Lecture 12: Offline RL

This lecture studies *offline reinforcement learning* (also known as *batch RL*), where the agent must learn a good policy from a fixed dataset collected a priori, without any further interaction with the environment. The central principle is *pessimism in the face of uncertainty*—the mirror image of the optimism principle studied in Lecture 11. We cover three main ideas: the suboptimality decomposition that explains why pessimism is needed, *Pessimistic Value Iteration* (PEVI) for *linear MDPs* (Jin et al., 2021), and *Bellman-consistent pessimism* (BCP) for general function classes (Xie et al., 2021). The PEVI portion of the lecture remains in the *finite-horizon episodic* setting of Lecture 11, while the BCP portion later switches to a *discounted infinite-horizon* setting following Xie et al. (2021).

Setup. For the PEVI part of the lecture, we continue with the *finite-horizon episodic MDP* $M = (\mathcal{S}, \mathcal{A}, H, \{P_h^\star\}, \{r_h\}, s_0)$ from Lecture 11, with state space $\mathcal{S}$, action space $\mathcal{A}$, horizon H, transition kernel $P_h^\star(\cdot \mid s, a)$, deterministic reward $r_h(s, a) \in [0, 1]$, and fixed initial state s_0. Value functions and Q-functions are defined as before. For any Q-function $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, write $V_g(s) \coloneqq \max_{a \in \mathcal{A}} g(s, a)$. In the PEVI sections, we use the Bellman optimality operator on next-step Q-functions:

$$(\mathcal{T}_h g)(s, a) \;\coloneqq\; r_h(s, a) + (P_h^\star V_g)(s, a),$$

where $(P_h^\star u)(s, a) = \mathbb{E}_{s' \sim P_h^\star(\cdot|s,a)}[u(s')]$ for any state-value function $u : \mathcal{S} \to \mathbb{R}$.

Finite-horizon Bellman equations (used in the PEVI sections): For any deterministic policy π,

$$Q_h^\pi(s, a) = r_h(s, a) + P_h^\star V_{h+1}^\pi(s, a), \qquad V_h^\pi(s) = Q_h^\pi(s, \pi_h(s)), \qquad V_H^\pi \equiv 0.$$

At the optimal policy $\pi^\star$, these specialize to the *Bellman optimality equations*:

$$Q_h^\star(s, a) = r_h(s, a) + P_h^\star V_{h+1}^\star(s, a) = (\mathcal{T}_h Q_{h+1}^\star)(s, a), \qquad V_h^\star(s) = \max_{a \in \mathcal{A}} Q_h^\star(s, a), \qquad V_H^\star \equiv 0.$$

The performance metric in offline RL is the *suboptimality* of a learned policy π:

$$\mathrm{SubOpt}(\pi; s_0) \;\coloneqq\; V_0^\star(s_0) - V_0^\pi(s_0).$$

The Offline RL Problem

Setting and Data Collecting Process

In the offline (or batch) setting, the learner has access only to a pre-collected dataset $\mathcal{D}$—there is no ability to interact with the environment or collect new data. This is in stark contrast to the online exploration problem of Lecture 11, where the agent adaptively collects trajectories over K episodes.

Definition 1 (Offline dataset)**.** *The offline dataset consists of K trajectories:*

$$\mathcal{D} \;=\; \left\{(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau)\right\}_{\tau=0,\,h=0}^{K-1,\,H-1},$$

where for each trajectory $\tau \in \{0, \ldots, K-1\}$, the experimenter takes action a_h^τ at state s_h^τ, observes reward $r_h^\tau = r_h(s_h^\tau, a_h^\tau)$, and the next state is sampled $s_{h+1}^\tau \sim P_h^\star(\cdot \mid s_h^\tau, a_h^\tau)$. The actions a_h^τ are chosen by a (possibly adaptive) experimenter and are not *under the learner's control.*

The key assumption on the dataset is *compliance*: the conditional distribution of (r_h^τ, s_{h+1}^τ) given the current and past data is determined by the true MDP.

Assumption 1 (Compliance)**.** *The dataset $\mathcal{D}$ is* compliant *with the underlying MDP $(\mathcal{S}, \mathcal{A}, H, P^\star, r)$: for each trajectory τ and step h, conditioning on (s_h^τ, a_h^τ) and all data from earlier trajectories and earlier steps within trajectory τ, the tuple (r_h^τ, s_{h+1}^τ) is generated by the reward function r_h and transition kernel $P_h^\star$ of the underlying MDP.*

Generality of compliance. Assumption 1 is deliberately permissive: the actions a_h^τ can be chosen by any mechanism—a fixed behavior policy, an adaptive strategy, or even an adversary. The K trajectories need not be independent or identically distributed. All that is required is that the *transitions and rewards* follow the true MDP. This captures diverse practical scenarios: data pooled from multiple operators, data collected by an evolving policy, or data from a simulator with scripted behaviors.

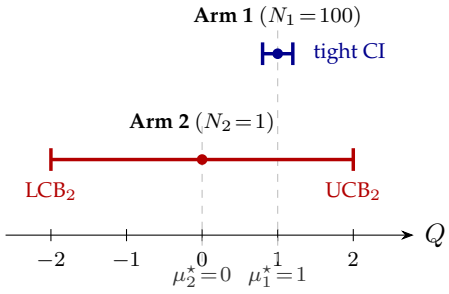
Online vs. Offline: From Optimism to Pessimism

In Lecture 11, we saw that the online exploration problem is solved by the *optimism* principle: the agent inflates uncertain Q-values with bonuses, which drives it to explore under-visited states. The key structural feature of the analysis is:

$$\textit{Online regret} \;\le\; \textit{sum of bonuses along the } \textbf{\textit{learned policy's}} \textit{ trajectory.}$$

This works because the agent controls which states to visit, and optimism ensures it visits informative ones.

In the offline setting, the agent has *no control* over the data. A natural question is whether we can simply run an online optimistic algorithm (e.g., UCBVI or Lin-UCBVI) on the offline dataset. Figure 1 shows why the answer is *no*: optimism's bonus mechanism picks in exactly the wrong direction when revisiting an arm is impossible. The remedy is to flip the sign—subtract a penalty (pessimism) instead of adding a bonus—so that poorly-covered state-action pairs are assigned *low* Q-values and the learned policy avoids them.



Greedy: $\arg\max \widehat{\mu}$ — • Arm 1 ✓ (but fragile to noise)

UCB: $\arg\max(\widehat{\mu} + \Gamma)$ — • Arm 2 ✗ (systematically wrong)

LCB: $\arg\max(\widehat{\mu} - \Gamma)$ — • Arm 1 ✓ (robust)

Figure 1: **Why optimism fails offline.** Bandit with $\mu_1^\star = 1$, $\mu_2^\star = 0$, data $N_1 = 100$, $N_2 = 1$, and point estimates equal to the true means ($\widehat{\mu}_1 = 1$, $\widehat{\mu}_2 = 0$; *no bad luck*). The penalty/bonus $\Gamma(a) = c/\sqrt{N(a)}$ makes Arm 2's confidence interval much wider than Arm 1's. **UCB** picks the upper endpoint of the CI and thus *deterministically* selects the under-sampled Arm 2—its bonus rewards data scarcity, which is exactly the wrong signal offline. **LCB** picks the lower endpoint and robustly selects Arm 1. Online, UCB would revisit Arm 2 and the bonus would shrink; offline, no revisits are possible, and the wrong-direction correction is permanent.

The pessimism principle:

$$\underbrace{\widehat{Q}_h = r_h + \widehat{P}_h \widehat{V}_{h+1} + b_h}_{\text{optimism (online, Lecture 11)}} \quad \longrightarrow \quad \underbrace{\widehat{Q}_h = r_h + \widehat{P}_h \widehat{V}_{h+1} - \Gamma_h}_{\text{pessimism (offline, this lecture)}}.$$

The penalty $\Gamma_h(s, a) \geq 0$ has the *same form* as the exploration bonus, but the sign is flipped.

What Causes Suboptimality?

Before presenting the PEVI algorithm, we analyze *what causes suboptimality* for any algorithm that constructs estimated Q-functions and a greedy policy from the dataset. This decomposition, due to Jin et al. (2021), is the finite-horizon offline analogue of the Perfor-

mance Difference via Bellman Error lemma from Lecture 7, and also the same standard policy-loss decomposition reused in the episodic regret analysis of Lecture 11. It reveals why pessimism is the right principle in the offline setting.

Decomposition of Suboptimality

Suppose an algorithm produces estimated Q-functions $\widehat{Q}_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ from the dataset $\mathcal{D}$. Define

$$\widehat{Q}_H(\cdot,\cdot) \;\coloneqq\; 0, \qquad \widehat{V}_h(s) \;\coloneqq\; \max_{a\in\mathcal{A}} \widehat{Q}_h(s,a), \qquad \widehat{\pi}_h(s) \in \operatorname*{argmax}_{a\in\mathcal{A}} \widehat{Q}_h(s,a),$$

and note that the *Bellman error* at step h is

$$(\mathcal{T}_h \widehat{Q}_{h+1})(s,a) - \widehat{Q}_h(s,a),$$

which measures how well $\widehat{Q}_h$ approximates the *true* Bellman backup of $\widehat{Q}_{h+1}$.

