\|W^* - \hat{W}\|_\infty + \|T\bar{J}\hat{W} - \bar{J}\hat{W}\|_\infty \quad \text{(non-expansiveness of of } \bar{J}^{-1}) \\ &\leq \gamma\|W^* - \hat{W}\|_\infty + \|T\bar{J}\hat{W} - J^*\|_\infty + \epsilon \\ &\leq \gamma\|W^* - \hat{W}\|_\infty + \gamma\epsilon + \epsilon\end{aligned}$$

Therefore:

$$\|W^* - \hat{W}\|_\infty \leq \frac{1+\gamma}{1-\gamma}\epsilon.$$

Next, note that by

$$\begin{aligned}\|\bar{J}(W^*) - J^*\|_\infty &\leq \|\bar{J}(W^*) - \bar{J}(\hat{W})\|_\infty + \epsilon \\ &\leq \|W^* - \hat{W}\|_\infty + \epsilon \\ &\leq \left(\frac{1+\gamma}{1-\gamma} + \frac{1-\gamma}{1-\gamma}\right)\epsilon \\ &= \frac{\|\mathbf{e}\|_\infty}{1-\gamma}.\end{aligned}$$

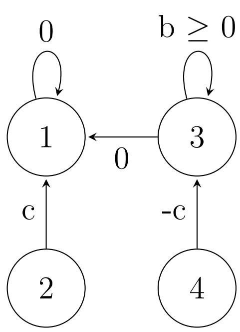
(iii) (Performance) $J^{\pi W^*}$ is the performance of the policy greedy w.r.t. $\bar{J}^{W^*}$, $T_{\pi W^*}$ is like T_μ with μ being this greedy policy.

$$\begin{aligned}\|J^{\pi W^*} - J^*\|_\infty &\leq \|T_{\pi W^*}J^{\pi W^*} - T\bar{J}(W^*)\|_\infty + \|T\bar{J}(W^*) - J^*\|_\infty \\ &= \|T_{\pi W^*}J^{\pi W^*} - T_{\pi W^*}\bar{J}(W^*)\|_\infty + \|T\bar{J}(W^*) - TJ^*\|_\infty \\ &\leq \gamma\|J^{\pi W^*} - J^* + J^* - \bar{J}(W^*)\| + \gamma\|\bar{J}(W^*) - J^*\|_\infty,\end{aligned}$$

$$\Rightarrow (1-\gamma)\|J^{\pi W^*} - J^*\|_\infty \leq 2\gamma\|\bar{J}W^* - J^*\|_\infty$$

$$\Rightarrow \|J^{\pi W^*} - J^*\|_\infty \leq \frac{2\gamma\|\mathbf{e}\|_\infty}{(1-\gamma)^2}.$$

(iv) The following MDP example shows that (ii), (iii) are tight.



is an MDP with four states.

Let the aggregated states be $S_A = \{1,2\}$, $S_B = \{3,4\}$ and suppose we always sample $\{2\}$ for S_A and $\{4\}$ for S_B.

Therefore, $J^*(1) = 0, J^*(2) = c, J^*(3) = 0, J^*(4) = -c, \|\mathbf{e}\|_\infty = c$.

By (i): $W^*(A) = c + \gamma W^*(A)$

$W^*(B) = -c + \gamma W^*(B)$

$$\Rightarrow W^* = \left(\frac{c}{1-\gamma}, \frac{-c}{1+\gamma}\right).$$

So, the maximum approximation error of $\bar{J}W^*$ is $\dfrac{c}{1-\gamma} = \dfrac{\|\mathbf{e}\|_\infty}{1-\gamma}$.

Let $b = \frac{2\gamma c}{1-\gamma}$ and consider π^{W^*}. At state 3: staying gives a cost of $b + \gamma(\frac{-c}{1-\gamma})$, and going to state 1 gives a cost of $0 + \gamma(\frac{c}{1-\gamma})$. They are equal.

Suppose the policy chooses to stay, which gives a cost of

$$b + \gamma b + \gamma^2 b + \cdots = \frac{b}{1-\gamma} = \frac{2\gamma\|\mathbf{e}\|_\infty}{(1-\gamma)^2}.$$

□

10.1.2 A Representative State Approach

A more general linear approach: $\bar{J}(W)(i) = \sum_{i=1}^{k} W_k f_k(i) = W^T F(i)$ where $F(i) = (f_1(i), f_2(i) \cdots f_k(i))$.

Main idea: Choose $(i_1, i_2, \cdots, i_k)$ representative states to perform V.I.

Special Assumptions:

1. $F(i_1), F(i_2), \ldots, F(i_k)$ are linearly independent.
2. There exists $\gamma' \in [\gamma, 1)$ such that $\forall\, i \in S$ there exists scalars $\theta_1(i), \theta_2(i), \cdots, \theta_k(i)$ with $\sum_{k=1}^{K} |\theta_k(i)| \leq 1$ and $F(i) = \frac{\gamma'}{\gamma} \sum_{k=1}^{K} \theta_k(i) F(i_k)$.

$\Rightarrow \|\bar{J}(W)\|_\infty \leq \frac{\gamma'}{\gamma} \max_i |\bar{J}(W)(i)|$.

Note that aggregation is a special case, which has $\theta_k(i) = 0$ or 1.

