Lecture 8: Policy Evaluation and Temporal Difference Learning

In the previous lectures, we developed a thorough understanding of MDPs and their solutions: we proved the existence and uniqueness of $V^\star$ and $Q^\star$ via the Bellman equations (Lectures 1–2), designed efficient planning algorithms such as Value Iteration and Policy Iteration (Lectures 3–4), established the concentration inequalities needed for learning (Lecture 5), analyzed the certainty-equivalence approach for tabular RL (Lecture 6), and studied Fitted Q-Iteration with function approximation in the batch setting (Lecture 7). A common thread through all these methods is that they are *value-based*: they aim to compute or estimate the optimal value function $Q^\star$ or $V^\star$ directly. In this lecture, we ask: *what are the fundamental difficulties of this value-based approach, and how can we build more stable alternatives?*

We will see that minimizing the Bellman error for the optimality operator is inherently **non-convex** due to the $\max$ operator, and that the combination of function approximation, bootstrapping, and off-policy data can lead to **divergence**. These difficulties motivate the study of **policy evaluation**—the problem of estimating Q^π for a *fixed* policy π—which removes the $\max$ operator and yields a clean, convex structure. We develop three increasingly sophisticated approaches: **Monte Carlo** policy evaluation (regressing returns—the simplest method), **Fitted Q-Evaluation** (FQE) with general function approximation (a bootstrapping method using only single-step transitions), and **linear TD algorithms** (TD(0) for on-policy and GTD2 for off-policy learning). Throughout, we work on the Q**-function space** $\mathcal{F} \subset \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, because the downstream application of policy evaluation is **policy optimization** (policy gradient, actor-critic), which requires $Q^\pi(s,a)$ to compute the advantage $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$.

Non-Convexity and Instability of Value-Based Methods

Non-Convexity of Bellman Error Minimization

Recall from Lecture 7 that Fitted Q-Iteration approximates $Q^\star$ using a function class $\mathcal{F}$. A natural objective for finding $Q^\star$ within a parametric class is the **mean squared Bellman error** (MSBE). Consider linear function approximation $Q_w(s,a) = \phi(s,a)^\mathsf{T} w$ with features $\phi(s,a) \in \mathbb{R}^d$ and weight vector $w \in \mathbb{R}^d$. The MSBE under the Bellman *optimality* operator $\mathcal{T}$ is

$$\mathrm{MSBE}(w) := \mathbb{E}_{(s,a)\sim\mu}\big[\big(\phi(s,a)^\mathsf{T} w - (\mathcal{T}Q_w)(s,a)\big)^2\big],$$

where μ is a distribution over $\mathcal{S}\times\mathcal{A}$ and $(\mathcal{T}Q_w)(s,a) = \mathbb{E}[r \mid s,a] + \gamma\mathbb{E}_{s'\sim\mathcal{P}(s,a)}\big[\max_{a'\in\mathcal{A}} \phi(s',a')^\mathsf{T} w\big]$.

Proposition 1 (Non-Convexity of MSBE)**.** *The MSBE objective under the Bellman optimality operator $\mathcal{T}$ is non-convex in w, even for linear function approximation.*

Proof. The Bellman target $(\mathcal{T}Q_w)(s,a)$ contains the term $\max_{a'} \phi(s',a')^\top w$, which is **convex** in w (as a pointwise maximum of linear functions). The Bellman residual $\phi(s,a)^\top w - (\mathcal{T}Q_w)(s,a)$ is therefore the difference of a linear function and a convex function, which is concave. The square of a concave function is generally non-convex, so $\mathrm{MSBE}(w)$ is non-convex. □

Contrast with supervised learning. In standard regression, the target y is fixed and $\min_w \mathbb{E}[(f_w(x)-y)^2]$ is a convex quadratic for linear f_w. In value-based RL, the target $\mathcal{T}Q_w$ depends on w itself through the $\max$ operator, creating the "moving target" problem. This non-convexity means that gradient-based methods may get stuck in local minima, and there is no guarantee of finding the global optimum.

FQI is iteration, not optimization. Beyond non-convexity, direct minimization of MSBE faces a statistical challenge: the operator $\mathcal{T}Q_w$ involves an expectation $\mathbb{E}_{s'}[\cdot]$ *inside* the squared loss, so estimating $\nabla\, \mathrm{MSBE}(w)$ from a single transition (s,a,r,s') yields a biased gradient (the "double sampling" problem). FQI (Lecture 7) avoids both difficulties by solving a *sequence* of regression problems: at iteration k, it fits $f_k \in \mathcal{F}$ to the targets $r_i + \gamma \max_{a'} f_{k-1}(s'_i, a')$. Each regression is a standard supervised learning problem—convex and with unbiased targets. However, FQI does **not** minimize any single objective across iterations: its convergence relies on the *contraction* property of $\mathcal{T}$, not on optimization theory. This makes iteration-based methods potentially unstable—even when the true value function lies in $\mathcal{F}$, the iterates can **diverge** if the function class or data distribution disrupts the contraction (see Theorem 13 for a concrete example in the policy evaluation setting).

Policy Evaluation: A Convex Alternative

The non-convexity of MSBE arises entirely from the $\max$ operator in $\mathcal{T}$. What happens if we remove it?

Key observation. For a **fixed** policy π, the Bellman operator $\mathcal{T}^\pi$ is **affine**. On Q-functions:

$$(\mathcal{T}^\pi Q)(s,a) = R(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \sum_{a'} \pi(a'|s')\, Q(s',a').$$

There is no $\max$ operator. The MSBE under $\mathcal{T}^\pi$ involves fitting to an *affine* target—a fundamentally easier problem. With linear function approximation $Q_w(s,a) = \phi(s,a)^\mathsf{T} w$:

$$\mathrm{MSBE}^\pi(w) = \|(I - \gamma \mathcal{P}^\pi)\Phi w - R\|_\mu^2,$$

which is a **convex quadratic** in w, where $\mathcal{P}^\pi$ is the state-action transition operator under π.

This observation motivates the study of **policy evaluation**—estimating Q^π for a given policy π—as a cleaner, more tractable problem than directly seeking $Q^\star$. Moreover, policy evaluation is not merely a simplification: it is the essential subroutine for **policy optimization** methods, where one alternates between evaluating and improving the current policy.

Preview: The Deadly Triad

Before diving into the theory, we preview a second fundamental difficulty of value-based methods with function approximation.

The deadly triad (Sutton and Barto, 2018): the combination of **function approximation**, **bootstrapping**, and **off-policy learning** can cause learning algorithms to **diverge**, even when the true value function lies within the function class. Any two of these ingredients can coexist safely; we will formalize this and present a concrete 2-state counterexample later in this lecture.

For the remainder of this lecture, we fix a policy π and study how to estimate Q^π using function approximation and data.

Monte Carlo Policy Evaluation

The simplest approach to estimating Q^π is to follow π, collect complete trajectory returns, and perform standard regression. This method uses **no bootstrapping**: the regression targets are unbiased estimates of Q^π.

Setting. Suppose we collect n independent episodes by executing policy π: at each episode, we start from $(s_0, a_0) \sim d_0 \times \pi(\cdot|s_0)$, observe $r_0, s_1, a_1, r_1, s_2, \ldots$, and compute the

discounted return

$$G_t := \sum_{h=0}^{\infty} \gamma^h \, r_{t+h}.$$

Since $\mathbb{E}[G_t \mid s_t, a_t] = Q^\pi(s_t, a_t)$, the return G_t is an **unbiased** estimate of $Q^\pi(s_t, a_t)$, and $|G_t| \le V_{\max} = R_{\max}/(1-\gamma)$.

Dataset. Let $D_{\mathrm{MC}} = \{(s_i, a_i, G_i)\}_{i=1}^n$ be a dataset of i.i.d. samples where $(s_i, a_i) \sim d^\pi$ (the discounted state-action occupancy of π) and G_i is the return starting from (s_i, a_i).

Algorithm 1 Monte Carlo Policy Evaluation

```
Require: Dataset $D_{\mathrm{MC}} = \{(s_i, a_i, G_i)\}_{i=1}^n$, function class $\mathcal{F} \subset \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$
$\widehat{f} \leftarrow \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \big(f(s_i, a_i) - G_i\big)^2$    ▷ least-squares regression
return $\widehat{f}$
```

Theorem 2 (Monte Carlo Policy Evaluation)**.** *Assume realizability:* $Q^\pi \in \mathcal{F}$*, and* $\|f\|_\infty \le V_{\max}$ *for all* $f \in \mathcal{F}$*. Then with probability at least* $1-\delta$*:*

$$\|\widehat{f} - Q^\pi\|_{d^\pi}^2 \le O\left(\frac{V_{\max}^2 \log(|\mathcal{F}|/\delta)}{n}\right).$$

Proof sketch. This is standard well-specified regression. The target G_i satisfies $\mathbb{E}[G_i \mid s_i, a_i] = Q^\pi(s_i, a_i)$ and $|G_i| \le V_{\max}$. Since $Q^\pi \in \mathcal{F}$, the squared-loss decomposition gives $\|\widehat{f} - Q^\pi\|_{d^\pi}^2 = \mathcal{L}_{d^\pi}(\widehat{f}; \mathrm{MC}) - \mathcal{L}_{d^\pi}(Q^\pi; \mathrm{MC})$, where $\mathcal{L}_{d^\pi}$ is the population squared loss. Applying Bernstein's inequality with a union bound over $\mathcal{F}$, and using the optimality of $\widehat{f}$ (which minimizes the empirical loss and $Q^\pi \in \mathcal{F}$), gives the fast rate. □

Properties of Monte Carlo policy evaluation:

- **No bootstrapping:** the targets G_i are unbiased estimates of Q^π that do *not* depend on the function class $\mathcal{F}$. There is no moving target problem.
- **Only realizability is needed:** Bellman completeness is *not* required, since we never apply $\mathcal{T}^\pi$ to any function in $\mathcal{F}$.
- **Always converges:** this is standard supervised learning.
- **Rate $O(V_{\max}/\sqrt{n})$:** note the absence of the $1/(1-\gamma)$ factor that will appear in FQE. This is because MC solves a single regression problem, while FQE accumulates error over K bootstrap iterations.
- **Limitations:** requires **complete trajectories** (to compute G_t) and **on-policy data** (must follow π). In many applications, we only have access to single-step transitions (s, a, r, s'), possibly collected by a different behavior policy.

