CSE 542: Statistical Reinforcement Learning

Lecture 3: Offline Reinforcement Learning

Kevin Jamieson

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

Today's Plan

The Data-Driven World

Modern decision-making is awash in **historical data**:

- **Healthcare:** years of electronic health records documenting treatments and outcomes
- **Autonomous driving:** millions of miles of logged human driving behavior
- **Recommender systems:** logs of user interactions with content
- **Finance:** decades of market microstructure data
- **Robotics:** teleoperated demonstrations and prior deployments

A natural question

Can we extract a *good policy* from this data alone, **without further interaction**?

Why Not Just Roll Out More?

Online RL (upcoming lectures) learns by interacting with the environment. It works well — but interaction can be:

- **Dangerous:** an ICU dosing agent that explores could harm patients
- **Expensive:** a real robot arm exploring costs time and hardware
- **Irreversible:** bad decisions in financial markets cannot be undone
- **Impossible:** the environment from which data was collected no longer exists

Why Not Just Roll Out More?

Online RL (upcoming lectures) learns by interacting with the environment. It works well — but interaction can be:

- **Dangerous:** an ICU dosing agent that explores could harm patients
- **Expensive:** a real robot arm exploring costs time and hardware
- **Irreversible:** bad decisions in financial markets cannot be undone
- **Impossible:** the environment from which data was collected no longer exists

Offline (Batch) RL

Learn a policy from a fixed dataset $\mathcal{D}$ of historical transitions.
No further interaction allowed.

The Core Challenge: Distribution Shift

Offline RL is fundamentally harder than the generative model (Lecture 2).

Generative Model (Lec. 2)

Query *any* (s, a) on demand.
Data coverage is **our choice**.
Sample complexity: $\frac{|\mathcal{S}||\mathcal{A}|H^3}{\varepsilon^2}$.

Offline RL

Data fixed by a **behavior policy**.
Some (s, a) pairs may be **unobserved**.
Cannot query missing transitions.

The Core Challenge: Distribution Shift

Offline RL is fundamentally harder than the generative model (Lecture 2).

Generative Model (Lec. 2)

Query *any* (s, a) on demand.
Data coverage is **our choice**.
Sample complexity: $\frac{|\mathcal{S}||\mathcal{A}|H^3}{\varepsilon^2}$.

Offline RL

Data fixed by a **behavior policy**.
Some (s, a) pairs may be **unobserved**.
Cannot query missing transitions.

The danger: a policy that deviates from data support enters regions where the empirical model $\widehat{P}$ is unreliable — potentially catastrophically wrong.

Main message of this lecture

Two algorithms for offline RL:

1. **Plug-in MLE** — natural but suboptimal; pays for uncertainty under *both* π^* and $\widehat{\pi}$.
2. **Pessimistic Value Iteration** — optimal; pays only for uncertainty under π^*.

Today's Plan

Finite-Horizon MDP

Definition

A finite-horizon MDP is $(\mathcal{S}, \mathcal{A}, \{P_h\}_h, \{r_h\}_h, H, \mu)$ where

- $\mathcal{S}$: finite state space, $S = |\mathcal{S}|$; $\mathcal{A}$: finite action space, $A = |\mathcal{A}|$
- $P_h : \mathcal{S} \times \mathcal{A} \to \triangle_{\mathcal{S}}$: transition kernel at stage h
- $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$: reward function at stage h
- $H \in \mathbb{N}$: horizon; $\mu \in \triangle_{\mathcal{S}}$: initial state distribution

A policy $\pi = \{\pi_h\}_h$ prescribes actions at each stage. An episode generates a trajectory $\tau = (s_1, a_1, r_1, s_2, \ldots, s_H, a_H, r_H)$ where $s_1 \sim \mu$, $a_h = \pi_h(s_h)$, $s_{h+1} \sim P_h(\cdot|s_h, a_h)$.

Objective: maximize expected cumulative reward $V_0^\pi = \mathbb{E}_\pi \left[\sum_{h=1}^{H} r_h(s_h, a_h) \right]$.

Value Functions

State value function

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \,\middle|\, \pi,\ s_h = s\right]$$

State-action value (Q) function

$$Q_h^\pi(s, a) = \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \,\middle|\, \pi,\ s_h = s,\ a_h = a\right]$$

Value Functions

State value function

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \,\middle|\, \pi,\ s_h = s\right]$$

State-action value (Q) function

$$Q_h^\pi(s, a) = \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \,\middle|\, \pi,\ s_h = s,\ a_h = a\right]$$

Key facts:

- $V_h^\pi(s) \in [0, H - h + 1]$ and $Q_h^\pi(s, a) \in [0, H - h + 1]$ $\quad V_H(s) = 0$
- $V_h^\pi(s) = Q_h^\pi(s, \pi_h(s))$
- **Bellman backup:** $Q_h^\pi(s, a) = r_h(s, a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)}\left[V_{h+1}^\pi(s')\right]$

We write $V_0^\pi = \mathbb{E}_{s_1 \sim \mu}[V_1^\pi(s_1)]$ for the overall value.

Optimal Policy and Value Iteration

Bellman Optimality Equations

Define $Q_h^*(s,a) = \sup_\pi Q_h^\pi(s,a)$. Then $Q_h^* = Q_h^\star$ for all h iff

$$Q_h(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)}\Big[\max_{a'} Q_{h+1}(s',a')\Big], \quad Q_{H+1} \equiv 0.$$

The greedy policy $\pi_h^*(s) = \arg\max_a Q_h^*(s,a)$ is optimal.

