Action-Gap Phenomenon in Reinforcement Learning

Amir-massoud Farahmand∗ School of Computer Science, McGill University Montreal, Quebec, Canada 2011 Neurips

Contributions

- Smaller Performance loss than the estimated optimal action-value function with action gap regularity
- action gap regularity affects approximate policy iteration algorithms

Markov Chain

- limiting distribution
- stationary distribution
- irreducible
- aperiodic
- ergodicity
- mixing time

Limiting Distribution

$$1.\ \pi_j = lim_{n\to\infty} P_{ij}^n$$
$$2.\ \Sigma_j \pi_j = 1$$

$$T = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0.1 & 0.9 \\ 0.6 & 0.4 & 0 \end{bmatrix} \qquad T^{\infty} = \begin{bmatrix} 0.22 & 0.41 & 0.37 \\ 0.22 & 0.41 & 0.37 \\ 0.22 & 0.41 & 0.37 \end{bmatrix}$$

$$\mu(x^{(1)})T^t \to p(x) = (0.22, 0.41, 0.37)$$

http://www.stat.uchicago.edu/~yibi/teaching/stat317/2014/Lectures
/Lecture4_6up.pdf

Stationary Distribution

$$\pi P = \pi$$

$$\pi = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \qquad P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad P^2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Staionary Distribution may not be unique.
For Example, P = I

Limiting Distribution is Stationary Distribution

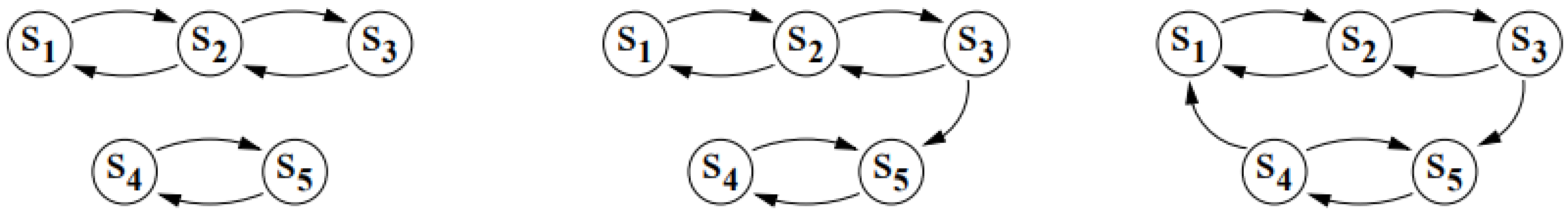
$$T = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0.1 & 0.9 \\ 0.6 & 0.4 & 0 \end{bmatrix} \qquad T^{\infty} = \begin{bmatrix} 0.22 & 0.41 & 0.37 \\ 0.22 & 0.41 & 0.37 \\ 0.22 & 0.41 & 0.37 \end{bmatrix}$$

$$[0.22, 0.41, 0.37]\, T = [0.22, 0.41, 0.37]$$

Irreducible

$$P^t(x, y) > 0 \quad for\ some\ t$$

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad P = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$



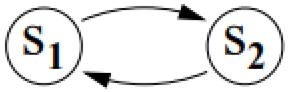
https://pages.dataiku.com/hubfs/Dataiku%20Dec%202016/Files/lecture3.pdf

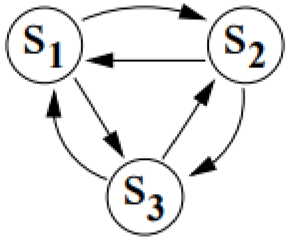
Aperiodic

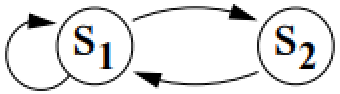
$$k = gcd\{n > 0 : Pr(X_n = i | X_0 = i) > 0)\}$$

state X is aperiodic if k>1

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad P^2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$







https://pages.dataiku.com/hubfs/Dataiku%20Dec%202016/Files/lecture3.pdf

Ergodicity

Markov Chain is ergodic if both irreducible and aperiodic

An ergodic Markov Chain has a unique stationary distribution
and has limit distribution

