Fitted Q-Iteration

(most references can be found on paper list for project topics)

Generalization for value-based batch RL

- We studied using abstractions to generalize in large state spaces
- Abstractions correspond to “histogram regression” in supervised learning—the most trivial form of generalization
- Can I use XXX for value-based RL?
  - Linear predictors?
  - Kernel machines?
  - Random forests?
  - Neural nets???
  - …
- **What you really want**: *Reduction* of RL to supervised learning.

$\arg\min_f \mathbb{E}[(Y-f(x))^2]$.

(s,a,r,s').

$\downarrow$

(s,a), Y.

Revisiting value iteration

- Recall the value iteration algorithm: $\boxed{f_k \leftarrow \mathcal{T} f_{k-1}}$
  - where $(\mathcal{T} f)(s,a) = \mathbb{E}_{r \sim R(s,a), s' \sim P(\cdot|s,a)}[r + \gamma \max_{a'} f(s',a')]$
  - i.e., $\mathcal{T} f_{k-1} = \mathbb{E}[r + \gamma \max_{a'} f_{k-1}(s',a') \mid s, a]$
- What we want: a function in the form of $\boxed{\mathbb{E}[Y \mid X]} = \mathcal{T} f_{k-1}$
  - $\boxed{Y = r + \gamma \max_{a'} f_{k-1}(s',a')}$, $X = (s,a)$
  - How to obtain $\mathbb{E}[Y \mid X]$? Squared-loss regression!!!
- Fitted-Q Iteration [Ernst et al'05]

$f^* = \mathcal{T} f_{k-1}$

$$f_k = \arg\min_{f \in \mathcal{F}} \sum_{(s,a,r,s') \in D} \left(f(s,a) - \left(r + \gamma \max_{a' \in \mathcal{A}} f_{k-1}(s',a') \right) \right)^2$$

  - F = all functions: FQI = VI in the estimated tabular model
  - F = all piece-wise const functions under abstraction ϕ: FQI = VI in the estimated abstract model

$$(X_1, Y_1), (X_2, Y_2) \cdots (X_n, Y_n)$$

$$\boxed{(X, Y) \sim P.}$$

$$f^* = \underset{f: \mathcal{X} \to \mathbb{R}}{\operatorname{argmin}} \; \mathbb{E}_P[(f(X) - Y)^2].$$

$$\forall x \in \mathcal{X} \qquad f(x) = f_x.$$

$$\forall x. \quad \underset{f_x}{\operatorname{argmin}} \; \mathbb{E}_{Y \sim P(\cdot|x)}[(f_x - Y)^2].$$

$$\parallel$$

$$\mathbb{E}_{Y \sim P(\cdot|X=x)}\Big[(f_x - \mathbb{E}[Y|X=x])^2 + (\mathbb{E}[Y|X=x] - Y)^2\Big].$$

$$f_x = \mathbb{E}[Y|X=x].$$

$$\boxed{f^*(X) = \mathbb{E}[Y|X].}$$

Special case: MBRL (CE) with ϕ

$D_{x,a} = \bigcup_{s \in \phi^{-1}(x)} D_{s,a}.$

- Algorithm: estimate $\widehat{M}_\phi$, and do planning

$$\hat{R}_\phi(x,a) = \frac{1}{|D_{x,a}|}\sum_{(r,s')\in D_{x,a}} r, \quad \hat{P}_\phi(x,a) = \frac{1}{|D_{x,a}|}\sum_{(r,s')\in D_{x,a}} \mathbf{e}_{\phi(s')}$$

- Use Value Iteration as the planning algorithm:
  - Initialize g_0 as any function in $\mathbb{R}^{|\mathcal{S}_\phi \times \mathcal{A}|}$
  - $g_t \leftarrow \mathcal{T}_{\widehat{M}_\phi} g_{t-1}$. That is, for each $x \in S_\phi$, $a \in A$:

$\max_{a'} g_{t-1}(x', a')$

$$\begin{aligned} g_t(x,a) &= \widehat{R}_\phi(x,a) + \gamma\langle \widehat{P}_\phi(x,a), V_{g_{t-1}}\rangle \\ &= \frac{1}{|D_{x,a}|}\sum_{(r,s')\in D_{x,a}} \left(r + \gamma\langle \mathbf{e}_{\phi(s')}, V_{g_{t-1}}\rangle\right) \\ &= \frac{1}{|D_{x,a}|}\sum_{(r,s')\in D_{x,a}} \left(r + \gamma V_{g_{t-1}}(\phi(s'))\right) \end{aligned}$$

Rewrite in the original S

- Rewrite the algorithm so that $f_t = [g_t]_M$
- Define $\mathcal{F}^\phi \subset \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ as the space of all functions over $S\times A$ that are piece-wise constant under ϕ with value in $[0, V_{\max}]$
- Initialize f_0 as any function in F^ϕ
- For each $s \in S$, $a \in A$: essentially $f_t \leftarrow \mathcal{T}_{\widehat{M}'_\phi} f_{t-1}$

$g_0, g_1, g_2, g_3, \ldots \in \mathbb{R}^{|S_\phi \times A|}$

$$f_t(s,a) = \widehat{R}_\phi(\phi(s),a) + \gamma\langle \widehat{P}_\phi(\phi(s),a), [V_{f_{t-1}}]_\phi\rangle$$
$$= \frac{1}{|D_{\phi(s),a}|}\sum_{(r,s')\in D_{\phi(s),a}} (r + \gamma\langle \mathbf{e}_{\phi(s')}, [V_{f_{t-1}}]_\phi\rangle)$$
$$= \frac{1}{|D_{\phi(s),a}|}\sum_{(r,s')\in D_{\phi(s),a}} \underbrace{(r + \gamma V_{f_{t-1}}(s'))}_{\text{"Empirical Bellman update" (based on 1 data point)}}$$

$$g_t(x,a) = \widehat{R}_\phi(x,a) + \gamma\langle \widehat{P}_\phi(x,a), V_{g_{t-1}}\rangle$$
$$= \frac{1}{|D_{x,a}|}\sum_{(r,s')\in D_{x,a}} (r + \gamma\langle \mathbf{e}_{\phi(s')}, V_{g_{t-1}}\rangle)$$
$$= \frac{1}{|D_{x,a}|}\sum_{(r,s')\in D_{x,a}} (r + \gamma V_{g_{t-1}}(\phi(s')))$$

$$f_t(s,a) = \frac{1}{|D_{\phi(s),a}|} \sum_{(r,s') \in D_{\phi(s),a}} \left(r + \gamma \max_{a' \in \mathcal{A}} f_{t-1}(s',a') \right)$$

Alternative interpretation of the above step

- Dataset $D = \{(s, a, r, s')\}$
- Apply emp. Bellman up. to $f_{t\text{-}1}$ based on each data point:

$$\left\{ \left((s,a), (r + \gamma \max_{a' \in \mathcal{A}} f_{t-1}(s',a')) \right) \right\}$$

