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?
  - Kernal machines?
  - Random forests?
  - Neural nets???
  - …
- **What you really want**: *Reduction* of RL to supervised learning.
- And YES! We have already seen a special case of the algorithm…

Revisit MBRL (CE) with ϕ

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

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

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

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

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 $[0, R_{\max}/(1\text{-}\gamma)]$
- Initialize f_0 as any function in F^ϕ
- For each $s \in S$, $a \in A$: essentially $f_t \leftarrow \mathcal{T}_{\widehat{M}'_\phi} f_{t-1}$

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

"Empirical Bellman update" (based on 1 data point)

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

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

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}
  - Reduction done! Plug in your favorite SL method

- 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]

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

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

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

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

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

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

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

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 be manageable => 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: $f_t \leftarrow \prod_F \mathcal{T} f_{t-1}$

realizability.

↑

if $Q^* \in F$.

Convergence?

$f_t \leftarrow \Pi_F \mathcal{T} f_{t-1}$.

$f_{t-1} = Q^*$.

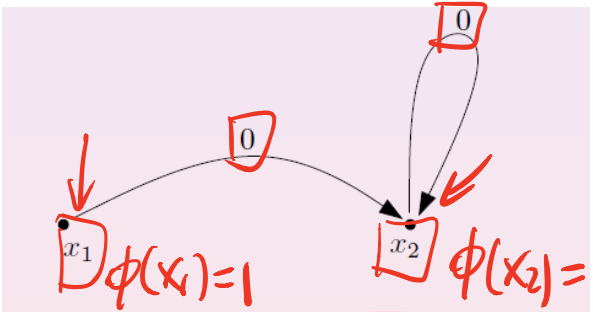
$\Pi_F \mathcal{T} Q^*$

$= \Pi_F Q^* = Q^*$.

- First: with infinite data, Q^* is a fixed point (as long as $Q^* \in F$)
- We know: CE w/ Q*-irrelevant abstractions is a special case of FQI—convergence guaranteed. [Gordon'95].
- But: not necessarily with general function spaces! Can diverge even with linear functions, under realizability (i.e., $Q^* \in F$)

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 $\tilde{V}_\theta(x) = \theta f_x$.



uncontrolled.

linear. 1-d.

$Q^* \equiv 0$

$Q^*(x_1) = \phi(x_1) \cdot 0$

credit: course notes from Shipra Agrawal

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

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

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

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

$Q^*(x_2) = \phi(x_2) \cdot 0$

$\forall x$

$Q^*(x) = \theta^T \phi(x)$

$\theta = 0$.

Bellman error minimization

- Standard VI: $f_t \leftarrow \mathcal{T} f_{t-1}$
- FQI keeps things tractable by: $f_t \leftarrow \prod_F \mathcal{T} f_{t-1}$
  - We know $\mathcal{T}$ is γ-contraction.
  - If $\prod_F$ is non-expansion, we are good — not always true though
    (You should have seen an example of non-expansion $\prod_F$. Which one?)
  - Still an iterated algorithm for fixed point eq. => optimization objective changes as current f changes!
- Alternative: minimize $\| f - \mathcal{T} f \|$ over $f \in F$
  - Is it equivalent to minimizing: $\mathbb{E}_{\substack{(s,a)\sim\mu \\ r\sim R(s,a) \\ s'\sim P(s,a)}} \left[\left(f(s,a) - \left(r + \gamma \max_{a'} f(s',a')\right)\right)^2 \right]$? (omitted in the rest of slides)

Bellman error minimization

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

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

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

Unbiased estimate "double sampling"

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/ resetting functionality…

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:

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

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: Use G=F; has poly sample complexity guarantee under this assumption

Bellman error minimization (residual) (BRM)

$\mathcal{T} f \in F$ for all $f \in F$

"closedness" "completeness"

"small inherent Bellman error".

