CSE 542: Statistical Reinforcement Learning

Lecture 14: Fitted Q-Iteration

Kevin Jamieson

Paul G. Allen School of Computer Science & Engineering
University of Washington

Outline

Today's Plan

Beyond Linear Function Approximation

Lectures 12–13: Linear class $\mathcal{F} = \{\langle \phi, w \rangle\}$. Strong structure (Bellman completeness or Linear MDP) buys $\mathrm{poly}(d, H)$ rates.

This lecture: a more flexible style of analysis.

- $\mathcal{F}$ is an arbitrary class of Q-functions.
- No structural assumption on P, r.
- Offline data, no choice in queries.
- Algorithm: Fitted Q-Iteration (FQI) — regression onto Bellman targets.

Beyond Linear Function Approximation

Lectures 12–13: Linear class $\mathcal{F} = \{\langle \phi, w \rangle\}$. Strong structure (Bellman completeness or Linear MDP) buys $\mathrm{poly}(d, H)$ rates.

This lecture: a more flexible style of analysis.

- $\mathcal{F}$ is an arbitrary class of Q-functions.
- No structural assumption on P, r.
- Offline data, no choice in queries.
- Algorithm: Fitted Q-Iteration (FQI) — regression onto Bellman targets.

Key idea

Reduce RL to a sequence of *supervised regression* problems. Trade structural assumptions on P, r for two quantitative assumptions on the *data distribution* and the *function class*.

Beyond Linear Function Approximation

Lectures 12–13: Linear class $\mathcal{F} = \{\langle \phi, w \rangle\}$. Strong structure (Bellman completeness or Linear MDP) buys $\mathrm{poly}(d, H)$ rates.

This lecture: a more flexible style of analysis.

- $\mathcal{F}$ is an arbitrary class of Q-functions.
- No structural assumption on P, r.
- Offline data, no choice in queries.
- Algorithm: Fitted Q-Iteration (FQI) — regression onto Bellman targets.

Key idea

Reduce RL to a sequence of *supervised regression* problems. Trade structural assumptions on P, r for two quantitative assumptions on the *data distribution* and the *function class*.

Plan today. Start with *finite horizon* (cleaner, mirrors Lectures 12–13), then extend to the *infinite-horizon discounted* setting (book's main treatment, with one extra wrinkle).

Two Settings

Finite horizon (Lectures 12–13)

- Horizon H.
- Stage-dependent P_h, r_h.
- Per-stage operators $\mathcal{T}_h$.
- Backward induction: $V_{H+1} \equiv 0$, then go back.
- Values in $[0, H]$.

Infinite horizon discounted (book)

- Discount $\gamma \in (0, 1)$.
- Stationary P, r.
- Single Bellman operator $\mathcal{T}$, fixed point $Q^\star$.
- Iterate K times until γ^K tail is small.
- Values in $[0, V_{\max}]$ where $V_{\max} := 1/(1-\gamma)$.

Two Settings

Finite horizon (Lectures 12–13)

- Horizon H.
- Stage-dependent P_h, r_h.
- Per-stage operators $\mathcal{T}_h$.
- Backward induction: $V_{H+1} \equiv 0$, then go back.
- Values in $[0, H]$.

Infinite horizon discounted (book)

- Discount $\gamma \in (0, 1)$.
- Stationary P, r.
- Single Bellman operator $\mathcal{T}$, fixed point $Q^\star$.
- Iterate K times until γ^K tail is small.
- Values in $[0, V_{\max}]$ where $V_{\max} := 1/(1-\gamma)$.

What's similar. Both use offline data, both rely on concentrability (data covers what optimal policies visit) and inherent Bellman error (the class is approximately closed under Bellman backups).

What's different. Infinite horizon has a fixed-point structure and an extra "iteration" error term $\gamma^K V_{\max}$. Finite horizon does not iterate — one regression per stage, H stages, done.

Today's Plan

Setup: Finite-Horizon MDP

Finite-horizon MDP with horizon H, stage-dependent transitions P_h, rewards $r_h \in [0,1]$, initial-state distribution ν. Stage-dependent Bellman operator

$$(\mathcal{T}_h f)(s,a) \;:=\; r_h(s,a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)}\big[\max_{a'} f(s',a')\big].$$

Optimal Q-functions $Q_h^\star$ satisfy $Q_h^\star = \mathcal{T}_h Q_{h+1}^\star$ with $Q_{H+1}^\star \equiv 0$.

Stagewise function class. $\mathcal{F}_h \subseteq \{f : \mathcal{S} \times \mathcal{A} \to [0,H]\}$ for $h \in [H]$; write $\mathcal{F} := \bigcup_h \mathcal{F}_h$.

Per-stage offline data. For each $h \in [H]$,

$$\mathcal{D}_h \;=\; \{(s_h^i, a_h^i, r_h^i, s_{h+1}^i)\}_{i=1}^n, \quad (s_h^i, a_h^i) \sim \nu_h,$$

collected by some unknown behavior policy.

For a distribution ρ on $\mathcal{S} \times \mathcal{A}$, $\|f\|_{2,\rho}^2 := \mathbb{E}_{(s,a)\sim\rho}[f(s,a)^2]$.

Two Requirements (Finite Horizon)

Let $d_h^\pi \in \Delta(\mathcal{S} \times \mathcal{A})$ be the visitation distribution of π at stage h, starting from $s_1 \sim \nu$.

Assumption (Concentrability)

There exists $C \geq 1$ such that for every policy π and every $h \in [H]$,

$$\sup_{(s,a)} \frac{d_h^\pi(s,a)}{\nu_h(s,a)} \leq C.$$

Two Requirements (Finite Horizon)

Let $d_h^\pi \in \Delta(\mathcal{S} \times \mathcal{A})$ be the visitation distribution of π at stage h, starting from $s_1 \sim \nu$.

Assumption (Concentrability)

There exists $C \geq 1$ such that for every policy π and every $h \in [H]$,

$$\sup_{(s,a)} \frac{d_h^\pi(s,a)}{\nu_h(s,a)} \leq C.$$

Assumption (Inherent Bellman error)

$$\varepsilon_{\mathrm{approx},h} := \max_{f \in \mathcal{F}_{h+1}} \min_{f' \in \mathcal{F}_h} \left\| f' - \mathcal{T}_h f \right\|_{2,\nu_h}^2, \qquad \varepsilon_{\mathrm{approx},\nu} := \max_h \varepsilon_{\mathrm{approx},h}.$$

Two Requirements (Finite Horizon)

Let $d_h^\pi \in \Delta(\mathcal{S} \times \mathcal{A})$ be the visitation distribution of π at stage h, starting from $s_1 \sim \nu$.