These limitations motivate a bootstrapping approach: can we achieve good guarantees using *only single-step transitions*, potentially from an off-policy distribution? This leads to Fitted Q-Evaluation.

Fitted Q-Evaluation with General Function Approximation

We now develop a general function approximation theory for policy evaluation on the Q-function space. The framework parallels Lecture 7's analysis of FQI but is *simpler in every dimension*, because the policy evaluation operator $\mathcal{T}^\pi$ is affine (no $\max$).

Setup and Structural Assumptions

We work with a finite MDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma, d_0)$ with $|\mathcal{S}| = S$ states, $|\mathcal{A}| = A$ actions, discount factor $\gamma \in (0, 1)$, and $V_{\max} = R_{\max}/(1-\gamma)$. Fix a policy π. We represent candidate Q-functions using a **function class** $\mathcal{F} \subset \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$—a finite set of functions $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, with $\|f\|_\infty \le V_{\max}$ for all $f \in \mathcal{F}$. The class $\mathcal{F}$ may be **exponentially large**, and its statistical complexity is measured by $\log |\mathcal{F}|$.

Same function class space as FQI. Both FQE and FQI use $\mathcal{F} \subset \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. The *only* difference is the operator: FQI uses the Bellman optimality operator $\mathcal{T}$ (which contains $\max_a$), while FQE uses the policy evaluation operator $\mathcal{T}^\pi$ (which averages over $a' \sim \pi$). This makes the comparison particularly clean.

Definition 1 (Realizability)**.** *We say $\mathcal{F}$ satisfies* ***realizability*** *if $Q^\pi \in \mathcal{F}$.*

The Bellman operator for policy evaluation on Q-functions is

$$(\mathcal{T}^\pi Q)(s,a) = R(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \sum_{a'} \pi(a'|s')\, Q(s',a'). \tag{1}$$

We write $(\mathcal{P}^\pi Q)(s,a) := \sum_{s'} \mathcal{P}(s'|s,a) \sum_{a'} \pi(a'|s')\, Q(s',a')$ for the transition operator on Q-functions, so that $\mathcal{T}^\pi Q = R + \gamma\, \mathcal{P}^\pi Q$.

Definition 2 (Bellman Completeness for $\mathcal{T}^\pi$)**.** *We say $\mathcal{F}$ satisfies **Bellman completeness** for $\mathcal{T}^\pi$ if for all $f \in \mathcal{F}$, $\mathcal{T}^\pi f \in \mathcal{F}$.*

Completeness is easier for $\mathcal{T}^\pi$ than for $\mathcal{T}$. Since $\mathcal{T}^\pi$ is *affine*—$\mathcal{T}^\pi Q = R + \gamma \mathcal{P}^\pi Q$—Bellman completeness requires the function class to be closed under an affine transformation. For the Bellman optimality operator $\mathcal{T}$, completeness requires closure under $\max_a$, which is much more restrictive (it breaks linearity). As a consequence:

- **Linear function classes** $\mathcal{F} = \{(s,a) \mapsto \phi(s,a)^\top w\}$ satisfy completeness for $\mathcal{T}^\pi$ whenever $\mathcal{P}^\pi \Phi$ and R lie in $\mathrm{col}(\Phi)$ (e.g., in linear MDPs). But they may fail completeness for $\mathcal{T}$ because $\max_a \phi(s,a)^\top w$ is generally not linear.
- Any class closed under affine operations (e.g., certain reproducing kernel Hilbert spaces) satisfies completeness for $\mathcal{T}^\pi$.

This is the **first simplification** compared to Lecture 7's FQI analysis.

Batch setting and data distribution. We consider a **batch** setting: we are given a fixed dataset $D = \{(s_i, a_i, r_i, s'_i, a'_i)\}_{i=1}^n$ generated i.i.d. as follows:

$$(s,a) \sim \mu, \quad r \sim R(s,a), \quad s' \sim \mathcal{P}(s,a), \quad a' \sim \pi(\cdot|s'),$$

where $\mu \in \Delta(\mathcal{S} \times \mathcal{A})$ is a **state-action distribution**. Note that (s,a) may be drawn from any distribution μ (possibly off-policy), while a' is always drawn from π (which is known). Since $\mathbb{E}[r + \gamma\, f(s',a') \mid s,a] = (\mathcal{T}^\pi f)(s,a)$, the regression target $r_i + \gamma\, f(s'_i, a'_i)$ is an unbiased estimate of $(\mathcal{T}^\pi f)(s_i,a_i)$ given (s_i,a_i).

Definition 3 (Concentrability Coefficient for Policy Evaluation)**.** *Let $d_h^\pi(s,a) := \Pr[s_h = s, a_h = a \mid s_0 \sim d_0, \pi]$ be the marginal state-action distribution at step h under policy π. The **concentrability coefficient** for policy evaluation is*

$$C^\pi := \max_{h \geq 0} \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{d_h^\pi(s,a)}{\mu(s,a)}.$$

The concentrability coefficient C^π measures how well the data distribution μ covers the state-action distributions that arise along trajectories of π. A key consequence is the **change-of-measure inequality**: for any admissible distribution ν (meaning $\nu = d_h^\pi$ for some $h \geq 0$) and any function g,

$$\|g\|_\nu \leq \sqrt{C^\pi}\, \|g\|_\mu.$$

Single-policy coverage suffices. In the FQI analysis (Lecture 7), the concentrability coefficient C required coverage over distributions d_h^π for all steps h and *all possible policies* π, because the error propagation introduced arbitrary "witness policies" at each step (via the V-to-Q reduction lemma). For policy evaluation, we only evaluate a *single* fixed policy π, so C^π only depends on coverage of one policy's trajectory distribution. This is a dramatically weaker requirement—the **second simplification** compared to FQI.

The Fitted Q-Evaluation Algorithm

Define the **empirical loss** for fitting f to the Bellman backup of f' under $\mathcal{T}^\pi$:

$$\mathcal{L}_D(f; f') \coloneqq \frac{1}{n} \sum_{(s_i, a_i, r_i, s_i', a_i') \in D} \big(f(s_i, a_i) - r_i - \gamma\, f'(s_i', a_i')\big)^2.$$

The regression targets are $y_i = r_i + \gamma\, f'(s_i', a_i')$. Since $a_i' \sim \pi(\cdot|s_i')$ and $s_i' \sim \mathcal{P}(s_i, a_i)$, the conditional mean of y_i given (s_i, a_i) is $(\mathcal{T}^\pi f')(s_i, a_i)$.

Algorithm 2 Fitted Q-Evaluation (FQE)

```
Require: Dataset $D = \{(s_i, a_i, r_i, s_i', a_i')\}_{i=1}^n$, function class $\mathcal{F}$, number of iterations $K$, policy $\pi$
Initialize $f_0 \equiv \mathbf{0}$                                        ▷ assuming $\mathbf{0} \in \mathcal{F}$
for $k = 1, 2, \ldots, K$ do
    $f_k \leftarrow \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \big(f(s_i, a_i) - r_i - \gamma\, f_{k-1}(s_i', a_i')\big)^2$        ▷ least-squares regression
end for
return $f_K$
```

Comparison with FQI. FQE is the policy evaluation analogue of FQI. In FQI (Lecture 7), the regression target is $r_i + \gamma \max_{a'} f_{k-1}(s_i', a')$, which involves the $\max$ operator and depends on the Q-function at *all* actions. In FQE, the target is $r_i + \gamma\, f_{k-1}(s_i', a_i')$ with $a_i' \sim \pi(\cdot|s_i')$—the $\max$ is replaced by sampling from π. Each iteration is a standard least-squares regression problem, and under Bellman completeness, the Bayes-optimal regressor $\mathcal{T}^\pi f_{k-1}$ lies in $\mathcal{F}$.