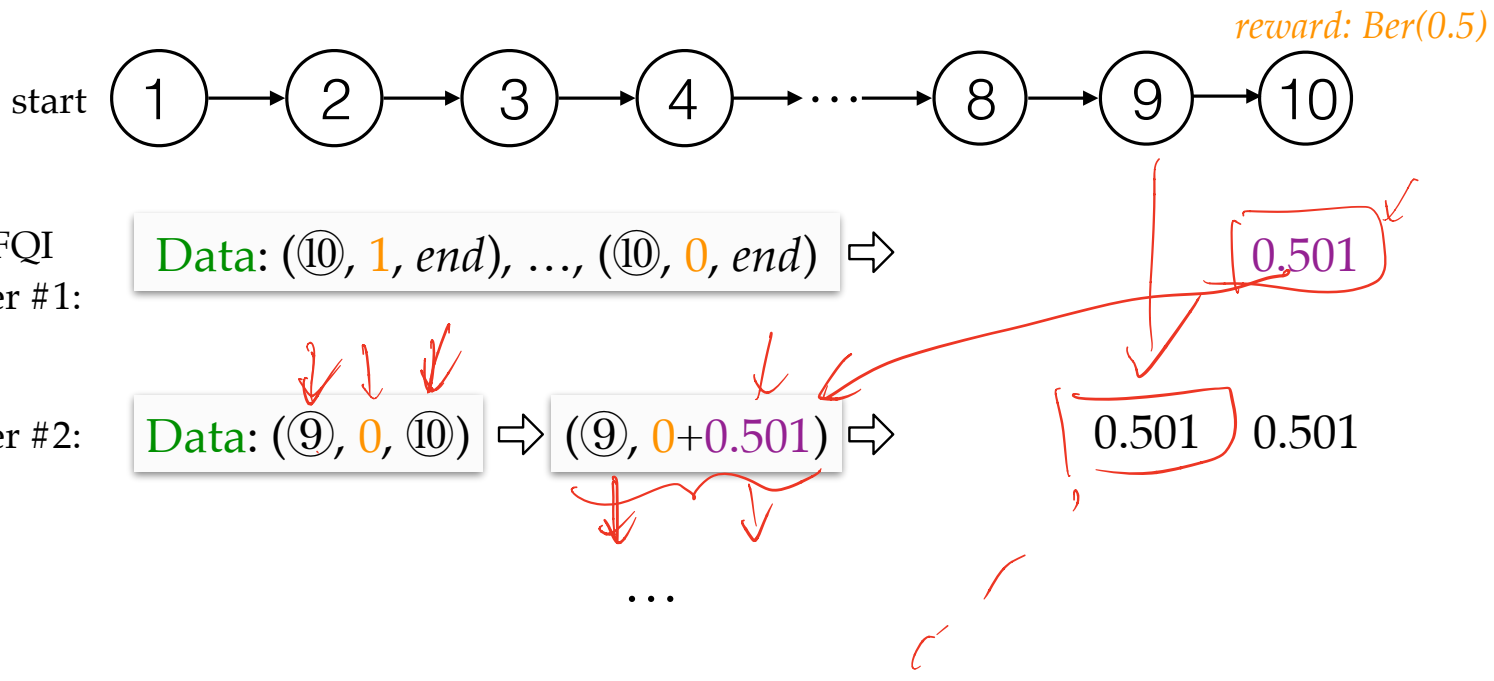
- A kind of "completeness": F is closed under $\mathcal{T}$
- For finite class F, this implies realizability. Why?
- What does it mean for abstractions?
  - For piece-wise constant function spaces, completeness = bisimulation (homework)

$Q^* \in F$.

if $f \in F \Rightarrow \mathcal{T} f \in F \Rightarrow \mathcal{T}^2 f \in F \Rightarrow \cdots \Rightarrow \mathcal{T}^\infty f \in F$