Lemma 1 (Decomposition of suboptimality)**.** *Let $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=0}^{H-1}$ be the greedy policy induced by $\widehat{Q}$. Then for any initial state s_0,*

$$\begin{aligned}
\mathrm{SubOpt}(\widehat{\pi}; s_0) \;=\; & \underbrace{\sum_{h=0}^{H-1} \mathbb{E}_{\widehat{\pi}}\Big[\widehat{Q}_h(s_h,a_h) - (\mathcal{T}_h\widehat{Q}_{h+1})(s_h,a_h) \;\Big|\; s_0\Big]}_{\text{(i): Pessimism-controlled term}} \\
& + \underbrace{\sum_{h=0}^{H-1} \mathbb{E}_{\pi^\star}\Big[(\mathcal{T}_h\widehat{Q}_{h+1})(s_h,a_h) - \widehat{Q}_h(s_h,a_h) \;\Big|\; s_0\Big]}_{\text{(ii): Optimism-controlled term}} \\
& + \underbrace{\sum_{h=0}^{H-1} \mathbb{E}_{\pi^\star}\Big[\widehat{Q}_h(s_h,\pi_h^\star(s_h)) - \widehat{Q}_h(s_h,\widehat{\pi}_h(s_h)) \;\Big|\; s_0\Big]}_{\text{(iii): Optimization Error}}. \qquad (1)
\end{aligned}$$

Here $\mathbb{E}_{\widehat{\pi}}$ and $\mathbb{E}_{\pi^\star}$ denote expectations over trajectories induced by $\widehat{\pi}$ and $\pi^\star$ in the true MDP, given the fixed functions $\widehat{Q}_{h+1}$ and $\widehat{Q}_h$.

Proof. The proof is the same telescoping argument behind the *Performance Difference via Bellman Error* lemma from Lecture 7, applied once with $\pi = \pi^\star$ and once with $\pi = \widehat{\pi}$ and then subtracted. Fix s_0 and abbreviate $\mathbb{E}_\pi[\cdot] = \mathbb{E}_\pi[\cdot \mid s_0]$. For any policy π, telescoping the Bellman equation $V_h^\pi(s) = Q_h^\pi(s,\pi_h(s)) = r_h(s,\pi_h(s)) + P_h^\star V_{h+1}^\pi(s,\pi_h(s))$ and using

$(\mathcal{T}_h\widehat{Q}_{h+1})(s,a) = r_h(s,a) + P_h^\star \widehat{V}_{h+1}(s,a)$ gives

$$V_0^\pi(s_0) - \widehat{V}_0(s_0) = \sum_{h=0}^{H-1} \mathbb{E}_\pi\left[(\mathcal{T}_h\widehat{Q}_{h+1})(s_h,a_h) - \widehat{Q}_h(s_h,a_h) + \widehat{Q}_h(s_h,\pi_h(s_h)) - \widehat{V}_h(s_h)\right]. \quad (2)$$

Since $\widehat{V}_h(s) = \widehat{Q}_h(s,\widehat{\pi}_h(s))$, applying (2) to $\pi = \pi^\star$ yields

$$V_0^\star(s_0) - \widehat{V}_0(s_0) = \sum_{h=0}^{H-1} \mathbb{E}_{\pi^\star}\left[(\mathcal{T}_h\widehat{Q}_{h+1})(s_h,a_h) - \widehat{Q}_h(s_h,a_h) + \widehat{Q}_h(s_h,\pi_h^\star(s_h)) - \widehat{Q}_h(s_h,\widehat{\pi}_h(s_h))\right],$$

and applying (2) to $\pi = \widehat{\pi}$ yields

$$V_0^{\widehat{\pi}}(s_0) - \widehat{V}_0(s_0) = \sum_{h=0}^{H-1} \mathbb{E}_{\widehat{\pi}}\left[(\mathcal{T}_h\widehat{Q}_{h+1})(s_h,a_h) - \widehat{Q}_h(s_h,a_h)\right],$$

since $\widehat{Q}_h(s_h,\widehat{\pi}_h(s_h)) - \widehat{V}_h(s_h) = 0$. Subtracting the second from the first and rewriting the first difference as $\widehat{Q}_h - \mathcal{T}_h\widehat{Q}_{h+1}$ gives (1). □

Interpreting the three sources of suboptimality.

- **Term (i): pessimism-controlled term.** This is the Bellman error along the trajectory of $\widehat{\pi}$, written in the order $\widehat{Q}_h - \mathcal{T}_h\widehat{Q}_{h+1}$. If $\widehat{Q}_h \le (\mathcal{T}_h\widehat{Q}_{h+1})$ pointwise, equivalently $\widehat{Q}_h - (\mathcal{T}_h\widehat{Q}_{h+1}) \le 0$, then term (i) is non-positive. Without such sign control, this term can be badly biased because $\widehat{\pi}$ is chosen from the same data used to estimate $\widehat{Q}$.
- **Term (ii): optimism-controlled term.** This is the Bellman error along the trajectory of $\pi^\star$. It is the comparator-policy term that already appeared in the telescoping / Bellman-error decompositions from earlier lectures. In the symmetric optimistic direction, if $\widehat{Q}_h \ge (\mathcal{T}_h\widehat{Q}_{h+1})$ pointwise, then this term is controlled with the favorable sign. In offline RL, we instead upper-bound it by uncertainty along $\pi^\star$'s trajectory.
- **Term (iii): Optimization error.** This term measures the gap $\widehat{Q}_h(s_h,\pi_h^\star(s_h)) - \widehat{Q}_h(s_h,\widehat{\pi}_h(s_h))$ under $\pi^\star$'s trajectory. It is non-positive when $\widehat{\pi}$ is greedy with respect to $\widehat{Q}_h$ (since then $\widehat{Q}_h(s,\widehat{\pi}_h(s)) \ge \widehat{Q}_h(s,a)$ for all a), and hence can be dropped.

The key insight: for the greedy policy $\widehat{\pi}$, term (iii) ≤ 0, and pessimism eliminates the pessimism-controlled term (i), so the remaining burden is to control the optimism-controlled term (ii). This depends on how well $\mathcal{D}$ covers the *optimal policy's* trajectory, not on uniform coverage of the entire state-action space.

Illustration via a Special Case: Multi-Armed Bandit

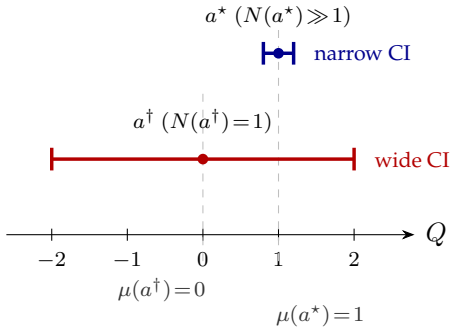
To build intuition, consider the multi-armed bandit (MAB), a special case with $|\mathcal{S}| = 1$ and $H = 1$. There are A arms with mean rewards $\mu(a)$, $a \in \mathcal{A}$. The dataset consists of K pulls $\mathcal{D} = \{(a^\tau, r^\tau)\}_{\tau=0}^{K-1}$, and the sample average estimator is

$$\widehat{\mu}(a) = \frac{1}{N(a)} \sum_{\tau:\, a^\tau = a} r^\tau, \qquad N(a) = \sum_{\tau=0}^{K-1} \mathbb{1}\{a^\tau = a\}.$$

Bandit Decomposition. Let $\widehat{Q} : \mathcal{A} \to \mathbb{R}$ be any bandit score and let $\widehat{\pi} = \operatorname{argmax}_a \widehat{Q}(a)$. Since the Bellman error is $\mu(a) - \widehat{Q}(a)$ and the optimization term is $\widehat{Q}(\pi^\star) - \widehat{Q}(\widehat{\pi}) \le 0$, the decomposition (1) becomes

$$\mathrm{SubOpt}(\widehat{\pi}) \le \underbrace{\mathbb{E}_{\widehat{\pi}}\left[\widehat{Q}(\widehat{\pi}) - \mu(\widehat{\pi})\right]}_{\text{(i)}} + \underbrace{\mathbb{E}_{\pi^\star}\left[\mu(\pi^\star) - \widehat{Q}(\pi^\star)\right]}_{\text{(ii)}}.$$

For naive greedy, $\widehat{Q} = \widehat{\mu}$, so $\widehat{\pi} = \operatorname{argmax}_a \widehat{\mu}(a)$ and the estimation error $\widehat{\mu}(a) - \mu(a)$ is mean zero for each arm. However, term (i) becomes $\widehat{\mu}(\widehat{\pi}) - \mu(\widehat{\pi})$, which is *positively biased* because $\widehat{\pi}$ is selected to maximize $\widehat{\mu}$.



