A Theory of Regularized MDPs

Peng Lingwei

September 16, 2019

Contents

1 Regularized MDPs

1. Regularized function: $\Omega(\pi)$ is strongly convex;

2. Regularized value functions: $V^{\pi,\Omega}(s) = V^{\pi} - \Omega(\pi(s))$

$$\begin{aligned}
V^{\pi,\Omega}(s) =&\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t}(r(S_t,A_t)-(1-\lambda)\Omega(\pi(s)))|S_0=s\right]\\
=&\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t}(r(S_t,A_t))|S_0=s\right]-\sum_{t=0}^{\infty}(1-\gamma)\Omega(\pi(s))\\
=&V^{\pi}(s)-\Omega(\pi(s))
\end{aligned}$$

In MDP, $Q^{\pi}(s,a) = r(s,a) + \gamma\mathbb{E}_{P(s'|s,a)}[V^{\pi}(s')]$. And $V^{\pi} = T^{\pi}V^{\pi} = (\langle \pi(s), Q^{\pi}(s,\cdot)\rangle)_{s\in S}$. Then,let $Q^{\pi,\Omega}(s,a) = r(s,a)+\gamma\mathbb{E}_{P(s'|s,a)}\left[V^{\pi,\Omega}(s')\right]$,

$$V^{\pi,\Omega}(s) = \langle \pi(s), Q^{\pi,\Omega}(s,\cdot)\rangle - (1-\gamma)\Omega(\pi(s))$$

3. Regularized optimal value function: $V^{*,\Omega}(s) = \max_{\pi\in\Pi^{MR}} V^{\pi}(s) - \Omega(\pi(s))$
Let $Q^{*,\Omega}(s,\cdot) = r(s,a) + \gamma\mathbb{E}_{P(s'|s,a)}\left[V^{*,\Omega}(s')\right]$.

$$\begin{aligned}
V^{*,\Omega}(s) =& \max_{\pi\in\Pi^{MR}} V^{\pi}(s) - \Omega(\pi(s))\\
=& \max_{\pi\in\Pi^{MR}} \langle \pi(s), Q^{\pi,\Omega}(s,\cdot)\rangle - (1-\gamma)\Omega(\pi(s))\\
=& \max_{\pi\in\Pi^{MR}} \langle \pi(s), Q^{*,\Omega}(s,\cdot)\rangle - (1-\gamma)\Omega(\pi(s)) \quad \text{(proof is trivial)}\\
=&\Omega^{*}_{\gamma}(Q^{*,\Omega}(s,\cdot))
\end{aligned}$$

where Ω^{*}_{γ} is Legendre-Fenchel transform of $(1-\gamma)\Omega$. More specifically,

$$\forall q_s \in \mathbb{R}^{|A|}, \Omega^{*}_{\gamma}(q_s) = \max_{\pi\in\Pi^{MR}}\langle \pi_s, q_s\rangle - (1-\gamma)\Omega(\pi_s)$$

4. Regularized Bellman operator: $T^{\pi,\Omega}V = T^{\pi}V - (1-\gamma)\Omega(\pi)$

   - Let $Q_V(s,a) = r(s,a) + \gamma\mathbb{E}_{P(s'|s,a)}[V(s')]$,

   $$T^{\pi,\Omega}V(s) = \langle \pi_s, Q_V(s,\cdot)\rangle - (1-\lambda)\Omega(\pi_s)$$

   - Monotonicity:$V_1 \succeq V_2 \Rightarrow T^{\pi,\Omega}V_1 \succeq T^{\pi,\Omega}V_2$

   $$T^{\pi,\Omega}V_1 - T^{\pi,\Omega}V_2 = T^{\pi}V_1 - T^{\pi}V_2 \succeq \vec{0}$$

   - Distributivity: $T^{\pi,\Omega}(V + c\vec{1}) = T^{\pi,\Omega}(V) + \gamma c\vec{1}$

   $$\begin{aligned}
   T^{\pi,\Omega}(V+c\vec{1}) =&T^{\pi}(V+c\vec{1}) - (1-\gamma)\Omega(\pi)\\
   =&T^{\pi}(V) + \gamma c\vec{1} - (1-\gamma)\Omega(\pi) = T^{\pi\Omega}V + \gamma c\vec{1}
   \end{aligned}$$