Population loss. The **population loss** (the expected version of $\mathcal{L}_D$) is

$$\mathcal{L}_\mu(f;f') \coloneqq \mathbb{E}_{(s,a)\sim\mu,\, r,\, s',\, a'\sim\pi(\cdot|s')}\big[(f(s,a)-r-\gamma\, f'(s',a'))^2\big].$$

Since $(\mathcal{T}^\pi f')(s,a) = \mathbb{E}[r+\gamma\, f'(s',a') \mid s,a]$ is the conditional mean of the target given (s,a), the standard bias-variance decomposition gives

$$\mathcal{L}_\mu(f;f') = \|f-\mathcal{T}^\pi f'\|_\mu^2 + \mathcal{L}_\mu(\mathcal{T}^\pi f';f'), \tag{2}$$

where the second term is the irreducible noise, independent of f.

Derivation of (2). Let $Y \coloneqq r+\gamma\, f'(s',a')$ denote the random target. By definition, $\mathbb{E}[Y \mid s,a] = (\mathcal{T}^\pi f')(s,a)$. Insert the conditional mean:

$$f(s,a)-Y = \big(f(s,a)-\mathcal{T}^\pi f'(s,a)\big) + \big(\mathcal{T}^\pi f'(s,a)-Y\big).$$

Squaring and taking the conditional expectation given (s,a):

$$\begin{aligned}&\mathbb{E}\big[(f(s,a)-Y)^2 \mid s,a\big]\\ &= \big(f(s,a)-\mathcal{T}^\pi f'(s,a)\big)^2 + 2\big(f(s,a)-\mathcal{T}^\pi f'(s,a)\big)\underbrace{\mathbb{E}\big[\mathcal{T}^\pi f'(s,a)-Y \mid s,a\big]}_{=0} + \mathbb{E}\big[(\mathcal{T}^\pi f'(s,a)-Y)^2 \mid s,a\big].\end{aligned}$$

The cross term vanishes because $f(s,a)$ and $\mathcal{T}^\pi f'(s,a)$ are deterministic given (s,a), and $\mathbb{E}[Y \mid s,a] = \mathcal{T}^\pi f'(s,a)$. Taking the expectation over $(s,a)\sim\mu$:

$$\mathcal{L}_\mu(f;f') = \underbrace{\|f-\mathcal{T}^\pi f'\|_\mu^2}_{\text{approximation error}} + \underbrace{\mathcal{L}_\mu(\mathcal{T}^\pi f';f')}_{\text{irreducible noise}},$$

where the second term equals $\mathcal{L}_\mu(\mathcal{T}^\pi f';f')$ by substituting $f=\mathcal{T}^\pi f'$ into the definition of $\mathcal{L}_\mu$. □

Analysis of FQE

We now analyze the approximation error $\|f_K - Q^\pi\|_\nu$ of FQE, following the same analytical framework as Lecture 7 but highlighting where the absence of $\max$ simplifies the analysis.

Bellman contraction for $\mathcal{T}^\pi$. The key structural property is that $\mathcal{T}^\pi$ contracts *directly*, without the need for any auxiliary lemma.

Lemma 3 (Bellman Contraction for $\mathcal{T}^\pi$)**.** *For any state-action distribution $\nu \in \Delta(\mathcal{S}\times\mathcal{A})$, define $\mathcal{P}^\pi(\nu)$ as the next-state-action distribution: $(s',a')\sim\mathcal{P}^\pi(\nu)$ means $(s,a)\sim\nu$, $s'\sim\mathcal{P}(s,a)$,*

$a' \sim \pi(\cdot|s')$. *Then*

$$\|\mathcal{T}^\pi f - \mathcal{T}^\pi g\|_\nu \le \gamma \, \|f - g\|_{\mathcal{P}^\pi(\nu)}, \qquad \forall \, f, g \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}.$$

Proof. Since $(\mathcal{T}^\pi f)(s,a) - (\mathcal{T}^\pi g)(s,a) = \gamma \sum_{s',a'} \mathcal{P}(s'|s,a)\, \pi(a'|s')\, (f(s',a') - g(s',a'))$:

$$\begin{aligned}
\|\mathcal{T}^\pi f - \mathcal{T}^\pi g\|_\nu^2 &= \sum_{s,a} \nu(s,a) \left(\gamma \sum_{s',a'} \mathcal{P}(s'|s,a)\, \pi(a'|s')\, (f(s',a') - g(s',a')) \right)^2 \\
&\le \gamma^2 \sum_{s,a} \nu(s,a) \sum_{s',a'} \mathcal{P}(s'|s,a)\, \pi(a'|s')\, (f(s',a') - g(s',a'))^2 \\
&\hspace{20em} \text{(Jensen's inequality)} \\
&= \gamma^2 \, \|f - g\|_{\mathcal{P}^\pi(\nu)}^2. \qquad \square
\end{aligned}$$

No V-to-Q reduction needed. In the FQI analysis (Lecture 7), the analogous contraction step required the V-to-Q reduction lemma (Lecture 7, Lemma 2) to handle the $\max$ operator: $\|V_f - V_{Q^\star}\|_\nu \le \|f - Q^\star\|_{\nu\times\pi_{f,Q^\star}}$, where $\pi_{f,Q^\star}$ is an artificial "witness policy." For $\mathcal{T}^\pi$, there is no $\max$, so the contraction follows directly from Jensen's inequality in two lines. This is the **third simplification** compared to FQI.

Recursive error bound. We now bound $\|f_k - Q^\pi\|_\nu$ for any admissible state-action distribution ν (i.e., $\nu = d_h^\pi$ for some $h \ge 0$). By the triangle inequality:

$$\begin{aligned}
\|f_k - Q^\pi\|_\nu &\le \|f_k - \mathcal{T}^\pi f_{k-1}\|_\nu + \|\mathcal{T}^\pi f_{k-1} - \mathcal{T}^\pi Q^\pi\|_\nu \\
&\le \sqrt{C^\pi}\, \|f_k - \mathcal{T}^\pi f_{k-1}\|_\mu + \gamma \, \|f_{k-1} - Q^\pi\|_{\mathcal{P}^\pi(\nu)},
\end{aligned}$$

where the first term uses the change-of-measure inequality, and the second uses Lemma 3. Since $\mathcal{P}^\pi(d_h^\pi) = d_{h+1}^\pi$ is also admissible, we can apply the same bound recursively. Expanding k times and using $\|f_0 - Q^\pi\|_\nu \le V_{\max}$:

$$\|f_k - Q^\pi\|_\nu \le \sqrt{C^\pi} \sum_{j=0}^{k-1} \gamma^j \, \|f_{k-j} - \mathcal{T}^\pi f_{k-j-1}\|_\mu + \gamma^k \, V_{\max}. \tag{3}$$

Per-iteration error bound. It remains to bound $\|f_k - \mathcal{T}^\pi f_{k-1}\|_\mu$ for each iteration. We use the same uniform deviation framework as Lecture 7.

Assumption 1 (Uniform Deviation)**.** *For the dataset D of size n:*

$$\forall \, f, f' \in \mathcal{F}, \qquad |\mathcal{L}_D(f; f') - \mathcal{L}_\mu(f; f')| \le \varepsilon_{\text{stat}}.$$

This can be justified by Hoeffding's inequality and a union bound over $|\mathcal{F}|^2$ pairs, giving $\varepsilon_{\text{stat}} = O(V_{\max}^2 \sqrt{\log|\mathcal{F}|/n})$ with high probability.

Under Bellman completeness ($\mathcal{T}^\pi f_{k-1} \in \mathcal{F}$) and Assumption 1, the same argument as in Lecture 7 yields:

$$\begin{aligned} \|f_k - \mathcal{T}^\pi f_{k-1}\|_\mu^2 &= \mathcal{L}_\mu(f_k; f_{k-1}) - \mathcal{L}_\mu(\mathcal{T}^\pi f_{k-1}; f_{k-1}) && \text{(by (2))} \\ &\le [\mathcal{L}_D(f_k; f_{k-1}) + \varepsilon_{\text{stat}}] - [\mathcal{L}_D(\mathcal{T}^\pi f_{k-1}; f_{k-1}) - \varepsilon_{\text{stat}}] && \text{(Assumption 1)} \\ &\le 2\varepsilon_{\text{stat}}. && (f_k \text{ minimizes } \mathcal{L}_D \text{ and } \mathcal{T}^\pi f_{k-1} \in \mathcal{F}) \end{aligned}$$

Substituting $\|f_{k-j} - \mathcal{T}^\pi f_{k-j-1}\|_\mu \le \sqrt{2\varepsilon_{\text{stat}}}$ into (3):

Theorem 4 (FQE Guarantee — Slow Rate)**.** *Under realizability, Bellman completeness for $\mathcal{T}^\pi$, bounded concentrability C^π, and Assumption 1, the output f_K of FQE satisfies: for any admissible $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$,*

$$\|f_K - Q^\pi\|_\nu \le \frac{1-\gamma^K}{1-\gamma} \sqrt{2C^\pi\,\varepsilon_{\text{stat}}} + \gamma^K V_{\max}.$$

FQE vs. MC: the price of bootstrapping. Comparing with Theorem 2, FQE's rate has an extra $1/(1-\gamma)$ factor from accumulating error over K bootstrap iterations. However, FQE has two advantages: it only needs *single-step transitions* (not full trajectories), and it can work with *off-policy* data (any μ with bounded C^π). The tradeoff is: MC is statistically better by $1/(1-\gamma)$ but requires more structured data.

