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

Lecture 11: Approximate Policy Iteration

Lecturer: Daniel Jiang *Scribes: Boyuan Lai, Ibrahim El Shar*

References:

D. P. Bertsekas. *Dynamic Programming and Optimal Control: Approximate Dynamic Programming*, Vol. 2, 4th ed., Athena Scientific, Belmont MA, 2012. (§6.2)

D. P. Bertsekas. *Approximate policy iteration: A survey and some new methods.* Journal of Control Theory and Applications, 9(3), 310-335, 2011.

R. Munos and C. Szepesvari. *Finite-time bounds for fitted value iteration.* Journal of Machine Learning Research 9.May (2008): 815-857.

11.1 Approximate Policy Iteration

Recall the exact PI algorithm; on iteration k,

- Policy evaluation: compute the value of μ^k, i.e. solve $T_{\mu^k} J = J$ to get J_{μ^k};
- Policy improvement: find μ^{k+1} that is greedy with respective to J_{μ^k}.

The policy improvement property is at the core of why PI works. It was proven in a previous lecture.

Theorem 11.1 (Policy Improvement Property)**.** *μ^{k+1} is improved from μ^k in the sense that $J_{\mu^{k+1}} \leq J_{\mu^k}$ and if μ^k is not optimal, we have a strict improvement.*

11.1.1 Optimistic Policy Iteration

- A first approximation to exact PI: do partial evaluation of μ^k using a few steps of VI: on iteration k, do m_k steps of VI so that $J_{\mu^k} \approx T_{\mu^k}^{m_k} J$ for $1 \leq m_k \leq \infty$.
- But do the improvement step exactly.
- The algorithm can be written:

$$T_{\mu^k} J_k = T J_k, \quad J_{k+1} = T_{\mu^k}^{m_k} J_k.$$

- Note the special case $m_k = 1 \,\forall\, k$ reduces to standard VI because $J_{k+1} = TJ_k$.
- The special case $m_k = \infty \;\forall\, k$ is PI. So, optimistic PI can be viewed as a "scaling" between PI and VI.

Proposition 11.2 (Optimistic PI)**.** *$J_k \to J^*$ and μ_k is optimal for large enough k.*

Proof. Omitted, but see B&T book. □

11.1.2 Approximate PI, Bounded Max Norm Error

Assumption 11.3. *Suppose both steps of PI are performed with some error bounded in max norm: $\|J_k - J_{\mu^k}\|_\infty \leq \delta$ and $\|T_{\mu^{k+1}} J_k - TJ_k\| \leq \epsilon$.*

Proposition 11.4 (Error bound for API)**.** *This procedure admits the following bound:*

$$\limsup_{k\to\infty} \|J_{\mu^k} - J^*\|_\infty \leq \frac{\epsilon + 2\gamma\delta}{(1-\gamma)^2}.$$

Proof. We have: $T_{\mu^{k+1}} J_k \leq TJ_k + \epsilon e$ and

$$J_{\mu_k} - \delta e \leq J_k \Rightarrow T_{\mu^{k+1}} J_{\mu_k} - \gamma\delta e \leq T_{\mu^{k+1}} J_k \Rightarrow T_{\mu^{k+1}} J_{\mu_k} \leq T_{\mu^{k+1}} J_k + \gamma\delta e,$$

so $T_{\mu^{k+1}} J_{\mu^k} \leq TJ_k + (\epsilon + \gamma\delta)e$. But since $J_k \leq J_{\mu^k} + \delta e \Rightarrow TJ_k \leq TJ_{\mu^k} + \gamma\delta e$, we have

$$T_{\mu^{k+1}} J_{\mu^k} \leq TJ_{\mu^k} + (\epsilon + 2\gamma\delta)e. \qquad (*)$$

Since $TJ_{\mu^k} \leq T_{\mu^k} J_{\mu^k} = J_{\mu^k}$, we also have:

$$T_{\mu^{k+1}} J_{\mu^k} \leq J_{\mu^k} + (\epsilon + 2\gamma\delta)e. \qquad (**)$$

We pause here for a small lemma.

