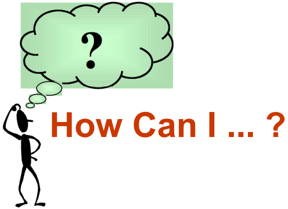
Statistical Learning Theory in Reinforcement Learning & Approximate Dynamic Programming

A. Lazaric & M. Ghavamzadeh (*INRIA Lille – Team SequeL*)

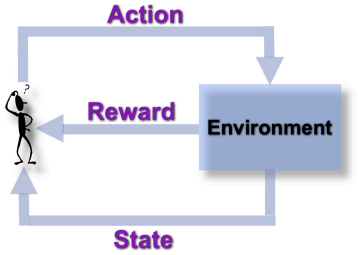
ICML 2012

Sequential Decision-Making under Uncertainty



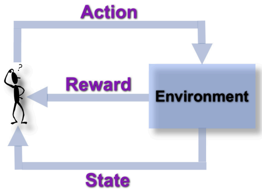
- Move around in the physical world (e.g. driving, navigation)
- Play and win a game
- Retrieve information over the web
- Medical diagnosis and treatment
- Maximize the throughput of a factory
- Optimize the performance of a rescue team

Reinforcement Learning (RL)



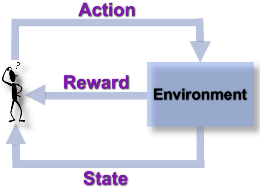
- **RL:** A class of learning problems in which an agent interacts with a dynamic, stochastic, and incompletely known environment
- **Goal:** Learn an action-selection strategy, or policy, to optimize some measure of its long-term performance
- **Interaction:** Modeled as a MDP or a POMDP

Reinforcement Learning (RL)



Goal: Learn an action-selection strategy, or policy, to optimize some measure of its long-term performance

Reinforcement Learning (RL)



Goal: Learn an action-selection strategy, or policy, to optimize some measure of its long-term performance

Algorithms: are based on the two celebrated *dynamic programming* algorithms: **policy iteration** and **value iteration**

A Bit of History

- formulation of the problem: *optimal control, state, value function, Bellman equations, etc.*
- dynamic programming algorithms: *policy iteration and value iteration + proof of convergence to an optimal policy*
- approximate dynamic programming
  - performance evaluation: *how close is the obtained solution to an optimal one?*
    - asymptotic analysis: *the performance with infinite number of samples*

A Bit of History

- formulation of the problem: *optimal control, state, value function, Bellman equations, etc.*
- dynamic programming algorithms: *policy iteration and value iteration + proof of convergence to an optimal policy*
- approximate dynamic programming
  - performance evaluation: *how close is the obtained solution to an optimal one?*
    - asymptotic analysis: *the performance with infinite number of samples*

$$\boxed{\textbf{in real problems we always have a finite number of samples}}$$

Motivation

$\boxed{\textbf{what about the performance with finite number of samples?}}$

- approximate dynamic programming (ADP)
  - asymptotic analysis
  - finite sample analysis

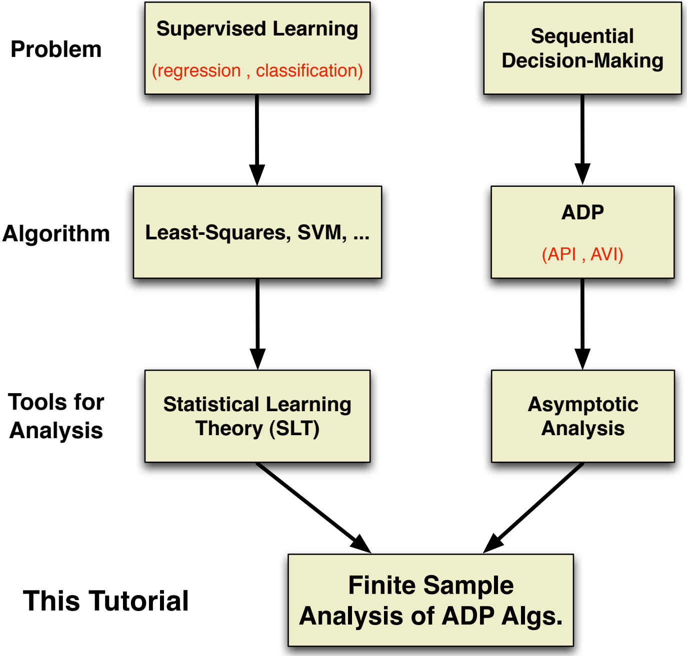
Motivation

what about the performance with finite number of samples?

- approximate dynamic programming (ADP)
  - asymptotic analysis
  - finite sample analysis
- finite sample analysis of ADP algorithms
  - error at each iteration of the alg.
  - how the error propagates through the iterations of the alg.

Motivation

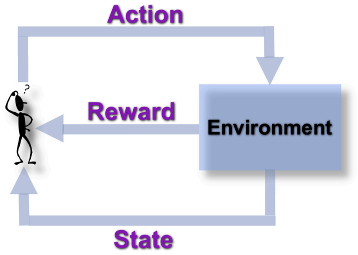
- finite sample analysis of ADP algorithms
  - error at each iteration of the alg.
    - the problem is formulated as *regression*, *classification*, or *fixed point*
    - tools from *statistical learning theory* are used to bound the error of these problems
  - how the error propagates through the iterations of the alg.


Problem
Supervised Learning
(regression , classification)
Sequential Decision-Making
Algorithm
Least-Squares, SVM, ...
ADP
(API , AVI)
Tools for Analysis
Statistical Learning Theory (SLT)
Asymptotic Analysis
This Tutorial
Finite Sample Analysis of ADP Algs.

Outline

Outline

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

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

- 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

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

$\mathcal{B}(\mathcal{X}; V_{\max})$ is the space of functions on $\mathcal{X}$ bounded by $V_{\max}$

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

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

- 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, then

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

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

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

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

Outline

Dynamic Programming Algorithms

Value Iteration

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

Convergence

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

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

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 the optimal policy π^*.

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

Outline

Approximate Dynamic Programming
&
Batch Reinforcement Learning

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)

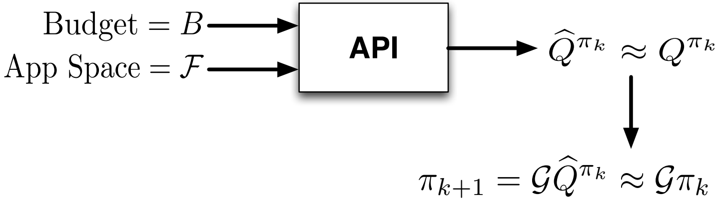
$$\pi_{k+1}(x) = \arg\max_{a \in \mathcal{A}} \widehat{Q}^{\pi_k}(x,a) \neq (\mathcal{G}\pi_k)(x) \longrightarrow V^{\pi_{k+1}} \overset{?}{\geq} V^{\pi_k}$$

Error at each Iteration (AVI)

Error at iteration k

$$||\mathcal{T} Q_k - Q_{k+1}||_{p,\rho} \leq f(B, \mathcal{F}) \qquad \text{w.h.p.}$$

Error at each Iteration (API)



Error at iteration k

$$||Q^{\pi_k} - \widehat{Q}^{\pi_k}||_{p,\rho} \leq f(B, \mathcal{F}) \qquad \text{w.h.p.}$$

Final Performance Bound

Final Objective: Bound the error after K iteration of the alg.

$$||V^* - V^{\pi_K}||_{p,\mu} \leq f(B, \mathcal{F}, K) \qquad \text{w.h.p.}$$

π_K is the policy computed by the algorithm after K iterations

Final Performance Bound

Final Objective: Bound the error after K iteration of the alg.

$$||V^* - V^{\pi_K}||_{p,\mu} \leq f(B, \mathcal{F}, K) \qquad \text{w.h.p.}$$

π_K is the policy computed by the algorithm after K iterations

Error Propagation: How the error at each iteration propagates through the iterations of the algorithm

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)
- the choice of the sampling distribution ρ (exploration problem)

Outline

Objective of the section

- Introduce the theoretical tools used to derive the *error bounds* at each iteration

Objective of the section

- Introduce the theoretical tools used to derive the *error bounds* at each iteration
- Understand the relationship between *accuracy*, *number of samples*, and *confidence*

Outline

The Chernoff–Hoeffding Bound

Remark: all the learning algorithms use ***random*** samples instead of ***actual*** distributions.

The Chernoff–Hoeffding Bound

Remark: all the learning algorithms use ***random*** samples instead of ***actual*** distributions.

Question: how *reliable* is the solution learned from *finite random* samples?

The Chernoff–Hoeffding Bound

Theorem

Let $X_1, \ldots, X_n$ be i.i.d. samples from a distribution $\mathcal{P}$ bounded in $[a, b]$, then for any $\varepsilon > 0$

$$\mathbb{P}\left[\Big|\frac{1}{n}\sum_{t=1}^{n} X_t - \mathbb{E}_{\mathcal{P}}[X_1]\Big| > \varepsilon\right] \leq 2\exp\Big(-\frac{2n\varepsilon^2}{(b-a)^2}\Big)$$

The Chernoff–Hoeffding Bound

Theorem

Let $X_1, \ldots, X_n$ *be i.i.d. samples from a distribution bounded in* $[a, b]$*, then for any* $\varepsilon > 0$

$$\mathbb{P}\Bigg[\underbrace{\Big|\frac{1}{n}\sum_{t=1}^{n} X_t - \mathbb{E}[X_1]\Big|}_{deviation} > \underbrace{\varepsilon}_{accuracy}\Bigg] \leq \underbrace{2\exp\Big(-\frac{2n\varepsilon^2}{(b-a)^2}\Big)}_{confidence}$$

The Chernoff–Hoeffding Bound (Cont.d)

Theorem

Let $X_1, \ldots, X_n$ *be i.i.d. samples from a distribution bounded in* $[a, b]$, *then for any* $\delta \in (0, 1)$

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

The Chernoff–Hoeffding Bound (Cont.d)

Theorem

Let $X_1, X_2, \ldots$ be i.i.d. samples from a distribution bounded in $[a, b]$, then for any $\delta \in (0, 1)$ and $\varepsilon > 0$

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{t=1}^{n} X_t - \mathbb{E}[X_1]\right| > \varepsilon\right] \leq \delta$$

if

$$n \geq \frac{(b-a)^2 \log 2/\delta}{2\varepsilon^2}.$$

The Azuma Bound

Remark: in ADP and RL, the samples are ***not*** necessarily ***i.i.d.*** but may be generated from ***trajectories***

The Azuma Bound

Remark: in ADP and RL, the samples are ***not*** necessarily ***i.i.d.*** but may be generated from ***trajectories***

Question: how is it possible to extend the previous results to *non–i.i.d.* samples?

