Finite-Sample Analysis in Reinforcement Learning

Mohammad Ghavamzadeh

INRIA Lille – Nord Europe, Team SequeL

Outline

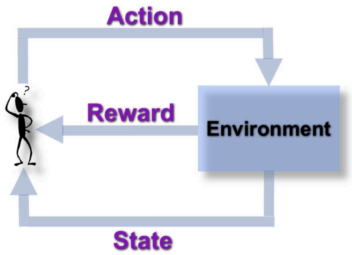
1. Introduction to RL and DP
2. Approximate Dynamic Programming (AVI & API)
3. How does Statistical Learning Theory come to the picture?
4. Error Propagation (AVI & API Error Propagation)
5. An AVI Algorithm (Fitted Q-Iteration)
   - FQI: error at each iteration
   - Final performance bound of FQI
6. An API Algorithm (Least-Squares Policy Iteration)
   - Error at each iteration (LSTD error)
   - Final performance bound of LSPI
7. Discussion

Sequential Decision-Making under Uncertainty

How Can I ... ?

- Move around in the physical world (e.g. driving, navigation)
- Play and win a game
- Retrieve information over the web
- Medical diagnosis and treatment
- Maximize the throughput of a factory
- Optimize the performance of a rescue team

Reinforcement Learning (RL)



- **RL:** A class of learning problems in which an agent interacts with a dynamic, stochastic, and incompletely known environment
- **Goal:** Learn an action-selection strategy, or policy, to optimize some measure of its long-term performance
- **Interaction:** Modeled as a MDP or a POMDP

Markov Decision Process

MDP

- An MDP $\mathcal{M}$ is a tuple $\langle \mathcal{X}, \mathcal{A}, r, p, \gamma \rangle$.
- The state space $\mathcal{X}$ is a bounded closed subset of $\mathbb{R}^d$.
- The set of actions $\mathcal{A}$ is finite ($|\mathcal{A}| < \infty$).
- The reward function $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is bounded by $R_{\max}$.
- The transition model $p(\cdot|x, a)$ is a distribution over $\mathcal{X}$.
- $\gamma \in (0, 1)$ is a discount factor.

- **Policy:** a mapping from states to actions $\pi(x) \in \mathcal{A}$

Value Function

For a policy π

- **Value function** $V^\pi : \mathcal{X} \to \mathbb{R}$

$$V^\pi(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r\big(X_t, \pi(X_t)\big) | X_0 = x\right]$$

- **Action-value function** $Q^\pi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$

$$Q^\pi(x,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(X_t, A_t) | X_0 = x,\ A_0 = a\right]$$

Notation

Bellman Operator

- Bellman operator for policy π

$$\mathcal{T}^\pi : \mathcal{B}^V(\mathcal{X}; V_{\max}) \to \mathcal{B}^V(\mathcal{X}; V_{\max})$$

- V^π is the unique fixed-point of the Bellman operator

$$(\mathcal{T}^\pi V)(x) = r\big(x, \pi(x)\big) + \gamma \int_{\mathcal{X}} p\big(dy|x, \pi(x)\big) V(y)$$

- The action-value function Q^π is defined as

$$Q^\pi(x, a) = r(x, a) + \gamma \int_{\mathcal{X}} p(dy|x, a) V^\pi(y)$$

Optimal Value Function and Optimal Policy

- **Optimal value function**

$$V^*(x) = \sup_\pi V^\pi(x) \qquad \forall x \in \mathcal{X}$$

- **Optimal action-value function**

$$Q^*(x,a) = \sup_\pi Q^\pi(x,a) \qquad \forall x \in \mathcal{X},\ \forall a \in \mathcal{A}$$

- A policy π is **optimal** if

$$V^\pi(x) = V^*(x) \qquad \forall x \in \mathcal{X}$$

Notation

Bellman Optimality Operator

- Bellman optimality operator

$$\mathcal{T} : \mathcal{B}^V(\mathcal{X}; V_{\max}) \to \mathcal{B}^V(\mathcal{X}; V_{\max})$$

- V^* is the unique fixed-point of the Bellman optimality operator

$$(\mathcal{T}V)(x) = \max_{a \in \mathcal{A}} \left[r(x,a) + \gamma \int_{\mathcal{X}} p(dy|x,a) V(y) \right]$$

- Optimal action-value function Q^* is defined as

$$Q^*(x,a) = r(x,a) + \gamma \int_{\mathcal{X}} p(dy|x,a) V^*(y)$$

Properties of Bellman Operators

- **Monotonicity:** if $V_1 \leq V_2$ component-wise

$$\mathcal{T}^\pi V_1 \leq \mathcal{T}^\pi V_2 \qquad \text{and} \qquad \mathcal{T} V_1 \leq \mathcal{T} V_2$$

- **Max-Norm Contraction:** $\forall V_1, V_2 \in \mathcal{B}^V(\mathcal{X}; V_{\max})$

$$||\mathcal{T}^\pi V_1 - \mathcal{T}^\pi V_2||_\infty \leq \gamma ||V_1 - V_2||_\infty$$

$$||\mathcal{T} V_1 - \mathcal{T} V_2||_\infty \leq \gamma ||V_1 - V_2||_\infty$$

Dynamic Programming Algorithms

Value Iteration

- start with an arbitrary action-value function Q_0
- at each iteration k $\qquad Q_{k+1} = \mathcal{T} Q_k$

Convergence

- $\lim_{k \to \infty} V_k = V^*$.

$$||V^* - V_{k+1}||_\infty = ||\mathcal{T} V^* - \mathcal{T} V_k||_\infty \leq \gamma ||V^* - V_k||_\infty \leq \gamma^{k+1} ||V^* - V_0||_\infty \stackrel{k \to \infty}{\longrightarrow} 0$$

Dynamic Programming Algorithms

Policy Iteration

- start with an arbitrary policy π_0
- at each iteration k
  - Policy Evaluation: Compute Q^{π_k}
  - Policy Improvement: Compute the *greedy* policy w.r.t. Q^{π_k}

$$\pi_{k+1}(x) = (\mathcal{G}\pi_k)(x) = \arg\max_{a \in \mathcal{A}} Q^{\pi_k}(x, a)$$

Convergence

PI generates a sequence of policies with increasing performance ($V^{\pi_{k+1}} \geq V^{\pi_k}$) and stops after a finite number of iterations with an optimal policy π^*.

$$V^{\pi_k} = \mathcal{T}^{\pi_k} V^{\pi_k} \leq \mathcal{T} V^{\pi_k} = \mathcal{T}^{\pi_{k+1}} V^{\pi_k} \leq \lim_{n \to \infty} (\mathcal{T}^{\pi_{k+1}})^n V^{\pi_k} = V^{\pi_{k+1}}$$

Approximate Dynamic Programming

Approximate Dynamic Programming Algorithms

Value Iteration

- start with an arbitrary action-value function Q_0
- at each iteration k $\qquad Q_{k+1} = \mathcal{T} Q_k$

What if $Q_{k+1} \approx \mathcal{T} Q_k$?