Value Iteration (planning algorithm)

1. Set $Q_H(s,a) = r_H(s,a)$ for all (s,a).
2. For $h = H-1, \ldots, 1$: $\quad Q_h(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)}\Big[\max_{a'} Q_{h+1}(s',a')\Big]$.

Value iteration is **exact** when P_h is known. Today: P_h is **unknown** and we only have a fixed offline dataset.

Hoeffding's Inequality

The central concentration inequality we will use throughout.

Theorem (Hoeffding's Inequality)

Let $X_1, \ldots, X_N$ be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $a \leq X_i \leq b$ almost surely. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$\left| \frac{1}{N} \sum_{i=1}^{N} X_i \; - \; \mu \right| \; \leq \; (b - a) \sqrt{\frac{\log(2/\delta)}{2N}}.$$

Hoeffding's Inequality

The central concentration inequality we will use throughout.

Theorem (Hoeffding's Inequality)

Let $X_1, \ldots, X_N$ be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $a \leq X_i \leq b$ almost surely. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$\left| \frac{1}{N} \sum_{i=1}^{N} X_i \; - \; \mu \right| \; \leq \; (b - a) \sqrt{\frac{\log(2/\delta)}{2N}}.$$

Key features:

- Deviation scales as $\frac{b-a}{\sqrt{N}}$ — the *range* of X_i divided by $\sqrt{N}$.
- No assumption on the distribution beyond bounded support.
- **Union bound over K events:** replace $\log(2/\delta)$ with $\log(2K/\delta)$.

Hoeffding Applied to Transition Estimation

Fix a function $f : \mathcal{S} \to [0, H]$. Given $n_h(s, a)$ i.i.d. transitions from $P_h(\cdot|s, a)$:

$$\left|(\widehat{P}_h - P_h)f\right|(s, a) = \left|\frac{1}{n_h(s, a)} \sum_{i=1}^{n_h(s,a)} f(s'_i) - \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[f(s')]\right|.$$

Hoeffding Applied to Transition Estimation

Fix a function $f : \mathcal{S} \to [0, H]$. Given $n_h(s, a)$ i.i.d. transitions from $P_h(\cdot|s, a)$:

$$\left|(\widehat{P}_h - P_h)f\right|(s, a) = \left|\frac{1}{n_h(s, a)} \sum_{i=1}^{n_h(s,a)} f(s'_i) - \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[f(s')]\right|.$$

Since $f(s') \in [0, H]$, Hoeffding gives with probability $\geq 1 - \delta$:

$$\left|(\widehat{P}_h - P_h)f\right|(s, a) \;\leq\; H\sqrt{\frac{\log(2/\delta)}{2\, n_h(s, a)}}.$$

Hoeffding Applied to Transition Estimation

Fix a function $f : \mathcal{S} \to [0, H]$. Given $n_h(s, a)$ i.i.d. transitions from $P_h(\cdot|s, a)$:

$$\left|(\widehat{P}_h - P_h)f\right|(s, a) = \left|\frac{1}{n_h(s, a)} \sum_{i=1}^{n_h(s,a)} f(s'_i) - \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[f(s')]\right|.$$

Since $f(s') \in [0, H]$, Hoeffding gives with probability $\geq 1 - \delta$:

$$\left|(\widehat{P}_h - P_h)f\right|(s, a) \;\leq\; H\sqrt{\frac{\log(2/\delta)}{2\, n_h(s, a)}}.$$

Union bounding over all (h, s, a) triples (at most SAH of them):

$$\left|(\widehat{P}_h - P_h)f\right|(s, a) \;\leq\; H\sqrt{\frac{\log(2SAH/\delta)}{2\, n_h(s, a)}} \quad \text{simultaneously for all } (h, s, a).$$

$:= \iota$

This is the engine behind almost every result today.

Today's Plan

The Offline Data Model

We are given a fixed dataset of historical transitions.

Offline data assumption

For each stage $h \in \{1, \ldots, H-1\}$ and pair $(s, a) \in \mathcal{S} \times \mathcal{A}$:

- We observe $n_h(s,a) \geq 1$ i.i.d. samples $s' \sim P_h(\cdot|s,a)$.
- Sample sets are **independent** across triples (h, s, a).

The empirical kernel is $\widehat{P}_h(\cdot|s,a) = \frac{1}{n_h(s,a)} \sum_i \mathbf{1}\{s'_i = \cdot\}$.

Remarks:

- The counts $\{n_h(s,a)\}$ are fixed (determined by a behavior policy); we cannot choose them.
- This stylized model highlights the role of **data coverage**: pairs with small $n_h(s,a)$ are unreliable.
- We also assume deterministic rewards $r_h(s,a)$ are known (or can be estimated trivially).

Uncertainty Bonus

Define the **log factor** and **uncertainty bonus**:

$$\iota \triangleq \log\left(\frac{8SAH}{\delta}\right), \qquad b_h(s,a) \triangleq H\sqrt{\frac{\iota}{2\,n_h(s,a)}}.$$

Interpretation: $b_h(s,a)$ is the Hoeffding confidence interval width for transition-weighted values at (s,a,h), holding simultaneously over all pairs after a union bound.

Uncertainty Bonus

Define the **log factor** and **uncertainty bonus**:

$$\iota \triangleq \log\left(\frac{8SAH}{\delta}\right), \qquad b_h(s,a) \triangleq H\sqrt{\frac{\iota}{2\,n_h(s,a)}}.$$

Interpretation: $b_h(s,a)$ is the Hoeffding confidence interval width for transition-weighted values at (s,a,h), holding simultaneously over all pairs after a union bound.

