Linear MDPs

Kianté Brantley & Sham Kakade

Outline

Recap: From Bandits to MDPs

- **Tabular MDPs (UCBVI):** We achieved $\tilde{O}(\mathrm{poly}(H)\sqrt{SAK})$ regret.
  - Great, but scales poorly when the state space $\mathcal{S}$ is enormous.
- **Linear Bandits (LinUCB):** We achieved $\tilde{O}(d\sqrt{K})$ regret.
  - No dependence on the number of arms! We exploit the feature dimension d.
- **Linear Bellman Completeness:** $\mathrm{poly}(d, H)$ sample size with gen model.
  - What about exploration/episodic setting?
    Random exploration is bad with $H > 1$.
  - What about an instantiation?

Recap: From Bandits to MDPs

- **Tabular MDPs (UCBVI):** We achieved $\tilde{O}(\mathrm{poly}(H)\sqrt{SAK})$ regret.
  - Great, but scales poorly when the state space $\mathcal{S}$ is enormous.
- **Linear Bandits (LinUCB):** We achieved $\tilde{O}(d\sqrt{K})$ regret.
  - No dependence on the number of arms! We exploit the feature dimension d.
- **Linear Bellman Completeness:** $\mathrm{poly}(d, H)$ sample size with gen model.
  - What about exploration/episodic setting?
    Random exploration is bad with $H > 1$.
  - What about an instantiation?
- **Today's Goal:** Can we combine these ideas?
  - We want an algorithm for episodic MDPs with $\mathrm{poly}(d, H, \sqrt{K})$ regret.
  - Crucially: Zero explicit dependence on $|\mathcal{S}|$ or $|\mathcal{A}|$.

Episodic MDP Setting

- **Horizon:** H steps per episode, total K episodes.
- **State/Action:** Large state space $\mathcal{S}$, action space $\mathcal{A}$, fixed start s_0.
- **Rewards:** Deterministic, known, and bounded: $r_h(s, a) \in [0, 1]$.
  - *Note: Unknown rewards just require an extra Hoeffding bonus. We assume them known to isolate the difficulty of learning transitions.*

The Interaction Loop: In episode k, the learner chooses a policy $\pi^k = \{\pi_h^k\}_{h=0}^{H-1}$, and generates a trajectory:

$$s_0^k = s_0, \qquad a_h^k = \pi_h^k(s_h^k), \qquad s_{h+1}^k \sim P_h^\star(\cdot \mid s_h^k, a_h^k)$$

Regret:

$$R_K := \sum_{k=0}^{K-1} \left(V^\star - V^{\pi^k}\right)$$

Outline

The Linear Transition Model

We assume the transition kernel is *linear in known features*.

Assumption (Linear MDP)

There exists a known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ *with* $\sup_{s,a} \|\phi(s,a)\|_2 \leq 1$, *and unknown parameters* $\{\mu_h^\star\}_{h=0}^{H-1}$ *with* $\mu_h^\star \in \mathbb{R}^{|\mathcal{S}| \times d}$ *such that for all* h *and* (s,a),

$$P_h^\star(\cdot \mid s,a) \;=\; \mu_h^\star \, \phi(s,a).$$

Equivalently, for each $s' \in \mathcal{S}$, $P_h^\star(s' \mid s,a) \;=\; \mu_h^\star(s')^\top \phi(s,a)$.

The Linear Transition Model

We assume the transition kernel is *linear in known features*.

Assumption (Linear MDP)

There exists a known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ *with* $\sup_{s,a} \|\phi(s,a)\|_2 \le 1$, *and unknown parameters* $\{\mu_h^\star\}_{h=0}^{H-1}$ *with* $\mu_h^\star \in \mathbb{R}^{|\mathcal{S}| \times d}$ *such that for all* h *and* (s,a),

$$P_h^\star(\cdot \mid s,a) \;=\; \mu_h^\star \, \phi(s,a).$$

Equivalently, for each $s' \in \mathcal{S}$, $P_h^\star(s' \mid s,a) \;=\; \mu_h^\star(s')^\top \phi(s,a)$.

Low-rank view (matrix form). View $P_h^\star$ as a matrix in $\mathbb{R}^{(SA) \times S}$ and $\Phi \in \mathbb{R}^{(SA) \times d}$. Then

$$P_h^\star \;=\; \Phi \, (\mu_h^\star)^\top, \qquad \text{so} \qquad \operatorname{rank}(P_h^\star) \le d.$$

The Linear Transition Model

We assume the transition kernel is *linear in known features*.

Assumption (Linear MDP)

There exists a known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ *with* $\sup_{s,a} \|\phi(s,a)\|_2 \le 1$, *and unknown parameters* $\{\mu_h^\star\}_{h=0}^{H-1}$ *with* $\mu_h^\star \in \mathbb{R}^{|\mathcal{S}| \times d}$ *such that for all* h *and* (s,a),

$$P_h^\star(\cdot \mid s,a) \;=\; \mu_h^\star \, \phi(s,a).$$

Equivalently, for each $s' \in \mathcal{S}$, $P_h^\star(s' \mid s,a) \;=\; \mu_h^\star(s')^\top \phi(s,a)$.