$$||Q^* - Q_{k+1}|| \overset{?}{\leq} \gamma ||Q^* - Q_k||$$

Approximate Dynamic Programming Algorithms

Policy Iteration

- start with an arbitrary policy π_0
- at each iteration k
  - Policy Evaluation: Compute Q^{π_k}
  - Policy Improvement: Compute the *greedy* policy w.r.t. Q^{π_k}

$$\pi_{k+1}(x) = (\mathcal{G}\pi_k)(x) = \arg\max_{a \in \mathcal{A}} Q^{\pi_k}(x, a)$$

What if we cannot compute Q^{π_k} exactly? (Compute $\widehat{Q}^{\pi_k} \approx Q^{\pi_k}$ instead)

$$\pi_{k+1}(x) = \arg\max_{a \in \mathcal{A}} \widehat{Q}^{\pi_k}(x, a) \neq (\mathcal{G}\pi_k)(x) \longrightarrow V^{\pi_{k+1}} \overset{?}{\geq} V^{\pi_k}$$

Statistical Learning Theory in RL & ADP

Approximate Value Iteration (AVI) $Q_{k+1} \approx \mathcal{T} Q_k$

- finding a function that best approximates $\mathcal{T} Q_k$ $\quad Q = \min_f ||f - \mathcal{T} Q_k||_\mu$
- only noisy observations of $\mathcal{T} Q_k$ are available $\quad \widehat{\mathcal{T}} Q_k$

  Target Function $= \mathcal{T} Q_k$ Noisy Observation $= \widehat{\mathcal{T}} Q_k$

- we minimize the ***empirical error*** $\quad Q_{k+1} = \widehat{Q} = \min_f ||f - \widehat{\mathcal{T}} Q_k||_{\widehat{\mu}}$

  with the target of minimizing the ***true error*** $\quad Q = \min_f ||f - \mathcal{T} Q_k||_\mu$

- **Objective:** $||\widehat{Q} - \mathcal{T} Q_k||_\mu \leq \underbrace{||\widehat{Q} - Q||_\mu}_{\text{estimation error}} + \underbrace{||Q - \mathcal{T} Q_k||_\mu}_{\text{approximation error}}$ to be small

regression

Approximate Value Iteration (AVI) $Q_{k+1} \approx \mathcal{T} Q_k$

- finding a function that best approximates $\mathcal{T} Q_k$ $\quad Q = \min_f ||f - \mathcal{T} Q_k||_\mu$
- only noisy observations of $\mathcal{T} Q_k$ are available $\quad \widehat{\mathcal{T}} Q_k$

  Target Function $= \mathcal{T} Q_k$ $\qquad$ Noisy Observation $= \widehat{\mathcal{T}} Q_k$

- we minimize the ***empirical error*** $\quad Q_{k+1} = \widehat{Q} = \min_f ||f - \widehat{\mathcal{T}} Q_k||_{\widehat{\mu}}$

  with the target of minimizing the ***true error*** $\quad Q = \min_f ||f - \mathcal{T} Q_k||_\mu$

- **Objective:** $||\widehat{Q} - \mathcal{T} Q_k||_\mu \le \underbrace{||\widehat{Q} - Q||_\mu}_{\text{estimation error}} + \underbrace{||Q - \mathcal{T} Q_k||_\mu}_{\text{approximation error}}$ to be small

regression

Statistical Learning Theory in RL & ADP

Approximate Value Iteration (AVI) $Q_{k+1} \approx \mathcal{T} Q_k$

- finding a function that best approximates $\mathcal{T} Q_k$ $Q = \min_f ||f - \mathcal{T} Q_k||_\mu$
- only noisy observations of $\mathcal{T} Q_k$ are available $\widehat{\mathcal{T}} Q_k$

Target Function $= \mathcal{T} Q_k$ Noisy Observation $= \widehat{\mathcal{T}} Q_k$

- we minimize the ***empirical error*** $Q_{k+1} = \widehat{Q} = \min_f ||f - \widehat{\mathcal{T}} Q_k||_{\widehat{\mu}}$

  with the target of minimizing the ***true error*** $Q = \min_f ||f - \mathcal{T} Q_k||_\mu$

- **Objective:** $||\widehat{Q} - \mathcal{T} Q_k||_\mu \leq \underbrace{||\widehat{Q} - Q||_\mu}_{\text{estimation error}} + \underbrace{||Q - \mathcal{T} Q_k||_\mu}_{\text{approximation error}}$ to be small

regression

Statistical Learning Theory in RL & ADP

Approximate Value Iteration (AVI) $Q_{k+1} \approx \mathcal{T} Q_k$

- finding a function that best approximates $\mathcal{T} Q_k$ $Q = \min_f ||f - \mathcal{T} Q_k||_\mu$
- only noisy observations of $\mathcal{T} Q_k$ are available $\widehat{\mathcal{T}} Q_k$

Target Function $= \mathcal{T} Q_k$ Noisy Observation $= \widehat{\mathcal{T}} Q_k$

- we minimize the ***empirical error*** $Q_{k+1} = \widehat{Q} = \min_f ||f - \widehat{\mathcal{T}} Q_k||_{\widehat{\mu}}$

with the target of minimizing the ***true error*** $Q = \min_f ||f - \mathcal{T} Q_k||_\mu$

- **Objective:** $||\widehat{Q} - \mathcal{T} Q_k||_\mu \leq \underbrace{||\widehat{Q} - Q||_\mu}_{\text{estimation error}} + \underbrace{||Q - \mathcal{T} Q_k||_\mu}_{\text{approximation error}}$ to be small

regression

Statistical Learning Theory in RL & ADP

Approximate Value Iteration (AVI) $Q_{k+1} \approx \mathcal{T} Q_k$

- finding a function that best approximates $\mathcal{T} Q_k$ $\quad Q = \min_f ||f - \mathcal{T} Q_k||_\mu$
- only noisy observations of $\mathcal{T} Q_k$ are available $\quad \widehat{\mathcal{T}} Q_k$

  Target Function $= \mathcal{T} Q_k$ $\qquad$ Noisy Observation $= \widehat{\mathcal{T}} Q_k$

- we minimize the ***empirical error*** $\quad Q_{k+1} = \widehat{Q} = \min_f ||f - \widehat{\mathcal{T}} Q_k||_{\widehat{\mu}}$

  with the target of minimizing the ***true error*** $\quad Q = \min_f ||f - \mathcal{T} Q_k||_\mu$

- **Objective:** $||\widehat{Q} - \mathcal{T} Q_k||_\mu \leq \underbrace{||\widehat{Q} - Q||_\mu}_{\text{estimation error}} + \underbrace{||Q - \mathcal{T} Q_k||_\mu}_{\text{approximation error}}$ to be small

regression

Statistical Learning Theory in RL & ADP

Approximate Value Iteration (AVI)

$Q_{k+1} \approx \mathcal{T} Q_k$

- finding a function that best approximates $\mathcal{T} Q_k$ $\quad Q = \min_f ||f - \mathcal{T} Q_k||_\mu$
- only noisy observations of $\mathcal{T} Q_k$ are available $\quad \widehat{\mathcal{T}} Q_k$

Target Function $= \mathcal{T} Q_k$ $\qquad$ Noisy Observation $= \widehat{\mathcal{T}} Q_k$

- we minimize the ***empirical error*** $\quad Q_{k+1} = \widehat{Q} = \min_f ||f - \widehat{\mathcal{T}} Q_k||_{\widehat{\mu}}$

  with the target of minimizing the ***true error*** $\quad Q = \min_f ||f - \mathcal{T} Q_k||_\mu$