The Azuma Bound

A sequence of random variables $X_1, X_2, \ldots$ is a *martingale difference sequence* if for any t

$$\mathbb{E}[X_{t+1} | X_1, \ldots, X_t] = 0$$

The Azuma Bound

Theorem

Let $X_1, \ldots, X_n$ be a martingale difference sequence bounded in $[a, b]$, then for any $\varepsilon > 0$

$$\mathbb{P}\Big[\Big|\frac{1}{n}\sum_{t=1}^{n} X_t\Big| > \varepsilon\Big] \leq 2\exp\Big(-\frac{2n\varepsilon^2}{(b-a)^2}\Big)$$

Outline

The Functional Chernoff–Hoeffding Bound

Remark: the learning algorithm returns the *empirical* best *hypothesis* from a *hypothesis set* (e.g., a value function, a policy).

Question: how do the previous results extend to the case of *random* hypotheses in a hypothesis set?

The Functional Chernoff–Hoeffding Bound

Theorem

Let $X_1, \ldots, X_n$ be i.i.d. samples from an arbitrary distribution $\mathcal{P}$ in $\mathcal{X}$ and $f : \mathcal{X} \to [a, b]$ a bounded function, then for any $\varepsilon > 0$

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

The Functional Chernoff–Hoeffding Bound (Cont.d)

Theorem

*Let $X_1, \ldots, X_n$ be i.i.d. samples from an arbitrary distribution $\mathcal{P}$ in $\mathcal{X}$ and $\mathcal{F}$ a set of functions bounded in $[a, b]$, then for any **fixed** $f \in \mathcal{F}$ and any $\varepsilon > 0$*

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

The Functional Chernoff–Hoeffding Bound (Cont.d)

Remark: usually we do not know which function f the learning algorithm will return (it is *random*!)

The Functional Chernoff–Hoeffding Bound (Cont.d)

Remark: usually we do not know which function f the learning algorithm will return (it is *random*!)

Theorem

Let $X_1, \ldots, X_n$ be i.i.d. samples from an arbitrary distribution $\mathcal{P}$ in $\mathcal{X}$ and $\mathcal{F}$ a set of functions bounded in $[a, b]$, then for any $\varepsilon > 0$

$$\mathbb{P}\left[\exists f \in \mathcal{F} : \left|\frac{1}{n}\sum_{t=1}^{n} f(X_t) - \mathbb{E}[f(X_1)]\right| > \varepsilon\right] \leq \quad ???$$

The Union Bound

Also known as: Boole's inequality, Bonferroni inequality, etc.

Theorem

Let $A_1, A_2, \ldots$ be a countable set of events, then

$$\mathbb{P}\Big[\bigcup_i A_i\Big] \le \sum_i \mathbb{P}[A_i].$$

The Union Bound

Also known as: Boole's inequality, Bonferroni inequality, etc.

Theorem

Let $A_1, A_2, \dots$ be a countable set of events, then

$$\mathbb{P}\Big[\bigcup_i A_i\Big] \leq \sum_i \mathbb{P}[A_i].$$

The Functional Chernoff–Hoeffding Bound (Cont.d)

$$\mathbb{P}\Big[\exists f \in \mathcal{F} : \Big|\frac{1}{n}\sum_{t=1}^{n} f(X_t) - \mathbb{E}[f(X_1)]\Big| > \varepsilon\Big]$$

The Functional Chernoff–Hoeffding Bound (Cont.d)

$$\mathbb{P}\Big[\exists f \in \mathcal{F} : \Big|\frac{1}{n}\sum_{t=1}^{n} f(X_t) - \mathbb{E}[f(X_1)]\Big| > \varepsilon\Big]$$

$$= \mathbb{P}\Bigg[\Big\{\Big|\frac{1}{n}\sum_{t=1}^{n} f_1(X_t) - \mathbb{E}[f_1(X_1)]\Big| > \varepsilon\Big\} \bigcup$$

$$\Big\{\Big|\frac{1}{n}\sum_{t=1}^{n} f_2(X_t) - \mathbb{E}[f_2(X_1)]\Big| > \varepsilon\Big\} \bigcup$$

$$\dots$$

$$\Big\{\Big|\frac{1}{n}\sum_{t=1}^{n} f_N(X_t) - \mathbb{E}[f_N(X_1)]\Big| > \varepsilon\Big\} \bigcup$$

$$\dots\Bigg]$$

The Functional Chernoff–Hoeffding Bound (Cont.d)

Theorem

*Let $X_1, \ldots, X_n$ be i.i.d. samples from an arbitrary distribution $\mathcal{P}$ in $\mathcal{X}$ and $\mathcal{F}$ a **finite** set of functions bounded in $[a, b]$ with $|\mathcal{F}| = N$, then for any $f_1 \in \mathcal{F}$ and any $\delta \in (0,1)$*

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

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

The Functional Chernoff–Hoeffding Bound (Cont.d)

Theorem

*Let $X_1, \ldots, X_n$ be i.i.d. samples from an arbitrary distribution $\mathcal{P}$ in $\mathcal{X}$ and $\mathcal{F}$ a **finite** set of functions bounded in $[a, b]$ with $|\mathcal{F}| = N$, then for any $f_1 \in \mathcal{F}$ and any $\delta \in (0,1)$*

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

The Functional Chernoff–Hoeffding Bound (Cont.d)

Problem: In general $\mathcal{F}$ contains an infinite number of functions (e.g., a linear classifier)

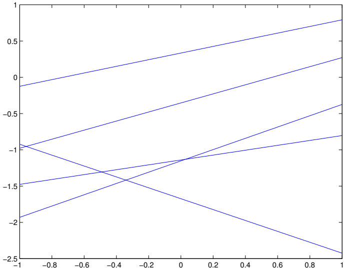
The Symmetrization Trick

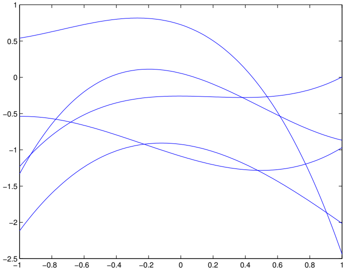
$$\mathbb{P}\left[\exists f \in \mathcal{F} : \left|\frac{1}{n}\sum_{t=1}^{n} f(X_t) - \mathbb{E}\big[f(X_1)\big]\right| > \varepsilon\right]$$
$$\leq 2\mathbb{P}\left[\exists f \in \mathcal{F} : \left|\frac{1}{n}\sum_{t=1}^{n} f(X_t) - \frac{1}{n}\sum_{t=1}^{n} f(X'_t)\right| > \frac{\varepsilon}{2}\right]$$

with the ghost samples $\{X'_t\}_{t=1}^n$ independently drawn from $\mathcal{P}$.

The VC dimension

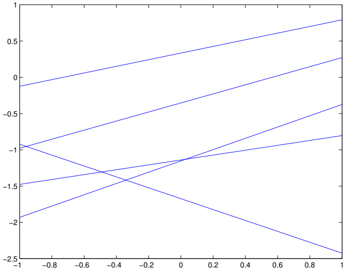
Not all the *infinities* are the same...

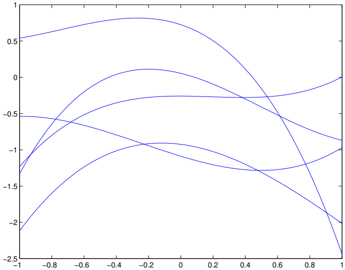




The VC dimension

Not all the *infinities* are the same...



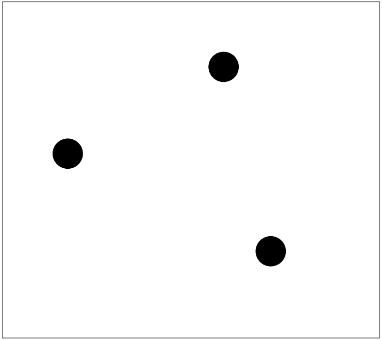


Let's consider a binary space $\mathcal{F} = \{f : \mathcal{X} \rightarrow \{0, 1\}\}$

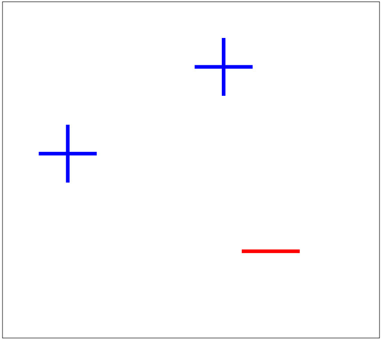
The VC dimension (cont'd)

How many *different predictions* can a space $\mathcal{F}$ produce over n distinct inputs?