Assumption (Concentrability)

There exists $C \geq 1$ such that for every policy π and every $h \in [H]$,

$$\sup_{(s,a)} \frac{d_h^\pi(s,a)}{\nu_h(s,a)} \leq C.$$

Assumption (Inherent Bellman error)

$$\varepsilon_{\mathrm{approx},h} := \max_{f \in \mathcal{F}_{h+1}} \min_{f' \in \mathcal{F}_h} \left\| f' - \mathcal{T}_h f \right\|_{2,\nu_h}^2, \qquad \varepsilon_{\mathrm{approx},\nu} := \max_h \varepsilon_{\mathrm{approx},h}.$$

Interpretation. $\varepsilon_{\mathrm{approx},\nu} = 0$ means $\mathcal{F}$ is $L_2(\nu)$-Bellman complete; otherwise an irreducible approximation floor.

Caveat. Concentrability here is *all-policy* — the proof will use occupancies of both $\pi^\star$ and $\widehat{\pi}$.

Backward Fitted Q-Iteration

Set $f_{H+1} \equiv 0$. For $h = H, H-1, \ldots, 1$:

Finite-horizon FQI

$$f_h \in \arg\min_{f \in \mathcal{F}_h} \sum_{i=1}^{n} \left(f(s_h^i, a_h^i) - r_h^i - \max_{a'} f_{h+1}(s_{h+1}^i, a') \right)^2.$$

Output greedy policy $\widehat{\pi}_h(s) := \arg\max_a f_h(s, a)$.

Backward Fitted Q-Iteration

Set $f_{H+1} \equiv 0$. For $h = H, H-1, \ldots, 1$:

Finite-horizon FQI

$$f_h \in \arg\min_{f \in \mathcal{F}_h} \sum_{i=1}^{n} \left(f(s_h^i, a_h^i) - r_h^i - \max_{a'} f_{h+1}(s_{h+1}^i, a') \right)^2.$$

Output greedy policy $\widehat{\pi}_h(s) := \arg\max_a f_h(s, a)$.

Conditional mean of the target. For each sample,

$$\mathbb{E}\left[r_h^i + \max_{a'} f_{h+1}(s_{h+1}^i, a') \,\middle|\, s_h^i, a_h^i \right] = (\mathcal{T}_h f_{h+1})(s_h^i, a_h^i).$$

So we regress onto $\mathcal{T}_h f_{h+1}$, which is realizable up to $\sqrt{\varepsilon_{\text{approx},h}}$ by Assumption 2.

Backward Fitted Q-Iteration

Set $f_{H+1} \equiv 0$. For $h = H, H-1, \ldots, 1$:

Finite-horizon FQI

$$f_h \in \arg\min_{f \in \mathcal{F}_h} \sum_{i=1}^{n} \left(f(s_h^i, a_h^i) - r_h^i - \max_{a'} f_{h+1}(s_{h+1}^i, a') \right)^2.$$

Output greedy policy $\widehat{\pi}_h(s) := \arg\max_a f_h(s, a)$.

Conditional mean of the target. For each sample,

$$\mathbb{E}\left[r_h^i + \max_{a'} f_{h+1}(s_{h+1}^i, a') \,\middle|\, s_h^i, a_h^i \right] = (\mathcal{T}_h f_{h+1})(s_h^i, a_h^i).$$

So we regress onto $\mathcal{T}_h f_{h+1}$, which is realizable up to $\sqrt{\varepsilon_{\text{approx},h}}$ by Assumption 2.

Why this is a clean regression. At stage h, the target depends on f_{h+1}, which is *already determined* from the previous (later-stage) backward step. The variable f we are optimizing affects only the left side of the squared error — so the regression target is fixed and unbiased.

One regression per stage. No outer "iteration" loop. H regressions total.

Lemma A: Error Propagation (Finite Horizon)

Per-stage Bellman residual. $\Delta_h := f_h - \mathcal{T}_h f_{h+1}, \quad \varepsilon_h := \|\Delta_h\|_{2,\nu_h}$.

Recall Lecture 12, Lemma 1. We proved $V_0^\star - V_0^{\widehat{\pi}} \leq 2\sum_h \|\Delta_h\|_\infty$. But that's in ℓ_∞ — too loose when we only control $\|\Delta_h\|_{2,\nu_h}$.

Lemma A: Error Propagation (Finite Horizon)

Per-stage Bellman residual. $\Delta_h := f_h - \mathcal{T}_h f_{h+1}, \quad \varepsilon_h := \|\Delta_h\|_{2,\nu_h}$.

Recall Lecture 12, Lemma 1. We proved $V_0^\star - V_0^{\widehat{\pi}} \le 2\sum_h \|\Delta_h\|_\infty$. But that's in ℓ_∞ — too loose when we only control $\|\Delta_h\|_{2,\nu_h}$.

Sharper version (rerun Lecture 12's proof without taking sup). Two telescopes give

$$V_0^\star - \widehat{V}_0 \le \sum_h \|\Delta_h\|_{1,d_h^{\pi^\star}},$$

$$\widehat{V}_0 - V_0^{\widehat{\pi}} \le \sum_h \|\Delta_h\|_{1,d_h^{\widehat{\pi}}}.$$

So $V_0^\star - V_0^{\widehat{\pi}} \le \sum_h \left[\|\Delta_h\|_{1,d_h^{\pi^\star}} + \|\Delta_h\|_{1,d_h^{\widehat{\pi}}} \right]$.

Lemma A: Error Propagation (Finite Horizon)

Per-stage Bellman residual. $\Delta_h := f_h - \mathcal{T}_h f_{h+1}, \quad \varepsilon_h := \|\Delta_h\|_{2,\nu_h}$.

Recall Lecture 12, Lemma 1. We proved $V_0^\star - V_0^{\widehat{\pi}} \le 2\sum_h \|\Delta_h\|_\infty$. But that's in ℓ_∞ — too loose when we only control $\|\Delta_h\|_{2,\nu_h}$.

Sharper version (rerun Lecture 12's proof without taking sup). Two telescopes give

$$V_0^\star - \widehat{V}_0 \le \sum_h \|\Delta_h\|_{1,d_h^{\pi^\star}},$$

$$\widehat{V}_0 - V_0^{\widehat{\pi}} \le \sum_h \|\Delta_h\|_{1,d_h^{\widehat{\pi}}}.$$

So $V_0^\star - V_0^{\widehat{\pi}} \le \sum_h \left[\|\Delta_h\|_{1,d_h^{\pi^\star}} + \|\Delta_h\|_{1,d_h^{\widehat{\pi}}} \right]$.

Cauchy–Schwarz + concentrability. For any policy π and any g:
$\|g\|_{1,d_h^\pi} \le \|g\|_{2,d_h^\pi} \le \sqrt{C}\, \|g\|_{2,\nu_h}$. The change-of-measure uses *all-policy* concentrability, since we have both $\pi^\star$ and $\widehat{\pi}$.

