Sample Complexity of ADP Algorithms

A. LAZARIC (SequeL Team @INRIA-Lille)
ENS Cachan - Master 2 MVA
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 $\gamma$
- Study the impact of estimation error
**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

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

⇒ worst-case w.r.t. to all the distributions bounded in \([a, b]\), loose for other distributions.
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 \( \mathcal{F} = \{ f : f(\cdot) = \sum_{j=1}^{d} \alpha_j \varphi_j(\cdot) \} \)
- \( V^\pi \) is the fixed-point of \( \mathcal{T}^\pi \)
- \( V^\pi \) may not belong to \( \mathcal{F} \)
- Best approximation of \( V^\pi \) in \( \mathcal{F} \) is

\[
\Pi V^\pi = \arg \min_{f \in \mathcal{F}} || V^\pi - f || \\
(\Pi \text{ is the projection onto } \mathcal{F})
\]
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. $L_?\text{-norm}$)

- $\Pi_\infty \mathcal{T}^\pi$ is a **contraction** in $L_\infty\text{-norm}$
  - $L_\infty\text{-projection}$ 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 $V_{TD} = \Pi_\rho \mathcal{T}^\pi V_{TD}$

$$\langle \mathcal{T}^\pi V_{TD} - V_{TD}, \varphi_i \rangle_\rho = 0, \quad i = 1, \ldots, d$$
$$\langle r^\pi + \gamma P^\pi V_{TD} - V_{TD}, \varphi_i \rangle_\rho = 0$$

$$\langle r^\pi, \varphi_i \rangle_\rho - \sum_{i=1}^{d} \langle \varphi_j - \gamma P^\pi \varphi_j, \varphi_i \rangle_\rho \cdot \alpha^{(j)}_{TD} = 0 \quad \rightarrow \quad A \alpha_{TD} = b$$
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 $\pi$, then $\Pi_\rho \mathcal{T}^\pi$ has a unique fixed-point.

Proposition (LSTD Performance)

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

Empirical LSTD

- We observe a trajectory \((X_0, R_0, X_1, R_1, \ldots, X_N)\) where 
  \(X_{t+1} \sim P(\cdot \mid X_t, \pi(X_t))\) and 
  \(R_t = r(X_t, \pi(X_t))\)

- We build estimators of the matrix \(A\) and vector \(b\)

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

- \(\hat{A}\hat{\alpha}_{TD} = \hat{b}\) \quad \(\hat{V}_{TD}(\cdot) = \phi(\cdot)^\top \hat{\alpha}_{TD}\)

when \(n \to \infty\) then \(\hat{A} \to A\) and \(\hat{b} \to b\), and thus, \(\hat{\alpha}_{TD} \to \alpha_{TD}\) and 
\(\hat{V}_{TD} \to V_{TD}\)
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 $\pi$ has a stationary distribution $\rho^\pi$ (Markov chain is ergodic - e.g. $\beta$-mixing), then

**Theorem (LSTD Error Bound)**

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

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

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

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

- **Approximation error**: it depends on how well the function space \( \mathcal{F} \) can approximate the value function \( V^\pi \)

- **Estimation error**: it depends on the number of samples \( n \), the dim of the function space \( d \), the smallest eigenvalue of the Gram matrix \( \nu \), the mixing properties of the Markov chain (hidden in \( O \))
Theorem (LSPI Error Bound)

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

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

**Theorem (LSPI Error Bound)**

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

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

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

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

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

- **Approximation error:** $E_0(F) = \sup_{\pi \in G(\tilde{F})} \inf_{f \in F} \|V^\pi - f\|_{\rho^\pi}$
- **Estimation error:** depends on $n, d, \nu_\rho, K$
LSPI Error Bound

Theorem (LSPI Error Bound)

Let \( V_{-1} \in \tilde{\mathcal{F}} \) be an arbitrary initial value function, \( \tilde{V}_0, \ldots, \tilde{V}_{K-1} \) be the sequence of truncated value functions generated by LSPI after \( K \) iterations, and \( \pi_K \) be the greedy policy w.r.t. \( \tilde{V}_{K-1} \). 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:** \( E_0(\mathcal{F}) = \sup_{\pi \in \mathcal{G}(\tilde{\mathcal{F}})} \inf_{f \in \mathcal{F}} \| V^\pi - f \|_{\rho^\pi} \)

