A Theory of Regularized MDPs

Peng Lingwei

September 18, 2019

Contents

1 Regularized MDPs

1. Regularized function: $\Omega(\pi)$ is strongly convex;
2. Regularized value functions: $V^{\pi,\Omega}(s) = V^{\pi} - \Omega(\pi(s))$

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

In MDP, $Q^{\pi}(s,a) = r(s,a) + \gamma\mathbb{E}_{P(s'|s,a)}\left[V^{\pi}(s')\right]$. 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 $Q_V(s,a) = r(s,a) + \gamma\mathbb{E}_{P(s'|s,a)}\left[V(s')\right]$,

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

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

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

- $T^{\pi,\Omega}$'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 $V \succeq T^{*,\Omega}V \Rightarrow V \succeq V^{*,\Omega}$:

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

  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 Regularized Modified Policy Iteration

Definition 1. ***(Regularized modified policy iteration).***

$$\pi_{k+1} = \arg\max_{\pi_k\in\Pi^{MR}} T_{\pi,\Omega}V_k, \quad V_{k+1} = T^m_{\pi_{k+1},\Omega}V_k$$

Related algorithms:

1. Soft Q-learning: $\hat{q}_{k+1}(s,a) = r(s,a) + \gamma\hat{\mathbb{E}}_{s'|s,a}\left[\Omega^*(q_k(s',\cdot))\right], J(\theta) = \mathbb{E}\left[\|\hat{q}_{k+1} - q_\theta\|_2^2\right]$
2. SAC: $\hat{\pi}_{k+1}(\cdot|s) = \nabla\Omega^*(q_k(s,\cdot)), \quad J(w) = \hat{\mathbb{E}}\left[KL(\pi_w(\cdot|s_i)\|\hat{\pi}_{k+1}(\cdot|s))\right]$

3.1 Analysis

Two errors is introduced in AMPI.

- We only can get ϵ'_{k+1}-optimal policy π'_{k+1}: $T_{\pi_{k+1},\Omega}V_k \preceq T_{\hat{\pi}_{k+1},\Omega}V_k + \epsilon'_{k+1}$;
- $V_{k+1} = T^m_{\hat{\pi}_{k+1},\Omega}V_k + \epsilon_{k+1}$.

We want bound $l_{k,\Omega} = V^{*,\Omega} - V^{\pi_k,\Omega}$. We also denote $d_k = V^{*,\Omega} - V_k$ and $b_k = V_k - T^{\pi_{k+1},\Omega}V_k$.

Denote $\frac{1}{q} + \frac{1}{q'} = 1$, and

$$C^i_q = \frac{1-\gamma}{\gamma^i}\sum_{j=i}^{\infty}\gamma^j \max_{\pi_1,\dots,\pi_j} \|\frac{\rho P_{\pi_1}P_{\pi_2}\dots P_{\pi_j}}{\mu}\|_{q,\mu}$$

- $l_{k,\Omega} \leq 2\sum_{i=1}^{k-1}\sum_{j=i}^{\infty}\Gamma^j\left|\epsilon_{k-i}\right| + \sum_{i=0}^{k-1}\sum_{j=i}^{\infty}\Gamma^i\left|\epsilon'_{k-i}\right| + 2\sum_{j=k}^{\infty}\Gamma^j\min\{\left|d_0\right|, \left|b_0\right|\}$;
- $\|l_{k,\Omega}\|_{p,\rho} \leq 2\sum_{i=1}^{k-1}\frac{\gamma^i}{1-\gamma}(C_q^i)^{\frac{1}{p}}\|\epsilon_{k-i}\|_{pq',\mu} + \sum_{i=0}^{k-1}\frac{\gamma^i}{1-\gamma}(C_q^i)^{\frac{1}{p}}\|\epsilon'_{k-i}\|_{pq',\mu} + \frac{2\gamma^k}{1-\gamma}(C_q^i)^{\frac{1}{p}}\min(\|d_0\|_{pq',\mu}, \|b_0\|_{pq',\mu})$

The bound does no explain the good empirical results of related algorithms.