Low-rank view (matrix form). View $P_h^\star$ as a matrix in $\mathbb{R}^{(SA) \times S}$ and $\Phi \in \mathbb{R}^{(SA) \times d}$. Then

$$P_h^\star \;=\; \Phi \, (\mu_h^\star)^\top, \qquad \text{so} \qquad \operatorname{rank}(P_h^\star) \le d.$$

Normalization (used to bound coefficients). For all $f : \mathcal{S} \to \mathbb{R}$,

$$\left\| (\mu_h^\star)^\top f \right\|_2 \;\le\; \sqrt{d}\, \|f\|_\infty.$$

Examples & Intuition: What is a Linear MDP?

Warm-up (tabular): take $\phi(s, a)$ one-hot in $\mathbb{R}^{SA}$, so $d = SA$.

Examples & Intuition: What is a Linear MDP?

Warm-up (tabular): take $\phi(s, a)$ one-hot in $\mathbb{R}^{SA}$, so $d = SA$.

Example 1: Latent mixture / "topic" model

- Known mixture weights: $\phi(s, a) \in \Delta(d)$.
- Unknown components: $\mu^\star_{h,1}, \ldots, \mu^\star_{h,d} \in \Delta(\mathcal{S})$.
- Mixture of d base transition kernels:

$$P^\star_h(\cdot \mid s, a) = \sum_{i=1}^{d} \phi_i(s, a)\, \mu^\star_{h,i}(\cdot).$$

Examples & Intuition: What is a Linear MDP?

Warm-up (tabular): take $\phi(s,a)$ one-hot in $\mathbb{R}^{SA}$, so $d = SA$.

Example 1: Latent mixture / "topic" model

- Known mixture weights: $\phi(s,a) \in \Delta(d)$.
- Unknown components: $\mu^\star_{h,1}, \ldots, \mu^\star_{h,d} \in \Delta(\mathcal{S})$.
- Mixture of d base transition kernels:

$$P^\star_h(\cdot \mid s,a) = \sum_{i=1}^{d} \phi_i(s,a)\, \mu^\star_{h,i}(\cdot).$$

Example 2: Known state embedding + unknown linear latent dynamics

- Suppose we are given a *state embedding* $\psi : \mathcal{S} \to \mathbb{R}^d$ (learned or engineered).
- Unknown dynamics act linearly in the embedding space: $W^\star_h \in \mathbb{R}^{d \times d}$,

$$P^\star_h(s' \mid s,a) = \psi(s')^\top W^\star_h \phi(s,a) \qquad \text{(with normalization so this is a distribution).}$$

- Interpretation: next-state probabilities depend on (s,a) only through $\phi(s,a)$, and the dependence on s' is summarized by $\psi(s')$.

The "Magic" Property: Linear Backups

Key consequence: for any bounded $f : \mathcal{S} \to \mathbb{R}$, $(P_h^\star f)(s,a)$ is linear in $\phi(s,a)$.

Lemma (Linearization of $P_h^\star f$)

Fix h and bounded f. Let $w_{h,f}^\star := (\mu_h^\star)^\top f \in \mathbb{R}^d$. Then

$$\mathbb{E}_{s' \sim P_h^\star(\cdot|s,a)}[f(s')] \;=\; (P_h^\star f)(s,a) \;=\; \phi(s,a)^\top w_{h,f}^\star.$$

If $\|f\|_\infty \le H$, then $\|w_{h,f}^\star\|_2 \le H\sqrt{d}$.

$\mu^\star \in \mathbb{R}^{S \times d}$

$(P^\star f)_{s,a}$

$(T V)(s,a) = r(s,a) + \mathbb{E}_{s' \sim P_{sa}}[V(s')]$

Pf: $P = \Phi \mu^\top$, $Pf = \Phi \mu^\top f := \Phi w_f$ ($w_f = \mu^\top f$)

The "Magic" Property: Linear Backups

Key consequence: for any bounded $f : \mathcal{S} \to \mathbb{R}$, $(P_h^\star f)(s,a)$ is linear in $\phi(s,a)$.

Lemma (Linearization of $P_h^\star f$)

Fix h and bounded f. Let $w_{h,f}^\star := (\mu_h^\star)^\top f \in \mathbb{R}^d$. Then

$$\mathbb{E}_{s' \sim P_h^\star(\cdot|s,a)}[f(s')] \;=\; (P_h^\star f)(s,a) \;=\; \phi(s,a)^\top w_{h,f}^\star.$$

If $\|f\|_\infty \le H$, then $\|w_{h,f}^\star\|_2 \le H\sqrt{d}$.

$Pf = \Phi w$

Proof sketch:

$$(P_h^\star f)(s,a) = \sum_{s'} f(s')\, P_h^\star(s'|s,a) = \sum_{s'} f(s')\, \mu_h^\star(s')^\top \phi(s,a),$$

$$= \Big(\sum_{s'} f(s')\, \mu_h^\star(s')\Big)^\top \phi(s,a) = ((\mu_h^\star)^\top f)^\top \phi(s,a).$$

Implication: A Strong Form of Completeness

Recall (Bellman completeness): $f \in \mathcal{F} \Rightarrow \mathcal{T}_h f \in \mathcal{F}$ (closure only for f in the hypothesis class).