Conclusion

$V_0^\star - V_0^{\widehat{\pi}} \le 2\sqrt{C} \sum_{h=1}^H \varepsilon_h$.

Lemma B: Per-Stage Regression Bound

Lemma (Stagewise regression error, finite $\mathcal{F}_h$)

With probability $\geq 1 - \delta$, simultaneously for all $h \in [H]$:

$$\varepsilon_h^2 \leq \frac{20\, H^2 \ln(2|\mathcal{F}_h|H/\delta)}{n} + 3\, \varepsilon_{\text{approx},h}.$$

Lemma B: Per-Stage Regression Bound

Lemma (Stagewise regression error, finite $\mathcal{F}_h$)

With probability $\geq 1-\delta$, simultaneously for all $h \in [H]$:

$$\varepsilon_h^2 \;\leq\; \frac{20\, H^2 \ln(2|\mathcal{F}_h|H/\delta)}{n} \;+\; 3\,\varepsilon_{\mathrm{approx},h}.$$

Proof sketch. Fix h and treat f_{h+1} as given (and independent of stage-h samples). Each target $y_i = r_h^i + \max_{a'} f_{h+1}(s_{h+1}^i, a')$ satisfies

- $\mathbb{E}[y_i \mid s_h^i, a_h^i] = (\mathcal{T}_h f_{h+1})(s_h^i, a_h^i)$.
- $y_i \in [0, H]$, $(\mathcal{T}_h f_{h+1}) \in [0, H]$.

By Assumption 2, $\min_{f \in \mathcal{F}_h} \|f - \mathcal{T}_h f_{h+1}\|_{2,\nu_h}^2 \leq \varepsilon_{\mathrm{approx},h}$.

Standard ERM analysis for bounded targets (AJKS Lemma A.11) gives the per-stage bound. Union bound over $h \in [H]$. □

Lemma B: Per-Stage Regression Bound

Lemma (Stagewise regression error, finite $\mathcal{F}_h$)

With probability $\geq 1-\delta$, simultaneously for all $h \in [H]$:

$$\varepsilon_h^2 \;\leq\; \frac{20\, H^2 \ln(2|\mathcal{F}_h|H/\delta)}{n} \;+\; 3\,\varepsilon_{\mathrm{approx},h}.$$

Proof sketch. Fix h and treat f_{h+1} as given (and independent of stage-h samples). Each target $y_i = r_h^i + \max_{a'} f_{h+1}(s_{h+1}^i, a')$ satisfies

- $\mathbb{E}[y_i \mid s_h^i, a_h^i] = (\mathcal{T}_h f_{h+1})(s_h^i, a_h^i)$.
- $y_i \in [0, H]$, $(\mathcal{T}_h f_{h+1}) \in [0, H]$.

By Assumption 2, $\min_{f \in \mathcal{F}_h} \|f - \mathcal{T}_h f_{h+1}\|_{2,\nu_h}^2 \leq \varepsilon_{\mathrm{approx},h}$.

Standard ERM analysis for bounded targets (AJKS Lemma A.11) gives the per-stage bound. Union bound over $h \in [H]$. □

Independence note. Use fresh data $\mathcal{D}_h$ at each stage, so f_{h+1} (built from $\mathcal{D}_{h+1}, \ldots, \mathcal{D}_H$) is independent of $\mathcal{D}_h$.

Theorem: Finite-Horizon FQI

Theorem

Under all-policy concentrability C and inherent Bellman error $\varepsilon_{\text{approx},\nu}$, with probability $\geq 1-\delta$:

$$V_0^\star - V_0^{\widehat{\pi}} \;\leq\; 2H\sqrt{C}\left[\sqrt{\frac{20\,H^2\ln(2|\mathcal{F}|H/\delta)}{n}} \;+\; \sqrt{3\,\varepsilon_{\text{approx},\nu}}\right].$$

Theorem: Finite-Horizon FQI

Theorem

Under all-policy concentrability C and inherent Bellman error $\varepsilon_{\text{approx},\nu}$, with probability $\geq 1-\delta$:

$$V_0^\star - V_0^{\widehat{\pi}} \;\leq\; 2H\sqrt{C}\left[\sqrt{\frac{20\,H^2\ln(2|\mathcal{F}|H/\delta)}{n}} \;+\; \sqrt{3\,\varepsilon_{\text{approx},\nu}}\right].$$

Proof. Lemma A gives $V_0^\star - V_0^{\widehat{\pi}} \leq 2\sqrt{C}\sum_h \varepsilon_h$. Lemma B bounds each ε_h. Using $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$ and summing over h: done. □

Theorem: Finite-Horizon FQI

Theorem

Under all-policy concentrability C and inherent Bellman error $\varepsilon_{\text{approx},\nu}$, with probability $\geq 1-\delta$:

$$V_0^\star - V_0^{\widehat{\pi}} \;\leq\; 2H\sqrt{C}\left[\sqrt{\frac{20\,H^2\ln(2|\mathcal{F}|H/\delta)}{n}} \;+\; \sqrt{3\,\varepsilon_{\text{approx},\nu}}\right].$$

Proof. Lemma A gives $V_0^\star - V_0^{\widehat{\pi}} \leq 2\sqrt{C}\sum_h \varepsilon_h$. Lemma B bounds each ε_h. Using $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$ and summing over h: done. □

Two error sources

1. **Statistical:** $H^2\sqrt{C\log|\mathcal{F}|/n}$. Vanishes as $n\to\infty$.
2. **Approximation:** $H\sqrt{C\,\varepsilon_{\text{approx},\nu}}$. The irreducible model floor.

No iteration term — backward FQI uses exactly H stages.

Today's Plan

Setup: Discounted MDP

Switch to a discounted MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$ with $\gamma \in (0,1)$, $r \in [0,1]$, initial distribution μ. $V_{\max} := 1/(1-\gamma)$.

Bellman optimality operator (stationary).

$$(\mathcal{T} f)(s,a) \;:=\; r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\Big[\max_{a'} f(s',a')\Big].$$

$Q^\star$ is the unique fixed point: $\mathcal{T} Q^\star = Q^\star$.

Function class. $\mathcal{F} \subseteq \{f : \mathcal{S} \times \mathcal{A} \to [0, V_{\max}]\}$.

Offline data. Single distribution ν, $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^n$ with $(s_i, a_i) \sim \nu$.

Setup: Discounted MDP

Switch to a discounted MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$ with $\gamma \in (0,1)$, $r \in [0,1]$, initial distribution μ. $V_{\max} := 1/(1-\gamma)$.

Bellman optimality operator (stationary).

$$(\mathcal{T} f)(s,a) \;:=\; r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\Big[\max_{a'} f(s',a')\Big].$$

$Q^\star$ is the unique fixed point: $\mathcal{T} Q^\star = Q^\star$.