Term (ii)	Oracle-controlled: only uses the well-covered arm $a^\star$.
Term (i)	Selection-biased: $\widehat{\pi}$ may jump to the poorly covered arm $a^\dagger$.
Pessimism	Pointwise pessimism: if $\widehat{Q}(a) \le \mu(a)$ for all a, then $\mathbb{E}_{\widehat{\pi}}[\widehat{Q}(\widehat{\pi}) - \mu(\widehat{\pi})] \le 0$.

Figure 2: **Bandit view of the offline decomposition.** Term (ii) is small when the optimal arm $a^\star$ is well covered. Term (i) is not oracle-controlled because the data-dependent policy $\widehat{\pi}$ can jump to the wide-CI arm $a^\dagger$. Pessimism penalizes $a^\dagger$ and makes term (i) non-positive.

Figure 2 isolates the asymmetry in (1): term (ii) is oracle-controlled by the coverage of $a^\star$, whereas term (i) is evaluated at the data-dependent arm $\widehat{\pi}$ and can therefore be driven by the poorly covered arm $a^\dagger$. Pessimism is stronger than concentration: if $\widehat{Q}(a) = \widehat{\mu}(a) - \Gamma(a)$

satisfies $\widehat{Q}(a) \le \mu(a)$ for all a, then

$$\mathbb{E}_{\widehat{\pi}}\Big[\widehat{Q}(\widehat{\pi}) - \mu(\widehat{\pi})\Big] \le 0,$$

so only the oracle-controlled term remains.

PEVI for Linear MDPs

We now study Pessimistic Value Iteration (PEVI) as a *linear-MDP algorithm*. The presentation is intentionally parallel to Lecture 11's treatment of Lin-UCBVI: the Bellman backup is estimated by linear regression, uncertainty is measured by an ellipsoidal norm, and the only conceptual change is that the uncertainty term is *subtracted* rather than added.

The Linear MDP

We reuse the linear transition model from Lecture 11 and add a linear reward model, restated here for convenience.

Definition 2 (Linear MDP)**.** *An episodic MDP satisfies the* linear MDP *structure with feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ if there exist unknown signed measures $\mu_h^\star = (\mu_h^{\star(1)}, \dots, \mu_h^{\star(d)})$ over $\mathcal{S}$ and an unknown vector $\theta_h \in \mathbb{R}^d$ such that*

$$P_h^\star(\cdot \mid s,a) = \langle \phi(s,a), \mu_h^\star(\cdot)\rangle, \qquad r_h(s,a) = \langle \phi(s,a), \theta_h\rangle.$$

We assume $\|\phi(s,a)\|_2 \le 1$ for all (s,a) and $\max\{\|\mu_h^\star(\mathcal{S})\|_2, \|\theta_h\|_2\} \le \sqrt{d}$.

The key property is that the Bellman backup remains linear in ϕ. By the linearization of the transition term from Lecture 11, for any state-value function $f : \mathcal{S} \to \mathbb{R}$,

$$(P_h^\star f)(s,a) = \langle \phi(s,a),\ (\mu_h^\star)^\mathsf{T} f\rangle.$$

Therefore, for *any* Q-function $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,

$$(\mathcal{T}_h g)(s,a) = r_h(s,a) + (P_h^\star V_g)(s,a) = \langle \phi(s,a),\ w_{h,g}^\star\rangle, \quad \text{where } w_{h,g}^\star \coloneqq \theta_h + (\mu_h^\star)^\mathsf{T} V_g \in \mathbb{R}^d.$$

This linearization holds regardless of whether g is itself linear—the Bellman backup of any bounded next-step Q-function remains linear in ϕ.

Algorithm 1 PEVI (episodic, linear MDP)

Require: Dataset $\mathcal{D} = \{(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau)\}_{\tau,h=0}^{K-1,H-1}$, regularization $\lambda \geq 1$, confidence $\beta > 0$

Initialize: $\widehat{Q}_H(\cdot,\cdot) \leftarrow 0$, $\widehat{V}_H(\cdot) \leftarrow 0$
for $h = H-1, H-2, \ldots, 0$ **do**
  $\Lambda_h \leftarrow \lambda I + \sum_{\tau=0}^{K-1} \phi(s_h^\tau, a_h^\tau)\, \phi(s_h^\tau, a_h^\tau)^\mathsf{T}$ ▷ *Design matrix*
  $\widehat{w}_h \leftarrow \Lambda_h^{-1} \sum_{\tau=0}^{K-1} \phi(s_h^\tau, a_h^\tau)(r_h^\tau + \widehat{V}_{h+1}(s_{h+1}^\tau))$ ▷ *Ridge regression*
  $\Gamma_h(\cdot,\cdot) \leftarrow \beta \cdot \big(\phi(\cdot,\cdot)^\mathsf{T} \Lambda_h^{-1} \phi(\cdot,\cdot)\big)^{1/2}$ ▷ *Uncertainty*
  $\widehat{Q}_h(\cdot,\cdot) \leftarrow \min\{\langle \phi(\cdot,\cdot), \widehat{w}_h\rangle - \Gamma_h(\cdot,\cdot),\ H-h\}^+$ ▷ *Pessimistic clipping*
  $\widehat{\pi}_h(\cdot) \leftarrow \operatorname{argmax}_a \widehat{Q}_h(\cdot, a)$ ▷ *Optimization*
  $\widehat{V}_h(\cdot) \leftarrow \widehat{Q}_h(\cdot, \widehat{\pi}_h(\cdot))$ ▷ *Evaluation*
end for
Output: $\texttt{Pess}(\mathcal{D}) = \{\widehat{\pi}_h\}_{h=0}^{H-1}$

Algorithm: Linear PEVI

Ridge regression for the Bellman backup. At each step h, define the regularized design matrix

$$\Lambda_h \;\coloneqq\; \lambda I + \sum_{\tau=0}^{K-1} \phi(s_h^\tau, a_h^\tau)\, \phi(s_h^\tau, a_h^\tau)^\mathsf{T}, \qquad \lambda \geq 1, \tag{3}$$

and the ridge regression estimate of $w^\star_{h,\widehat{Q}_{h+1}}$:

$$\widehat{w}_h \;\coloneqq\; \Lambda_h^{-1} \sum_{\tau=0}^{K-1} \phi(s_h^\tau, a_h^\tau)\big(r_h^\tau + \widehat{V}_{h+1}(s_{h+1}^\tau)\big), \tag{4}$$

and, more generally, for any Q-function g define the empirical Bellman backup

$$(\widehat{\mathcal{T}}_h g)(s,a) \;\coloneqq\; \phi(s,a)^\mathsf{T} \Lambda_h^{-1} \sum_{\tau=0}^{K-1} \phi(s_h^\tau, a_h^\tau)\big(r_h^\tau + V_g(s_{h+1}^\tau)\big).$$

In particular, $(\widehat{\mathcal{T}}_h \widehat{Q}_{h+1})(s,a) = \langle \phi(s,a), \widehat{w}_h\rangle$.

Ellipsoidal penalty. The penalty function is the Mahalanobis norm:

$$\Gamma_h(s,a) \;\coloneqq\; \beta \cdot \big(\phi(s,a)^\mathsf{T} \Lambda_h^{-1} \phi(s,a)\big)^{1/2} \;=\; \beta \cdot \|\phi(s,a)\|_{\Lambda_h^{-1}}, \tag{5}$$

where $\beta > 0$ is a scaling parameter set below. This is the *same form* as the Lin-UCBVI bonus from Lecture 11, but now it is *subtracted* from the estimate rather than added.

PEVI for linear MDP:

$$\widehat{Q}_h(s,a) = \left[\underbrace{\langle \phi(s,a), \widehat{w}_h \rangle}_{\text{ridge regression estimate}} - \underbrace{\beta \cdot \|\phi(s,a)\|_{\Lambda_h^{-1}}}_{\text{ellipsoidal penalty}} \right]^{+}_{H-h}.$$

Lin-UCBVI vs. PEVI.