- What does it mean to take average over $D_{\phi(s),a}$?
  - Recall: average minimizes mean squared error (MSE)
  - *Projection* onto F^{ϕ}! (think of functions over D)

$$f_t = \arg\min_{f \in \mathcal{F}^{\phi}} \sum_{(s,a,r,s') \in D} \left(f(s,a) - \left(r + \gamma \max_{a' \in \mathcal{A}} f_{t-1}(s',a') \right) \right)^2$$

  - … which is, solving a SL regression problem with histogram regression F^{ϕ}

Fitted Q-Iteration (FQI): $f_t = \arg\min_{f \in \mathcal{F}} \sum_{(s,a,r,s') \in D} \left(f(s,a) - \left(r + \gamma \max_{a' \in \mathcal{A}} f_{t-1}(s',a') \right) \right)^2$

[Ernst et al'05]; see also [Gordon'95]

Δ $\mathcal{T} f_{t-1}$

We simplified a "regression algorithm" to its corresponding function space F

- Empirical Risk Minimization (ERM); assume optimization is exact; does not consider regularization, etc.
- Will also assume finite (but exponentially large) F
  - continuous spaces are often handled by discretization in SLT (e.g., growth function, covering number)
  - methods like regression trees have dynamic function spaces (and often need SRM); not accommodated
- A minimal but (hopefully) insightful simplification of supervised learning

Fitted Q-Iteration (FQI): $f_t = \arg\min_{f\in\mathcal{F}} \sum_{(s,a,r,s')\in D} \left(f(s,a) - \left(r + \gamma \max_{a'\in\mathcal{A}} f_{t-1}(s',a') \right)\right)^2$

[Ernst et al'05]; see also [Gordon'95]

$f_{t-1}. \longrightarrow f_t.$ f_θ

Asynchronous update + stochastic approximation?

- Assume parameterized & differentiable function: $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$
- Online regression: randomly pick a data point and do a stochastic gradient update:

Treat as constant; don't pass gradient

$$\theta \leftarrow \theta - \frac{\alpha}{2} \cdot \nabla_\theta \left(f_\theta(s,a) - \left(r + \gamma \max_{a'\in\mathcal{A}} \overbrace{f_\theta(s',a')} \right)\right)^2$$

$$= \theta - \alpha \left(f_\theta(s,a) - \left(r + \gamma \max_{a'\in\mathcal{A}} f_\theta(s',a') \right)\right) \nabla_\theta f_\theta(s,a)$$

- If f_θ is the tabular function, it's (tabular) Q-learning
- If f_θ is a neural net, it's (almost) DQN (Mnih et al.'15)
  - Using a target network is even more similar to FQI

Fitted Q-Iteration (FQI): $f_t = \arg\min_{f \in \mathcal{F}} \sum_{(s,a,r,s') \in D} \left(f(s,a) - \left(r + \gamma \max_{a' \in \mathcal{A}} f_{t-1}(s', a') \right) \right)^2$

[Ernst et al'05]; see also [Gordon'95]

The argmin step plays two roles:

1. Denoise the emp update $r + \gamma V_f(s')$ to $(\mathcal{T} f)(s, a)$ (w/ inf data)
   - This happens even in tabular setting
2. $\mathcal{T} f$ may not have a succinct representation => find the closest approximation in F *(i.e., projection)*
   - Denote $\prod_F$ as the projection. Dependence on weights over state-action pairs omitted—determined by data distribution
   - With infinite data, FQI becomes: $f_t \leftarrow \prod_F \mathcal{T} f_{t-1}$

Convergence and Stability

- With infinite data, Q^* is a fixed point (as long as $Q^* \in F$)
  - $Q^* \in F$ is called (Q^*-)"*realizability*"
- CE w/ Q*-irrelevant ϕ is a special case of FQI—convergence guaranteed
- Doesn't hold in general: FQI may diverge under $\boxed{Q^* \in F,}$ even with
  - Infinite data
  - Fully exploratory data
  - Linear function class F
  - MDP has no actions (just policy evaluation)

$(s, a) \sim \mu, \; r, s$

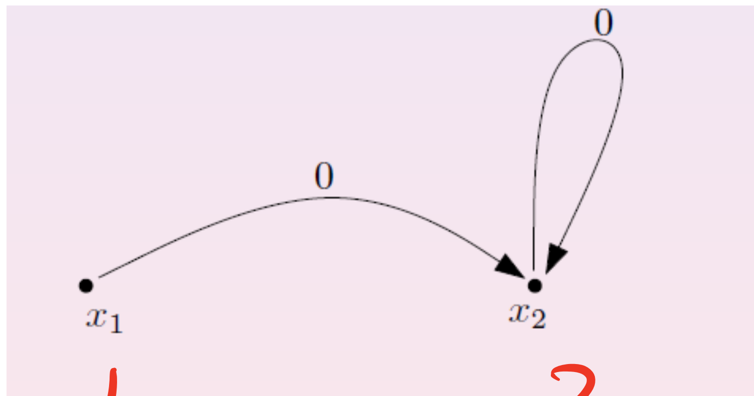
$\mu(s, a) > 0.$

2.1 Counter-example for least-square regression [Tsitsiklis and van Roy, 1996]

An MDP with two states x_1, x_2, 1-d features for the two states: $f_{x_1} = 1, f_{x_2} = 2$. Linear Function approximation with $\tilde{V}_\theta(x) = \theta f_x$.

$V(x_2) = \phi(x_2)\theta$

$\downarrow$ 2



1 2

$V(x_1)$

$= \phi(x_1) \cdot \theta.$

$\downarrow$

credit: course notes from Shipra Agrawal

1.

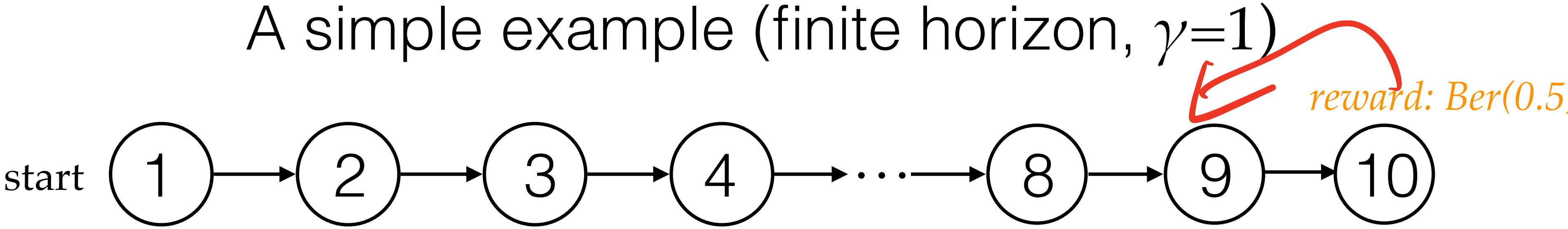
$$
\begin{aligned}
\theta_k &:= \arg\min_\theta \frac{1}{2}(\theta - \text{target}_1)^2 + (2\theta - \text{target}_2)^2 \\
&= \arg\min_\theta \frac{1}{2}(\theta - \gamma\theta^{k-1} f_{x_2})^2 + (2\theta - \gamma\theta^{k-1} f_{x_2})^2 \\
&= \arg\min_\theta \frac{1}{2}(\theta - \gamma 2\theta^{k-1})^2 + (2\theta - \gamma 2\theta^{k-1})^2
\end{aligned}
$$

$$(\theta - \gamma 2\theta^{k-1}) + 2(2\theta - \gamma 2\theta^{k-1}) = 0 \Rightarrow 5\theta = 6\gamma\theta^{k-1}$$

$$\theta_k = \frac{6}{5}\gamma\theta_{k-1}$$

This diverges if $\gamma \geq 5/6$.