The "Good Event" (Lemma)

With probability at least $1-\delta$, simultaneously for every $h \in \{1,\dots,H-1\}$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$:

$$\left|(\widehat{P}_h - P_h)V_{h+1}^{\pi^*}\right|(s,a) \le b_h(s,a), \qquad \left|(\widehat{P}_h - P_h)V_{h+1}^{\widehat{\pi}}\right|(s,a) \le b_h(s,a),$$

$$\left|(\widehat{P}_h - P_h)\Delta_{h+1}\right|(s,a) \le b_h(s,a), \qquad \left|(\widehat{P}_h - P_h)\Gamma_{h+1}\right|(s,a) \le b_h(s,a),$$

where $\Delta_h \triangleq V_h^* - \widehat{V}_h$ and $\Gamma_h \triangleq \widehat{V}_h - V_h^{\widehat{\pi}}$.

Why Four Inequalities?

The good event controls four error quantities simultaneously:

1. $|(\widehat{P}_h - P_h)V^{\pi^*}_{h+1}|$ — error in estimating transitions weighted by **optimal values**
2. $|(\widehat{P}_h - P_h)V^{\widehat{\pi}}_{h+1}|$ — error weighted by **learned policy's values**
3. $|(\widehat{P}_h - P_h)\Delta_{h+1}|$ — error in the gap $V^* - \widehat{V}$ (used in MLE proof)
4. $|(\widehat{P}_h - P_h)\Gamma_{h+1}|$ — error in the gap $\widehat{V} - V^{\widehat{\pi}}$ (used in MLE proof)

Why Four Inequalities?

The good event controls four error quantities simultaneously:

1. $|(\widehat{P}_h - P_h)V_{h+1}^{\pi^*}|$ — error in estimating transitions weighted by **optimal values**
2. $|(\widehat{P}_h - P_h)V_{h+1}^{\widehat{\pi}}|$ — error weighted by **learned policy's values**
3. $|(\widehat{P}_h - P_h)\Delta_{h+1}|$ — error in the gap $V^* - \widehat{V}$ (used in MLE proof)
4. $|(\widehat{P}_h - P_h)\Gamma_{h+1}|$ — error in the gap $\widehat{V} - V^{\widehat{\pi}}$ (used in MLE proof)

Why can we union bound over all four?
Each function depends only on empirical kernels from *future* stages $h+1, \ldots, H-1$. After conditioning on that future-stage data, each function is **fixed and independent** of the samples at stage h — enabling Hoeffding.

There are at most $4SAH$ such bounds; a union bound at level $\delta/(4SAH)$ each concludes the proof.

Occupancy Measures

A key quantity in the bounds: how often does policy π visit (s, a) at stage t?

Occupancy measure

$w_t^\pi(s, a) \triangleq \Pr_\pi[s_t = s,\ a_t = a]$ (starting from $s_1 \sim \mu$).

Properties:

- $\sum_{s,a} w_t^\pi(s, a) = 1$ for each t (it is a distribution over (s, a) at stage t).
- The bound we will derive depends on $\sum_{t,s,a} w_t^{\pi^*}(s, a)\, b_t(s, a)$.
- If π^* concentrates on well-covered pairs (large n_t), this sum is small.

Occupancy Measures

A key quantity in the bounds: how often does policy π visit (s, a) at stage t?

Occupancy measure

$w_t^\pi(s,a) \triangleq \Pr_\pi[s_t = s,\ a_t = a]$ (starting from $s_1 \sim \mu$).

$w_1^\pi(s,a) = \mu(s) \cdot \mathbb{1}\{\pi(s) = a\}$

$\mathbb{E}_{s\sim\mu}[V_0^\pi(s)] = \mathbb{E}_\pi\left[\sum_{t=0}^{H} r(s_t,a_t) \mid s_1 \sim \mu\right] = \sum_{t=1}^{H}\sum_{s,a} r(s,a)\, w_t^\pi(s,a)$

Properties:

- $\sum_{s,a} w_t^\pi(s,a) = 1$ for each t (it is a distribution over (s,a) at stage t).
- The bound we will derive depends on $\sum_{t,s,a} w_t^{\pi^*}(s,a)\, b_t(s,a)$.
- If π^* concentrates on well-covered pairs (large n_t), this sum is small.

Key intuition

We can only do as well as our data *covers the optimal policy's trajectory*. Pairs visited by π^* but unseen in data are the fundamental bottleneck.

Today's Plan

Natural Algorithm: Plug-in MLE

The most natural approach: **estimate P_h, then plan**.

Plug-in MLE algorithm

1. Build empirical transitions: $\widehat{P}_h(s'|s,a) = \dfrac{\#\{s' \text{ observed at } (s,a,h)\}}{n_h(s,a)}$.
2. Run value iteration on the **empirical MDP** $\widehat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \{\widehat{P}_h\}, \{r_h\}, H, \mu)$.
3. Output: $\widehat{\pi} = \{\widehat{\pi}_h\}$ where $\widehat{\pi}_h(s) = \arg\max_a \widehat{Q}_h(s,a)$.

Natural Algorithm: Plug-in MLE

The most natural approach: **estimate P_h, then plan**.

Plug-in MLE algorithm

1. Build empirical transitions: $\widehat{P}_h(s'|s,a) = \dfrac{\#\{s' \text{ observed at } (s,a,h)\}}{n_h(s,a)}$.
2. Run value iteration on the **empirical MDP** $\widehat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \{\widehat{P}_h\}, \{r_h\}, H, \mu)$.
3. Output: $\widehat{\pi} = \{\widehat{\pi}_h\}$ where $\widehat{\pi}_h(s) = \arg\max_a \widehat{Q}_h(s,a)$.