Definition 10.2. $M \in \mathbb{R}^{n \times k}$:

$$M = \begin{bmatrix} F^T(1) \\ F^T(2) \\ \vdots \\ F^T(n) \end{bmatrix} \tag{10.1}$$

Then $\bar{J}(W) = MW$. Assume $i_1 = 1, i_2 = 2, \cdots, i_k = k$ (w.l.o.g).

Definition 10.3. *$L \in \mathbb{R}^{k \times k}$ to be M restricted to the representative states:*

$$L = \begin{bmatrix} F^T(1) \\ F^T(2) \\ \vdots \\ F^T(k) \end{bmatrix} \tag{10.2}$$

Let $M^{-1} = [L^{-1},\ 0]$, so $M^{-1}M = L^{-1}L = I$, $T' = M^{-1} \circ T \circ M$. Let's consider the algorithm $W^{n+1} = T'\,W^n$.

Theorem 10.4. *Assume special assumptions hold, then:*

(i): $W^n \to W^*$,

(ii): T' *is a* γ'*-contraction w.r.t* $\|\cdot\|_M$*, where* $\|W\|_M = \|MW\|_\infty$.

Let $\epsilon = \inf_W \|J^* - \bar{J}W\|$. *We also have:*

(iii): $\|J^* - \bar{J}W\|_\infty \le \dfrac{\gamma + \gamma'}{\gamma(1-\gamma')}\epsilon$,

(iv): $\|J^{\pi W^*} - J^*\|_\infty \le \dfrac{2(\gamma + \gamma')}{(1-\gamma)(1-\gamma')}\epsilon$.

We will most of the proof as it is similar to the aggregation case. However, the following is to prove the contraction of M^{-1} : $\|M^{-1}J - M^{-1}J'\|_M \le \frac{\gamma'}{\gamma}\|J - J'\|_\infty$.

Proof. Let $D = M(M^{-1}J - M^{-1}J')$, so $\|D\|_\infty = \|M^{-1}J - M^{-1}J'\|_M$

$$
\begin{aligned}
|D(i)| &= |[M(M^{-1}J - M^{-1}J')](i)| \\
&= |F(i)^T(M^{-1}J - M^{-1}J')| \\
&= \frac{\gamma'}{\gamma}\left|\sum_{k=1}^{K}\theta_k(i)F^T(i_k)(M^{-1}J - M^{-1}J')\right| \\
&\le \frac{\gamma'}{\gamma}\max_k\left|F^T(i_k)(M^{-1}J - M^{-1}J')\right| \cdot \left|\sum_{k=1}^{K}\theta_k(i)\right| \\
&\le \frac{\gamma'}{\gamma}\max_k|D(i_k)| \\
&= \frac{\gamma'}{\gamma}\max_k|J(i_k) - J'(i_k)| \quad (\because M^{-1} = [L^{-1},\ 0]) \\
&\le \frac{\gamma'}{\gamma}\|J - J'\|_\infty.
\end{aligned}
$$

10.2 Paper Discussion: Natural Policy Gradient, Actor-Critic

10.2.1 Natural Descent Derivation by KL-divergence

Consider a parameter vector $\theta = \left[\theta_1, \theta_2, \ldots, \theta_N\right]^{\mathrm{T}}$, Fisher information matrix (FIM):

$$\left[\mathcal{I}(\theta)\right]_{i,j} = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta_i}\log f(X;\theta)\right)\left(\frac{\partial}{\partial\theta_j}\log f(X;\theta)\right)\middle|\,\theta\right],$$

is positive semidefinite.

Consider a KL-divergence between $p(\theta)$ and $p(\theta + \Delta\theta)$:

$$\mathrm{KL}(p(\theta{+}\Delta\theta)\|p(\theta)) = \int \ln\left(\frac{p(x|\theta + \Delta\theta)}{p(x|\theta)}\right) p(x|\theta{+}\Delta\theta)dx \approx \frac{1}{2}\sum_i\sum_j \left[\mathcal{I}(\theta)\right]_{i,j}\Delta\theta_i\Delta\theta_j,$$

Construct a Lagrangian function:

$$L(\Delta\theta, \lambda) = \sum_i \frac{\partial\mathbb{E}[f|\theta]}{\partial\theta_i}\Delta\theta_i + \lambda\left(\epsilon - \frac{1}{2}\sum_i\sum_j \left[\mathcal{I}(\theta)\right]_{i,j}\Delta\theta_i\Delta\theta_j,\right)$$

where $\mathrm{KL} \leq \epsilon$. So the matrix form is: $\nabla_\theta{}^T\mathbb{E}[f|\theta]\Delta\theta + \lambda\left(\epsilon - \frac{1}{2}\boldsymbol{\Delta}\theta^T\left[\mathcal{I}(\theta)\right]\ \boldsymbol{\Delta}\theta\right)$.

Then the optimal solution for the Lagragian function is: $-\left[\mathcal{I}(\theta)\right]^{-1}\nabla_\theta\mathbb{E}[f|\theta]$.

10.2.2 Actor-Critic

Policy gradient can be written as: $\mathbb{E}\left[\sum_{t=0}^{\infty}\Psi_t\nabla_\theta\log\pi_\theta\left(a_t|s_t\right)\right]$, where

- the policy $\pi_\theta\left(a|s\right)$ is the "actor,"
- and the value function approximation Ψ_t is the "critic." Many options to learn, usually use $\mathrm{TD}(\lambda)$.