4 Mirror Descent Modified Policy Iteration

- $\Omega_{\pi'_s}(\pi_s) = D_\Omega(\pi_s\|\pi'_s) = \Omega(\pi_s) - \Omega(\pi'_s) - \langle\nabla\Omega(\pi'_s), \pi_s - \pi'_s\rangle$;
- $\pi_{k+1} = \arg\max_\pi\langle q_k, \pi\rangle - D_\Omega(\pi\|\pi_k)$;

Definition 2. ***(Mirror Descent MPI).***

1. *Type1:* $\pi_{k+1} = \arg\max_\pi T_{\pi,\Omega_{\pi_k}}V_k, V_{k+1} = T^m_{\pi_{k+1},\Omega_{\pi_k}}V_k$*;*
2. *Type2:* $\pi_{k+1} = \arg\max_\pi T_{\pi,\Omega_{\pi_k}}V_k, V_{k+1} = T^m_{\pi_{k+1},\Omega_{\pi_{k+1}}}V_k = T^m_{\pi_{k+1}}V_k$*.*

The tool used to analyse error bound is very complicated.

5 Error Bounds for Approximate Policy Iteration

5.1 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.

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

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

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*

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

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

5.2 APPROXIMATE POLICY EVALUATION

5.2.1 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: $\min_\theta \|V_\theta - T^{\pi_k} V_\theta\|^2_{\rho_k}$;

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

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.

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

5.2.2 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. $(P_1P_2)(i,j) = \sum_k P_1(i,k)P_2(k,j) \leq C\mu(j)\sum_k P_1(i,k) = C\mu(j)$. 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. $\|P^{\pi_k}\|^2_{\rho_k^\lambda} \leq \min\left(\frac{C}{1-\lambda}, \frac{1}{\lambda}\right)$:

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

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

3. $\|P^{\pi_k}\|^2_{\rho_k^\delta} \leq C$.

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

4.

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

□

5.2.3 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.

6 Finite-Time Bounds for Fitted Value Iteration

6.1 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.

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

3. Condition of $\mathbb{P}\left\{\|\hat{V}-TV\|_p \leq \epsilon\right\} \geq 1-\delta$

   Proof.

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

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

   $$\mathbb{P}\left\{\max_\pi \|\hat{V}-\mathbb{E}\hat{V}\|_\infty \geq \epsilon\right\} \leq \delta$$

   Because $\max_x f(x) - \max_x g(x) = f(x_f) - g(x_g) \leq f(x_f) - g(x_f) \leq \max_x(f(x)-g(x))$, therefore

   $$\|TV-\hat{V}\|_p \leq \|TV-\hat{V}\|_\infty \leq \max_\pi \|\mathbb{E}\hat{V}-\hat{V}\|_\infty$$

   □

4. Condition of $\mathbb{P}\left\{\sup_{f\in\mathcal{F}} \left|\|f-TV\|_{p,\mu} - \|f-TV\|_{p,\hat{\mu}}\right| \leq \epsilon\right\} \geq 1-\delta$, where $\|f\|_{p,\hat{\mu}}^p = \frac{1}{N}\sum_{i=1}^{N}|f_i|^p$. ($\hat{\mu}$ is sample distribution.) We can use Rademacher complexities to get function $N \geq C_N(\epsilon,\delta)$ to guarantees these. The paper has some problem here, so I skip the proof.

   - Rademacher complexity;
   - Covering numbers.

5. We need bound $\mathbb{P}\left\{\|V_{k+1}-TV_k\|_{p,\mu} \leq \epsilon\right\} \geq 1-\delta$. (The preceeding conditions are sufficient.)

   $$\|V_{k+1}-TV_k\|_{p,\mu} \leq \|V_{k+1}-TV_k\|_{p,\mu} + \epsilon \leq \|V_{k+1}-\hat{V}\|_{p,\hat{\mu}} + 2\epsilon \leq \inf_f \|f-\hat{V}\|_{p,\hat{\mu}} + 2\epsilon$$
   $$\leq \inf_f \|f-TV\|_{p,\hat{\mu}} + 3\epsilon \leq \inf_f \|f-TV\|_{p,\mu} + 4\epsilon$$