Notation: $\widehat{V}_h^{\pi}$ is the value of policy π in $\widehat{\mathcal{M}}$ (using $\widehat{P}$). We write $\widehat{V}_h = \widehat{V}_h^{\widehat{\pi}}$ for the value of the MLE policy in its own MDP.

Question: how suboptimal is $\widehat{\pi}$ in the *true* MDP?

$$V_0^* - V_0^{\widehat{\pi}} = ?$$

MLE Bound: Main Result

Theorem (MLE bound)

On the good event, for every stage h and state s:

$$V_h^*(s) - V_h^{\widehat{\pi}}(s) \le 2\sum_{t=h}^{H-1}\sum_{s',a'} w_t^{\pi^*}(s',a')\, b_t(s',a') + 2\sum_{t=h}^{H-1}\sum_{s',a'} w_t^{\widehat{\pi}}(s',a')\, b_t(s',a').$$

$\le \max_{\pi}$

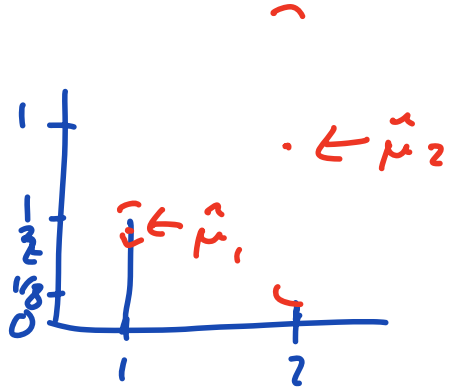
MLE $\operatorname{argmax}_i \hat{\mu}_i = 2$ $\quad \Delta_h \quad +$ $\quad \Gamma_h$

Pessimism $\operatorname{argmax}_i \hat{\mu}_i - \frac{1}{\sqrt{n_i}} = 1$

$\mu_1 = \frac{1}{2}$, $\mu_2 = \frac{1}{8}$ Bernoulli arms

$n_1 = 10000$, $n_2 = 1$

$\hat{\mu}_1 \approx \frac{1}{2}$ $\quad \hat{\mu}_2 = 1$



MLE Bound: Main Result

Theorem (MLE bound)

On the good event, for every stage h and state s:

$$V_h^*(s) - V_h^{\widehat{\pi}}(s) \;\le\; 2\sum_{t=h}^{H-1}\sum_{s',a'} w_t^{\pi^*}(s',a')\, b_t(s',a') \;+\; 2\sum_{t=h}^{H-1}\sum_{s',a'} w_t^{\widehat{\pi}}(s',a')\, b_t(s',a').$$

Two terms, two occupancies:

- First sum: uncertainty under the **optimal policy** π^* trajectory.
- Second sum: uncertainty under the **learned policy** $\widehat{\pi}$ trajectory.

MLE Bound: Main Result

Theorem (MLE bound)

On the good event, for every stage h and state s:

$$V_h^*(s) - V_h^{\widehat{\pi}}(s) \;\le\; 2\sum_{t=h}^{H-1}\sum_{s',a'} w_t^{\pi^*}(s',a')\, b_t(s',a') \;+\; 2\sum_{t=h}^{H-1}\sum_{s',a'} w_t^{\widehat{\pi}}(s',a')\, b_t(s',a').$$

Two terms, two occupancies:

- First sum: uncertainty under the **optimal policy** π^* trajectory.
- Second sum: uncertainty under the **learned policy** $\widehat{\pi}$ trajectory.

The problem with MLE

The bound depends on where $\widehat{\pi}$ goes — but $\widehat{\pi}$ could stray into poorly-covered regions where $b_t(s,a)$ is large. The second term is hard to control without strong coverage assumptions on *all* policies.

Decomposing the Sub-optimality

Write $\widehat{V}_h = \widehat{V}_h^{\widehat{\pi}}$ and decompose:

$$V_h^* - V_h^{\widehat{\pi}} = \underbrace{(V_h^* - \widehat{V}_h)}_{\Delta_h} + \underbrace{(\widehat{V}_h - V_h^{\widehat{\pi}})}_{\Gamma_h}.$$

Decomposing the Sub-optimality

Write $\widehat{V}_h = \widehat{V}_h^{\widehat{\pi}}$ and decompose:

$$V_h^* - V_h^{\widehat{\pi}} = \underbrace{(V_h^* - \widehat{V}_h)}_{\Delta_h} + \underbrace{(\widehat{V}_h - V_h^{\widehat{\pi}})}_{\Gamma_h}.$$

$\Delta_h = V_h^* - \widehat{V}_h$

How much does the empirical MDP *underestimate* the optimal value? (Optimism of $\widehat{V}$ would make this negative.)

$\Gamma_h = \widehat{V}_h - V_h^{\widehat{\pi}}$

How much does $\widehat{\pi}$'s empirical value *overestimate* its true value? (Evaluation error.)

We bound each term separately by unrolling Bellman recursions along the respective policy trajectories.

