Sample Complexity of ADP Algorithms

A. LAZARIC (*SequeL Team @INRIA-Lille*)
ENS Cachan - Master 2 MVA

SequeL – INRIA Lille

Sources of Error

- **Approximation error**. If X is *large* or *continuous*, value functions V cannot be *represented* correctly
  $\Rightarrow$ use an *approximation space* $\mathcal{F}$
- **Estimation error**. If the reward r and dynamics p are *unknown*, the Bellman operators $\mathcal{T}$ and $\mathcal{T}^\pi$ cannot be *computed* exactly
  $\Rightarrow$ *estimate* the Bellman operators from *samples*

In This Lecture

- Infinite horizon setting with discount γ
- Study the impact of estimation error

In This Lecture: *Warning!!*

Problem: are these performance bounds accurate/useful?

Answer: of course not! :)

Reason: upper bounds, non-tight analysis, worst case.

In This Lecture: *Warning!!*

Chernoff-Hoeffding inequality

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{t=1}^{n} X_t - \mathbb{E}[X_1]\right| > (b-a)\sqrt{\frac{\log 2/\delta}{2n}}\right] \leq \delta$$

$\Rightarrow$ worst-case w.r.t. to all the distributions bounded in $[a, b]$, loose for other distributions.

In This Lecture: *Warning!!*

Question: so why should we derive/study these bounds?

Answer:

- General guarantees
- Rates of convergence (not always available in asymptotic analysis)
- Explicit dependency on the design parameters
- Explicit dependency on the problem parameters
- First guess on how to tune parameters
- Better understanding of the algorithms

Outline

Sample Complexity of LSTD
The Algorithm
LSTD and LSPI Error Bounds

Sample Complexity of Fitted Q-iteration

Outline

Sample Complexity of LSTD

The Algorithm

LSTD and LSPI Error Bounds

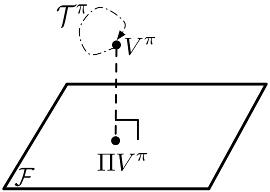
Sample Complexity of Fitted Q-iteration

Least-Squares Temporal-Difference Learning (LSTD)

- Linear function space $\mathcal{F} = \{f : f(\cdot) = \sum_{j=1}^{d} \alpha_j \varphi_j(\cdot)\}$
- V^π is the fixed-point of $\mathcal{T}^\pi$ $\qquad V^\pi = \mathcal{T}^\pi V^\pi$
- V^π may not belong to $\mathcal{F}$ $\qquad V^\pi \notin \mathcal{F}$
- Best approximation of V^π in $\mathcal{F}$ is

$$\Pi V^\pi = \arg\min_{f \in \mathcal{F}} ||V^\pi - f||$$

(Π is the projection onto $\mathcal{F}$)



Least-Squares Temporal-Difference Learning (LSTD)

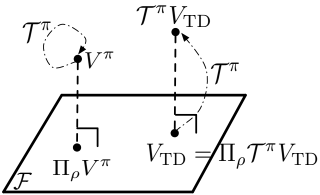
- LSTD searches for the fixed-point of $\Pi_? \mathcal{T}^\pi$ instead ($\Pi_?$ is a projection into $\mathcal{F}$ w.r.t. $L_?$-norm)
- $\Pi_\infty \mathcal{T}^\pi$ is a contraction in L_∞-norm
  - L_∞-projection is numerically expensive when the number of states is large or infinite
- LSTD searches for the fixed-point of $\Pi_{2,\rho} \mathcal{T}^\pi$

$$\Pi_{2,\rho}\, g = \arg\min_{f \in \mathcal{F}} ||g - f||_{2,\rho}$$

Least-Squares Temporal-Difference Learning (LSTD)

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

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



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

LSTD Algorithm

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

Proposition (LSTD Performance)

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

LSTD Algorithm

Empirical LSTD

- We observe a trajectory $(X_0, R_0, X_1, R_1, \ldots, X_N)$ 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

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

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

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 $\widehat{V}_{\text{TD}} \to V_{\text{TD}}$.

Outline

Sample Complexity of LSTD
- The Algorithm
- LSTD and LSPI Error Bounds

Sample Complexity of Fitted Q-iteration

LSTD Error Bound

When the Markov chain induced by the policy under evaluation π has a stationary distribution ρ^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

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

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

LSTD Error Bound

LSTD Error Bound

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

- **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, $\widetilde{V}_0, \ldots, \widetilde{V}_{K-1}$ 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

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

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

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

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

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

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

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

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

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

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

Lower-Bounding Distribution

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

LSPI Error Bound

LSPI Error Bound

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

Lower-Bounding Distribution

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

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

Outline

Sample Complexity of LSTD

Sample Complexity of Fitted Q-iteration

- Error at Each Iteration
- Error Propagation
- The Final Bound

Linear Fitted Q-iteration

Input: space $\mathcal{F}$, iterations K, sampling distribution ρ, num of samples n

Initial function $\widetilde{Q}^0 \in \mathcal{F}$
For $k = 1, \dots, K$

- Draw n samples $(x_i, a_i) \overset{\text{i.i.d}}{\sim} \rho$
- Sample $x_i' \sim p(\cdot|x_i, a_i)$ and $r_i = r(x_i, a_i)$
- Compute $y_i = r_i + \gamma \max_a \widetilde{Q}^{k-1}(x_i', a)$
- Build training set $\big\{\big((x_i, a_i), y_i\big)\big\}_{i=1}^n$
- Solve the *least squares problem*

$$f_{\hat{\alpha}_k} = \arg\min_{f_\alpha \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \big(f_\alpha(x_i, a_i) - y_i\big)^2$$