$$\pi P = \pi$$

$$P = \begin{pmatrix} 0.1 & 0 & 0.9 \\ 0.1 & 0.5 & 0.4 \\ 0 & 0.3 & 0.7 \end{pmatrix}$$

```
import numpy as np
```

```
P = np.array([[0.1,0,0.9],[0.1,0.5,0.4],[0,0.3,0.7]])
```

```
C = P
for i in range(100000):
    C = np.dot(P,C)
print(C)
```

```
[[0.04 0.36 0.6 ]
 [0.04 0.36 0.6 ]
 [0.04 0.36 0.6 ]]
```

Mixing Time

- Markov chain is the time until the Markov chain is "close" to its steady state distribution.

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2} \sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)| \,.$$

$$d(t) = \sup_{\mu \in \mathcal{P}} \left\| \mu P^t - \pi \right\|_{\mathrm{TV}} ,$$

$$t_{\mathrm{mix}}(\varepsilon) := \min\{t \;:\; d(t) \leq \varepsilon\}$$

https://pages.uoregon.edu/dlevin/MARKOV/mcmt2e.pdf

Concentration Coefficient

Hypothesis 2 (Uniform stochasticity) *Let* $\overline{\mu}$ *be some distribution, for example a uniform distribution. There exists a constant* C*, such that for all policies* π*, for all* $i, j \in X$*,*

$$P^{\pi}(i,j) \leq C\overline{\mu}(j) \tag{8}$$

Remi Munos. Error bounds for approximate policy iteration. In ´ ICML 2003: Proceedings of the 20th Annual International Conference on Machine Learning, pages 560–567, 2003.

Concentration Coefficients

$$C(\mu) = \max_{x,y \in X,\, a \in A} \frac{p(x,a,y)}{\mu(y)}$$

Relative smoothness of the immediate transition probabilities w.r.t the denominator μ

$$c(m) = \max_{\pi_1,\ldots,\pi_m,\, y \in X} \frac{(\nu P^{\pi_1} P^{\pi_2} \ldots P^{\pi_m})(y)}{\mu(y)},$$

how much the future state distributions may possibly differ from μ

$$C_1(\nu,\mu) := (1-\gamma) \sum_{m \geq 0} \gamma^m c(m),$$

$$C_2(\nu,\mu) := (1-\gamma)^2 \sum_{m \geq 1} m \gamma^{m-1} c(m).$$

Concentration Coefficient

- N states and μ is uniform distribution

$$C(\mu) = \max_{x,y \in X, a \in A} \frac{p(x,a,y)}{\mu(y)} \qquad C(\mu) = N$$

$$c(m) = \max_{\pi_1,\ldots,\pi_m,\, y\in X} \frac{(\nu P^{\pi_1} P^{\pi_2} \ldots P^{\pi_m})(y)}{\mu(y)},$$

$$C_1(\nu,\mu) := (1-\gamma)\sum_{m\geq 0} \gamma^m c(m),$$

$$(1-\gamma)\sum_{m\geq 0}\gamma^m \Pr(x_m = y | x_0 \sim \nu, \pi_1, \ldots, \pi_m) \leq C_1(\nu,\mu)\mu(y),$$

$$C_2(\nu,\mu) := (1-\gamma)^2 \sum_{m\geq 1} m\gamma^{m-1} c(m).$$

$$(1-\gamma)^2 \sum_{m\geq 1} m\gamma^{m-1}\Pr(x_m = y | x_0 \sim \nu, \pi_1, \ldots, \pi_m) \leq C_2(\nu,\mu)\mu(y).$$

Theorem 5.2. *Let μ and ν be two probability measures on X. Consider the AVI algorithm defined by (1.1), write π_n a policy greedy w.r.t. V_n, and $\varepsilon_n = V_{n+1} - \mathcal{T}V_n \in \mathbb{R}^N$ the approximation error. Let $\varepsilon > 0$ and assume that $\mathcal{A}$ returns $\varepsilon-$approximations V_{n+1} in $L_{p,\mu}$-norm ($p \geq 1$) of $\mathcal{T}V_n$, i.e. $||\varepsilon_n||_{p,\mu} \leq \varepsilon$, for $n \geq 0$. Then:*