	Lin-UCBVI (Lecture 11)	**PEVI (this lecture)**
Data	Adaptive online trajectories	Fixed offline dataset
Regression target	$r_h + \widehat{V}_{h+1}(s')$	$r_h + \widehat{V}_{h+1}(s')$
Correction	$+\beta\|\phi\|_{(\Lambda_h^k)^{-1}}$	$-\beta\|\phi\|_{\Lambda_h^{-1}}$
Effect	Encourages exploration	Avoids uncertain regions
Performance term	Bonuses along π^k	Penalties along $\pi^\star$

Main Result

Theorem 2 (PEVI suboptimality for linear MDP)**.** *Suppose Assumption 1 holds and the underlying MDP is a linear MDP (Definition 2). In Algorithm 1, set*

$$\lambda = 1, \qquad \beta = c \cdot dH\sqrt{\zeta}, \qquad \zeta = \log(2dHK/\xi).$$

Then with probability at least $1-\xi$*, for any* $s_0 \in \mathcal{S}$*,*

$$\mathrm{SubOpt}\big(\mathtt{Pess}(\mathcal{D}); s_0\big) \;\le\; 2\beta \sum_{h=0}^{H-1} \mathbb{E}_{\pi^\star}\left[\left(\phi(s_h,a_h)^\mathsf{T} \Lambda_h^{-1} \phi(s_h,a_h)\right)^{1/2} \;\middle|\; s_0 \right]. \tag{6}$$

PEVI suboptimality (linear MDP):

$$\mathrm{SubOpt}\big(\mathtt{Pess}(\mathcal{D}); s_0\big) \;\le\; \widetilde{O}(dH) \cdot \sum_{h=0}^{H-1} \mathbb{E}_{\pi^\star}\left[\|\phi(s_h,a_h)\|_{\Lambda_h^{-1}} \;\middle|\; s_0 \right].$$

The bound depends on the penalty evaluated along $\pi^\star$'s trajectory, with the design matrix Λ_h determined by the dataset.

Analysis of Linear PEVI

As in Lecture 11, the key step is a confidence bound for the linear Bellman backup. The only extra issue is that the target $\widehat{V}_{h+1}$ is fitted from the same offline dataset, so the usual fixed-target ridge bound must be uniformized over the PEVI iterate class.

Proposition 3 (Fixed-target model confidence for linear PEVI)**.** *Fix $h \in \{0, \dots, H-1\}$ and a deterministic stage-$(h+1)$ Q-function Q with $V_Q(\cdot) \in [0, H-h-1]$. Under the conditions of Theorem 2, with probability at least $1-\xi$, simultaneously for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\left|(\widehat{\mathcal{T}}_h Q)(s,a) - (\mathcal{T}_h Q)(s,a)\right| \;\le\; \beta \cdot \|\phi(s,a)\|_{\Lambda_h^{-1}}.$$

Proof sketch. Fix h and first fix a deterministic stage-$(h+1)$ Q-function Q. Write $V \coloneqq V_Q$ and $\phi_\tau \coloneqq \phi(s_h^\tau, a_h^\tau)$. By Bellman closure, there exists $w_{h,Q}^\star \in \mathbb{R}^d$ such that

$$r_h(s,a) + \big(P_h^\star V_Q\big)(s,a) = \langle \phi(s,a), w_{h,Q}^\star \rangle, \qquad \|w_{h,Q}^\star\|_2 \le H\sqrt{d}.$$

Also define

$$\eta_\tau(Q) \coloneqq V_Q(s_{h+1}^\tau) - \big(P_h^\star V_Q\big)(s_h^\tau, a_h^\tau).$$

Then

$$r_h^\tau + V_Q(s_{h+1}^\tau) = r_h(s_h^\tau, a_h^\tau) + \big(P_h^\star V_Q\big)(s_h^\tau, a_h^\tau) + \eta_\tau(Q),$$

where $\{\phi_\tau \eta_\tau(Q)\}_{\tau=0}^{K-1}$ is a martingale difference sequence and $|\eta_\tau(Q)| \le H$. Since $\Lambda_h = \lambda I + \sum_{\tau=0}^{K-1} \phi_\tau \phi_\tau^\mathsf{T}$,

$$\begin{aligned}
&(\widehat{\mathcal{T}}_h Q)(s,a) - (\mathcal{T}_h Q)(s,a) \\
&\qquad = \phi(s,a)^\mathsf{T} \Lambda_h^{-1} \sum_{\tau=0}^{K-1} \phi_\tau \eta_\tau(Q) - \lambda\, \phi(s,a)^\mathsf{T} \Lambda_h^{-1} w_{h,Q}^\star.
\end{aligned}$$

Hence, for every fixed Q,

$$\begin{aligned}
&\left|(\widehat{\mathcal{T}}_h Q)(s,a) - (\mathcal{T}_h Q)(s,a)\right| \\
&\qquad \le \|\phi(s,a)\|_{\Lambda_h^{-1}} \left\| \sum_{\tau=0}^{K-1} \phi_\tau \eta_\tau(Q) \right\|_{\Lambda_h^{-1}} + \lambda\, |\phi(s,a)^\mathsf{T} \Lambda_h^{-1} w_{h,Q}^\star|.
\end{aligned}$$

By a self-normalized bound (Abbasi-Yadkori et al., 2011),

$$\left\| \sum_{\tau=0}^{K-1} \phi_\tau \eta_\tau(Q) \right\|_{\Lambda_h^{-1}} = O(H\sqrt{d\zeta}),$$

and since $\Lambda_h \succeq \lambda I$,

$$\lambda \left|\phi(s,a)^\mathsf{T} \Lambda_h^{-1} w^\star_{h,Q}\right| \le \sqrt{\lambda}\, \|w^\star_{h,Q}\|_2 \cdot \|\phi(s,a)\|_{\Lambda_h^{-1}} \le H\sqrt{\lambda d}\, \|\phi(s,a)\|_{\Lambda_h^{-1}}.$$

Therefore,

$$\begin{aligned}&\left|(\widehat{\mathcal{T}}_h Q)(s,a) - (\mathcal{T}_h Q)(s,a)\right| \\ &\quad \le O(H\sqrt{d\zeta} + H\sqrt{\lambda d}) \cdot \|\phi(s,a)\|_{\Lambda_h^{-1}}.\end{aligned}$$

This proves the fixed-target bound. □

Corollary 4 (Confidence for the PEVI iterate)**.** *Under the conditions of Theorem 2, with probability at least $1-\xi$, simultaneously for all $h \in \{0,\dots,H-1\}$ and all $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$\left|(\widehat{\mathcal{T}}_h \widehat{Q}_{h+1})(s,a) - (\mathcal{T}_h \widehat{Q}_{h+1})(s,a)\right| \le \beta \|\phi(s,a)\|_{\Lambda_h^{-1}}.$$

External ingredient. Proposition 3 gives the bound for every fixed deterministic Q. To instantiate it at the data-dependent PEVI iterate $\widehat{Q}_{h+1}$, Jin et al. (2021) uniformize over an ℓ_∞-cover of the PEVI iterate class. The log-covering number is $\widetilde{O}(d^2)$, which is absorbed by the choice $\beta = c\, dH\sqrt{\zeta}$. □

Lemma 5 (Pointwise pessimism of $\widehat{Q}$)**.** *On the event of Corollary 4, simultaneously for all $h \in \{0,\dots,H-1\}$ and all $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$\big(\mathcal{T}_h \widehat{Q}_{h+1}\big)(s,a) - 2\Gamma_h(s,a) \;\le\; \widehat{Q}_h(s,a) \;\le\; \big(\mathcal{T}_h \widehat{Q}_{h+1}\big)(s,a).$$

Equivalently,

$$0 \;\le\; (\mathcal{T}_h \widehat{Q}_{h+1})(s,a) - \widehat{Q}_h(s,a) \;\le\; 2\Gamma_h(s,a) \;=\; 2\beta \|\phi(s,a)\|_{\Lambda_h^{-1}}.$$

Proof.

$$\widehat{Q}_h(s,a) = \min\Big\{\big(\widehat{\mathcal{T}}_h \widehat{Q}_{h+1}\big)(s,a) - \Gamma_h(s,a),\; H-h\Big\}^+.$$

By construction, $\widehat{V}_{h+1}(\cdot) \in [0, H-h-1]$, so

$$0 \;\le\; (\mathcal{T}_h \widehat{Q}_{h+1})(s,a) = r_h(s,a) + P_h^\star \widehat{V}_{h+1}(s,a) \le H-h.$$

On the event of Corollary 4,

$$\big(\mathcal{T}_h \widehat{Q}_{h+1}\big)(s,a) - 2\Gamma_h(s,a) \;\le\; \big(\widehat{\mathcal{T}}_h \widehat{Q}_{h+1}\big)(s,a) - \Gamma_h(s,a) \;\le\; \big(\mathcal{T}_h \widehat{Q}_{h+1}\big)(s,a) \;\le\; H-h.$$

Therefore the upper truncation is inactive, and

$$\widehat{Q}_h(s,a) = \max\Big\{ \big(\widehat{\mathcal{T}}_h \widehat{Q}_{h+1}\big)(s,a) - \Gamma_h(s,a),\ 0 \Big\}.$$