Choosing the number of iterations. As in FQI, the bound has two terms: the first grows with K (statistical error accumulation), and the second decays with K (optimization error from initialization). Balancing gives $K \approx \frac{1}{1-\gamma} \log\left(\frac{V_{\max}(1-\gamma)}{\sqrt{2C^\pi \varepsilon_{\text{stat}}}}\right)$.

Fast Rate via Bernstein's Inequality

As in Lecture 7, the slow rate arises from a crude uniform deviation bound. By exploiting the self-bounding variance structure via Bernstein's inequality, we achieve a faster rate.

The Y-variable trick. Define the excess loss random variable:

$$Y(f; f') := \big(f(s,a) - r - \gamma f'(s',a')\big)^2 - \big((\mathcal{T}^\pi f')(s,a) - r - \gamma f'(s',a')\big)^2,$$

where $(s,a,r,s',a') \sim \mu \times R \times \mathcal{P} \times \pi$. The empirical average is $\frac{1}{n}\sum_i Y_i(f; f') = \mathcal{L}_D(f; f') - \mathcal{L}_D(\mathcal{T}^\pi f'; f')$, and the population mean is $\mathbb{E}[Y(f; f')] = \|f - \mathcal{T}^\pi f'\|_\mu^2$.

Lemma 5 (Variance-Mean Relationship). *For any $f, f' \in \mathcal{F}$,* $\mathrm{Var}[Y(f; f')] \leq 4V_{\max}^2 \cdot \mathbb{E}[Y(f; f')]$.

Proof. Using $A^2 - B^2 = (A - B)(A + B)$ with $A = f(s,a) - r - \gamma f'(s', a')$ and $B = (\mathcal{T}^\pi f')(s,a) - r - \gamma f'(s', a')$:

$$Y(f; f') = \big(f(s,a) - (\mathcal{T}^\pi f')(s,a)\big) \cdot \big(f(s,a) + (\mathcal{T}^\pi f')(s,a) - 2r - 2\gamma f'(s', a')\big).$$

Since $|f(s,a) + (\mathcal{T}^\pi f')(s,a) - 2r - 2\gamma f'(s',a')| \leq 4V_{\max}$, we have $\mathrm{Var}[Y] \leq \mathbb{E}[Y^2] \leq 4V_{\max}^2 \cdot \|f - \mathcal{T}^\pi f'\|_\mu^2 = 4V_{\max}^2 \cdot \mathbb{E}[Y]$. □

Applying Bernstein's inequality with a union bound over all $f \in \mathcal{F}$ (exactly as in Lecture 7), using the optimality of $f_k = \mathrm{argmin}_{f \in \mathcal{F}} \mathcal{L}_D(f; f_{k-1})$ and the self-bounding structure (Lemma 5), we obtain:

Fast rate for empirical Bellman update: With probability at least $1-\delta$, for all $k = 1, \ldots, K$:

$$\|f_k - \mathcal{T}^\pi f_{k-1}\|_\mu^2 \leq O\left(\frac{V_{\max}^2 \log(|\mathcal{F}|/\delta)}{n}\right).$$

Comparing with the slow rate: Hoeffding's inequality gave a uniform deviation bound $\varepsilon_{\text{stat}} = O(V_{\max}^2 \sqrt{\log|\mathcal{F}|/n})$, leading to $\|f_k - \mathcal{T}^\pi f_{k-1}\|_\mu^2 \leq 2\varepsilon_{\text{stat}}$. The Bernstein approach yields an improved *fast-rate* uniform deviation:

$$\varepsilon_{\text{stat}}^{\text{fast}} := O\left(\frac{V_{\max}^2 \log(|\mathcal{F}|/\delta)}{n}\right),$$

so that $\|f_k - \mathcal{T}^\pi f_{k-1}\|_\mu^2 \leq \varepsilon_{\text{stat}}^{\text{fast}}$. The improvement is from $1/\sqrt{n}$ to $1/n$ in the per-iteration bound, thanks to the self-bounding variance structure.

Substituting $\|f_{k-j} - \mathcal{T}^\pi f_{k-j-1}\|_\mu \leq \sqrt{\varepsilon_{\text{stat}}^{\text{fast}}}$ into the recursive error bound (3):

Theorem 6 (FQE Guarantee — Fast Rate). *Under realizability, Bellman completeness for $\mathcal{T}^\pi$, bounded concentrability C^π, and n i.i.d. samples, with probability at least $1 - \delta$:*

$$\|f_K - Q^\pi\|_\nu \leq \widetilde{O}\left(\frac{\sqrt{C^\pi}\, V_{\max}}{(1-\gamma)\sqrt{n}}\right),$$

choosing $K = O\big(\frac{1}{1-\gamma}\log n\big)$.

The Bernstein trick is operator-agnostic. The self-bounding variance property (Lemma 5) comes from the squared-loss structure, not from the specific Bellman operator. The same technique applies identically to FQI ($\mathcal{T}$) and FQE ($\mathcal{T}^\pi$). This is a satisfying unifying principle: the *statistical* analysis is modular and operator-agnostic, while the *structural* simplifications (contraction, concentrability) come from properties of the specific operator.

Comparison with FQI

The following table summarizes the simplifications that arise from replacing the Bellman optimality operator $\mathcal{T}$ with the policy evaluation operator $\mathcal{T}^\pi$.

FQE vs. FQI: simplifications from removing the $\max$ operator

Aspect	**FQI (Lecture 7, $\mathcal{T}$)**	**FQE (this lecture, $\mathcal{T}^\pi$)**
Function class	$\mathcal{F} \subset \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$	$\mathcal{F} \subset \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ (**same!**)
Completeness	Closure under $\max_a$ (hard)	Affine closure (easier)
Contraction	Needs V-to-Q reduction lemma	Direct (Jensen only)
Concentrability	All policies' distributions	Single policy π only
Fast rate	$\widetilde{O}\left(\dfrac{\sqrt{C}\,V_{\max}}{(1-\gamma)^2\sqrt{n}}\right)$	$\widetilde{O}\left(\dfrac{\sqrt{C^\pi}\,V_{\max}}{(1-\gamma)\sqrt{n}}\right)$

The additional $1/(1-\gamma)$ factor in the FQI rate comes from converting Q-function error to policy suboptimality via the Performance Difference Lemma. In FQE, the output is $\|f_K - Q^\pi\|_\nu$ directly.

Specialization: Linear Function Approximation and TD Learning

When the state-action space has known structure, a natural choice is **linear function approximation**: $\mathcal{F} = \{(s,a) \mapsto \phi(s,a)^\top w : w \in \mathbb{R}^d\}$. In this section, we specialize the general theory to the linear setting, where richer structural results are available, and introduce two foundational algorithms: **TD(0)** for on-policy learning and **GTD2** for off-policy learning.

Linear Q-Function Approximation

Definition 4 (Linear Q-Function Approximation)**.** *We approximate Q^π using a linear model:*

- *A **feature map** $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ with feature matrix $\Phi \in \mathbb{R}^{(S \cdot A) \times d}$, where $\phi(s,a) \in \mathbb{R}^d$ is the feature vector for state-action pair (s,a), and $d \ll S \cdot A$. We assume Φ has full column rank.*
- *A **weight vector** $w \in \mathbb{R}^d$, so $Q_w(s,a) = \phi(s,a)^\top w$.*

We write Φw for the $(S \cdot A)$-dimensional vector $[\phi(s,a)^\top w]_{(s,a) \in \mathcal{S} \times \mathcal{A}}$.

We define the **weighted L_2 norm** $\|Q\|_\mu := (Q^\top D_\mu Q)^{1/2}$ for $D_\mu = \operatorname{diag}(\mu) \in \mathbb{R}^{(S \cdot A) \times (S \cdot A)}$, and the **projection operator** $\Pi_\mu Q := \operatorname{argmin}_{\Phi w} \|\Phi w - Q\|_\mu = \Phi(\Phi^\top D_\mu \Phi)^{-1} \Phi^\top D_\mu Q$. The projection Π_μ is linear, idempotent, self-adjoint under $\langle \cdot, \cdot \rangle_\mu$, and satisfies the non-expansion property $\|\Pi_\mu Q\|_\mu \le \|Q\|_\mu$. A useful matrix identity: for any $g \in \mathbb{R}^{S \cdot A}$,

$$\mathbb{E}_{(s,a) \sim \mu}\big[g(s,a)\,\phi(s,a)\big] = \sum_{s,a} \mu(s,a)\, g(s,a)\, \phi(s,a) = \Phi^\top D_\mu\, g.$$

The Projected Bellman Equation

Since Q^π generally does not lie in $\operatorname{col}(\Phi)$, we seek the best approximation within the linear subspace. A natural approach is the **projected Bellman equation**: find $w^\star \in \mathbb{R}^d$ such that

$$\Phi w^\star = \Pi_\mu \mathcal{T}^\pi(\Phi w^\star) = \Pi_\mu\big(R + \gamma \mathcal{P}^\pi \Phi w^\star\big). \tag{4}$$

Projected Bellman Equation:

$$\Phi w^\star = \Pi_\mu \mathcal{T}^\pi(\Phi w^\star).$$

Does a solution exist, and is it unique? Recall the **discounted state-action occupancy distribution**:

$$d^\pi(s,a) := (1-\gamma) \sum_{h=0}^{\infty} \gamma^h \Pr(s_h = s, a_h = a \mid s_0 \sim d_0, \pi),$$

which satisfies the **flow balance equation**: $d^\pi = (1-\gamma)\, d_0 \otimes \pi + \gamma\, (\mathcal{P}^\pi)^\top d^\pi$, where $(d_0 \otimes \pi)(s,a) = d_0(s)\, \pi(a|s)$.