Lemma 11.5. *For $c \geq 0$ and some J such that $T_\mu J \leq J + ce$, we have*

$$J_\mu \leq J + \frac{ce}{1-\gamma}.$$

Proof. Note that if we iteratively apply T_μ:

$$T_\mu^2 J \leq T_\mu J + \gamma ce,\; T_\mu^3 J \leq T_\mu^2 J + \gamma^2 ce, \ldots, T_\mu^l J \leq T_\mu^{l-1} J + \gamma^{l-1} ce.$$

Now, observe that:

$$T_\mu^k J - J = \sum_{l=1}^{k} (T_\mu^l J - T_\mu^{l-1} J)$$

$$\leq \sum_{l=1}^{k} \gamma^{l-1} ce.$$

Letting $k \to \infty$ gives the desired result. □

Applying the lemma to $(**)$, we have

$$J_{\mu^{k+1}} \leq J_{\mu^k} + \frac{\epsilon + 2\gamma\delta}{1-\gamma} e.$$

Next, note that

$$\begin{aligned} J_{\mu^{k+1}} - J^* &= T_{\mu^{k+1}} J_{\mu^{k+1}} - J^* \\ &= T_{\mu^{k+1}} J_{\mu^k} + (T_{\mu^{k+1}} J_{\mu^{k+1}} - T_{\mu^{k+1}} J_{\mu^k}) - J^* \\ &\leq T_{\mu^{k+1}} J_{\mu^k} - J^* + \gamma \frac{\epsilon + 2\gamma\delta}{1-\gamma} e \qquad (\#). \end{aligned}$$

Using

$$\begin{aligned} TJ_{\mu^{k+1}} - J^* &= TJ_{\mu^{k+1}} - TJ^* \\ &\leq \gamma \|J_{\mu^{k+1}} - J^*\| e, \end{aligned}$$

we conclude from $(*)$ that

$$\begin{aligned} T_{\mu^{k+1}} J_{\mu^k} - J^* &\leq TJ_{\mu^k} - J^* + (\epsilon + 2\gamma\delta) e \\ &\leq \gamma \|J_{\mu^k} - J^*\| e + (\epsilon + 2\gamma\delta) e \end{aligned}$$

From $(\#)$,

$$\begin{aligned} J_{\mu^{k+1}} - J^* &\leq \gamma \|J_{\mu^k} - J^*\| e + (\epsilon + 2\gamma\delta) e + \gamma \frac{\epsilon + 2\gamma\delta}{1-\gamma} e \\ &= \gamma \|J_{\mu^k} - J^*\| e + \frac{\epsilon + 2\gamma\delta}{1-\gamma} e \end{aligned}$$

Therefore, $\|J_{\mu^{k+1}} - J^*\| \leq \gamma \|J_{\mu^k} - J^*\| + \dfrac{\epsilon + 2\gamma\delta}{1-\gamma}$.

Take limsup to conclude that: $\limsup\limits_{k\to\infty} \|J_{\mu^k} - J^*\| \leq \dfrac{\epsilon + 2\gamma\delta}{(1-\gamma)^2}$. □

In the limit, we get policies oscillate (or sometimes converge) in a region such that the worst case performance is not too far from J^*. In some cases, policies converge. We now suppose they do. We can show an improved error bound.