- **Estimation error:** depends on \( n, d, \nu_\rho, K \)

- **Initialization error:** error due to the choice of the initial value function or initial policy \( | V^* - V^{\pi_0} | \)
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 \( \rho \) such that for any policy \( \pi \in \mathcal{G}(\tilde{\mathcal{F}}) \), we have \( \rho \leq C \rho^\pi \), where \( C < \infty \) is a constant and \( \rho^\pi \) is the stationary distribution of \( \pi \). Furthermore, we can define the concentrability coefficient \( C_{\mu,\rho} \) as before.
LSPI Error Bound

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

Lower-Bounding Distribution

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

\[ \nu_\rho = \text{the smallest eigenvalue of the Gram matrix } (\int \varphi_i \varphi_j \, d\rho)_{i,j} \]
Sample Complexity of Fitted Q-iteration

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 $\rho$, num of samples $n$

Initial function $\tilde{Q}^0 \in \mathcal{F}$

For $k = 1, \ldots, K$

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

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

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

Return $\pi_K(\cdot) = \arg\max_a \tilde{Q}^K(\cdot, a)$ (greedy policy)
Theoretical Objectives

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

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

Initial function $\tilde{Q}^0 \in \mathcal{F}$

For $k = 1, \ldots, K$

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

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

- Return $\tilde{Q}^k = \text{Trunc}(f_{\alpha_k})$

Return $\pi_K(\cdot) = \arg\max_a \tilde{Q}^K(\cdot, a)$ (greedy policy)
Linear Fitted Q-iteration

- Draw $n$ samples $(x_i, a_i)^{i.i.d} \sim \rho$
- Sample $x_i' \sim p(\cdot | x_i, a_i)$ and $r_i = r(x_i, a_i)$
- Compute $y_i = r_i + \gamma \max_a \tilde{Q}^{k-1}(x_i', a)$
- Build training set $\{(x_i, a_i), y_i\}_{i=1}^n$
- Solve the least squares problem
  
  $$f_{\hat{\alpha}_k} = \arg \min_{f_{\alpha} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f_{\alpha}(x_i, a_i) - y_i)^2$$

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

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

**Objective 2:** derive an *intermediate* bound on the prediction error

[random design]

$$||Q^k - \tilde{Q}^k||_\rho \leq ???$$
Theoretical Objectives

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

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

\[
\frac{1}{n} \sum_{i=1}^{n} \left( Q^k(x_i, a_i) - \tilde{Q}^k(x_i, a_i) \right)^2 = \| Q^k - \tilde{Q}^k \|_{\hat{\rho}}^2 \leq ???
\]
Theoretical Objectives

Obj 3

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

⇒ Obj 2

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

⇒ Obj 1

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

**Returned** solution

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

**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 \( \rho \).

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

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

Empirical orthogonal projection

\[ \hat{\Pi} y = \arg \min_{z \in \mathcal{F}_n} ||y - z||_n \]
Additional Notation

- **Target vector:**
  \[ q_i = Q^k(x_i, a_i) = T \tilde{Q}^{k-1}(x_i, a_i) \]
  \[ = r(x_i, a_i) + \gamma \max_a \int_X \tilde{Q}^{k-1}(dx', a)p(dx'|x_i, a_i) \]

- **Observed target vector:**
  \[ y_i = r_i + \gamma \max_a \tilde{Q}^{k-1}(x'_i, a) \]

- **Noise vector (zero–mean and bounded):**
  \[ \xi_i = q_i - y_i \]
  \[ |\xi_i| \leq V_{\text{max}} \quad \mathbb{E}[\xi_i|x_i] = 0 \]
Additional Notation
Additional Notation

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

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

- Returned vector

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

$$\hat{q} = \hat{\Pi}y = \arg \min_{z \in \mathcal{F}_n} ||y - z||_n$$
Additional Notation
Theoretical Analysis

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

$$\| q - \hat{q} \|_n \leq \| q - \hat{\Pi} q \|_n + \| \hat{\Pi} q - \hat{q} \|_n = \| q - \hat{\Pi} q \|_n + \| \hat{\xi} \|_n$$
Theoretical Analysis

\[ \|q - \hat{\pi}q\|_n \leq \|q - \hat{\Pi}q\|_n + \|\xi\|_n \]