Theorem 7 (Contraction of the Projected Bellman Operator)**.** *Let $\mu = d^\pi$. Then $\Pi_{d^\pi} \mathcal{T}^\pi$ is a*

$\sqrt{\gamma}$-contraction under $\|\cdot\|_{d^\pi}$:

$$\|\Pi_{d^\pi}\mathcal{T}^\pi Q - \Pi_{d^\pi}\mathcal{T}^\pi Q'\|_{d^\pi} \le \sqrt{\gamma}\,\|Q - Q'\|_{d^\pi}, \qquad \forall\, Q, Q' \in \mathbb{R}^{S\cdot A}.$$

Consequently, by the Banach fixed point theorem, there exists a unique $w^\star$ satisfying (4) with $\mu = d^\pi$.

Proof. By the general contraction (Lemma 3), $\|\mathcal{T}^\pi Q - \mathcal{T}^\pi Q'\|_{d^\pi} \le \gamma\,\|Q - Q'\|_{\mathcal{P}^\pi(d^\pi)}$. Now, $\mathcal{P}^\pi(d^\pi)(s', a') = [(\mathcal{P}^\pi)^\top d^\pi](s', a')$. By the flow balance equation, $(\mathcal{P}^\pi)^\top d^\pi = (d^\pi - (1-\gamma)d_0 \otimes \pi)/\gamma$, so $\mathcal{P}^\pi(d^\pi)(s', a') \le d^\pi(s', a')/\gamma$. Therefore:

$$\|Q - Q'\|^2_{\mathcal{P}^\pi(d^\pi)} \le \frac{1}{\gamma}\,\|Q - Q'\|^2_{d^\pi},$$

giving $\|\mathcal{T}^\pi Q - \mathcal{T}^\pi Q'\|_{d^\pi} \le \gamma \cdot \gamma^{-1/2}\,\|Q - Q'\|_{d^\pi} = \sqrt{\gamma}\,\|Q - Q'\|_{d^\pi}$. Since Π_{d^π} is a non-expansion (Pythagorean theorem), composing gives $\|\Pi_{d^\pi}\mathcal{T}^\pi Q - \Pi_{d^\pi}\mathcal{T}^\pi Q'\|_{d^\pi} \le \sqrt{\gamma}\,\|Q - Q'\|_{d^\pi}$. □

The critical role of d^π. The flow balance equation $(\mathcal{P}^\pi)^\top d^\pi = (d^\pi - (1-\gamma)d_0 \otimes \pi)/\gamma$ strengthens the general contraction rate from γ (Lemma 3) to $\sqrt{\gamma}$, by bounding $\mathcal{P}^\pi(d^\pi)(s', a') \le d^\pi(s', a')/\gamma$. For an arbitrary distribution $\mu \ne d^\pi$, this bound fails, and $\Pi_\mu \mathcal{T}^\pi$ may **not be a contraction**—the root cause of the deadly triad.

Approximation Quality of the Projected Bellman Fixed Point

Theorem 8 (Approximation Error)**.** *Let $w^\star$ be the unique solution to the projected Bellman equation (4) with $\mu = d^\pi$. Then*

$$\|Q^\pi - \Phi w^\star\|_{d^\pi} \le \frac{1}{\sqrt{1-\gamma}}\,\|Q^\pi - \Pi_{d^\pi} Q^\pi\|_{d^\pi}.$$

Proof. By the Pythagorean theorem, $\|Q^\pi - \Phi w^\star\|^2_{d^\pi} = \|Q^\pi - \Pi_{d^\pi} Q^\pi\|^2_{d^\pi} + \|\Pi_{d^\pi} Q^\pi - \Phi w^\star\|^2_{d^\pi}$. For the second term, using $\Phi w^\star = \Pi_{d^\pi}\mathcal{T}^\pi(\Phi w^\star)$, $Q^\pi = \mathcal{T}^\pi Q^\pi$, and the $\sqrt{\gamma}$-contraction (Theorem 7):

$$\|\Pi_{d^\pi} Q^\pi - \Phi w^\star\|_{d^\pi} = \|\Pi_{d^\pi}\mathcal{T}^\pi Q^\pi - \Pi_{d^\pi}\mathcal{T}^\pi(\Phi w^\star)\|_{d^\pi} \le \sqrt{\gamma}\,\|Q^\pi - \Phi w^\star\|_{d^\pi}.$$

Substituting back: $(1-\gamma)\,\|Q^\pi - \Phi w^\star\|^2_{d^\pi} \le \|Q^\pi - \Pi_{d^\pi} Q^\pi\|^2_{d^\pi}$. □

The price of bootstrapping. The term $\|Q^\pi - \Pi_{d^\pi} Q^\pi\|_{d^\pi}$ is the *best possible* approximation error within $\mathrm{col}(\Phi)$. If $Q^\pi \in \mathrm{col}(\Phi)$ (realizability), then $\Phi w^\star = Q^\pi$ exactly. The factor $1/\sqrt{1-\gamma}$ is the price of solving a projected fixed-point equation rather than directly computing the best approximation.

TD(0) for On-Policy Q-Evaluation

We now turn to the algorithmic question: *how do we find $w^\star$ from data in an online fashion?* The answer is **Temporal Difference (TD) learning**[1], a foundational algorithm that learns $w^\star$ from a single stream of on-policy experience.

Consider the **online** setting: an agent follows policy π and observes $s_0, a_0, r_0, s_1, a_1, r_1, \ldots$ where $a_t \sim \pi(\cdot|s_t)$. The **temporal difference error** at time t is

$$\delta_t(w) := r_t + \gamma\, \phi(s_{t+1}, a_{t+1})^\top w - \phi(s_t, a_t)^\top w,$$

where $a_{t+1} \sim \pi(\cdot|s_{t+1})$.

Algorithm 3 TD(0) for Q-Function Evaluation (On-Policy)

```
Require: Policy $\pi$, feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, step sizes $\{\alpha_t\}_{t \geq 0}$
Initialize $w_0 \in \mathbb{R}^d$ arbitrarily
Sample $s_0 \sim d_0$, $a_0 \sim \pi(\cdot|s_0)$
for $t = 0, 1, 2, \ldots$ do
    Observe reward $r_t$ and next state $s_{t+1}$
    Sample $a_{t+1} \sim \pi(\cdot|s_{t+1})$
    $\delta_t \leftarrow r_t + \gamma\, \phi(s_{t+1}, a_{t+1})^\top w_t - \phi(s_t, a_t)^\top w_t$    ▷ TD error
    $w_{t+1} \leftarrow w_t + \alpha_t\, \delta_t\, \phi(s_t, a_t)$    ▷ semi-gradient update
end for
```

Why "semi-gradient". The TD update $w_{t+1} = w_t + \alpha_t\, \delta_t\, \phi(s_t, a_t)$ resembles a stochastic gradient step, but it is *not* the gradient of any fixed objective. A stochastic gradient of the mean squared Bellman *residual* $\mathbb{E}[(Q_w(s,a) - r - \gamma Q_w(s', a'))^2]$ would yield $\delta_t \cdot (\phi(s_t, a_t) - \gamma\, \phi(s_{t+1}, a_{t+1}))$, but TD(0) uses $\delta_t \cdot \phi(s_t, a_t)$, dropping the $-\gamma\, \phi(s_{t+1}, a_{t+1})$ term. Note that the mean squared Bellman residual differs from the true MSBE $\mathbb{E}[(\Phi w - \mathcal{T}^\pi \Phi w)^2]$, which has an expectation *inside* the square; estimating $\nabla\,\mathrm{MSBE}$ from a single sample is biased because (s_{t+1}, a_{t+1}) would appear in both δ_t and the gradient direction (the "double sampling" problem).