- Return $\widetilde{Q}^k = \text{Trunc}(f_{\hat{\alpha}_k})$

Return $\pi_K(\cdot) = \arg\max_a \widetilde{Q}^K(\cdot, a)$ (*greedy policy*)

Theoretical Objectives

Objective 1: derive a bound on the performance (*quadratic*) loss w.r.t. a *testing* distribution μ

$$||Q^* - Q^{\pi_K}||_\mu \leq \ ???$$

Outline

Sample Complexity of LSTD

Sample Complexity of Fitted Q-iteration
- Error at Each Iteration
- Error Propagation
- The Final Bound

Linear Fitted Q-iteration

Input: space $\mathcal{F}$, iterations K, sampling distribution ρ

Initial function $\widetilde{Q}^0 \in \mathcal{F}$
For $k = 1, \ldots, K$

- Draw n samples $(x_i, a_i) \overset{\text{i.i.d}}{\sim} \rho$
- Sample $x_i' \sim p(\cdot|x_i, a_i)$ and $r_i = r(x_i, a_i)$
- Compute $y_i = r_i + \gamma \max_a \widetilde{Q}^{k-1}(x_i', a)$
- Build training set $\big\{\big((x_i, a_i), y_i\big)\big\}_{i=1}^n$
- Solve the *least squares problem*

$$f_{\hat{\alpha}_k} = \arg\min_{f_\alpha \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \big(f_\alpha(x_i, a_i) - y_i\big)^2$$

- Return $\widetilde{Q}^k = \text{Trunc}(f_{\hat{\alpha}_k})$

Return $\pi_K(\cdot) = \arg\max_a \widetilde{Q}^K(\cdot, a)$ (*greedy policy*)

Linear Fitted Q-iteration

- Draw n samples $(x_i, a_i) \overset{\text{i.i.d}}{\sim} \rho$
- Sample $x'_i \sim p(\cdot|x_i, a_i)$ and $r_i = r(x_i, a_i)$
- Compute $y_i = r_i + \gamma \max_a \widetilde{Q}^{k-1}(x'_i, a)$
- Build training set $\big\{\big((x_i, a_i), y_i\big)\big\}_{i=1}^n$
- Solve the *least squares problem*

$$f_{\hat{\alpha}_k} = \arg\min_{f_\alpha \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \big(f_\alpha(x_i, a_i) - y_i\big)^2$$

- Return $\widetilde{Q}^k = \text{Trunc}(f_{\hat{\alpha}_k})$

Theoretical Objectives

Target: at each iteration we want to approximate $Q^k = \mathcal{T}\widetilde{Q}^{k-1}$

Objective 2: derive an *intermediate* bound on the prediction error [*random design*]

$$||Q^k - \widetilde{Q}^k||_\rho \leq \text{ ???}$$

Theoretical Objectives

Target: at each iteration we have samples $\{(x_i, a_i)\}_{i=1}^n$ (from ρ)

Objective 3: derive an *intermediate* bound on the prediction error ***on the samples*** [*deterministic design*]

$$\frac{1}{n}\sum_{i=1}^{n}\Big(Q^k(x_i, a_i) - \widetilde{Q}^k(x_i, a_i)\Big)^2 = ||Q^k - \widetilde{Q}^k||_{\hat{\rho}}^2 \leq \ ???$$

Theoretical Objectives

Obj 3

$$||Q^k - \widetilde{Q}^k||_{\hat{\rho}} \leq \ ???$$

$\Rightarrow$ **Obj 2**

$$||Q^k - \widetilde{Q}^k||_{\rho} \leq \ ???$$

$\Rightarrow$ **Obj 1**

$$||Q^* - Q^{\pi_K}||_{\mu} \leq \ ???$$

Theoretical Objectives

Returned solution

$$f_{\hat{\alpha}_k} = \arg\min_{f_\alpha \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \big(f_\alpha(x_i, a_i) - y_i \big)^2$$

Best solution

$$f_{\alpha_k^*} = \arg\inf_{f_\alpha \in \mathcal{F}} ||f_\alpha - Q^k||_\rho$$

Additional Notation

Given the set of inputs $\{(x_i, a_i)\}_{i=1}^n$ drawn from ρ.
Vector space

$$\mathcal{F}_n = \{z \in \mathbb{R}^n, z_i = f_\alpha(x_i, a_i); f_\alpha \in \mathcal{F}\} \subset \mathbb{R}^n$$

Empirical L_2-norm

$$||f_\alpha||_{\hat{\rho}}^2 = \frac{1}{n}\sum_{i=1}^n f_\alpha(x_i, a_i)^2 = \frac{1}{n}\sum_{i=1}^n z_i^2 = ||z||_n^2$$

Empirical orthogonal projection

$$\widehat{\Pi} y = \arg\min_{z \in \mathcal{F}_n} ||y - z||_n$$

Additional Notation

- Target vector:

$$q_i = Q^k(x_i, a_i) = \mathcal{T}\widetilde{Q}^{k-1}(x_i, a_i)$$
$$= r(x_i, a_i) + \gamma \max_a \int_X \widetilde{Q}^{k-1}(dx', a)p(dx'|x_i, a_i)$$