- **Prediction error**: distance between learned function and target function
- **Approximation error**: distance between the best function in \( F \) and the target function \( \Rightarrow \) depends on \( F \)
- **Estimation error**: distance between the best function in \( F \) and the learned function \( \Rightarrow \) depends on the samples
Theoretical Analysis

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

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

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

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

By definition of inner product

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

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

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

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

$$\Rightarrow \text{we can use concentration inequalities}$$
Theoretical Analysis

**Problem:** $f_\beta$ is a *random function*

**Solution:** we need *functional concentration inequalities*
Theoretical Analysis

Define the space of *normalized functions*

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

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

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

Application of Pollard’s inequality for space $G$

For any $g \in G$

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

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

By definition of $g$

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

For the specific $f_\beta$ equivalent to $\hat{\xi}$

$$\Rightarrow \langle \hat{\xi}, \xi \rangle \leq 4V_{\text{max}} \| \hat{\xi} \| n \sqrt{\frac{2}{n} \log \left( \frac{3}{\delta} \right)}$$

Recalling the objective

$$\Rightarrow \| \hat{\xi} \|_n^2 \leq 4V_{\text{max}} \| \hat{\xi} \| n \sqrt{\frac{2}{n} \log \left( \frac{3}{\delta} \right)}$$

$$\Rightarrow \| \hat{\Pi} q - \hat{q} \|_n \leq 4V_{\text{max}} \sqrt{\frac{2}{n} \log \left( \frac{3}{\delta} \right)}$$
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 \( \hat{q} \) such that

\[
\|q - \hat{q}\|_n \leq \|q - \hat{\Pi} q\|_n + \|\hat{\Pi} q - \hat{q}\|_n \\
\leq \|q - \hat{\Pi} q\|_n + O \left( V_{\max} \sqrt{\frac{d \log n}{\delta n}} \right)
\]
Theoretical Analysis

Moving back from vectors to functions

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

\[ \Rightarrow || Q^k - f_{\alpha_k} ||_{\hat{\rho}} \leq || Q^k - f_{\alpha^*_k} ||_{\hat{\rho}} + O\left( V_{\text{max}} \sqrt{\frac{d \log n}{n}} \right) \]
Theoretical Analysis

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

**Theorem**

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

$$||Q^k - \tilde{Q}^k||_{\hat{\rho}} \leq ||Q^k - f_{\hat{\alpha}_k}||_{\hat{\rho}}$$

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

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

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

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

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

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

**Theorem**

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

\[ ||Q^k - \tilde{Q}^k||_\rho \leq 4||Q^k - f_{\alpha^*_k}||_\rho \]

\[ + O\left((V_{\text{max}} + L||\alpha^*_k||)\sqrt{\frac{\log 1/\delta}{n}}\right) \]

\[ + O\left(V_{\text{max}} \sqrt{\frac{d \log n/\delta}{n}}\right), \]

with probability $1 - \delta$. 
Theoretical Analysis

\[ \|Q^k - \tilde{Q}^k\|_\rho \leq 4\|Q^k - f_{\alpha_k^*}\|_\rho \]

\[ + O\left( (V_{\text{max}} + L\|\alpha_k^*\|) \sqrt{\log \frac{1}{\delta}} \frac{\log 1/\delta}{n} \right) \]

\[ + O\left( V_{\text{max}} \sqrt{\frac{d \log n/\delta}{n}} \right) \]
Theoretical Analysis

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

Remarks

▶ No algorithm can do better
▶ Constant 4
▶ Depends on the space \( F \)
▶ Changes with the iteration \( k \)
Theoretical Analysis

\[ \| Q^k - \tilde{Q}^k \|_\rho \leq 4 \| Q^k - f_{\alpha_k^*} \|_\rho \]

\[ + O \left( (V_{\text{max}} + L \| \alpha_k^* \|) \sqrt{\frac{\log 1/\delta}{n}} \right) \]

\[ + O \left( V_{\text{max}} \sqrt{\frac{d \log n/\delta}{n}} \right) \]

Remarks

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

\[ \| Q^k - \tilde{Q}^k \|_\rho \leq 4 \| Q^k - f_{\alpha^*_k} \|_\rho \]

\[ + O \left( (V_{\text{max}} + L \| \alpha^*_k \|) \sqrt{\frac{\log 1/\delta}{n}} \right) \]

\[ + O \left( V_{\text{max}} \sqrt{\frac{d \log n/\delta}{n}} \right) \]