Theorem 9 (TD Fixed Point = Projected Bellman Fixed Point)**.** *The expected TD(0) update*

[1] Due to Sutton (1988).

satisfies $\mathbb{E}_{(s,a)\sim d^\pi,r,s',a'}[\delta_t(w)\,\phi(s_t,a_t)] = \Phi^\mathsf{T} D_{d^\pi}(\mathcal{T}^\pi(\Phi w) - \Phi w)$*. This is zero if and only if* Φw *solves the projected Bellman equation* (4) *with* $\mu = d^\pi$*.*

Proof. $\mathbb{E}[\delta_t(w) \mid s_t = s, a_t = a] = R(s,a) + \gamma(\mathcal{P}^\pi \Phi w)(s,a) - \phi(s,a)^\mathsf{T} w = (\mathcal{T}^\pi \Phi w)(s,a) - \phi(s,a)^\mathsf{T} w$. So $\mathbb{E}_{(s,a)\sim d^\pi}[\delta_t(w)\,\phi(s,a)] = \Phi^\mathsf{T} D_{d^\pi}(\mathcal{T}^\pi(\Phi w) - \Phi w)$. Setting this to zero gives $\Phi^\mathsf{T} D_{d^\pi}\mathcal{T}^\pi(\Phi w^\star) = \Phi^\mathsf{T} D_{d^\pi}\Phi w^\star$. Left-multiplying by $(\Phi^\mathsf{T} D_{d^\pi}\Phi)^{-1}$ and then by Φ yields $\Phi w^\star = \Pi_{d^\pi}\mathcal{T}^\pi(\Phi w^\star)$. □

The linear system formulation. Define:

$$A := \Phi^\mathsf{T} D_{d^\pi}(I - \gamma\mathcal{P}^\pi)\Phi \;\in\; \mathbb{R}^{d\times d}, \tag{5}$$

$$b := \Phi^\mathsf{T} D_{d^\pi} R \;\in\; \mathbb{R}^{d}. \tag{6}$$

The TD fixed point satisfies $Aw^\star = b$, i.e., $w^\star = A^{-1}b$. The matrix A is positive definite: for any $w \neq 0$ with $x = \Phi w \neq 0$, $w^\mathsf{T} A w = \|x\|_{d^\pi}^2 - \gamma\langle x, \mathcal{P}^\pi x\rangle_{d^\pi} \geq (1-\sqrt{\gamma})\,\|x\|_{d^\pi}^2 > 0$, where we used $|\langle x, \mathcal{P}^\pi x\rangle_{d^\pi}| \leq \|x\|_{d^\pi}\|\mathcal{P}^\pi x\|_{d^\pi} \leq \|x\|_{d^\pi}^2/\sqrt{\gamma}$ (Cauchy–Schwarz and the bound from the contraction proof).

Finite-Sample Convergence of TD(0)

For i.i.d. sampling from d^π with bounded features ($\|\phi(s,a)\|_2 \leq 1$) and rewards ($R \in [0, R_{\max}]$), the following convergence guarantee holds.[2]

Theorem 10 (Finite-Sample Convergence of TD(0))**.** *With constant step size* $\alpha \leq \omega/\|A\|_{\mathrm{op}}^2$*, where* $\omega := \lambda_{\min}(\frac{A+A^\mathsf{T}}{2}) \geq (1-\sqrt{\gamma})\,\lambda_{\min}(\Phi^\mathsf{T} D_{d^\pi}\Phi)$*:*

$$\mathbb{E}\big[\|w_T - w^\star\|_2^2\big] \leq (1-\alpha\omega)^T\,\|w_0 - w^\star\|_2^2 + \frac{2\alpha\,\sigma^2}{\omega},$$

where $\sigma^2 := \mathbb{E}[\|\delta_t(w^\star)\phi(s_t,a_t)\|_2^2] \leq V_{\max}^2$ *is the variance at the fixed point.*

The proof follows from standard linear stochastic approximation analysis; see Bhandari et al. (2018) for details. The bound exhibits a fundamental **bias-variance tradeoff**: the first term (bias) decays geometrically, while the second term (variance floor) is $O(\alpha\sigma^2/\omega)$. Smaller α reduces the floor but slows convergence.

Corollary 11 (Sample Complexity)**.** *To achieve* $\mathbb{E}[\|w_T - w^\star\|_2^2] \leq \varepsilon$*, set* $\alpha = \Theta(\varepsilon\omega/\sigma^2)$ *and run for* $T = \widetilde{O}(\sigma^2/(\varepsilon\omega^2))$ *iterations. The sample complexity scales as* $\widetilde{O}\big(\sigma^2/(\varepsilon(1-\sqrt{\gamma})^2\lambda_{\min}(\Sigma)^2)\big)$*, where* $\Sigma = \Phi^\mathsf{T} D_{d^\pi}\Phi$*.*

[2] The extension to Markovian sampling requires bounding the mixing time; see Bhandari et al. (2018).

GTD2 for Off-Policy Q-Evaluation

The convergence guarantee for TD(0) relies on on-policy sampling ($\mu = d^\pi$), which ensures $\Pi_{d^\pi}\mathcal{T}^\pi$ is a contraction (Theorem 7). For general off-policy μ, this may fail. In many applications, we only have access to data from a **behavior policy** $b \neq \pi$. As we will see in the Deadly Triad section, naive semi-gradient TD can **diverge** under off-policy sampling. Can we design a convergent off-policy algorithm?

The MSPBE objective. The key idea (Sutton et al., 2009) is to minimize the **mean squared projected Bellman error** (MSPBE):

$$\mathrm{MSPBE}(w) := \|\Phi w - \Pi_\mu \mathcal{T}^\pi(\Phi w)\|_\mu^2, \tag{7}$$

where μ is the data distribution (from the behavior policy b). Unlike the semi-gradient TD update, the gradient of MSPBE yields a *true* gradient descent algorithm that converges regardless of the sampling distribution μ.

Matrix formulation. Define the off-policy analogues of the TD(0) matrices (5)–(6):

$$A_\mu := \Phi^\top D_\mu (I - \gamma \mathcal{P}^\pi)\Phi \;\in\; \mathbb{R}^{d\times d}, \tag{8}$$

$$b_\mu := \Phi^\top D_\mu R \;\in\; \mathbb{R}^d, \tag{9}$$

$$C_\mu := \Phi^\top D_\mu \Phi \;\in\; \mathbb{R}^{d\times d}. \tag{10}$$

Using the identity $\Phi^\top D_\mu(\mathcal{T}^\pi\Phi w - \Phi w) = b_\mu - A_\mu w$, the MSPBE (7) becomes

$$\mathrm{MSPBE}(w) = (b_\mu - A_\mu w)^\top C_\mu^{-1}(b_\mu - A_\mu w). \tag{11}$$

This is a **convex quadratic** in w with minimizer $w_\mu^\star = A_\mu^{-1} b_\mu$ (assuming A_μ is invertible), which is the projected Bellman fixed point under μ: $\Phi w_\mu^\star = \Pi_\mu \mathcal{T}^\pi(\Phi w_\mu^\star)$. When $\mu = d^\pi$, we recover the on-policy TD fixed point: $A_{d^\pi} = A$, $b_{d^\pi} = b$, so $w_\mu^\star = w^\star = A^{-1}b$.

Why not run semi-gradient TD off-policy?. One might attempt to use the same semi-gradient update $w_{t+1} = w_t + \alpha_t \delta_t \phi(s_t, a_t)$ with off-policy data from μ. The expected update direction is $b_\mu - A_\mu w$, making this a linear stochastic approximation with matrix A_μ. For convergence, we need A_μ to be positive definite. When $\mu = d^\pi$, we proved $A = A_{d^\pi}$ is positive definite using the flow balance equation. For general off-policy μ, A_μ **need not be positive definite**, and the semi-gradient iteration can diverge (see the deadly triad example below). In contrast, the MSPBE (11) is always convex regardless of the spectrum of A_μ, so gradient descent on MSPBE always converges.

The double sampling challenge. Computing $\nabla\mathrm{MSPBE}(w)$ from samples is nontrivial: a naive stochastic gradient requires two *independent* samples of the next state to avoid bias (the "double sampling" problem). GTD2 circumvents this by introducing **auxiliary weights** $\theta \in \mathbb{R}^d$ via a saddle-point reformulation.

Algorithm 4 GTD2 for Off-Policy Q-Evaluation (Sutton et al., 2009)

```
Require: Behavior policy $b$, target policy $\pi$, feature map $\phi$, step sizes $\{\alpha_t, \beta_t\}$
Initialize $w_0, \theta_0 \in \mathbb{R}^d$ arbitrarily
for $t = 0, 1, 2, \dots$ do
    Observe $(s_t, a_t)$ from $b$, reward $r_t$, next state $s_{t+1}$
    Sample $a'_{t+1} \sim \pi(\cdot|s_{t+1})$                                        ▷ next action from target policy $\pi$
    $\delta_t \leftarrow r_t + \gamma\,\phi(s_{t+1}, a'_{t+1})^\top w_t - \phi(s_t, a_t)^\top w_t$           ▷ TD error
    $\theta_{t+1} \leftarrow \theta_t + \beta_t\,(\delta_t - \phi(s_t, a_t)^\top \theta_t)\,\phi(s_t, a_t)$          ▷ auxiliary update
    $w_{t+1} \leftarrow w_t + \alpha_t\,(\phi(s_t, a_t) - \gamma\,\phi(s_{t+1}, a'_{t+1}))\,(\phi(s_t, a_t)^\top \theta_t)$   ▷ primary update
end for
```

Key properties of GTD2:

- **True gradient descent:** GTD2 performs stochastic gradient descent on MSPBE (11) via a primal-dual (saddle-point) reformulation. The auxiliary weight θ tracks the "dual variable" $\theta \approx C_\mu^{-1}(b_\mu - A_\mu w)$.
- **Off-policy convergence:** Under standard two-timescale stochastic approximation conditions ($\beta_t \gg \alpha_t \to 0$, $\sum \alpha_t = \sum \beta_t = \infty$, $\sum \alpha_t^2, \sum \beta_t^2 < \infty$), GTD2 converges to $w_\mu^\star = A_\mu^{-1} b_\mu$, the projected Bellman fixed point under μ, regardless of whether $\mu = d^\pi$ or not.
- **Cost:** requires maintaining the auxiliary weights θ (doubling the parameter count) and a slower two-timescale learning rate schedule.