Function class. $\mathcal{F} \subseteq \{f : \mathcal{S} \times \mathcal{A} \to [0, V_{\max}]\}$.

Offline data. Single distribution ν, $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^n$ with $(s_i, a_i) \sim \nu$.

New feature vs. finite horizon

There's no natural "stop after H stages" — backward FQI is undefined for an infinite-horizon MDP.

The algorithm will iterate $\mathcal{T}$ explicitly K times. K is a tuning parameter; the analysis will have a $\gamma^K V_{\max}$ tail.

Two Requirements (Discounted)

Discounted state-action occupancy:
$d_\mu^\pi(s,a) := (1-\gamma)\sum_{t\geq 0}\gamma^t \mathbb{P}_\pi(s_t = s, a_t = a \mid s_0 \sim \mu)$.

Assumption 4.1 (Concentrability)

There exists $C \geq 1$ such that for every (possibly non-stationary) policy π,
$\sup_{(s,a)} \frac{d_\mu^\pi(s,a)}{\nu(s,a)} \leq C$.

Two Requirements (Discounted)

Discounted state-action occupancy:
$d_\mu^\pi(s,a) := (1-\gamma)\sum_{t\geq 0}\gamma^t \mathbb{P}_\pi(s_t = s, a_t = a \mid s_0 \sim \mu)$.

Assumption 4.1 (Concentrability)

There exists $C \geq 1$ such that for every (possibly non-stationary) policy π,
$\sup_{(s,a)} \frac{d_\mu^\pi(s,a)}{\nu(s,a)} \leq C$.

Assumption 4.2 (Inherent Bellman error under ν)

$\varepsilon_{\text{approx},\nu} := \max_{f\in\mathcal{F}} \min_{f'\in\mathcal{F}} \|f' - \mathcal{T}f\|_{2,\nu}^2$.

Two Requirements (Discounted)

Discounted state-action occupancy:
$d_\mu^\pi(s,a) := (1-\gamma)\sum_{t\geq 0}\gamma^t \mathbb{P}_\pi(s_t = s, a_t = a \mid s_0 \sim \mu)$.

Assumption 4.1 (Concentrability)

There exists $C \geq 1$ such that for every (possibly non-stationary) policy π,
$\sup_{(s,a)} \frac{d_\mu^\pi(s,a)}{\nu(s,a)} \leq C$.

Assumption 4.2 (Inherent Bellman error under ν)

$\varepsilon_{\text{approx},\nu} := \max_{f\in\mathcal{F}} \min_{f'\in\mathcal{F}} \|f' - \mathcal{T}f\|_{2,\nu}^2$.

Same shape as finite-horizon. The only difference is the use of discounted occupancy d_μ^π in concentrability, and a single $\mathcal{F}$ and single $\mathcal{T}$ operator (no stage index).

Fitted Q-Iteration (Discounted)

Starting from arbitrary $f_0 \in \mathcal{F}$, iterate for $t = 1, 2, \ldots, K$:

FQI iteration (AJKS eq. 4.1)

$$f_t \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \left(f(s_i, a_i) - r_i - \gamma \max_{a'} f_{t-1}(s_i', a') \right)^2.$$

Output greedy policy $\pi^K(s) := \arg\max_a f_K(s, a)$.

Fitted Q-Iteration (Discounted)

Starting from arbitrary $f_0 \in \mathcal{F}$, iterate for $t = 1, 2, \ldots, K$:

FQI iteration (AJKS eq. 4.1)

$$f_t \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \left(f(s_i, a_i) - r_i - \gamma \max_{a'} f_{t-1}(s_i', a') \right)^2.$$

Output greedy policy $\pi^K(s) := \arg\max_a f_K(s, a)$.

Difference from finite-horizon.

- No backward sweep — the operator $\mathcal{T}$ is stationary, so we can iterate it on a single function class.
- Iteration count K is a tuning parameter. More iterations $\Rightarrow$ smaller γ^K tail, but also a larger union bound penalty in the regression lemma.

Why Iterate? Why Not Minimize $\|f - \mathcal{T}f\|^2$ Directly?

In the discounted setting, $Q^\star$ is a single function satisfying $Q^\star = \mathcal{T}Q^\star$. A natural alternative to FQI is to look for any $f \in \mathcal{F}$ such that $f \approx \mathcal{T}f$ — i.e., minimize the empirical Bellman residual:

$$L_n(f) := \frac{1}{n}\sum_{i=1}^{n}\Big(f(s_i, a_i) - r_i - \gamma \max_{a'} f(s_i', a')\Big)^2.$$

(This question doesn't arise in finite horizon: each f_h is targeting the next-stage iterate f_{h+1}, not itself.)

Why Iterate? Why Not Minimize $\|f - \mathcal{T}f\|^2$ Directly?

In the discounted setting, $Q^\star$ is a single function satisfying $Q^\star = \mathcal{T}Q^\star$. A natural alternative to FQI is to look for any $f \in \mathcal{F}$ such that $f \approx \mathcal{T}f$ — i.e., minimize the empirical Bellman residual:

$$L_n(f) \;:=\; \frac{1}{n}\sum_{i=1}^{n}\Big(f(s_i,a_i) - r_i - \gamma \max_{a'} f(s_i',a')\Big)^2.$$

(This question doesn't arise in finite horizon: each f_h is targeting the next-stage iterate f_{h+1}, not itself.)

Compute the expectation. Let $Y(s,a) := r + \gamma \max_{a'} f(s',a')$, so $\mathbb{E}[Y \mid s,a] = (\mathcal{T}f)(s,a)$:

$$\mathbb{E}[L_n(f)] \;=\; \underbrace{\|f - \mathcal{T}f\|_{2,\nu}^2}_{\text{what we want}} + \underbrace{\mathbb{E}_\nu\big[\mathrm{Var}(Y \mid s,a)\big]}_{\text{depends on } f\,!}.$$

Why Iterate? Why Not Minimize $\|f - \mathcal{T}f\|^2$ Directly?

In the discounted setting, $Q^\star$ is a single function satisfying $Q^\star = \mathcal{T}Q^\star$. A natural alternative to FQI is to look for any $f \in \mathcal{F}$ such that $f \approx \mathcal{T}f$ — i.e., minimize the empirical Bellman residual:

$$L_n(f) \;:=\; \frac{1}{n}\sum_{i=1}^{n}\Big(f(s_i,a_i) - r_i - \gamma \max_{a'} f(s_i',a')\Big)^2.$$

(This question doesn't arise in finite horizon: each f_h is targeting the next-stage iterate f_{h+1}, not itself.)