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \big[C(\mu)\big]^{1/p} \varepsilon,$$

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_{p,\nu} \leq \frac{2\gamma}{(1-\gamma)^2} \big[C_2(\nu,\mu)\big]^{1/p} \varepsilon.$$

bellman operator

$$(T_\mu J)(i) = \sum_{j=0}^{n} p_{ij}\big(\mu(i)\big)\Big(g\big(i, \mu(i), j\big) + J(j)\Big), \qquad i = 1, \ldots, n.$$

$$(TJ)(i) = \min_{u \in U(i)} \sum_j p_{ij}(u)\big(g(i, u, j) + \alpha J(j)\big), \qquad i = 1, \ldots, n,$$

$$lim_{k \leftarrow \infty} T^{\mu_k} J = J^{\mu}$$

Properties of Bellman Operator

1. Monotonicity property

$$(TJ)(x) \le (TJ')(x), \qquad \forall\ x,$$

$$(T_\mu J)(x) \le (T_\mu J')(x), \qquad \forall\ x.$$

2. Constant shift property

$$\big(T(J + re)\big)(x) = (TJ)(x) + \alpha r,$$

$$\big(T_\mu(J + re)\big)(x) = (T_\mu J)(x) + \alpha r,$$

Contraction Property of Bellman Operator

$$\max_x \left|(TJ)(x) - (TJ')(x)\right| \le \alpha \max_x \left|J(x) - J'(x)\right|,$$

$$\max_x \left|(T_\mu J)(x) - (T_\mu J')(x)\right| \le \alpha \max_x \left|J(x) - J'(x)\right|.$$

Proof)

$$c = max_{x \in S}|J(x) - J'(x)|$$

$$J(x) - c \le J'(x) \le J(x) + c$$

Approximate Value Iteration

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\varepsilon.$$

Proof)

$$\epsilon_n = TJ_n - J_{n+1}$$

$$||J^* - J^{\pi_n}||_\infty \leq ||TJ^* - T^{\pi_n}J_n||_\infty + ||T^{\pi_n}J_n - T^{\pi_n}J^{\pi_n}||_\infty$$

$$\leq \gamma|||J^* - J_n|| + \gamma||J_n - J^{\pi_n}||$$

$$\leq \gamma|||J^* - J_n|| + \gamma(||J_n - J^*|| + ||J_* - J^{\pi_n}||)$$

$$||J^* - J_{n+1}|| \leq ||JV^* - TV_n|| + ||TV_n - V_{n+1}|| \leq \gamma||J^* - J_n|| + \epsilon$$

Approximate Policy Iteration

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

Proof)

$$||J - J^*|| = \epsilon$$

$$\begin{aligned}
\|J^\mu - J^*\| &= \|T_\mu J^\mu - J^*\| \\
&\leq \|T_\mu J^\mu - T_\mu J\| + \|T_\mu J - J^*\| \\
&\leq \alpha\|J^\mu - J\| + \|TJ - J^*\| \\
&\leq \alpha\|J^\mu - J^*\| + \alpha\|J^* - J\| + \alpha\|J - J^*\| \\
&= \alpha\|J^\mu - J^*\| + 2\alpha\epsilon,
\end{aligned}$$