- **Objective:** $||\widehat{Q} - \mathcal{T} Q_k||_\mu \leq \underbrace{||\widehat{Q} - Q||_\mu}_{\text{estimation error}} + \underbrace{||Q - \mathcal{T} Q_k||_\mu}_{\text{approximation error}}$ to be small

regression

Statistical Learning Theory in RL & ADP

Approximate Policy Iteration (API) - policy evaluation

- finding a function that best approximates Q^{π_k} $\qquad Q = \min_f ||f - Q^{\pi_k}||_\mu$
- only noisy observations of Q^{π_k} are available $\qquad \widehat{Q}^{\pi_k}$

  Target Function $= Q^{\pi_k}$ $\qquad$ Noisy Observation $= \widehat{Q}^{\pi_k}$

- we minimize the ***empirical error*** $\qquad \widehat{Q} = \min_f ||f - \widehat{Q}^{\pi_k}||_{\widehat{\mu}}$

  with the target of minimizing the ***true error*** $\qquad Q = \min_f ||f - Q^{\pi_k}||_\mu$

- **Objective:** $||\widehat{Q} - Q^{\pi_k}||_\mu \leq \underbrace{||\widehat{Q} - Q||_\mu}_{\text{estimation error}} + \underbrace{||Q - Q^{\pi_k}||_\mu}_{\text{approximation error}}$ to be small

regression

Statistical Learning Theory in RL & ADP

Approximate Policy Iteration (API) $\pi_{k+1} \approx \mathcal{G}\pi_k$

- finding a policy that best approximates $\mathcal{G}\pi_k$ $\pi = \min_f \mathcal{L}(f, \pi_k; \mu)$

- we minimize the ***empirical error*** $\pi_{k+1} = \widehat{\pi} = \min_f \widehat{\mathcal{L}}(f, \pi_k; \widehat{\mu})$

  with the target of minimizing the ***true error*** $\pi = \min_f \mathcal{L}(f, \pi_k; \mu)$

- **Objective:** $\mathcal{L}(\widehat{\pi}, \pi_k; \mu) \leq \underbrace{\mathcal{L}(\widehat{\pi}, \pi; \mu)}_{\text{estimation error}} + \underbrace{\mathcal{L}(\pi, \pi_k; \mu)}_{\text{approximation error}}$ to be small

classification (*we do not discuss it in this talk*)

Approximate Policy Iteration (API) - policy evaluation

- finding the fixed-point of $\mathcal{T}^{\pi_k}$
- only noisy observations of $\mathcal{T}^{\pi_k}$ are available $\widehat{\mathcal{T}}^{\pi_k}$

a fixed-point problem

SLT in RL & ADP

- supervised learning methods (regression, classification) appear in the inner-loop of ADP algorithms (performance at each iteration)
- tools from SLT that are used to analyze supervised learning methods can be used in RL and ADP (e.g., how many samples are required to achieve a certain performance)

What makes RL more challenging?

- the objective is not always to recover a target function from its noisy observations (fixed-point vs. regression)
- the target sometimes has to be approximated given sample trajectories (non i.i.d. samples)
- propagation of error (control problem)

is there any hope?

Inria
INVENTEURS DU MONDE NUMÉRIQUE

Approximate Value Iteration (AVI)

$$V_{k+1} = \mathcal{T} V_k + \epsilon_k \qquad \text{or} \qquad ||V_{k+1} - \mathcal{T} V_k||_\infty = \epsilon_k$$

Proposition (AVI Error Propagation)

We run AVI for K iterations and $\pi_K = \mathcal{G} V_K$

$$||V^* - V^{\pi_K}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{0 \leq k < K} \epsilon_k + \frac{2\gamma^{K+1}}{1-\gamma} ||V^* - V_0||_\infty.$$

Proof

$$||V^* - V_{k+1}||_\infty \leq ||\mathcal{T} V^* - \mathcal{T} V_k||_\infty + ||\mathcal{T} V_k - V_{k+1}||_\infty = \gamma ||V^* - V_k||_\infty + \epsilon_k$$

so

$$||V^* - V_K||_\infty \leq \sum_{k=0}^{K-1} \gamma^{K-1-k} \epsilon_k + \gamma^K ||V^* - V_0||_\infty \leq \frac{1}{1-\gamma} \max_{0 \leq k < K} \epsilon_k + \gamma^K ||V^* - V_0||_\infty$$

the result follows by the fact that $||V^* - V^{\pi_K}||_\infty \leq \dfrac{2\gamma}{1-\gamma} ||V^* - V_K||_\infty.$

Approximate Policy Iteration (API)

1. $V_k = V^{\pi_k} + \epsilon_k$ or $||V_k - V^{\pi_k}||_\infty = \epsilon_k$ *(Policy Evaluation Error)*

2. $V_k = \mathcal{T}^{\pi_k} V_k + \epsilon_k$ or $||V_k - \mathcal{T}^{\pi_k} V_k||_\infty = \epsilon_k$ *(Bellman Residual)*

Proposition (API Asymptotic Performance)

$$(1) \qquad \limsup_{k\to\infty} ||V^* - V^{\pi_k}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} \underbrace{||V_k - V^{\pi_k}||_\infty}_{\epsilon_k}$$

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

Approximate Dynamic Programming (ADP)

Proposition (AVI Asymptotic Performance)

$$\limsup_{k\to\infty} ||V^* - V^{\pi_k}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} \underbrace{||V_{k+1} - \mathcal{T} V_k||_\infty}_{\epsilon_k}$$

Proposition (API Asymptotic Performance)

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

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

Error Propagation

AVI Error Propagation

Error at each iteration k: $\epsilon_k = \mathcal{T} V_k - V_{k+1}$

π_K is a greedy policy w.r.t. V_{K-1} $\pi_K = \mathcal{G}(V_{K-1})$

Proposition (AVI Pointwise Error Bound)

$$V^* - V^{\pi_K} \leq (I - \gamma P^{\pi_K})^{-1} \Big\{ \sum_{k=0}^{K-1} \gamma^{K-k} \big[(P^{\pi^*})^{K-k} + P^{\pi_K} P^{\pi_{K-1}} \dots P^{\pi_{k+1}} \big] |\epsilon_k|$$

$$+ \gamma^{K+1} \big[(P^{\pi^*})^{K+1} + (P^{\pi_K} P^{\pi_{K-1}} \dots P^{\pi_0}) \big] |V^* - V_0| \Big\}$$

AVI Error Propagation

Proposition (AVI L_p Error Bound) $\epsilon_k = \mathcal{T} V_k - V_{k+1}$

$$||V^* - V^{\pi_K}||_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2}\left[C_{\rho,\mu}^{1/p} \max_{0\leq k<K} ||\epsilon_k||_{p,\mu} + 2\gamma^{K/p} V_{\max}\right] \qquad \text{(A1)}$$

$$||V^* - V^{\pi_K}||_{\infty} \leq \frac{2\gamma}{(1-\gamma)^2}\left[C_{\mu}^{1/p} \max_{0\leq k<K} ||\epsilon_k||_{p,\mu} + 2\gamma^{K/p} V_{\max}\right] \qquad \text{(A2)}$$