Assumption 11.6. *$\mu^{\bar{k}+1} = \mu^{\bar{k}} = \bar{\mu}$ for some k.*

Proposition 11.7 (Error bound when policies converge)**.** *Assume the same errors related to ϵ and δ. We have:*

$$\|J_{\bar{\mu}} - J^*\| \leq \frac{\epsilon + 2\gamma\delta}{1-\gamma}.$$

Proof. Let $\bar{J}$ be an approximate policy evaluation of $\bar{\mu}$. Therefore, $\|\bar{J} - J_{\bar{\mu}}\| \leq \delta$ and by our assumption of convergence, $\|T_{\bar{\mu}}\bar{J} - T\bar{J}\| \leq \epsilon$. Then:

$$\begin{aligned}\|TJ_{\bar{\mu}} - J_{\bar{\mu}}\| &\leq \|TJ_{\bar{\mu}} - T\bar{J} + T\bar{J} - T_{\bar{\mu}}\bar{J} + T_{\bar{\mu}}\bar{J} - J_{\bar{\mu}}\| \\ &\leq \gamma\|J_{\bar{\mu}} - \bar{J}\| + \epsilon + \gamma\|\bar{J} - J_{\bar{\mu}}\| \\ &= \epsilon + 2\gamma\delta\end{aligned}$$

The rest follows as we did many times before. □

11.2 Convergence of API

Can we give conditions under which API converges? There are many ways to do evaluation step, but could lead to oscillation. Consider the following special API algorithm:

- Evaluation: solve $WT_\mu J = J$, with fixed point $\tilde{J}_\mu$. Here we think of W as an approximation step; e.g., $W = \Pi$ corresponds to the basis function case we did previously.

- Improvement: exactly solve $T_{\bar{\mu}}\tilde{J}_\mu = T\tilde{J}_\mu$

Assumption 11.8. *Assume the following:*

1. *$WJ \leq W\bar{J}$ for $J \leq \bar{J}$ (monotonicity)*

2. *For each μ, there is a unique $\tilde{J}_\mu$ such that*

$$\tilde{J}_\mu = \lim_{k\to\infty}(WT_\mu)^k J.$$

The first part implies that $(WT_\mu)(J) \leq (WT_\mu)(\bar{J})$ for each μ and all $J \leq \bar{J}$. The second is a VI-like property.

Theorem 11.9. *This special API converges in finite number of iterations to a policy $\bar{\mu}$. $\tilde{J}_{\bar{\mu}}$, the vector obtained upon termination, is the fixed point of WT.*

Proof. First, note that

$$(WT_{\bar{\mu}})(\tilde{J}_\mu) = (WT)(\tilde{J}_\mu) \leq (WT_\mu)(\tilde{J}_\mu) = \tilde{J}_\mu$$

by monotonicity. Applying $WT_{\bar{\mu}}$ iteratively,

$$\tilde{J}_\mu \geq (WT_{\bar{\mu}})(\tilde{J}_\mu) \geq (WT_{\bar{\mu}})^2 \tilde{J}_\mu \geq \cdots \geq \lim_{k\to\infty} (WT_{\bar{\mu}})^k J = \tilde{J}_{\bar{\mu}}.$$

This is the policy improvement property. Suppose there is no improvement, i.e., $\tilde{J}_\mu = \tilde{J}_{\bar{\mu}}$. Then, by the policy improvement step, $T_{\bar{\mu}}\tilde{J}_{\bar{\mu}} = T\tilde{J}_{\bar{\mu}}$. So $WT_{\bar{\mu}}\tilde{J}_{\bar{\mu}} = WT\tilde{J}_{\bar{\mu}}$ and we conclude by seeing that $WT_{\bar{\mu}}\tilde{J}_{\bar{\mu}} = \tilde{J}_{\bar{\mu}}$. □

11.2.1 Policy Oscillations

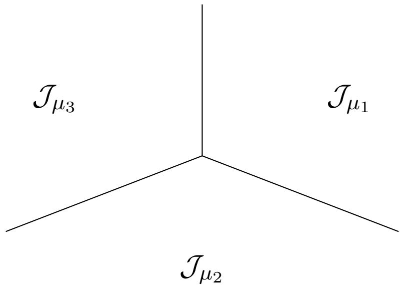
Some intuition about policy oscillations. A *greedy partition* is defined as "all J's such that greedy$(J) = \mu$," i.e.,

$$\begin{aligned} \mathcal{J}_\mu &= \{J : T_\mu J = TJ\} \\ &= \{J : \mu(i) = \text{argmin}_u \{g(i,u) + \mathbf{E}\, J(f(i,u,w))\}\}. \end{aligned}$$