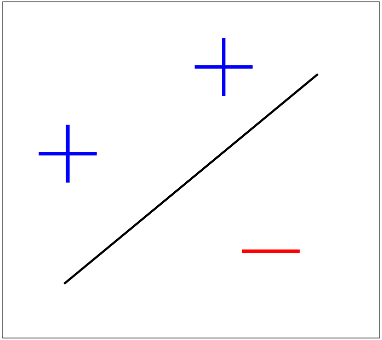
The VC dimension (cont'd)



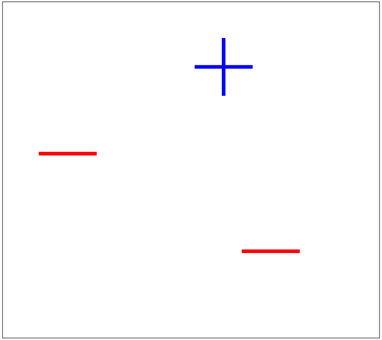
The VC dimension (cont'd)



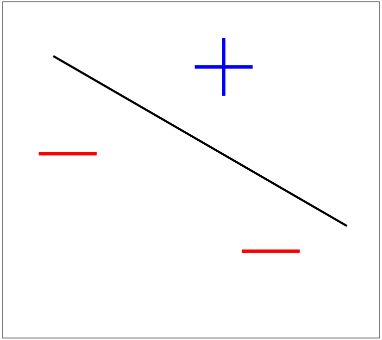
The VC dimension (cont'd)



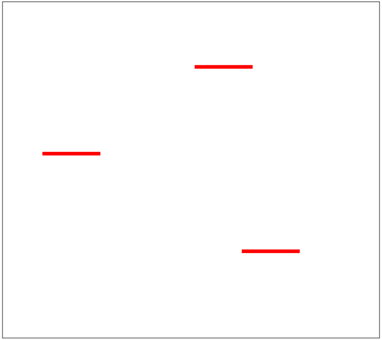
The VC dimension (cont'd)



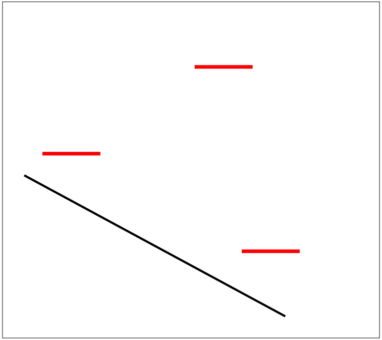
The VC dimension (cont'd)



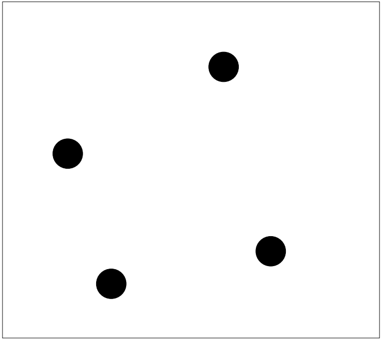
The VC dimension (cont'd)



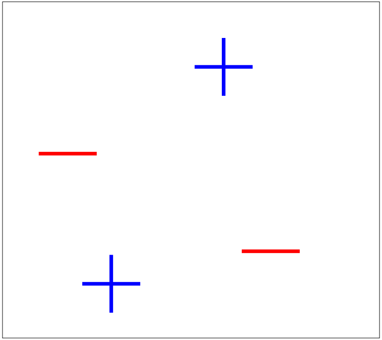
The VC dimension (cont'd)



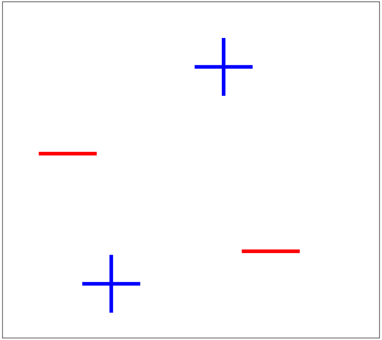
The VC dimension (cont'd)



The VC dimension (cont'd)



The VC dimension (cont'd)



The *VC dimension* of a linear classifier in dim. 2 is $VC(\mathcal{F}) = 3$.

The VC dimension (cont'd)

Let $S = (x_1, \ldots, x_d)$ be an arbitrary sequence of points, then

$$\Pi_S(\mathcal{F}) = \{(f(x_1), \ldots, f(x_d)), h \in \mathcal{F}\}$$

is the set of all the possible ways the d points can be classified by hypothesis in $\mathcal{F}$.

The VC dimension (cont'd)

Let $S = (x_1, \ldots, x_d)$ be an arbitrary sequence of points, then

$$\Pi_S(\mathcal{F}) = \{(f(x_1), \ldots, f(x_d)), h \in \mathcal{F}\}$$

is the set of all the possible ways the d points can be classified by hypothesis in $\mathcal{F}$.

Definition

A set S is shattered by a hypothesis space $\mathcal{F}$ if $|\Pi_S(\mathcal{F})| = 2^d$.

The VC dimension (cont'd)

Definition (VC Dimension)

The VC dimension of a hypothesis space $\mathcal{F}$ is

$$\mathsf{VC}(\mathcal{F}) = \max\{d | \, \exists |S| = d, |\Pi_S(\mathcal{F})| = 2^d\}$$

The VC dimension (cont'd)

Definition (VC Dimension)

The VC dimension of a hypothesis space $\mathcal{F}$ is

$$\mathsf{VC}(\mathcal{F}) = \max\{d \mid \exists |S| = d, |\Pi_S(\mathcal{F})| = 2^d\}$$

Lemma (Sauer's Lemma)

Let $\mathcal{F}$ be a hypothesis space with VC dimension d, then for any sequence of n points $S = (x_1, \ldots, x_n)$ with $n > d$

$$|\Pi_S(\mathcal{F})| \leq \sum_{i=0}^{d} \binom{n}{i} \leq n^d$$

The Functional CH Bound for Binary Spaces

Question: how many values can $f \in \mathcal{F}$ (with $\mathcal{F}$ a *binary* space) take on $2n$ samples?

$$2\mathbb{P}\left[\exists f \in \mathcal{F} : \left|\frac{1}{n}\sum_{t=1}^{n} f(X_t) - \frac{1}{n}\sum_{t=1}^{n} f(X'_t)\right| > \frac{\varepsilon}{2}\right]$$

The Functional CH Bound for Binary Spaces

Question: how many values can $f \in \mathcal{F}$ (with $\mathcal{F}$ a *binary* space) take on $2n$ samples?

$$2\mathbb{P}\left[\exists f \in \mathcal{F} : \left|\frac{1}{n}\sum_{t=1}^{n} f(X_t) - \frac{1}{n}\sum_{t=1}^{n} f(X'_t)\right| > \frac{\varepsilon}{2}\right]$$

If VC$(\mathcal{F}) = d$ and $2n > d$, then the answer is ***at most*** $(2n)^d$!

The Functional CH Bound for Binary Spaces

Theorem

*Let $X_1, \ldots, X_n$ be i.i.d. samples from an arbitrary distribution $\mathcal{P}$ in $\mathcal{X}$ and $\mathcal{F}$ a **finite** set of binary functions with VC $= d$, then for any $\delta \in (0,1)$*

$$\mathbb{P}\left[\exists f \in \mathcal{F} : \left|\frac{1}{n}\sum_{t=1}^{n} f(X_t) - \mathbb{E}[f(X_1)]\right| > \sqrt{\frac{\log 2N/\delta}{2n}}\right] \leq 2\delta$$

with $N = (2n)^d$.

The Functional CH Bound for Binary Spaces

A simplified reading of the previous bound

For any set of n i.i.d. samples and any binary function $f \in \mathcal{F}$ with VC$(\mathcal{F}) = d$

$$\Big|\frac{1}{n}\sum_{t=1}^{n} f(X_t) - \mathbb{E}[f(X_1)]\Big| \leq O\Big(\sqrt{\frac{d \log n/\delta}{n}}\Big)$$

with probability $1 - \delta$ (w.r.t. to the randomness of the samples)

The Pollard's Inequality

Extension: how does the previous result extend to the case of a real–valued space $\mathcal{F}$?

The Pollard's Inequality

Question: how many values can $f \in \mathcal{F}$ (with $\mathcal{F}$ a *real–valued* space) take on $2n$ samples?

$$2\mathbb{P}\left[\exists f \in \mathcal{F} : \Big|\frac{1}{n}\sum_{t=1}^{n} f(X_t) - \frac{1}{n}\sum_{t=1}^{n} f(X'_t)\Big| > \frac{\varepsilon}{2}\right]$$

Answer: an infinite number of values...

Covering

Observation: we only need an *accuracy* of order ε.

Covering

Observation: we only need an *accuracy* of order ε.

Question: *how many functions* from $\mathcal{F}$ do we need to achieve an accuracy of order ε on $2n$ samples?