Using the display above together with $0 \le (\mathcal{T}_h \widehat{Q}_{h+1})(s,a)$, we obtain

$$\widehat{Q}_h(s,a) \le \big(\mathcal{T}_h \widehat{Q}_{h+1}\big)(s,a),$$

because $\max\{u,0\} \le x$ whenever $u \le x$ and $0 \le x$. Likewise,

$$\widehat{Q}_h(s,a) \ge \big(\widehat{\mathcal{T}}_h \widehat{Q}_{h+1}\big)(s,a) - \Gamma_h(s,a) \ge \big(\mathcal{T}_h \widehat{Q}_{h+1}\big)(s,a) - 2\Gamma_h(s,a),$$

since $\max\{u,0\} \ge u$. This proves the sandwich bound, and the equivalent Bellman-error bound follows by rearranging. □

Proof of Theorem 2. Let $\widehat{\pi} = \texttt{Pess}(\mathcal{D})$, and let E denote the event of Corollary 4. On E, Lemma 5 yields pointwise control of the Bellman error:

$$0 \le (\mathcal{T}_h \widehat{Q}_{h+1})(s,a) - \widehat{Q}_h(s,a) \le 2\beta \|\phi(s,a)\|_{\Lambda_h^{-1}}.$$

Now fix the dataset on the event E, so that $\widehat{Q}$ and $\widehat{\pi}$ are deterministic functions and Lemma 1 applies directly. Lemma 1 gives

$$\begin{aligned}
\mathrm{SubOpt}(\widehat{\pi}; s_0) = &\sum_{h=0}^{H-1} \mathbb{E}_{\widehat{\pi}}\Big[\widehat{Q}_h(s_h,a_h) - (\mathcal{T}_h \widehat{Q}_{h+1})(s_h,a_h) \mid s_0 \Big] \\
&+ \sum_{h=0}^{H-1} \mathbb{E}_{\pi^\star}\Big[(\mathcal{T}_h \widehat{Q}_{h+1})(s_h,a_h) - \widehat{Q}_h(s_h,a_h) \mid s_0 \Big] \\
&+ \sum_{h=0}^{H-1} \mathbb{E}_{\pi^\star}\Big[\widehat{Q}_h(s_h,\pi_h^\star(s_h)) - \widehat{Q}_h(s_h,\widehat{\pi}_h(s_h)) \mid s_0 \Big].
\end{aligned}$$

We now bound the three terms on the right-hand side:

- Since $\widehat{\pi}_h(s) \in \mathrm{argmax}_a\, \widehat{Q}_h(s,a)$, we have

  $$\widehat{Q}_h(s,\pi_h^\star(s)) - \widehat{Q}_h(s,\widehat{\pi}_h(s)) \le 0 \qquad \text{for every } (h,s),$$

  so the optimization error term (iii) is non-positive.

- Since $\widehat{Q}_h(s,a) - (\mathcal{T}_h\widehat{Q}_{h+1})(s,a) \le 0$ pointwise on E, the pessimism-controlled term

$$\sum_{h=0}^{H-1} \mathbb{E}_{\widehat{\pi}}\left[\widehat{Q}_h(s_h,a_h) - (\mathcal{T}_h\widehat{Q}_{h+1})(s_h,a_h) \mid s_0\right]$$

  is also non-positive.

- Finally, using the upper bound

$$(\mathcal{T}_h\widehat{Q}_{h+1})(s,a) - \widehat{Q}_h(s,a) \le 2\beta\|\phi(s,a)\|_{\Lambda_h^{-1}}$$

  pointwise on E and taking expectation along $\pi^\star$ yields

$$\begin{aligned}\sum_{h=0}^{H-1} \mathbb{E}_{\pi^\star}\left[(\mathcal{T}_h\widehat{Q}_{h+1})(s_h,a_h) - \widehat{Q}_h(s_h,a_h) \mid s_0\right] &\le \sum_{h=0}^{H-1} \mathbb{E}_{\pi^\star}\left[2\beta\|\phi(s_h,a_h)\|_{\Lambda_h^{-1}} \mid s_0\right] \\ &= 2\beta \sum_{h=0}^{H-1} \mathbb{E}_{\pi^\star}\left[\|\phi(s_h,a_h)\|_{\Lambda_h^{-1}} \,\middle|\, s_0\right].\end{aligned}$$

Combining the three displays above, we obtain

$$\mathrm{SubOpt}(\widehat{\pi}; s_0) \le 2\beta \sum_{h=0}^{H-1} \mathbb{E}_{\pi^\star}\left[\|\phi(s_h,a_h)\|_{\Lambda_h^{-1}} \,\middle|\, s_0\right].$$

Since $\Pr(E) \ge 1-\xi$ by Corollary 4, the theorem follows. □

How linear PEVI works.

1. The linear MDP structure turns the Bellman backup into a linear regression problem in the same feature map ϕ used by Lin-UCBVI.
2. The ellipsoidal penalty $\beta\|\phi\|_{\Lambda_h^{-1}}$ makes the estimated Q-function a *lower* bound on the Bellman backup, so the pessimism-controlled term in Lemma 1 has the favorable sign.
3. The remaining suboptimality is paid only along $\pi^\star$'s trajectory. This is the offline analog of "paying for the bonus" in Lecture 11, with the comparator trajectory switching from the learned policy to the optimal policy.

The Oracle Property

A remarkable feature of the bound (6) is that the suboptimality only depends on how well the dataset $\mathcal{D}$ covers the trajectory of $\pi^\star$. We call this the *oracle property*: the algorithm automatically adapts to the optimal policy, even though $\pi^\star$ is unknown.

Illustration via tabular MDP. Consider the tabular MDP as a special case of the linear MDP with $\phi(s,a) = e_{(s,a)} \in \mathbb{R}^{SA}$ (the one-hot encoding). Then

$$\Lambda_h = \lambda I + \sum_\tau \phi_\tau \phi_\tau^\mathsf{T} = \mathrm{diag}(\lambda + N_h(s,a) : (s,a) \in \mathcal{S} \times \mathcal{A}),$$

where $N_h(s,a) = \sum_\tau \mathbb{1}\{(s_h^\tau, a_h^\tau) = (s,a)\}$ is the number of visits to (s,a) at step h. The penalty becomes

$$\Gamma_h(s,a) = \frac{\beta}{\sqrt{\lambda + N_h(s,a)}},$$

and the bound (6) specializes to

$$\mathrm{SubOpt} \;\le\; 2\beta \sum_{h=0}^{H-1} \mathbb{E}_{\pi^\star}\left[\frac{1}{\sqrt{\lambda + N_h(s_h,a_h)}}\right].$$

The suboptimality only depends on the counts $N_h(s_h^\star, a_h^\star)$ along the optimal trajectory. State-action pairs (s,a) that $\pi^\star$ never visits contribute zero to the bound, regardless of how often (or rarely) they appear in the dataset.

Oracle property. Suppose the transition is deterministic, so $\pi^\star$ follows a unique trajectory $\{(s_h^\star, a_h^\star)\}_{h=0}^{H-1}$. Then

$$\mathrm{SubOpt} \;\le\; \sum_{h=0}^{H-1} \frac{2\beta}{\sqrt{\lambda + N_h(s_h^\star, a_h^\star)}}.$$

This depends only on how many times the dataset visits the states on $\pi^\star$'s path. A behavior policy π^b that visits many "wrong" states does not hurt; and a behavior policy that focuses on the optimal trajectory leads to low suboptimality. Meanwhile, a state-action pair (s°, a°) off $\pi^\star$'s path—even one visited millions of times—does not affect the bound at all.

This is analogous to *local* sample complexity: the difficulty of learning $\pi^\star$ depends on the *local* coverage near $\pi^\star$, not on global coverage.

Sufficient Coverage and Sample Complexity

To translate the data-dependent bound (6) into a sample complexity guarantee, we need to quantify how well the dataset covers $\pi^\star$'s trajectory.

Corollary 6 (Well-explored dataset). *Suppose there exists $c^\dagger > 0$ such that the event*

$$\mathcal{E}^\dagger = \left\{\Lambda_h \succeq I + c^\dagger \cdot K \cdot \mathbb{E}_{\pi^\star}\left[\phi(s_h,a_h)\phi(s_h,a_h)^\mathsf{T} \;\middle|\; s_0\right] \text{ for all } h\right\}$$

holds with probability at least $1 - \xi/2$. Set $\lambda = 1$, $\beta = c \cdot dH\sqrt{\zeta}$, $\zeta = \log(4dHK/\xi)$. Then with

probability at least $1-\xi$,

$$\mathrm{SubOpt}\big(\texttt{Pess}(\mathcal{D}); s_0\big) \;\le\; c' \cdot d^{3/2} H^2 K^{-1/2} \sqrt{\zeta},$$

where c' *depends only on* $c^\dagger$ *and* c.

Under stronger conditions on the behavior policy (e.g., $\lambda_{\min}(\Sigma_h) \ge \underline{c}/d$ for all h, where $\Sigma_h = \mathbb{E}_{\pi^b}[\phi(s_h,a_h)\phi(s_h,a_h)^\mathsf{T}]$), Corollary 4.6 of Jin et al. (2021) shows that $K = \widetilde{O}(d^3H^4/\varepsilon^2)$ i.i.d. trajectories suffice for ε-suboptimality.