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 \( \mu \) is different from the sampling norm \( \rho \)
- **Problem 2**: we have bounds for \( \tilde{Q}^k \) not for the performance of the corresponding \( \pi_k \)
- **Problem 3**: we have bounds for one single iteration
Propagation of Errors

- Bellman operators
  \[ T Q(x, a) = r(x, a) + \gamma \int_X \max_{a'} Q(dx', a') p(dx'|x, a) \]
  \[ T^\pi Q(x, a) = r(x, a) + \gamma \int_X Q(dx', \pi(dx')) p(dx'|x, a) \]

- Optimal action–value function
  \[ Q^* = T Q^* \]

- Greedy policy
  \[ \pi(x) = \arg \max_a Q(x, a) \]
  \[ \pi^*(x) = \arg \max_a Q^*(x, a) \]

- Prediction error
  \[ \epsilon^k = Q^k - \tilde{Q}^k \]
Propagation of Errors

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

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

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

$$Q^* - \tilde{Q}^{k+1} = \mathcal{T} \pi^* Q^* - \mathcal{T} \pi^* \tilde{Q}^k + \mathcal{T} \pi^* \tilde{Q}^k - \mathcal{T} \tilde{Q}^k + \epsilon_k$$

$$Q^* - \tilde{Q}^{k+1} = \mathcal{T} \pi^* Q^* - \mathcal{T} \pi^* \tilde{Q}^k + \mathcal{T} \pi^* \tilde{Q}^k - \mathcal{T} \tilde{Q}^k + \epsilon_k$$

$$\leq \gamma P \pi^* (Q^* - \tilde{Q}^k) + \epsilon_k$$

$$Q^* - \tilde{Q}^K \leq \sum_{k=0}^{K-1} \gamma^{K-k-1} (P \pi^*)^{K-k-1} \epsilon_k + \gamma^K (P \pi^*)^K (Q^* - \tilde{Q}^0)$$
Propagation of Errors

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

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

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

$$Q^* - \tilde{Q}^{k+1} = \underbrace{\mathcal{T}Q^* - \mathcal{T}^{\pi_k}Q^*}_0 + \underbrace{\mathcal{T}^{\pi_k}Q^* - \mathcal{T}\tilde{Q}^k + \epsilon_k}_{\text{greedy pol.}} + \underbrace{\epsilon_k}_{\text{error}}$$

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

$$Q^* - \tilde{Q}^{k+1} \geq \gamma P^{\pi_k}(Q^* - \tilde{Q}^k) + \epsilon_k$$
Propagation of Errors

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

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

$Q^* - Q^{\pi_K} = T^{\pi^*} Q^* - T^{\pi^*} \tilde{Q}^K + T^{\pi^*} \tilde{Q}^K - T^{\pi_K} \tilde{Q}^K + T^{\pi_K} \tilde{Q}^K - T^{\pi_K} \tilde{Q}^K$

- fixed point
- 0
- fixed point

$Q^* - Q^{\pi_K} = T^{\pi^*} Q^* - T^{\pi^*} \tilde{Q}^K + T^{\pi^*} \tilde{Q}^K - T^{\pi_K} \tilde{Q}^K + T^{\pi_K} \tilde{Q}^K - T^{\pi_K} \tilde{Q}^K$

- error
- $\leq 0$
- function vs policy

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

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

- error
- error
- policy performance

$(I - \gamma P^{\pi_K})(Q^* - Q^{\pi_K}) \leq \gamma (P^{\pi^*} - P^{\pi_K})(Q^* - \tilde{Q}^K)$
Step 3: plugging the error propagation (problem 2)

\[
Q^* - Q^π K \leq (I - \gamma P^π K)^{-1} \left\{ \sum_{k=0}^{K-1} \gamma^{K-k} \left[ (P^π)^{K-k} - P^π K P^π K-1 \ldots P^π k+1 \right] \epsilon_k \right. \\
+ \left. \left[ (P^π)^{K+1} - (P^π K P^π K-1 \ldots P^π 0) \right] (Q^* - \tilde{Q}^0) \right\}
\]
Propagation of Errors