- Observed target vector:

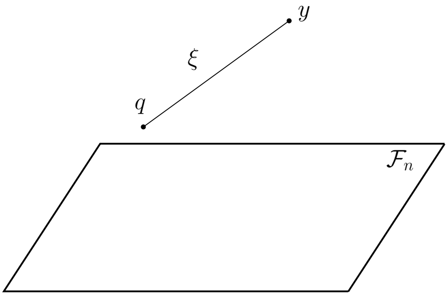
$$y_i = r_i + \gamma \max_a \widetilde{Q}^{k-1}(x_i', a)$$

- Noise vector (zero–mean and bounded):

$$\xi_i = q_i - y_i$$

$$|\xi_i| \leq V_{\max} \qquad \mathbb{E}[\xi_i | x_i] = 0$$

Additional Notation



Additional Notation

- Optimal solution in $\mathcal{F}_n$

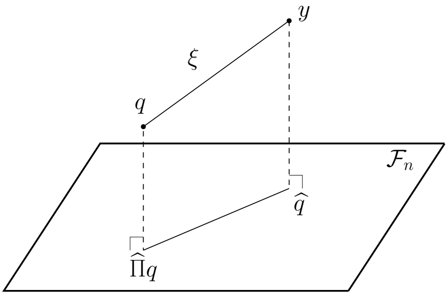
$$\widehat{\Pi}q = \arg\min_{z \in \mathcal{F}_n} ||q - z||_n$$

- Returned vector

$$\widehat{q}_i = f_{\hat{\alpha}_k}(x_i, a_i)$$

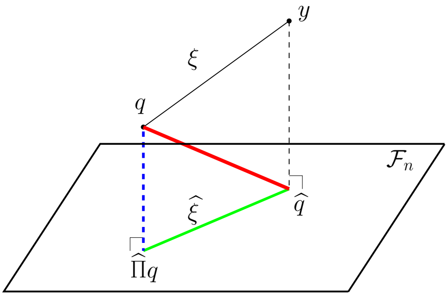
$$\widehat{q} = \widehat{\Pi}y = \arg\min_{z \in \mathcal{F}_n} ||y - z||_n$$

Additional Notation



Theoretical Analysis

$$||Q^k - f_{\hat{\alpha}^k}||_{\hat{\rho}}^2 = ||q - \widehat{q}||_n^2$$



$$||q - \widehat{q}||_n \leq ||q - \widehat{\Pi}q||_n + ||\widehat{\Pi}q - \widehat{q}||_n = ||q - \widehat{\Pi}q||_n + ||\widehat{\xi}||_n$$

Theoretical Analysis

$$\underbrace{||q - \widehat{q}||_n}_{\text{prediction err}} \leq \underbrace{||q - \widehat{\Pi} q||_n}_{\text{approx. err}} + \underbrace{||\widehat{\xi}||_n}_{\text{estim. err}}$$

- ***Prediction error***: distance between *learned* function and *target* function
- ***Approximation error***: distance between the *best* function in $\mathcal{F}$ and the *target* function $\Rightarrow$ depends on $\mathcal{F}$
- ***Estimation error***: distance between the *best* function in $\mathcal{F}$ and the *learned* function $\Rightarrow$ depends on the samples

Theoretical Analysis

The noise $\widehat{\xi} = \widehat{\Pi}\xi$

$$\Rightarrow ||\widehat{\xi}||_n = \langle \widehat{\xi}, \widehat{\xi} \rangle = \langle \widehat{\xi}, \xi \rangle$$

The projected noise belongs to $\mathcal{F}_n$

$$\Rightarrow \exists f_\beta \in \mathcal{F} : f_\beta(x_i, a_i) = \widehat{\xi}_i, \quad \forall (x_i, a_i)$$

By definition of inner product

$$\Rightarrow ||\widehat{\xi}||_n = \frac{1}{n} \sum_{i=1}^{n} f_\beta(x_i, a_i)\xi_i$$

Theoretical Analysis

The noise ξ has zero mean and it is bounded in $[-V_{\max}, V_{\max}]$
Thus for any ***fixed*** $f_\beta \in \mathcal{F}$ (the expectation is *conditioned* on (x_i, a_i))

$$\Rightarrow \mathbb{E}_\xi\Big[\frac{1}{n}\sum_{i=1}^{n} f_\beta(x_i, a_i)\xi_i\Big] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\xi\big[f_\beta(x_i, a_i)\xi_i\big] = 0$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n}\big(f_\beta(x_i, a_i)\xi_i\big)^2 \leq 4{V_{\max}}^2\frac{1}{n}\sum_{i=1}^{n} f_\beta(x_i, a_i)^2 = 4V_{\max}||f_\beta||_{\hat{\rho}}^2$$

$\Rightarrow$ we can use *concentration inequalities*

Theoretical Analysis

Problem: f_β is a *random function*
Solution: we need *functional concentration inequalities*

Theoretical Analysis

Define the space of *normalized functions*

$$\mathcal{G} = \left\{ g(\cdot) = \frac{f_\alpha(\cdot)}{||f_\alpha||_{\hat{\rho}}}, f_\alpha \in \mathcal{F} \right\}$$

[by definition] $\Rightarrow \forall g \in \mathcal{G}, ||g||_{\hat{\rho}} \leq 1$

[$\mathcal{F}$ is a linear space] $\Rightarrow \mathcal{V}(\mathcal{G}) = d + 1$

Theoretical Analysis

Application of Pollard's inequality for space $\mathcal{G}$

For any $g \in \mathcal{G}$

$$\left|\frac{1}{n}\sum_{i=1}^{n} g(x_i, a_i)\xi_i\right| \leq 4V_{\max}\sqrt{\frac{2}{n}\log\left(\frac{3(9ne^2)^{d+1}}{\delta}\right)}$$

with probability $1-\delta$ (w.r.t., the realization of the noise ξ).