Covering

A space $\mathcal{F}_\varepsilon \subset \mathcal{F}$ is an *ε-cover* of $\mathcal{F}$ on the states $\{x_t\}_{t=1}^n$ if

$$\forall f \in \mathcal{F}, \exists f' \in \mathcal{F}_\varepsilon : \Big| \frac{1}{n} \sum_{t=1}^n f(x_t) - \frac{1}{n} \sum_{t=1}^n f'(x_t) \Big| \leq \varepsilon$$

Covering

A space $\mathcal{F}_\varepsilon \subset \mathcal{F}$ is an *ε-cover* of $\mathcal{F}$ on the states $\{x_t\}_{t=1}^n$ if

$$\forall f \in \mathcal{F}, \exists f' \in \mathcal{F}_\varepsilon : \Big| \frac{1}{n} \sum_{t=1}^n f(x_t) - \frac{1}{n} \sum_{t=1}^n f'(x_t) \Big| \leq \varepsilon$$

Covering

A space $\mathcal{F}_\varepsilon \subset \mathcal{F}$ is an *ε-cover* of $\mathcal{F}$ on the states $\{x_t\}_{t=1}^n$ if

$$\forall f \in \mathcal{F}, \exists f' \in \mathcal{F}_\varepsilon : \Big| \frac{1}{n} \sum_{t=1}^n f(x_t) - \frac{1}{n} \sum_{t=1}^n f'(x_t) \Big| \leq \varepsilon$$

Covering

A space $\mathcal{F}_\varepsilon \subset \mathcal{F}$ is an *ε-cover* of $\mathcal{F}$ on the states $\{x_t\}_{t=1}^n$ if

$$\forall f \in \mathcal{F}, \exists f' \in \mathcal{F}_\varepsilon : \Big|\frac{1}{n}\sum_{t=1}^n f(x_t) - \frac{1}{n}\sum_{t=1}^n f'(x_t)\Big| \leq \varepsilon$$

The *cover number* of $\mathcal{F}$ is

$$\mathcal{N}\big(\mathcal{F}, \varepsilon, \{x_t\}_{t=1}^n\big) = |\mathcal{F}_\varepsilon|$$

The Pollard's Inequality

We build an $(\varepsilon/8)$-cover of $\mathcal{F}$ on states $\{X_t\}_{t=1}^n \cup \{X'_t\}_{t=1}^n$, thus we have

$$\mathbb{P}\Big[\exists f \in \mathcal{F} : \Big|\frac{1}{n}\sum_{t=1}^n f(X_t) - \frac{1}{n}\sum_{t=1}^n f(X'_t)\Big| > \frac{\varepsilon}{2}\Big]$$

$$\leq \mathbb{P}\Big[\exists f \in \mathcal{F}_{\varepsilon/8} : \Big|\frac{1}{n}\sum_{t=1}^n f(X_t) - \frac{1}{n}\sum_{t=1}^n f(X'_t)\Big| > \frac{\varepsilon}{4}\Big]$$

$$\leq \mathbb{E}\Big[\mathcal{N}\big(\mathcal{F}, \varepsilon/8, \{X_t \cup X'_t\}_{t=1}^n\big)\Big]\mathbb{P}\Big[\Big|\frac{1}{n}\sum_{t=1}^n f(X_t) - \frac{1}{n}\sum_{t=1}^n f(X'_t)\Big| > \frac{\varepsilon}{4}\Big]$$

The Pollard's Inequality

Theorem

Let $X_1, \ldots, X_n$ be i.i.d. samples from an arbitrary distribution $\mathcal{P}$ in $\mathcal{X}$ and $\mathcal{F}$ a set of bounded functions in $[0, B]$, then for any $\varepsilon > 0$

$$\mathbb{P}\left[\exists f \in \mathcal{F} : \Big|\frac{1}{n}\sum_{t=1}^{n} f(X_t) - \mathbb{E}[f(X_1)]\Big| > \varepsilon\right] \leq 8\mathbb{E}\Big[\mathcal{N}\big(\mathcal{F}, \varepsilon/8, \{X_t \cup X'_t\}_{t=1}^{n}\big)\Big] \exp\Big(-\frac{n\varepsilon^2}{64B^2}\Big).$$

The Pseudo–Dimension

Question: how is it possible to *compute the cover number*?
A real–valued space $\mathcal{F}$ has a *psuedo–dimension* d if

$$\mathcal{V}(\mathcal{F}) = \mathsf{VC}\Big(\big\{(x,y) \rightarrow \mathsf{sign}(f(x)-y), f \in \mathcal{F}\big\}\Big) = d$$

The Pseudo–Dimension

Question: how is it possible to *compute the cover number*?
A real–valued space $\mathcal{F}$ has a *psuedo–dimension* d if

$$\mathcal{V}(\mathcal{F}) = \mathsf{VC}\Big(\big\{(x,y) \to \mathsf{sign}(f(x) - y), f \in \mathcal{F}\big\}\Big) = d$$

For any $\{x_t\}_{t=1}^n$

$$\mathcal{N}\big(\mathcal{F}, \varepsilon, \{x_t\}_{t=1}^n\big) \leq O\left(\left(\frac{B}{\varepsilon}\right)^{d}\right)$$

Functional Concentration Inequality for L_2-norm

Remark: In some cases we want to consider the *deviations* between *different norms*.

Functional Concentration Inequality for L_2-norm

Remark: In some cases we want to consider the *deviations* between *different norms*.

Example: in *least–squares regression*, the error is measured with L_2*-norms*, so we want to bound the deviation between

$$\left(\frac{1}{n}\sum_{t=1}^{n} f(X_t)^2\right)^{1/2} \qquad \left(\mathbb{E}\big[f(X)^2\big]\right)^{1/2}$$

Functional Concentration Inequality for L_2-norm

Theorem

Let $X_1, \ldots, X_n$ be i.i.d. samples from an arbitrary distribution $\mathcal{P}$ in $\mathcal{X}$ and $\mathcal{F}$ a set of bounded functions in $[0, B]$, then for any ε

$$\mathbb{P}\Bigg[\exists f \in \mathcal{F} : \Bigg|\Big(\frac{1}{n}\sum_{t=1}^{n} f(X_t)^2\Big)^{1/2} - 2\Big(\mathbb{E}\big[f(X)^2\big]\Big)^{1/2}\Bigg| > \varepsilon\Bigg]$$

$$\leq 3\mathbb{E}\Big[\mathcal{N}_2\big(\mathcal{F}, \frac{\sqrt{2}}{24}\varepsilon, \{X_t \cup X'_t\}_{t=1}^{n}\big)\Big] \exp\Big(-\frac{n\varepsilon^2}{288B^2}\Big).$$

Summary

- Learning algorithms use *finite random* samples
  $\Rightarrow$ *concentration* of averages to expectations

Summary

- Learning algorithms use *finite random* samples
  $\Rightarrow$ *concentration* of averages to expectations
- Learning algorithms use *spaces of functions*
  $\Rightarrow$ *concentration* of averages to expectations for *any* function

Outline

Objective of the section

- *Step-by-step* derivation a *performance bound* for a popular algorithm

Objective of the section

- *Step-by-step* derivation a *performance bound* for a popular algorithm
- Show the *interplay* between *prediction error* and *propagation*

Linear Fitted Q-iteration

Linear space (used to approximate action–value functions)

$$\mathcal{F} = \Big\{ f(x,a) = \sum_{j=1}^{d} \alpha_j \varphi_j(x,a), \quad \alpha \in \mathbb{R}^{d} \Big\}$$

Linear Fitted Q-iteration

Linear space (used to approximate action–value functions)

$$\mathcal{F} = \Big\{ f(x,a) = \sum_{j=1}^{d} \alpha_j \varphi_j(x,a), \quad \alpha \in \mathbb{R}^d \Big\}$$

with features

$$\varphi_j : \mathcal{X} \times \mathcal{A} \rightarrow [0, L] \qquad \phi(x,a) = [\varphi_1(x,a) \ldots \varphi_d(x,a)]^\top$$

Linear Fitted Q-iteration

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

Linear Fitted Q-iteration

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

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

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 $k = 1, \dots, K$

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 $k = 1, \ldots, K$

- Draw n samples $(x_i, a_i) \overset{\text{i.i.d}}{\sim} \rho$

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 $k = 1, \ldots, K$

- 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 $r_i = r(x_i, a_i)$

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 $k = 1, \ldots, K$

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

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 $k = 1, \ldots, K$

- 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 $\big\{\big((x_i, a_i), y_i\big)\big\}_{i=1}^n$

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 $k = 1, \ldots, K$

- 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*

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

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*

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

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

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 $k = 1, \ldots, K$

- 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*

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

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

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

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*

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

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

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

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*

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

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

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

Linear Fitted Q-iteration

- Draw n samples $(x_i, a_i) \stackrel{\text{i.i.d}}{\sim} \rho$
- 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*

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

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

Theoretical Objectives

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

Theoretical Objectives

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

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

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

Theoretical Objectives

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

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*]

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

Theoretical Objectives

Obj 3

$$||Q^k - \widetilde{Q}^k||_{\widehat{\rho}}^2 \leq \ ???$$

Theoretical Objectives

Obj 3

$$||Q^k - \widetilde{Q}^k||_{\widehat{\rho}}^2 \leq \; ???$$

$\Rightarrow$ **Obj 2**

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

Theoretical Objectives

Obj 3

$$||Q^k - \widetilde{Q}^k||_{\widehat{\rho}}^2 \leq \ ???$$

$\Rightarrow$ **Obj 2**

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

$\Rightarrow$ **Obj 1**

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

Theoretical Objectives

Returned solution

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

Theoretical Objectives

Returned solution

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

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.

Additional Notation

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

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

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

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

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

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

Empirical orthogonal projection

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

Additional Notation

- Target vector:

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

Additional Notation

- Target vector:

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

- Observed target vector:

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

Additional Notation

- Target vector:

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

- Observed target vector:

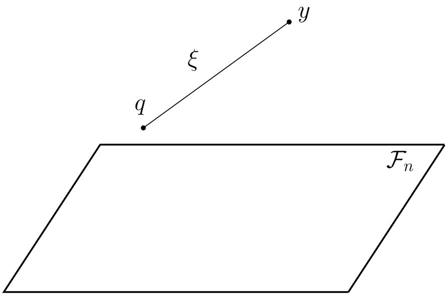
$$y_i = r_i + \gamma \max_a \widetilde{Q}^{k-1}(x_i', a)$$

- 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 $\mathcal{F}_n$

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

Additional Notation

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

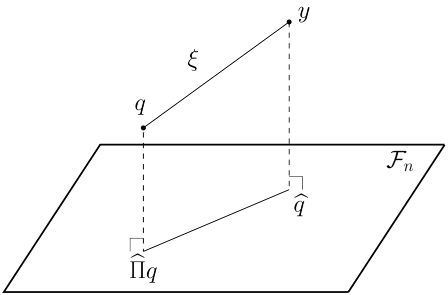
$$\widehat{\Pi} q = \arg\min_{z \in \mathcal{F}_n} ||q - z||_n$$