Compute the expectation. Let $Y(s,a) := r + \gamma \max_{a'} f(s',a')$, so $\mathbb{E}[Y \mid s,a] = (\mathcal{T}f)(s,a)$:

$$\mathbb{E}[L_n(f)] \;=\; \underbrace{\|f - \mathcal{T}f\|_{2,\nu}^2}_{\text{what we want}} + \underbrace{\mathbb{E}_\nu\big[\mathrm{Var}(Y \mid s,a)\big]}_{\text{depends on } f\,!}.$$

The double-sampling problem (Baird 1995; Antos–Szepesvári–Munos 2008)

The conditional variance is generally non-zero (stochastic transitions) and *itself a function of* f. Minimizing L_n biases toward f's with low $\mathrm{Var}(Y \mid s,a)$ — "smooth" or "flat" f — unrelated to the Bellman fixed point. Debiasing would require *two* independent draws of s' per (s,a) — almost never available offline.

How FQI Sidesteps the Problem

Key move: use the *previous* iterate f_{t-1} for the target, not f. FQI minimizes

$$\tilde{L}_n(f; f_{t-1}) := \frac{1}{n} \sum_i \Big(f(s_i, a_i) - \underbrace{[r_i + \gamma \max_{a'} f_{t-1}(s_i', a')]}_{Y_i\text{: function of } f_{t-1}\text{, not } f} \Big)^2.$$

Since Y_i is independent of the variable f:

$$\mathbb{E}[\tilde{L}_n(f; f_{t-1})] \;=\; \|f - \mathcal{T} f_{t-1}\|_{2,\nu}^2 + \underbrace{\mathbb{E}_\nu[\mathrm{Var}(Y \mid s, a)]}_{\text{constant in } f}.$$

The variance term is an additive constant — it drops out of the optimization.

How FQI Sidesteps the Problem

Key move: use the *previous* iterate f_{t-1} for the target, not f. FQI minimizes

$$\tilde{L}_n(f; f_{t-1}) := \frac{1}{n} \sum_i \Big(f(s_i, a_i) - \underbrace{[r_i + \gamma \max_{a'} f_{t-1}(s_i', a')]}_{Y_i\text{: function of } f_{t-1}\text{, not } f} \Big)^2.$$

Since Y_i is independent of the variable f:

$$\mathbb{E}[\tilde{L}_n(f; f_{t-1})] \;=\; \|f - \mathcal{T} f_{t-1}\|_{2,\nu}^2 + \underbrace{\mathbb{E}_\nu[\mathrm{Var}(Y \mid s, a)]}_{\text{constant in } f}.$$

The variance term is an additive constant — it drops out of the optimization.

What we gained

- Each step is a clean regression with a realizable target $\mathcal{T} f_{t-1}$.
- No double sampling required — single transitions (s, a, r, s') suffice.
- Each step is a vanilla least-squares problem.

The price: an outer loop over t, with error propagation controlled by Lemma 4.4 (next).

Another Fix: Modified BRM (Antos–Szepesvári–Munos 2008)

Recall the obstacle: $\mathbb{E}[L_n(f)] = \|f - \mathcal{T}f\|_{2,\nu}^2 + \mathbb{E}_\nu[\mathrm{Var}(Y \mid s, a)]$, with the variance term depending on f.

Idea (Antos et al.): explicitly *subtract off* the bias. For each candidate f, define the best in-class fit to the bootstrap target:

$$h_f \;:=\; \arg\min_{h \in \mathcal{F}} \; \frac{1}{n} \sum_{i=1}^{n} \big(h(s_i, a_i) - r_i - \gamma \max_{a'} f(s_i', a')\big)^2.$$

Then define the modified empirical Bellman residual

$$\tilde{L}_n(f) \;:=\; \frac{1}{n} \sum_{i=1}^{n} \Big[\big(f(s_i, a_i) - Y_i\big)^2 \;-\; \big(h_f(s_i, a_i) - Y_i\big)^2\Big].$$

where $Y_i := r_i + \gamma \max_{a'} f(s_i', a')$.

Another Fix: Modified BRM (Antos–Szepesvári–Munos 2008)

Recall the obstacle: $\mathbb{E}[L_n(f)] = \|f - \mathcal{T}f\|_{2,\nu}^2 + \mathbb{E}_\nu[\mathrm{Var}(Y \mid s, a)]$, with the variance term depending on f.

Idea (Antos et al.): explicitly *subtract off* the bias. For each candidate f, define the best in-class fit to the bootstrap target:

$$h_f \;:=\; \arg\min_{h \in \mathcal{F}} \; \frac{1}{n} \sum_{i=1}^{n} \big(h(s_i, a_i) - r_i - \gamma \max_{a'} f(s_i', a')\big)^2.$$

Then define the modified empirical Bellman residual

$$\tilde{L}_n(f) \;:=\; \frac{1}{n} \sum_{i=1}^{n} \Big[\big(f(s_i, a_i) - Y_i\big)^2 \;-\; \big(h_f(s_i, a_i) - Y_i\big)^2\Big].$$

where $Y_i := r_i + \gamma \max_{a'} f(s_i', a')$.

Why this is (essentially) unbiased. Each squared loss has the same conditional variance term, so it cancels:

$$\mathbb{E}[\tilde{L}_n(f)] \;=\; \|f - \mathcal{T}f\|_{2,\nu}^2 \;-\; \|h_f - \mathcal{T}f\|_{2,\nu}^2 .$$

The subtracted term is $\leq \min_{h \in \mathcal{F}} \|h - \mathcal{T}f\|_{2,\nu}^2 \leq \varepsilon_{\mathrm{approx},\nu}$ — the inherent Bellman error. So minimizing $\tilde{L}_n$ approximately minimizes $\|f - \mathcal{T}f\|_{2,\nu}^2$, up to an irreducible $\varepsilon_{\mathrm{approx},\nu}$ floor, with *no* variance bias.

Modified BRM: Algorithm and Guarantees

Modified BRM

$\hat{f} := \arg\min_{f \in \mathcal{F}} \tilde{L}_n(f); \qquad \widehat{\pi}(s) := \arg\max_a \hat{f}(s, a).$

Modified BRM: Algorithm and Guarantees

Modified BRM

$\hat{f} := \arg\min_{f \in \mathcal{F}} \tilde{L}_n(f); \qquad \widehat{\pi}(s) := \arg\max_a \hat{f}(s, a).$

Guarantee (Antos–Szepesvári–Munos 2008, informal). Under concentrability C and inherent Bellman error $\varepsilon_{\text{approx},\nu}$, with probability $\geq 1 - \delta$,

$$V^\star - V^{\widehat{\pi}} \lesssim \frac{\sqrt{C}}{(1-\gamma)^2} \left[\sqrt{V_{\max}^2 \log |\mathcal{F}|/n} + \sqrt{\varepsilon_{\text{approx},\nu}} \right].$$

Same $1/(1-\gamma)^2$ propagation, same dependence on $C, \varepsilon_{\text{approx},\nu}, \log|\mathcal{F}|, n$ as FQI.