Action Gap

$$Loss = \int (V^*(x) - V^\pi(x))d\rho(x)$$

$$g_{Q_*(x)} = |Q^*(x,1) - Q^*(x,2)|$$

remi

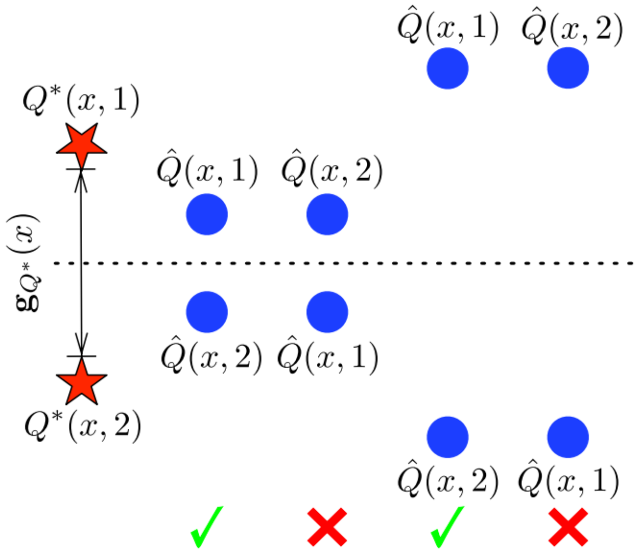


Figure 2: The action-gap function $\mathbf{g}_{Q^*}(x)$ and the relative ordering of the optimal and the estimated action-value functions for a single state x. Depending on the ordering of the estimates, the greedy action is the same as (✓) or different from (X) the optimal action. This figure does not show all possible configurations.

Peformance Loss : $E[V^* - V^{\hat{\pi}}]$

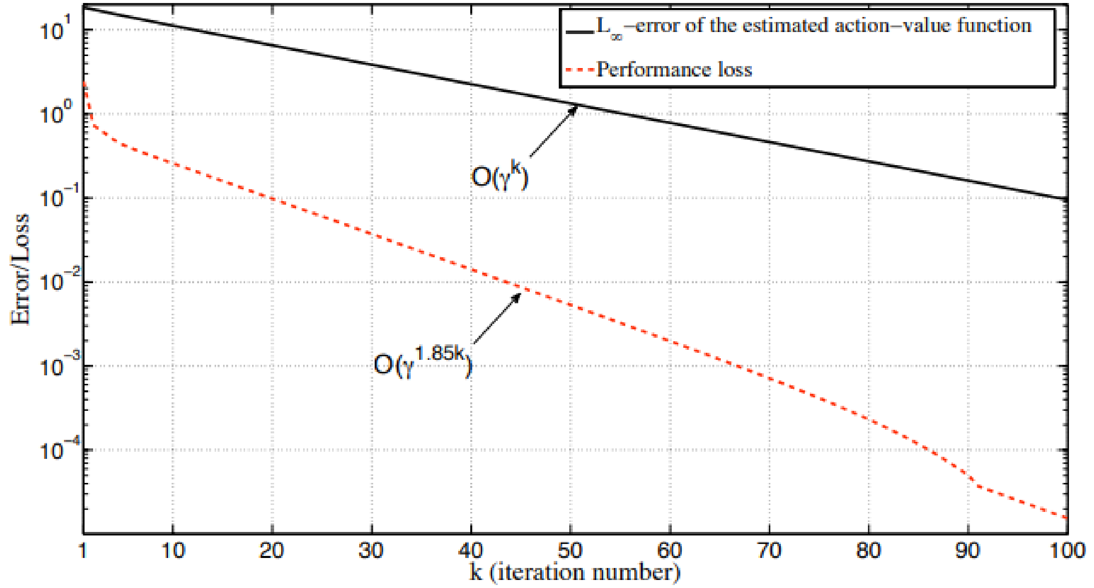


Figure 1: Comparison of the action-value estimation error $\|\hat{Q} - Q^*\|_\infty$ and the performance loss $\|V^* - V^{\hat{\pi}}\|_1$ ($\hat{\pi}$ is the greedy policy with respect to $\hat{Q}$) at different iterations of the value iteration algorithm. The rate of decrease for the performance loss is considerably faster than that of the estimation error. The problem is a 1D stochastic chain walk with 500 states and $\gamma = 0.95$.

