RL with continuous action spaces
General function approximation

Aarti Singh

Machine Learning 10-734
Nov 6, 2025

Slides courtesy: Wen Sun

Continuous state, action spaces

Bandits:

Reward is linear, Lipshitz, GP, NN, …

e.g. $r^*(x) = x^T\theta^*$ x, θ^* are d-dimensional $x^* \to r^*$

MDP:

Linear MDP - Reward is linear, Transition is low rank $-d$ $\phi \in \mathbb{R}^d$

$$r_h(s,a) = w_h^T\ \phi(s,a),\ P_h(s'|s,a) = \mu_h(s')^T\phi(s,a)$$

LSVI-UCB algorithm has low regret $\tilde{O}(H^2\sqrt{d^3N})$

Linear Q* - $Q^*(s,a) = \theta^{*\,T}\phi(s,a)$ $\pi^* \to Q^*$

Doesn't work! $e^{S,A}$

LSVI-UCB: Least Square Value Iteration with UCB

Value iteration at episode n using $\{s_h^i, a_h^i, r_h^i, s_{h+1}^i\}_{h=1,i=1}^{H-1,n-1}$

$$\widehat{V}_H^n(s) = 0, \forall s$$

For h = H-1, H-2, …, 1

least square Bellman consistency

$$\theta_h^n \leftarrow \arg\min_{\theta} \sum_{i=1}^{n-1} \Big(\overbrace{\langle \phi(s_h^i, a_h^i), \theta \rangle}^{\bar{Q}(s,a)} - \overbrace{r_h^i - \widehat{V}_{h+1}^n(s_{h+1}^i)}^{TQ} \Big)^2 + \lambda \|\theta\|_2^2$$

$$\widehat{Q}_h^n(s,a) = \min\left\{ b_h^n(s,a) + \langle \phi(s,a), \theta_h^n \rangle, \quad H \right\}, \forall s, a$$

$$\widehat{V}_h^n(s) = \max_a \widehat{Q}_h^n(s,a), \qquad \pi_h^n(s) = \arg\max_a \widehat{Q}_h^n(s,a), \forall s$$

Bellman error

Consider $f(s,a) = Q(s,a)$.

Bellman error = $f(s,a) - T\,f(s,a)$ $\equiv Q - TQ$

f linear

Tf linear

$$= \overset{\phi(s,a)}{f(s,a)} - \underbrace{\left(\overset{\phi(s,a)}{r(s,a)} + \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} \overset{\phi(s,a)}{f(s',a')} \right)}_{Tf}$$

$$= \sum_{s'} P(s'|s,a) \max_{a'} f(s',a')$$

If Bellman error $\neq 0$, then $f \neq Q^*$

Why does linear Q^* not suffice?

Even if f is linear, $T\,f$ may not be linear T unless transitions P (and reward r) is also linear!

Bellman completeness

Bellman completeness: For any Q function in $\mathcal{F}$, its Bellman update is also in $\mathcal{F}$

Implies Bellman error

Bellman completeness: For any Q function in $\mathcal{F}$, its Bellman update is also in $\mathcal{F}$

Implies Bellman error

can be 0 for $f = Q^*$

can be 0 for $f = Q^*$

LSVI-UCB: Least Square Value Iteration with UCB

Value iteration at episode n using $\{s_h^i, a_h^i, r_h^i, s_{h+1}^i\}_{h=1,i=1}^{H-1,n-1}$

$$\widehat{V}_H^n(s) = 0, \forall s$$

For h = H-1, H-2, …, 1

$Q = TQ$

$$\theta_h^n \leftarrow \arg\min_{\theta} \sum_{i=1}^{n-1} \left(\langle \phi(s_h^i, a_h^i), \theta \rangle - r_h^i - \widehat{V}_{h+1}^n(s_{h+1}^i) \right)^2 + \lambda \|\theta\|_2^2$$

$$\widehat{Q}_h^n(s,a) = \min \left\{ b_h^n(s,a) + \langle \phi(s,a), \theta_h^n \rangle, \quad H \right\}, \forall s, a$$

$\frac{1}{\sqrt{N_h^n(s,a)}}$

$$\widehat{V}_h^n(s) = \max_a \widehat{Q}_h^n(s,a), \qquad \pi_h^n(s) = \arg\max_a \widehat{Q}_h^n(s,a), \forall s$$

LSVI-UCB does not work under Bellman completeness

$$\widehat{Q}_h^n(s,a) = \min\left\{ b_h^n(s,a) + \langle \phi(s,a), \theta_h^n \rangle, \quad H \right\}, \forall s, a$$

Issue: Adding bonus which may be non-realizable
(e.g. in linear case, bonus may be nonlinear in s)

Recall $b_h^n(s,a) = \|\phi\|_{\Lambda_h^{n\,-1}} \beta$ $\longrightarrow$ nonlinear in $\phi(s,a)$

Need different algorithm (no bonus on Q)
– how to achieve optimism?

