

Offline RL: Fitted Q Iteration

Wen Sun

CS 6789: Foundations of Reinforcement Learning

Recap: Value Iteration (Planning)

$$f_{t+1} = \mathcal{T} f_t$$

Recap: Value Iteration (Planning)

$$f_{t+1} = \mathcal{T}f_t$$

1. We have point-wise accuracy (via the contraction property):

$$\|f_t - Q^*\|_\infty \leq \gamma^k / (1 - \gamma)$$

Recap: Value Iteration (Planning)

$$f_{t+1} = \mathcal{T}f_t$$

1. We have point-wise accuracy (via the contraction property):

$$\|f_t - Q^*\|_\infty \leq \gamma^k / (1 - \gamma)$$

2. Turn f_t 's **point-wise** approximation error to policy's performance (error amplification):

$$\pi^t(s) = \arg \max_a f_t(s, a), \forall s$$

$$V^* - V^{\pi^t} \leq \frac{2}{1 - \gamma} \frac{\gamma^k}{1 - \gamma}$$

Recap: Linear Bellman Completion

Given feature ϕ , take any linear function $w^\top \phi(s, a)$:

$$\forall h, \exists \theta \in \mathbb{R}^d, \text{ s.t. }, \theta^\top \phi(s, a) = r(s, a) + \mathbb{E}_{s' \sim P_h(s, a)} \max_{a'} w^\top \phi(s', a'), \forall s, a$$

(It implies that Q_h^\star is linear in ϕ : $Q_h^\star = (\theta_h^\star)^\top \phi, \forall h$)

Theorem: There exists a way to construct datasets $\{\mathcal{D}_h\}_{h=0}^{H-1}$, such that with probability at least $1 - \delta$, we have:

$$V^{\hat{\pi}} - V^\star \leq \epsilon$$

w/ total number of samples in these datasets scaling $\tilde{O}(d^2 + H^6 d^2 / \epsilon^2)$

Recap: Least-Square Value Iteration

Using D-optimal design, we construct a linear regression dataset such that at all h:

$$\max_{s,a} \left| \theta_h^\top \phi(s,a) - \mathcal{T}_h(\theta_{h+1})^\top \phi(s,a) \right| \leq O\left(Hd/\sqrt{N}\right)$$

Which implies that $Q_t := \theta_t^\top \phi$ is **point-wise accurate**:

$$\|Q_t - Q^*\|_\infty \leq H^2d/\sqrt{N}$$

Today's Question:

what happens when we do nonlinear function regression?

Point-wise prediction error guarantee is not possible anymore

Today's Question:

what happens when we do nonlinear function regression?

Point-wise prediction error guarantee is not possible anymore

Instead of aiming for point-wise guarantee,
We will focus on the average case (i.e., average over some distributions)

Outline

1. Setting: Assumptions

2. Algorithm: Fitted Q Iteration

2. Guarantee and Proof sketch

Setting

1. Infinite horizon Discounted MDPs $\gamma \in (0,1)$
2. A given offline distribution $\nu \in \Delta(S \times A)$ from which we sample offline data
3. Function class $\mathcal{F} = \{f : S \times A \mapsto [0, 1/(1 - \gamma)]\}$

Key Assumptions

1. offline distribution ν has full coverage (i.e., diverse):

$$\max_{\pi} \max_{s,a} \frac{d^{\pi}(s,a)}{\nu(s,a)} \leq C < \infty$$

Key Assumptions

1. offline distribution ν has full coverage (i.e., diverse):

Necessary for
today's Alg

$$\max_{\pi} \max_{s,a} \frac{d^{\pi}(s,a)}{\nu(s,a)} \leq C < \infty$$

Key Assumptions

1. offline distribution ν has full coverage (i.e., diverse):

Necessary for
today's Alg

$$\max_{\pi} \max_{s,a} \frac{d^{\pi}(s,a)}{\nu(s,a)} \leq C < \infty$$

2. Small inherent Bellman error, i.e., near Bellman

Completion (note it's averaged over ν):

$$\max_{g \in \mathcal{F}} \min_{f \in \mathcal{F}} \mathbb{E}_{s,a \sim \nu} (f(s,a) - \mathcal{T}g(s,a))^2 \leq \epsilon_{approx,\nu}$$

Key Assumptions

1. offline distribution ν has full coverage (i.e., diverse):