AVI Error Propagation

Proposition (AVI L_p Error Bound) $\epsilon_k = \mathcal{T}V_k - V_{k+1}$

$$||V^* - V^{\pi_K}||_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2}\left[C_{\rho,\mu}^{1/p} \max_{0\leq k<K} ||\epsilon_k||_{p,\mu} + 2\gamma^{K/p} V_{\max}\right] \qquad \text{(A1)}$$

$$||V^* - V^{\pi_K}||_{\infty} \leq \frac{2\gamma}{(1-\gamma)^2}\left[C_{\mu}^{1/p} \max_{0\leq k<K} ||\epsilon_k||_{p,\mu} + 2\gamma^{K/p} V_{\max}\right] \qquad \text{(A2)}$$

- $||\epsilon_k||_{p,\mu}$: error at each iteration k, **note that** $\epsilon_k = \mathcal{T}V_k - V_{k+1}$

AVI Error Propagation

Proposition (AVI L_p Error Bound) $\epsilon_k = \mathcal{T}V_k - V_{k+1}$

$$||V^* - V^{\pi_K}||_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2}\left[C_{\rho,\mu}^{1/p} \max_{0\leq k<K} ||\epsilon_k||_{p,\mu} + 2\gamma^{K/p} V_{\max}\right] \qquad \text{(A1)}$$

$$||V^* - V^{\pi_K}||_{\infty} \leq \frac{2\gamma}{(1-\gamma)^2}\left[C_{\mu}^{1/p} \max_{0\leq k<K} ||\epsilon_k||_{p,\mu} + 2\gamma^{K/p} V_{\max}\right] \qquad \text{(A2)}$$

- $||\epsilon_k||_{p,\mu}$: error at each iteration k, **note that** $\epsilon_k = \mathcal{T}V_k - V_{k+1}$
- $2\gamma^{K/p} V_{\max}$: initialization error $|V^* - V_0|$

AVI Error Propagation

Proposition (AVI L_p Error Bound) $\epsilon_k = \mathcal{T} V_k - V_{k+1}$

$$||V^* - V^{\pi_K}||_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2}\left[C_{\rho,\mu}^{1/p} \max_{0\leq k<K} ||\epsilon_k||_{p,\mu} + 2\gamma^{K/p} V_{\max}\right] \quad \text{(A1)}$$

$$||V^* - V^{\pi_K}||_{\infty} \leq \frac{2\gamma}{(1-\gamma)^2}\left[C_{\mu}^{1/p} \max_{0\leq k<K} ||\epsilon_k||_{p,\mu} + 2\gamma^{K/p} V_{\max}\right] \quad \text{(A2)}$$

- $||\epsilon_k||_{p,\mu}$: error at each iteration k, **note that** $\epsilon_k = \mathcal{T} V_k - V_{k+1}$
- $2\gamma^{K/p} V_{\max}$: initialization error $|V^* - V_0|$
- $C_{\rho,\mu}, C_{\mu}$: final performance is evaluated w.r.t. a measure $\rho \neq \mu$

AVI Error Propagation (Concentrability Coefficients)

Final performance is evaluated w.r.t. a measure $\rho \neq \mu, \; ||V^* - V^{\pi_K}||_{p,\rho}$

Assumption 1. (Uniformly Stochastic Transitions)

For all $x \in \mathcal{X}$ and $a \in \mathcal{A}$, there exists a constant $C_\mu < \infty$ such that $P(\cdot|x,a) \leq C_\mu \mu(\cdot)$.

Assumption 2. (Discounted-Average Concentrability of Future-State Distribution)

For any sequence of policies $\{\pi_m\}_{m\geq 1}$, there exists a constant $c_{\rho,\mu}(m) < \infty$ such that $\rho P^{\pi_1} P^{\pi_2} \ldots P^{\pi_m} \leq c_{\rho,\mu}(m)\mu$. We define

$$C_{\rho,\mu} = (1-\gamma)^2 \sum_{m\geq 1} m\gamma^{m-1} c_{\rho,\mu}(m)$$

- Note that $C_{\rho,\mu} \leq C_\mu$.

API Error Propagation

Error at each iteration k: $\qquad \epsilon_k = V_k - \mathcal{T}^{\pi_k} V_k$

π_K is a greedy policy w.r.t. V_{K-1} $\qquad \pi_K = \mathcal{G}(V_{K-1})$

Proposition (API Pointwise Error Bound)

$$V^* - V^{\pi_K} \leq \gamma \sum_{k=0}^{K-1} (\gamma P^{\pi^*})^{K-k-1} E_k |\epsilon_k| + (\gamma P^{\pi^*})^K |V^* - V^{\pi_0}|$$

where $\quad E_k = P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1} - P^{\pi^*}(I - \gamma P^{\pi_k})^{-1}$

API Error Propagation

Proposition (API L_p Error Bound) $\epsilon_k = V_k - \mathcal{T}^{\pi_k} V_k$

$$||V^* - V^{\pi_K}||_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} \left[C_{\rho,\mu}^{1/p} \max_{0 \leq k < K} ||\epsilon_k||_{p,\mu} + 2\gamma^{K/p} V_{\max} \right] \qquad \text{(A1)}$$

$$||V^* - V^{\pi_K}||_{\infty} \leq \frac{2\gamma}{(1-\gamma)^2} \left[C_{\mu}^{1/p} \max_{0 \leq k < K} ||\epsilon_k||_{p,\mu} + 2\gamma^{K/p} V_{\max} \right] \qquad \text{(A2)}$$

API Error Propagation

Proposition (API L_p Error Bound) $\epsilon_k = V_k - \mathcal{T}^{\pi_k} V_k$

$$||V^* - V^{\pi_K}||_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2}\left[C_{\rho,\mu}^{1/p} \max_{0\leq k<K} ||\epsilon_k||_{p,\mu} + 2\gamma^{K/p} V_{\max}\right] \qquad \text{(A1)}$$

$$||V^* - V^{\pi_K}||_{\infty} \leq \frac{2\gamma}{(1-\gamma)^2}\left[C_{\mu}^{1/p} \max_{0\leq k<K} ||\epsilon_k||_{p,\mu} + 2\gamma^{K/p} V_{\max}\right] \qquad \text{(A2)}$$

- $||\epsilon_k||_{p,\mu}$: error at each iteration k, **note that** $\epsilon_k = V_k - \mathcal{T}^{\pi_k} V_k$
- $2\gamma^{K/p} V_{\max}$: initialization error $|V^* - V^{\pi_0}|$
- $C_{\rho,\mu}, C_{\mu}$: final performance is evaluated w.r.t. a measure $\rho \neq \mu$

Inria

API Error Propagation (Concentrability Coefficients)

Final performance is evaluated w.r.t. a measure $\rho \neq \mu$, $||V^* - V^{\pi_K}||_{p,\rho}$

Assumption 1. (Uniformly Stochastic Transitions)

For all $x \in \mathcal{X}$ and $a \in \mathcal{A}$, there exists a constant $C_\mu < \infty$ such that $P(\cdot|x,a) \leq C_\mu \mu(\cdot)$.