A simple example (finite horizon, γ=1)



FQI Iter #1: Data: (⑩, 1, *end*), …, (⑩, 0, *end*) ⇨ $f_1(10) =$ 0.501

Iter #2: Data: (⑨, 0, ⑩) ⇨ (⑨, 0+0.501) ⇨ 0.501 0.501

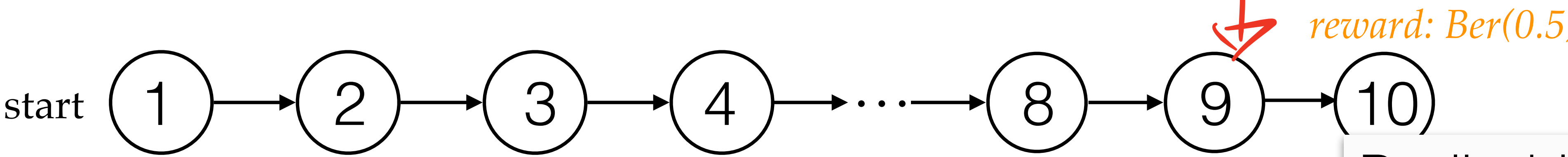
$r + f_{k-1}(s')$

…

Iter #10: 0.501 0.501 0.501 0.501 … 0.501 0.501 0.501

- Dataset $D = \{(s, r, s')\}$ looks like (action omitted):
  {(①, 0, ②), (②, 0,③), …, (⑩, 1, *end*), …, (⑩, 0, *end*)}

How things go wrong (w/ restricted class)



Realizable

Function class: 0.5 × 1.012 | 0.5 × 0.756 | 0.5 × 0.628 | … × | 0.5 × 0.504 | 0.5 × 0.502 | 0.5 × 0.501

$2^H \Delta$.

FQI Iter #1: Data: (⑩, 1, *end*), …, (⑩, 0, *end*) ⇨ 0.501

Iter #2: Data: (⑨, 0, ⑩) ⇨ (⑨, 0+0.501) ⇨ 0.502 0.501

… (8, 0.502)

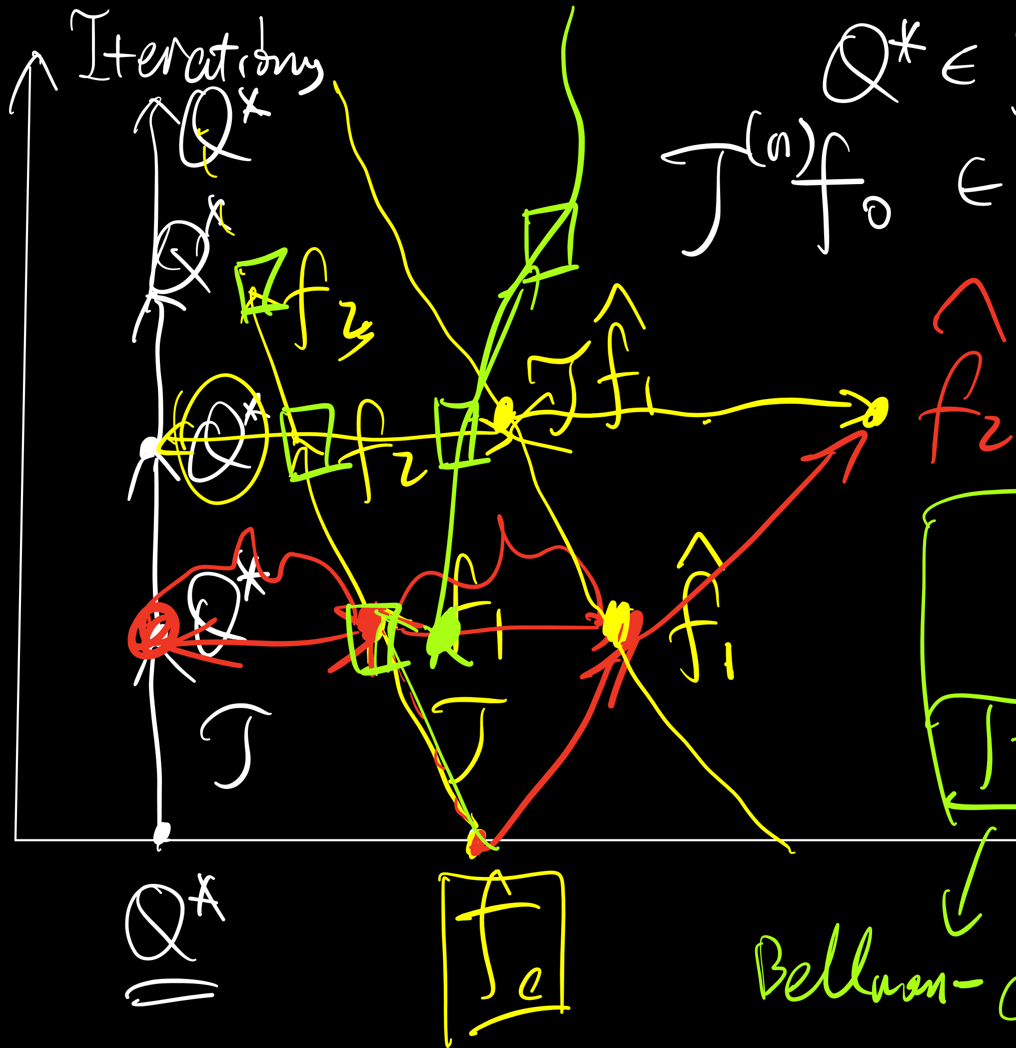
!!!

Iter #10: 1.012 0.756 0.628 … 0.502 0.501

Example given in Dann et al’18

$Q^* \in \mathcal{F}$.

$\mathcal{T}^{(n)} f_0 \in \mathcal{F}, \ \forall n.$



$\forall f \in \mathcal{F}$.
$\mathcal{T}f \in \mathcal{F}$.

Bellman-completeness/closure.