Theoretical Analysis

By definition of g

$$\Rightarrow \left|\frac{1}{n}\sum_{i=1}^{n} f_\alpha(x_i, a_i)\xi_i\right| \leq 4V_{\max}||f_\alpha||_{\hat{\rho}}\sqrt{\frac{2}{n}\log\left(\frac{3(9ne^2)^{d+1}}{\delta}\right)}$$

For the specific f_β equivalent to $\widehat{\xi}$

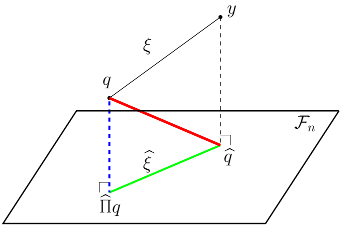
$$\Rightarrow \langle\widehat{\xi}, \xi\rangle \leq 4V_{\max}||\widehat{\xi}||_n\sqrt{\frac{2}{n}\log\left(\frac{3(9ne^2)^{d+1}}{\delta}\right)}$$

Recalling the objective

$$\Rightarrow ||\widehat{\xi}||_n^2 \leq 4V_{\max}||\widehat{\xi}||_n\sqrt{\frac{2}{n}\log\left(\frac{3(9ne^2)^{d+1}}{\delta}\right)}$$

$$\Rightarrow ||\widehat{\Pi}q - \widehat{q}||_n \leq 4V_{\max}\sqrt{\frac{2}{n}\log\left(\frac{3(9ne^2)^{d+1}}{\delta}\right)}$$

Theoretical Analysis



Theorem (see e.g. Lazaric et al.,'11)

At each iteration k and given a set of state–action pairs $\{(x_i, a_i)\}$, *LinearFQI returns an approximation* $\widehat{q}$ *such that*

$$||q - \widehat{q}||_n \leq ||q - \widehat{\Pi}q||_n + ||\widehat{\Pi}q - \widehat{q}||_n$$

$$\leq ||q - \widehat{\Pi}q||_n + O\left(V_{\max}\sqrt{\frac{d \log n/\delta}{n}}\right)$$

Theoretical Analysis

Moving back from vectors to functions

$$||q - \widehat{q}||_n = ||Q^k - f_{\hat{\alpha}_k}||_{\hat{\rho}}$$
$$||q - \widehat{\Pi} q||_n \leq ||Q^k - f_{\alpha_k^*}||_{\hat{\rho}}$$

$$\Rightarrow ||Q^k - f_{\hat{\alpha}_k}||_{\hat{\rho}} \leq ||Q^k - f_{\alpha_k^*}||_{\hat{\rho}} + O\left(V_{\max} \sqrt{\frac{d \log n/\delta}{n}} \right)$$

Theoretical Analysis

By definition of truncation ($\widetilde{Q}^k = \text{Trunc}(f_{\hat{\alpha}_k})$)

Theorem

*At each iteration k and given a set of state–action pairs $\{(x_i, a_i)\}$, LinearFQI returns an approximation $\widehat{Q}^k$ such that (**Objective 3**)*

$$
\begin{aligned}
||Q^k - \widetilde{Q}^k||_{\hat{\rho}} &\leq ||Q^k - f_{\hat{\alpha}_k}||_{\hat{\rho}} \\
&\leq ||Q^k - f_{\alpha_k^*}||_{\hat{\rho}} + O\left(V_{\max}\sqrt{\frac{d \log n/\delta}{n}}\right)
\end{aligned}
$$

Theoretical Analysis

Remark: in order to move from **Obj3** to **Obj2** we need to move from empirical to expected L_2-norms

Since $\widetilde{Q}^k$ is truncated, it is bounded in $[-V_{\max}, V_{\max}]$

$$2||Q^k - \widetilde{Q}^k||_{\hat{\rho}} \geq ||Q^k - \widetilde{Q}^k||_{\rho} - O\Big(V_{\max}\sqrt{\frac{d \log n/\delta}{n}}\Big)$$

The best solution $f_{\alpha_k^*}$ is a fixed function in $\mathcal{F}$

$$||Q^k - f_{\alpha_k^*}||_{\hat{\rho}} \leq 2||Q^k - f_{\alpha_k^*}||_{\rho} + O\Big(\big(V_{\max} + L||\alpha_k^*||\big)\sqrt{\frac{\log 1/\delta}{n}}\Big)$$

Theoretical Analysis

Theorem

*At each iteration k, LinearFQI returns an approximation $\widetilde{Q}^k$ such that (**Objective 2**)*

$$
\begin{aligned}
||Q^k - \widetilde{Q}^k||_\rho \leq 4||Q^k - f_{\alpha_k^*}||_\rho \\
&+ O\left(\left(V_{\max} + L||\alpha_k^*||\right) \sqrt{\frac{\log 1/\delta}{n}} \right) \\
&+ O\left(V_{\max} \sqrt{\frac{d \log n/\delta}{n}} \right),
\end{aligned}
$$

with probability $1 - \delta$.