Assumption 2. (Discounted-Average Concentrability of Future-State Distribution)

For any policy π and any non-negative integers s and t, there exists a constant $c_{\rho,\mu}(s,t) < \infty$ such that $\rho(P^*)^s(P^\pi)^t \leq c_{\rho,\mu}(s,t)\mu$. We define

$$C_{\rho,\mu} = (1-\gamma)^2 \sum_{s=0}^{\infty} \sum_{t=0}^{\infty} \gamma^{s+t} c_{\rho,\mu}(s,t)$$

- Note that $C_{\rho,\mu} \leq C_\mu$.

Finite-Sample Performance Bound of an AVI Algorithm

Approximate Value Iteration (AVI)

- if $\mathcal{F}$ is a function space, then V_{k+1} can be defined as

$$V_{k+1} = \inf_{V \in \mathcal{F}} ||V - \mathcal{T} V_k||_{?} = \Pi_{?} \mathcal{T} V_k$$

(*projection of $\mathcal{T} V_k$ into $\mathcal{F}$ according to the norm $L_?$*)

- if $V_{k+1} = \Pi_\infty \mathcal{T} V_k$ then AVI converges to the unique fixed-point of $\Pi_\infty \mathcal{T}$, i.e. $\widehat{V} \in \mathcal{F} : \ \widehat{V} = \Pi_\infty \mathcal{T} \widehat{V}$

(*$\mathcal{T}$ is a contraction in L_∞-norm and Π_∞ is non-expansive*)

- if we consider another norm, e.g. $L_2(\mu)$, then AVI does not necessarily converge

($\Pi_{2,\mu} \mathcal{T}$ *is not necessarily a contraction*)

An Approximate Value Iteration Algorithm

- Linear function space $\mathcal{F} = \{f : f(\cdot) = \sum_{j=1}^{d} \alpha_j \varphi_j(\cdot)\}$

$$\{\varphi_j\}_{j=1}^{d} \in \mathcal{B}\big((\mathcal{X}, \mathcal{A}); L\big) \quad , \quad \phi : (\mathcal{X}, \mathcal{A}) \to \mathbb{R}^d, \ \phi(\cdot) = \big(\varphi_1(\cdot), \ldots, \varphi_d(\cdot)\big)^\top$$

Fitted Q-Iteration (FQI)

At each iteration k:

- Generate N samples of the form (X_i, A_i, X'_i, R_i), where $(X_i, A_i) \sim \mu, \quad X'_i \sim p(\cdot|X_i, A_i), \quad R_i \sim r(X_i, A_i)$
- Build the training set $\mathcal{D}_k = \Big\{\big((X_i, A_i), \widehat{\mathcal{T}}Q_k(X_i, A_i)\big)\Big\}_{i=1}^{N}$, where $\widehat{\mathcal{T}}Q_k(X_i, A_i) = R_i + \gamma \max_{a \in \mathcal{A}} Q_k(X'_i, a)$
- $Q_{k+1} = \arg\min_{f \in \mathcal{F}} ||f - \widehat{\mathcal{T}}Q_k||_N^2 = \arg\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \big[f(X_i, A_i) - \widehat{\mathcal{T}}Q_k(X_i, A_i)\big]^2$ **(regression)**

Inria

FQI - Error at Each Iteration

Theorem (FQI - Error at Iteration k)

Let $\mathcal{F}$ be a d-dim linear space, $\mathcal{D}_k = \{(X_i, A_i, X'_i, R_i)\}_{i=1}^N$, $(X_i, A_i) \overset{\text{iid}}{\sim} \mu$, $X'_i \sim p(\cdot|X_i, A_i)$, $R_i = r(X_i, A_i)$, and $\widetilde{Q}$ be the *training set* and the *truncated solution* at the k'th iteration of FQI. Then with probability $1 - \delta$, we have

$$||\widetilde{Q} - \mathcal{T}Q_k||_\mu \leq 4 \inf_{f \in \mathcal{F}} ||f - \mathcal{T}Q_k||_\mu + O\left(||\alpha_k^*|| \sqrt{\frac{\log(1/\delta)}{N}}\right) + O\left(\sqrt{\frac{d \log(N/\delta)}{N}}\right).$$

- Note that $Q_{k+1} = \widetilde{Q}$.

FQI - Error at Each Iteration

Theorem (FQI - Error at Iteration k)

Let $\mathcal{F}$ be a d-dim linear space, $\mathcal{D}_k = \{(X_i, A_i, X'_i, R_i)\}_{i=1}^N$, $(X_i, A_i) \overset{\text{iid}}{\sim} \mu$, $X'_i \sim p(\cdot|X_i, A_i)$, $R_i = r(X_i, A_i)$, and $\widetilde{Q}$ be the *training set* and the *truncated solution* at the k'th iteration of FQI. Then with probability $1 - \delta$, we have

$$||\widetilde{Q} - \mathcal{T} Q_k||_\mu \leq 4 \inf_{f \in \mathcal{F}} ||f - \mathcal{T} Q_k||_\mu + O\left(||\alpha_k^*|| \sqrt{\frac{\log(1/\delta)}{N}} \right) + O\left(\sqrt{\frac{d \log(N/\delta)}{N}} \right).$$

- Note that $Q_{k+1} = \widetilde{Q}$.
- $N = \#$ of samples , $d =$ dimension of the linear function space $\mathcal{F}$

FQI - Error at Each Iteration

Theorem (FQI - Error at Iteration k)

Let $\mathcal{F}$ be a d-dim linear space, $\mathcal{D}_k = \{(X_i, A_i, X'_i, R_i)\}_{i=1}^N$, $(X_i, A_i) \overset{\text{iid}}{\sim} \mu$, $X'_i \sim p(\cdot|X_i, A_i)$, $R_i = r(X_i, A_i)$, and $\widetilde{Q}$ be the *training set* and the *truncated solution* at the k'th iteration of FQI. Then with probability $1-\delta$, we have

$$||\widetilde{Q} - \mathcal{T}Q_k||_\mu \le 4 \inf_{f \in \mathcal{F}} ||f - \mathcal{T}Q_k||_\mu + O\left(||\alpha_k^*|| \sqrt{\frac{\log(1/\delta)}{N}}\right) + O\left(\sqrt{\frac{d \log(N/\delta)}{N}}\right).$$

- Note that $Q_{k+1} = \widetilde{Q}$.
- $N = \#$ of samples , $d =$ dimension of the linear function space $\mathcal{F}$
- $\alpha_k^* \longrightarrow f_{\alpha_k^*} = \Pi_{2,\mu}\mathcal{T}Q_k$: the best approximation of $\mathcal{T}Q_k$ in $\mathcal{F}$ w.r.t. μ

FQI - Error at Each Iteration

FQI - Error at Iteration *k*

$$||\widetilde{Q} - \mathcal{T} Q_k||_\mu \leq \underbrace{4 \inf_{f \in \mathcal{F}} ||f - \mathcal{T} Q_k||_\mu}_{\text{approximation error}} + \underbrace{O\left(||\alpha_k^*|| \sqrt{\frac{\log(1/\delta)}{N}} \right) + O\left(\sqrt{\frac{d \log(N/\delta)}{N}} \right)}_{\text{estimation error}}$$

- **Approximation error:** it depends on how well the function space $\mathcal{F}$ can approximate $\mathcal{T} Q_k$
- **Estimation error:** it depends on the number of samples N, the dim of the function space d, and $||\alpha_k^*||$

FQI Error Bound

Theorem (FQI Error Bound)