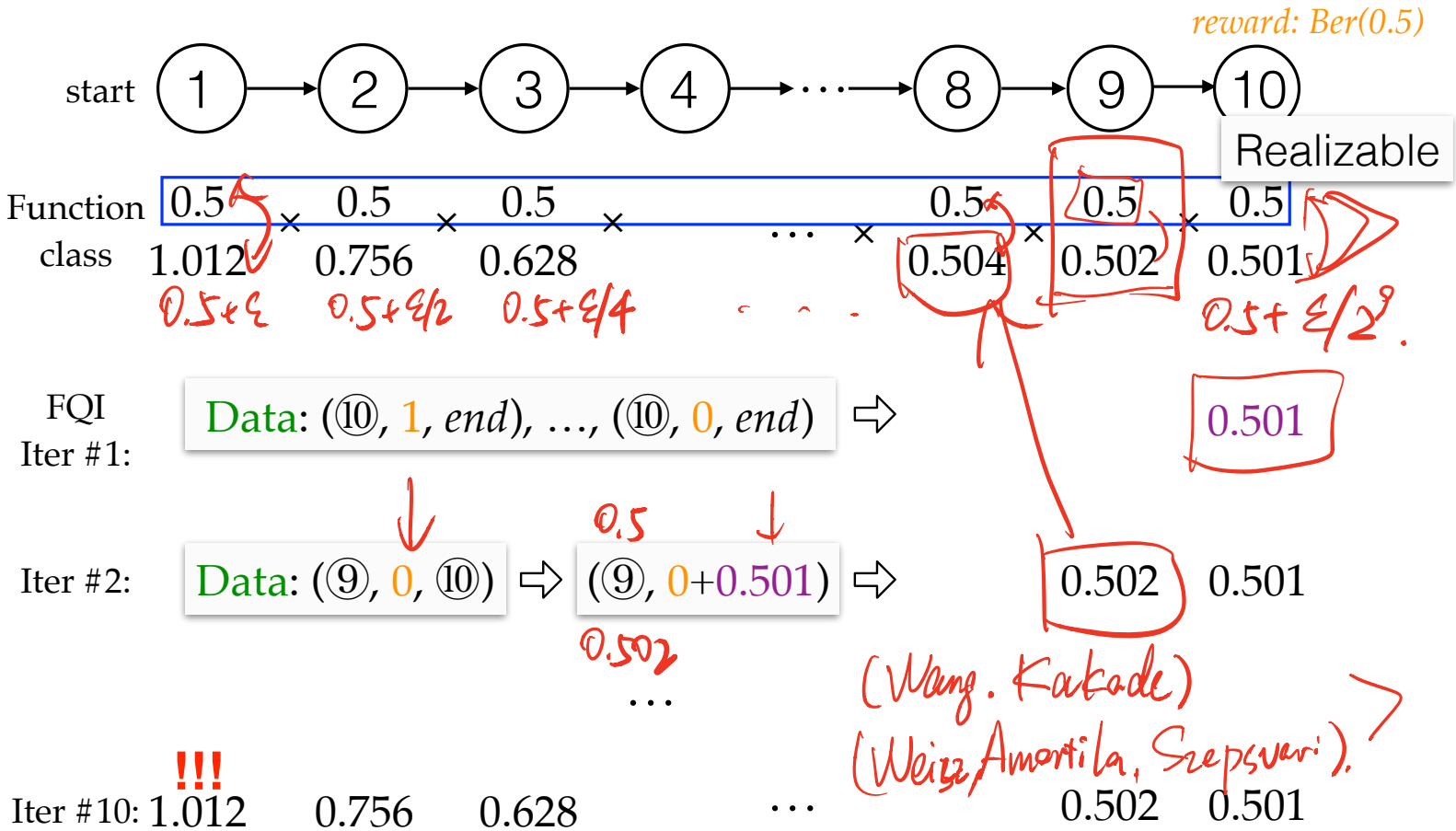
$\| \; Q^*$.

A simple example (finite horizon, γ=1)



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

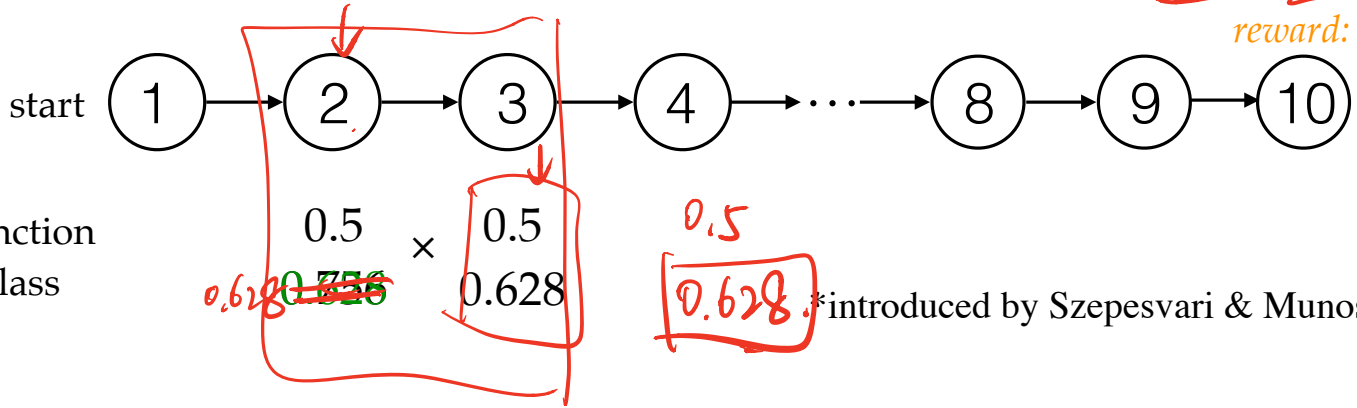
How things go wrong (w/ restricted class)