Theoretical Analysis

$$||Q^k - \widetilde{Q}^k||_\rho \leq 4||Q^k - f_{\alpha_k^*}||_\rho + O\left(\left(V_{\max} + L||\alpha_k^*||\right)\sqrt{\frac{\log 1/\delta}{n}}\right) + O\left(V_{\max}\sqrt{\frac{d \log n/\delta}{n}}\right)$$

Theoretical Analysis

$$
\begin{aligned}
||Q^k - \widetilde{Q}^k||_\rho \leq 4||Q^k - f_{\alpha_k^*}||_\rho
&+ O\Big(\big(V_{\max} + L||\alpha_k^*||\big)\sqrt{\frac{\log 1/\delta}{n}}\Big) \\
&+ O\Big(V_{\max}\sqrt{\frac{d \log n/\delta}{n}}\Big)
\end{aligned}
$$

Remarks

- No algorithm can do better
- Constant 4
- Depends on the space $\mathcal{F}$
- Changes with the iteration k

Theoretical Analysis

$$
\begin{aligned}
||Q^k - \widetilde{Q}^k||_\rho \leq 4||Q^k - f_{\alpha_k^*}||_\rho \\
&+ O\Big(\big(V_{\max} + L||\alpha_k^*||\big) \sqrt{\frac{\log 1/\delta}{n}} \Big) \\
&+ O\Big(V_{\max} \sqrt{\frac{d \log n/\delta}{n}} \Big)
\end{aligned}
$$

Remarks

- Vanishing to zero as $O(n^{-1/2})$
- Depends on the features (L) and on the best solution ($||\alpha_k^*||$)

Theoretical Analysis

$$||Q^k - \widetilde{Q}^k||_\rho \le 4||Q^k - f_{\alpha_k^*}||_\rho + O\Big((V_{\max} + L||\alpha_k^*||)\sqrt{\frac{\log 1/\delta}{n}}\Big) + O\Big(V_{\max}\sqrt{\frac{d \log n/\delta}{n}}\Big)$$

Remarks

- Vanishing to zero as $O(n^{-1/2})$
- Depends on the dimensionality of the space (d) and the number of samples (n)

Outline

Sample Complexity of LSTD

Sample Complexity of Fitted Q-iteration

Error at Each Iteration

Error Propagation

The Final Bound

Theoretical Analysis

Objective 1

$$||Q^* - Q^{\pi_K}||_\mu$$

- **Problem 1**: the test norm μ is different from the sampling norm ρ
- **Problem 2**: we have bounds for $\widetilde{Q}^k$ not for the performance of the corresponding π_k
- **Problem 3**: we have bounds for one single iteration

Propagation of Errors

- Bellman operators

$$\mathcal{T}Q(x,a) = r(x,a) + \gamma \int_X \max_{a'} Q(dx', a') p(dx'|x,a)$$
$$\mathcal{T}^\pi Q(x,a) = r(x,a) + \gamma \int_X Q(dx', \pi(dx')) p(dx'|x,a)$$

- Optimal action–value function

$$Q^* = \mathcal{T}Q^*$$

- Greedy policy

$$\pi(x) = \arg\max_a Q(x,a)$$
$$\pi^*(x) = \arg\max_a Q^*(x,a)$$

- Prediction error

$$\epsilon^k = Q^k - \widetilde{Q}^k$$

Propagation of Errors

Step 1: upper-bound on the propagation (problem 3)

By definition $\mathcal{T} Q^k \geq \mathcal{T}^{\pi^*} Q^k$

$$Q^* - \widetilde{Q}^{k+1} = \underbrace{\mathcal{T}^{\pi^*} Q^*}_{\text{fixed point}} \underbrace{- \mathcal{T}^{\pi^*} \widetilde{Q}^k + \mathcal{T}^{\pi^*} \widetilde{Q}^k}_{0} \underbrace{- \mathcal{T} \widetilde{Q}^k + \epsilon_k}_{\widetilde{Q}^{k+1}}$$

$$Q^* - \widetilde{Q}^{k+1} = \underbrace{\mathcal{T}^{\pi^*} Q^* - \mathcal{T}^{\pi^*} \widetilde{Q}^k}_{\text{recursion}} + \underbrace{\mathcal{T}^{\pi^*} \widetilde{Q}^k - \mathcal{T} \widetilde{Q}^k}_{\leq 0} + \underbrace{\epsilon_k}_{\text{error}}$$

$$\begin{aligned} Q^* - \widetilde{Q}^{k+1} &= \mathcal{T}^{\pi^*} Q^* - \mathcal{T}^{\pi^*} \widetilde{Q}^k + \mathcal{T}^{\pi^*} \widetilde{Q}^k - \mathcal{T} \widetilde{Q}^k + \epsilon_k \\ &\leq \gamma P^{\pi^*} (Q^* - \widetilde{Q}^k) + \epsilon_k \end{aligned}$$

$$Q^* - \widetilde{Q}^K \leq \sum^{K-1} \gamma^{K-k-1} (P^{\pi^*})^{K-k-1} \epsilon_k + \gamma^K (P^{\pi^*})^K (Q^* - \widetilde{Q}^0)$$