Implication: A Strong Form of Completeness

Recall (Bellman completeness): $f \in \mathcal{F} \Rightarrow \mathcal{T}_h f \in \mathcal{F}$ (closure only for f in the hypothesis class).

Linear MDP is stronger (for the transition part): for each stage h,

$$\forall\, f : \mathcal{S} \to \mathbb{R} \text{ bounded}, \qquad (P_h^\star f)(s,a) = \mathbb{E}[f(s') \mid s,a] \in \operatorname{span}(\phi).$$

Equivalently, there exists $w_{h,f}^\star \in \mathbb{R}^d$ such that

$$(P_h^\star f)(s,a) = \phi(s,a)^\top w_{h,f}^\star \qquad \forall (s,a).$$

Implication: A Strong Form of Completeness

Recall (Bellman completeness): $f \in \mathcal{F} \Rightarrow \mathcal{T}_h f \in \mathcal{F}$ (closure only for f in the hypothesis class).

Linear MDP is stronger (for the transition part): for each stage h,

$$\forall\, f : \mathcal{S} \to \mathbb{R} \text{ bounded}, \qquad (P_h^\star f)(s,a) = \mathbb{E}[f(s') \mid s,a] \in \operatorname{span}(\phi).$$

Equivalently, there exists $w_{h,f}^\star \in \mathbb{R}^d$ such that

$$(P_h^\star f)(s,a) = \phi(s,a)^\top w_{h,f}^\star \qquad \forall (s,a).$$

Why this matters for DP / least squares:

- $V_{h+1}^\star$ is nonlinear (max over actions), and $\widehat{V}_{h+1}^k$ is even messier (bonus + clipping).
- Still: the *regression target* $(s,a) \mapsto (P_h^\star \widehat{V}_{h+1}^k)(s,a)$ is always exactly linear in $\phi(s,a)$.
- So each stage reduces to a *well-specified* linear regression problem for $w_{h,\widehat{V}_{h+1}^k}^\star$.

Planning with a Known Model

Suppose we knew the transition parameters $\{\mu_h^\star\}_{h=0}^{H-1}$.

Backward DP: $V_H \equiv 0$, and for $h = H-1, \ldots, 0$,

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

Planning with a Known Model

Suppose we knew the transition parameters $\{\mu_h^\star\}_{h=0}^{H-1}$.

Backward DP: $V_H \equiv 0$, and for $h = H-1, \dots, 0$,

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

Linear MDP plug-in: for any f,

$$(P_h^\star f)(s,a) = \phi(s,a)^\top (\mu_h^\star)^\top f.$$

So the Bellman backup can be written as

$$Q_h^\star(s,a) = r_h(s,a) + \phi(s,a)^\top w_h^\star, \qquad w_h^\star := (\mu_h^\star)^\top V_{h+1}^\star \in \mathbb{R}^d.$$

Planning with a Known Model

Suppose we knew the transition parameters $\{\mu_h^\star\}_{h=0}^{H-1}$.

Backward DP: $V_H \equiv 0$, and for $h = H-1, \dots, 0$,

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

Linear MDP plug-in: for any f,

$$(P_h^\star f)(s,a) = \phi(s,a)^\top (\mu_h^\star)^\top f.$$

So the Bellman backup can be written as

$$Q_h^\star(s,a) = r_h(s,a) + \phi(s,a)^\top w_h^\star, \qquad w_h^\star := (\mu_h^\star)^\top V_{h+1}^\star \in \mathbb{R}^d.$$

Key takeaway: we never need the full $P_h^\star(\cdot \mid s,a)$. To plan, we can estimate the *linear functional*:

$$(s,a) \mapsto (P_h^\star V_{h+1})(s,a) = \phi(s,a)^\top w_{h,V_{h+1}}^\star.$$

Lin-UCBVI does this with ridge regression + an elliptical bonus.

Stage-wise Ridge Regression (Estimating $P_h^\star f$)

At episode k, stage h, we have past transitions $D_h^k = \{(s_h^i, a_h^i, s_{h+1}^i)\}_{i<k}$ and features $x_h^i := \phi(s_h^i, a_h^i) \in \mathbb{R}^d$.

Design matrix:

$$\Lambda_h^k := \lambda I + \sum_{i=0}^{k-1} x_h^i (x_h^i)^\top.$$

given f

what is a "good" estimates of P_f

$\mathbb{E}_{s' \sim P_{sa}}[f(s')] \neq \phi^\top w_f$

Ridge fit for a given target function $f : \mathcal{S} \to \mathbb{R}$:

$$\widehat{w}_{h,f}^k := (\Lambda_h^k)^{-1} \sum_{i=0}^{k-1} x_h^i f(s_{h+1}^i), \qquad (\widehat{P}_h^k f)(s,a) := \phi(s,a)^\top \widehat{w}_{h,f}^k.$$

$\sum_{s,a,s'} \left(\phi(s,a) \cdot w - f(s') \right)^2$

$\mathbb{E}[f(s') | s,a) = [Pf]_{s,a}$

Stage-wise Ridge Regression (Estimating $P_h^\star f$)

At episode k, stage h, we have past transitions $D_h^k = \{(s_h^i, a_h^i, s_{h+1}^i)\}_{i<k}$ and features $x_h^i := \phi(s_h^i, a_h^i) \in \mathbb{R}^d$.