Example given in Dann et al'18

- similar conclusion known for decades (e.g., Van Roy'94)

How completeness fixes the issue



*introduced by Szepesvari & Munos [2005]

- More generally: issue goes away if the regression problem

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

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

- 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
- With this assumption, *poly*(|A|, log|F|, H, $1/\varepsilon$, $1/\delta$, C) achievable

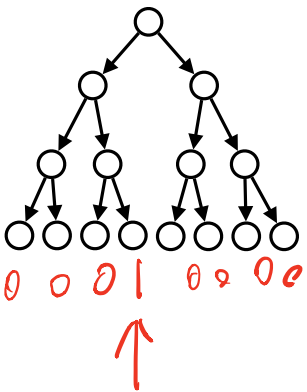
One last assumption: data

- 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??
- Bad news: in the worst case, **no distribution would work**
- Formally:

> There exists a family of large MDPs, F that is always realizable, G that includes T_f for any $f \in F$, s.t.: for any data distribution and any batch RL algorithm with F, G as input, the worst-case sample complexity *cannot* be *poly*($|A|$, H, $\log|F|$, $\log|G|$).

Proof of lower bound

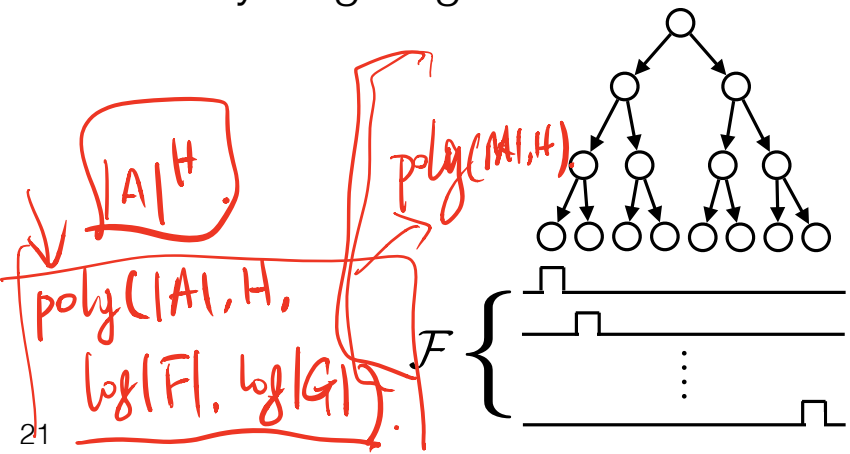
- Idea: we are allowed unbounded # of states — use a depth-H complete tree to essentially emulate MAB w/ $|A|^H$ arms
- Recall that sample complexity lower bound for MAB is #arms/ε^2
- Without function approximation: exponential sample complexity for exploration algorithms
- Remain to show: (1) function approx. does not help. (2) transfer the claim to batch algorithms



Proof of lower bound

Show: func. approx. does not help:

- Let F be the collection of Q^* from all MDPs in family
- $\log|F| = H \log|A|$, always realizable
- Lef G be the collection of applying $\mathcal{T}$ in every MDP to the Q^* in every other MDP
- $\log|G| = \log|F|^2 = 2\log|F|$
- In lower bound proof, alg is allowed to specialize to the problem family — giving F and G does not help

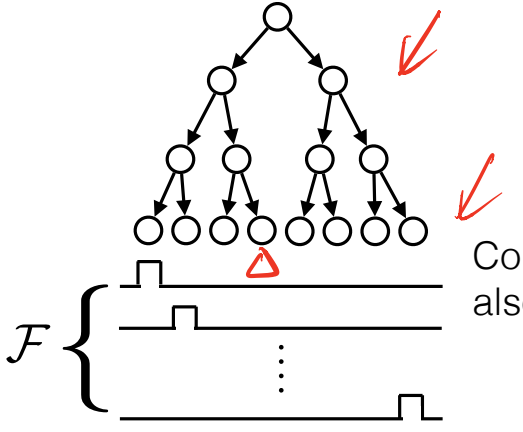


Construction from [Krishnamurthy et al'16]; also used in [Jiang et al'17]

Proof of lower bound

Exploration to batch

- Transition is deterministic (only leaf reward is stochastic)
- An exploration algorithm can obtain any data distribution as it likes
- Any fixed data distribution + batch algorithm = special case of exploration algorithm!
- Lower bound for a set of algorithms automatically applies to its subset — done.