Propagation of Errors

Step 2: lower-bound on the propagation (problem 3)

By definition $\mathcal{T}Q^* \geq \mathcal{T}^{\pi_k}Q^*$

$$Q^* - \widetilde{Q}^{k+1} = \underbrace{\mathcal{T}Q^*}_{\text{fixed point}} \underbrace{-\mathcal{T}^{\pi_k}Q^* + \mathcal{T}^{\pi_k}Q^*}_{0} \underbrace{-\mathcal{T}\widetilde{Q}^k + \epsilon_k}_{\widetilde{Q}^{k+1}}$$

$$Q^* - \widetilde{Q}^{k+1} = \underbrace{\mathcal{T}Q^* - \mathcal{T}^{\pi_k}Q^*}_{\geq 0} + \underbrace{\mathcal{T}^{\pi_k}Q^* - \mathcal{T}\widetilde{Q}^k}_{\text{greedy pol.}} + \underbrace{\epsilon_k}_{\text{error}}$$

$$Q^* - \widetilde{Q}^{k+1} \geq \underbrace{\mathcal{T}^{\pi_k}Q^* - \mathcal{T}^{\pi_k}\widetilde{Q}^k}_{\text{recursion}} + \underbrace{\epsilon_k}_{\text{error}}$$

$$Q^* - \widetilde{Q}^{k+1} \geq \gamma P^{\pi_k}(Q^* - \widetilde{Q}^k) + \epsilon_k$$

$K-1$

Propagation of Errors

Step 3: from $\widetilde{Q}^K$ to π_K (problem 2)

By definition $\mathcal{T}^{\pi_K}\widetilde{Q}^K = \mathcal{T}\widetilde{Q}^K \geq \mathcal{T}^{\pi^*}Q^K$

$$Q^* - Q^{\pi_K} = \underbrace{\mathcal{T}^{\pi^*}Q^*}_{\text{fixed point}} \underbrace{-\mathcal{T}^{\pi^*}\widetilde{Q}^K + \mathcal{T}^{\pi^*}\widetilde{Q}^K}_{0} \underbrace{-\mathcal{T}^{\pi_K}\widetilde{Q}^K + \mathcal{T}^{\pi_K}\widetilde{Q}^K}_{0} \underbrace{-\mathcal{T}^{\pi_K}\widetilde{Q}^K}_{\text{fixed point}}$$

$$Q^* - Q^{\pi_K} = \underbrace{\mathcal{T}^{\pi^*}Q^* - \mathcal{T}^{\pi^*}\widetilde{Q}^K}_{\text{error}} + \underbrace{\mathcal{T}^{\pi^*}\widetilde{Q}^K - \mathcal{T}^{\pi_K}\widetilde{Q}^K}_{\leq 0} + \underbrace{\mathcal{T}^{\pi_K}\widetilde{Q}^K - \mathcal{T}^{\pi_K}\widetilde{Q}^K}_{\text{function vs policy}}$$

$$Q^* - Q^{\pi_K} \leq \gamma P^{\pi^*}(Q^* - \widetilde{Q}^K) + \gamma P^{\pi_K}(\widetilde{Q}^K \underbrace{-Q^* + Q^*}_{0} - Q^{\pi_K})$$

$$Q^* - Q^{\pi_K} \leq \gamma P^{\pi^*}(\underbrace{Q^* - \widetilde{Q}^K}_{\text{error}}) + \gamma P^{\pi_K}(\underbrace{\widetilde{Q}^K - Q^*}_{\text{error}} + \underbrace{Q^* - Q^{\pi_K}}_{\text{policy performance}})$$

$$(I - \gamma P^{\pi_K})(Q^* - Q^{\pi_K}) \leq \gamma(P^{\pi^*} - P^{\pi_K})(Q^* - \widetilde{Q}^K)$$

Propagation of Errors

Step 3: plugging the error propagation (problem 2)

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

$$+ \Big[(P^{\pi^*})^{K+1} - (P^{\pi_K} P^{\pi_{K-1}} \dots P^{\pi_0}) \Big] (Q^* - \widetilde{Q}^0) \Big\}$$

Propagation of Errors

Step 4: rewrite in compact form

$$Q^* - Q^{\pi_K} \leq \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\left[\sum_{k=0}^{K-1}\alpha_k A_k|\epsilon_k| + \alpha_K A_K|Q^* - \widetilde{Q}^0|\right]$$

- α_k: weights ($\sum_k \alpha_k = 1$)
- A_k: summarize the P^{π_i} terms

Propagation of Errors

Step 5: take the norm w.r.t. to the test distribution μ

$$||Q^* - Q^{\pi_K}||_\mu^2 = \int \mu(dx, da)(Q^*(x,a) - Q^{\pi_K}(x,a))^2$$

$$\leq \left[\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\right]^2 \int \mu(dx, da)\left[\sum_{k=0}^{K-1} \alpha_k A_k |\epsilon_k| + \alpha_K A_K |Q^* - \widetilde{Q}^0|\right]^2(x,a)$$

$$\leq \left[\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\right]^2 \int \mu(dx, da)\left[\sum_{k=0}^{K-1} \alpha_k A_k \epsilon_k^2 + \alpha_K A_K (Q^* - \widetilde{Q}^0)^2\right](x,a)$$