Modified BRM: Algorithm and Guarantees

Modified BRM

$\hat{f} := \arg\min_{f \in \mathcal{F}} \tilde{L}_n(f); \qquad \hat{\pi}(s) := \arg\max_a \hat{f}(s,a).$

Guarantee (Antos–Szepesvári–Munos 2008, informal). Under concentrability C and inherent Bellman error $\varepsilon_{\text{approx},\nu}$, with probability $\geq 1-\delta$,

$$V^\star - V^{\hat{\pi}} \lesssim \frac{\sqrt{C}}{(1-\gamma)^2}\left[\sqrt{V_{\max}^2 \log|\mathcal{F}|/n} + \sqrt{\varepsilon_{\text{approx},\nu}}\right].$$

Same $1/(1-\gamma)^2$ propagation, same dependence on $C, \varepsilon_{\text{approx},\nu}, \log|\mathcal{F}|, n$ as FQI.

Trade-off vs. FQI

	FQI	**Modified BRM**
Optimization	K standard LSQ steps	one *nested* min over f, h_f
Iteration tail	$\gamma^K V_{\max}$ ($K \sim \log(1/\varepsilon)$)	none — one-shot
Convexity	convex per step	non-convex (saddle/bilevel)

Take-away. Two principled fixes for the double-sampling problem: *iterate* (FQI) or *debias* (Antos et al.). Statistically similar; computationally quite different.

Lemma 4.4: Error Propagation (Discounted)

Lemma (Approximate VI bound)

Suppose $\|f_{t+1} - \mathcal{T} f_t\|_{2,\nu} \leq \varepsilon$ for all $t = 0, 1, \ldots, K-1$. Let π^t be greedy w.r.t. f_t. Then for every $k \in \{0, 1, \ldots, K\}$:

$$V^\star - V^{\pi^k} \leq \frac{2\sqrt{C}\,\varepsilon}{(1-\gamma)^2} + \frac{2\gamma^k V_{\max}}{1-\gamma}.$$

Lemma 4.4: Error Propagation (Discounted)

Lemma (Approximate VI bound)

Suppose $\|f_{t+1} - \mathcal{T} f_t\|_{2,\nu} \leq \varepsilon$ for all $t = 0, 1, \ldots, K-1$. Let π^t be greedy w.r.t. f_t. Then for every $k \in \{0, 1, \ldots, K\}$:

$$V^\star - V^{\pi^k} \leq \frac{2\sqrt{C}\,\varepsilon}{(1-\gamma)^2} + \frac{2\gamma^k V_{\max}}{1-\gamma}.$$

Factors.

- $\sqrt{C}$: concentrability transfers regression error from ν to $d_\mu^{\pi^t}$.
- $1/(1-\gamma)^2$: one factor from per-step Bellman residual to value error; another from summing geometric weights γ^{k-1-t}.
- $2\gamma^k V_{\max}/(1-\gamma)$: new term — truncation after K iterations. Decays geometrically.

Lemma 4.4: Error Propagation (Discounted)

Lemma (Approximate VI bound)

Suppose $\|f_{t+1} - \mathcal{T} f_t\|_{2,\nu} \le \varepsilon$ for all $t = 0, 1, \ldots, K-1$. Let π^t be greedy w.r.t. f_t. Then for every $k \in \{0, 1, \ldots, K\}$:

$$V^\star - V^{\pi^k} \le \frac{2\sqrt{C}\,\varepsilon}{(1-\gamma)^2} + \frac{2\gamma^k V_{\max}}{1-\gamma}.$$

Factors.

- $\sqrt{C}$: concentrability transfers regression error from ν to $d_\mu^{\pi^t}$.
- $1/(1-\gamma)^2$: one factor from per-step Bellman residual to value error; another from summing geometric weights γ^{k-1-t}.
- $2\gamma^k V_{\max}/(1-\gamma)$: new term — truncation after K iterations. Decays geometrically.

Reference. Munos (2005); Antos, Szepesvári, Munos (2008).

Lemma 4.4: Proof Idea

The finite-horizon proof (Lemma A) ran two telescopes: one under $d_h^{\pi^\star}$ for $V_0^\star - \widehat{V}_0$, one under $d_h^{\widehat{\pi}}$ for $\widehat{V}_0 - V_0^{\widehat{\pi}}$. The infinite-horizon proof has the same skeleton, but the telescoping is now over the *iteration index* t rather than the stage index h:

- Track pointwise inequalities for $h_t := Q^\star - f_t$ and $g_t := f_t - Q^{\pi^t}$ under operator products $P^{\pi^\star} \cdots$ and $P^{\pi^t} \cdots$.
- Use $\mathcal{T}^{\pi^\star} g \leq \mathcal{T} g$ and $\mathcal{T}^{\pi^t} f_t = \mathcal{T} f_t$ to chain the two policies' transition operators together.
- Iterating K steps replaces "sum over h" with "sum over t with geometric weights γ^{K-1-t}."

Lemma 4.4: Proof Idea

The finite-horizon proof (Lemma A) ran two telescopes: one under $d_h^{\pi^\star}$ for $V_0^\star - \widehat{V}_0$, one under $d_h^{\widehat{\pi}}$ for $\widehat{V}_0 - V_0^{\widehat{\pi}}$. The infinite-horizon proof has the same skeleton, but the telescoping is now over the *iteration index* t rather than the stage index h:

- Track pointwise inequalities for $h_t := Q^\star - f_t$ and $g_t := f_t - Q^{\pi^t}$ under operator products $P^{\pi^\star} \cdots$ and $P^{\pi^t} \cdots$.
- Use $\mathcal{T}^{\pi^\star} g \le \mathcal{T} g$ and $\mathcal{T}^{\pi^t} f_t = \mathcal{T} f_t$ to chain the two policies' transition operators together.
- Iterating K steps replaces "sum over h" with "sum over t with geometric weights γ^{K-1-t}."

From distribution to $L_2(\nu)$. Operator products of $P^{\pi^\star}, P^{\pi^t}$ act on the pivot distribution μ to produce d^π-type occupancies. Taking L_1 at the end and applying $\|g\|_{1,d_\mu^\pi} \le \|g\|_{2,d_\mu^\pi} \le \sqrt{C}\,\|g\|_{2,\nu}$ (Cauchy–Schwarz + concentrability) gives the bound.