Average Bellman error

Weaker notion of Bellman error:

Evaluate g-approximation of Q using a policy π_f

$$\mathcal{E}(g; f, h) = \mathbb{E}_{s_h, a_h \sim d_h^{\pi_f}} \left[\underbrace{g(s_h, a_h) - r(s_h, a_h) - \mathbb{E}_{s_{h+1} \sim P(\cdot | s_h, a_h)} \left[\max_{a \in \mathcal{A}} g(s_{h+1}, a) \right]}_{\text{Bellman error } Q - TQ} \right]$$

f: defines roll-in distribution over s_h, a_h

We know that $\mathcal{E}(Q^\star; f, h) = 0, \forall f$ if $g = Q^*$ $g = Tg$ pointwise (s, a)

Hence, any g such that $\mathcal{E}(g; f, h) \neq 0$, is an incorrect $Q^\star$ approximator

Average Bellman error

Evaluate average Bellman wrt V function induced by g as well:

V - TV

$$\mathscr{E}(g; f, h) = \mathbb{E}_{s_h \sim d_h^{\pi_f}} \left[V_g(s_h) - r(s_h, \pi_g(s_h)) - \mathbb{E}_{s_{h+1} \sim P(\cdot | s_h, \pi_g(s_h))} \left[V_g(s_{h+1}) \right] \right]$$

f: defines roll-in distribution over s_h, a_h

Again we have $\mathscr{E}(Q^\star; f, h) = 0, \forall f$ if $g = Q^*$ $V_g = V^*$

(because: $V_{Q^\star}(s) - r(s, \pi_{Q^\star}(s)) - \mathbb{E}_{s' \sim P_h(.|s, \pi_{Q^\star}(s))} V_{Q^\star}(s') = 0$)

Hence, any g such that $\mathscr{E}(g; f, h) \neq 0$, is an incorrect $Q^\star$ approximator

$Q-\mathcal{T}Q$ (s,a)

avg $(Q-\mathcal{T}Q)$

f g

Bellman rank

approx

$g \approx Q$ $f = \text{avg}$

$\exists$ two mappings $W_h : \mathcal{F} \mapsto \mathbb{R}^d$, $X_h : \mathcal{F} \mapsto \mathbb{R}^d$ (d = Bellman-rank)

s.t. $\forall f, g \in \mathcal{F} : \mathcal{E}(g; f, h) = \langle W_h(g), X_h(f) \rangle$

$\forall h : \mathcal{E}_h \in \mathbb{R}^{|\mathcal{F}| \times |\mathcal{F}|}$

approx

		g	f			
π_f		$\mathcal{E}_{g;f,h}$	$\mathcal{E}_{f;f,h}$			
avg						

Rank of this matrix = Bellman rank

Note: we just assume the existence of W, X, but they are unknown

Examples of Bellman rank

- **Linear Bellman completeness**: For any linear Q function, its Bellman update is also linear

Given feature ϕ, take any linear function $\theta^\top \phi(s, a)$:

$$\forall h, \exists w \in \mathbb{R}^d, s.t., w^\top \phi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} \theta^\top \phi(s', a'), \forall s, a$$

$Q \leftarrow TQ$

Claim: it has Q-Bellman rank d

$\forall g(s,a) := \theta^\top \phi(s,a)$, we have:

avg (Q − TQ)

$$\mathcal{E}(g; f, h) = \mathbb{E}_{s_h, a_h \sim d_h^{\pi_f}} \left[\theta^\top \phi(s_h, a_h) - r(s_h, a_h) - \mathbb{E}_{s_{h+1} \sim P_h(\cdot | s_h, a_h)} \left[\max_{a \in \mathcal{A}} \theta^\top \phi(s_{h+1}, a) \right] \right]$$

$$= \mathbb{E}_{s_h, a_h \sim d_h^{\pi_f}} \left[\theta^\top \phi(s_h, a_h) - w^T \phi(s_h, a_h) \right]$$ Bellman completeness

$$= \left\langle \theta - w, \mathbb{E}_{s_h, a_h \sim d_h^{\pi_f}} [\phi(s_h, a_h)] \right\rangle$$

Examples of Bellman rank

- **Linear Bellman completeness**: For any linear Q function, its Bellman update is also linear.