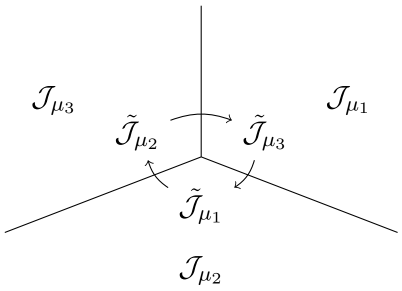
The parameter space $\Re^n$ is partitioned by these greedy sets: $\Re^n = \bigcup_\mu \mathcal{J}_\mu$

Policy iteration:

1. Start with μ_0;
2. Evaluate it to get $\tilde{J}_{\mu_0}$ (unique);
3. Improve by finding μ_1 s.t. $\tilde{J}_{\mu_0} \in \mathcal{J}_{\mu_1}$, and so on.



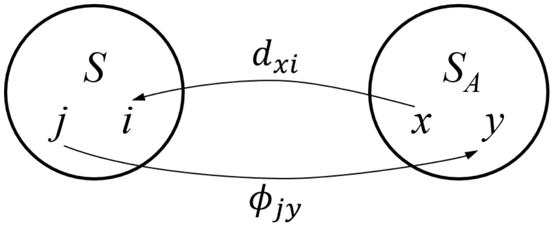
If $\tilde{J}_{\mu_k} \in \mathcal{J}_{\mu_k}$, then it keeps generating μ_k (by definition of $\mathcal{J}_{\mu_k}$). This is convergence. What about oscillation? Having a finite number of policies means we have a finite number of $\tilde{J}_\mu$'s, so oscillation must look like:


$\mathcal{J}_{\mu_3}$
$\mathcal{J}_{\mu_1}$
$\tilde{\mathcal{J}}_{\mu_2}$
$\tilde{\mathcal{J}}_{\mu_3}$
$\tilde{\mathcal{J}}_{\mu_1}$
$\mathcal{J}_{\mu_2}$

Why did the previous theorem get rid of the possibility of oscillation? Note that $\tilde{J}_{\bar{\mu}} \leq \tilde{J}_{\mu}$ where $\bar{\mu}$ is improvement of μ. So $\tilde{J}_{\mu_1} \geq \tilde{J}_{\mu_2} \geq \tilde{J}_{\mu_3} = \tilde{J}_{\mu_3} \cdots$ (because one of them is the optimal policy). Monotonicity assumption gives better error bound plus no oscillation.

Example 11.10 (Case of W that satisfies monotonicity condition is aggregation)**.** *Let S be the state space. Let S_A be the aggregated space. Define the following probability distributions.*

- *Aggregation: Φ_{jy} $\quad j \in S, y \in S_A$ "degree of membership of j in y"*
- *Disaggregation: d_{xi} $\quad i \in S, x \in S_A$ "degree to which x is represented by i"*



Here, $r = DT\Phi r$ and $F = DT\Phi$ is a monotone operator and a contraction (fixed point exists). Thus, API converges.

11.3 Paper Discussion: Finite-time Bounds for Fitted VI

11.3.1 Problem Setup

- Discounted reward MDP, continuous or very large state-space
- Finite number of actions, infinite horizon
- Planning i.e. simulator of the model is available

- Value function approximation
- Using a fitted value iteration algorithm (FVI)
- Objective: Finite-time bounds for FVI
  - Allow a better understanding of RL algorithms and function approximation methods.
  - Develop a theory explaining when and why sampling-based ADP can be expected to perform well.