6.2 MAIN RESULT

- Single-sample: $V_{k+1} = \arg\min_{f\in\mathbb{F}} \sum_{i=1}^{N}\left|f(s_i) - \max_{a\in A}\frac{1}{M}\sum_{j=1}^{M}\left[R_j(s_i,a)+\gamma V_k(s'_j)\right]\right|^p$
- Multi-sample: $V_{k+1} = \arg\min_{f\in\mathbb{F}} \sum_{i=1}^{N}\left|f(s_i^k) - \max_{a\in A}\frac{1}{M}\sum_{j=1}^{M}\left[R_j(s_i^k,a)+\gamma V_k(s_j'^k)\right]\right|^p$

We want bound $L_k = \|V^* - V^{\pi_k}\|_{p,\rho}$.

I lost my patience.

7 Approximate Modified Policy Iteration

1. Modified policy ieration:$\pi_{k+1} = \arg\max_\pi T^\pi v_k, v_{k+1} = (T^{\pi_{k+1}})^m v_k$.
2. $c_q(m) = \max_{\pi_1,\ldots,\pi_m} \|\frac{d(\rho P^{\pi_1} P^{\pi_2} \ldots P^{\pi_m})}{d\mu}\|_{q,\mu}$

7.1 Approximate MPI Algorithms

1. AMPI-V:
   - $\pi_{k+1}(s) \in \arg\max_{a\in A} \frac{1}{M}\sum_{j=1}^{M}(r^{(j)}(s,a) + \gamma v_k(s_a^{(j)}))$;
   - $\hat{v}_{k+1}(s^{(i)}) = \sum_{t=0}^{m-1}\gamma^t r_t^{(i)} + \gamma^m v_k(s_m^{(i)}), i = 1, 2, \ldots, N$;
   - Empirical error: $\hat{L}_k^{\mathcal{F}}(\hat{\mu}; v) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{v}_{k+1}(s^{(i)}) - v_{k+1}(s^{(i)})\right)^2$, which is used to get v_{k+1} with any regression algorithm;
   - True error: $L_k^{\mathcal{F}}(\mu; v) = \|T_{\pi_{k+1}}^m v_k - v\|_{2,\mu}^2 = \int \left(T_{\pi_{k+1}}^m v_k(s) - v(s)\right)^2 \mu(ds)$
2. AMPI-Q:
   - $\pi_{k+1}(s) \in \arg\max_{a\in A} Q_k(s,a)$;
   - $\hat{Q}_{k+1}(s^{(i)}, a^{(i)}) = \sum_{t=0}^{m-1}\gamma^t r_t^{(i)} + \gamma^m Q_k(s_m^{(i)}, a_m^{(i)})$;
   - Empirical error: $\hat{L}_k^{\mathcal{F}}(\hat{\mu}; Q) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{Q}_{k+1}(s^{(i),a^{(i)}} - Q(s^{(i)}, a^{(i)}))\right)^2$ (regression).
   - True error: $L_k^{\mathcal{F}}(\mu; Q) = \|T_{\pi_{k+1}}^m Q_k - Q\|_{2,\mu}^2 = \int \left(T_{\pi_{k+1}}^m Q_k(s,a) - Q(s,a)\right)^2 \mu(dsda)$.
3. Classification-Based MPI:
   - Rewrite $v_k = T_{\pi_k}^m v_{k-1}, \pi_{k+1} = \arg\max_\pi T^\pi(T_{\pi_k}^m v_{k-1})$;
   - $\hat{v}_k(s^{(i)}) = \sum_{t=0}^{m-1}\gamma^t r_t^{(i)} + \gamma^m v_{k-1}(s_m^{(i)})$;
   - $\hat{L}_k^{\mathcal{F}}(\hat{\mu}; v) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{v}_k(s^{(i)}) - v(s^{(i)})\right)^2$; (regression)
   - $L_k^{\mathcal{F}}(\mu; v) = \|T_{\pi_k}^m v_{k-1} - v\|_{2,\mu}^2 = \int \left(T_{\pi_k}^m v_{k-1}(s) - v(s)\right)^2 \mu(ds)$;
   - $\hat{Q}_k(s^{(i)}, a) = \frac{1}{M}\sum_{j=1}^{M} R_k^j(s^{(i)}, a), R_k^j(s^{(i)}, a) = \sum_{t=0}^{m}\gamma^t r_t^{(i,j)} + \gamma^{m+1} v_{k-1}(s_{m+1}^{(i,j)})$;
   - $\hat{L}_k^{\Pi}(\hat{\mu}; \pi) = \frac{1}{N'}\sum_{i=1}^{N'}\left[\max_{a\in A}\hat{Q}_k(s^{(i)}, a) - \hat{Q}_k(s^{(i)}, \pi(s^{(i)}))\right]$ (classification)