Necessary for
today's Alg

$$\max_{\pi} \max_{s,a} \frac{d^{\pi}(s,a)}{\nu(s,a)} \leq C < \infty$$

2. Small inherent Bellman error, i.e., near Bellman

Completion (note it's averaged over ν):

$$\max_{g \in \mathcal{F}} \min_{f \in \mathcal{F}} \mathbb{E}_{s,a \sim \nu} (f(s,a) - \mathcal{T}g(s,a))^2 \leq \epsilon_{approx,\nu}$$

Necessary in
general (we saw
realizability itself
is not enough)

Outline



1. Setting: Assumptions

2. Algorithm: Fitted Q Iteration

2. Guarantee and Proof sketch

The FQI Algorithm

1. offline data points obtained from ν :

$$\mathcal{D} = \{s, a, r, s'\}, \quad (s, a) \sim \nu, r = r(s, a), s' \sim P(\cdot | s, a)$$

The FQI Algorithm

1. offline data points obtained from ν :

$$\mathcal{D} = \{s, a, r, s'\}, \quad (s, a) \sim \nu, r = r(s, a), s' \sim P(\cdot | s, a)$$

2. Initialize $f_0 \in \mathcal{F}$, and iterate:

$$f_{t+1} = \arg \min_{f \in \mathcal{F}} \sum_{s, a, r, s' \in \mathcal{D}} \left(f(s, a) - r - \gamma \max_{a'} f_t(s', a') \right)^2$$

The FQI Algorithm

1. offline data points obtained from ν :

$$\mathcal{D} = \{s, a, r, s'\}, \quad (s, a) \sim \nu, r = r(s, a), s' \sim P(\cdot | s, a)$$

2. Initialize $f_0 \in \mathcal{F}$, and iterate:

$$f_{t+1} = \arg \min_{f \in \mathcal{F}} \sum_{s, a, r, s' \in \mathcal{D}} \left(f(s, a) - r - \gamma \max_{a'} f_t(s', a') \right)^2$$

3. After K iterations, return $\pi(s) = \arg \max_a f_K(s, a), \forall s$

The FQI Algorithm

1. offline data points obtained from ν :

$$\mathcal{D} = \{s, a, r, s'\}, \quad (s, a) \sim \nu, r = r(s, a), s' \sim P(\cdot | s, a)$$

2. Initialize $f_0 \in \mathcal{F}$, and iterate:

$$f_{t+1} = \arg \min_{f \in \mathcal{F}} \sum_{s, a, r, s' \in \mathcal{D}} \left(f(s, a) - r - \gamma \max_{a'} f_t(s', a') \right)^2$$

3. After K iterations, return $\pi(s) = \arg \max_a f_K(s, a), \forall s$

(Note: the algorithmic idea here is similar to DQNs [Deepmind 15])

Why we could expect it to work...

$$f_{t+1} = \arg \min_{f \in \mathcal{F}} \sum_{s,a,r,s' \in \mathcal{D}} \left(f(s,a) - r - \gamma \max_{a'} f_t(s',a') \right)^2$$

Why we could expect it to work...

$$f_{t+1} = \arg \min_{f \in \mathcal{F}} \sum_{s,a,r,s' \in \mathcal{D}} \left(f(s,a) - r - \gamma \max_{a'} f_t(s',a') \right)^2$$

$y := r(s,a) + \gamma \max_{a'} f_t(s',a')$

Why we could expect it to work...

$$f_{t+1} = \arg \min_{f \in \mathcal{F}} \sum_{s,a,r,s' \in \mathcal{D}} \left(f(s,a) - r - \gamma \max_{a'} f_t(s',a') \right)^2$$

$y := r(s,a) + \gamma \max_{a'} f_t(s',a')$

Bayes optimal: $r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_a f_t(s',a')$

$\underbrace{\hspace{15em}}_{(\mathcal{T}f_t)(s,a)}$

Why we could expect it to work...

$$f_{t+1} = \arg \min_{f \in \mathcal{F}} \sum_{s,a,r,s' \in \mathcal{D}} \left(f(s,a) - r - \gamma \max_{a'} f_t(s',a') \right)^2$$

$y := r(s,a) + \gamma \max_{a'} f_t(s',a')$

Bayes optimal: $r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_a f_t(s',a')$