- Returned vector

$$\widehat{q}_i = f_{\widehat{\alpha}^k}(x_i, a_i)$$

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

Additional Notation

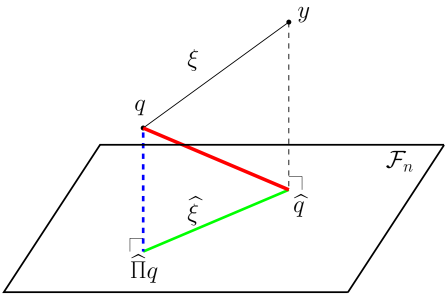


Theoretical Analysis

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

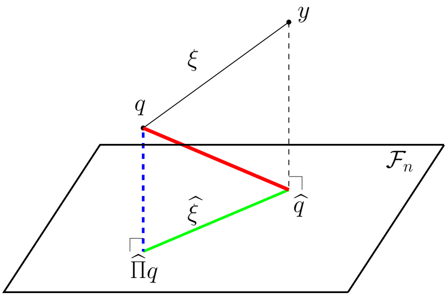
Theoretical Analysis

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



Theoretical Analysis

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



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

Theoretical Analysis

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

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

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 $\mathcal{F}$

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

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

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 $\mathcal{F}_n$

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

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 $[-V_{\max}, V_{\max}]$

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))

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

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))

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

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

$\Rightarrow$ we can use *concentration inequalities*

Theoretical Analysis

Problem: f_β is a *random variable*

Theoretical Analysis

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

Theoretical Analysis

Define the space of *normalized functions*

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

Theoretical Analysis

Define the space of *normalized functions*

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

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

Theoretical Analysis

Define the space of *normalized functions*

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

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

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

Theoretical Analysis

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

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

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

Theoretical Analysis

By definition of g

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

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

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

Theoretical Analysis

By definition of g

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

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

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

Theoretical Analysis

By definition of g

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

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

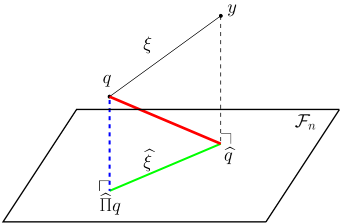
$$\Rightarrow \langle\widehat{\xi},\xi\rangle \leq 4V_{\max}||\widehat{\xi}||_n\sqrt{\frac{2}{n}\log\left(\frac{3(9ne^2)^{d+1}}{\delta}\right)}$$

Recalling the objective

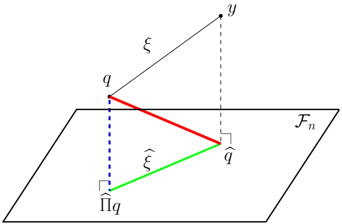
$$\Rightarrow ||\widehat{\xi}||_n^2 \leq 4V_{\max}||\widehat{\xi}||_n\sqrt{\frac{2}{n}\log\left(\frac{3(9ne^2)^{d+1}}{\delta}\right)}$$

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

Theoretical Analysis



Theoretical Analysis



Theorem

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

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

Theoretical Analysis

Moving back from vectors to functions

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

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

Theoretical Analysis

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

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**)*

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

Theoretical Analysis

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

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 $[-V_{\max}, V_{\max}]$

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

Theoretical Analysis

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

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

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

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

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

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

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

Theoretical Analysis

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

Remarks

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

Theoretical Analysis

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

Remarks

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

Theoretical Analysis

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

Remarks

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

Outline

Theoretical Analysis

Objective 1

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

Theoretical Analysis

Objective 1

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

- **Problem 1**: the test norm μ is different from the sampling norm ρ

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

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

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

- 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

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

Propagation of Errors

Step 1: upper-bound on the propagation (problem 3)
By definition $\mathcal{T}Q^k \geq \mathcal{T}^{\pi^*}Q^k$

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

Propagation of Errors

Step 1: upper-bound on the propagation (problem 3)
By definition $\mathcal{T} Q^k \geq \mathcal{T}^{\pi^*} Q^k$

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

Propagation of Errors

Step 1: upper-bound on the propagation (problem 3)
By definition $\mathcal{T}Q^k \geq \mathcal{T}^{\pi^*}Q^k$

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

Propagation of Errors

Step 1: upper-bound on the propagation (problem 3)
By definition $\mathcal{T}Q^k \geq \mathcal{T}^{\pi^*}Q^k$

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

Propagation of Errors

Step 2: lower-bound on the propagation (problem 3)
By definition $\mathcal{T}Q^* \geq \mathcal{T}^{\pi_k}Q^*$

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

Propagation of Errors

Step 2: lower-bound on the propagation (problem 3)
By definition $\mathcal{T}Q^* \geq \mathcal{T}^{\pi_k}Q^*$

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

Propagation of Errors

Step 2: lower-bound on the propagation (problem 3)
By definition $\mathcal{T}Q^* \geq \mathcal{T}^{\pi_k}Q^*$

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

Propagation of Errors

Step 2: lower-bound on the propagation (problem 3)
By definition $\mathcal{T}Q^* \geq \mathcal{T}^{\pi_k}Q^*$

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

Propagation of Errors

Step 2: lower-bound on the propagation (problem 3)
By definition $\mathcal{T}Q^* \geq \mathcal{T}^{\pi_k} Q^*$

$$Q^* - \widetilde{Q}^{k+1} \geq \sum_{k=0}^{K-1} \gamma^{K-k-1} (P^{\pi_{K-1}} P^{\pi_{K-2}} \dots P^{\pi_{k+1}}) \varepsilon_k$$
$$+ \gamma^K (P^{\pi_{K-1}} P^{\pi_{K-2}} \dots P^{\pi_0})(Q^* - \widetilde{Q}^0)$$

Propagation of Errors

Step 3: from $\widetilde{Q}^K$ to π_K (problem 2)
By definition $\mathcal{T}^{\pi_K}\widetilde{Q}^K = \mathcal{T}\widetilde{Q}^K \geq \mathcal{T}^{\pi^*}Q^K$

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

Propagation of Errors

Step 3: from $\widetilde{Q}^K$ to π_K (problem 2)
By definition $\mathcal{T}^{\pi_K}\widetilde{Q}^K = \mathcal{T}\widetilde{Q}^K \geq \mathcal{T}^{\pi^*}Q^K$

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

Propagation of Errors

Step 3: from $\widetilde{Q}^K$ to π_K (problem 2)
By definition $\mathcal{T}^{\pi_K}\widetilde{Q}^K = \mathcal{T}\widetilde{Q}^K \geq \mathcal{T}^{\pi^*}Q^K$

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

Propagation of Errors

Step 3: from $\widetilde{Q}^K$ to π_K (problem 2)
By definition $\mathcal{T}^{\pi_K}\widetilde{Q}^K = \mathcal{T}\widetilde{Q}^K \geq \mathcal{T}^{\pi^*}Q^K$

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

Propagation of Errors

Step 3: from $\widetilde{Q}^K$ to π_K (problem 2)
By definition $\mathcal{T}^{\pi_K}\widetilde{Q}^K = \mathcal{T}\widetilde{Q}^K \geq \mathcal{T}^{\pi^*}Q^K$

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

Propagation of Errors

Step 3: from $\widetilde{Q}^K$ to π_K (problem 2)
By definition $\mathcal{T}^{\pi_K}\widetilde{Q}^K = \mathcal{T}\widetilde{Q}^K \geq \mathcal{T}^{\pi^*}Q^K$

$$Q^* - Q^{\pi_K} \leq \gamma(I - \gamma P^{\pi_K})^{-1}(P^{\pi^*} - P^{\pi_K})(Q^* - \widetilde{Q}^k)$$

Propagation of Errors

Step 3: from $\widetilde{Q}^K$ to π_K (problem 2)
By definition $\mathcal{T}^{\pi_K}\widetilde{Q}^K = \mathcal{T}\widetilde{Q}^K \geq \mathcal{T}^{\pi^*}Q^K$

$$Q^* - Q^{\pi_K} \leq \gamma(I - \gamma P^{\pi_K})^{-1}(P^{\pi^*} - P^{\pi_K})(Q^* - \widetilde{Q}^k)$$

Propagation of Errors

Step 3: plugging the error propagation (problem 2)

$$Q^* - Q^{\pi_K} \leq (I - \gamma P^{\pi_K})^{-1} \Big\{ \sum_{k=0}^{K-1} \gamma^{K-k} \Big[(P^{\pi^*})^{K-k} - P^{\pi_K} P^{\pi_{K-1}} \ldots P^{\pi_{k+1}} \Big] \varepsilon_k$$
$$+ \Big[(P^{\pi^*})^{K+1} - (P^{\pi_K} P^{\pi_{K-1}} \ldots P^{\pi_0}) \Big] (Q^* - \widetilde{Q}^0) \Big\}$$

Propagation of Errors

Step 4: rewrite in compact form

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

- α_k: weights
- A_k: summarize the P^{π_i} terms

Propagation of Errors

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

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

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

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

Propagation of Errors

Focusing on one single term

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

Propagation of Errors

Assumption: concentrability terms

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

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

Propagation of Errors

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

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

Propagation of Errors

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

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

Outline

Plugging Per–Iteration Regret

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

Plugging Per–Iteration Regret

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

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

Plugging Per–Iteration Regret

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 Per–Iteration Regret

$f_{\alpha_k^*}$ 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 Per–Iteration Regret

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

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 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

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

The *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

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

The *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 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

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

The error decreases exponentially in K

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

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

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

The 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

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

Outline

Finite-Sample Performance Bound of Least-Squares Policy Iteration (LSPI)

Least-Squares Policy Iteration (LSPI)

LSPI: is an approximate policy iteration algorithm that uses

Least-Squares Temporal-Difference Learning (LSTD)

for *policy evaluation*.