Let $Q_{-1} \in \widetilde{\mathcal{F}}$ be an arbitrary initial value function, $\widetilde{Q}_0, \ldots, \widetilde{Q}_{K-1}$ be the sequence of truncated action-value functions generated by FQI after K iterations, and π_K be the greedy policy w.r.t. $\widetilde{Q}_{K-1}$. Then with probability $1 - \delta$, we have

$$||V^* - V^{\pi_K}||_\rho \leq \frac{2\gamma}{(1-\gamma)^2} \left\{ \sqrt{C_{\rho,\mu}} \left[d_\mu(\mathcal{T}\mathcal{F}, \mathcal{F}) + O\left(Q_{\max} \sqrt{\frac{\log(K/\delta)}{N\ \nu_\mu}} \right) + O\left(\sqrt{\frac{d \log(NK/\delta)}{N}} \right) \right] + 2\gamma^{K/2} Q_{\max} \right\}$$

FQI Error Bound

Theorem (FQI Error Bound)

Let $Q_{-1} \in \widetilde{\mathcal{F}}$ be an arbitrary initial value function, $\widetilde{Q}_0, \ldots, \widetilde{Q}_{K-1}$ be the sequence of truncated action-value functions generated by FQI after K iterations, and π_K be the greedy policy w.r.t. $\widetilde{Q}_{K-1}$. Then with probability $1 - \delta$, we have

$$||V^* - V^{\pi_K}||_\rho \leq \frac{2\gamma}{(1-\gamma)^2}\left\{\sqrt{C_{\rho,\mu}}\left[d_\mu(\mathcal{T}\mathcal{F}, \mathcal{F}) + O\left(Q_{\max}\sqrt{\frac{\log(K/\delta)}{N\ \nu_\mu}}\right)\right.\right.$$

$$\left.\left. + O\left(\sqrt{\frac{d\log(NK/\delta)}{N}}\right)\right] + 2\gamma^{K/2}Q_{\max}\right\}$$

- **Approximation error:** $d_\mu(\mathcal{T}\mathcal{F}, \mathcal{F}) = \sup_{f\in\widetilde{\mathcal{F}}} \inf_{g\in\mathcal{F}} ||g - \mathcal{T}f||_\mu$

FQI Error Bound

Theorem (FQI Error Bound)

Let $Q_{-1} \in \widetilde{\mathcal{F}}$ be an arbitrary initial value function, $\widetilde{Q}_0, \ldots, \widetilde{Q}_{K-1}$ be the sequence of truncated action-value functions generated by FQI after K iterations, and π_K be the greedy policy w.r.t. $\widetilde{Q}_{K-1}$. Then with probability $1-\delta$, we have

$$||V^* - V^{\pi_K}||_\rho \leq \frac{2\gamma}{(1-\gamma)^2}\left\{\sqrt{C_{\rho,\mu}}\left[d_\mu(\mathcal{T}\mathcal{F},\mathcal{F}) + O\left(Q_{\max}\sqrt{\frac{\log(K/\delta)}{N\ \nu_\mu}}\right) + O\left(\sqrt{\frac{d\log(NK/\delta)}{N}}\right)\right] + 2\gamma^{K/2}Q_{\max}\right\}$$

- **Approximation error:** $d_\mu(\mathcal{T}\mathcal{F},\mathcal{F}) = \sup_{f\in\widetilde{\mathcal{F}}}\inf_{g\in\mathcal{F}}||g - \mathcal{T}f||_\mu$
- **Estimation error:** depends on N, d, ν_μ, K. Note that $||\alpha_k^*|| \leq \frac{Q_{\max}}{\nu_\mu}$
  $\nu_\mu =$ the smallest eigenvalue of the Gram matrix $(\int \varphi_i\ \varphi_j\ d\mu)_{i,j}$

FQI Error Bound

Theorem (FQI Error Bound)

Let $Q_{-1} \in \widetilde{\mathcal{F}}$ be an arbitrary initial value function, $\widetilde{Q}_0, \ldots, \widetilde{Q}_{K-1}$ be the sequence of truncated action-value functions generated by FQI after K iterations, and π_K be the greedy policy w.r.t. $\widetilde{Q}_{K-1}$. Then with probability $1-\delta$, we have

$$||V^* - V^{\pi_K}||_\rho \leq \frac{2\gamma}{(1-\gamma)^2}\left\{\sqrt{C_{\rho,\mu}}\left[d_\mu(\mathcal{T}\mathcal{F},\mathcal{F}) + O\left(Q_{\max}\sqrt{\frac{\log(K/\delta)}{N\,\nu_\mu}}\right) + O\left(\sqrt{\frac{d\log(NK/\delta)}{N}}\right)\right] + 2\gamma^{K/2}Q_{\max}\right\}$$

- **Approximation error:** $d_\mu(\mathcal{T}\mathcal{F},\mathcal{F}) = \sup_{f\in\widetilde{\mathcal{F}}}\inf_{g\in\mathcal{F}}||g - \mathcal{T}f||_\mu$
- **Estimation error:** depends on N, d, ν_μ, K. Note that $||\alpha_k^*|| \leq \frac{Q_{\max}}{\nu_\mu}$
  ν_μ = the smallest eigenvalue of the Gram matrix $(\int \varphi_i\,\varphi_j\,d\mu)_{i,j}$
- **Initialization error:** error due to the choice of the initial action-value function $|Q^* - Q_0|$

Finite-Sample Performance Bound of an API Algorithm

Least-Squares Temporal-Difference Learning (LSTD)

- Linear function space $\mathcal{F} = \{f : f(\cdot) = \sum_{j=1}^{d} \alpha_j \varphi_j(\cdot)\}$

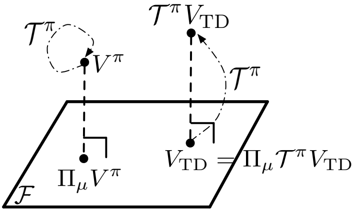
$$\{\varphi_j\}_{j=1}^{d} \in \mathcal{B}(\mathcal{X}; L) \quad , \quad \phi : \mathcal{X} \to \mathbb{R}^d,\ \phi(\cdot) = \big(\varphi_1(\cdot), \ldots, \varphi_d(\cdot)\big)^\top$$

- V^π is the fixed-point of $\mathcal{T}^\pi$ $\qquad \mathcal{T}^\pi V^\pi = V^\pi$
- V^π may not belong to $\mathcal{F}$ $\qquad V^\pi \notin \mathcal{F}$
- LSTD searches for the fixed-point of $\Pi_? \mathcal{T}^\pi$ instead ($\Pi_?$ is a projection into $\mathcal{F}$ w.r.t. $L_?$-norm)
- $\Pi_\infty \mathcal{T}^\pi$ is a contraction in L_∞-norm
  - L_∞-projection is numerically expensive when the number of states is large or infinite
- LSTD searches for the fixed-point of $\Pi_{2,\mu} \mathcal{T}^\pi$
  $\Pi_{2,\mu} g = \arg\min_{f \in \mathcal{F}} ||f - g||_{2,\mu}$

Least-Squares Temporal-Difference Learning (LSTD)

When the fixed-point of $\Pi_\mu \mathcal{T}^\pi$ exists, we call it the LSTD solution

$V_{\text{TD}} = \Pi_\mu \mathcal{T}^\pi V_{\text{TD}}$



$$\langle \mathcal{T}^\pi V_{\text{TD}} - V_{\text{TD}}, \varphi_i \rangle_\mu = 0, \qquad i = 1, \ldots, d$$

$$\langle r^\pi + \gamma P^\pi V_{\text{TD}} - V_{\text{TD}}, \varphi_i \rangle_\mu = 0$$

$$\underbrace{\langle r^\pi, \varphi_i \rangle_\mu}_{b_i} - \sum_{i=1}^{d} \underbrace{\langle \varphi_j - \gamma P^\pi \varphi_j, \varphi_i \rangle_\mu}_{A_{ij}} \cdot \alpha_{\text{TD}}^{(j)} = 0 \quad \longrightarrow \quad A\, \alpha_{\text{TD}} = b$$

- In general, $\Pi_\mu \mathcal{T}^\pi$ is not a contraction and does not have a fixed-point.
- If $\mu = \mu^\pi$, the stationary dist. of π, then $\Pi_{\mu^\pi} \mathcal{T}^\pi$ has a unique fixed-point.