$\underbrace{\hspace{15em}}_{(\mathcal{T}f_t)(s,a)}$

1. **Near Bellman completion** means regression target $\mathcal{T}f_t$ nearly belongs to \mathcal{F}

$$\mathbb{E}_{s,a \sim \nu} \left(f_{t+1}(s,a) - \mathcal{T}f_t(s,a) \right)^2 \approx \frac{1}{N} + \epsilon_{approx,\nu}$$

Why we could expect it to work...

$$f_{t+1} = \arg \min_{f \in \mathcal{F}} \sum_{s,a,r,s' \in \mathcal{D}} \left(f(s,a) - r - \gamma \max_{a'} f_t(s',a') \right)^2$$

$y := r(s,a) + \gamma \max_{a'} f_t(s',a')$

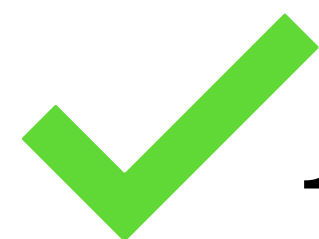
Bayes optimal: $r(s,a) + \gamma \underbrace{\mathbb{E}_{s' \sim P(\cdot|s,a)} \max_a f_t(s',a')}_{(\mathcal{T}f_t)(s,a)}$

1. **Near Bellman completion** means regression target $\mathcal{T}f_t$ nearly belongs to \mathcal{F}

$$\mathbb{E}_{s,a \sim \nu} (f_{t+1}(s,a) - \mathcal{T}f_t(s,a))^2 \approx \frac{1}{N} + \epsilon_{approx,\nu}$$

2. $f_{t+1} \approx \mathcal{T}f_t$ (under **the diverse ν**), i.e., it's like Value Iteration, we could hope for a convergence

Outline



1. Setting: Assumptions



2. Algorithm: Fitted Q Iteration

2. Guarantee and Proof sketch

Theorem

Theorem: Fix iteration number K , w/ probability at least $1 - \delta$,

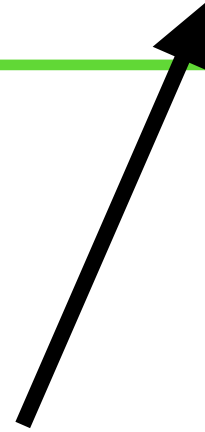
$$V^{\star} - V^{\pi} \leq O \left(\frac{1}{(1 - \gamma)^4} \sqrt{\frac{C \ln(|\mathcal{F}| K / \delta)}{N}} + \frac{1}{(1 - \gamma)^3} \sqrt{C \epsilon_{approx, \nu}} \right) + \frac{2\gamma^K}{(1 - \gamma)^2}$$

Theorem

Theorem: Fix iteration number K , w/ probability at least $1 - \delta$,

$$V^* - V^\pi \leq O \left(\frac{1}{(1 - \gamma)^4} \sqrt{\frac{C \ln(|\mathcal{F}| K / \delta)}{N}} + \frac{1}{(1 - \gamma)^3} \sqrt{C \epsilon_{approx, \nu}} \right) + \frac{2\gamma^K}{(1 - \gamma)^2}$$

Statistical error related to
regression



Theorem

Theorem: Fix iteration number K , w/ probability at least $1 - \delta$,

$$V^* - V^\pi \leq O \left(\frac{1}{(1-\gamma)^4} \sqrt{\frac{C \ln(|\mathcal{F}| K/\delta)}{N}} + \frac{1}{(1-\gamma)^3} \sqrt{C \epsilon_{approx,\nu}} \right) + \frac{2\gamma^K}{(1-\gamma)^2}$$

Statistical error related to
regression

Inherent Bellman error

Theorem

Theorem: Fix iteration number K , w/ probability at least $1 - \delta$,

$$V^* - V^\pi \leq O \left(\frac{1}{(1-\gamma)^4} \sqrt{\frac{C \ln(|\mathcal{F}| K/\delta)}{N}} + \frac{1}{(1-\gamma)^3} \sqrt{C \epsilon_{approx,\nu}} \right) + \frac{2\gamma^K}{(1-\gamma)^2}$$

Statistical error related to
regression

Inherent Bellman error

VI-style
Convergence rate

Step 1:

Least Squares regression ensure near Bellman consistency (averaged over ν)

Standard Generalization Bound for regression:

Step 1:

Least Squares regression ensure near Bellman consistency (averaged over ν)

Standard Generalization Bound for regression:

Given $\{x_i, y_i\}_{i=1}^N$, $(x_i, y_i) \sim \nu$, $y_i = f^*(x_i) + \epsilon_i$, where $|y_i| \leq Y$, $\|f^*\|_\infty \leq Y$,

Step 1:

Least Squares regression ensure near Bellman consistency (averaged over ν)

Standard Generalization Bound for regression:

Given $\{x_i, y_i\}_{i=1}^N$, $(x_i, y_i) \sim \nu$, $y_i = f^\star(x_i) + \epsilon_i$, where $|y_i| \leq Y$, $\|f^\star\|_\infty \leq Y$,
a function class $\mathcal{F} = \{f : \mathcal{X} \mapsto [-Y, Y]\}$, where $\min_{f \in \mathcal{F}} \mathbb{E}_{x \sim \nu} (f(x) - f^\star(x))^2 \leq \epsilon$

Step 1:

Least Squares regression ensure near Bellman consistency (averaged over ν)

Standard Generalization Bound for regression:

Given $\{x_i, y_i\}_{i=1}^N$, $(x_i, y_i) \sim \nu$, $y_i = f^\star(x_i) + \epsilon_i$, where $|y_i| \leq Y$, $\|f^\star\|_\infty \leq Y$,

a function class $\mathcal{F} = \{f : \mathcal{X} \mapsto [-Y, Y]\}$, where $\min_{f \in \mathcal{F}} \mathbb{E}_{x \sim \nu} (f(x) - f^\star(x))^2 \leq \epsilon$

Denote $\hat{f} := \arg \min_{f \in \mathcal{F}} \sum_{i=1}^N (f(x_i) - y_i)^2$ as the least square minimizer, then w/ prob $1 - \delta$:

$$\mathbb{E}_{x \sim \nu} \left(\hat{f}(x) - f^\star(x) \right)^2 \leq O \left(\frac{Y^2 \ln(|\mathcal{F}|/\delta)}{N} + \epsilon \right)$$

Step 1:

Least Squares regression ensure near Bellman consistency (averaged over ν)

1. Recall FQI's regression problem:

$$f_{t+1} = \arg \min_{f \in \mathcal{F}} \sum_{s, a, r, s' \in \mathcal{D}} \left(f(s, a) - r - \gamma \max_{a'} f_t(s', a') \right)^2$$

Step 1:

Least Squares regression ensure near Bellman consistency (averaged over ν)

1. Recall FQI's regression problem:

$$f_{t+1} = \arg \min_{f \in \mathcal{F}} \sum_{s,a,r,s' \in \mathcal{D}} \left(f(s,a) - r - \gamma \max_{a'} f_t(s',a') \right)^2$$

2. Here Bayes optimal is $f^\star := \mathcal{T} f_t$,

Step 1:

Least Squares regression ensure near Bellman consistency (averaged over ν)

1. Recall FQI's regression problem:

$$f_{t+1} = \arg \min_{f \in \mathcal{F}} \sum_{s,a,r,s' \in \mathcal{D}} \left(f(s,a) - r - \gamma \max_{a'} f_t(s',a') \right)^2$$

2. Here Bayes optimal is $f^* := \mathcal{T} f_t$,

3. Via small inherent BE, we know that $\min_{f \in \mathcal{F}} \mathbb{E}_{s,a \sim \nu} (f(s,a) - \mathcal{T} f_t(s,a))^2 \leq \epsilon_{approx,\nu}$

Step 1:

Least Squares regression ensure near Bellman consistency (averaged over ν)

1. Recall FQI's regression problem:

$$f_{t+1} = \arg \min_{f \in \mathcal{F}} \sum_{s,a,r,s' \in \mathcal{D}} \left(f(s,a) - r - \gamma \max_{a'} f_t(s',a') \right)^2$$

2. Here Bayes optimal is $f^* := \mathcal{T} f_t$,

3. Via small inherent BE, we know that $\min_{f \in \mathcal{F}} \mathbb{E}_{s,a \sim \nu} (f(s,a) - \mathcal{T} f_t(s,a))^2 \leq \epsilon_{approx,\nu}$

$$1+2+3 \Rightarrow \mathbb{E}_{s,a \sim \nu} (f_{t+1}(s,a) - \mathcal{T} f_t(s,a))^2 \leq \frac{1}{(1-\gamma)^2} \frac{\ln(|\mathcal{F}|/\delta)}{N} + \epsilon_{approx,\nu}$$