   - Contraction: $\|T^{\pi,\Omega}V_1 - T^{\pi,\Omega}V_2\|_{\infty} \le \gamma\|V_1 - V_2\|_{\infty}$

   $$\|T^{\pi,\Omega}V_1 - T^{\pi,\Omega}V_2\|_{\infty} = \|T^{\pi}V_1 - T^{\pi}V_2\|_{\infty} \le \gamma\|V_1 - V_2\|_{\infty}$$

- $T^{\pi,\Omega}$'s unique fixed point is $V^{\pi,\Omega}$;

$$
\begin{aligned}
T^{\pi,\Omega}V^{\pi,\Omega} =& T^{\pi}V^{\pi,\Omega} - (1-\gamma)\Omega(\pi) \\
=& T^{\pi}\left(V^{\pi} - \Omega(\pi)\right) - (1-\gamma)\Omega(\pi) \\
=& T^{\pi}(V^{\pi}) - \gamma\Omega(\pi) - (1-\gamma)\Omega(\pi) \\
=& V^{\pi} - \Omega(\pi) = V^{\pi,\Omega}
\end{aligned}
$$

5. Regularized optimal Bellman operator: $T^{*,\Omega}V = \max_{\pi\in\Pi^{MR}} T^{\pi,\Omega}V$;

$$
T^{*,\Omega}V = \max_{\pi\in\Pi^{MR}} \langle \pi_s, Q_V(s,\cdot)\rangle - (1-\lambda)\Omega(\pi_s) = \Omega^{*}_{\gamma}(Q_V(s,\cdot))
$$

- Monotonicity: $V_1 \succeq V_2 \Rightarrow T^{*,\Omega}V_1 \succeq T^{*,\Omega}V_2$.
  Let V_1's optimal policy be π_1, and V_2's be π_2.

$$
\begin{aligned}
T^{*,\Omega}V_1 - T^{*,\Omega}V_2 =& \max_{\pi\in\Pi^{MR}} T^{\pi,\Omega}V_1 - \max_{\pi\in\Pi^{MR}} T^{\pi,\Omega}V_2 \\
\succeq& T^{\pi_2,\Omega}V_1 - T^{\pi_2,\Omega}V_2 \succeq P^{\pi_2}(V_1 - V_2) \succeq \vec{0}
\end{aligned}
$$

- Distributivity: $T^{*,\Omega}(V + c\vec{1}) = T^{*,\Omega}V + \gamma c\vec{1}$.
- Contraction: $\|T^{*,\Omega}V_1 - T^{*,\Omega}V_2\|_\infty \preceq \gamma\|V_1 - V_2\|_\infty$

$$
\|T^{*,\Omega}V_1 - T^{*,\Omega}V_2\|_\infty \leq \|T^{\pi_1,\Omega}V_1 - T^{\pi_1,\Omega}V_2\|_\infty \leq \|T^{\pi_1}V_1 - T^{\pi_1}V_2\|_\infty \leq \gamma\|V_1 - V_2\|_\infty
$$

- $T^{*,\Omega}$'s unique fixed point is $V^{*,\Omega}$. (We talk about sup instead of min)
  First we proof $V \succeq T^{*,\Omega}V \Rightarrow V \succeq V^{*,\Omega}$:

$$
\begin{aligned}
&\forall\pi, \quad V \succeq \sup_{\pi'\in\Pi^{MR}} T^{\pi',\Omega}V \succeq r^{\pi} + \gamma P^{\pi}V - (1-\gamma)\Omega(\pi) \\
\Rightarrow& V \succeq (I - \gamma P^{\pi})(r^{\pi} - (1-\gamma)\Omega(\pi)) = V^{\pi,\Omega} \qquad \Rightarrow V \succeq V^{*,\Omega}
\end{aligned}
$$

  Second we proof $V \preceq T^{*,\Omega}V \Rightarrow V \preceq V^{*,\Omega}$: By definition of sup,

$$
\forall\epsilon, \exists\pi\in\Pi^{MR}, V \preceq T^{\pi,\Omega}V + \epsilon\cdot\vec{1} \Rightarrow V \preceq (I - \lambda P^{\pi})^{-1}[r^{\pi} - (1-\gamma)\Omega(\pi) + \epsilon\cdot\vec{1}]
$$

$$
V \preceq (I - \lambda P^{\pi})^{-1}[r^{\pi} - (1-\gamma)\Omega(\pi)] + \frac{\epsilon}{1-\gamma}\vec{1} \preceq V^{*,\Omega} + \frac{\epsilon}{1-\gamma}\vec{1}
$$

6. Assume that $\Omega_L \leq \Omega \leq \Omega_U$, then $V^{\pi} - \Omega_U \leq V^{\pi,\Omega} \leq V^{\pi} - \Omega_L$.