LSTD Algorithm

Proposition (LSTD Performance)

$$||V^\pi - V_{\text{TD}}||_{\mu^\pi} \leq \frac{1}{\sqrt{1-\gamma^2}} \inf_{V \in \mathcal{F}} ||V^\pi - V||_{\mu^\pi}$$

LSTD Algorithm

- We observe a trajectory generated by following the policy π $(X_0, R_0, X_1, R_1, \ldots, X_N)$ where $X_{t+1} \sim P(\cdot|X_t, \pi(X_t))$ and $R_t = r(X_t, \pi(X_t))$
- We build estimators of the matrix A and vector b

$$\widehat{A}_{ij} = \frac{1}{N}\sum_{t=0}^{N-1} \varphi_i(X_t)\big[\varphi_j(X_t) - \gamma\varphi_j(X_{t+1})\big] \quad , \quad \widehat{b}_i = \frac{1}{N}\sum_{t=0}^{N-1} \varphi_i(X_t) R_t$$

- $\widehat{A}\widehat{\alpha}_{\text{TD}} = \widehat{b} \quad , \quad \widehat{V}_{\text{TD}}(\cdot) = \phi(\cdot)^\top \widehat{\alpha}_{\text{TD}}$

when $n \to \infty$ then $\widehat{A} \to A$ and $\widehat{b} \to b$, and thus, $\widehat{\alpha}_{\text{TD}} \to \alpha_{\text{TD}}$ and $\widehat{V}_{\text{TD}} \to V_{\text{TD}}$.

LSTD Error Bound

When the Markov chain induced by the policy under evaluation π has a stationary distribution μ^π (Markov chain is ergodic - e.g. β-mixing), then

Theorem (LSTD Error Bound)

Let $\widetilde{V}$ be the truncated LSTD solution computed using n samples along a trajectory generated by following the policy π. Then with probability $1 - \delta$, we have

$$||V^\pi - \widetilde{V}||_{\mu^\pi} \leq \frac{c}{\sqrt{1-\gamma^2}} \inf_{f \in \mathcal{F}} ||V^\pi - f||_{\mu^\pi} + O\left(\sqrt{\frac{d \log(d/\delta)}{n\, \nu}}\right)$$

- $n = \#$ of samples , $d =$ dimension of the linear function space $\mathcal{F}$
- $\nu =$ the smallest eigenvalue of the Gram matrix $(\int \varphi_i \, \varphi_j \, d\mu^\pi)_{i,j}$
(**Assume:** *eigenvalues of the Gram matrix are strictly positive - existence of the model-based LSTD solution*)
- β-mixing coefficients are hidden in O notation

LSTD Error Bound

LSTD Error Bound

$$||V^\pi - \widetilde{V}||_{\mu^\pi} \leq \frac{c}{\sqrt{1-\gamma^2}} \underbrace{\inf_{f \in \mathcal{F}} ||V^\pi - f||_{\mu^\pi}}_{\text{approximation error}} + \underbrace{O\left(\sqrt{\frac{d \log(d/\delta)}{n\,\nu}}\right)}_{\text{estimation error}}$$

- **Approximation error:** it depends on how well the function space $\mathcal{F}$ can approximate the value function V^π
- **Estimation error:** it depends on the number of samples n, the dim of the function space d, the smallest eigenvalue of the Gram matrix ν, the mixing properties of the Markov chain (hidden in O)

LSPI Error Bound

Theorem (LSPI Error Bound)

Let $V_{-1} \in \widetilde{\mathcal{F}}$ be an arbitrary initial value function, $\widetilde{V}_0, \dots, \widetilde{V}_{K-1}$ be the sequence of truncated value functions generated by LSPI after K iterations, and π_K be the greedy policy w.r.t. $\widetilde{V}_{K-1}$. Then with probability $1-\delta$, we have

$$||V^* - V^{\pi_K}||_\rho \leq \frac{4\gamma}{(1-\gamma)^2} \left\{ \sqrt{CC_{\rho,\mu}} \left[cE_0(\mathcal{F}) + O\left(\sqrt{\frac{d \log(dK/\delta)}{n\ \nu_\mu}} \right) \right] + \gamma^{\frac{K-1}{2}} R_{\max} \right\}$$

LSPI Error Bound

Theorem (LSPI Error Bound)

Let $V_{-1} \in \widetilde{\mathcal{F}}$ be an arbitrary initial value function, $\widetilde{V}_0, \ldots, \widetilde{V}_{K-1}$ be the sequence of truncated value functions generated by LSPI after K iterations, and π_K be the greedy policy w.r.t. $\widetilde{V}_{K-1}$. Then with probability $1-\delta$, we have

$$||V^* - V^{\pi_K}||_\rho \le \frac{4\gamma}{(1-\gamma)^2}\left\{\sqrt{CC_{\rho,\mu}}\left[cE_0(\mathcal{F}) + O\left(\sqrt{\frac{d\log(dK/\delta)}{n\,\nu_\mu}}\right)\right] + \gamma^{\frac{K-1}{2}} R_{\max}\right\}$$

- **Approximation error:** $E_0(\mathcal{F}) = \sup_{\pi \in \mathcal{G}(\widetilde{\mathcal{F}})} \inf_{f \in \mathcal{F}} ||V^\pi - f||_{\mu^\pi}$

LSPI Error Bound

Theorem (LSPI Error Bound)

Let $V_{-1} \in \widetilde{\mathcal{F}}$ be an arbitrary initial value function, $\widetilde{V}_0, \ldots, \widetilde{V}_{K-1}$ be the sequence of truncated value functions generated by LSPI after K iterations, and π_K be the greedy policy w.r.t. $\widetilde{V}_{K-1}$. Then with probability $1-\delta$, we have

$$||V^* - V^{\pi_K}||_\rho \leq \frac{4\gamma}{(1-\gamma)^2} \left\{ \sqrt{CC_{\rho,\mu}} \left[cE_0(\mathcal{F}) + O\left(\sqrt{\frac{d \log(dK/\delta)}{n\, \nu_\mu}} \right) \right] + \gamma^{\frac{K-1}{2}} R_{\max} \right\}$$