Intuition for the instability

- Standard VI: $f_t \leftarrow \mathscr{T} f_{k-1}$
- FQI keeps things tractable by: $f_t \leftarrow \Pi_{\mathscr{F}}(\mathscr{T} f_{k-1})$
  - $\prod_F$ can destroy contraction of $\mathscr{T}$!
  - Preserved only in special cases (e.g., Q*-irrelevant ϕ)
- A sufficient condition that fixes the issue

Bellman completeness (closure)

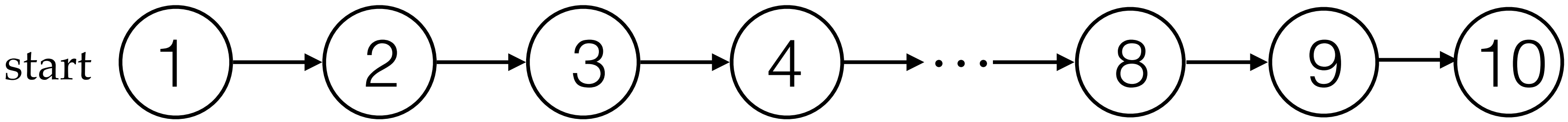
$$\mathscr{T} f \in \mathscr{F}, \forall f \in \mathscr{F}$$

*introduced by Szepesvari & Munos [2005]

  - whatever f_{k-1} is used, regression is always well-specified
  - Implies realizability for finite class (why?)
  - For piecewise const F, completeness = bisimulation (hw)
  - Not necessarily converge, but will get close to a good solution (under additional data assumptions)

How completeness fixes the issue

reward: Ber(0.5)



Function class: 0.5, 0.5

0.628 × $\boxed{0.628}$

0.1

0.628

$\mathcal{T}f_{t-1}$

- More generally: issue goes away if the regression problem

$$\left\{\left((s,a),\left(r+\gamma \max_{a'\in\mathcal{A}} f_{t-1}(s',a')\right)\right)\right\}$$

is realizable with F, for any $f_{t-1} \in F$

$\forall f \in \mathcal{F}$, $\mathcal{T}f \in \mathcal{F}$.

- In finite-horizon setting: the richer function class you use at a lower level, the more difficult to satisfy realizability at higher level
- In discounted setting: F closed under Bellman update—adding functions can hurt representation

Alternative approach

$R(s,a) + \gamma \mathbb{E}_{s'|s,a}[\max_{a'} f(s',a')]$.

$\min_\theta L(\theta; D)$

$\cancel{(f - Q^*)^2}$

- FQI is an iterative alg in its nature
  - not optimizing a fixed objective function!
  - objective changes as current f changes
- Alternative: minimize $\boxed{\| f - \boxed{\mathcal{T}} f \|}$ over $f \in F$

$f = Q^* \Leftrightarrow f = \mathcal{T} f$.

  - Is it equivalent to minimizing?

$(s,a) \sim \mu$.

r, s'.

$$\mathbb{E}_{\boxed{(s,a)\sim\mu,\; r\sim R(s,a),\; s'\sim P(s,a)}} \left[\left(f(s,a) - \underline{(r + \gamma \max_{a'} f(s',a'))}\right)^2\right]$$

(omitted in the rest of slides)

$\mathbb{E}[\,\cdot\,|\, s,a] = \mathcal{T} f$.

$$\underbrace{\mathbb{E}[(Y - f(X))^2]}_{=} = \mathbb{E}[(Y - \mathbb{E}[Y|X])^2] + \underbrace{\mathbb{E}[(\mathbb{E}[Y|X] - f(X))^2]}.$$

Bellman error minimization

$$\mathbb{E}_{(s,a)\sim\mu}\left[\left(f(s,a) - (r + \gamma \max_{a'} f(s',a'))\right)^2\right]$$

$$= \mathbb{E}_{(s,a)\sim\mu}\left[(f(s,a) - (\mathcal{T}f)(s,a))^2\right] + \mathbb{E}_{(s,a)\sim\mu}\left[\left((\mathcal{T}f)(s,a) - (r + \gamma \max_{a'} f(s',a'))\right)^2\right]$$

γ $Q^*(s,a) - (r + \gamma \cdots)$

This part is what we want: $\|f - \mathcal{T}f\|$, with a weighted 2-norm defined w/ ν

This part is annoying!

- Prefer "flat" f
- Q^* is not necessarily flat!
- 0 for deterministic transitions. Issue is only serious when env highly stochastic

Unbiased estimate "double sampling"

Baird '95

(s, a, r_A, s'_A)
s, a, r_B, s'_B

Workaround #1

- For $(s, a)\sim\mu$, if we can obtain **2** i.i.d. copies of (r, s') (copy A & B):

$$\left(f(s,a) - \left(r_A + \gamma \max_{a'\in\mathcal{A}} f(s'_A, a')\right)\right)\left(f(s,a) - \left(r_B + \gamma \max_{a'\in\mathcal{A}} f(s'_B, a')\right)\right)$$

- Only doable in simulators w/ resets…

if $\mathcal{T}f \in \mathcal{G}$ $\quad \min_{g \in \mathcal{G}} \mathbb{E}[(g(s,a) - (r + \gamma \max_{a'} f(s',a')))^2]$

Bellman error minimization

$$\mathbb{E}_{(s,a)\sim\mu}\left[\left(f(s,a) - (r + \gamma \max_{a'} f(s',a'))\right)^2\right]$$

$$= \mathbb{E}_{(s,a)\sim\mu}\left[(f(s,a) - (\mathcal{T}f)(s,a))^2\right] + \mathbb{E}_{(s,a)\sim\mu}\left[\left((\mathcal{T}f)(s,a) - (r + \gamma \max_{a'} f(s',a'))\right)^2\right]$$

This part is what we want: $\|f - \mathcal{T}f\|$, with a weighted 2-norm defined w/ ν

This part is annoying!

- Prefer "flat" f
- Q^* is not necessarily flat!
- 0 for deterministic transitions. Issue is only serious when env highly stochastic

Workaround #2

- Estimate the 2nd part, and subtract it from LHS
- Antos et al'08: $\arg\min_{f \in \mathcal{F}} L(f)$. where $L(f) :=$

$$\mathbb{E}_{(s,a)\sim\mu}\left[\left(f(s,a) - \left(r + \gamma \max_{a' \in \mathcal{A}} f(s',a')\right)\right)^2\right] - \min_{g \in \mathcal{G}} \mathbb{E}_{(s,a)\sim\mu}\left[\left(g(s,a) - \left(r + \gamma \max_{a' \in \mathcal{A}} f(s',a')\right)\right)^2\right]$$

argmin max

$Q^* \in \mathcal{F}$, $\forall f \in \mathcal{F},\ \mathcal{T}f \in \mathcal{G}$.