11.3.2 Definitions

- Space of bounded functions: $B(\mathcal{X})$
- $L^p(\mu)$-norms : μ distribution over $\mathcal{X}, p \geq 1 : \|f\|_{p,\mu} := \Big(\displaystyle\int \|f(x)\|^p \mu(dx) \Big)^{1/p}$
- Space of $L^p(\mu)$-norms bounded functions: $L^p(\mathcal{X}, \mu)$
- $d_{p,\mu}(TV, \mathcal{F}) = \inf\limits_{f \in \mathcal{F}} \|f - TV\|_{p,\mu}$
- $d_{p,\mu}(T\mathcal{F}, \mathcal{F}) = \sup\limits_{f \in \mathcal{F}} d_{p,\mu}(Tf, \mathcal{F})$

11.3.3 Fitted Value Iteration

- If the state space is large or continuous then the value iteration iterates cannot be computed exactly anymore. In this case, we use function approximation.
- Fitted value iteration (Boyan 1995, Gordon 1995, Tsitsiklis & Van Roy 1996)

  $$V_{k+1} = \text{argmin}_{f \in \mathcal{F}} \|f - TV_k\|$$

  Where $\mathcal{F}$ is an appropriate function space and $\|\cdot\|$ is an appropriate norm.

11.3.4 Sampling Based Fitted Value Iteration

- Input: $\mathcal{F}$-function space, N, M, K integers, μ- distribution over the state space.
- Algorithm (stage k):
  1. Sample basis points: $X_1, \ldots, X_N \in \mathcal{X}, X_i \sim \mu$

2. For each action $a \in \mathcal{A}$ and state X_i, sample next states and rewards:

$$Y_j^{X_i,a} \sim P(\cdot \mid X_i, a), R_j^{X_i,a} \sim S(\cdot \mid X_i, a),\ j = 1, \ldots, M$$

3. Calculate the Monte-Carlo approximation of backed up values:

$$\hat{V}(X_i) = \max_{a \in \mathcal{A}} \frac{1}{M} \sum_{j=1}^{M} \left[R_j^{X_i,a} + \gamma V_k(Y_j^{X_i,a})\right],\ i = 1, \ldots, N$$

4. Solve for $V' = V_{k+1}$:

$$V_{k+1} = \underset{f \in \mathcal{F}}{\text{argmin}} \frac{1}{N} \sum_{i=1}^{N} |f(X_i) - \hat{V}(X_i)|^p \tag{11.1}$$

11.3.5 Finite-Sample Bounds

Assumption 11.11 (MDP Regularity). *The MDP (X, A, P, S, γ) satisfies the following conditions: X is a bounded, closed subset of some Euclidean space, A is finite and the discount factor γ satisfies $0 < \gamma < 1$. The reward kernel S is such that the immediate reward function r is a bounded measurable function with bound R_{max}. Further, the support of $S(\cdot \mid x, a)$ is included in $[-\hat{R}_{max}, \hat{R}_{max}]$ independently of $(x, a) \in \mathcal{X} \times \mathcal{A}$*

Assumption 11.12 (Uniformly stochastic transitions). *For all $x \in \mathcal{X}$ and $a \in \mathcal{A}$, assume that $P(\cdot \mid x, a)$ is absolutely continuous w.r.t. μ and the Radon-Nikodym derivative of P w.r.t. μ is bounded uniformly with bound C_μ:*

$$C_\mu := \underset{x \in \mathcal{X}, a \in \mathcal{A}}{sup} \left\| \frac{dP(.|x, a)}{d\mu} \right\|_\infty < +\infty$$

Assumption 11.12 can be written as $P(\cdot \mid x, a) \leq C_\mu \mu(\cdot)$. The noisier the dynamics the smaller the constant C_μ. This assumption certainly excludes deterministic MDPs.