$$
\max_{\pi\in\Pi^{MR}} V^{\pi} - \Omega_U \leq \max_{\pi\in\Pi^{MR}} V^{\pi,\Omega} \leq \max_{\pi\in\Pi^{MR}} V^{\pi} - \Omega_L \Rightarrow V^{*} - \Omega_U \leq V^{*,\Omega} \leq V^{*} - \Omega_L
$$

   Furthermore,

$$
\begin{aligned}
&V^{*} \leq V^{*,\Omega} + \Omega_U = V^{\pi^{*,\Omega},\Omega} + \Omega_U \leq V^{\pi^{*,\Omega}} + \Omega_U - \Omega_L \\
\Rightarrow& V^{*} - (\Omega_U - \Omega_L) \leq V^{\pi^{*,\Omega}} \leq V^{*}
\end{aligned}
$$

2 Negative entropy

A classical example is the negative entropy $\Omega(\pi_s) = (1-\gamma)^{-1} \sum_a \pi_s(a) \ln \pi_s(a)$.

$$\Omega_\gamma^*(q_s) = \max_{\pi \in \Pi^{MR}} \langle \pi_s, q_s \rangle - \sum_a \pi_s(a) \ln \pi_s(a)$$

We change it into

$$\begin{aligned}
-\Omega_\gamma^*(q_s) &= \min_{\pi_s \succeq \vec{0}} \max_{\alpha \neq 0} \alpha \left(\sum_a \pi_s(a) - 1 \right) - \langle \pi_s, q_s \rangle + \sum_a \pi_s(a) \ln \pi_s(a) \\
&= \max_{\alpha \neq 0} \min_{\pi_s \succeq \vec{0}} \alpha \left(\sum_a \pi_s(a) - 1 \right) - \langle \pi_s, q_s \rangle + \sum_a \pi_s(a) \ln \pi_s(a) \\
&\Rightarrow \alpha - q_s(a) + \ln \pi_s(a) + 1 = 0, \quad \sum_a \pi_s(a) = 1 \\
&\Rightarrow \sum_a \exp\{-1 + q_s(a) - \alpha\} = 1 \Rightarrow \alpha + 1 = \ln \sum_a \exp\{q_s(a)\} \\
&\Rightarrow \pi_s(a) = \frac{\exp\{q_s(a)\}}{\sum_a \exp\{q_s(a)\}}
\end{aligned}$$

$$\Omega_\gamma^*(q_s) = \ln \sum_a \exp q_s(a) \Rightarrow \nabla \Omega_\gamma^*(q_s) = \frac{\exp\{q_s(a)\}}{\sum_a \exp\{q_s(a)\}} = \pi_s^*(a)$$

3 Error Bounds for Approximate Policy Iteration

3.1 KEY BOUND THEOREM

1. $e_k = V_k - V^{\pi_k}$
2. $g_k = V^{\pi_{k+1}} - V^{\pi_k}$
3. $l_k = V^* - V^{\pi_k}$
4. $b_k = V_k - T^{\pi_k} V_k$
5. $\pi_{k+1} = \max_\pi T^\pi V_k$

Target: bound l_k.

Lemma 1.

$$l_{k+1} \preceq \gamma P^{\pi^*} l_k + \gamma \left\{ P^{\pi_{k+1}} (I - \gamma P^{\pi_{k+1}})^{-1} - P^{\pi^*} (I - \gamma P^{\pi_k})^{-1} \right\} b_k$$

$$l_{k+1} \preceq \gamma P^{\pi^*} l_k + \gamma \left\{ P^{\pi_{k+1}} (I - \gamma P^{\pi_{k+1}})^{-1} (I - \gamma P^{\pi_k}) - P^{\pi^*} \right\} e_k$$

Proof.

$$
\begin{aligned}
g_k =& T^{\pi_{k+1}}V^{\pi_{k+1}} - T^{\pi_{k+1}}V^{\pi_k} + T^{\pi_{k+1}}V^{\pi_k} - T^{\pi_{k+1}}V_k \\
&+ T^{\pi_{k+1}}V_k - T^{\pi_k}V_k + T^{\pi_k}V_k - T^{\pi_k}V^{\pi_k} \\
\succeq& \gamma P^{\pi_{k+1}}(V^{\pi_{k+1}} - V^{\pi_k}) + \gamma P^{\pi_{k+1}}(V^{\pi_k} - V_k) + \gamma P^{\pi_k}(V_k - V^{\pi_k}) \\
\succeq& -\gamma(I - \gamma P^{\pi_{k+1}})^{-1}(P^{\pi_{k+1}} - P^{\pi_k})e_k
\end{aligned}
$$