Given feature ϕ, take any linear function $\theta^\top \phi(s,a)$:

$$\forall h, \exists w \in \mathbb{R}^d, s.t., w^\top \phi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} \theta^\top \phi(s',a'), \forall s, a$$

r linear, P low rank $r = \theta^\top \phi(s,a)$ $P = \mu(s')^\top \phi(s,a)$

- **Linear MDP**

$\Rightarrow$ linear Bellman completeness $\Rightarrow$ Q-Bellman rank d

Examples of Bellman rank

- **Linear Q* and V*** $Q^\star(s,a) = (w^\star)^\top \phi(s,a), \quad V^\star(s) = (\theta^\star)^\top \psi(s), \forall s, a$

Claim: it has Q-Bellman rank 2d

$$\mathscr{F}_h = \left\{ (w,\theta) : \max_a \overbrace{w^\top \phi(s,a)}^{Q(s,a)} = \overbrace{\theta^\top \psi(s)}^{V(s)}, \forall s \right\}$$

avg(Q-TQ)

$$\mathscr{E}(g; f, h) = \mathbb{E}_{s_h, a_h \sim d_h^{\pi_f}} \left[w^\top \phi(s_h, a_h) - r(s_h, a_h) - \mathbb{E}_{s_{h+1} \sim P_h(\cdot | s_h, a_h)} \left[\theta^\top \psi(s_{h+1}) \right] \right]$$

$$= \mathbb{E}_{s_h, a_h \sim d_h^{\pi_f}} \Big[w^\top \phi(s_h, a_h) - \underbrace{(w^\star)^\top \phi(s_h, a_h)}_{Q^\star(s_h,a_h)} + \mathbb{E}_{s_{h+1} \sim P_h(\cdot | s_h, a_h)} \overbrace{\left[(\theta^\star)^\top \psi(s_{h+1}) \right]}^{V^\star(s_{h+1})}$$

$$- \mathbb{E}_{s_{h+1} \sim P_h(\cdot | s_h, a_h)} \left[\theta^\top \psi(s_{h+1}) \right] \Big]$$

$$= \left\langle \begin{bmatrix} w - w^\star \\ \theta - \theta^\star \end{bmatrix}, \ \mathbb{E}_{s_h, a_h \sim d_h^{\pi_f}} \begin{bmatrix} \phi(s_h, a_h) \\ -\mathbb{E}_{s' \sim P_h(s_h, a_h)}[\psi(s')] \end{bmatrix} \right\rangle$$

$\psi(s') \neq \eta^\top \phi(s,a)$

Examples of Bellman rank

- **Linear Q* and V*** $Q^\star(s,a) = (w^\star)^\top \phi(s,a), \quad V^\star(s) = (\theta^\star)^\top \psi(s), \forall s, a$

Claim: it has Q-Bellman rank 2d

Note that $\psi(s_h, a_h) := \mathbb{E}_{s' \sim P_h(s_h, a_h)}[\psi(s')]$ is in general not linear in $\phi(s_h, a_h)$ if transition dynamics are not linear

But V* linear *inherently* implies transition dynamics are linear:

Since V* = TV*, we have

$\mathbb{E}_{s' \sim P(\cdot|s,a)}[\psi(s')]$

$$\underbrace{\theta^{*T}\psi(s)}_{V^*} = \max_a \left(r(s,a) + \theta^{*T} \underbrace{\psi(s,a)}_{\mathbb{E}_{s' \sim P(\cdot|s,a)}[\psi(s')]} \right)$$

which implies transition dynamics are linear (given definition of $\psi(s,a)$).

Linear Q*, V* suffices, though linear Q* doesn't!