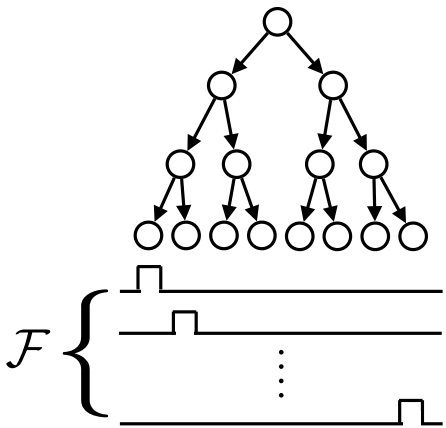


Construction from [Krishnamurthy et al'16]; also used in [Jiang et al'17]

Proof of lower bound

Summarize:

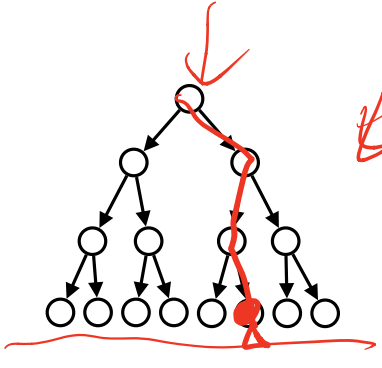
- Sample complexity is $\Omega(|A|^H/\varepsilon^2)$ (because F and G are useless)
- $|A|$, H, $\log|F|$ and $\log|G|$ are all poly in $|A|$ and H
- $poly(|A|, H, \log|F|, \log|G|)$ is impossible



Construction from [Krishnamurthy et al'16]; also used in [Jiang et al'17]