Step 1:

Least Squares regression ensure near Bellman consistency (averaged over ν)

1. Recall FQI's regression problem:

$$f_{t+1} = \arg \min_{f \in \mathcal{F}} \sum_{s,a,r,s' \in \mathcal{D}} \left(f(s,a) - r - \gamma \max_{a'} f_t(s',a') \right)^2$$

2. Here Bayes optimal is $f^* := \mathcal{T} f_t$,

3. Via small inherent BE, we know that $\min_{f \in \mathcal{F}} \mathbb{E}_{s,a \sim \nu} (f(s,a) - \mathcal{T} f_t(s,a))^2 \leq \epsilon_{approx,\nu}$

$$1+2+3 \Rightarrow \mathbb{E}_{s,a \sim \nu} (f_{t+1}(s,a) - \mathcal{T} f_t(s,a))^2 \leq \frac{1}{(1-\gamma)^2} \frac{\ln(|\mathcal{F}|/\delta)}{N} + \epsilon_{approx,\nu}$$

$$\mathbb{E}_{s,a \sim \nu} |f_{t+1}(s,a) - \mathcal{T} f_t(s,a)| \leq \underbrace{\sqrt{\frac{1}{(1-\gamma)^2} \frac{\ln(|\mathcal{F}|/\delta)}{N} + \epsilon_{approx,\nu}}}_{:= \epsilon_{regress}}$$

Step 2:

Near Bellman consistency implies convergence

Consider any state-action distribution $\beta(s, a)$ (induced by some policy)

$$\sqrt{\mathbb{E}_{s,a \sim \beta} (f_t(s, a) - Q^*(s, a))^2} := \|f_t - Q^*\|_{\beta, 2}$$

Step 2:

Near Bellman consistency implies convergence

Consider any state-action distribution $\beta(s, a)$ (induced by some policy)

$$\begin{aligned} & \sqrt{\mathbb{E}_{s,a \sim \beta} (f_t(s, a) - Q^*(s, a))^2} := \|f_t - Q^*\|_{\beta, 2} \\ & \leq \|f_t - \mathcal{T}f_{t-1}\|_{2, \beta} + \|\mathcal{T}f_{t-1} - Q^*\|_{2, \beta} \end{aligned}$$

Step 2:

Near Bellman consistency implies convergence

Consider any state-action distribution $\beta(s, a)$ (induced by some policy)

$$\begin{aligned} & \sqrt{\mathbb{E}_{s,a \sim \beta} (f_t(s, a) - Q^*(s, a))^2} := \|f_t - Q^*\|_{\beta, 2} \\ & \leq \|f_t - \mathcal{T}f_{t-1}\|_{2, \beta} + \|\mathcal{T}f_{t-1} - Q^*\|_{2, \beta} \end{aligned}$$

Dist-change and
Coverage condition

Step 2:

Near Bellman consistency implies convergence

Consider any state-action distribution $\beta(s, a)$ (induced by some policy)

$$\begin{aligned} & \sqrt{\mathbb{E}_{s,a \sim \beta} (f_t(s, a) - Q^*(s, a))^2} := \|f_t - Q^*\|_{\beta, 2} \\ & \leq \|f_t - \mathcal{T}f_{t-1}\|_{2, \beta} + \|\mathcal{T}f_{t-1} - Q^*\|_{2, \beta} \\ & \leq \sqrt{C} \|f_t - \mathcal{T}f_{t-1}\|_{2, \nu} + \|\mathcal{T}f_{t-1} - Q^*\|_{2, \beta} \end{aligned}$$

Dist-change and
Coverage condition

Step 2:

Near Bellman consistency implies convergence

Consider any state-action distribution $\beta(s, a)$ (induced by some policy)

$$\sqrt{\mathbb{E}_{s,a \sim \beta} (f_t(s, a) - Q^*(s, a))^2} := \|f_t - Q^*\|_{\beta, 2}$$

$$\leq \|f_t - \mathcal{T}f_{t-1}\|_{2, \beta} + \|\mathcal{T}f_{t-1} - Q^*\|_{2, \beta}$$

$$\leq \sqrt{C} \|f_t - \mathcal{T}f_{t-1}\|_{2, \nu} + \|\mathcal{T}f_{t-1} - Q^*\|_{2, \beta}$$