- **Approximation error:** $E_0(\mathcal{F}) = \sup_{\pi \in \mathcal{G}(\widetilde{\mathcal{F}})} \inf_{f \in \mathcal{F}} ||V^\pi - f||_{\mu^\pi}$
- **Estimation error:** depends on n, d, ν_μ, K

LSPI Error Bound

Theorem (LSPI Error Bound)

Let $V_{-1} \in \widetilde{\mathcal{F}}$ be an arbitrary initial value function, $\widetilde{V}_0, \ldots, \widetilde{V}_{K-1}$ be the sequence of truncated value functions generated by LSPI after K iterations, and π_K be the greedy policy w.r.t. $\widetilde{V}_{K-1}$. Then with probability $1-\delta$, we have

$$||V^* - V^{\pi_K}||_\rho \leq \frac{4\gamma}{(1-\gamma)^2} \left\{ \sqrt{CC_{\rho,\mu}} \left[cE_0(\mathcal{F}) + O\left(\sqrt{\frac{d \log(dK/\delta)}{n\, \nu_\mu}} \right) \right] + \gamma^{\frac{K-1}{2}} R_{\max} \right\}$$

- **Approximation error:** $E_0(\mathcal{F}) = \sup_{\pi \in \mathcal{G}(\widetilde{\mathcal{F}})} \inf_{f \in \mathcal{F}} ||V^\pi - f||_{\mu^\pi}$
- **Estimation error:** depends on n, d, ν_μ, K
- **Initialization error:** error due to the choice of the initial value function or initial policy $|V^* - V^{\pi_0}|$

LSPI Error Bound

LSPI Error Bound

$$||V^* - V^{\pi_K}||_\rho \leq \frac{4\gamma}{(1-\gamma)^2} \left\{ \sqrt{C C_{\rho,\mu}} \left[c E_0(\mathcal{F}) + O\left(\sqrt{\frac{d \log(dK/\delta)}{n\, \nu_\mu}} \right) \right] + \gamma^{\frac{K-1}{2}} R_{\max} \right\}$$

Lower-Bounding Distribution

There exists a distribution μ such that for any policy $\pi \in \mathcal{G}(\widetilde{\mathcal{F}})$, we have $\mu \leq C\mu^\pi$, where $C < \infty$ is a constant and μ^π is the stationary distribution of π. Furthermore, we can define the concentrability coefficient $C_{\rho,\mu}$ as before.

LSPI Error Bound

LSPI Error Bound

$$||V^* - V^{\pi_K}||_\rho \leq \frac{4\gamma}{(1-\gamma)^2}\left\{\sqrt{CC_{\rho,\mu}}\left[cE_0(\mathcal{F}) + O\left(\sqrt{\frac{d\log(dK/\delta)}{n\,\nu_\mu}}\right)\right] + \gamma^{\frac{K-1}{2}}R_{\max}\right\}$$

Lower-Bounding Distribution

There exists a distribution μ such that for any policy $\pi \in \mathcal{G}(\widetilde{\mathcal{F}})$, we have $\mu \leq C\mu^\pi$, where $C < \infty$ is a constant and μ^π is the stationary distribution of π. Furthermore, we can define the concentrability coefficient $C_{\rho,\mu}$ as before.

- $\nu_\mu =$ the smallest eigenvalue of the Gram matrix $(\int \varphi_i \, \varphi_j \, d\mu)_{i,j}$

Discussion

we obtain the optimal rate of regression and classification for RL (ADP) algorithms

What makes RL more challenging then?

- the propagation of error (control problem)
- the approximation error is more complex
- the sampling problem (how to choose μ - exploration problem)

Other Finite-Sample Analysis Results in RL

- Approximate Value Iteration ***[MS08]***
- Approximate Policy Iteration
  - LSTD and LSPI ***[LGM10, LGM11]***
  - Bellman Residual Minimization ***[MMLG10]***
  - Modified Bellman Residual Minimization ***[ASM08]***
  - Classification-based Policy Iteration ***[FYG06, LGM10, GLGS11]***
- Regularized Approximate Dynamic Programming
  - L_2-Regularization
    - L_2-Regularized Policy Iteration ***[FGSM08]***
    - L_2-Regularized Fitted Q-Iteration ***[FGSM09]***
  - L_1-Regularization and High-Dimensional RL
    - Lasso-TD ***[GLMH11]***
    - LSTD (LSPI) with Random Projections ***[GLMM10]***

Bibliography I

Antos, A., Szepesvári, Cs., and Munos, R.
Learning Near-Optimal Policies with Bellman Residual Minimization-based Fitted Policy Iteration and a Single Sample Path.
Machine Learning Journal, 71:89–129, 2008.

Farahmand, A., Ghavamzadeh, M., Szepesvári Cs., and Mannor, S.
Regularized Policy Iteration.
Proceedings of Advances in Neural Information Processing Systems 21, pp. 441–448, 2008.

Farahmand, A., Ghavamzadeh, M., Szepesvári Cs., and Mannor, S.
Regularized Fitted Q-iteration for Planning in Continuous-Space Markovian Decision Problems.
Proceedings of the American Control Conference, pp. 725–730, 2009.

Fern, A., Yoon, S., and Givan, R.
Approximate Policy Iteration with a Policy Language Bias: Solving Relational Markov Decision Processes.
Journal of Artificial Intelligence Research, 25:85–118, 2006.

Gabillon, V., Lazaric, A., Ghavamzadeh, M., and Scherrer, B.
Classification-based Policy Iteration with a Critic.
Proceedings of the Twenty-Eighth International Conference on Machine Learning, pp. 1049–1056, 2011.

Ghavamzadeh, M., Lazaric A., Munos, R., and Hoffman, M.
Finite-Sample Analysis of Lasso-TD.
Proceedings of the Twenty-Eighth International Conference on Machine Learning, pp. 1177–1184, 2011.

Ghavamzadeh, M., Lazaric, A., Maillard, O., and Munos, R.
LSTD with Random Projections.
Proceedings of Advances in Neural Information Processing Systems 23, pp. 721–729, 2010.

Bibliography II

Lazaric A., Ghavamzadeh, M., and Munos, R.
Analysis of a Classification-based Policy Iteration Algorithm.
Proceedings of the Twenty-Seventh International Conference on Machine Learning, pp. 607–614, 2010.

Lazaric A., Ghavamzadeh, M., and Munos, R.
Finite-Sample Analysis of LSTD.
Proceedings of the Twenty-Seventh International Conference on Machine Learning, pp. 615–622, 2010.

Lazaric A., Ghavamzadeh, M., and Munos, R.
Finite-Sample Analysis of Least-Squares Policy Iteration.
Accepted at the Journal of Machine Learning Research, 2011.

Maillard, O., Munos, R., Lazaric A., and Ghavamzadeh, M.
Finite-Sample Analysis of Bellman Residual Minimization.
Proceedings of the Second Asian Conference on Machine Learning, pp. 299–314, 2010.

Munos, R. and Szepesvári, Cs.
Finite-Time Bounds for Fitted Value Iteration.
Journal of Machine Learning Research, 9:815–857, 2008.

Munos, R.
Performance Bounds in Lp-norm for Approximate Value Iteration.
SIAM Journal of Control and Optimization, 2007.

Munos, R.
Error Bounds for Approximate Policy Iteration.
Proceedings of the Nineteenth International Conference on Machine Learning, pp. 560–567, 2003.