$$
\begin{aligned}
e_k - g_k \preceq& \left[I + \gamma(I - \gamma P^{\pi_{k+1}})^{-1}(P^{\pi_{k+1}} - P^{\pi_k})\right] e_k \\
=& (I - \gamma P^{\pi_{k+1}})^{-1}(I - \gamma P^{\pi_k})e_k
\end{aligned}
$$

$$
\begin{aligned}
l_{k+1} =& T^{\pi^*}V^* - T^{\pi^*}V^{\pi_k} + T^{\pi^*}V^{\pi_k} - T^{\pi^*}V_k + T^{\pi^*}V_k - T^{\pi_{k+1}}V_k \\
&+ T^{\pi_{k+1}}V_k - T^{\pi_{k+1}}V^{\pi_k} + T^{\pi_{k+1}}V^{\pi_k} - T^{\pi_{k+1}}V^{\pi_{k+1}} \\
\preceq& \gamma P^{\pi^*}(V^* - V^{\pi_k}) + \gamma P^{\pi^*}(V^{\pi_k} - V_k) + \gamma P^{\pi_{k+1}}(V_k - V^{\pi_k}) + \gamma P^{\pi_{k+1}}(V^{\pi_k} - V^{\pi_{k+1}}) \\
=& \gamma P^{\pi^*} l_k + \gamma(P^{\pi_{k+1}} - P^{\pi^*})e_k - \gamma P^{\pi_{k+1}} g_k \\
\preceq& \gamma P^{\pi^*} l_k + \gamma \left\{P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1}(I - \gamma P^{\pi_k}) - P^{\pi^*}\right\} e_k
\end{aligned}
$$

For $(I - \gamma P^{\pi_k})e_k = b_k$,

$$
l_{k+1} \preceq \gamma P^{\pi^*} l_k + \gamma \left\{P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1} - P^{\pi^*}(I - \gamma P^{\pi_k})^{-1}\right\} b_k
$$

□

Theorem 1.

$$
\limsup_{k\to\infty} \|V^* - V^{\pi_k}\|_{\mu_0} \le \limsup_{k\to\infty} \gamma\mu_0(I - \gamma P^{\pi^*})^{-1} \left\{P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1} + P^{\pi^*}(I - \gamma P^{\pi_k})^{-1}\right\} |b_k|
$$

$$
\limsup_{k\to\infty} \|V^* - V^{\pi_k}\|_{\mu_0} \le \limsup_{k\to\infty} \gamma\mu_0(I - \gamma P^{\pi^*})^{-1} \left\{P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1}(I + \gamma P^{\pi_k}) + P^{\pi^*}\right\} |e_k|
$$

After normalization,let

$$
Q_k = \frac{(1-\gamma)^2}{2}(I - \gamma P^{\pi^*})^{-1} \left\{P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1} + P^{\pi^*}(I - \gamma P^{\pi_k})^{-1}\right\},
$$

and

$$
\tilde{Q}_k = \frac{(1-\gamma)^2}{2}(I - \gamma P^{\pi^*})^{-1} \left\{P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1}(I + \gamma P^{\pi_k}) + P^{\pi^*}\right\}.
$$

Then, write $\mu_k = \mu_0 Q_k$ *and* $\tilde{\mu}_k = \mu_0 \tilde{Q}_k$*, we have*

$$
\limsup_{k\to\infty} \|V^* - V^{\pi_k}\|_{\mu_0} \le \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} \|V_k - T^{\pi_k}V_k\|_{\mu_k}
$$

$$
\limsup_{k\to\infty} \|V^* - V^{\pi_k}\|_{\mu_0} \le \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} \|V_k - V^{\pi_k}\|_{\tilde{\mu}_k}
$$

3.2 APPROXIMATE POLICY EVALUATION

3.2.1 Linear Feature-based approximation

1. Monte-Carlo simulations and regression: $\min_\theta \|\Phi\theta - V^{\pi_k}\|^2_{\rho_k}$;
2. Minimal quadratic residual solution: $\min_\theta \|V_\theta - T^{\pi_k} V_\theta\|^2_{\rho_k}$;

$$A\theta = b \ with \ \begin{cases} A = \Phi^T (I - \gamma P^{\pi_k})^T D_{\rho_k} (I - \gamma P^{\pi_k})\Phi \\ b = \Phi^T (I - \gamma P^{\pi_k})^T D_{\rho_k} r^{\pi_k} \end{cases}$$