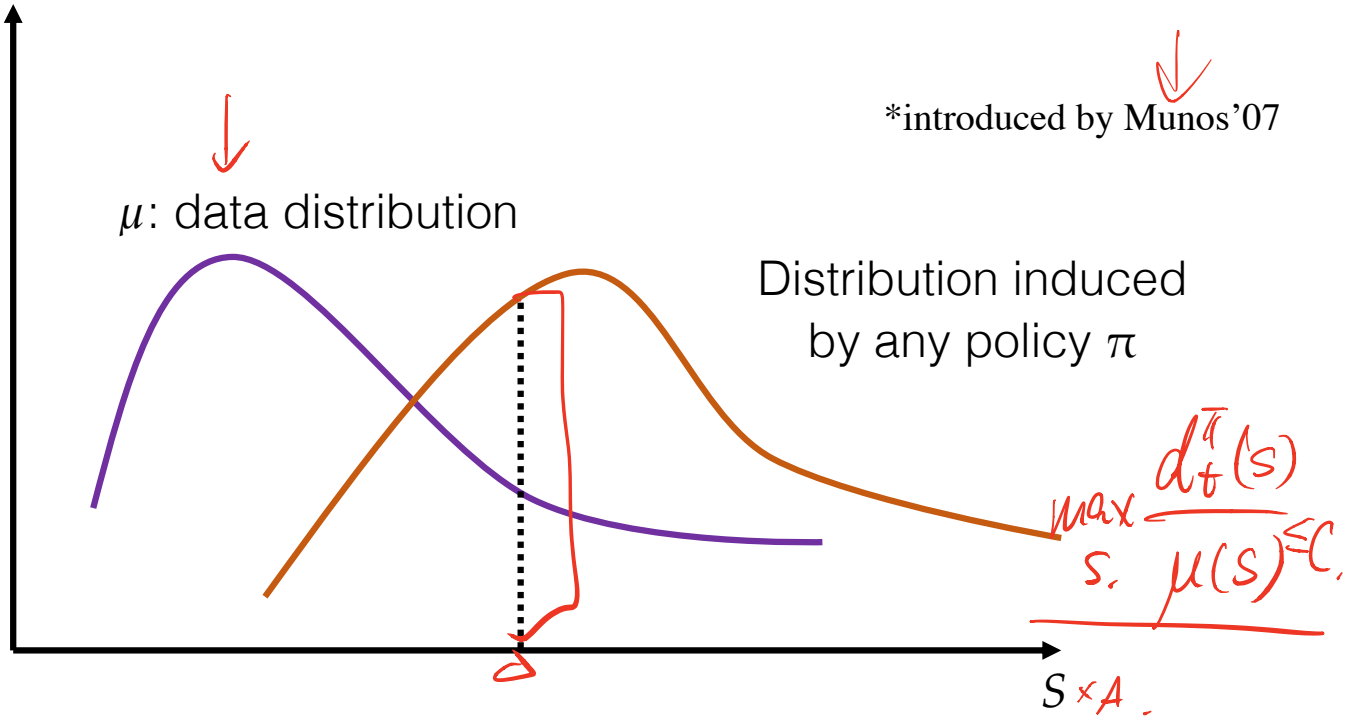
What's causing this: Distribution shift

- All paths are symmetric — training data should be uniform (ideally)
- We incur error "along" the the computed policy, which is 1 path
- States that matter in "test phase" are only given exponentially small weights in training
- Need assumption on the environment: distribution shift is not severe



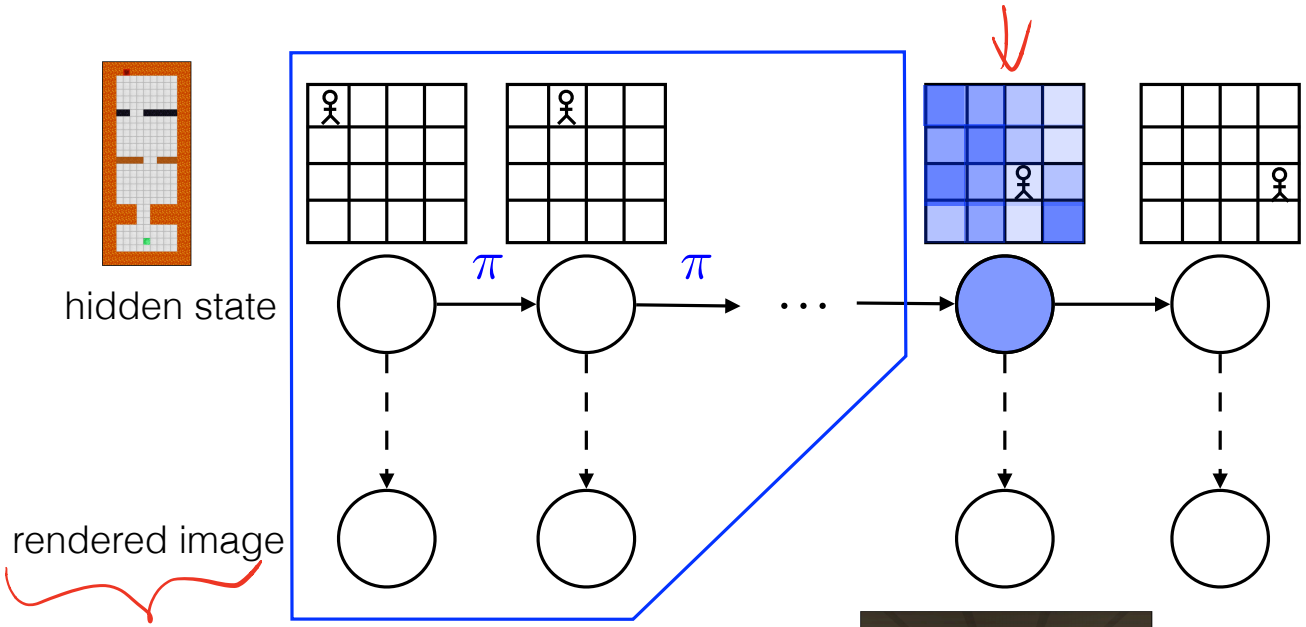
$C = A^H$

Assumption on data: “Concentrability”



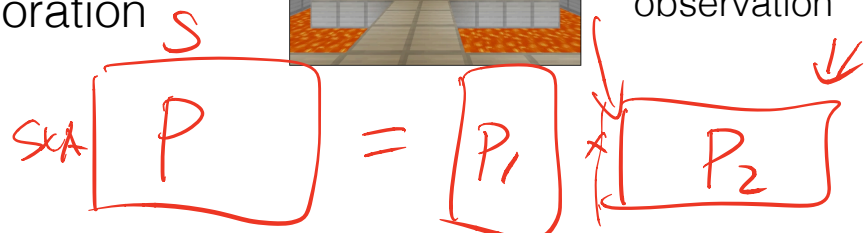
- Let C be a uniform upper bound on the density ratio
- Assumption: C is small (= allow polynomial dependence on C)
- Previous exponential lower bound is “explained away” by an exponentially large C

Concentrability: when is it small?