Controlling Δ_h: Proof Sketch

Let $a^* = \pi_h^*(s)$. Since $\widehat{V}_h(s) = \max_a \{r_h(s,a) + (\widehat{P}_h \widehat{V}_{h+1})(s,a)\}$:

$$\Delta_h(s) \leq r_h(s,a^*) + (P_h V_{h+1}^*)(s,a^*) - r_h(s,a^*) - (\widehat{P}_h \widehat{V}_{h+1})(s,a^*).$$

$$\Delta_h = V^* - \hat{V} = r_h(s,a^*) + P_h V_{h+1}^*(s,a^*) - \underbrace{\left(r_h(s,\hat{\pi}_h(s)) + \hat{P}_h \hat{V}_{h+1}(s,\hat{\pi}(s))\right)}$$

$$= \max_a r_h(s,a) + \hat{P}_h \hat{V}_{h+1}(s,a)$$

$$\geq r_h(s,a_h) + \hat{P} \hat{V}_{h+1}(s,a_h)$$

$$\hat{\pi}_h(s) = \arg\max_a \hat{Q}_h(s,a)$$

$$\hat{Q}_h(s,a) = r_h(s,a) + (\widehat{P}_h \widehat{V}_{h+1})(s,a)$$

Controlling Δ_h: Proof Sketch

Let $a^* = \pi_h^*(s)$. Since $\widehat{V}_h(s) = \max_a\{r_h(s,a) + (\widehat{P}_h \widehat{V}_{h+1})(s,a)\}$:

$$\Delta_h(s) \;\leq\; r_h(s,a^*) + (P_h V_{h+1}^*)(s,a^*) - r_h(s,a^*) - (\widehat{P}_h \widehat{V}_{h+1})(s,a^*).$$

Add and subtract $(\widehat{P}_h V_{h+1}^*)(s,a^*)$:

$$\Delta_h(s) \;\leq\; \underbrace{(P_h - \widehat{P}_h)V_{h+1}^*}_{|\cdot| \leq b_h(s,a^*)}(s,a^*) + (\widehat{P}_h \Delta_{h+1})(s,a^*).$$

$\widehat{P}_h \Delta_{h+1}(s,a) = V_{h+1}^{\pi_h} - \widehat{V}_{h+1}$

Add and subtract $(P_h \Delta_{h+1})(s,a^*)$:

$$\Delta_h(s) \;\leq\; (P_h \Delta_{h+1})(s,\pi_h^*(s)) + 2b_h(s,\pi_h^*(s)).$$

$(\widehat{P}_h - P_h)\Delta_{h+1}(s,a) \leq b_h(s,a)$

$$(P_h \Delta_{h+1})(s,\pi_h) = \mathbb{E}\left[\Delta_{h+1}(S_{h+1}) \mid S_h = s, a_h = \pi_h(s_h)\right]$$

$S_{h+1} \sim P(\cdot \mid s_h, \pi_h(s_h))$

$$\leq \mathbb{E}\left[(P_{h+1}\Delta_{h+2})(S_{h+1}, \pi_{h+1}(S_{h+1})) + 2b_{h+1}(S_{h+1}, \pi_{h+1}(S_{h+1})) \,\middle|\, S_h = s, a_h = \pi\right] + 2b_h(s,\pi_h(s))$$

$$= \mathbb{E}\left[\Delta_{h+2}(S_{h+2}) \mid S_{h+1}, \pi_{h+1}(S_{h+1}), S_h = s \right] + \mathbb{E}\left[2b_{h+1}(S_{h+1}, \pi_{h+1}(S_{h+1})) \mid S_h = s, a_h = \pi_h(s) \right]$$

$$+ 2b_h(s, \pi_h(s)) \qquad = \sum_{s',a'} b_{h+1}(s',a') \mathbb{P}^{\pi}(S_{h+1} = s', a_{h+1} = a' \mid S_h = s)$$

$\Delta_H = 0$

$$\leq \sum_{t=h}^{H} \mathbb{E}_{\pi}\left[2b_t(S_t, \pi_t(S_t)) \mid S_h = s \right]$$

Controlling Δ_h: Proof Sketch

Let $a^* = \pi_h^*(s)$. Since $\widehat{V}_h(s) = \max_a\{r_h(s,a) + (\widehat{P}_h \widehat{V}_{h+1})(s,a)\}$:

$$\Delta_h(s) \;\leq\; r_h(s,a^*) + (P_h V_{h+1}^*)(s,a^*) - r_h(s,a^*) - (\widehat{P}_h \widehat{V}_{h+1})(s,a^*).$$

Add and subtract $(\widehat{P}_h V_{h+1}^*)(s,a^*)$:

$$\Delta_h(s) \;\leq\; \underbrace{(P_h - \widehat{P}_h) V_{h+1}^*}_{|\cdot| \leq b_h(s,a^*)}(s,a^*) + (\widehat{P}_h \Delta_{h+1})(s,a^*).$$

Add and subtract $(P_h \Delta_{h+1})(s,a^*)$:

$$\Delta_h(s) \;\leq\; (P_h \Delta_{h+1})(s,\pi_h^*(s)) + 2b_h(s,\pi_h^*(s)).$$

Unroll along π^*'s trajectory (using $\Delta_{H+1} \equiv 0$):

$$\Delta_h(s) \;\leq\; 2\sum_{t=h}^{H-1} \sum_{s',a'} w_t^{\pi^*}(s',a')\, b_t(s',a').$$

Controlling Γ_h: Proof Sketch

Since $\widehat{\pi}$ is greedy for $\widehat{\mathcal{M}}$ and $V_h^{\widehat{\pi}}$ satisfies the true Bellman equation:

$$\widehat{V}_h(s) = r_h(s, \widehat{\pi}_h(s)) + (\widehat{P}_h \widehat{V}_{h+1})(s, \widehat{\pi}_h(s)),$$
$$V_h^{\widehat{\pi}}(s) = r_h(s, \widehat{\pi}_h(s)) + (P_h V_{h+1}^{\widehat{\pi}})(s, \widehat{\pi}_h(s)).$$

Subtracting:

$$\Gamma_h(s) = \underbrace{(\widehat{P}_h - P_h) V_{h+1}^{\widehat{\pi}}}_{|\cdot| \leq b_h(s, \widehat{\pi}_h(s))}(s, \widehat{\pi}_h(s)) + (\widehat{P}_h \Gamma_{h+1})(s, \widehat{\pi}_h(s)).$$