**Step 4: rewrite in compact form**

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

- \( \alpha_k \): weights (\( \sum_k \alpha_k = 1 \))
- \( A_k \): summarize the \( P^{\pi_i} \) terms
Proposition of Errors

**Step 5**: take the norm w.r.t. to the test distribution $\mu$

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

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

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

Focusing on one single term

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

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

\[
= \frac{1 - \gamma}{2} \left[ \sum_{m \geq 0} \gamma^m \mu (P^{\pi_K})^m (P^{\pi^*_K})^{K-k} + \sum_{m \geq 0} \gamma^m \mu (P^{\pi_K})^m P^{\pi_K} P^{\pi_K-1} \ldots P^{\pi_{k+1}} \right]
\]
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 \textit{weak} stability condition
Propagation of Errors

**Step 5:** take the norm w.r.t. to the test distribution $\mu$

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

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

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

Sample Complexity of Fitted Q-iteration
  Error at Each Iteration
  Error Propagation
  The Final Bound
Plugging the Per–Iteration Error

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

\[ \| \epsilon_k \|_\rho = \| Q^k - \tilde{Q}^k \|_\rho \leq 4 \| Q^k - f_{\alpha_k^*} \|_\rho \]

\[ + O\left( (V_{\max} + L \|\alpha_k^*\|) \sqrt{\frac{\log 1/\delta}{n}} \right) \]

\[ + O\left( V_{\max} \sqrt{\frac{d \log n/\delta}{n}} \right) \]
Plugging the Per–Iteration Error

The inherent Bellman error

\[
\|Q^k - f_{\alpha_k^*}\|_\rho = \inf_{f \in \mathcal{F}} \|Q^k - f\|_\rho
\]

\[
= \inf_{f \in \mathcal{F}} \|T\tilde{Q}^{k-1} - f\|_\rho
\]

\[
\leq \inf_{f \in \mathcal{F}} \|Tf_{\alpha_{k-1}} - f\|_\rho
\]

\[
\leq \sup_{g \in \mathcal{F}} \inf_{f \in \mathcal{F}} \|Tg - f\|_\rho = d(\mathcal{F}, T\mathcal{F})
\]
Plugging the Per–Iteration Error

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

\[
\Rightarrow \| f_{\alpha_k}^* \|_\rho \leq \| Q^k \|_\rho = \| T \tilde{Q}^{k-1} \|_\rho \leq \| \tilde{Q}^{k-1} \|_\infty \leq V_{\max}
\]
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 \( \omega \)

\[ ||f_\alpha||^2_\rho = ||\varphi^\top \alpha||^2_\rho = \alpha^\top G \alpha \geq \omega \alpha^\top \alpha = \omega ||\alpha||^2 \]

\[ \max_k ||\alpha_k^*|| \leq \max_k \frac{||f_{\alpha_k^*}||_\rho}{\sqrt{\omega}} \leq \frac{V_{\max}}{\sqrt{\omega}} \]
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 $\pi_K$ 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 Final Bound

**Theorem**

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

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

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 $\pi^K$ after $K$ iterations such that

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

$$+ O\left(\frac{\gamma^K}{(1-\gamma)^3} V_{\text{max}}^2 \right)$$

The approximation error is worse than in regression $\Rightarrow$ 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 $\pi_K$ after $K$ iterations such that

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

$$+ O\left(\frac{\gamma^K}{(1 - \gamma)^3} V_{\text{max}}^2\right)$$

The dependency on $\gamma$ 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 $\pi_K$ after $K$ iterations such that

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

The error decreases exponentially in $K$

$$
\Rightarrow K \approx \frac{\epsilon}{(1-\gamma)}
$$
The Final Bound

Theorem

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

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

$$+ O\left( \frac{\gamma^K}{(1 - \gamma)^3} V_{\max}^2 \right)$$

The smallest eigenvalue of the Gram matrix

$\Rightarrow$ design the features so as to be orthogonal w.r.t. $\rho$
The Final Bound

**Theorem**

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

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

$$
+ O\left( \frac{\gamma^K}{(1 - \gamma)^3} V_{\text{max}}^2 \right)
$$

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

- At each iteration FQI solves a regression problem
  ⇒ least–squares prediction error bound

- The error is propagated through iterations
  ⇒ propagation of any error
Reinforcement Learning

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