Information-Theoretic Lower Bound

PEVI's suboptimality bound is *minimax optimal* in the linear MDP setting, as established by the following information-theoretic lower bound. Recall that $P_{\mathcal{D}}$ denotes the joint distribution of the data collecting process.

Theorem 7 (Information-theoretic lower bound)**.** *There exists an absolute constant* $c > 0$ *such that for the output* $\texttt{Algo}(\mathcal{D})$ *of* any *algorithm, there exist a linear MDP* $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P^\star, r)$*, an initial state* s_0*, and a compliant dataset* $\mathcal{D}$ *such that*

$$\mathbb{E}_{\mathcal{D}\sim P_{\mathcal{D}}}\left[\frac{\mathrm{SubOpt}(\texttt{Algo}(\mathcal{D}); s_0)}{\sum_{h=0}^{H-1} \mathbb{E}_{\pi^\star}\left[\|\phi(s_h,a_h)\|_{\Lambda_h^{-1}} \;\middle|\; s_0\right]}\right] \;\ge\; c.$$

Minimax optimality of PEVI: Up to the scaling parameter $\beta = \widetilde{O}(dH)$, the suboptimality of PEVI (Theorem 2) matches the lower bound (Theorem 7). In other words, the data-dependent quantity $\sum_h \mathbb{E}_{\pi^\star}[\|\phi\|_{\Lambda_h^{-1}}]$ is the *fundamental statistical difficulty* of the offline RL problem in linear MDPs.

Beyond Pointwise Pessimism: Bellman-Consistent Pessimism

The linear PEVI analysis above relies on a *pointwise* penalty $\Gamma_h(s,a)$ such that $\widehat{Q}_h \le \mathcal{T}_h \widehat{Q}_{h+1}$ at every state-action pair. In structured settings such as tabular and linear MDPs, these penalties can be written explicitly. For general function approximation, however, no comparable closed-form pointwise penalty is available.

This mirrors online RL (Lecture 11): pointwise bonuses work for tabular (UCBVI) and linear (Lin-UCBVI) MDPs, but GOLF uses *global* optimism for general function classes. We now develop the offline analog: *Bellman-consistent pessimism* (BCP) (Xie et al., 2021).

Pointwise vs. global pessimism (cf. Lecture 11).

	Pointwise pessimism (PEVI)	**Global pessimism (BCP)**
Goal	$\widehat{Q}_h \le (\mathcal{T}_h \widehat{Q}_{h+1})$ for all (s,a,h)	Pessimistic evaluation at s_0
Mechanism	Subtract penalty $\Gamma_h(s,a)$	Version space + min
Scope	Requires closed-form penalty	General $\mathcal{F}$
Online analog	UCBVI / Lin-UCBVI	GOLF

Setup

We now switch notation and follow the discounted infinite-horizon framework of Xie et al. (2021): an MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma, s_0)$ with discount factor $\gamma \in [0,1)$, transition $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, reward $R : \mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$, and value bound $V_{\max} = R_{\max}/(1-\gamma)$. We write $J(\pi) = V^\pi(s_0)$ for the performance of policy π, and $f(s_0, \pi) \coloneqq \sum_a f(s_0,a)\,\pi(a \mid s_0)$ for a function f evaluated under a (possibly stochastic) policy. The policy-specific Bellman operator and transition operator are

$$(\mathcal{T}^\pi f)(s,a) = R(s,a) + \gamma\,(\mathcal{P}^\pi f)(s,a), \qquad (\mathcal{P}^\pi f)(s,a) = \mathbb{E}_{s' \sim P(\cdot|s,a)}\big[f(s', \pi(s'))\big].$$

Discounted Bellman equation (used in the BCP section):

$$Q^\pi = \mathcal{T}^\pi Q^\pi, \qquad \text{equivalently} \qquad Q^\pi(s,a) = R(s,a) + \gamma\,(\mathcal{P}^\pi Q^\pi)(s,a).$$

The *discounted state-action occupancy* is $d_\pi(s,a) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s, a_t = a \mid s_0, \pi)$.

The offline dataset $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^n$ consists of n i.i.d. tuples: $(s,a) \sim \mu$, $r = R(s,a)$, $s' \sim P(\cdot \mid s,a)$, for a *data distribution* μ over $\mathcal{S} \times \mathcal{A}$. We write $\|g\|_{2,\nu}^2 = \mathbb{E}_\nu[g(s,a)^2]$ for the ν-weighted L^2 norm.

We are given a finite value-function class $\mathcal{F} \subset (\mathcal{S} \times \mathcal{A} \to [0, V_{\max}])$ and policy class Π. For clarity, we present the theory under exact realizability and completeness; the full result in Xie et al. (2021) handles approximate versions of both assumptions (with errors $\varepsilon_{\mathcal{F}}, \varepsilon_{\mathcal{F},\mathcal{F}} \ge 0$) and decomposes the bound into on-support and off-support errors that automatically adapt to the best bias-variance tradeoff.

Assumption 2 (Realizability). *$Q^\pi \in \mathcal{F}$ for all $\pi \in \Pi$.*

Assumption 3 (Completeness). *$\mathcal{T}^\pi f \in \mathcal{F}$ for all $\pi \in \Pi$ and $f \in \mathcal{F}$.*

Definition 3 (Distribution shift coefficient)**.**

$$\mathcal{C}(\nu;\mu,\mathcal{F},\pi) \;:=\; \max_{f\in\mathcal{F}} \frac{\big(\mathbb{E}_\nu[|f-\mathcal{T}^\pi f|]\big)^2}{\mathbb{E}_\mu[(f-\mathcal{T}^\pi f)^2]}.$$

This is the *performance-relevant* shift coefficient for our proof: it measures how well squared Bellman error under μ (data) controls *average* Bellman error under ν (target). When $\mathcal{C}(d_\pi;\mu,\mathcal{F},\pi) \le C_2$, any function with small Bellman error on the data also has small average Bellman error along π's trajectory.

Remark 1 (Relation to density ratio and covariance shift)**.** *This is the weakest coefficient needed here, because Lemma 11 reduces performance to the* average *Bellman error* $\mathbb{E}_{d_\pi}[|f-\mathcal{T}^\pi f|]$*. If the Bellman-error class* $\mathcal{G}_{\mathcal{F},\pi} := \{f-\mathcal{T}^\pi f : f\in\mathcal{F}\}$ *is linear in some feature map* ψ*, say* $g=\langle\psi,w\rangle$*, then*

$$\mathcal{C}(\nu;\mu,\mathcal{F},\pi) \le \sup_{w\neq 0} \frac{\big(\mathbb{E}_\nu[|\langle\psi,w\rangle|]\big)^2}{w^\mathsf{T}\Sigma_\mu w}, \qquad \Sigma_\mu = \mathbb{E}_\mu[\psi\psi^\mathsf{T}].$$

Indeed, each $f\in\mathcal{F}$ *induces some* $g=f-\mathcal{T}^\pi f\in\mathcal{G}_{\mathcal{F},\pi}$*, and by linearity* $g=\langle\psi,w\rangle$ *for some* w*. Moreover,*

$$\mathbb{E}_\mu[g^2] = \mathbb{E}_\mu[\langle\psi,w\rangle^2] = w^\mathsf{T}\Sigma_\mu w.$$

Substituting this into Definition 3 gives the displayed upper bound. Moreover,

$$\begin{aligned}
\sup_{w\neq 0} \frac{\big(\mathbb{E}_\nu[|\langle\psi,w\rangle|]\big)^2}{w^\mathsf{T}\Sigma_\mu w} &= \sup_{w\neq 0}\Big(\mathbb{E}_\nu\Big[\frac{|\langle\psi,w\rangle|}{\sqrt{w^\mathsf{T}\Sigma_\mu w}}\Big]\Big)^2 \\
&= \sup_{w\neq 0}\Big(\mathbb{E}_\nu\Big[\Big|\Big\langle\Sigma_\mu^{-1/2}\psi,\ \frac{\Sigma_\mu^{1/2}w}{\sqrt{w^\mathsf{T}\Sigma_\mu w}}\Big\rangle\Big|\Big]\Big)^2 \\
&\le \sup_{w\neq 0}\Big(\mathbb{E}_\nu\Big[\|\psi\|_{\Sigma_\mu^{-1}}\cdot\Big\|\frac{\Sigma_\mu^{1/2}w}{\sqrt{w^\mathsf{T}\Sigma_\mu w}}\Big\|_2\Big]\Big)^2 \\
&= \sup_{w\neq 0}\Big(\mathbb{E}_\nu[\|\psi\|_{\Sigma_\mu^{-1}}]\Big)^2 \\
&= \Big(\mathbb{E}_\nu[\|\psi\|_{\Sigma_\mu^{-1}}]\Big)^2,
\end{aligned}$$

Here the inequality is Cauchy–Schwarz, and the penultimate equality uses that the second factor has unit norm. Thus this current form is weaker *than the PEVI-style average local uncertainty. A simpler sufficient condition is*

$$\Sigma_\nu \preceq C\Sigma_\mu, \qquad \Sigma_\nu = \mathbb{E}_\nu[\psi\psi^\mathsf{T}],\ \Sigma_\mu = \mathbb{E}_\mu[\psi\psi^\mathsf{T}].$$

Indeed, for any w,

$$\big(\mathbb{E}_\nu[|\langle \psi, w\rangle|]\big)^2 \le \mathbb{E}_\nu[\langle \psi, w\rangle^2] = w^\mathsf{T}\Sigma_\nu w \le C\, w^\mathsf{T}\Sigma_\mu w.$$

Finally, the raw density ratio bound $\|\nu/\mu\|_\infty \le C$ is an even stronger pointwise sufficient condition.