Lemma 4.4: Proof Idea

The finite-horizon proof (Lemma A) ran two telescopes: one under $d_h^{\pi^\star}$ for $V_0^\star - \widehat{V}_0$, one under $d_h^{\widehat{\pi}}$ for $\widehat{V}_0 - V_0^{\widehat{\pi}}$. The infinite-horizon proof has the same skeleton, but the telescoping is now over the *iteration index* t rather than the stage index h:

- Track pointwise inequalities for $h_t := Q^\star - f_t$ and $g_t := f_t - Q^{\pi^t}$ under operator products $P^{\pi^\star} \cdots$ and $P^{\pi^t} \cdots$.
- Use $\mathcal{T}^{\pi^\star} g \le \mathcal{T} g$ and $\mathcal{T}^{\pi^t} f_t = \mathcal{T} f_t$ to chain the two policies' transition operators together.
- Iterating K steps replaces "sum over h" with "sum over t with geometric weights γ^{K-1-t}."

From distribution to $L_2(\nu)$. Operator products of $P^{\pi^\star}, P^{\pi^t}$ act on the pivot distribution μ to produce d^π-type occupancies. Taking L_1 at the end and applying $\|g\|_{1,d_\mu^\pi} \le \|g\|_{2,d_\mu^\pi} \le \sqrt{C}\,\|g\|_{2,\nu}$ (Cauchy–Schwarz + concentrability) gives the bound.

Two new features versus finite horizon.

- The initial-error term $\|h_0\| \le V_{\max}$ becomes the $\gamma^K V_{\max}/(1-\gamma)$ *iteration tail*.
- The geometric weights $\sum_t \gamma^{K-1-t} \le 1/(1-\gamma)$ contribute a second factor of $1/(1-\gamma)$ on top of the per-step $1/(1-\gamma)$ from converting Bellman residual to value error. Hence $1/(1-\gamma)^2$.

Full proof: Munos (2005, 2007); Antos–Szepesvári–Munos (2008). □

Lemma 4.5: Uniform Regression Bound (Discounted)

Lemma (Per-iteration regression error)

For a finite class $\mathcal{F}$, with probability $\geq 1-\delta$, simultaneously for all $t = 0, \ldots, K-1$:

$$\|f_{t+1} - \mathcal{T} f_t\|_{2,\nu}^2 \;\leq\; \frac{20\, V_{\max}^2 \ln(2|\mathcal{F}|^2 K/\delta)}{n} \;+\; 3\,\varepsilon_{\text{approx},\nu}.$$

Lemma 4.5: Uniform Regression Bound (Discounted)

Lemma (Per-iteration regression error)

For a finite class $\mathcal{F}$, with probability $\geq 1-\delta$, simultaneously for all $t = 0, \ldots, K-1$:

$$\|f_{t+1} - \mathcal{T} f_t\|_{2,\nu}^2 \;\leq\; \frac{20\, V_{\max}^2 \ln(2|\mathcal{F}|^2 K/\delta)}{n} \;+\; 3\,\varepsilon_{\text{approx},\nu}.$$

Proof. Same as the finite-horizon regression bound (Lemma B), with two changes:

- Bounded range becomes $[0, V_{\max}]$ instead of $[0, H]$.
- Union bound now over $g \in \mathcal{F}$ *and* $t = 0, \ldots, K-1$ (i.e. $\delta \to \delta/(|\mathcal{F}|K)$).

The extra $\ln K$ factor comes from the iteration loop — not present in finite-horizon FQI.

Theorem 4.3: Main Performance Guarantee

Theorem (FQI, infinite horizon)

Fix $K \in \mathbb{N}_+$. Under Assumptions 4.1, 4.2, for a finite class $\mathcal{F}$, FQI satisfies: with probability $\geq 1-\delta$,

$$V^\star - V^{\pi^K} \leq \frac{2\sqrt{C}}{(1-\gamma)^2}\left[\sqrt{\frac{20\, V_{\max}^2 \ln(2|\mathcal{F}|^2 K/\delta)}{n}} + \sqrt{3\,\varepsilon_{\text{approx},\nu}}\right] + \frac{2\gamma^K V_{\max}}{1-\gamma}.$$

Theorem 4.3: Main Performance Guarantee

Theorem (FQI, infinite horizon)

Fix $K \in \mathbb{N}_+$. Under Assumptions 4.1, 4.2, for a finite class $\mathcal{F}$, FQI satisfies: with probability $\geq 1-\delta$,

$$V^\star - V^{\pi^K} \leq \frac{2\sqrt{C}}{(1-\gamma)^2}\left[\sqrt{\frac{20\, V_{\max}^2 \ln(2|\mathcal{F}|^2 K/\delta)}{n}} + \sqrt{3\,\varepsilon_{\text{approx},\nu}}\right] + \frac{2\gamma^K V_{\max}}{1-\gamma}.$$

Proof. Combine Lemma 4.5 (take ε as the per-step regression error) and Lemma 4.4 at $k = K$. □

Theorem 4.3: Main Performance Guarantee

Theorem (FQI, infinite horizon)

Fix $K \in \mathbb{N}_+$. Under Assumptions 4.1, 4.2, for a finite class $\mathcal{F}$, FQI satisfies: with probability $\geq 1-\delta$,

$$V^\star - V^{\pi^K} \leq \frac{2\sqrt{C}}{(1-\gamma)^2}\left[\sqrt{\frac{20\, V_{\max}^2 \ln(2|\mathcal{F}|^2 K/\delta)}{n}} + \sqrt{3\,\varepsilon_{\text{approx},\nu}}\right] + \frac{2\gamma^K V_{\max}}{1-\gamma}.$$

Proof. Combine Lemma 4.5 (take ε as the per-step regression error) and Lemma 4.4 at $k = K$. □

Three sources of error

1. **Statistical:** $\sqrt{V_{\max}^2 \ln|\mathcal{F}|/n}/(1-\gamma)^2$. Vanishes as $n \to \infty$.
2. **Approximation:** $\sqrt{C\,\varepsilon_{\text{approx},\nu}}/(1-\gamma)^2$. Irreducible model floor.
3. **Iteration:** $2\gamma^K V_{\max}/(1-\gamma)$. Vanishes as $K \to \infty$. New vs. finite horizon.

Choose $K = \Theta(\log(V_{\max}/\varepsilon)/(1-\gamma))$ to make the iteration term $O(\varepsilon)$.

Reading the Bound

Coverage price ($\sqrt{C}$). Regression accuracy under ν must transfer to occupancies of every policy we might compare against.

Approximation price ($\sqrt{\varepsilon_{\text{approx},\nu}}$). Even with infinite data and infinite iterations, this remains. Vanishes iff $\mathcal{F}$ is $L_2(\nu)$-Bellman complete.

Iteration price (γ^K). Choose $K = \Theta(\log(V_{\max}/\varepsilon)/(1-\gamma))$ to make it $O(\varepsilon)$.

Reading the Bound

Coverage price ($\sqrt{C}$). Regression accuracy under ν must transfer to occupancies of every policy we might compare against.

Approximation price ($\sqrt{\varepsilon_{\text{approx},\nu}}$). Even with infinite data and infinite iterations, this remains. Vanishes iff $\mathcal{F}$ is $L_2(\nu)$-Bellman complete.