Bellman error minimization

$$\arg\min_{f\in\mathcal{F}}\max_{g\in\mathcal{G}}\left(\mathbb{E}_{(s,a)\sim\mu}\left[\left(f(s,a)-\left(r+\gamma\max_{a'\in\mathcal{A}}f(s',a')\right)\right)^2-\left(g(s,a)-\left(r+\gamma\max_{a'\in\mathcal{A}}f(s',a')\right)\right)^2\right]\right)$$

- Fix any f, the first squared error is constant; second square is a regression problem w/ Bayes optimal being $\mathcal{T}f$
- So, if G is rich enough to contain $\mathcal{T}f$ for all f, this works!
  - and w/ a consistent optimization objective, unlike FQI
- If G is not rich enough, may under-estimate the Bellman error of some f (subtracting too much)
- FQI: When G=F, this is just Bellman completeness again!

train. $\hat{Q}_1$, $\hat{Q}_2$

Val data $\hat{Q}_1$? $\hat{Q}_2$

$\hat{Q}_1 - \mathcal{T}\hat{Q}_1$?

$\hat{Q}_2 - \mathcal{T}\hat{Q}_2$?

f_1 f_2.

$\exists f \in \mathcal{F}$ $\forall f \in \mathcal{F}$.

One last assumption: data

- Recall that data needs to be exploratory for batch RL
- What does it actually mean?
  - tabular: relatively uniform over state space
  - abstraction: relatively uniform over abstract state space
  - large/continuous state space: uniform? in what measure??

iid (s, a, r, s').

$(s, a) \sim \boxed{\mu}$.

$r \sim R(s, a)$ $s' \sim P(\cdot \mid s, a)$

$J(\pi^*) - J(\hat{\pi})$

Assumption on data: "Concentrability"

$\boxed{d_0}$

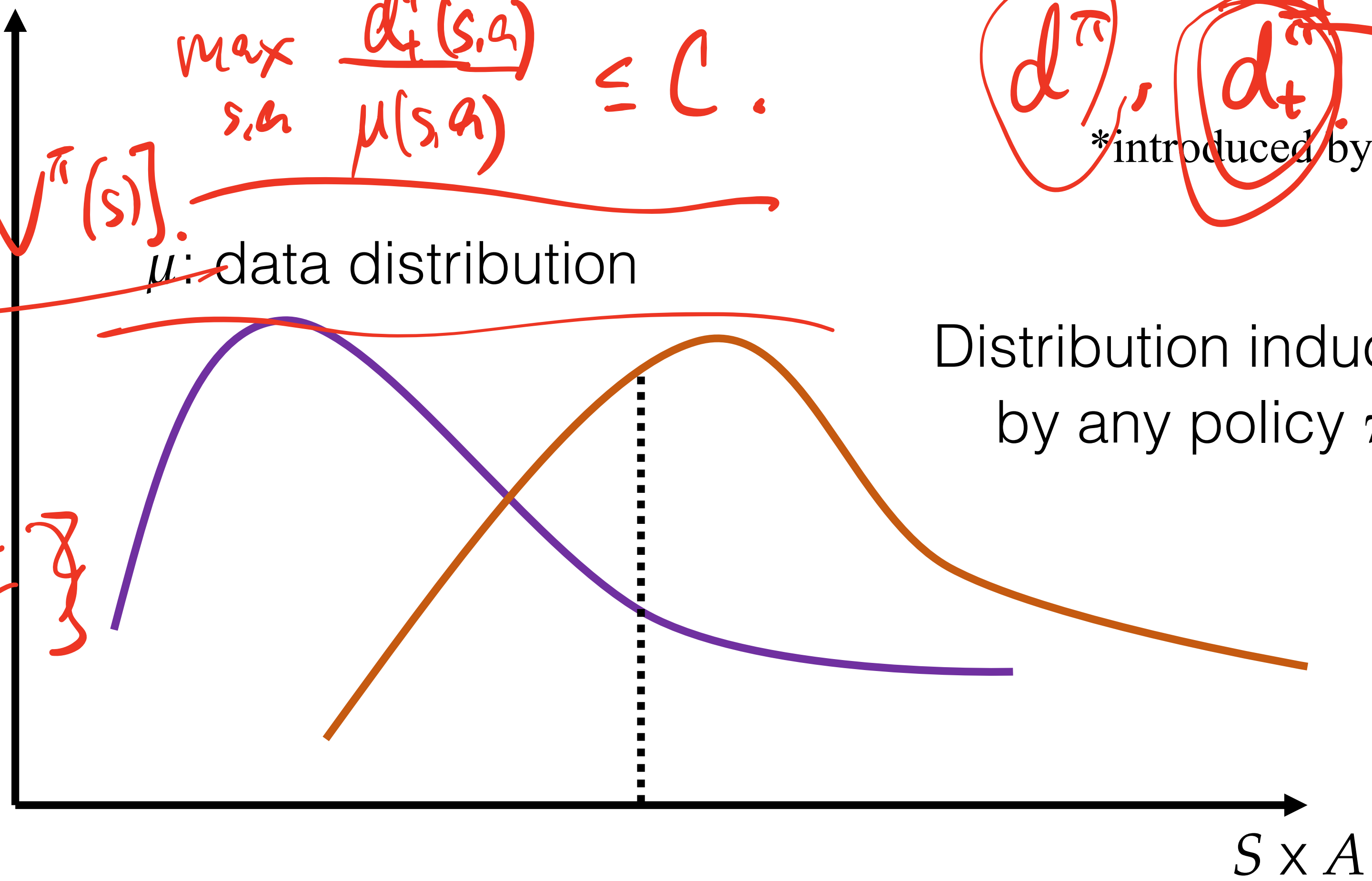
$d^\pi,\ d_t^\pi$

*introduced by Munos'05

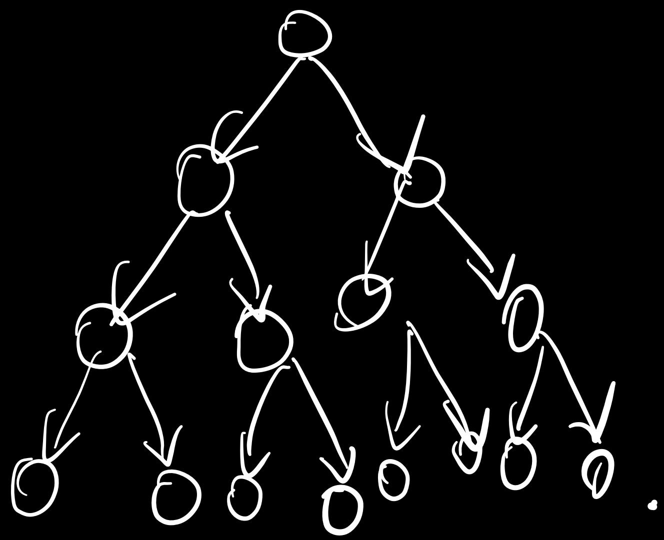
$J(\pi) = \mathbb{E}_{s \sim d_0}[V^\pi(s)]$