Assumption

$$\mathbb{P}_{\rho^*}\left(0 < \mathbf{g}_{Q^*}(X) \leq t\right) \triangleq \int_{\mathcal{X}} \mathbb{I}\{0 < \mathbf{g}_{Q^*}(x) \leq t\}\, \mathrm{d}\rho^*(x) \leq c_g\, t^{\zeta}.$$

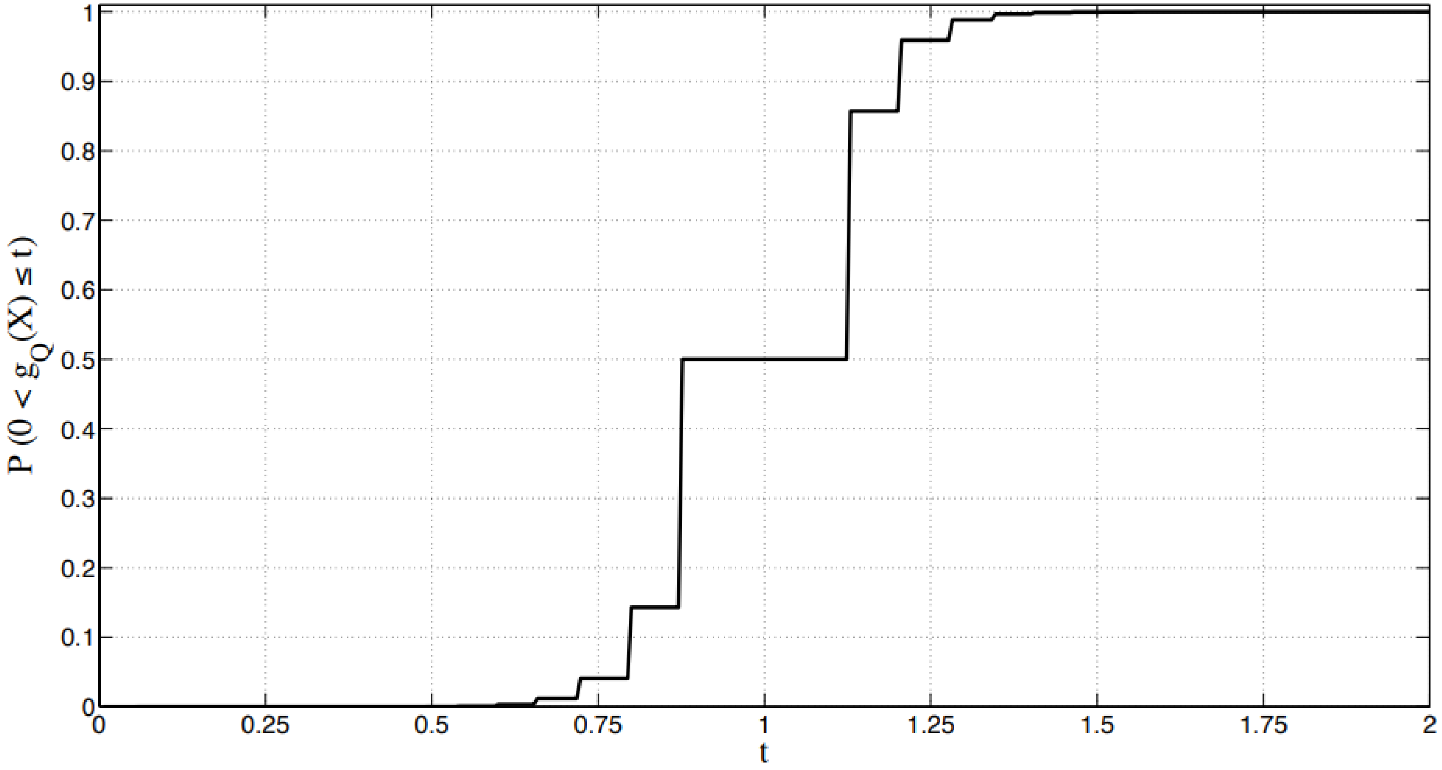


Figure 3: The probability distribution $\mathbb{P}_{\rho^*}\left(0 < \mathbf{g}_{Q^*}(X) \leq t\right)$ for a 1D stochastic chain walk with 500 states and $\gamma = 0.95$. Here the probability of the action-gap being close to zero is small.