Controlling Γ_h: Proof Sketch

Since $\widehat{\pi}$ is greedy for $\widehat{\mathcal{M}}$ and $V_h^{\widehat{\pi}}$ satisfies the true Bellman equation:

$$\widehat{V}_h(s) = r_h(s, \widehat{\pi}_h(s)) + (\widehat{P}_h \widehat{V}_{h+1})(s, \widehat{\pi}_h(s)),$$
$$V_h^{\widehat{\pi}}(s) = r_h(s, \widehat{\pi}_h(s)) + (P_h V_{h+1}^{\widehat{\pi}})(s, \widehat{\pi}_h(s)).$$

Subtracting:

$$\Gamma_h(s) = \underbrace{(\widehat{P}_h - P_h) V_{h+1}^{\widehat{\pi}}}_{|\cdot| \le b_h(s, \widehat{\pi}_h(s))}(s, \widehat{\pi}_h(s)) + (\widehat{P}_h \Gamma_{h+1})(s, \widehat{\pi}_h(s)).$$

Insert and subtract $(P_h \Gamma_{h+1})(s, \widehat{\pi}_h(s))$:

$$\Gamma_h(s) \;\le\; (P_h \Gamma_{h+1})(s, \widehat{\pi}_h(s)) + 2 b_h(s, \widehat{\pi}_h(s)).$$

Unroll along $\widehat{\pi}$'s trajectory:

$$\Gamma_h(s) \;\le\; 2 \sum_{t=h}^{H-1} \sum_{s', a'} w_t^{\widehat{\pi}}(s', a')\, b_t(s', a').$$

Adding the bounds for Δ_h and Γ_h proves the theorem.

Today's Plan

Why MLE Fails: The Exploitation Trap

The MLE policy maximizes value in the **empirical MDP**, ignoring estimation uncertainty.

Illustrative failure

Suppose state s_{rare} is visited very rarely in the data (small $n_h(s_{\text{rare}}, a)$ for all a).

- $\widehat{P}_h(\cdot|s_{\text{rare}}, a)$ is noisy — could erroneously suggest high reward.
- The MLE policy may route through s_{rare} to exploit this "phantom" high reward.
- In truth, the policy gets poor reward and the bound's second term (under $\widehat{\pi}$) is large.

Why MLE Fails: The Exploitation Trap

The MLE policy maximizes value in the **empirical MDP**, ignoring estimation uncertainty.

Illustrative failure

Suppose state s_{rare} is visited very rarely in the data (small $n_h(s_{\text{rare}}, a)$ for all a).

- $\widehat{P}_h(\cdot|s_{\text{rare}}, a)$ is noisy — could erroneously suggest high reward.
- The MLE policy may route through s_{rare} to exploit this "phantom" high reward.
- In truth, the policy gets poor reward and the bound's second term (under $\widehat{\pi}$) is large.

The fix: penalize transitions into poorly-covered regions.
An agent should be **pessimistic** about poorly-covered state-action pairs — it should avoid them, not exploit imaginary rewards there.

This mirrors the **optimism** principle from online RL, but in the opposite direction.

Pessimistic Value Iteration (PEVI)

Algorithm: Pessimistic Value Iteration

Initialize $\underline{V}_{H+1} \equiv 0$. Backward for $h = H, H-1, \ldots, 1$:

$$\underline{Q}_h(s,a) \triangleq r_h(s,a) + (\widehat{P}_h \underline{V}_{h+1})(s,a) \; - \; b_h(s,a),$$

$$\underline{V}_h(s) \triangleq \max_{a \in \mathcal{A}} \underline{Q}_h(s,a).$$

Output: $\widetilde{\pi}_h(s) = \arg\max_a \underline{Q}_h(s,a)$ (greedy with respect to $\underline{V}$).

Pessimistic Value Iteration (PEVI)

Algorithm: Pessimistic Value Iteration

Initialize $\underline{V}_{H+1} \equiv 0$. Backward for $h = H, H-1, \ldots, 1$:

$$\underline{Q}_h(s,a) \triangleq r_h(s,a) + (\widehat{P}_h \underline{V}_{h+1})(s,a) \; - \; b_h(s,a),$$

$$\underline{V}_h(s) \triangleq \max_{a \in \mathcal{A}} \underline{Q}_h(s,a).$$

Output: $\widetilde{\pi}_h(s) = \arg\max_a \underline{Q}_h(s,a)$ (greedy with respect to $\underline{V}$).

Key difference from MLE: subtract the bonus $b_h(s,a)$ at each backup.

- High-uncertainty pairs (small $n_h(s,a)$, large b_h) are **penalized**.
- The policy is discouraged from visiting regions not well-covered by data.
- Think of $\underline{Q}_h(s,a)$ as a **lower confidence bound** on the true $Q_h^*(s,a)$.

Key Lemma: Sandwiching

Lemma (Pessimistic sandwiching)

On the good event, for all h and s:

$$\underline{V}_h(s) \leq V_h^{\widetilde{\pi}}(s) \leq V_h^*(s).$$

Key Lemma: Sandwiching

Lemma (Pessimistic sandwiching)

On the good event, for all h and s:

$$\underline{V}_h(s) \;\leq\; V_h^{\widetilde{\pi}}(s) \;\leq\; V_h^*(s).$$

Proof: The upper bound ($V_h^{\widetilde{\pi}} \leq V_h^*$) is immediate (no policy beats optimal).