Connection to the General Theory

The linear setting reveals a deeper connection between FQE, TD(0), and GTD2 through the general FQE population loss $\mathcal{L}_\mu$.

Population-level FQE as projected Bellman iteration. With the linear class $\mathcal{F} = \{Q_w = \Phi w : w \in \mathbb{R}^d\}$, each FQE iteration at the population level (minimizing $\mathcal{L}_\mu(f; f_{k-1})$ exactly over $f \in \mathcal{F}$) computes

$$Q_{w_k} = \operatorname*{argmin}_{\Phi w} \underbrace{\|Q_w - \mathcal{T}^\pi Q_{w_{k-1}}\|_\mu^2 + \sigma^2(w_{k-1})}_{\mathcal{L}_\mu(Q_w; Q_{w_{k-1}})} \; = \; \Pi_\mu \mathcal{T}^\pi Q_{w_{k-1}},$$

using the population loss decomposition (2) (the noise $\sigma^2(w_{k-1})$ is independent of w). This is **projected Bellman iteration**: $Q_{w_k} = \Pi_\mu \mathcal{T}^\pi Q_{w_{k-1}}$. Its fixed point is the projected Bellman equation (4), and the contraction of $\Pi_{d^\pi}\mathcal{T}^\pi$ (Theorem 7) is a specialization of the general contraction (Lemma 3), strengthened from γ to $\sqrt{\gamma}$ by the flow balance equation. TD(0) finds this fixed point via on-policy stochastic approximation (Theorem 9).

MSPBE as a minimax FQE loss. The connection to GTD2 is more subtle: the MSPBE (7) can be expressed as a *gap* in the FQE population loss.

Proposition 12 (MSPBE as an FQE Loss Gap)**.** *For any $w \in \mathbb{R}^d$:*

$$\mathrm{MSPBE}(w) = \mathcal{L}_\mu(Q_w;\, Q_w) \;-\; \min_{\theta\in\mathbb{R}^d} \mathcal{L}_\mu(Q_\theta;\, Q_w).$$

Proof. By the population loss decomposition (2), $\mathcal{L}_\mu(f;\, Q_w) = \|f - \mathcal{T}^\pi Q_w\|_\mu^2 + \sigma^2(w)$ for any f, where $\sigma^2(w) \coloneqq \mathcal{L}_\mu(\mathcal{T}^\pi Q_w;\, Q_w)$ does not depend on f. The noise cancels in the difference:

$$\begin{aligned}\mathcal{L}_\mu(Q_w;\, Q_w) - \min_\theta \mathcal{L}_\mu(Q_\theta;\, Q_w) &= \|Q_w - \mathcal{T}^\pi Q_w\|_\mu^2 - \min_\theta \|Q_\theta - \mathcal{T}^\pi Q_w\|_\mu^2 \\ &= \|Q_w - \mathcal{T}^\pi Q_w\|_\mu^2 - \|\mathcal{T}^\pi Q_w - \Pi_\mu \mathcal{T}^\pi Q_w\|_\mu^2.\end{aligned}$$

Since $Q_w \in \mathrm{col}(\Phi)$ and $\mathcal{T}^\pi Q_w - \Pi_\mu \mathcal{T}^\pi Q_w \perp \mathrm{col}(\Phi)$ under $\langle\cdot,\cdot\rangle_\mu$, the Pythagorean theorem gives

$$\|Q_w - \mathcal{T}^\pi Q_w\|_\mu^2 = \underbrace{\|Q_w - \Pi_\mu \mathcal{T}^\pi Q_w\|_\mu^2}_{\mathrm{MSPBE}(w)} + \|\mathcal{T}^\pi Q_w - \Pi_\mu \mathcal{T}^\pi Q_w\|_\mu^2. \qquad \square$$

Interpretation and noise cancellation. The MSPBE measures how much the FQE loss *decreases* when we fit a fresh Q_θ to the Bellman backup of Q_w, compared to evaluating the loss at Q_w itself. In sample form:

$$\mathrm{MSPBE}(w) \approx \frac{1}{n}\sum_{i=1}^n \Big[\big(Q_w(s_i,a_i) - y_i\big)^2 - \big(Q_{\widehat{\theta}}(s_i,a_i) - y_i\big)^2\Big], \qquad y_i = r_i + \gamma\, Q_w(s_i', a_i'),$$

where $\widehat{\theta} = \mathrm{argmin}_\theta\, \mathcal{L}_D(Q_\theta; Q_w)$. Both terms share the **same target** y_i, so the irreducible noise $\sigma^2(w)$ cancels exactly—this is the key to avoiding the double sampling problem.

Saddle-point reformulation and GTD2. Since $-\min_\theta = \max_\theta(-\cdot)$, Proposition 12 yields a minimax problem. Using the variational identity $\|\Pi_\mu v\|_\mu^2 = \max_\theta\,[2\langle\Phi\theta,\, v\rangle_\mu - \|\Phi\theta\|_\mu^2]$ with $v = \mathcal{T}^\pi Q_w - Q_w$ (noting $\Pi_\mu v = \Pi_\mu \mathcal{T}^\pi Q_w - Q_w$ since $Q_w \in \mathrm{col}(\Phi)$):

Saddle-point formulation of MSPBE minimization:

$$\min_w \mathrm{MSPBE}(w) \;=\; \min_w \max_\theta \; \Big[2\,\mathbb{E}_\mu\big[\phi(s,a)^\top\theta \;\cdot\; \delta_w\big] - \mathbb{E}_\mu\big[(\phi(s,a)^\top\theta)^2\big]\Big],$$

where $\delta_w := r + \gamma\,\phi(s',a')^\top w - \phi(s,a)^\top w$ is the TD error.

Each expectation can be estimated unbiasedly from a single transition (s,a,r,s',a'), because θ and w are held fixed during sampling—the saddle-point *decouples* the two appearances of the next-state expectation that cause the double sampling problem. The GTD2 updates (Algorithm 4) follow directly: the auxiliary update (line 6) performs stochastic gradient *ascent* on θ, and the primary update (line 7) performs stochastic gradient *descent* on w.

The Deadly Triad

We now return to the instability question previewed at the beginning of this lecture. Our convergence theory for TD(0) (Theorem 10) crucially relies on on-policy sampling ($\mu = d^\pi$), which ensures the positive definiteness of A and the contraction of $\Pi_{d^\pi}\mathcal{T}^\pi$ (Theorem 7). What happens when we use off-policy data?

Three Dangerous Ingredients

The *deadly triad* (Sutton and Barto, 2018) refers to the simultaneous use of:

1. **Function approximation** — representing value functions parametrically.
2. **Bootstrapping** — using current estimates as regression targets (as in TD).
3. **Off-policy learning** — learning about a target policy π from data generated by a different behavior policy b.

Ingredients present	**Example method**	**Converges?**
FA + Bootstrap (**on-policy**)	On-policy TD(0)	✓ (Theorem 10)
FA + Off-policy (**no bootstrap**)	Off-policy Monte Carlo	✓ (standard regression)
Bootstrap + Off-policy (**tabular**)	Tabular off-policy TD	✓ (finite state space)
FA + Bootstrap + Off-policy	Off-policy TD with FA	✗ (can diverge!)

Removing any one ingredient prevents divergence.

- **Without function approximation** (tabular): the projection Π_μ is the identity, and $\mathcal{T}^\pi$ is a γ-contraction under $\|\cdot\|_\infty$ regardless of μ (Lecture 2).
- **Without bootstrapping** (Monte Carlo): we use complete returns $G_t = \sum_{h=0}^{\infty} \gamma^h r_{t+h}$ as targets. This is standard supervised learning with unbiased targets, which always converges (Theorem 2).
- **Without off-policy** (on-policy TD): convergence is guaranteed by our theory (Theorem 10), since the distribution d^π ensures the contraction property.

A Simple Divergence Example

We now present a minimal example where projected Bellman iteration—the population-level version of FQE—**diverges** when the projection distribution differs from the on-policy distribution.[3]

The MDP. Consider an MDP with 2 states $\mathcal{S} = \{1, 2\}$ and a single action (so there is only one policy π):

- **Transitions**: state 1 transitions to state 2; state 2 transitions to itself (absorbing).
- **Rewards**: $R(s, a) = 0$ for all s. Thus $V^\pi(s) = 0$ for all s.

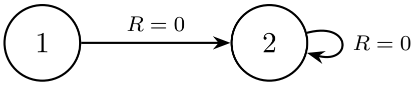

Figure 1: A 2-state MDP for the divergence example. State 2 is absorbing.

Linear function approximation. We use a one-dimensional feature: $\phi(1) = 1$ and $\phi(2) = 2$, so the approximate value function is $V_w(s) = \phi(s) \cdot w$ for a scalar parameter $w \in \mathbb{R}$. The true value $V^\pi = 0$ is realizable at $w^\star = 0$.