Design matrix:

$$\Lambda_h^k := \lambda I + \sum_{i=0}^{k-1} x_h^i (x_h^i)^\top.$$

Ridge fit for a given target function $f : \mathcal{S} \to \mathbb{R}$:

$$\widehat{w}_{h,f}^k := (\Lambda_h^k)^{-1} \sum_{i=0}^{k-1} x_h^i f(s_{h+1}^i), \qquad (\widehat{P}_h^k f)(s,a) := \phi(s,a)^\top \widehat{w}_{h,f}^k.$$

Interpretation: we never estimate $P_h^\star(\cdot \mid s,a)$; we only estimate the *linear functional* $(s,a) \mapsto (P_h^\star f)(s,a)$ for the specific f used in planning (namely $f = \widehat{V}_{h+1}^k$).

UCB Bonus: Uncertainty in Direction $\phi(s, a)$

Under the linear MDP model, for each fixed f there is a true coefficient $w^\star_{h,f} = (\mu^\star_h)^\top f$ with $(P^\star_h f)(s, a) = \phi(s, a)^\top w^\star_{h,f}$. Ridge gives $\widehat{w}^k_{h,f}$.

UCB Bonus: Uncertainty in Direction $\phi(s, a)$

Under the linear MDP model, for each fixed f there is a true coefficient $w^\star_{h,f} = (\mu^\star_h)^\top f$ with $(P^\star_h f)(s, a) = \phi(s, a)^\top w^\star_{h,f}$. Ridge gives $\widehat{w}^k_{h,f}$.

High-probability form (LinUCB-style): for all (s, a),

$$\left|\phi(s, a)^\top (\widehat{w}^k_{h,f} - w^\star_{h,f})\right| \;\le\; \beta \, \|\phi(s, a)\|_{(\Lambda^k_h)^{-1}}.$$

UCB Bonus: Uncertainty in Direction $\phi(s, a)$

Under the linear MDP model, for each fixed f there is a true coefficient $w^\star_{h,f} = (\mu^\star_h)^\top f$ with $(P^\star_h f)(s, a) = \phi(s, a)^\top w^\star_{h,f}$. Ridge gives $\widehat{w}^k_{h,f}$.

High-probability form (LinUCB-style): for all (s, a),

$$\left|\phi(s, a)^\top (\widehat{w}^k_{h,f} - w^\star_{h,f})\right| \;\le\; \beta \, \|\phi(s, a)\|_{(\Lambda^k_h)^{-1}}.$$

So we use the exploration bonus:

$$b^k_h(s, a) := \beta \, \|\phi(s, a)\|_{(\Lambda^k_h)^{-1}}.$$

This is large when $\phi(s, a)$ lies in a poorly-sampled direction (small eigenvalues of Λ^k_h), and shrinks as we collect more data.

Optimistic Planning (Lin-UCBVI)

Exploration bonus:

$$b_h^k(s,a) := \beta \, \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}.$$

Optimistic Planning (Lin-UCBVI)

Exploration bonus:

$$b_h^k(s,a) := \beta \, \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}.$$

Backward induction: set $\widehat{V}_H^k \equiv 0$. For $h = H-1, \dots, 0$,

$$\widehat{Q}_h^k(s,a) := \mathrm{clip}_{[0,H-h]}\Big(r_h(s,a) + b_h^k(s,a) + (\widehat{P}_h^k \widehat{V}_{h+1}^k)(s,a) \Big)$$

Optimistic Planning (Lin-UCBVI)

Exploration bonus:

$$b_h^k(s,a) := \beta \, \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}.$$

Backward induction: set $\widehat{V}_H^k \equiv 0$. For $h = H-1, \dots, 0$,

$$\widehat{Q}_h^k(s,a) := \mathrm{clip}_{[0,H-h]}\Big(r_h(s,a) + b_h^k(s,a) + (\widehat{P}_h^k \widehat{V}_{h+1}^k)(s,a)\Big)$$

$$\widehat{V}_h^k(s) := \max_{a \in \mathcal{A}} \widehat{Q}_h^k(s,a)$$

Optimistic Planning (Lin-UCBVI)

Exploration bonus:

$$b_h^k(s,a) := \beta \, \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}.$$

Backward induction: set $\widehat{V}_H^k \equiv 0$. For $h = H-1, \dots, 0$,

$$\widehat{Q}_h^k(s,a) := \mathrm{clip}_{[0,H-h]}\Big(r_h(s,a) + b_h^k(s,a) + (\widehat{P}_h^k \widehat{V}_{h+1}^k)(s,a)\Big)$$

$$\widehat{V}_h^k(s) := \max_{a \in \mathcal{A}} \widehat{Q}_h^k(s,a)$$

$$\pi_h^k(s) \in \operatorname*{argmax}_{a \in \mathcal{A}} \widehat{Q}_h^k(s,a).$$

Optimistic Planning (Lin-UCBVI)

Exploration bonus:

$$b_h^k(s,a) := \beta \, \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}.$$

Backward induction: set $\widehat{V}_H^k \equiv 0$. For $h = H-1, \ldots, 0$,

$$\widehat{Q}_h^k(s,a) := \mathrm{clip}_{[0,H-h]}\Big(r_h(s,a) + b_h^k(s,a) + (\widehat{P}_h^k \widehat{V}_{h+1}^k)(s,a)\Big)$$

$$\widehat{V}_h^k(s) := \max_{a \in \mathcal{A}} \widehat{Q}_h^k(s,a)$$

$$\pi_h^k(s) \in \operatorname*{argmax}_{a \in \mathcal{A}} \widehat{Q}_h^k(s,a).$$

Execute π^k for one episode, then update $\Lambda_h^{k+1} = \Lambda_h^k + \phi(s_h^k, a_h^k)\phi(s_h^k, a_h^k)^\top$.