Theorem 1. *Consider an MDP* $(\mathcal{X}, \mathcal{A}, P, \mathcal{R}, \gamma)$ *with* $|\mathcal{A}| = 2$ *and an estimate* $\hat{Q}$ *of the optimal action-value function. Let Assumption A1 hold and* $C(\rho, \rho^*) < \infty$*. Denote* $\hat{\pi}$ *as the greedy policy w.r.t.* $\hat{Q}$*. We then have*

$$\mathrm{Loss}(\hat{\pi}; \rho) \le \begin{cases} 2^{1+\zeta}\, c_g\, C(\rho, \rho^*) \left\| \hat{Q} - Q^* \right\|_{\infty}^{1+\zeta}, \\ 2^{1+\frac{p(1+\zeta)}{p+\zeta}}\, c_g^{\frac{p-1}{p+\zeta}} C(\rho, \rho^*) \left\| \hat{Q} - Q^* \right\|_{p,\rho^*}^{\frac{p(1+\zeta)}{p+\zeta}}. \qquad (1 \le p < \infty) \end{cases}$$

Proof)

$$F(x) = \left(Q^{\pi^*}(x, \pi^*(x)) - Q^{\pi^*}(x, \hat{\pi}(x))\right) + \left(Q^{\pi^*}(x, \hat{\pi}(x)) - Q^{\hat{\pi}}(x, \hat{\pi}(x))\right) = F_1(x) + F_2(x).$$

$$\begin{aligned} F_2(x) = &\left[r(x, \hat{\pi}(x)) + \gamma \int_{\mathcal{X}} P(dy|x, \hat{\pi}(x)) Q^{\pi^*}(y, \pi^*(y)) \right] - \\ &\left[r(x, \hat{\pi}(x)) + \gamma \int_{\mathcal{X}} P(dy|x, \hat{\pi}(x)) Q^{\hat{\pi}}(y, \hat{\pi}(y)) \right] \\ = \gamma P^{\hat{\pi}}(\cdot|x) F(\cdot). \end{aligned}$$

$$\rho F = \sum_{m\geq 0} \rho(\gamma P^{\hat{\pi}})^m F_1 = \sum_{m\geq 0} \gamma^m \int_{\mathcal{X}} \left(\rho(P^{\hat{\pi}})^m\right)(\mathrm{d}y) F_1(y)$$

$$= \sum_{m\geq 0} \gamma^m \int_{\mathcal{X}} \frac{\mathrm{d}\left(\rho(P^{\hat{\pi}})^m\right)}{\mathrm{d}\rho^*}(y)\mathrm{d}\rho^*(y) F_1(y)$$

$$\leq \sum_{m\geq 0} \gamma^m c(m;\hat{\pi})\rho^* F_1 \leq C(\rho,\rho^*)\,\rho^* F_1.$$

$$\hat{\pi}(x) \neq \pi^*(x) \qquad |Q^{\pi^*}(x,a) - \hat{Q}(x,a)| \leq \varepsilon$$

$$\mathbf{g}_{Q^*}(x) = |Q^{\pi^*}(x,1) - Q^{\pi^*}(x,2)| \leq 2\varepsilon$$

$$\hat{\pi}(x) \neq \pi^*(x) \qquad |Q^{\pi^*}(x,a) - \hat{Q}(x,a)| \leq \varepsilon$$

$$\varepsilon_0 = \|Q^{\pi^*} - \hat{Q}\|_\infty$$

$$\mathbf{g}_{Q^*}(x) = |Q^{\pi^*}(x,1) - Q^{\pi^*}(x,2)| \leq 2\varepsilon$$

$$F_1(x) = \left[Q^{\pi^*}(x,\pi^*(x)) - Q^{\pi^*}(x,\hat{\pi}(x))\right][\mathbb{I}\{\hat{\pi}(x) = \pi^*(x)\} + \mathbb{I}\{\hat{\pi}(x) \neq \pi^*(x)\}]$$