3. Temporal Difference solution: $\min_\theta \|V_\theta - \Pi_{\pi_k} T^{\pi_k} V_\theta\|^2_{\rho_k}$. For $TD(0)$:

$$A\theta = b \ with \ \begin{cases} A = \Phi^T D_{\rho_k} (I - \gamma P^{\pi_k})\Phi \\ b = \Phi^T D_{\rho_k} r^{\pi_k} \end{cases}$$

Because these method depends on the distribution ρ_k used in the minimization problem, which usually depends on the policy π_k, therefore we have to consider the choice of ρ_k.

- Steady-state distribution $\bar{\rho}_{\pi_k}$: $\bar{\rho}_{\pi_k} = \bar{\rho}_{\pi_k} P^{\pi_k}$;
- Constant distribution ρ_0;
- Mixed distribution $\rho^\lambda_{\pi_k} = \rho_0 (I - \lambda P^{\pi_k})^{-1}(1 - \lambda)$;
- Convex combination mixed distribution: $\rho^\delta_{\pi_k} = (1 - \delta)\rho_0 + \delta \bar{\rho}_{\pi_k}$.

Assumption 1.

$$\inf_\theta \|V_\theta - V^\pi\|_{\rho_\pi} \leq \epsilon$$

3.2.2 The Quadratic Residual Soluction

$$\|V_k - T^{\pi_k} V_k\|_{\rho_k} = \inf_\theta \|V_\theta - T^{\pi_k} V_\theta\|_{\rho_k} = \inf_\theta \|(I - \gamma P^{\pi_k})(V_\theta - V^{\pi_k})\|_{\rho_k} \leq \|I - \gamma P^{\pi_k}\|_{\rho_k} \epsilon$$

$$\|V_k - T^{\pi_k} V_k\|^2_{\mu_k} \leq \|\mu_k / \rho_k\|_\infty \|V_k - T^{\pi_k} V_k\|^2_{\rho_k}$$

So we need a new assumption.

Assumption 2.

$$\forall \pi, \exists \mu, C, have \ P^\pi(i,j) \leq C\mu(j).$$

If $\bar{\mu}(j) = 1/N$ and $C = N$, it always satisfies. However, we are actually interested in finding a constant $C \ll N$.

Lemma 2. *In preceeding section, $\mu_k = \mu_0 Q_k$. If assumption2 exists, we have $\mu_k \leq C\mu$.*

Proof. $(P_1 P_2)(i,j) = \sum_k P_1(i,k) P_2(k,j) \leq C\mu(j) \sum_k P_1(i,k) = C\mu(j)$. So $Q_k(i,j) \leq C\mu(j) \Rightarrow \mu_k(j) \leq C\mu(j)$ □

Theorem 2. *Assume two assumption hold with some distribution μ_0 and C.*

- $\rho^{\lambda}_{\pi_k} = \mu_0(I - \lambda P^{\pi_k})^{-1}(1-\lambda)$, *then*

$$\limsup \|V^* - V^{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\sqrt{\frac{C}{1-\lambda}}\left(1 + \gamma\sqrt{\min\left(\frac{C}{1-\lambda}, \frac{1}{\lambda}\right)}\right)\epsilon$$

- $\rho^{\delta}_{\pi_k} = (1-\delta)\mu_0 + \delta\bar{\rho}_{\pi_k}$.

$$\limsup \|V^* - V^{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\sqrt{\frac{C}{1-\delta}}(1+\gamma\sqrt{C})\epsilon$$

Proof. 1. $\rho^\lambda_k \succeq (1-\lambda)\mu_0$ and $\rho^\delta_k \geq (1-\delta)\mu_0$.

2. $\|P^{\pi_k}\|^2_{\rho^\lambda_k} \leq \min\left(\frac{C}{1-\lambda}, \frac{1}{\lambda}\right)$:

$$\|P^{\pi_k}h\|^2_{\rho^\lambda_k} = \rho^\lambda_k (P^{\pi_k}h)^2 \leq \rho^\lambda_k P^{\pi_k} h^2 \leq C\mu_0 h^2 \leq \frac{C}{1-\lambda}\rho^\lambda_k h^2 = \frac{C}{1-\lambda}\|h\|^2_{\rho^\lambda_k}$$