For the lower bound, argue **backward induction**. Trivially $\underline{V}_{H+1} \leq V_{H+1}^{\widetilde{\pi}}$.
Suppose $\underline{V}_{h+1} \leq V_{h+1}^{\widetilde{\pi}}$. By greediness of $\widetilde{\pi}$:

$$\underline{V}_h(s) = r_h(s, \widetilde{\pi}_h(s)) + (\widehat{P}_h \underline{V}_{h+1})(s, \widetilde{\pi}_h(s)) - b_h(s, \widetilde{\pi}_h(s)).$$

On the good event:

$$(\widehat{P}_h \underline{V}_{h+1})(s, \widetilde{\pi}_h(s)) - b_h(s, \widetilde{\pi}_h(s)) \;\leq\; (P_h \underline{V}_{h+1})(s, \widetilde{\pi}_h(s)) \;\leq\; (P_h V_{h+1}^{\widetilde{\pi}})(s, \widetilde{\pi}_h(s)).$$

Therefore $\underline{V}_h(s) \leq V_h^{\widetilde{\pi}}(s)$. □

Pessimism Bound: Main Result

Theorem (Pessimism bound)

On the good event, for every stage h and state s:

$$V_h^*(s) - V_h^{\widetilde{\pi}}(s) \;\leq\; 2 \sum_{t=h}^{H-1} \sum_{s',a'} w_t^{\pi^*}(s',a')\, b_t(s',a').$$

Consequently, the same bound holds with probability at least $1-\delta$.

Pessimism Bound: Main Result

Theorem (Pessimism bound)

On the good event, for every stage h and state s:

$$V_h^*(s) - V_h^{\widetilde{\pi}}(s) \;\le\; 2 \sum_{t=h}^{H-1} \sum_{s',a'} w_t^{\pi^*}(s',a')\, b_t(s',a').$$

Consequently, the same bound holds with probability at least $1-\delta$.

Crucial improvement over MLE

Only one occupancy term — under π^*, not $\widetilde{\pi}$.
The second term (evaluation error under the learned policy) has **vanished**.

Pessimism Bound: Proof

By the sandwiching lemma: $V_h^* - V_h^{\widetilde{\pi}} \leq V_h^* - \underline{V}_h =: \Delta_h$.

Let $a^* = \pi_h^*(s)$. Since $\underline{V}_h(s) \geq \underline{Q}_h(s, a^*)$:

$$\Delta_h(s) \;\leq\; r_h(s, a^*) + (P_h V_{h+1}^*)(s, a^*) - r_h(s, a^*) - (\widehat{P}_h \underline{V}_{h+1})(s, a^*) + b_h(s, a^*).$$

Pessimism Bound: Proof

By the sandwiching lemma: $V_h^* - V_h^{\widetilde{\pi}} \leq V_h^* - \underline{V}_h =: \Delta_h$.

Let $a^* = \pi_h^*(s)$. Since $\underline{V}_h(s) \geq \underline{Q}_h(s, a^*)$:

$$\Delta_h(s) \;\leq\; r_h(s, a^*) + (P_h V_{h+1}^*)(s, a^*) - r_h(s, a^*) - (\widehat{P}_h \underline{V}_{h+1})(s, a^*) + b_h(s, a^*).$$

Add and subtract $(P_h \underline{V}_{h+1})(s, a^*)$:

$$\Delta_h(s) \;\leq\; (P_h \Delta_{h+1})(s, a^*) + \underbrace{(P_h - \widehat{P}_h)\underline{V}_{h+1}}_{|\cdot| \leq b_h(s,a^*)}(s, a^*) + b_h(s, a^*).$$

On the good event, the middle term is at most $b_h(s, a^*)$, hence:

$$\Delta_h(s) \;\leq\; (P_h \Delta_{h+1})(s, \pi_h^*(s)) + 2b_h(s, \pi_h^*(s)).$$

Pessimism Bound: Proof

By the sandwiching lemma: $V_h^* - V_h^{\widetilde{\pi}} \leq V_h^* - \underline{V}_h =: \Delta_h$.

Let $a^* = \pi_h^*(s)$. Since $\underline{V}_h(s) \geq \underline{Q}_h(s, a^*)$:

$$\Delta_h(s) \;\leq\; r_h(s, a^*) + (P_h V_{h+1}^*)(s, a^*) - r_h(s, a^*) - (\widehat{P}_h \underline{V}_{h+1})(s, a^*) + b_h(s, a^*).$$

Add and subtract $(P_h \underline{V}_{h+1})(s, a^*)$:

$$\Delta_h(s) \;\leq\; (P_h \Delta_{h+1})(s, a^*) + \underbrace{(P_h - \widehat{P}_h)\underline{V}_{h+1}}_{|\cdot| \leq b_h(s,a^*)}(s, a^*) + b_h(s, a^*).$$

On the good event, the middle term is at most $b_h(s, a^*)$, hence:

$$\Delta_h(s) \;\leq\; (P_h \Delta_{h+1})(s, \pi_h^*(s)) + 2b_h(s, \pi_h^*(s)).$$

Unroll along π^*'s trajectory (using $\Delta_{H+1} \equiv 0$):

$$\Delta_h(s) \;\leq\; 2 \sum_{t=h}^{H-1} \sum_{s', a'} w_t^{\pi^*}(s', a')\, b_t(s', a'). \qquad \square$$

Side-by-Side Comparison

MLE bound

$$V_h^* - V_h^{\widehat{\pi}} \leq 2 \sum_{t,s',a'} w_t^{\pi^*}(s',a')\, b_t(s',a') + 2 \sum_{t,s',a'} w_t^{\widehat{\pi}}(s',a')\, b_t(s',a').$$