$$= \left[Q^{\pi^*}(x,\pi^*(x)) - Q^{\pi^*}(x,1-\pi^*(x))\right]\mathbb{I}\{\hat{\pi}(x) \neq \pi^*(x)\}$$

$$\times [\mathbb{I}\{\mathbf{g}_{Q^*}(x) = 0\} + \mathbb{I}\{0 < \mathbf{g}_{Q^*}(x) \leq 2\varepsilon_0\} + \mathbb{I}\{\mathbf{g}_{Q^*}(x) > 2\varepsilon_0\}]$$

$$\leq 0 + 2\varepsilon_0\,\mathbb{I}\{0 < \mathbf{g}_{Q^*}(x) \leq 2\varepsilon_0\} + 0.$$

$$\rho^* F_1 \leq 2\varepsilon_0\,\mathbb{P}_{\rho^*}(0 < \mathbf{g}_{Q^*}(X) \leq 2\varepsilon_0) \leq 2\varepsilon_0\, c_g\,(2\varepsilon_0)^\zeta$$

Theorem 2 (Error Propagation for AVI). *Consider an MDP $(\mathcal{X}, \mathcal{A}, P, \mathcal{R}, \gamma)$ with $|\mathcal{A}| = 2$ that satisfies Assumption A1 and has $C(\rho, \rho^*) < \infty$. Let $p \geq 1$ be a real number and K be a positive integer. Then for any sequence $(\hat{Q}_k)_{k=0}^{K} \subset B(\mathcal{X} \times \mathcal{A}, Q_{max})$ and the corresponding sequence $(\varepsilon_k)_{k=0}^{K-1}$ defined in* (3), *we have*

$$\mathrm{Loss}(\hat{\pi}(\cdot, Q_K); \rho) \leq 2 \left(\frac{2}{1-\gamma} \right)^{\frac{p(1+\zeta)}{p+\zeta}} c_g^{\frac{p-1}{p+\zeta}} C(\rho, \rho^*) \left[\sum_{k=0}^{K-1} \alpha_k \, \|\varepsilon_k\|_{p,\rho^*}^{p} + \alpha_K (2Q_{max})^p \right]^{\frac{1+\zeta}{p+\zeta}} .$$