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

Markovian high-dimensional observation



Remainder of this part

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

$\exists f \in F, \forall f \in F.$

- $\mathcal{T} f \in F$ for all $f \in F$ (implies $Q^* \in F$ for finite class F)
- Dataset: $D = \{(s, a, r, s')\}$, where $s \sim \mu$, $a \sim \mathrm{U}$ (Uniform)
- Let C be a constant s.t. for any (non-stationary) π and time step h, $\max_{s \in S} d_h^{\pi}(s)/\mu(s) \le C$, where d_h^{π} is the h-step distribution of π
- Algorithm: $f_{k+1} \leftarrow \arg\min_{f \in \mathcal{F}} \mathcal{L}_D(f; f_k)$, where $\mathcal{L}_D(f; f') := \frac{1}{|D|} \sum_{(s,a,r,s') \in D} \left(f(s,a) - r - \gamma V_{f'}(s')\right)^2$. Also let: $\mathcal{L}_{\mu \times U}(f; f') := \mathbb{E}_D[\mathcal{L}_D(f; f')]$
- For any function g and distribution ν, define $\|g\|_{p,\nu} := (\mathbb{E}_{s \sim \nu}[|g(s)|^p])^{1/p}$
  - Motonone in p ("dim normalized"). $p = 2$ if omitted.
  - If $\max_{s \in S} \nu(s)/\mu(s) \le C$, $\|\cdot\|_\nu \le \sqrt{C}\|\cdot\|_\mu$
- Uniform deviation bound: $\forall f, f' \in \mathcal{F}, |\mathcal{L}_D(f; f') - \mathcal{L}_{\mu \times U}(f; f')| \le \epsilon$.
  - Can be obtained by Hoeffding's
  - Improved analysis with fast rates (see note)

Lemma 1. *Define* $\pi_{f,f_k}(s) := \arg\max_{a \in \mathcal{A}} \max\{f(s,a), f_k(s,a)\}$. *Then we have* $\forall \nu \in \Delta(\mathcal{S})$,

$$\|V_f - V_{f_k}\|_\nu \le \|f - f_k\|_{\nu \times \pi_{f,f_k}}.$$

- $D = \{(s,a,r,s')\}$. w/ $s \sim \mu$, $a \sim \mathrm{Unif}(A)$, $r \sim R(s,a)$, $s' \sim P(\cdot|s,a)$.
- $\mathcal{L}_D(f; f') := \frac{1}{|D|} \sum_{(s,a,r,s') \in D} \left(f(s,a) - r - \gamma \max_{a'} f(s',a')\right)^2$.

→ $\mathcal{L}_{\mu \times U}(f; f') := \mathbb{E}_D[\mathcal{L}_D(f; f')]$.

$\forall f, f' \in \mathcal{F}$. Hoeffding + union bound.
$|\mathcal{L}_D(f; f') - \mathcal{L}_{\mu \times U}(f; f')| \le \varepsilon$

- Assume $\exists f \in \mathcal{F}$ $\forall f \in \mathcal{F}$.
- Assume $\exists C < +\infty$, s.t. $\forall \pi, h$, $\max_s \frac{d_h^{\pi}(s)}{\mu(s)} \le C$
- $\forall g : S \to \mathbb{R}$, $\nu \in \Delta(S)$, $p \ge 1$, Let $\|g\|_{p,\nu} := \left(\mathbb{E}_{s \sim \nu}[|g(s)|^p]\right)^{1/p}$.

→ for ν s.t. $\max_s \frac{\nu(s)}{\mu(s)} \le C$. $\|g\|_{2,\nu} \le \sqrt{C} \cdot \|g\|_{2,\mu}$.

- FQI alg: $f_{k+1} \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{L}_D(f; f_k)$. (or $f_{k+1} \leftarrow \hat{\mathcal{T}}_{\mathcal{F}} f_k$).

Let $\hat{\pi} = \pi_{f_k}$.