Assumption 11.13 (Discounted-average concentrability of future-state distributions). *Given $\rho, \mu, m \geq 1$ and an arbitrary sequence of stationary policies $\{\pi_m\}_{m \geq 1}$, assume that the future-state distribution $\rho P^{\pi_1} P^{\pi_2} \cdots P^{\pi_m}$ is absolutely continuous w.r.t. μ. Assume that*

$$c(m) := \underset{\pi_1, \ldots, \pi_m}{sup} \left\| \frac{d(\rho P^{\pi_1} P^{\pi_2} \cdots P^{\pi_m})}{d\mu} \right\|_\infty$$

satisfies

$$C_{\rho,\mu} := (1-\gamma)^2 \sum_{m\geq 1} m\gamma^{m-1} c(m) < +\infty$$

$c(m)$ is the m-step concentrability of a future state distribution. $C_{\rho,\mu}$ is the discounted-average concentrability coefficient of the future state distribution. $C_{\rho,\mu}$ is a constant relating how quickly future state distributions can concentrate starting from ρ and relative to μ.

Lemma 11.14. *Consider an MDP satisfying Assumption 11.11, Let $V_{max} = R_{max}/(1-\gamma)$, fix a real number $p \geq 1$, integers $N, M \geq 1, \mu \in M(\mathcal{X})$ and $\mathcal{F} \subset B(\mathcal{X}; V_{max})$. Pick any $V \in B(\mathcal{X}; V_{max})$ and let $V' = V'(V, N, M, \mu, \mathcal{F})$ be defined by equation (11.1). Let $\mathcal{N}_0(N) = \mathcal{N}(\frac{1}{8}(\frac{\varepsilon}{4})^p, \mathcal{F}, N, \mu)$. Then for any $\varepsilon, \delta > 0$,*

$$\|V' - TV\|_{p,\mu} \leq d_{p,\mu}(TV, \mathcal{F}) + \varepsilon$$

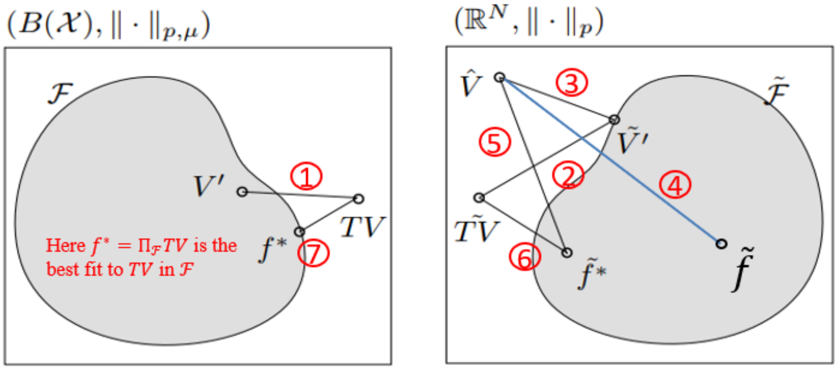
holds w.p. at least $1-\delta$ provided that

$$N > 128\Big(\frac{8V_{max}}{\epsilon}\Big)^{2p}\big(\log(1/\delta) + \log(32\mathcal{N}_0(N))\big)$$

and

$$M > \frac{8(\hat{R}_{max} + \gamma V_{max})^2}{\epsilon^2}\big(\log(1/\delta) + \log(8N|\mathcal{A}|)\big)$$

(Lemma 11.14 Proof illustration)


$(B(\mathcal{X}), \|\cdot\|_{p,\mu})$
$\mathcal{F}$
V'
1
TV
Here $f^* = \Pi_{\mathcal{F}}TV$ is the best fit to TV in $\mathcal{F}$
f^*
7
$(\mathbb{R}^N, \|\cdot\|_p)$
$\hat{V}$
3
$\tilde{\mathcal{F}}$
5
$\tilde{V}'$
2
4
$\tilde{TV}$
6
$\tilde{f}^*$
$\tilde{f}$

Space of bounded functions

Corresponding vector space

If N >> then 1 is close to 2
If M >> then $\tilde{TV}$ is close to $\hat{V}$ } => 2 close to 3