Dist-change and
Coverage condition

$$\leq \sqrt{C} \epsilon_{regress} + \gamma \sqrt{\mathbb{E}_{s,a \sim \beta} \left(\mathbb{E}_{s' \sim P(\cdot | s, a)} \left(\max_{a'} f_{t-1}(s', a') - \max_{a'} Q^*(s', a') \right) \right)^2}$$

Step 2:

Near Bellman consistency implies convergence

Consider any state-action distribution $\beta(s, a)$ (induced by some policy)

$$\sqrt{\mathbb{E}_{s,a \sim \beta} (f_t(s, a) - Q^*(s, a))^2} := \|f_t - Q^*\|_{\beta, 2}$$

$$\leq \|f_t - \mathcal{T}f_{t-1}\|_{2, \beta} + \|\mathcal{T}f_{t-1} - Q^*\|_{2, \beta}$$

$$\leq \sqrt{C} \|f_t - \mathcal{T}f_{t-1}\|_{2, \nu} + \|\mathcal{T}f_{t-1} - Q^*\|_{2, \beta}$$

Dist-change and
Coverage condition

$$\leq \sqrt{C} \epsilon_{regress} + \gamma \sqrt{\mathbb{E}_{s,a \sim \beta} \left(\mathbb{E}_{s' \sim P(\cdot | s, a)} \left(\max_{a'} f_{t-1}(s', a') - \max_{a'} Q^*(s', a') \right) \right)^2}$$

$$\leq \sqrt{C} \epsilon_{regress} + \gamma \sqrt{\underbrace{\mathbb{E}_{s,a \sim \beta} \mathbb{E}_{s' \sim P(\cdot | s, a)} \max_{a'} (f_{t-1}(s', a') - Q^*(s', a'))^2}_{:= \beta'(s', a')}}}$$

Step 2:

Near Bellman consistency implies convergence

Consider any state-action distribution $\beta(s, a)$ (induced by some policy)

$$\sqrt{\mathbb{E}_{s,a \sim \beta} (f_t(s, a) - Q^*(s, a))^2} := \|f_t - Q^*\|_{\beta, 2}$$

$$\leq \|f_t - \mathcal{T}f_{t-1}\|_{2, \beta} + \|\mathcal{T}f_{t-1} - Q^*\|_{2, \beta}$$

$$\leq \sqrt{C} \|f_t - \mathcal{T}f_{t-1}\|_{2, \nu} + \|\mathcal{T}f_{t-1} - Q^*\|_{2, \beta}$$

Dist-change and
Coverage condition

$$\leq \sqrt{C} \epsilon_{regress} + \gamma \sqrt{\mathbb{E}_{s,a \sim \beta} \left(\mathbb{E}_{s' \sim P(\cdot | s, a)} \left(\max_{a'} f_{t-1}(s', a') - \max_{a'} Q^*(s', a') \right) \right)^2}$$

$$\leq \sqrt{C} \epsilon_{regress} + \gamma \sqrt{\underbrace{\mathbb{E}_{s,a \sim \beta} \mathbb{E}_{s' \sim P(\cdot | s, a)} \max_{a'} (f_{t-1}(s', a') - Q^*(s', a'))^2}_{:= \beta'(s', a')}} = \sqrt{C} \epsilon_{regress} + \gamma \|f_{t-1} - Q^*\|_{2, \beta'}$$

Step 2:

Near Bellman consistency implies convergence

Consider **ANY** state-action distribution $\beta(s, a)$ (induced by some policy)

$$\sqrt{\mathbb{E}_{s,a \sim \beta} (f_t(s, a) - Q^*(s, a))^2} := \|f_t - Q^*\|_{\beta, 2} \leq \sqrt{C} \epsilon_{regress} + \gamma \|f_{t-1} - Q^*\|_{2, \beta'}$$

Step 2:

Near Bellman consistency implies convergence

Consider **ANY** state-action distribution $\beta(s, a)$ (induced by some policy)

$$\begin{aligned} \sqrt{\mathbb{E}_{s,a \sim \beta} (f_t(s, a) - Q^*(s, a))^2} &:= \|f_t - Q^*\|_{\beta, 2} \leq \sqrt{C} \epsilon_{regress} + \gamma \|f_{t-1} - Q^*\|_{2, \beta'} \\ &\leq \sqrt{C} \epsilon_{regress} + \gamma \left[\sqrt{C} \epsilon_{regress} + \gamma \|f_{t-2} - Q^*\|_{2, \beta''} \right] \end{aligned}$$