$$\begin{aligned}\|P^{\pi_k}h\|^2_{\rho^\lambda_k} =&(1-\lambda)\mu_0\sum_{t=0}^{\infty}\lambda^t (P^{\pi_k})^t P^{\pi_k} h^2 \leq (1-\lambda)\mu_0 \sum_{t=0}^{\infty}\lambda^t (P^{\pi_k})^{t+1}h^2 \\ =&\frac{1-\lambda}{\lambda}\mu_0\left\{\sum_{t=0}^{\infty}\lambda^t(P^{\pi_k})^t h^2 - h^2\right\} \leq \frac{1}{\lambda}\rho^\lambda_k h^2 = \frac{1}{\lambda}\|h\|^2_{\rho^\lambda_k}\end{aligned}$$

3. $\|P^{\pi_k}\|^2_{\rho^\delta_k} \leq C$.

$$\begin{aligned}\|P^{\pi_k}h\|^2_{\rho^\delta_k} =&\rho^\delta_k (P^{\pi_k h})^2 \leq (1-\delta)\mu_0 P^{\pi_k}h^2 + \delta\bar{\rho}_{\pi_k}P^{\pi_k}h^2 \leq C(1-\delta)\mu_0 h^2 + \delta\bar{\rho}_k h^2 \\ =&C(\rho^\delta_k - \delta\bar{\rho}_k)h^2 + \delta\bar{\rho}_k h^2 \leq C\rho^\delta_k h^2\end{aligned}$$

4.

$$\begin{aligned}\limsup_{k\to\infty}\|l_k\|_{\mu_0} \leq& \frac{2\gamma}{(1-\gamma)^2}\limsup_{k\to\infty}\sqrt{\|\mu_k/\rho_k\|_\infty}\|I - \gamma P^{\pi_k}\|_{\rho_{\pi_k}}\epsilon \\ \leq&\frac{2\gamma}{(1-\gamma)^2}\limsup_{k\to\infty}\sqrt{\|\mu_k/\rho_k\|_\infty}\left(1+\gamma\|P^\pi\|_{\rho_{\pi_k}}\right)\epsilon\end{aligned}$$

□

3.2.3 Temporal Difference Solution

1.

$$\begin{aligned}&(I - \gamma\Pi_{\pi_k}P^{\pi_k})(V_k - V^{\pi_k}) = V_k - \gamma\Pi_{\pi_k}P^{\pi_k}V_k - V^{\pi_k} + \gamma\Pi_{\pi_k}P^{\pi_k}V^{\pi_k} \\ =& - V^{\pi_k} + \Pi_{\pi_k}(r^{\pi_k} + \gamma P^{\pi_k}V^{\pi_k}) = \Pi_{\pi_k}V^{\pi_k} - V^{\pi_k} := \epsilon'_k\end{aligned}$$

I lose my patience again.

4 Finite-Time Bounds for Fitted Value Iteration

4.1 Approximating the Bellman Operator

1. Monte-Carlo estimate of TV_k:

$$\hat{V}(s) = \max_{a \in A} \frac{1}{M} \sum_{j=1}^{M} \left[R_j(s,a) + \gamma V_k(s'_j) \right], s = 1, 2, \ldots, N$$

$$V_{k+1} = \arg\min_{f \in \mathcal{F}} \|f - \hat{V}\|_p$$

2.

$$\begin{aligned} \mathbb{E}\left[\hat{V}(s)\right] =& \mathbb{E}\left[\max_{a \in A} \frac{1}{M} \sum_{j=1}^{M} \left[R_j(s,a) + \gamma V_k(s'_j) \right]\right] \\ &\geq \max_{a \in A} \mathbb{E}\left[\frac{1}{M} \sum_{j=1}^{M} \left[R_j(s,a) + \gamma V_k(s'_j) \right]\right] = TV_k \end{aligned}$$

3. Bound $\mathbb{P}\left\{\|\hat{V} - T^{\pi_k} V_k\|_\infty \geq \epsilon\right\} \leq \delta$

   Proof.

$$\mathbb{P}\left\{\|V^{\pi_k} - T^{\pi_k} V_k\|_\infty \geq \epsilon\right\} \leq 2e^{-\frac{2M\epsilon^2}{(R_{\max} + \gamma V_{\max})^2}}$$

   It's easy to find function $M \geq C_M(\epsilon, \delta)$, which guarantees

$$\mathbb{P}\left\{\|V^{\pi_k} - T^{\pi_k} V_k\|_\infty \geq \epsilon\right\} \leq \delta$$

□

5 Regularized Modified Policy Iteration

Wait