The Regret Guarantee

Theorem (Regret of Lin-UCBVI)

Assume the Linear MDP model and $r_h(s,a) \in [0,1]$. Run Lin-UCBVI with $\lambda \geq 1$ and a bonus scale $\beta = \widetilde{O}(Hd)$ (from a uniform confidence bound).
Then with probability at least $1-\delta$, the cumulative regret satisfies:

$$R_K = \sum_{k=0}^{K-1} \left(V^\star - V^{\pi^k}\right) \leq c\,\beta\,H\sqrt{K\,d\,\log\left(1+\frac{K}{\lambda}\right)} + \mathcal{O}\left(H\sqrt{K\log(1/\delta)}\right),$$

for a universal constant c.

The Regret Guarantee

Theorem (Regret of Lin-UCBVI)

Assume the Linear MDP model and $r_h(s,a) \in [0,1]$. Run Lin-UCBVI with $\lambda \geq 1$ and a bonus scale $\beta = \widetilde{O}(Hd)$ (from a uniform confidence bound).
Then with probability at least $1-\delta$, the cumulative regret satisfies:

$$R_K = \sum_{k=0}^{K-1} \left(V^\star - V^{\pi^k}\right) \leq c\,\beta\,H\sqrt{K\,d\,\log\left(1+\tfrac{K}{\lambda}\right)} + \mathcal{O}\left(H\sqrt{K\log(1/\delta)}\right),$$

for a universal constant c.

Takeaways:

- The $\mathcal{O}(\sqrt{K})$ martingale noise term is dominated by the primary β term.
- Absorbing lower-order terms gives: $R_K = \widetilde{O}\left(H^2\sqrt{d^3K}\right)$.
- No explicit dependence on $|\mathcal{S}|$ or $|\mathcal{A}|$ (structure captured entirely by d).

Proof Roadmap (What are the three moving parts?)

We reuse the same three ideas as UCBVI + LinUCB:

1. **(Confidence)** Uniformly control value-dependent regression targets:

$$\left|(\widehat{P}_h^k f)(s,a) - (P_h^\star f)(s,a)\right| \;\leq\; \beta \left\|\phi(s,a)\right\|_{(\Lambda_h^k)^{-1}} \quad \text{for } f = \widehat{V}_{h+1}^k.$$

Proof Roadmap (What are the three moving parts?)

We reuse the same three ideas as UCBVI + LinUCB:

1. **(Confidence)** Uniformly control value-dependent regression targets:

$$\left|(\widehat{P}_h^k f)(s,a) - (P_h^\star f)(s,a)\right| \;\le\; \beta \,\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} \quad \text{for } f = \widehat{V}_{h+1}^k.$$

2. **(Optimism)** Backward induction shows $\widehat{V}_h^k(s) \ge V_h^\star(s)$ for all k, h, s.

Proof Roadmap (What are the three moving parts?)

We reuse the same three ideas as UCBVI + LinUCB:

1. **(Confidence)** Uniformly control value-dependent regression targets:

$$\left|(\widehat{P}_h^k f)(s,a) - (P_h^\star f)(s,a)\right| \;\le\; \beta\,\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} \quad \text{for } f = \widehat{V}_{h+1}^k.$$

2. **(Optimism)** Backward induction shows $\widehat{V}_h^k(s) \ge V_h^\star(s)$ for all k, h, s.
3. **(Simulation + summation)** For each episode k,

$$V^\star - V^{\pi^k} \;\le\; 2\sum_{h=0}^{H-1} \mathbb{E}_{d_h^{\pi^k}}\left[b_h^k(s_h, a_h)\right],$$

then sum using an **elliptical potential** (log-det) argument.

Proof Roadmap (What are the three moving parts?)

We reuse the same three ideas as UCBVI + LinUCB:

1. **(Confidence)** Uniformly control value-dependent regression targets:

$$\left|(\widehat{P}_h^k f)(s,a) - (P_h^\star f)(s,a)\right| \;\le\; \beta\,\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} \quad \text{for } f = \widehat{V}_{h+1}^k.$$

2. **(Optimism)** Backward induction shows $\widehat{V}_h^k(s) \ge V_h^\star(s)$ for all k, h, s.
3. **(Simulation + summation)** For each episode k,

$$V^\star - V^{\pi^k} \;\le\; 2\sum_{h=0}^{H-1} \mathbb{E}_{d_h^{\pi^k}}\left[b_h^k(s_h,a_h)\right],$$

   then sum using an **elliptical potential** (log-det) argument.

Today's only new wrinkle vs LinUCB: the target $f = \widehat{V}_{h+1}^k$ is random, so we need *uniform* confidence over the whole optimistic value-function class.