Step 2:

Near Bellman consistency implies convergence

Consider **ANY** state-action distribution $\beta(s, a)$ (induced by some policy)

$$\begin{aligned} \sqrt{\mathbb{E}_{s,a \sim \beta} (f_t(s, a) - Q^*(s, a))^2} &:= \|f_t - Q^*\|_{\beta, 2} \leq \sqrt{C} \epsilon_{regress} + \gamma \|f_{t-1} - Q^*\|_{2, \beta'} \\ &\leq \sqrt{C} \epsilon_{regress} + \gamma \left[\sqrt{C} \epsilon_{regress} + \gamma \|f_{t-2} - Q^*\|_{2, \beta''} \right] \\ &\leq \sqrt{C} \epsilon_{regress} (1 + \gamma + \dots + \gamma^k) + \gamma^k \|f_0 - Q^*\|_{2, \tilde{\beta}} \end{aligned}$$

Step 2:

Near Bellman consistency implies convergence

Consider **ANY** state-action distribution $\beta(s, a)$ (induced by some policy)

$$\begin{aligned} \sqrt{\mathbb{E}_{s,a \sim \beta} (f_t(s, a) - Q^*(s, a))^2} &:= \|f_t - Q^*\|_{\beta, 2} \leq \sqrt{C} \epsilon_{regress} + \gamma \|f_{t-1} - Q^*\|_{2, \beta'} \\ &\leq \sqrt{C} \epsilon_{regress} + \gamma \left[\sqrt{C} \epsilon_{regress} + \gamma \|f_{t-2} - Q^*\|_{2, \beta''} \right] \\ &\leq \sqrt{C} \epsilon_{regress} (1 + \gamma + \dots + \gamma^k) + \gamma^k \|f_0 - Q^*\|_{2, \tilde{\beta}} \\ &\leq \frac{\sqrt{C} \epsilon_{regress}}{1 - \gamma} + \gamma^k / (1 - \gamma) \end{aligned}$$

Step 3:

Turn in error $\|f_k - Q^\star\|_{2,\beta}$ to policy π^k performance

Denote $\pi^k(s) = \arg \max_a f_k(s, a)$

We know f_k is close to Q^\star (averaged over any distribution):

$$V^\star - V^{\pi^k} = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi^k}} [Q^\star(s, \pi^\star(s)) - Q^\star(s, a)]$$

Step 3:

Turn in error $\|f_k - Q^\star\|_{2,\beta}$ to policy π^k performance

Denote $\pi^k(s) = \arg \max_a f_k(s, a)$

We know f_k is close to Q^\star (averaged over any distribution):

$$\begin{aligned} V^\star - V^{\pi^k} &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi^k}} [Q^\star(s, \pi^\star(s)) - Q^\star(s, a)] \\ &\leq \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi^k}} [Q^\star(s, \pi^\star(s)) - f_k(s, \pi^\star(s)) + f_k(s, \pi^k(s)) - Q^\star(s, a)] \end{aligned}$$

Step 3:

Turn in error $\|f_k - Q^\star\|_{2,\beta}$ to policy π^k performance

Denote $\pi^k(s) = \arg \max_a f_k(s, a)$

We know f_k is close to Q^\star (averaged over any distribution):

$$\begin{aligned} V^\star - V^{\pi^k} &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi^k}} [Q^\star(s, \pi^\star(s)) - Q^\star(s, a)] \\ &\leq \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi^k}} [Q^\star(s, \pi^\star(s)) - f_k(s, \pi^\star(s)) + f_k(s, \pi^k(s)) - Q^\star(s, a)] \\ &\leq \frac{1}{1-\gamma} \left[\sqrt{\mathbb{E}_{s \sim d^{\pi^k}} (Q^\star(s, \pi^\star(s)) - f_k(s, \pi^\star(s)))^2} + \sqrt{\mathbb{E}_{s \sim d^{\pi^k}} (f_k(s, \pi^k(s)) - Q^\star(s, \pi^k(s)))^2} \right] \end{aligned}$$