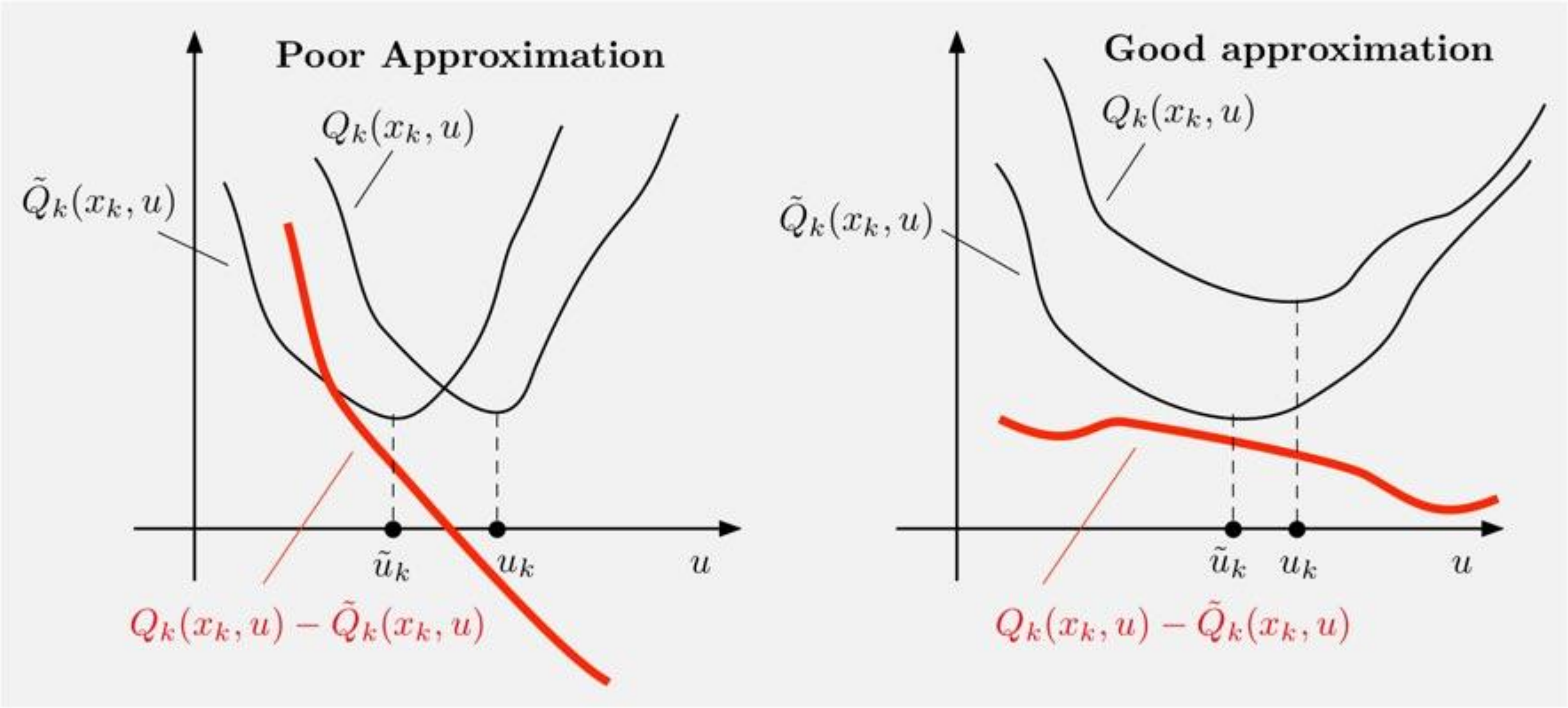
$$\varepsilon_k \triangleq T^* \hat{Q}_k - \hat{Q}_{k+1}$$

$$Q^* - \hat{Q}_{k+1} = T^{\pi^*} Q^* - T^{\pi^*} \hat{Q}_k + T^{\pi^*} \hat{Q}_k - T^* \hat{Q}_k + \varepsilon_k \leq \gamma P^{\pi^*} (Q^* - \hat{Q}_k) + \varepsilon_k$$

$$Q^* - \hat{Q}_K \leq \sum_{k=0}^{K-1} \gamma^{K-k-1} (P^{\pi^*})^{K-k-1} \varepsilon_k + \gamma^K (P^{\pi^*})^K (Q^* - \hat{Q}_0).$$

$$\|Q^* - \hat{Q}_K\|_{p,\rho^*} = \rho^* |Q^* - \hat{Q}_K|^p \qquad \alpha_k = \begin{cases} \frac{(1-\gamma)}{1-\gamma^{K+1}} \gamma^{K-k-1} & 0 \leq k < K, \\ \frac{(1-\gamma)}{1-\gamma^{K+1}} \gamma^K & k = K. \end{cases}$$

$$\begin{aligned} \rho^* |Q^* - \hat{Q}_K|^p &\leq \left(\frac{1-\gamma^{K+1}}{1-\gamma} \right)^p \left[\sum_{k=0}^{K-1} \alpha_k \rho^* (P^{\pi^*})^{K-k-1} |\varepsilon_k| + \alpha_K \rho^* (P^{\pi^*})^K |Q^* - \hat{Q}_0| \right]^p \\ &\leq \left(\frac{1-\gamma^{K+1}}{1-\gamma} \right)^p \left[\sum_{k=0}^{K-1} \alpha_k \|\varepsilon_k\|_{p,\rho^*}^p + \alpha_K (2Q_{\max})^p \right], \end{aligned}$$



REINFORCEMENT LEARNING AND OPTIMAL CONTROL BOOK, Athena Scientific, July 2019