The Statistical Hurdle: Why Uniform Confidence?

In LinUCB we estimate a fixed parameter $\mu^\star$. Here we estimate a *value-dependent* coefficient:

$$\widehat{w}^k_{h,\widehat{V}^k_{h+1}} = (\Lambda^k_h)^{-1} \sum_{i=0}^{k-1} \phi(s^i_h, a^i_h)\, \widehat{V}^k_{h+1}(s^i_{h+1}).$$

The Statistical Hurdle: Why Uniform Confidence?

In LinUCB we estimate a fixed parameter $\mu^\star$. Here we estimate a *value-dependent* coefficient:

$$\widehat{w}^k_{h,\widehat{V}^k_{h+1}} = (\Lambda^k_h)^{-1} \sum_{i=0}^{k-1} \phi(s^i_h, a^i_h)\, \widehat{V}^k_{h+1}(s^i_{h+1}).$$

The catch:

- $\widehat{V}^k_{h+1}$ is a random function built from past data (and bonuses).
- So the "labels" $\widehat{V}^k_{h+1}(s^i_{h+1})$ are correlated with the design/history.
- A fixed-function concentration bound is not enough; we need a bound that holds *simultaneously for all optimistic value functions the algorithm may produce*.

Step (a): Uniform Confidence over Optimistic Values

Proposition (Uniform model confidence (informal))

With probability at least $1-\delta$*, simultaneously for all episodes* k*, stages* h*, all* (s,a)*, and the specific* $f = \widehat{V}^k_{h+1}$ *produced by the algorithm,*

$$\left|(\widehat{P}^k_h f)(s,a) - (P^\star_h f)(s,a)\right| \;\le\; \beta\,\|\phi(s,a)\|_{(\Lambda^k_h)^{-1}}.$$

Step (a): Uniform Confidence over Optimistic Values

Proposition (Uniform model confidence (informal))

With probability at least $1-\delta$*, simultaneously for all episodes* k*, stages* h*, all* (s,a)*, and the specific* $f = \widehat{V}_{h+1}^{k}$ *produced by the algorithm,*

$$\left|(\widehat{P}_h^k f)(s,a) - (P_h^\star f)(s,a)\right| \;\le\; \beta\,\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}.$$

Where does β come from? Self-normalized martingale concentration + an ϵ-net over the optimistic value class (max + bonus + clipping), which yields $\beta = \widetilde{O}(Hd)$.

Step (b): Optimism (Backward Induction)

On the uniform confidence event, for every episode k, stage h, and state s,

$$\widehat{V}_h^k(s) \;\geq\; V_h^\star(s).$$

Step (b): Optimism (Backward Induction)

On the uniform confidence event, for every episode k, stage h, and state s,

$$\widehat{V}_h^k(s) \;\geq\; V_h^\star(s).$$

Reason (one line): for any (s, a),

$$(P_h^\star \widehat{V}_{h+1}^k)(s,a) \leq (\widehat{P}_h^k \widehat{V}_{h+1}^k)(s,a) + b_h^k(s,a),$$

so $r_h + P_h^\star V_{h+1}^\star \leq r_h + b_h^k + \widehat{P}_h^k \widehat{V}_{h+1}^k$ (and clipping preserves the inequality).

Step (c): Simulation Lemma (Per-episode bound)

Let $d_h^{\pi^k}(s,a) = \Pr(s_h = s, a_h = a \mid s_0, \pi^k, P^\star)$ be the stage-h occupancy.
On the uniform confidence event, for every episode k,

$$V^\star - V^{\pi^k} \;\le\; 2 \sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^k}} \big[b_h^k(s_h, a_h) \big].$$

Step (c): Simulation Lemma (Per-episode bound)

Let $d_h^{\pi^k}(s,a) = \Pr(s_h = s, a_h = a \mid s_0, \pi^k, P^\star)$ be the stage-h occupancy.
On the uniform confidence event, for every episode k,

$$V^\star - V^{\pi^k} \;\le\; 2 \sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^k}} \big[b_h^k(s_h, a_h) \big].$$

Plug in the LinUCB-style bonus:

$$b_h^k(s,a) = \beta \, \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} \quad \Longrightarrow \quad V^\star - V^{\pi^k} \;\le\; 2\beta \sum_{h=0}^{H-1} \mathbb{E}_{d_h^{\pi^k}} \Big[\|\phi(s_h, a_h)\|_{(\Lambda_h^k)^{-1}} \Big].$$

Summation: Elliptical Potential (Log-det)

Summing the simulation bound over episodes reduces regret to

$$R_K \leq 2\beta \sum_{h=0}^{H-1} \sum_{k=0}^{K-1} \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}}.$$

Summation: Elliptical Potential (Log-det)

Summing the simulation bound over episodes reduces regret to

$$R_K \;\leq\; 2\beta \sum_{h=0}^{H-1} \sum_{k=0}^{K-1} \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}}.$$