Iteration price (γ^K). Choose $K = \Theta(\log(V_{\max}/\varepsilon)/(1-\gamma))$ to make it $O(\varepsilon)$.

Sample complexity (when $\varepsilon_{\text{approx},\nu} = 0$)

To achieve $V^\star - V^{\pi^K} \leq \varepsilon$, it suffices to take
$n = \widetilde{O}\left(\frac{C\, V_{\max}^2 \ln|\mathcal{F}|}{(1-\gamma)^4\, \varepsilon^2}\right)$, $K = \widetilde{O}\left(\frac{\log(V_{\max}/\varepsilon)}{1-\gamma}\right)$.

Reading the Bound

Coverage price ($\sqrt{C}$). Regression accuracy under ν must transfer to occupancies of every policy we might compare against.

Approximation price ($\sqrt{\varepsilon_{\text{approx},\nu}}$). Even with infinite data and infinite iterations, this remains. Vanishes iff $\mathcal{F}$ is $L_2(\nu)$-Bellman complete.

Iteration price (γ^K). Choose $K = \Theta(\log(V_{\max}/\varepsilon)/(1-\gamma))$ to make it $O(\varepsilon)$.

Sample complexity (when $\varepsilon_{\text{approx},\nu} = 0$)

To achieve $V^\star - V^{\pi^K} \le \varepsilon$, it suffices to take
$n = \widetilde{O}\left(\frac{C\, V_{\max}^2 \ln|\mathcal{F}|}{(1-\gamma)^4\, \varepsilon^2}\right)$, $K = \widetilde{O}\left(\frac{\log(V_{\max}/\varepsilon)}{1-\gamma}\right)$.

Realizability still not enough. Just $Q^\star \in \mathcal{F}$ won't suffice — Lemma 4.5 needs $\mathcal{T} f$ approximately in $\mathcal{F}$ for *every* $f \in \mathcal{F}$, not just for $f = Q^\star$.

Side by Side

	Finite horizon	**Infinite-horizon (Thm 4.3)**
Effective horizon	H	$1/(1-\gamma) = V_{\max}$
# regressions	H (forced)	K (chosen)
Per-step value range	$[0, H]$	$[0, V_{\max}]$
Coverage term	$\sqrt{C}$	$\sqrt{C}$
Statistical	$H^2\sqrt{C \log \lvert\mathcal{F}\rvert / n}$	$\frac{\sqrt{C} V_{\max} \sqrt{\log \lvert\mathcal{F}\rvert}}{(1-\gamma)^2 \sqrt{n}}$
Approximation	$H\sqrt{C\, \varepsilon_{\mathrm{approx},\nu}}$	$\frac{\sqrt{C \varepsilon_{\mathrm{approx},\nu}}}{(1-\gamma)^2}$
Iteration error	—	$\gamma^K V_{\max}/(1-\gamma)$

Side by Side

	Finite horizon	**Infinite-horizon (Thm 4.3)**
Effective horizon	H	$1/(1-\gamma) = V_{\max}$
# regressions	H (forced)	K (chosen)
Per-step value range	$[0, H]$	$[0, V_{\max}]$
Coverage term	$\sqrt{C}$	$\sqrt{C}$
Statistical	$H^2\sqrt{C \log \lvert\mathcal{F}\rvert / n}$	$\frac{\sqrt{C} V_{\max} \sqrt{\log \lvert\mathcal{F}\rvert}}{(1-\gamma)^2 \sqrt{n}}$
Approximation	$H\sqrt{C\, \varepsilon_{\mathrm{approx},\nu}}$	$\frac{\sqrt{C \varepsilon_{\mathrm{approx},\nu}}}{(1-\gamma)^2}$
Iteration error	—	$\gamma^K V_{\max}/(1-\gamma)$

Identifying $H \sim 1/(1-\gamma)$: the two bounds agree up to constants, except infinite horizon has the extra γ^K tail that vanishes geometrically.

Summary

What we did today

- Defined Fitted Q-Iteration: solve RL as a sequence of regression problems.
- Two assumptions, no structural condition on P, r:
  - Concentrability (C): data covers what every policy visits.
  - Inherent Bellman error ($\varepsilon_{\text{approx},\nu}$): $\mathcal{F}$ is approximately closed under $\mathcal{T}$.
- **Finite horizon** (Lemmas A, B): $V_0^\star - V_0^{\widehat{\pi}} \leq H\sqrt{C}(H\sqrt{\log|\mathcal{F}|/n} + \sqrt{\varepsilon_{\text{approx},\nu}})$.
- **Infinite horizon** (Thm 4.3 = Lemmas 4.4 + 4.5): adds a $\gamma^K V_{\max}$ tail.
- Double-sampling problem motivates why FQI *iterates* rather than minimizing $\|f - \mathcal{T}f\|^2$ directly.

Summary

What we did today

- Defined Fitted Q-Iteration: solve RL as a sequence of regression problems.
- Two assumptions, no structural condition on P, r:
  - Concentrability (C): data covers what every policy visits.
  - Inherent Bellman error ($\varepsilon_{\text{approx},\nu}$): $\mathcal{F}$ is approximately closed under $\mathcal{T}$.
- **Finite horizon** (Lemmas A, B): $V_0^\star - V_0^{\widehat{\pi}} \leq H\sqrt{C}(H\sqrt{\log|\mathcal{F}|/n} + \sqrt{\varepsilon_{\text{approx},\nu}})$.
- **Infinite horizon** (Thm 4.3 = Lemmas 4.4 + 4.5): adds a $\gamma^K V_{\max}$ tail.
- Double-sampling problem motivates why FQI *iterates* rather than minimizing $\|f - \mathcal{T}f\|^2$ directly.

Compared to Lectures 12–13

Linear FA with (C) was the special case where $\varepsilon_{\text{approx},\nu} = 0$ exactly (strict Bellman completeness). FQI gives a noisier but far more flexible analysis: any $\mathcal{F}$, paying for misspecification with $\sqrt{\varepsilon_{\text{approx},\nu}}$.

Reading & References

Primary reference:

AJKS Agarwal, Jiang, Kakade, Sun, *Reinforcement Learning: Theory and Algorithms*, Chapter 4.

Foundational papers:

- Munos, *Error bounds for approximate value iteration*, AAAI 2005.
- Antos, Szepesvári, Munos, *Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration*, MLJ 2008.
- Szepesvári, Munos, *Finite time bounds for fitted value iteration*, JMLR 2005.
- Lazaric, Ghavamzadeh, Munos, *Analysis of classification-based policy iteration*, JMLR 2016.

Modern offline RL with concentrability:

- Chen, Jiang, *Information-theoretic considerations in batch RL*, ICML 2019.

Questions?