Side-by-Side Comparison

MLE bound

$$V_h^* - V_h^{\widehat{\pi}} \leq 2 \sum_{t,s',a'} w_t^{\pi^*}(s',a')\, b_t(s',a') + 2 \sum_{t,s',a'} w_t^{\widehat{\pi}}(s',a')\, b_t(s',a').$$

Pessimism bound

$$V_h^* - V_h^{\widetilde{\pi}} \leq 2 \sum_{t,s',a'} w_t^{\pi^*}(s',a')\, b_t(s',a').$$

Side-by-Side Comparison

MLE bound

$$V_h^* - V_h^{\widehat{\pi}} \leq 2 \sum_{t,s',a'} w_t^{\pi^*}(s',a')\, b_t(s',a') + 2 \sum_{t,s',a'} w_t^{\widehat{\pi}}(s',a')\, b_t(s',a').$$

Pessimism bound

$$V_h^* - V_h^{\widetilde{\pi}} \leq 2 \sum_{t,s',a'} w_t^{\pi^*}(s',a')\, b_t(s',a').$$

Both bounds share the first term. Pessimism **eliminates the second term**.

- The second MLE term can be *arbitrarily larger* than the first if $\widehat{\pi}$ explores poorly-covered regions.
- Pessimism automatically steers $\widetilde{\pi}$ away from such regions, so no second term arises.

Why Does Pessimism Win?

The real benefit of pessimism is not that concentration becomes easier.

Both algorithms face the same statistical difficulty in estimating transitions. The difference is **structural**:

Key insight

The lower bound $\underline{V}_h \leq V_h^{\widetilde{\pi}}$ (sandwiching lemma) prevents the evaluation-error term $(\widehat{V}_h - V_h^{\widetilde{\pi}})$ from appearing in the analysis.
We only need to track how V^* decomposes along π^*'s trajectory — a trajectory we can analyze using concentration on V^*, which is **fixed** (not data-dependent).

Why Does Pessimism Win?

The real benefit of pessimism is not that concentration becomes easier.

Both algorithms face the same statistical difficulty in estimating transitions. The difference is **structural**:

Key insight

The lower bound $\underline{V}_h \leq V_h^{\widetilde{\pi}}$ (sandwiching lemma) prevents the evaluation-error term $(\widehat{V}_h - V_h^{\widetilde{\pi}})$ from appearing in the analysis.
We only need to track how V^* decomposes along π^*'s trajectory — a trajectory we can analyze using concentration on V^*, which is **fixed** (not data-dependent).

Contrast with MLE: to bound $(\widehat{V}_h - V_h^{\widehat{\pi}})$, we must unroll along $\widehat{\pi}$'s trajectory, which itself depends on the data — creating the problematic second occupancy term.

Explicit Sample Complexity

Substituting $b_t(s,a) = H\sqrt{\iota/(2n_t(s,a))}$ into the pessimism bound and assuming **uniform coverage** $n_t(s,a) = n$ for all (t,s,a) visited by π^*:

$$V_0^* - V_0^{\widetilde{\pi}} \;\leq\; 2\sum_{t=1}^{H-1}\sum_{s,a} w_t^{\pi^*}(s,a)\, H\sqrt{\frac{\iota}{2n}} \;\leq\; 2H^2\sqrt{\frac{\iota}{2n}}.$$

Setting this $\leq \varepsilon$ and solving for n:

$$n \;\gtrsim\; \frac{H^4 \log(SAH/\delta)}{\varepsilon^2}.$$

Explicit Sample Complexity

Substituting $b_t(s,a) = H\sqrt{\iota/(2n_t(s,a))}$ into the pessimism bound and assuming **uniform coverage** $n_t(s,a) = n$ for all (t,s,a) visited by π^*:

$$V_0^* - V_0^{\widetilde{\pi}} \;\le\; 2\sum_{t=1}^{H-1}\sum_{s,a} w_t^{\pi^*}(s,a)\, H\sqrt{\frac{\iota}{2n}} \;\le\; 2H^2\sqrt{\frac{\iota}{2n}}.$$

Setting this $\le \varepsilon$ and solving for n:

$$n \;\gtrsim\; \frac{H^4 \log(SAH/\delta)}{\varepsilon^2}.$$

Sample complexity (offline RL, pessimistic policy)

Under uniform coverage with n samples per (s,a,h) in π^*'s support:

$$\text{total samples} \;\lesssim\; \frac{H^4 \cdot |\mathrm{supp}(\pi^*)| \cdot \log(SAH/\delta)}{\varepsilon^2},$$

where $|\mathrm{supp}(\pi^*)|$ counts the state-action pairs reachable under π^*.

Looking Ahead

Next lecture: Online RL and Exploration

In offline RL, coverage is *given* — we either have it or we don't.

In online RL, we **choose** our data collection policy, adapting based on what we see.

- The agent must balance **exploration** (reducing uncertainty) and **exploitation** (acting well now).
- The right lens: **regret** — total sub-optimality over K episodes.
- Key algorithm: **UCB-VI** — optimistic (not pessimistic!) value iteration.

Looking Ahead

Next lecture: Online RL and Exploration

In offline RL, coverage is *given* — we either have it or we don't.

In online RL, we **choose** our data collection policy, adapting based on what we see.

- The agent must balance **exploration** (reducing uncertainty) and **exploitation** (acting well now).
- The right lens: **regret** — total sub-optimality over K episodes.
- Key algorithm: **UCB-VI** — optimistic (not pessimistic!) value iteration.

Optimism vs. Pessimism

- **Offline RL** $\Rightarrow$ **Pessimism**: penalize uncertain actions to stay in-distribution.
- **Online RL** $\Rightarrow$ **Optimism**: bonus uncertain actions to incentivize exploration.