Fix a stage h and write $x_k := \phi(s_h^k, a_h^k)$, $\Lambda_{k+1} = \Lambda_k + x_k x_k^\top$. Matrix determinant lemma gives the standard potential bound:

$$\sum_{k=0}^{K-1} \|x_k\|^2_{\Lambda_k^{-1}} \;\leq\; 2 \log \frac{\det(\Lambda_K)}{\det(\Lambda_0)} \;\leq\; 2d \log\Big(1 + \frac{K}{\lambda}\Big).$$

Summation: Elliptical Potential (Log-det)

Summing the simulation bound over episodes reduces regret to

$$R_K \;\le\; 2\beta \sum_{h=0}^{H-1} \sum_{k=0}^{K-1} \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}}.$$

Fix a stage h and write $x_k := \phi(s_h^k, a_h^k)$, $\Lambda_{k+1} = \Lambda_k + x_k x_k^\top$. Matrix determinant lemma gives the standard potential bound:

$$\sum_{k=0}^{K-1} \|x_k\|^2_{\Lambda_k^{-1}} \;\le\; 2 \log \frac{\det(\Lambda_K)}{\det(\Lambda_0)} \;\le\; 2d \log\Big(1 + \frac{K}{\lambda}\Big).$$

By Cauchy–Schwarz,

$$\sum_{k=0}^{K-1} \|x_k\|_{\Lambda_k^{-1}} \;\le\; \sqrt{K \sum_{k=0}^{K-1} \|x_k\|^2_{\Lambda_k^{-1}}} \;\le\; \sqrt{2K\, d\, \log\Big(1 + \frac{K}{\lambda}\Big)}.$$

Sum over h and multiply by β to get the theorem.

The Question: What makes RL Generalize?

We have seen two successful "recipes" for avoiding $|\mathcal{S}|$-dependence:

1. **Tabular UCBVI:** Estimate the full local transition.
2. **Lin-UCBVI:** Estimate only the Bellman-relevant lookahead functionals.

The Question: What makes RL Generalize?

We have seen two successful "recipes" for avoiding $|\mathcal{S}|$-dependence:

1. **Tabular UCBVI:** Estimate the full local transition.
2. **Lin-UCBVI:** Estimate only the Bellman-relevant lookahead functionals.

The General Problem: What if we don't know the features $\phi(s, a)$? Or what if the transitions are non-linear, but some other structural bottleneck exists?

The Question: What makes RL Generalize?

We have seen two successful "recipes" for avoiding $|\mathcal{S}|$-dependence:

1. **Tabular UCBVI:** Estimate the full local transition.
2. **Lin-UCBVI:** Estimate only the Bellman-relevant lookahead functionals.

The General Problem: What if we don't know the features $\phi(s, a)$? Or what if the transitions are non-linear, but some other structural bottleneck exists?

The Goal: We seek a **structural complexity measure** that tells us when trajectory data collected under policy f can effectively "test" the quality of another hypothesis g.

The Core Object: Average Bellman Error

Define the one-step Bellman residual of a hypothesis $g \in \mathcal{H}$ as:

$$\ell_h(s, a, s'; g) \;:=\; Q_{h,g}(s, a) - r_h(s, a) - V_{h+1,g}(s')$$

The Core Object: Average Bellman Error

Define the one-step Bellman residual of a hypothesis $g \in \mathcal{H}$ as:

$$\ell_h(s, a, s'; g) \;:=\; Q_{h,g}(s, a) - r_h(s, a) - V_{h+1,g}(s')$$

We measure the quality of g using its **Average Bellman Error** under the roll-in distribution induced by the policy of f, denoted $d_h^{\pi_f}$:

$$\mathcal{E}_h^Q(f, g) \;:=\; \mathbb{E}_{(s,a)\sim d_h^{\pi_f},\; s'\sim P_h^\star}\Big[\ell_h(s, a, s'; g)\Big]$$

The Core Object: Average Bellman Error

Define the one-step Bellman residual of a hypothesis $g \in \mathcal{H}$ as:

$$\ell_h(s, a, s'; g) \;:=\; Q_{h,g}(s, a) - r_h(s, a) - V_{h+1,g}(s')$$

We measure the quality of g using its **Average Bellman Error** under the roll-in distribution induced by the policy of f, denoted $d_h^{\pi_f}$:

$$\mathcal{E}_h^Q(f, g) \;:=\; \mathbb{E}_{(s,a)\sim d_h^{\pi_f},\; s'\sim P_h^\star}\Big[\ell_h(s, a, s'; g)\Big]$$

Key Intuition:

- If $g = f^\star$ (realizability), then $\mathcal{E}_h(f, f^\star) = 0$ for *every* roll-in policy f.
- Finding a near-optimal policy is equivalent to finding a g that has low Bellman error under its own distribution ($\mathcal{E}_h(g, g) \approx 0$).

Defining Bellman Rank

Consider a massive matrix M_h where rows are "Roll-in Policies" $f \in \mathcal{H}$ and columns are "Tested Hypotheses" $g \in \mathcal{H}$. Each entry is the cross-error $\mathcal{E}_h(f, g)$.

Definition (Q-Bellman Rank)

The pair $(\mathcal{M}, \mathcal{H})$ has *Q-Bellman rank at most d* if there exist maps $X_h, W_h : \mathcal{H} \to \mathbb{R}^d$ such that:

$$\mathcal{E}_h^Q(f, g) = \langle X_h(f), W_h(g) \rangle \qquad \forall f, g \in \mathcal{H}$$

Defining Bellman Rank

Consider a massive matrix M_h where rows are "Roll-in Policies" $f \in \mathcal{H}$ and columns are "Tested Hypotheses" $g \in \mathcal{H}$. Each entry is the cross-error $\mathcal{E}_h(f, g)$.