$\max_{s,a} \frac{d_t^\pi(s,a)}{\mu(s,a)} \le C.$

$\{\pi_f : f \in \mathcal{F}\}$

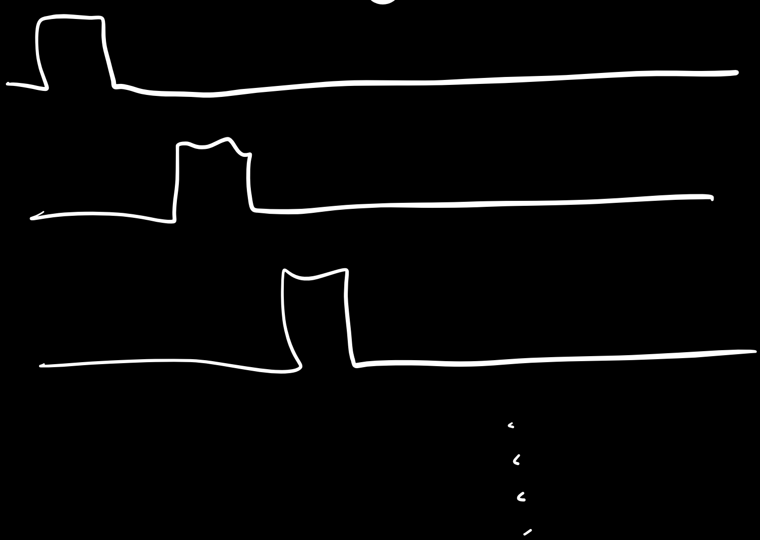


- Let C be a uniform upper bound on the density ratio
- Assumption: C is small (= allow polynomial dependence on C)
- Previous exponential lower bound is "explained away" by an exponentially large C



$H \qquad 2^H.$

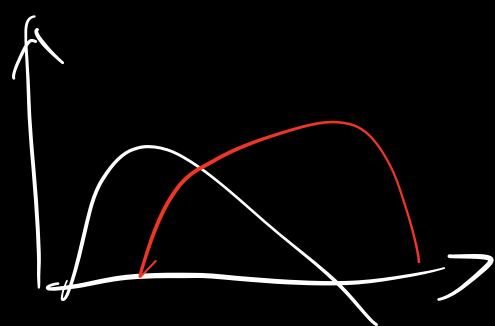
$\longrightarrow t = H.$



$d_H^{\pi}.$

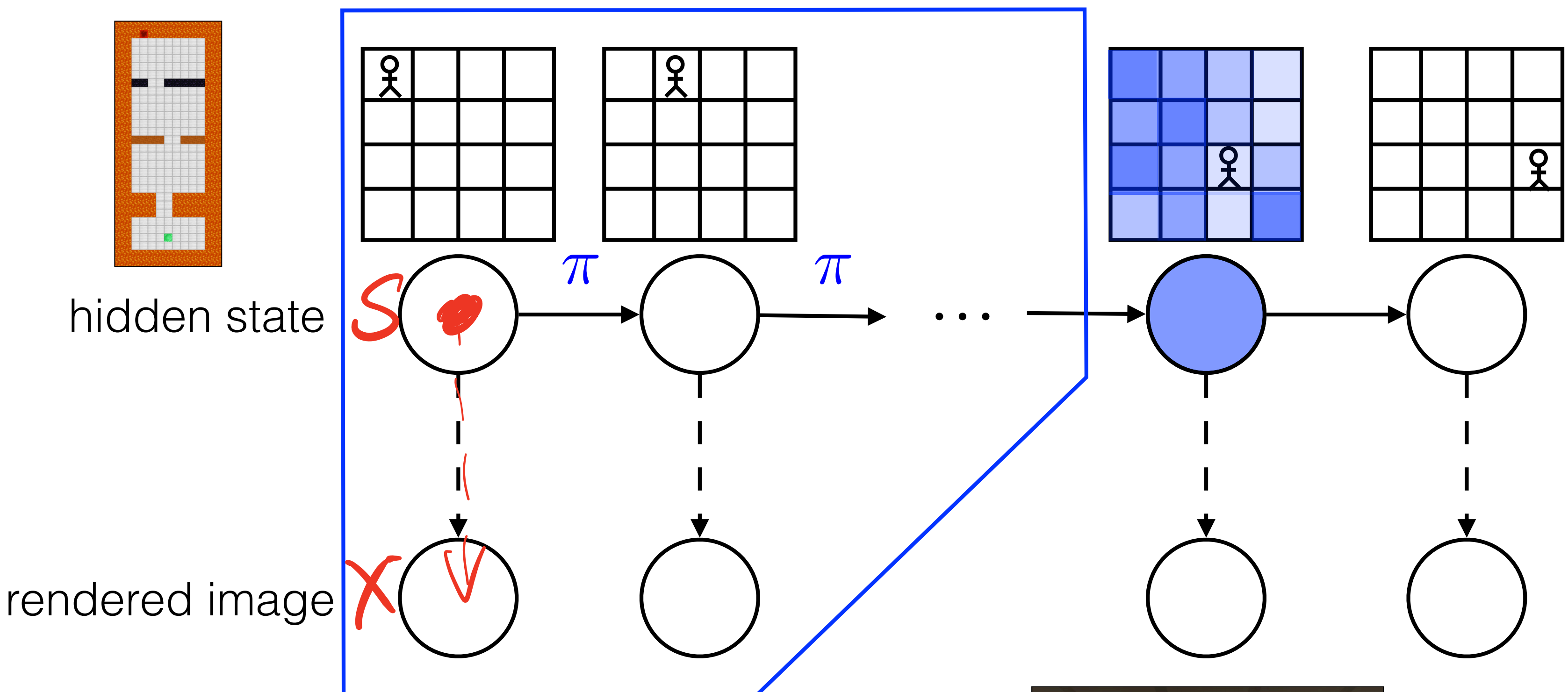
$$\max_{s,\pi} \frac{d_H^{\pi}(s)}{\mu_H(s)} \leq 2^H$$

$$= \max_{\pi} \left\| \frac{d_H^{\pi}}{\mu_H} \right\|_\infty.$$



$$\left\| \frac{d^{\pi^*}}{\mu} \right\|_\infty$$

Concentrability: when is it small?



Connections to the assumptions needed for efficient exploration [Jiang et al'17]

Markovian high-dimensional observation

$$\frac{d^{\pi}(x)}{\mu(x)} = \frac{d^{\pi}(s)\cdot e[x|s]}{\mu(s)\cdot e(x|s)}$$

Remainder of this part

Prove the $poly(H, \log|F|, C)$ result for FQI

Remainder of this part

Prove the *poly*(H, log|F|, C) result for FQI

FQE: $f_0 \equiv \vec{0}$.

$$\mathbb{E}[\, \mathcal{T} f_{k-1} \mid s, a)$$

$$f_k = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{|D|} \sum_{(s,a,r,s')} \left(f(s,a) - r - \gamma f_{k-1}(s', \pi) \right)^2.$$

$f_K \approx Q^\pi$.

$J(\pi) = \mathbb{E}_{s \sim d_0}[Q^\pi(s, \pi)]$.

$\wr\wr$

$\hat{J}(\pi) = \mathbb{E}_{s \sim d_0}[f_K(s, \pi)]$.

$|J(\pi) - \hat{J}(\pi)|$.

$\mathcal{T}^\pi f \in \mathcal{F}, \; \forall f \in \mathcal{F}$.