Algorithm: Pessimistic Policy Selection

Bellman consistency score. For $f, f' \in \mathcal{F}$ and $\pi \in \Pi$, define

$$\mathcal{L}(f', f, \pi; \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n}\big(f'(s_i, a_i) - r_i - \gamma f(s_i', \pi(s_i'))\big)^2,$$

and the *Bellman consistency score*:

$$\mathcal{E}(f, \pi; \mathcal{D}) \;\coloneqq\; \mathcal{L}(f, f, \pi; \mathcal{D}) - \min_{f' \in \mathcal{F}} \mathcal{L}(f', f, \pi; \mathcal{D}). \tag{7}$$

Version space and pessimistic selection. The version space is $\mathcal{F}_{\pi,\varepsilon} \coloneqq \{f \in \mathcal{F} : \mathcal{E}(f, \pi; \mathcal{D}) \le \varepsilon\}$, and BCP selects

$$\widehat{\pi} \;=\; \operatorname*{argmax}_{\pi \in \Pi} \min_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi). \tag{8}$$

BCP: For each π, pessimistically evaluate it as $\min_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi)$; then select the π with the highest pessimistic value.

BCP vs. GOLF. BCP is the offline mirror of GOLF (Lecture 11). GOLF selects the most *optimistic* f in a confidence set; BCP selects the most *pessimistic* evaluation of each policy. Both use Bellman-error-based version sets, replacing pointwise bonuses/penalties with a global selection principle.

Analysis of BCP

The analysis uses four ingredients: (1) show that the version space contains Q^π (so the pessimistic estimate lower-bounds $J(\pi)$), (2) show that every function in the version space has small Bellman error under μ, (3) use a simulation lemma to express $J(\pi) - f(s_0, \pi)$ as an average Bellman error along d_π, and (4) transfer that Bellman error from d_π back to μ via $\mathcal{C}$. Throughout, let $L \coloneqq \log(|\mathcal{F}||\Pi|/\delta)$.

Step 1: Version space containment.

Lemma 8 (Version space containment)**.** *Under Assumptions 2–3, setting* $\varepsilon = \varepsilon_r := c\,V_{\max}^2 L/n$ *for a sufficiently large constant* c*, we have with probability at least* $1-\delta$*:*

$$Q^\pi \in \mathcal{F}_{\pi,\varepsilon_r} \qquad \textit{for all } \pi \in \Pi.$$

Previously established fast-rate ingredient. This is the same squared-loss Bernstein fast-rate argument already used in Lecture 8 ("Fast Rate via Bernstein's Inequality"), which was presented there as the policy-evaluation analogue of the Lecture 7 fitted Q-iteration argument. For a fixed policy π, the score $\mathcal{E}(f,\pi;\mathcal{D})$ is an empirical *excess squared loss*, and Q^π is the population minimizer because $Q^\pi = \mathcal{T}^\pi Q^\pi$ by realizability. The same Y-variable trick, self-bounding variance bound, and Bernstein inequality from Lecture 8 therefore yield the fast rate

$$\mathcal{E}(Q^\pi,\pi;\mathcal{D}) \;\le\; c\,V_{\max}^2\frac{L}{n}$$

simultaneously for all $\pi \in \Pi$ with probability at least $1-\delta$. Hence $Q^\pi \in \mathcal{F}_{\pi,\varepsilon_r}$ for all π. □

Step 2: Bellman error control for version space members.

Lemma 9 (Bellman error bound)**.** *Under Assumptions 2–3, on the same high-probability event, for all* $f \in \mathcal{F}_{\pi,\varepsilon_r}$*:*

$$\mathbb{E}_\mu[(f-\mathcal{T}^\pi f)^2] \;\le\; c_2^2\,V_{\max}^2 L/n.$$

Previously established fast-rate ingredient. Again, this is the same operator-agnostic squared-loss fast-rate template from Lectures 7 and 8. For any fixed $f \in \mathcal{F}_{\pi,\varepsilon_r}$, completeness gives $g = \mathcal{T}^\pi f \in \mathcal{F}$, and the version-space condition says that the empirical excess loss of f relative to the Bellman target g is at most ε_r. Applying the same self-bounding Bernstein argument as in Lecture 8 to this excess loss yields $\mathbb{E}_\mu[(f-\mathcal{T}^\pi f)^2] \le c_2^2\,V_{\max}^2 L/n$. □

Step 3: Bellman error to decision loss. This is exactly the Performance Difference Identity from Lecture 6, specialized to the policy-specific operator $\mathcal{T}^\pi$ and the point-mass initial distribution at s_0.

Lemma 10 (Bellman error to decision loss)**.** *For any* $g : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ *and any policy* π*:*

$$g(s_0,\pi) - J(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{d_\pi}[g(s,a) - (\mathcal{T}^\pi g)(s,a)]\,.$$

Proof. Since $(\mathcal{T}^\pi g)(s,a) = R(s,a) + \gamma\,(\mathcal{P}^\pi g)(s,a)$, we have

$$\frac{1}{1-\gamma}\mathbb{E}_{d_\pi}[g - \mathcal{T}^\pi g] = \frac{1}{1-\gamma}\mathbb{E}_{d_\pi}[g - R - \gamma\,\mathcal{P}^\pi g].$$

Using $d_\pi(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t d_\pi^{(t)}(s,a)$ where $d_\pi^{(t)}$ is the step-t occupancy, this becomes

$$\sum_{t=0}^{\infty}\gamma^t\big(\mathbb{E}_{d_\pi^{(t)}}[g] - \mathbb{E}_{d_\pi^{(t)}}[R] - \gamma\,\mathbb{E}_{d_\pi^{(t+1)}}[g]\big).$$

The g-terms telescope to $\mathbb{E}_{d_\pi^{(0)}}[g] = g(s_0,\pi)$, while the reward terms sum to $J(\pi)$. Hence

$$\frac{1}{1-\gamma}\mathbb{E}_{d_\pi}[g - \mathcal{T}^\pi g] = g(s_0,\pi) - J(\pi).$$

□

Step 4: From pessimism to Bellman error. Compared with the Bellman-error-to-decision-loss lemmas from Lectures 6 and 7, BCP only needs the comparator-side Bellman error: pessimistic selection already replaces the learned-policy-side term.

For a comparator policy π, let

$$f_{\pi,\min} \coloneqq \operatorname*{argmin}_{f\in\mathcal{F}_{\pi,\varepsilon_r}} f(s_0,\pi),$$

which is well-defined on the event of Lemma 8 because $Q^\pi \in \mathcal{F}_{\pi,\varepsilon_r}$.

Lemma 11 (Pessimism reduces suboptimality to Bellman error)**.** *On the event of Lemma 8, for any comparator $\pi \in \Pi$:*

$$J(\pi) - J(\widehat{\pi}) \le \frac{1}{1-\gamma}\mathbb{E}_{d_\pi}\big[\mathcal{T}^\pi f_{\pi,\min} - f_{\pi,\min}\big] \le \frac{1}{1-\gamma}\mathbb{E}_{d_\pi}\big[|f_{\pi,\min} - \mathcal{T}^\pi f_{\pi,\min}|\big].$$

Proof. Since $Q^{\widehat{\pi}} \in \mathcal{F}_{\widehat{\pi},\varepsilon_r}$ (Lemma 8),

$$J(\widehat{\pi}) = Q^{\widehat{\pi}}(s_0,\widehat{\pi}) \ge \min_{f\in\mathcal{F}_{\widehat{\pi},\varepsilon_r}} f(s_0,\widehat{\pi}).$$

By the optimality of $\widehat{\pi}$ in (8),

$$\min_{f\in\mathcal{F}_{\widehat{\pi},\varepsilon_r}} f(s_0,\widehat{\pi}) \ge \min_{f\in\mathcal{F}_{\pi,\varepsilon_r}} f(s_0,\pi) = f_{\pi,\min}(s_0,\pi).$$