7.2 Error Propagation

General error:

- Greedy step error: $\pi_k = \hat{G}_{\epsilon_k'} v_{k-1} \Rightarrow \forall \pi', T_{\pi'} v_{k-1} \preceq T_{\pi_k} v_{k-1} + \epsilon_k'$;
- Evaluation step error: $v_k = T_{\pi_k}^m v_{k-1} + \epsilon_k$

Errors parameters:

- $d_k = V^* - T_{\pi_k}^m V_{k-1} = V^* - (V_k - \epsilon_k)$;

- $s_k = T^m_{\pi_k} V_{k-1} - V^{\pi_k} = (V_k - \epsilon_k) - V^{\pi_k}$;
- $b_k = V_k - T_{\pi_{k+1}} V_k$;
- $l_k = V^* - V^{\pi_k} = d_k + s_k$. (We want bound this.)

1. Bounding b_k:

$$\begin{aligned}
b_k =& V_k - \epsilon_k - T_{\pi_k}(v_k - \epsilon_k) + \epsilon_k - \gamma P_{\pi_k}\epsilon_k + T_{\pi_k} V_k - T_{\pi_{k+1}} V_k \\
\preceq& T^m_{\pi_k} V_{k-1} - T_{\pi_k} T^m_{\pi_k} V_{k-1} + (I - \gamma P_{\pi_k})\epsilon_k + \epsilon'_{k+1} \\
=& (\gamma P_{\pi_k})^m (V_{k-1} - T_{\pi_k} V_{k-1}) + (I - \gamma P_{\pi_k})\epsilon_k + \epsilon'_{k+1} \\
=& (\gamma P_{\pi_k})^m b_{k-1} + ((I - \gamma P_{\pi_k})\epsilon_k + \epsilon'_{k+1}) \\
=& (\gamma P_{\pi_k})^m b_{k-1} + x_k
\end{aligned}$$

2. Bounding d_k:

$$\begin{aligned}
d_{k+1} =& V^* - T^m_{\pi_{k+1}} V_k \\
=& T_{\pi^*} V^* - T_{\pi^*} V_k + T_{\pi^*} V_k - T_{\pi_{k+1}} V_k + \sum_{j=1}^{m-1} \left[T^j_{\pi_{k+1}} V_k - T^{j+1}_{\pi_{k+1}} V_k \right] \\
\preceq& \gamma P_{\pi^*}(V^* - V_k) + \epsilon'_{k+1} + \sum_{j=1}^{m-1} (\gamma P_{\pi_{k+1}})^j b_k \\
=& \gamma P_{\pi^*}(V^* - T^m_{\pi_k} V_{k-1} - \epsilon_k) + \epsilon'_{k+1} + \sum_{j=1}^{m-1} (\gamma P_{\pi_{k+1}})^j b_k \\
=& \gamma P_{\pi^*} d_k + (-\gamma P_{\pi^*}\epsilon_k + \epsilon'_{k+1}) + \sum_{j=1}^{m-1} (\gamma P_{\pi_{k+1}})^j b_k \\
=& \gamma P_{\pi^*} d_k + y_k + \sum_{j=1}^{m-1} (\gamma P_{\pi_{k+1}})^j b_k
\end{aligned}$$