Step 3:

Turn in error $\|f_k - Q^\star\|_{2,\beta}$ to policy π^k performance

Denote $\pi^k(s) = \arg \max_a f_k(s, a)$

We know f_k is close to Q^\star (averaged over any distribution):

$$\begin{aligned} V^\star - V^{\pi^k} &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi^k}} [Q^\star(s, \pi^\star(s)) - Q^\star(s, a)] \\ &\leq \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi^k}} [Q^\star(s, \pi^\star(s)) - f_k(s, \pi^\star(s)) + f_k(s, \pi^k(s)) - Q^\star(s, a)] \\ &\leq \frac{1}{1-\gamma} \left[\sqrt{\mathbb{E}_{s \sim d^{\pi^k}} (Q^\star(s, \pi^\star(s)) - f_k(s, \pi^\star(s)))^2} + \sqrt{\mathbb{E}_{s \sim d^{\pi^k}} (f_k(s, \pi^k(s)) - Q^\star(s, \pi^k(s)))^2} \right] \\ &\leq \frac{2}{1-\gamma} \left(\frac{\sqrt{C} \epsilon_{regress}}{1-\gamma} + \frac{\gamma^k}{1-\gamma} \right) \end{aligned}$$

To conclude:

$$V^\star - V^{\pi^k} \leq \frac{2}{1-\gamma} \left(\frac{\sqrt{C} \epsilon_{regress}}{1-\gamma} + \frac{\gamma^k}{1-\gamma} \right) \quad \text{where } \epsilon_{regress} = \sqrt{\frac{1}{(1-\gamma)^2} \frac{\ln(|\mathcal{F}|/\delta)}{N}} + \epsilon_{approx,\nu}$$

To conclude:

$$V^\star - V^{\pi^k} \leq \frac{2}{1-\gamma} \left(\frac{\sqrt{C} \epsilon_{regress}}{1-\gamma} + \frac{\gamma^k}{1-\gamma} \right) \quad \text{where } \epsilon_{regress} = \sqrt{\frac{1}{(1-\gamma)^2} \frac{\ln(|\mathcal{F}|/\delta)}{N}} + \epsilon_{approx,\nu}$$

1. Least square ensures we have near Bellman consistency under ν :

$$\|f_t - \mathcal{T}f_{t-1}\|_{2,\nu} \leq \epsilon_{regress}$$

To conclude:

$$V^\star - V^{\pi^k} \leq \frac{2}{1-\gamma} \left(\frac{\sqrt{C}\epsilon_{regress}}{1-\gamma} + \frac{\gamma^k}{1-\gamma} \right) \quad \text{where } \epsilon_{regress} = \sqrt{\frac{1}{(1-\gamma)^2} \frac{\ln(|\mathcal{F}|/\delta)}{N}} + \epsilon_{approx,\nu}$$

1. Least square ensures we have near Bellman consistency under ν :

$$\|f_t - \mathcal{T}f_{t-1}\|_{2,\nu} \leq \epsilon_{regress}$$

2. Near Bellman consistency under $\nu + \nu$ covers all other possible distributions β :

$$\|f_t - Q^\star\|_{2,\beta} \leq \left(\sqrt{C\epsilon_{regress}} + \gamma^k \right) \cdot \text{poly}(1/(1-\gamma))$$

To conclude:

$$V^\star - V^{\pi^k} \leq \frac{2}{1-\gamma} \left(\frac{\sqrt{C}\epsilon_{regress}}{1-\gamma} + \frac{\gamma^k}{1-\gamma} \right) \quad \text{where } \epsilon_{regress} = \sqrt{\frac{1}{(1-\gamma)^2} \frac{\ln(|\mathcal{F}|/\delta)}{N}} + \epsilon_{approx,\nu}$$

1. Least square ensures we have near Bellman consistency under ν :

$$\|f_t - \mathcal{T}f_{t-1}\|_{2,\nu} \leq \epsilon_{regress}$$

2. Near Bellman consistency under $\nu + \nu$ covers all other possible distributions β :

$$\|f_t - Q^\star\|_{2,\beta} \leq \left(\sqrt{C\epsilon_{regress}} + \gamma^k \right) \cdot \text{poly}(1/(1-\gamma))$$

3. Like what we did in VI, turn f_t 's approximation error to its policy's performance ($1/(1-\gamma)$ amplification):