Examples of Bellman rank

- **Low rank MDP**

$\in \mathbb{R}^d$ linear - ϕ known

$$P_h(s' \mid s, a) = \mu_h^\star(s')^\top \underbrace{\phi_h^\star(s, a)}$$ (neither $\mu^\star$ nor $\phi^\star$ is known)

Claim: this model has V-Bellman rank d

$$\mathcal{F}_h = \{\theta^\top \phi(\cdot, \cdot) : \|\theta\|_2 \le W, \phi \in \Phi\}$$

arg (V-TV)

$$\mathbb{E}_{s_h \sim d_h^{\pi_f}}\left[V_g(s_h) - r(s, \pi_g(s_h)) - \mathbb{E}_{s_{h+1} \sim P_h(\cdot|s_h, \pi_g(s_h))}[V_g(s_{h+1})]\right]$$

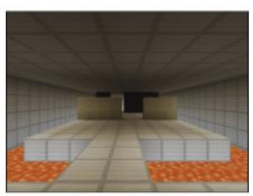
$$= \mathbb{E}_{\tilde{s}, \tilde{a} \sim d_{h-1}^{\pi_f}} \mathbb{E}_{s_h \sim P_{h-1}(\cdot|\tilde{s}, \tilde{a})}\left[V_g(s_h) - r(s, \pi_g(s_h)) - \mathbb{E}_{s_{h+1} \sim P_h(\cdot|s_h, \pi_g(s_h))}[V_g(s_{h+1})]\right]$$

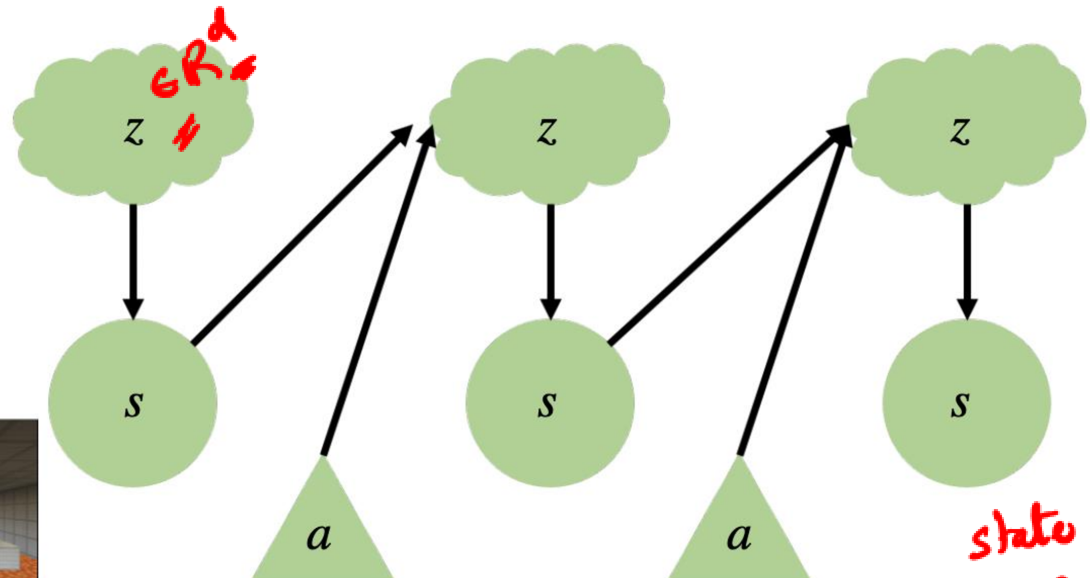
$$= \mathbb{E}_{\tilde{s}, \tilde{a} \sim d_{h-1}^{\pi_f}} \int_{s_h} \underbrace{\mu_{h-1}^\star(s_h)^\top \phi_{h-1}^\star(\tilde{s}, \tilde{a})}_{P_{h-1}(s_h|\tilde{s},\tilde{a})} \underbrace{\left[V_g(s_h) - r(s, \pi_g(s_h)) - \mathbb{E}_{s_{h+1} \sim P_h(\cdot|s_h, \pi_g(s_h))}[V_g(s_{h+1})]\right]}_{\text{scalar}} d(s_h)$$

$$= \left\langle \int_{s_h} \mu_{h-1}^\star(s_h) \underbrace{\left[V_g(s_h) - r(s, \pi_g(s_h)) - \mathbb{E}_{s_{h+1} \sim P_h(\cdot|s_h, \pi_g(s_h))}[V_g(s_{h+1})]\right]}_{\text{scalar}} d(s_h), \quad \mathbb{E}_{\tilde{s}, \tilde{a} \sim d_{h-1}^{\pi_f}}[\phi_{h-1}^\star(\tilde{s}, \tilde{a})] \right\rangle$$

Examples of Bellman rank

- **Latent variable MDP** V-Bellman rank = Number of latent states





Given s, a: $z \sim \phi^\star(s, a)$, $s' \sim \nu^\star(z)$

state transitions are governed by latent space dynamics

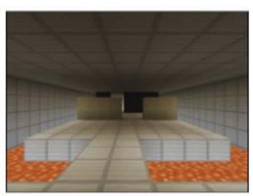
Latent variable MDP is captured by low-rank MDP, so it has small V-Bellman rank…