Objective of the Section

- a brief description of LSTD ***(policy evaluation)*** and LSPI ***(policy iteration)*** algorithms
- report final sample performance bounds for LSTD and LSPI
- describe the main components of these bounds

Outline

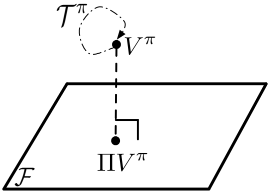
Least-Squares Temporal-Difference Learning (LSTD)

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

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

- V^π is the fixed-point of $\mathcal{T}^\pi$ $\qquad V^\pi = \mathcal{T}^\pi V^\pi$
- V^π may not belong to $\mathcal{F}$ $\qquad V^\pi \notin \mathcal{F}$
- Best approximation of V^π in $\mathcal{F}$ is

$$\Pi V^\pi = \arg\min_{f \in \mathcal{F}} ||V^\pi - f|| \qquad \text{(Π is the projection onto $\mathcal{F}$)}$$



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)

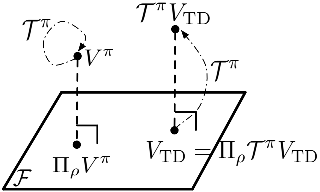
- $\Pi_\infty \mathcal{T}^\pi$ is a contraction in L_∞-norm
  - L_∞-projection is numerically expensive when the number of states is large or infinite

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



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

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

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

- 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.

LSTD Algorithm

Proposition (LSTD Performance)

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

LSTD Algorithm

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

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

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

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

Outline

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 $(\int \varphi_i\, \varphi_j\, d\rho)_{i,j}$

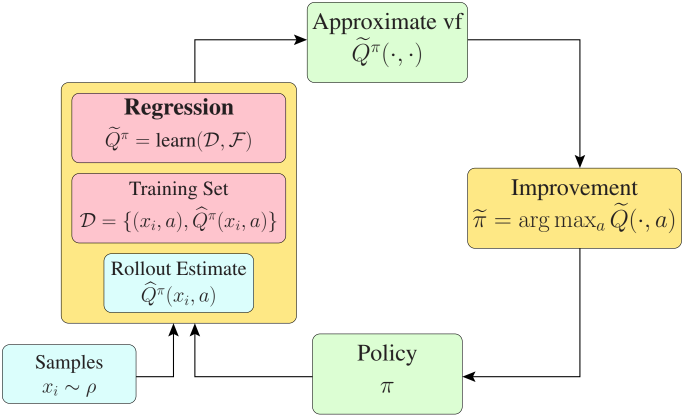
Outline

Finite-Sample Performance Bound of a Classification-based Policy Iteration Algorithm

Objective of the Section

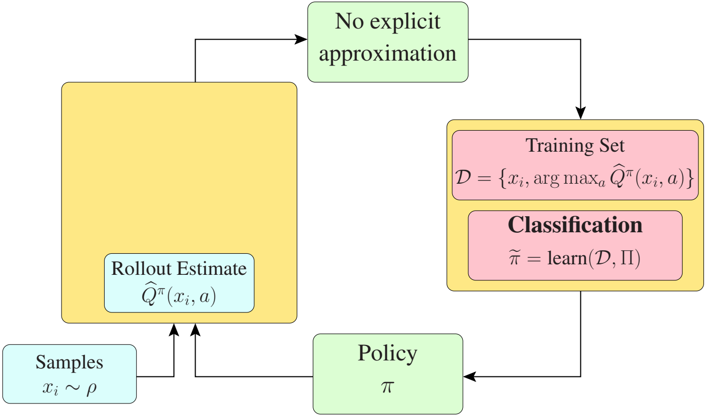
- classification-based vs. regression-based ***(value function-based)*** policy iteration
- describe a classification-based policy iteration algorithm
- report bounds on the error at each iteration and on the error after K iterations of the algorithm
- describe the main components of these bounds

Value-based (Approximate) Policy Iteration



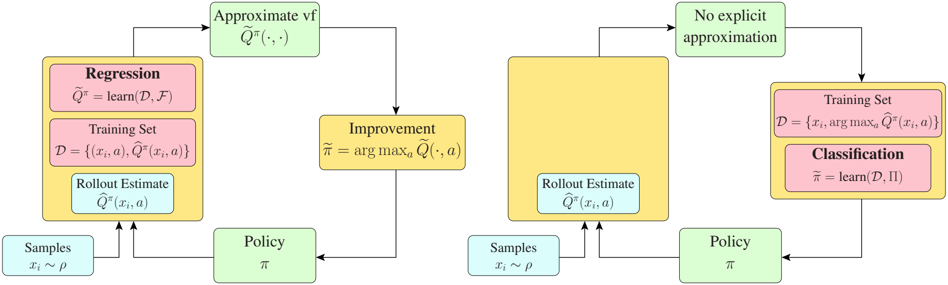
* We use Monte-Carlo estimation for illustration purposes.

Classification-based Policy Iteration



* First introduced by Lagoudakis & Parr (2003) and Fern et al. (2004,2006).

Value-based vs Classification-based Policy Iteration



Appealing Properties

- **Property 1.** More important to have a policy with a performance similar to the greedy policy w.r.t. Q^{π_k} than an accurate approximation of Q^{π_k}.

- **Property 2.** In some problems good policies are easier to represent and learn than their corresponding value functions.

Outline

Template of the Algorithm

Input: policy space $\Pi \subseteq \mathcal{B}^\pi(\mathcal{X})$, state distribution ρ, number of rollout states N, number of rollouts per state-action pair M, rollout horizon H

Initialize: Let $\pi_0 \in \Pi$ be an arbitrary policy

for $k = 0, 1, 2, \ldots$ **do**

  Construct the rollout set $\mathcal{D}_k = \{x_i\}_{i=1}^N,\ x_i \overset{\text{iid}}{\sim} \rho$

  for all states $x_i \in \mathcal{D}_k$ and actions $a \in \mathcal{A}$ **do**

    for $j = 1$ to M **do**

      Perform a rollout according to policy π_k and return

$$R_j^{\pi_k}(x_i, a) = r(x_i, a) + \sum_{t=1}^{H-1} \gamma^t r\big(x^t, \pi_k(x^t)\big),$$

      with $x^t \sim p\big(\cdot|x^{t-1}, \pi_k(x^{t-1})\big)$ and $x^1 \sim p(\cdot|x_i, a)$

    end for

    $\widehat{Q}^{\pi_k}(x_i, a) = \frac{1}{M}\sum_{j=1}^M R_j^{\pi_k}(x_i, a)$

  end for

  $\pi_{k+1} = \arg\min_{\pi \in \Pi} \widehat{\mathcal{L}}_{\pi_k}(\widehat{\rho}\,; \pi)$ **(classifier)**

end for

Template of the Algorithm

Input: policy space $\Pi \subseteq \mathcal{B}^\pi(\mathcal{X})$, state distribution ρ, number of rollout states N, number of rollouts per state-action pair M, rollout horizon H

Initialize: Let $\pi_0 \in \Pi$ be an arbitrary policy

for $k = 0, 1, 2, \ldots$ **do**

Construct the rollout set $\mathcal{D}_k = \{x_i\}_{i=1}^N$, $x_i \overset{\text{iid}}{\sim} \rho$

for all states $x_i \in \mathcal{D}_k$ and actions $a \in \mathcal{A}$ **do**

for $j = 1$ to M **do**

Perform a rollout according to policy π_k and return

$$R_j^{\pi_k}(x_i, a) = r(x_i, a) + \sum_{t=1}^{H-1} \gamma^t r\big(x^t, \pi_k(x^t)\big),$$

with $x^t \sim p\big(\cdot\,|x^{t-1}, \pi_k(x^{t-1})\big)$ and $x^1 \sim p(\cdot|x_i, a)$

end for

$\widehat{Q}^{\pi_k}(x_i, a) = \frac{1}{M} \sum_{j=1}^M R_j^{\pi_k}(x_i, a)$

end for

$\pi_{k+1} = \arg\min_{\pi \in \Pi} \widehat{\mathcal{L}}_{\pi_k}(\widehat{\rho}\,; \pi)$ **(classifier)**

end for

Template of the Algorithm

Input: policy space $\Pi \subseteq \mathcal{B}^\pi(\mathcal{X})$, state distribution ρ, number of rollout states N, number of rollouts per state-action pair M, rollout horizon H

Initialize: Let $\pi_0 \in \Pi$ be an arbitrary policy

for $k = 0, 1, 2, \dots$ **do**

  Construct the rollout set $\mathcal{D}_k = \{x_i\}_{i=1}^N,\ x_i \overset{\text{iid}}{\sim} \rho$

  for all states $x_i \in \mathcal{D}_k$ and actions $a \in \mathcal{A}$ **do**

    for $j = 1$ to M **do**

      Perform a rollout according to policy π_k and return

$$R_j^{\pi_k}(x_i, a) = r(x_i, a) + \sum_{t=1}^{H-1} \gamma^t r\big(x^t, \pi_k(x^t)\big),$$

      with $x^t \sim p\big(\cdot\,|x^{t-1}, \pi_k(x^{t-1})\big)$ and $x^1 \sim p(\cdot|x_i, a)$

    end for

    $\widehat{Q}^{\pi_k}(x_i, a) = \frac{1}{M}\sum_{j=1}^M R_j^{\pi_k}(x_i, a)$

  end for

  $\pi_{k+1} = \arg\min_{\pi \in \Pi} \widehat{\mathcal{L}}_{\pi_k}(\widehat{\rho}\,; \pi)$ **(classifier)**

end for

Template of the Algorithm

Input: policy space $\Pi \subseteq \mathcal{B}^\pi(\mathcal{X})$, state distribution ρ, number of rollout states N, number of rollouts per state-action pair M, rollout horizon H