FQE: $f_0 \equiv \vec{0}$.

$$f_k = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{|D|} \sum_{(s,a,r,s')} \left(f(s,a) - r - \gamma \, f_{k-1}(s',\pi) \right)^2 \qquad \mathcal{T}^\pi f_{k-1}$$

(1) Want to show $f_k \approx \mathcal{T}^\pi f_{k-1}$.

$$\| f_k - \mathcal{T}^\pi f_{k-1} \|_2 \leq \text{small}$$

$$\Rightarrow \mathbb{E}_\mu \left[\left(f_k - \mathcal{T}^\pi f_{k-1} \right)^2 \right] \leq \text{small} \quad \forall k.$$

$$= \mathbb{E}_\mu \left[\left(f_k(s,a) - r - \gamma f_{k-1}(s',\pi) \right)^2 \right]$$

$$- \mathbb{E}_\mu \left[\left((\mathcal{T} f_{k-1})(s,a) - r - \gamma f_{k-1}(s',\pi) \right)^2 \right].$$

$$L_\mu(\cdot\,; f_{k-1}) = \mathbb{E}_\mu \left[\left((\cdot) - r - \gamma f_{k-1}(s',\pi) \right)^2 \right].$$

$$L_D(\cdot\,; f_{k-1}) = \frac{1}{|D|} \sum_{(s,a,r,s')} \left(\qquad \right)^2.$$

fix any f, f'. w.p. $\geq 1-\delta$.

$$\left| L_\mu(f; f') - L_D(f; f') \right| \leq O\left(V_{max} \sqrt{\frac{1}{n} \ln \frac{1}{\delta}}\right).$$

$$f_k = \arg\min_{f \in \mathcal{F}} L_D(f; f_{k-1}).$$

$$\mathcal{T}^{a} f_{k-1} = \arg\min_{f \in \hat{\mathcal{F}}} L_\mu(f; f_{k-1}).$$

$$\mathbb{E}_\mu\left[(f_k - \mathcal{T} f_{k-1})^2\right]$$

$$\leq L_\mu(f_k; f_{k-1}) - L_\mu(\mathcal{T} f_{k-1}; f_{k-1}).$$

$$- L_D(f_k; f_{k-1}) + L_D(\mathcal{T} f_{k-1}; f_{k-1})$$

$$\leq 2 \cdot O\left(V_{max} \sqrt{\frac{1}{n} \ln \frac{|\mathcal{F}|^2}{\delta}}\right).$$

$$\mathbb{E}_{(s,a) \sim \mu}\left[\left(f_k(s,a) - (\mathcal{T}^{\pi_k} f_{k-1})(s,a)\right)^2\right] \leq$$

$$\mathbb{E}_\mu[(\cdot)^2] = \|\cdot\|_{2,\mu}^2 \qquad \|\cdot\|_{L^2(\mu)}$$

$$\left(\mathbb{E}_\mu[(\cdot)^p]\right)^{1/p} =: \boxed{\|\cdot\|_{p,\mu}}$$

$$\to \|\cdot\|_1 \geq \|\cdot\|_2 \geq \|\cdot\|_\infty .$$

$$\|\cdot\|_{1,\mu} \leq \|\cdot\|_{2,\mu} \leq \|\cdot\|_{\infty,\mu} .$$

$\|$ if fully supp.

if $v \in \mathbb{R}^d$, chose $\mu = \text{unif}([d])$.

$[d] = \{1, 2, \dots, d\}$.

$$\|v\|_1 = |v_1| + |v_2| + \dots + |v_d| .$$

$$\|v\|_{1,\mu} = \mu_1 |v_1| + \mu_2 |v_2| + \dots + \mu_d |v_d| .$$

$$= \frac{1}{d} \left(\sum_i |v_i| \right) .$$

$$\|v\|_{p,\mu} \leq \varepsilon . \qquad \|v\|_{p,\rho} \leq \left(\left\| \frac{\rho}{\mu} \right\|_\infty \right)^{1/p} \varepsilon .$$

$$= \left(\sum_{i=1}^{d} \mu(i) \, |v(i)|^p \right)^{1/p}$$

$$\left|\hat{J}(\bar{\pi}) - J(\pi)\right|$$

$$= \left| \mathbb{E}_{s\sim d_0}\left[f_k(s,\underline{\underline{\pi}})\right] - \mathbb{E}_{s\sim d_0}\left[Q^{\pi}(s,\pi)\right]\right|$$

$$= \Big| \underbrace{\mathbb{E}_{\substack{s\sim d_0 \\ a\sim\pi}}\left[f_k(s,a) - (\mathcal{T}^{\bar{\pi}} f_{k-1})(s,a)\right]}_{\mathrm{I}} - \underbrace{\mathbb{E}_{\substack{s\sim d_0 \\ a\sim\pi \\ r,s'}}\left[r + \gamma Q^{\pi}(s',\pi)\right]}_{\mathrm{III}} \Big|$$

$$d_2^{\pi}$$

$$\mathrm{II}: \quad + \mathbb{E}_{\substack{s\sim d_0 \\ a\sim\bar{\pi} \\ r,s'}}\left[r + \gamma f_{k-1}(s',\pi)\right]$$

$$\mathrm{I} \le \mathbb{E}_{d_0\times\pi}\left[\left|f_k - \mathcal{T}^{\bar{\pi}} f_{k-1}\right|\right]$$

$$\le \|f_k - \mathcal{T}^{\bar{\pi}} f_{k-1}\|_{2, d_0\times\pi}$$

$$\le \sqrt{\left\|\frac{d_0\times\pi}{\mu}\right\|_\infty} \cdot \|f_k - \mathcal{T}^{\bar{\pi}} f_{k-1}\|_{2,\mu}.$$

$$(s,a)$$

$$\boxed{\left\|\frac{d_t^{\bar{\pi}}}{\mu}\right\|_\infty \le C \quad \forall t.}$$

$$(\mathrm{II}) - (\mathrm{III}) = \gamma\, \mathbb{E}_{s'\sim d_2^{\pi}}\left[f_{k-1}(s',\pi) - Q^{\pi}(s',\pi)\right].$$

	Data	Function approximation
AVI	$\max_{\pi} \lVert d^{\pi}/d^{D} \rVert_{\infty} \le C$	$\mathcal{T} f \in \mathcal{F}, \forall f \in \mathcal{F}$ [Munos & Szepesvari'08]
API		$\mathcal{T}^{\pi} f \in \mathcal{F},\ \forall f \in \mathcal{F}, \pi \in \Pi$ [Antos et al '08]

- Assumption so far: data is exploratory (e.g., $\max_{\pi} \lVert d^{\pi}/\mu \rVert_{\infty} \le C$)
- Challenge: real-world data often lacks exploration!
  - Data may not contain all bad behaviors
  - Alg may over-estimate their performance

How to understand a driving behavior is unsafe, if all data are safe?