Projected Bellman iteration under $\mu = \text{Uniform}$. Suppose we run projected Bellman iteration (equivalently, population-level FQE) with the *uniform* distribution $\mu = (1/2, 1/2)$. Starting from parameter w_k:

[3] This example is closely related to the counterexamples in Tsitsiklis and Van Roy (1997) and Bertsekas and Tsitsiklis (1996). A more complex 7-state example due to Baird (1995) demonstrates the same phenomenon for semi-gradient TD.

Step 1: Bellman backup.

$$(\mathcal{T}^\pi V_{w_k})(1) = 0 + \gamma\, V_{w_k}(2) = 2\gamma w_k, \qquad (\mathcal{T}^\pi V_{w_k})(2) = 0 + \gamma\, V_{w_k}(2) = 2\gamma w_k.$$

Step 2: Least-squares projection onto $\mathcal{F} = \{\phi \cdot w : w \in \mathbb{R}\}$.

$$w_{k+1} = \operatorname*{argmin}_w \; \frac{1}{2}\Big[(w - 2\gamma w_k)^2 + (2w - 2\gamma w_k)^2\Big].$$

Setting the derivative to zero: $5w - 6\gamma w_k = 0$, giving

$$\boxed{w_{k+1} = \frac{6\gamma}{5}\, w_k.}$$

Theorem 13 (Divergence of Projected Bellman Iteration)**.** *For $\gamma > 5/6 \approx 0.833$, projected Bellman iteration under the uniform distribution diverges: $|w_k| \to \infty$ for any $w_0 \neq 0$, even though $V^\pi = 0$ is exactly representable.*

On-policy projection converges. Under the on-policy distribution d^π with $d_0 = \delta_1$ (start in state 1), the discounted occupancy is $d^\pi = (1-\gamma,\, \gamma)$, since state 1 is visited only at $h = 0$ and state 2 at all subsequent steps. The projection becomes:

$$w_{k+1} = \operatorname*{argmin}_w \; \Big[(1-\gamma)(w - 2\gamma w_k)^2 + \gamma(2w - 2\gamma w_k)^2\Big].$$

Setting the derivative to zero: $(1-\gamma)(w - 2\gamma w_k) + 4\gamma(w - \gamma w_k) = 0$, giving

$$w_{k+1} = \frac{2\gamma(1+\gamma)}{1+3\gamma}\, w_k.$$

Since $\frac{2\gamma(1+\gamma)}{1+3\gamma} < 1$ for all $\gamma \in (0,1)$ (as $(2\gamma+1)(\gamma-1) < 0$), this converges geometrically to $w^\star = 0$. This confirms that the contraction property of $\Pi_{d^\pi}\mathcal{T}^\pi$ (Theorem 7) breaks down when the projection distribution is changed from d^π to a mismatched μ.

Why does the projection distribution matter? The Bellman operator $\mathcal{T}^\pi$ is a γ-contraction in the $\|\cdot\|_\infty$ norm, and the projection Π_μ is non-expansive in the $\|\cdot\|_{2,\mu}$ norm. But composing two operators that are well-behaved under **different** norms need not yield a contraction under either norm. The on-policy distribution d^π is special: the flow balance equation (Theorem 7) ensures that $\Pi_{d^\pi}\mathcal{T}^\pi$ is a $\sqrt{\gamma}$-contraction in a **single** norm $\|\cdot\|_{d^\pi}$. For a mismatched $\mu \neq d^\pi$, this structural guarantee is lost, and the composition $\Pi_\mu\mathcal{T}^\pi$ can amplify errors—as the 2-state example demonstrates.

GTD2 converges on this example. We verify that GTD2, which minimizes $\mathrm{MSPBE}(w) = \|\Phi w - \Pi_\mu \mathcal{T}^\pi(\Phi w)\|_\mu^2$, converges to the correct answer $w^\star = 0$ even under the uniform distribution $\mu = (1/2,\, 1/2)$.

From the projected Bellman iteration derivation above, $\Pi_\mu \mathcal{T}^\pi(\Phi w) = \Phi \cdot \frac{6\gamma}{5}\, w$. Therefore:

$$\Phi w - \Pi_\mu \mathcal{T}^\pi(\Phi w) = \begin{pmatrix} w \\ 2w \end{pmatrix} - \begin{pmatrix} \frac{6\gamma}{5}\, w \\ \frac{12\gamma}{5}\, w \end{pmatrix} = \left(1 - \tfrac{6\gamma}{5}\right) w \begin{pmatrix} 1 \\ 2 \end{pmatrix},$$

giving

$$\mathrm{MSPBE}(w) = \left(1 - \tfrac{6\gamma}{5}\right)^2 w^2 \, \|\phi\|_\mu^2 = \frac{(5-6\gamma)^2}{10}\, w^2.$$

For any $\gamma \neq 5/6$, this is a strictly convex quadratic with unique minimizer $w^\star = 0 = V^\pi$. The gradient is $\nabla\, \mathrm{MSPBE}(w) = \frac{(5-6\gamma)^2}{5}\, w$, so gradient descent with step size α gives

$$w_{t+1} = \left(1 - \tfrac{\alpha(5-6\gamma)^2}{5}\right) w_t \;\xrightarrow{t\to\infty}\; 0,$$

for any sufficiently small $\alpha > 0$.

Iteration vs. optimization. Projected Bellman *iteration* ($w_{k+1} = \frac{6\gamma}{5}\, w_k$) applies the operator $\Pi_\mu \mathcal{T}^\pi$ repeatedly and diverges when $|\frac{6\gamma}{5}| > 1$. GTD2 instead performs gradient *descent* on the MSPBE, a convex objective whose minimizer is the fixed point $\Pi_\mu \mathcal{T}^\pi(\Phi w^\star) = \Phi w^\star$. These are fundamentally different algorithms: the instability of the operator as a dynamical system does not prevent its fixed point from being found by optimization.

Remedies

Several approaches address the deadly triad; we mention three:

1. **Stay on-policy.** Use on-policy methods such as TD(0) (Algorithm 3). Our theory guarantees convergence, but this sacrifices the ability to reuse off-policy data.

2. **Remove bootstrapping.** Monte Carlo policy evaluation (Algorithm 1) uses complete returns as targets. It always converges (Theorem 2) but requires full trajectories and may suffer from higher variance for large γ.

3. **Gradient corrections.** The GTD2 algorithm (Algorithm 4) (Sutton et al., 2009) minimizes the MSPBE (7) using true stochastic gradient descent, which converges under off-policy sampling. The cost is a more complex algorithm requiring auxiliary weights θ.

Summary

In this lecture, we studied policy evaluation—estimating Q^π for a fixed policy π—using three approaches of increasing sophistication.

Summary of Results:

Result	**Statement**	**Key Technique**
MC policy evaluation	$\|\widehat{f} - Q^\pi\|_{d^\pi} \le O\left(\frac{V_{\max}}{\sqrt{n}}\right)$	Standard regression
FQE (general FA, fast rate)	$\|f_K - Q^\pi\|_\nu \le \widetilde{O}\left(\frac{\sqrt{C^\pi} V_{\max}}{(1-\gamma)\sqrt{n}}\right)$	Contraction + Bernstein
Projected Bellman contraction	$\|\Pi_{d^\pi}\mathcal{T}^\pi Q - \Pi_{d^\pi}\mathcal{T}^\pi Q'\|_{d^\pi} \le \sqrt{\gamma}\,\|Q - Q'\|_{d^\pi}$	Jensen + flow balance
TD(0) convergence	On-policy $\to$ projected Bellman fixed point	Linear stoch. approx.
GTD2 convergence	Off-policy $\to$ MSPBE minimizer	Saddle-point + two-timescale
Deadly triad	Off-policy + FA + bootstrap $\Rightarrow$ divergence	2-state counterexample

Key Takeaways:

- **Monte Carlo is simplest but most demanding:** MC regression gives the best rate $O(V_{\max}/\sqrt{n})$ with only realizability, but requires complete on-policy trajectories.
- **FQE trades data flexibility for a $1/(1-\gamma)$ cost:** by bootstrapping, FQE works with single-step transitions (possibly off-policy), at the cost of needing Bellman completeness and accumulating error over K iterations.
- **Removing $\max$ from $\mathcal{T} \to \mathcal{T}^\pi$ simplifies everything:** completeness becomes affine closure (easier), contraction is direct (no V-to-Q reduction), and concentrability requires only single-policy coverage.
- **Linear FA: TD(0) for on-policy, GTD2 for off-policy:** TD(0) is simple and fast on-policy, but can diverge off-policy. GTD2 provides a gradient-based remedy via the MSPBE objective.
- **The deadly triad:** the combination of function approximation, bootstrapping, and off-policy learning can cause divergence. The on-policy distribution d^π plays a critical structural role in ensuring stability.
- **Foundation for policy optimization:** policy evaluation produces Q^π, which is needed to compute the advantage $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$ for policy gradient methods.

References

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Machine Learning*, pages 30–37, 1995.

Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on Learning Theory*, pages 1691–1692. PMLR, 2018.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th International Conference on Machine Learning*, pages 993–1000, 2009.

John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.