Propagation of Errors

Focusing on one single term

$$\mu A_k = \frac{1-\gamma}{2}\mu(I-\gamma P^{\pi_K})^{-1}\big[(P^{\pi^*})^{K-k} + P^{\pi_K}P^{\pi_{K-1}}\dots P^{\pi_{k+1}}\big]$$

$$= \frac{1-\gamma}{2}\sum_{m\geq 0}\gamma^m \mu(P^{\pi_K})^m\big[(P^{\pi^*})^{K-k} + P^{\pi_K}P^{\pi_{K-1}}\dots P^{\pi_{k+1}}\big]$$

$$= \frac{1-\gamma}{2}\Big[\sum_{m\geq 0}\gamma^m \mu(P^{\pi_K})^m(P^{\pi^*})^{K-k} + \sum_{m\geq 0}\gamma^m \mu(P^{\pi_K})^m P^{\pi_K}P^{\pi_{K-1}}\dots P^{\pi_{k+1}}\Big]$$

Propagation of Errors

Assumption: concentrability terms

$$c(m) = \sup_{\pi_1 \ldots \pi_m} \left\| \frac{d(\mu P^{\pi_1} \ldots P^{\pi_m})}{d\rho} \right\|_\infty$$

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

Remark: related to top-Lyapunov exponent $\Rightarrow C_{\mu,\rho} < \infty$ is a *weak stability* condition

Propagation of Errors

Step 5: take the norm w.r.t. to the test distribution μ

$$||Q^* - Q^{\pi_K}||_\mu^2$$
$$\leq \left[\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\right]^2 \left[\sum_{k=0}^{K-1} \alpha_k (1-\gamma) \sum_{m\geq 0} \gamma^m c(m+K-k) ||\epsilon_k||_\rho^2 + \alpha_K (2V_{\max})^2\right]$$

Propagation of Errors

Step 5: take the norm w.r.t. to the test distribution μ (problem 1)

$$||Q^* - Q^{\pi_K}||_\mu^2 \leq \left[\frac{2\gamma}{(1-\gamma)^2}\right]^2 C_{\mu,\rho} \max_k ||\epsilon_k||_\rho^2 + O\left(\frac{\gamma^K}{(1-\gamma)^3} {V_{\max}}^2\right)$$

Outline

Sample Complexity of LSTD

Sample Complexity of Fitted Q-iteration
- Error at Each Iteration
- Error Propagation
- The Final Bound

Plugging the Per–Iteration Error

$$||Q^* - Q^{\pi_K}||_\mu^2 \leq \left[\frac{2\gamma}{(1-\gamma)^2}\right]^2 C_{\mu,\rho} \max_k ||\epsilon_k||_\rho^2 + O\left(\frac{\gamma^K}{(1-\gamma)^3} {V_{\max}}^2\right)$$

$$\begin{aligned} ||\epsilon_k||_\rho = ||Q^k - \widetilde{Q}^k||_\rho \leq 4||Q^k - f_{\alpha_k^*}||_\rho \\ + O\left(\left(V_{\max} + L||\alpha_k^*||\right)\sqrt{\frac{\log 1/\delta}{n}}\right) \\ + O\left(V_{\max}\sqrt{\frac{d \log n/\delta}{n}}\right) \end{aligned}$$

Plugging the Per–Iteration Error

The inherent Bellman error

$$\begin{aligned}
||Q^k - f_{\alpha_k^*}||_\rho &= \inf_{f \in \mathcal{F}} ||Q^k - f||_\rho \\
&= \inf_{f \in \mathcal{F}} ||\mathcal{T}\widetilde{Q}^{k-1} - f||_\rho \\
&\leq \inf_{f \in \mathcal{F}} ||\mathcal{T} f_{\alpha_{k-1}} - f||_\rho \\
&\leq \sup_{g \in \mathcal{F}} \inf_{f \in \mathcal{F}} ||\mathcal{T} g - f||_\rho = d(\mathcal{F}, \mathcal{T}\mathcal{F})
\end{aligned}$$

Plugging the Per–Iteration Error

$f_{\alpha_k^*}$ is the orthogonal *projection* of Q^k onto $\mathcal{F}$ w.r.t. ρ

$$\Rightarrow ||f_{\alpha_k^*}||_\rho \leq ||Q^k||_\rho = ||\mathcal{T}\widetilde{Q}^{k-1}||_\rho \leq ||\widetilde{Q}^{k-1}||_\infty \leq V_{\max}$$

Plugging the Per–Iteration Error

Gram matrix

$$G_{i,j} = \mathbb{E}_{(x,a)\sim\rho}[\varphi_i(x,a)\varphi_j(x,a)]$$

Smallest eigenvalue of G is ω

$$||f_\alpha||_\rho^2 = ||\phi^\top \alpha||_\rho^2 = \alpha^\top G\alpha \geq \omega\alpha^\top\alpha = \omega||\alpha||^2$$

$$\max_k ||\alpha_k^*|| \leq \max_k \frac{||f_{\alpha_k^*}||_\rho}{\sqrt{\omega}} \leq \frac{V_{\max}}{\sqrt{\omega}}$$

The Final Bound

Theorem (see e.g., Munos,'03)