Data with insufficient coverage

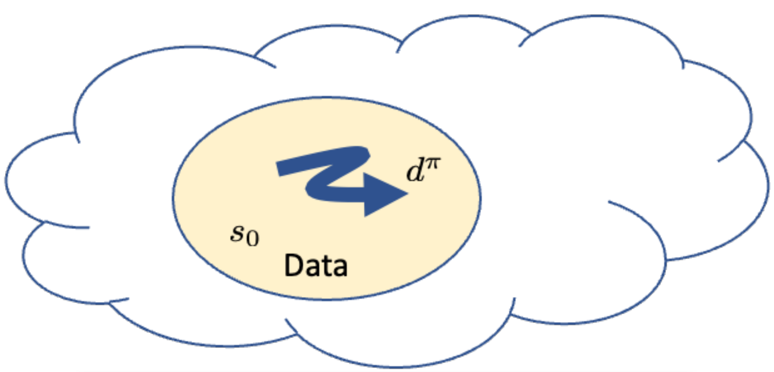
- Policy optimization: $\arg\max_{\pi \in \Pi} J(\pi) := Q^\pi(s_0, \pi)$
  - Q^π: value function; s_0: initial state; Π: policy class
- Considerations in estimating $\widehat{J}(\pi)$?

$$\arg\max_{\pi \in \Pi} \widehat{J}(\pi)$$

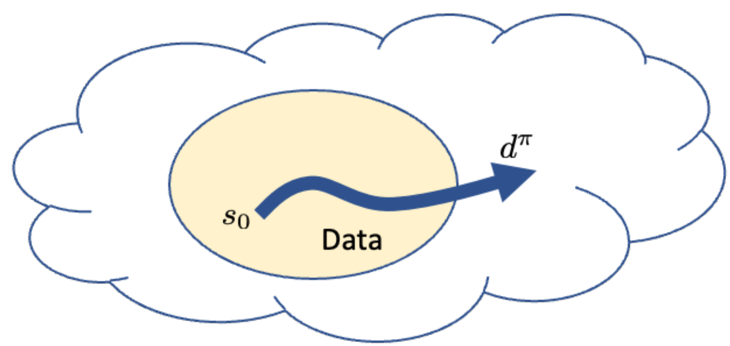
Pessimism in face of uncertainty

$$\widehat{J}(\pi) \approx J(\pi)$$

$$\widehat{J}(\pi) \leq J(\pi)$$



Policy covered by data



Policy not covered by data

Handle two cases simultaneously

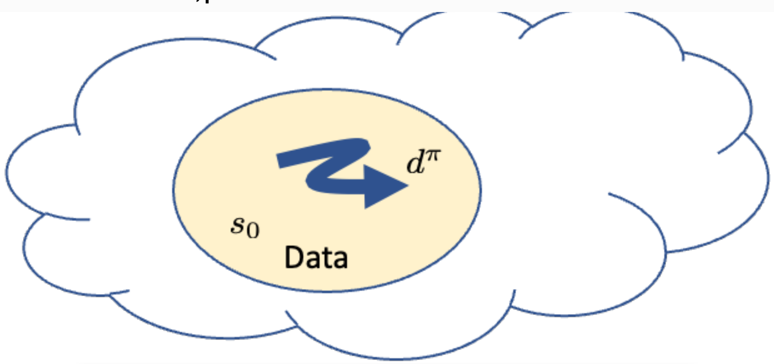
- Consider $\mathscr{F}^{\pi}_{\epsilon} := \{f \in \mathscr{F} : \|f - \mathscr{T}^{\pi} f\|_{2,\mu} \le \epsilon\}$ “Confidence set”/“Version space”
  - small $\|f - \mathscr{T}^{\pi} f\|_{2,\mu}$ implies $f(s_0, \pi) \approx J(\pi) = Q^{\pi}(s_0, \pi)$ if μ covers d^{π}
  - can estimate $|f(s_0,\pi) - J(\pi)| \le \frac{1}{1-\gamma}\|f - \mathcal{T}^{\pi} f\|_{2,d^{\pi}}$ ator) under “Bellman-completeness” $\mathscr{T}^{\pi} f \in \mathscr{F}, \forall f \in \mathscr{F}$
- Key observation: Q^{π} is in the set ($Q^{\pi} - \mathscr{T}^{\pi} Q^{\pi} \equiv 0$)
- **Pessimistic** policy evaluation

$$\hat{J}(\pi) := \min_{f \in \mathscr{F}^{\pi}_{\epsilon}} f(s_0, \pi) \le Q^{\pi}(s_0, \pi) = J(\pi)$$

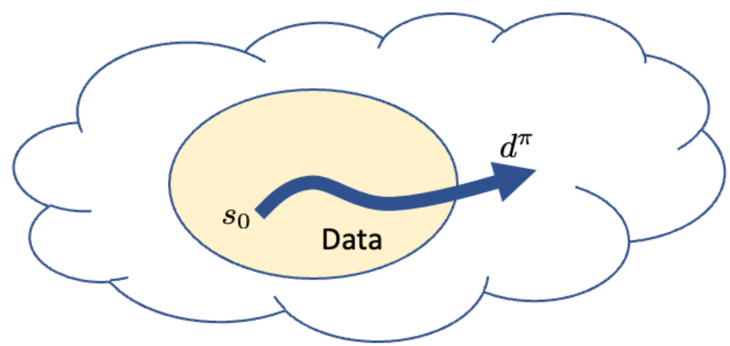
$$\Downarrow$$

$$\widehat{J}(\pi) \le J(\pi)$$

All members of $\mathscr{F}^{\pi}_{\epsilon}$ have small $\|f - \mathscr{T}^{\pi} f\|_{2,\mu}$, so $\hat{J}(\pi) \approx J(\pi)$ for covered π



Policy covered by data



Policy not covered by data

	Data	Function approximation
AVI	$\max_{\pi} \|d^\pi / d^D\|_\infty \leq C$	$\mathcal{T} f \in \mathcal{F}, \forall f \in \mathcal{F}$ [Munos & Szepesvari'08]
API		$\mathcal{T}^\pi f \in \mathcal{F}, \ \forall f \in \mathcal{F}, \pi \in \Pi$ [Antos et al '08]
Pessimism	$\|d^{\pi^\star} / d^D\|_\infty \leq C$	$\mathcal{T}^\pi f \in \mathcal{F}, \ \forall f \in \mathcal{F}, \pi \in \Pi$ [Xie et al '21]

- Guarantee: $\hat{\pi} = \arg\min_{\pi \in \Pi} \hat{J}(\pi)$ competes with any covered policy $\pi_{ref} \in \Pi$
  - $J(\hat{\pi}) \geq \hat{J}(\hat{\pi}) \geq \hat{J}(\pi_{ref}) \approx J(\pi_{ref})$
  - Near-optimality follows if $\pi^\star$ is covered
- Alternative: pointwise pessimism (construct $\widehat{Q}^\pi(s,a) \leq Q^\pi(s,a) \quad \forall s, a$)
  - Insert negative bonus in Bellman backup [Jin et al'21]
  - Density estimation + pessimistic in low-density area [Liu et al'20]