Initialize: Let $\pi_0 \in \Pi$ be an arbitrary policy

for $k = 0, 1, 2, \ldots$ **do**

  Construct the rollout set $\mathcal{D}_k = \{x_i\}_{i=1}^N$, $x_i \overset{\text{iid}}{\sim} \rho$

  for all states $x_i \in \mathcal{D}_k$ and actions $a \in \mathcal{A}$ **do**

    for $j = 1$ to M **do**

      Perform a rollout according to policy π_k and return

$$R_j^{\pi_k}(x_i, a) = r(x_i, a) + \sum_{t=1}^{H-1} \gamma^t r\big(x^t, \pi_k(x^t)\big),$$

      with $x^t \sim p\big(\cdot\,|x^{t-1}, \pi_k(x^{t-1})\big)$ and $x^1 \sim p(\cdot|x_i, a)$

    end for

    $\widehat{Q}^{\pi_k}(x_i, a) = \frac{1}{M} \sum_{j=1}^M R_j^{\pi_k}(x_i, a)$

  end for

  $\pi_{k+1} = \arg\min_{\pi \in \Pi} \widehat{\mathcal{L}}_{\pi_k}(\widehat{\rho}\,; \pi)$ **(classifier)**

end for

Template of the Algorithm

Input: policy space $\Pi \subseteq \mathcal{B}^{\pi}(\mathcal{X})$, state distribution ρ, number of rollout states N, number of rollouts per state-action pair M, rollout horizon H

Initialize: Let $\pi_0 \in \Pi$ be an arbitrary policy

for $k = 0, 1, 2, \dots$ **do**

  Construct the rollout set $\mathcal{D}_k = \{x_i\}_{i=1}^N,\ x_i \overset{\text{iid}}{\sim} \rho$

  for all states $x_i \in \mathcal{D}_k$ and actions $a \in \mathcal{A}$ **do**

    for $j = 1$ to M **do**

      Perform a rollout according to policy π_k and return

$$R_j^{\pi_k}(x_i, a) = r(x_i, a) + \sum_{t=1}^{H-1} \gamma^t r\big(x^t, \pi_k(x^t)\big),$$

      with $x^t \sim p\big(\cdot\,|x^{t-1}, \pi_k(x^{t-1})\big)$ and $x^1 \sim p(\cdot|x_i, a)$

    end for

    $\widehat{Q}^{\pi_k}(x_i, a) = \frac{1}{M} \sum_{j=1}^M R_j^{\pi_k}(x_i, a)$

  end for

  $\pi_{k+1} = \arg\min_{\pi \in \Pi} \widehat{\mathcal{L}}_{\pi_k}(\widehat{\rho}\,; \pi)$ **(classifier)**

end for

Template of the Algorithm

Empirical Error:

$$\widehat{\mathcal{L}}_{\pi_k}(\widehat{\rho}\,;\pi) = \frac{1}{N}\sum_{i=1}^{N}\Big[\max_{a\in\mathcal{A}}\widehat{Q}^{\pi_k}(x_i,a) - \widehat{Q}^{\pi_k}\big(x_i,\pi(x_i)\big)\Big],$$

($\widehat{\rho}$ is the empirical distribution induced by the samples in $\mathcal{D}_k$)

with the objective to minimize the **Expected Error**

$$\mathcal{L}_{\pi_k}(\rho;\pi) = \int_{\mathcal{X}}\Big[\max_{a\in\mathcal{A}} Q^{\pi_k}(x,a) - Q^{\pi_k}\big(x,\pi(x)\big)\Big]\rho(dx)$$

Mistake-based vs. Gap-based Errors

Mistake-based error

$$\mathcal{L}_{\pi_k}(\rho\,;\pi) = \mathbb{E}_{x\sim\rho}\Big[\mathbb{I}\left\{\pi(x) \neq (\mathcal{G}\pi_k)(x)\right\}\Big]$$
$$= \int_{\mathcal{X}} \mathbb{I}\left\{\pi(x) \neq \arg\max_{a\in\mathcal{A}} Q^{\pi_k}(x,a)\right\}\rho(dx)$$

Gap-based error

$$\mathcal{L}_{\pi_k}(\rho\,;\pi) = \int_{\mathcal{X}} \Big[\max_{a\in\mathcal{A}} Q^{\pi_k}(x,a) - Q^{\pi_k}\big(x,\pi(x)\big)\Big]\rho(dx)$$
$$= \int_{\mathcal{X}} \mathbb{I}\left\{\pi(x) \neq \arg\max_{a\in\mathcal{A}} Q^{\pi_k}(x,a)\right\}\Big[\max_{a\in\mathcal{A}} Q^{\pi_k}(x,a) - Q^{\pi_k}\big(x,\pi(x)\big)\Big]\rho(dx)$$

Mistake-based vs. Gap-based Errors

Mistake-based error

$$\mathcal{L}_{\pi_k}(\rho\,;\pi) = \mathbb{E}_{x\sim\rho}\Big[\mathbb{I}\left\{\pi(x) \neq (\mathcal{G}\pi_k)(x)\right\}\Big]$$
$$= \int_{\mathcal{X}} \underbrace{\mathbb{I}\left\{\pi(x) \neq \arg\max_{a\in\mathcal{A}} Q^{\pi_k}(x,a)\right\}}_{\text{mistake}} \rho(dx)$$

Gap-based error

$$\mathcal{L}_{\pi_k}(\rho\,;\pi) = \int_{\mathcal{A}} \Big[\max_{a\in\mathcal{A}} Q^{\pi_k}(x,a) - Q^{\pi_k}\big(x,\pi(x)\big)\Big]\rho(dx)$$
$$= \int_{\mathcal{X}} \underbrace{\mathbb{I}\left\{\pi(x) \neq \arg\max_{a\in\mathcal{A}} Q^{\pi_k}(x,a)\right\}}_{\text{mistake}} \Big[\max_{a\in\mathcal{A}} Q^{\pi_k}(x,a) - Q^{\pi_k}\big(x,\pi(x)\big)\Big]\rho(dx)$$

Mistake-based vs. Gap-based Errors

Mistake-based error

$$\mathcal{L}_{\pi_k}(\rho\,;\pi) = \mathbb{E}_{x\sim\rho}\Big[\mathbb{I}\left\{\pi(x) \neq (\mathcal{G}\pi_k)(x)\right\}\Big]$$
$$= \int_{\mathcal{X}} \underbrace{\mathbb{I}\Big\{\pi(x) \neq \arg\max_{a\in\mathcal{A}} Q^{\pi_k}(x,a)\Big\}}_{\text{mistake}} \rho(dx)$$

Gap-based error

$$\mathcal{L}_{\pi_k}(\rho\,;\pi) = \int_{\mathcal{X}} \Big[\max_{a\in\mathcal{A}} Q^{\pi_k}(x,a) - Q^{\pi_k}\big(x,\pi(x)\big)\Big]\rho(dx)$$
$$= \int_{\mathcal{X}} \underbrace{\mathbb{I}\Big\{\pi(x) \neq \arg\max_{a\in\mathcal{A}} Q^{\pi_k}(x,a)\Big\}}_{\text{mistake}} \underbrace{\Big[\max_{a\in\mathcal{A}} Q^{\pi_k}(x,a) - Q^{\pi_k}\big(x,\pi(x)\big)\Big]}_{\text{cost/regret}} \rho(dx)$$

Outline

Error at each Iteration

Theorem

Let Π *be a policy space with* $h = VC(\Pi) < \infty$ *and* ρ *be a distribution over* $\mathcal{X}$. *Let* N *be the number of states in* $\mathcal{D}_k$ *drawn i.i.d. from* ρ, H *be the rollout horizon, and* M *be the number of rollouts per state-action pair. Let*

$$\pi_{k+1} = \arg\min_{\pi \in \Pi} \widehat{\mathcal{L}}_{\pi_k}(\widehat{\rho}\,;\pi)$$

be the policy computed at the k*'th iteration of DPI . Then, for any* $\delta > 0$

$$\mathcal{L}_{\pi_k}(\rho\,;\pi_{k+1}) \leq \inf_{\pi \in \Pi} \mathcal{L}_{\pi_k}(\rho\,;\pi) + 2(\epsilon_1 + \epsilon_2 + \gamma^H Q_{\max}),$$

with probability $1 - \delta$, *where*

$$\epsilon_1 = 16 Q_{\max} \sqrt{\frac{2}{N}\left(h \log \frac{eN}{h} + \log \frac{32}{\delta}\right)} \qquad \textit{and}$$

$$\epsilon_2 = 8(1-\gamma^H) Q_{\max} \sqrt{\frac{2}{MN}\left(h \log \frac{eMN}{h} + \log \frac{32}{\delta}\right)}$$

Remarks

The bound

$$\mathcal{L}_{\pi_k}(\rho\,;\pi_{k+1}) \le \underbrace{\inf_{\pi\in\Pi} \mathcal{L}_{\pi_k}(\rho\,;\pi)}_{\text{approximation}} + 2\underbrace{\left(\epsilon_1(N) + \epsilon_2(N,M,H) + \gamma^H Q_{\max}\right)}_{\text{estimation}}$$

- **Approximation error:** it depends on how well the policy space Π can approximate greedy policies.
- **Estimation error:** it depends on the number of rollout states, number of rollouts, and the rollout horizon.