$\tilde{V}'$ is the best fit to $\hat{V}$ in $\tilde{\mathcal{F}}$ then 3 ≤ 4
Choose $\tilde{f} = \tilde{f}^*$ } => 3 ≤ 5

M >> then 5 is close to 6
N >> 6 is close to 7 } => If 7 is small then so is 1

- Lemma 1 however can't be readily used to bound the error committed when approximating TV_k starting from V_k based on a new sample.
- This is because Lemma 1 requires that V, the function whose Bellman image is approximated, is some fixed (non-random) function.
- This problem is easily mitigated in the multi-sample variant of FVI since the samples are independent between iterations.
- This is not the case in the single-sample variant of FVI. Lemma 11.15 gives conditions under which Lemma 11.14 continues to hold.

Lemma 11.15. *Denote by Ω the sample-space underlying the random variables $\{X_i\}$, $\{Y_j^{X_i,a}\}, \{R_j^{X_i,a}\}, i = 1, \dots, N, j = 1, \dots, M, a \in \mathcal{A}$. Then the result of Lemma 11.14 continues to hold if V is a random function satisfying $V(\omega) \in \mathcal{F}, \omega \in \Omega$ provided that*

$$N = O(V_{max}^2(1/\varepsilon)^{2p}\log(\mathcal{N}(c\epsilon, \mathcal{F}_{T-}, N, \mu)/\delta))$$

and

$$M = O((\hat{R}_{max} + \gamma V_{max})^2/\varepsilon^2 \log(N|\mathcal{A}|\mathcal{N}(c'\epsilon, \mathcal{F}, M, \mu)/\delta)),$$

where $c, c' > 0$ are constants independent of the parameters of the MDP and the function space $\mathcal{F}$.

Theorem 11.16. *Consider an MDP satisfying Assumption 11.11 and 11.13. Fix $p \geq 1, \mu \in M(\mathcal{X})$ and let $V_0 \in \mathcal{F} \subset B(\mathcal{X}; V_{max})$. Then for any ε, $\delta > 0$, there exist integers K, M and N such that K is linear in $\log(1/\varepsilon), \log V_{max}$ and $\log(1/(1-\gamma))$, N, M are polynomial in $1/\varepsilon, \log(1/\delta), \log(1/(1-\gamma)), V_{max}, \hat{R}_{max}, \log(|\mathcal{A}|)$, $\log(\mathcal{N}(c\varepsilon(1-\gamma)^2/(C_{\rho,\mu}^{1/p}\gamma), \mathcal{F}, N, \mu))$ for some constant $c > 0$, such that if the multi-sample variant of sampling-based FVI is run with parameters $(N, M, \mu, \mathcal{F})$ and π_k is a policy greedy w.r.t. the K^{th} iterate then w.p. at least $1 - \delta$,*

$$\|V^* - V^{\pi_k}\|_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} C_{\rho,\mu}^{1/p} d_{p,\mu}(T\mathcal{F}, \mathcal{F}) + \varepsilon.$$

If, instead of Assumption 11.13, Assumption 11.12 holds then w.p. at least $1 - \delta$,

$$\|V^* - V^{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} C_{\mu}^{1/p} d_{p,\mu}(T\mathcal{F}, \mathcal{F}) + \varepsilon.$$

Further, the results continue to hold for the single-sample variant of sampling-based FVI with the exception that N depends on $\log(\mathcal{N}(c\varepsilon, \mathcal{F}_{T-}, N, \mu))$ and M depends on $\log(\mathcal{N}(c'\varepsilon, \mathcal{F}, M, \mu))$ for appropriate $c, c' > 0$.

11.3.6 Conclusion

- The authors study continuous or very large state space MDP with generative model of the environment.
- Bounded the error of sample-based FVI with high probability. The bound is function of the concentration coefficient and approximation power and capacity of the underlying function space. Main condition was that the future state distributions do not concentrate fast.
- For a sufficiently rich function space $\mathcal{F}$, FVI will yield a good estimates of V^* as the number of samples and iterations go to infinity.