Therefore,

$$J(\pi) - J(\widehat{\pi}) \le Q^\pi(s_0,\pi) - f_{\pi,\min}(s_0,\pi).$$

Applying Lemma 10 with $g = f_{\pi,\min}$ gives

$$f_{\pi,\min}(s_0,\pi) - J(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{d_\pi}\big[f_{\pi,\min} - \mathcal{T}^\pi f_{\pi,\min}\big].$$

Rearranging yields

$$J(\pi) - f_{\pi,\min}(s_0,\pi) = \frac{1}{1-\gamma}\mathbb{E}_{d_\pi}\big[\mathcal{T}^\pi f_{\pi,\min} - f_{\pi,\min}\big].$$

Combining with the previous display, we obtain

$$J(\pi) - J(\widehat{\pi}) \le \frac{1}{1-\gamma}\mathbb{E}_{d_\pi}\big[\mathcal{T}^\pi f_{\pi,\min} - f_{\pi,\min}\big].$$

The final inequality follows from the triangle inequality. □

Main Result

Theorem 12 (BCP suboptimality)**.** *Under Assumptions 2–3, let $\widehat{\pi}$ be the output of (8) with $\varepsilon = \varepsilon_r = c\,V_{\max}^2 L/n$ and $L = \log(|\mathcal{F}||\Pi|/\delta)$. For any comparator $\pi \in \Pi$ with $\mathcal{C}(d_\pi;\mu,\mathcal{F},\pi) \le C_2$, with probability at least $1-\delta$:*

$$J(\pi) - J(\widehat{\pi}) \;\le\; O\left(\frac{V_{\max}\sqrt{C_2}}{1-\gamma}\sqrt{\frac{L}{n}}\right).$$

Proof. On the event of Lemma 8, the function $f_{\pi,\min}$ belongs to $\mathcal{F}_{\pi,\varepsilon_r}$, so Lemma 11 gives

$$J(\pi) - J(\widehat{\pi}) \le \frac{1}{1-\gamma}\mathbb{E}_{d_\pi}\big[|f_{\pi,\min} - \mathcal{T}^\pi f_{\pi,\min}|\big].$$

By the definition of $\mathcal{C}$,

$$\mathbb{E}_{d_\pi}\big[|f_{\pi,\min} - \mathcal{T}^\pi f_{\pi,\min}|\big] \le \sqrt{\mathcal{C}(d_\pi;\mu,\mathcal{F},\pi)\cdot\mathbb{E}_\mu[(f_{\pi,\min} - \mathcal{T}^\pi f_{\pi,\min})^2]} \le c_2\sqrt{C_2}\,V_{\max}\sqrt{L/n},$$

where the last step uses Lemma 9. Combining the two displays completes the proof. □

BCP suboptimality (exact realizability + completeness):

$$J(\pi) - J(\widehat{\pi}) \;\le\; O\left(\frac{V_{\max}\sqrt{C_2}}{1-\gamma}\sqrt{\frac{\log(|\mathcal{F}||\Pi|/\delta)}{n}}\right) \qquad \text{when } \mathcal{C}(d_\pi;\mu,\mathcal{F},\pi) \le C_2.$$

Key features of BCP.

- **No explicit density ratio.** BCP requires only the *function-class-aware* distribution shift coefficient $\mathcal{C}$, not the crude density ratio $\|d_\pi/\mu\|_\infty$.
- **Bounded degradation.** Setting $\pi = \pi_b$ (behavior policy) with $\mu = d_{\pi_b}$ gives $\mathcal{C} = 1$, and $J(\pi_b) - J(\widehat{\pi}) \leq \widetilde{O}\big(\frac{V_{\max}}{1-\gamma}\sqrt{1/n}\big)$: BCP is competitive with the data-collecting policy.
- **Competing with $\pi^\star$.** Setting $\pi = \pi^\star$ recovers $J(\pi^\star) - J(\widehat{\pi}) \leq \widetilde{O}\big(\frac{V_{\max}\sqrt{C_2}}{1-\gamma}\sqrt{1/n}\big)$ under $\mathcal{C}(d_{\pi^\star}; \mu, \mathcal{F}, \pi^\star) \leq C_2$.
- **Adaptive bias-variance tradeoff.** The general Theorem 3.1 in Xie et al. (2021) allows approximate realizability ($\varepsilon_{\mathcal{F}} > 0$) and completeness ($\varepsilon_{\mathcal{F},\mathcal{F}} > 0$), and decomposes the bound into on-support and off-support errors; the algorithm automatically adapts to the best decomposition.

A Remark on Computation

The constrained BCP selection (8) is stated as $\arg\max_{\pi\in\Pi} \min_{f\in\mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi)$, which is *information-theoretic* in nature—a direct implementation would search over Π and, for each π, solve a constrained minimization over $\mathcal{F}$. It is worth contrasting this with the online setting.

Online vs. offline computation. The main contrast is not that online RL is impossible while offline RL is easy. Rather, in our current theory line, online methods with general function approximation are primarily *value-based* and centered on the Bellman *optimality* operator $\mathcal{T}$. As discussed in Lecture 8, the $\max_{a' \in \mathcal{A}}$ inside $\mathcal{T}$ makes Bellman-error minimization a moving-target, non-convex problem. By contrast, Bellman-consistent pessimism is naturally paired with more *policy-based* relaxations: for a fixed policy π, the critic uses the policy-specific operator $\mathcal{T}^\pi$, so the inner Bellman-consistency objective is a regression problem rather than an optimization involving the $\max_a$ nonlinearity.
On top of this critic, one can then use standard policy-optimization primitives:

- **PSPI** (Section 4 of Xie et al., 2021) replaces (8) by a regularized objective and updates the actor via policy mirror descent / soft policy iteration (cf. Lecture 10).
- **ATAC** (Cheng et al., 2022) is a later actor–critic style implementation of offline pessimism based on adversarially trained critics.

This computational gap is further amplified by exploration: online optimism must couple value estimation with data collection (cf. GOLF in Lecture 11), whereas offline pessimism works with a fixed dataset and can be implemented through a regression-plus-policy-optimization loop.
The sample-complexity rates currently achieved by these implementable algorithms are not known to match the information-theoretic rate of (8), so we state no specific bound here; see the respective papers for the precise guarantees.

Summary

Summary of offline RL guarantees:

Algorithm	Suboptimality	Key Technique
Naive greedy	Can be $\Omega(H)$ (term (i) uncontrolled)	No pessimism
PEVI (linear)	$\widetilde{O}(dH) \cdot \sum_h \mathbb{E}_{\pi^\star}[\|\|\phi\|\|_{\Lambda^{-1}}]$	Ellipsoidal penalty
BCP (general FA)	$O\big(\frac{V_{\max}\sqrt{C_2}}{1-\gamma}\sqrt{\frac{\log \|\mathcal{F}\|\|\Pi\|}{n}}\big)$	Version space
Lower bound	$\Omega\big(\sum_h \mathbb{E}_{\pi^\star}[\|\|\phi\|\|_{\Lambda^{-1}}]\big)$	Info-theoretic

Online vs. Offline RL:

	Online RL (Lec 11)	**Offline RL (this lecture)**
Data	Adaptive (agent explores)	Fixed dataset (no interaction)
Principle	Optimism	Pessimism
Q-update	$\widehat{Q} = \widehat{\mathcal{T}}\widehat{V} + \text{bonus}$	$\widehat{Q} = \widehat{\mathcal{T}}\widehat{V} - \text{penalty}$
Effect	Explores uncertain regions	Avoids uncertain regions
Perf. depends on	Bonuses along *learned* policy	Penalties along *optimal* policy
General FA	GOLF: global optimism	BCP: global pessimism

Key takeaways.

- Offline RL is fundamentally different from online RL: without the ability to collect new data, the agent cannot explore, and *optimism* can be catastrophically wrong.
- The *pessimism principle* resolves this by penalizing uncertain state-action pairs, ensuring the learned policy avoids regions where the dataset provides insufficient coverage.
- The suboptimality decomposition (Lemma 1) reveals three sources of error: the pessimism-controlled term, the optimism-controlled term, and the optimization error. Pessimism eliminates the pessimism-controlled term, leaving only the optimism-controlled term.
- PEVI achieves *minimax optimal* suboptimality for linear MDPs, with a data-dependent bound that exhibits the *oracle property*: the suboptimality depends only on coverage of the *optimal* policy's trajectory.
- Bellman-consistent pessimism (BCP) extends the pessimism principle to general function approximation via version spaces, paralleling the GOLF algorithm's extension of optimism in the online setting.
- The online-offline duality is remarkably clean: the same mathematical objects (bonus = penalty, confidence set = version space, optimism = pessimism) appear in both settings with only a *sign change*. Understanding one setting deeply illuminates the other.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained

actor critic for offline reinforcement learning. In *International Conference on Machine Learning*, 2022.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096, 2021.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.