3. Bounding s_k:

$$\begin{aligned}
s_k =& T^m_{\pi_k} V_{k-1} - V_{\pi_k} = T^m_{\pi_k} V_{k-1} - T^\infty_{\pi_k} V_{k-1} = (\gamma P_{\pi_k})^m \left(V_{k-1} - T^\infty_{\pi_k} V_{k-1} \right) \\
=& (\gamma P_{\pi_k})^m \sum_{j=0}^{\infty} \left[T^j_{\pi_k} V_{k-1} - T^{j+1}_{\pi_k} V_{k-1} \right] = (\gamma P_{\pi_k})^m \sum_{j=0}^{\infty} (\gamma P_{\pi_k})^j (V_{k-1} - T_{\pi_k} V_{k-1}) \\
=& (\gamma P_{\pi_k})^m (I - \gamma P_{\pi_k})^{-1} b_{k-1}
\end{aligned}$$

Definition 3. *We define $\mathbb{P}_n$ as the smallest set of discounted transition kernels that are defined as follows:*

1. $\forall \{\pi_1, \dots, \pi_n\}, (\gamma P_{\pi_1})(\gamma P_{\pi_2}) \dots (\gamma P_{\pi_n}) \in \mathbb{P}_n$;
2. $\forall \alpha \in [0, 1]$ *and* $P_1, P_2 \in \mathbb{P}_n$, *we have* $\alpha P_1 + (1 - \alpha) P_2 \in \mathbb{P}_n$.

And we denote any element of $\mathbb{P}_n$, Γ^n.

1. $b_k \leq \sum_{i=1}^{k} \Gamma^{m(k-i)} x_i + \Gamma^{mk} b_0$;

2. $d_k \leq \sum_{i=0}^{k-1} \Gamma^{k-1-i} \left(y_i + \sum_{l=1}^{m-1} \Gamma^l b_i \right) + \Gamma^k d_0$;

$$d_k \leq \sum_{i=1}^{k} \Gamma^{i-1} y_{k-i} + \sum_{i=1}^{k-1} \sum_{j=i}^{mi-1} \Gamma^j x_{k-i} + \sum_{i=k}^{mk-1} \Gamma^i b_0 + \Gamma^k d_0.$$

3. $s_k = \Gamma^m \sum_{i=0}^{\infty} \Gamma^i b_{k-1} = \sum_{i=0}^{\infty} \Gamma^{m+i} b_{k-1} = \sum_{i=1}^{k-1} \sum_{j=mi}^{\infty} \Gamma^j x_{k-i} + \sum_{j=mk}^{\infty} \Gamma^j b_0$.

4. $l_k \leq \sum_{i=1}^{k} \Gamma^{i-1} y_{k-i} + \sum_{i=1}^{k-1} \sum_{j=i}^{\infty} \Gamma^j x_{k-i} + \sum_{j=k}^{\infty} \Gamma^j b_0 + \Gamma^k d_0$.

Lemma 3. *1. After k iterations, the losses of AMPI-V and AMPI-Q satisfy*

$$l_k \leq 2 \sum_{i=1}^{k-1} \sum_{j=i}^{\infty} \Gamma^j |\epsilon_{k-i}| + \sum_{i=0}^{k-1} \sum_{j=i}^{\infty} \Gamma^j \left|\epsilon'_{k-i}\right| + h(k);$$

2. The loss of CBMPI satisfies:

$$l_k \leq 2 \sum_{i=1}^{k-2} \sum_{j=i+m}^{\infty} \Gamma^j |\epsilon_{k-i-1}| + \sum_{i=0}^{k-1} \sum_{j=i}^{\infty} \Gamma^j \left|\epsilon'_{k-i}\right| + h(k);$$

3. $h(k) = 2 \sum_{j=k}^{\infty} \Gamma^j |d_0|$ *or* $h(h) = 2 \sum_{j=k}^{\infty} \Gamma^j |b_0|$.

It's easy to obtain $\limsup_{k\to\infty} \|l_k\|_\infty \leq \frac{2\gamma\epsilon+\epsilon'}{(1-\gamma)^2}$