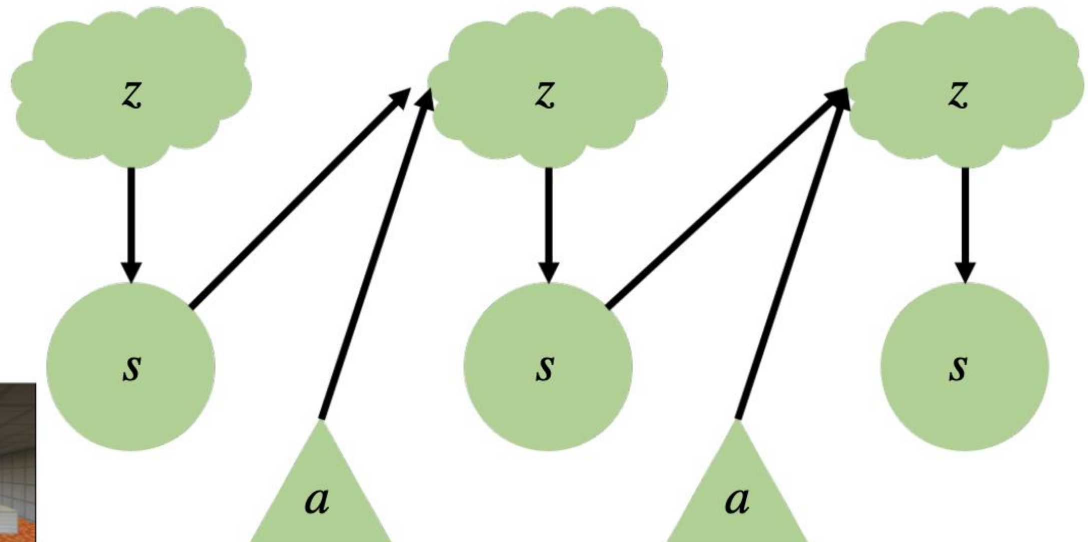
Examples of Bellman rank

Dim (z is continuous)

- **Latent variable MDP** V-Bellman rank = Number of latent states

(finite z)



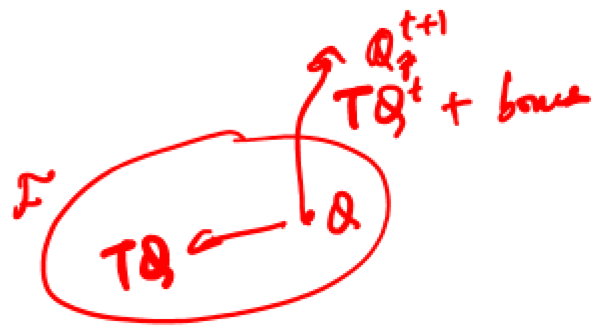


Given s, a: $z \sim \phi^\star(s,a), s' \sim \nu^\star(z)$

- **Block MDP -** Special case of latent variable MDP where a state can only be generated from one latent state i.e. one-to-one mapping, hence latent state is deterministically decodable $P(s|z) > 0 \Rightarrow P(s|z') = 0 \;\; \forall\, z' \neq z$

Algorithm under Bellman rank

- Works for general function approximation with low Bellman rank
- Gives optimism without adding (nonlinear) bonus



Bilinear-UCB

A general algorithm under Bellman rank that can learn an ϵ near optimal policy with number of samples

$$\text{e.g., poly}(\underline{H}, \underline{\text{b-rank}}, \ln(|\mathscr{F}|), 1/\underline{\epsilon^2})$$

e.g. $d \ll |S|, |A|$

Q-Bellman rank setting

Recall our hypothesis class $\mathscr{F}$, where each $g \in \mathscr{F}$ is in the form of $g(s, a)$

For Q-Bellman rank, we define Bellman error loss as:

$$\ell(s_h, a_h, s'_{h+1}, g) = g(s_h, a_h) - r(s_h, a_h) - \max_{a'} g(s_{h+1}, a')$$

generative

If we had a dataset $\mathscr{D} := \{s_h, a_h, s_{h+1}\}$ where $s_h, a_h \sim d_h^{\pi_f}, s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$

then $\forall g : \mathbb{E}_{\mathscr{D}}[\ell(s_h, a_h, s_{h+1}, g)]$ is an unbiased est of $\mathscr{E}(g; f, h)$