Remarks

The bound

$$\mathcal{L}_{\pi_k}(\rho\,;\pi_{k+1}) \leq \inf_{\pi\in\Pi} \mathcal{L}_{\pi_k}(\rho\,;\pi) + 2\Big(\epsilon_1(N) + \epsilon_2(N,M,H) + \gamma^H Q_{\max}\Big)$$

The approximation error

$$\inf_{\pi\in\Pi} \mathcal{L}_{\pi_k}(\rho\,;\pi) = \inf_{\pi\in\Pi} \int_{\mathcal{X}} \mathbb{I}\left\{\pi(x) \neq (\mathcal{G}\pi_k)(x)\right\} \Big[\max_{a\in\mathcal{A}} Q^{\pi_k}(x,a) - Q^{\pi_k}\big(x,\pi(x)\big)\Big]\rho(dx)$$

Remarks

The bound

$$\mathcal{L}_{\pi_k}(\rho\,;\pi_{k+1}) \leq \inf_{\pi\in\Pi} \mathcal{L}_{\pi_k}(\rho\,;\pi) + 2\big(\epsilon_1(N) + \epsilon_2(N,M,H) + \gamma^H Q_{\max}\big)$$

The estimation error

$$\epsilon_1 = 16 Q_{\max}\sqrt{\frac{2}{N}\left(h\log\frac{eN}{h} + \log\frac{32}{\delta}\right)}$$

$$\epsilon_2 = 8(1-\gamma^H) Q_{\max}\sqrt{\frac{2}{MN}\left(h\log\frac{eMN}{h} + \log\frac{32}{\delta}\right)}$$

Remarks

The bound

$$\mathcal{L}_{\pi_k}(\rho\,;\pi_{k+1}) \le \inf_{\pi\in\Pi} \mathcal{L}_{\pi_k}(\rho\,;\pi) + 2\big(\epsilon_1(N) + \epsilon_2(N,M,H) + \gamma^H Q_{\max}\big)$$

The estimation error

$$\epsilon_1 = 16 Q_{\max}\sqrt{\frac{2}{N}\left(h\log\frac{eN}{h} + \log\frac{32}{\delta}\right)}$$

$$\epsilon_2 = 8(1-\gamma^H)Q_{\max}\sqrt{\frac{2}{MN}\left(h\log\frac{eMN}{h} + \log\frac{32}{\delta}\right)}$$

- Avoid overfitting (ϵ_1) : take $N \gg h$
- Fixed budget of rollouts $B = MN$: take $M = 1$ and $N = B$
- Fixed budget $B = NMH$ and $M = 1$: take $H = O(\frac{\log B}{\log 1/\gamma})$ and $N = O(B/H)$

Final Bound

Theorem

Let Π be a policy space with VC-dimension h and π_K be the policy generated by DPI after K iterations. Then, for any $\delta > 0$

$$||V^* - V^{\pi_K}||_{1,\mu} \leq \frac{C_{\mu,\rho}}{(1-\gamma)^2}\Big[d(\Pi, \mathcal{G}\Pi) + 2(\epsilon_1 + \epsilon_2 + \gamma^H Q_{\max})\Big] + \frac{2\gamma^K R_{\max}}{1-\gamma}$$

with probability $1 - \delta$, where

$$\epsilon_1 = 16 Q_{\max}\sqrt{\frac{2}{N}\left(h \log\frac{eN}{h} + \log\frac{32K}{\delta}\right)} \quad \textit{and}$$

$$\epsilon_2 = 8(1-\gamma^H) Q_{\max}\sqrt{\frac{2}{MN}\left(h \log\frac{eMN}{h} + \log\frac{32K}{\delta}\right)}$$

Final Bound

Theorem

Let Π be a policy space with VC-dimension h and π_K be the policy generated by DPI after K iterations. Then, for any $\delta > 0$

$$||V^* - V^{\pi_K}||_{1,\mu} \leq \frac{C_{\mu,\rho}}{(1-\gamma)^2}\Big[d(\Pi, \mathcal{G}\Pi) + 2(\epsilon_1 + \epsilon_2 + \gamma^H Q_{\max})\Big] + \frac{2\gamma^K R_{\max}}{1-\gamma}$$

with probability $1-\delta$, where

$$\epsilon_1 = 16 Q_{\max}\sqrt{\frac{2}{N}\left(h \log\frac{eN}{h} + \log\frac{32K}{\delta}\right)} \quad \textit{and}$$

$$\epsilon_2 = 8(1-\gamma^H)Q_{\max}\sqrt{\frac{2}{MN}\left(h \log\frac{eMN}{h} + \log\frac{32K}{\delta}\right)}$$

Concentrability coefficient: $C_{\mu,\rho}$

Final Bound

Theorem

Let Π *be a policy space with VC-dimension* h *and* π_K *be the policy generated by DPI after* K *iterations. Then, for any* $\delta > 0$

$$||V^* - V^{\pi_K}||_{1,\mu} \leq \frac{C_{\mu,\rho}}{(1-\gamma)^2}\Big[d(\Pi, \mathcal{G}\Pi) + 2(\epsilon_1 + \epsilon_2 + \gamma^H Q_{\max})\Big] + \frac{2\gamma^K R_{\max}}{1-\gamma}$$

with probability $1-\delta$*, where*

$$\epsilon_1 = 16 Q_{\max}\sqrt{\frac{2}{N}\left(h\log\frac{eN}{h} + \log\frac{32K}{\delta}\right)} \qquad \textit{and}$$

$$\epsilon_2 = 8(1-\gamma^H)Q_{\max}\sqrt{\frac{2}{MN}\left(h\log\frac{eMN}{h} + \log\frac{32K}{\delta}\right)}$$

Estimation error: depends on M, N, H, h, and K

Final Bound

Theorem

Let Π *be a policy space with VC-dimension* h *and* π_K *be the policy generated by DPI after* K *iterations. Then, for any* $\delta > 0$

$$||V^* - V^{\pi_K}||_{1,\mu} \leq \frac{C_{\mu,\rho}}{(1-\gamma)^2}\Big[d(\Pi, \mathcal{G}\Pi) + 2(\epsilon_1 + \epsilon_2 + \gamma^H Q_{\max})\Big] + \frac{2\gamma^K R_{\max}}{1-\gamma}$$

with probability $1 - \delta$*, where*

$$\epsilon_1 = 16 Q_{\max}\sqrt{\frac{2}{N}\left(h \log\frac{eN}{h} + \log\frac{32K}{\delta}\right)} \qquad \textit{and}$$

$$\epsilon_2 = 8(1-\gamma^H) Q_{\max}\sqrt{\frac{2}{MN}\left(h \log\frac{eMN}{h} + \log\frac{32K}{\delta}\right)}$$

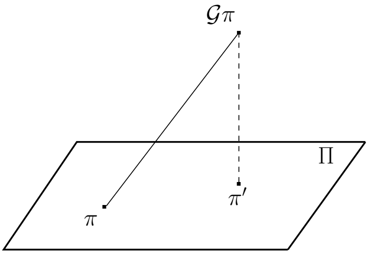
Initialization error: error due to the choice of the initial policy $|V^* - V^{\pi_0}|$

Remarks

Inherent Greedy Error $d(\Pi, \mathcal{G}\Pi)$ (approximation error)

$$d(\Pi, \mathcal{G}\Pi) = \sup_{\pi \in \Pi} \inf_{\pi' \in \Pi} \mathcal{L}_{\pi}(\rho\,; \pi')$$

$$= \sup_{\pi \in \Pi} \inf_{\pi' \in \Pi} \int_{\mathcal{X}} \mathbb{I}\left\{\pi'(x) \neq (\mathcal{G}\pi)(x)\right\} \Big[\max_{a \in \mathcal{A}} Q^{\pi}(x,a) - Q^{\pi}\big(x, \pi'(x)\big)\Big] \rho(dx)$$



Other Finite-Sample Analysis Results in Batch RL

- Approximate Value Iteration ***(Munos & Szepesvari 2008)***
- Approximate Policy Iteration
  - LSTD and LSPI ***(Lazaric et al. 2010, 2012)***
  - Bellman Residual Minimization ***(Maillard et al. 2010)***
  - Modified Bellman Residual Minimization ***(Antos et al. 2008)***
  - Classification-based Policy Iteration ***(Fern et al. 2006; Lazaric et al. 2010; Gabillon et al. 2011; Farahmand et al. 2012)***
  - Conservative Policy Iteration ***(Kakade & Langford 2002; Kakade 2003)***

Other Finite-Sample Analysis Results in Batch RL

- Approximate Modified Policy Iteration ***(Scherrer et al. 2012)***
- Regularized Approximate Dynamic Programming
  - L_2-Regularization
    - L_2-Regularized Policy Iteration ***(Farahmand et al. 2008)***
    - L_2-Regularized Fitted Q-Iteration ***(Farahmand et al. 2009)***
  - L_1-Regularization and High-Dimensional RL
    - Lasso-TD ***(Ghavamzadeh et al. 2011)***
    - LSTD (LSPI) with Random Projections ***(Ghavamzadeh et al. 2010)***

Outline

Discussion

Outline

Learned Lessons

Comparison to Supervised Learning

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

What makes RL more challenging then?

- dependency on $1/(1-\gamma)$ (sequential nature of the problem)
- the approximation error is more complex
- the propagation of error (control problem)
- the sampling problem (how to choose ρ – exploration problem)

Practical Lessons

- Tuning the parameters (given a fixed accuracy ϵ)
  - number of samples (inverting the bound) $n \geq \widetilde{\Omega}(\frac{d}{\epsilon})$
  - number of iterations (inverting the bound) $K \approx \epsilon/(1-\gamma)$
- choice of function $\mathcal{F}$ and/or policy space Π
  - features $\{\varphi_i\}_{i=1}^d$ to be linearly independent given the sampling distribution ρ (on-policy – off-policy sampling)
- tradeoff between approximation and estimation errors

Outline

Open Problems

Open Problems

- *High–dimensional spaces*: how to deal with MDPs with many state–action variables?
  - First example in *deterministic design for LSTD*
  - *Extension* to other algorithms

Open Problems

- *High–dimensional spaces*: how to deal with MDPs with many state–action variables?
  - First example in *deterministic design for LSTD*
  - *Extension* to other algorithms
- *Optimality*: how optimal are the current algorithms?
  - Improve the *sampling* distribution
  - Control the *concentrability* terms
  - Limit the *propagation* of error through iterations

Open Problems

- *High–dimensional spaces*: how to deal with MDPs with many state–action variables?
  - First example in *deterministic design for LSTD*
  - *Extension* to other algorithms
- *Optimality*: how optimal are the current algorithms?
  - Improve the *sampling* distribution
  - Control the *concentrability* terms
  - Limit the *propagation* of error through iterations
- *Off–policy learning for LSTD*

Statistical Learning Theory Meets Dynamic Programming

M. Ghavamzadeh, A. Lazaric

{mohammad.ghavamzadeh, alessandro.lazaric}@inria.fr

sequel.lille.inria.fr