Definition (Q-Bellman Rank)

The pair $(\mathcal{M}, \mathcal{H})$ has *Q-Bellman rank at most d* if there exist maps $X_h, W_h : \mathcal{H} \to \mathbb{R}^d$ such that:

$$\mathcal{E}_h^Q(f, g) = \langle X_h(f), W_h(g) \rangle \qquad \forall f, g \in \mathcal{H}$$

Conceptual Punchline:

- A Linear MDP assumes the **Transition Matrix** is low rank ($rank \leq d$).
- Bellman Rank assumes the **Error Matrix** is low rank ($rank \leq d$).
- We do not need to know the features X_h, W_h; we only assume they exist.

Example: LinMDPs have Bellman Rank d

Why did today's Linear MDP model work? Let's check its Bellman Rank:

- **LinMDP Hypothesis:** $Q_{h,g}(s,a) = \phi(s,a)^\top w_{h,g}$.
- **Bellman Error:** $\mathcal{E}_h^Q(f,g) = \mathbb{E}_{d_h^{\pi_f}}[\phi(s,a)^\top w_{h,g} - r_h(s,a) - \mathbb{E}_{P^\star}[V_{h+1,g}(s')]]$.

Example: LinMDPs have Bellman Rank d

Why did today's Linear MDP model work? Let's check its Bellman Rank:

- **LinMDP Hypothesis:** $Q_{h,g}(s,a) = \phi(s,a)^\top w_{h,g}$.
- **Bellman Error:** $\mathcal{E}_h^Q(f,g) = \mathbb{E}_{d_h^{\pi_f}}[\phi(s,a)^\top w_{h,g} - r_h(s,a) - \mathbb{E}_{P^\star}[V_{h+1,g}(s')]]$.

From our "Magic Property" Lemma: $r_h(s,a) + \mathbb{E}_{P^\star}[V_{h+1,g}]$ is linear in $\phi(s,a)$. Let its coefficient be $w^\star_{lookahead}$.

$$\begin{aligned}\mathcal{E}_h^Q(f,g) &= \mathbb{E}_{(s,a)\sim d_h^{\pi_f}}\left[\phi(s,a)^\top (w_{h,g} - w^\star_{lookahead})\right] \\ &= \Big\langle \underbrace{\mathbb{E}_{d_h^{\pi_f}}[\phi(s,a)]}_{X_h(f)}, \underbrace{w_{h,g} - w^\star_{lookahead}}_{W_h(g)} \Big\rangle\end{aligned}$$

Example: LinMDPs have Bellman Rank d

Why did today's Linear MDP model work? Let's check its Bellman Rank:

- **LinMDP Hypothesis:** $Q_{h,g}(s,a) = \phi(s,a)^\top w_{h,g}$.
- **Bellman Error:** $\mathcal{E}_h^Q(f,g) = \mathbb{E}_{d_h^{\pi_f}}[\phi(s,a)^\top w_{h,g} - r_h(s,a) - \mathbb{E}_{P^\star}[V_{h+1,g}(s')]]$.

From our "Magic Property" Lemma: $r_h(s,a) + \mathbb{E}_{P^\star}[V_{h+1,g}]$ is linear in $\phi(s,a)$. Let its coefficient be $w^\star_{lookahead}$.

$$\begin{aligned}\mathcal{E}_h^Q(f,g) &= \mathbb{E}_{(s,a)\sim d_h^{\pi_f}}\left[\phi(s,a)^\top (w_{h,g} - w^\star_{lookahead})\right] \\ &= \Big\langle \underbrace{\mathbb{E}_{d_h^{\pi_f}}[\phi(s,a)]}_{X_h(f)}, \underbrace{w_{h,g} - w^\star_{lookahead}}_{W_h(g)} \Big\rangle\end{aligned}$$

Result: LinMDPs are just one specific instance of the "Model Zoo" unified by Bellman Rank.

Algorithmic Teaser: Optimistic Elimination

How do we learn if the error matrix is low-rank but the features are unknown?

The PAC-RL Strategy:

1. **Eliminate:** If a hypothesis g shows a large empirical Bellman error on any past roll-in data, throw it out.
2. **Optimism:** Pick the f_t that predicts the highest value among survivors.
3. **Roll-in:** Collect new data using policy π_{f_t}.

Algorithmic Teaser: Optimistic Elimination

How do we learn if the error matrix is low-rank but the features are unknown?

The PAC-RL Strategy:

1. **Eliminate:** If a hypothesis g shows a large empirical Bellman error on any past roll-in data, throw it out.
2. **Optimism:** Pick the f_t that predicts the highest value among survivors.
3. **Roll-in:** Collect new data using policy π_{f_t}.

The Logic for Next Time: Since the error matrix has rank d, we can only "fail" to eliminate a sub-optimal f_t a maximum of d times before we have spanned the space of possible errors.

Next Lecture: Formalizing the "Simplicity" of RL and the PAC proof.