LinearFQI with a space $\mathcal{F}$ of d features, with n samples at each iteration returns a policy π_K after K iterations such that

$$||Q^* - Q^{\pi_K}||_\mu \leq \frac{2\gamma}{(1-\gamma)^2}\sqrt{C_{\mu,\rho}}\Bigg(4d(\mathcal{F},\mathcal{T}\mathcal{F}) + O\Big(V_{\max}\big(1+\frac{L}{\sqrt{\omega}}\big)\sqrt{\frac{d\log n/\delta}{n}}\Big)\Bigg) + O\Big(\frac{\gamma^K}{(1-\gamma)^3}{V_{\max}}^2\Big)$$

The Final Bound

Theorem

LinearFQI with a space $\mathcal{F}$ of d features, with n samples at each iteration returns a policy π_K after K iterations such that

$$||Q^* - Q^{\pi_K}||_\mu \leq \frac{2\gamma}{(1-\gamma)^2}\sqrt{C_{\mu,\rho}}\left(4d(\mathcal{F},\mathcal{T}\mathcal{F}) + O\left(V_{\max}\left(1+\frac{L}{\sqrt{\omega}}\right)\sqrt{\frac{d\log n/\delta}{n}}\right)\right) + O\left(\frac{\gamma^K}{(1-\gamma)^3}{V_{\max}}^2\right)$$

The *propagation* (and different norms) makes the problem *more complex*

$\Rightarrow$ how do we choose the *sampling distribution*?

The Final Bound

Theorem

LinearFQI with a space $\mathcal{F}$ of d features, with n samples at each iteration returns a policy π_K after K iterations such that

$$||Q^* - Q^{\pi_K}||_\mu \leq \frac{2\gamma}{(1-\gamma)^2}\sqrt{C_{\mu,\rho}}\left(4d(\mathcal{F},\mathcal{T}\mathcal{F}) + O\left(V_{\max}\left(1+\frac{L}{\sqrt{\omega}}\right)\sqrt{\frac{d\log n/\delta}{n}}\right)\right) + O\left(\frac{\gamma^K}{(1-\gamma)^3}{V_{\max}}^2\right)$$

The *approximation* error is *worse* than in regression ⇒ how do *adapt* to the Bellman operator?

The Final Bound

Theorem

LinearFQI with a space $\mathcal{F}$ of d features, with n samples at each iteration returns a policy π_K after K iterations such that

$$||Q^* - Q^{\pi_K}||_\mu \leq \frac{2\gamma}{(1-\gamma)^2}\sqrt{C_{\mu,\rho}}\left(4d(\mathcal{F},\mathcal{T}\mathcal{F}) + O\left(V_{\max}\left(1+\frac{L}{\sqrt{\omega}}\right)\sqrt{\frac{d\log n/\delta}{n}}\right)\right) + O\left(\frac{\gamma^K}{(1-\gamma)^3}{V_{\max}}^2\right)$$

The dependency on γ is worse than at each iteration

$\Rightarrow$ is it possible to *avoid* it?

The Final Bound

Theorem

LinearFQI with a space $\mathcal{F}$ of d features, with n samples at each iteration returns a policy π_K after K iterations such that

$$||Q^* - Q^{\pi_K}||_\mu \leq \frac{2\gamma}{(1-\gamma)^2}\sqrt{C_{\mu,\rho}}\left(4d(\mathcal{F},\mathcal{T}\mathcal{F}) + O\left(V_{\max}\left(1+\frac{L}{\sqrt{\omega}}\right)\sqrt{\frac{d\log n/\delta}{n}}\right)\right) + O\left(\frac{\gamma^K}{(1-\gamma)^3}{V_{\max}}^2\right)$$

The error decreases exponentially in K

$\Rightarrow K \approx \epsilon/(1-\gamma)$

The Final Bound

Theorem

LinearFQI with a space $\mathcal{F}$ of d features, with n samples at each iteration returns a policy π_K after K iterations such that

$$||Q^* - Q^{\pi_K}||_\mu \leq \frac{2\gamma}{(1-\gamma)^2}\sqrt{C_{\mu,\rho}}\left(4d(\mathcal{F},\mathcal{T}\mathcal{F}) + O\left(V_{\max}\left(1+\frac{L}{\sqrt{\omega}}\right)\sqrt{\frac{d\log n/\delta}{n}}\right)\right) + O\left(\frac{\gamma^K}{(1-\gamma)^3}{V_{\max}}^2\right)$$

The smallest eigenvalue of the Gram matrix

$\Rightarrow$ design the features so as to be *orthogonal* w.r.t. ρ

The Final Bound

Theorem

LinearFQI with a space $\mathcal{F}$ of d features, with n samples at each iteration returns a policy π_K after K iterations such that

$$||Q^* - Q^{\pi_K}||_\mu \leq \frac{2\gamma}{(1-\gamma)^2}\sqrt{C_{\mu,\rho}}\Bigg(4d(\mathcal{F},\mathcal{T}\mathcal{F}) + O\Bigg(V_{\max}\Big(1+\frac{L}{\sqrt{\omega}}\Big)\sqrt{\frac{d\log n/\delta}{n}}\Bigg)\Bigg) + O\Bigg(\frac{\gamma^K}{(1-\gamma)^3}{V_{\max}}^2\Bigg)$$

The asymptotic rate $O(d/n)$ is the same as for regression

Summary

- At each iteration FQI solves a regression problem
  $\Rightarrow$ *least–squares* prediction error bound
- The error is propagated through iterations
  $\Rightarrow$ *propagation* of *any* error

Bibliography I

Reinforcement Learning

Alessandro Lazaric
alessandro.lazaric@inria.fr
sequel.lille.inria.fr