↓ empirical average

V-Bellman rank setting

Recall our hypothesis class $\mathscr{F}$, where each $g \in \mathscr{F}$ is in the form of $g(s, a)$

For V-Bellman rank, we define Bellman error loss as: $V(s) = Q(s, \pi(s))$

$$\ell(s_h, a_h, s'_{h+1}, g) = \frac{\mathbf{1}\{a_h = \pi_g(s_h)\}}{1/A} \left(g(s_h, a_h) - r(s_h, a_h) - \max_{a'} g(s_{h+1}, a') \right)$$

If we had a dataset $\mathscr{D} := \{s_h, a_h, s_{h+1}\}$ where $s_h \sim d_h^{\pi_f}, a_h \sim U(\mathscr{A}),$
$s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$

then $\forall g : \mathbb{E}_{\mathscr{D}}[\ell(s_h, a_h, s_{h+1}, g)]$ is an unbiased est of $\mathscr{E}(g; f, h)$

Bilinear-UCB

At iteration t :

no bonus required

Select $f_t = \arg\max_{g \in \mathcal{F}} V_g(s_0)$

computationally efficient?

s.t., $\forall h : \sum_{i=0}^{t-1} \Big(\underbrace{\mathbb{E}_{\mathcal{D}_{h,i}}[\ell(s_h, a_h, s_{h+1}, g)]}_{\text{empirical Bellman error}} \Big)^2 \leq R^2$ $\equiv$ ellipsoid constraints for linear bandits

Bilinear-UCB

At iteration t :

avg Bellman error $\to 0$

Select $f_t = \arg\max_{g \in \mathscr{F}} V_g(s_0)$

s.t., $\forall h$: $\sum_{i=0}^{t-1} \left(\mathbb{E}_{\mathscr{D}_{h,i}}[\ell(s_h, a_h, s_{h+1}, g)] \right)^2 \leq R^2$

generative setting

For all h, create $\mathscr{D}_{h,t} = \{s_h, a_h, s_{h+1}\}$ w/ m triples, where:

- For Q-B rank case: $\underline{s_h, a_h} \sim \underline{d_h^{\pi_{f_t}}, s_{h+1}} \sim P_h(\cdot \mid s_h, a_h)$
- For V-B rank case: $s_h \sim d_h^{\pi_{f_t}}$, $a_h \sim U(A)$, $s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$

Bilinear-UCB

Select $f_t = \arg\max_{g \in \mathcal{F}} V_g(s_0)$ s.t., $\forall h : \sum_{i=0}^{t-1} \left(\mathbb{E}_{\mathcal{D}_{h,i}}[\ell(s_h, a_h, s_{h+1}, g)] \right)^2 \leq \underline{\underline{R^2}}$

bonus

1. When the batch size ($|\mathcal{D}_{h,i}|$) is large,

$$\mathbb{E}_{\mathcal{D}_{h,i}} \ell(s_h, a_h, s_{h+1}, g) \to \mathcal{E}(g; f_i, h)$$

2. We know that $\sum_{i=1}^{t-1} \mathcal{E}(f^\star; f_i, h) = 0$

linear bandit : $\theta^* \in C_t$ ellipsoid

3. By properly setting batch size and R, we eliminate wrong hypothesis, but keep $f^\star$

4. This gives optimism: $V_{f_t}(s_0) \geq V_{f^\star}(s_0) := V^\star(s_0)$

Optimism allows explore and exploit tradeoff!

Analysis of Bilinear-UCB

Uniform convergence style assumption on our hypothesis class $\mathscr{F}$:

Given any distribution $\nu \in \Delta(S \times A \times S)$, and m i.i.d samples $\{s_i, a_i, s_i'\}$ from ν,
w/ probability at least $1 - \delta$,

$$\forall g : \left| \underbrace{\mathbb{E}_\nu}_{\text{true}} \ell(s, a, s', g) - \underbrace{\mathbb{E}_{\mathscr{D}}}_{\text{empirical}} \ell(s, a, s', g) \right| \leq \varepsilon_{gen}(m, \mathscr{F}, \delta)$$

Example: when $\mathscr{F}$ is discrete (for B-rank loss), Hoeffding + union bound over $\mathscr{F}$ implies:

$$\varepsilon_{gen}(m, \mathscr{F}, \delta) := 2H\sqrt{\frac{\ln(|\mathscr{F}|/\delta)}{m}}$$