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

$J(\pi^*) - J(\hat{\pi})$ P-D Lemma.

$$= \sum_{h=1}^{\infty} \gamma^{h-1} \mathbb{E}_{s \sim d_h^{\hat{\pi}}}[V^*(s) - Q^*(s, \hat{\pi})]$$

$$\le \sum_{h=1}^{\infty} \gamma^{h-1} \mathbb{E}_{s \sim d_h^{\hat{\pi}}}[Q^*(s,\pi^*) - f_k(s,\pi^*) + f_k(s,\hat{\pi}) - Q^*(s,\hat{\pi})]$$

$$\le \sum_{h=1}^{\infty} \gamma^{h-1} \left(\mathbb{E}_{s \sim d_h^{\hat{\pi}}}[|Q^*(s,\pi^*) - f_k(s,\pi^*)|] + \mathbb{E}[\dots]\right)$$

$$= \sum_{h=1}^{\infty} \gamma^{h-1} \left(\|Q^* - f_k\|_{1, d_h^{\hat{\pi}} \times \pi^*} + \|Q^* - f_k\|_{1, d_h^{\hat{\pi}} \times \hat{\pi}}\right)$$

$s \sim d_h^{\hat{\pi}}$, $a \sim \pi^*(\cdot|s)$.

$\|g\|_{2,\nu}^2 = \mathbb{E}_{s \sim \nu}[g(s)^2]$.

$= \sum_{s \in S} \nu(s) \cdot g(s)^2$.

$\le \sum_{s \in S} C \cdot \mu(s) \cdot g(s)^2$

$= C \cdot \sum_{s \in S} \mu(s) \cdot g(s)^2$.

$= C \cdot \|g\|_{2,\mu}^2$.

P-D Lemma:

$J(\pi_1) - J(\pi_2) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi_1}}[A^{\pi_2}(s, \pi_1)]$.

$\pi_1 = \hat{\pi}$, $\pi_2 = \pi^*$

$$\le \sum_{n=1}^{\infty} \gamma^{n-1} \left(\|Q^* - f_k\|_{2, d_n^{\hat{\pi}} \times \pi^*} + \|Q^* - f_k\|_{2, d_n^{\hat{\pi}} \times \hat{\pi}} \right).$$

$\gamma^{n-1} \to \frac{1}{1-\gamma}$.

$\forall p \le q.\ \forall \nu.$

$\|\cdot\|_{p,\nu} \le \|\cdot\|_{q,\nu}.$

$\|\cdot\|_1 \ge \|\cdot\|_2 \ge \|\cdot\|_\infty$

$\sum$ $\quad$ max

Remark: this is refined ver. of Singh & Yee '94.

$$J(\pi^*) - J(\pi_f) \le \|V^* - V^{\pi_f}\|_\infty \le \frac{\boxed{2}\|f - Q^*\|_\infty}{1-\gamma}.$$

It remains to show: $\|f_k - Q^*\|_{2, \nu \times \pi}$ is small

$\nu \times \pi \to d_n^{\hat{\pi}} \times \pi^*$, or $d_n^{\hat{\pi}} \times \hat{\pi}$.

$\forall \nu, \pi$, let $P(\nu \times \pi) \in \Delta(S)$ defined as:

$s' \sim P(\nu \times \pi) \iff s \sim \nu,\ a \sim \pi(\cdot|s),\ s' \sim P(\cdot|s,a).$

$$\to \boxed{\|f_k - Q^*\|_{2, \nu \times \pi}} = \|f_k - \mathcal{T} f_{k-1} + \mathcal{T} f_{k-1} - Q^*\|_{2, \nu \times \pi}.$$

$$\le \underbrace{\|f_k - \mathcal{T} f_{k-1}\|_{2, \nu \times \pi}}_{\text{the k-th iter of FQI.}} + \underbrace{\|\mathcal{T} f_{k-1} - \mathcal{T} Q^*\|_{2, \nu \times \pi}}_{\text{"contraction".}}.$$

"contraction": $\|\mathcal{T} f_{k-1} - \mathcal{T} Q^*\|_{2, \nu \times \pi}^{\boxed{2}}$.

$$= \mathbb{E}_{s \sim \nu, a \sim \pi}\left[\gamma^2 \left(\mathbb{E}_{s' \sim P(s,a)}\left[V_{f_{k-1}}(s') - V_{Q^*}(s') \right] \right)^2 \right].$$

$(\mathcal{T} f)(s,a) = R(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)}[V_f(s')]$.

$V_f(s') = \max_{a'} f(s', a')$.

Jensen's

$$\le \gamma^2 \mathbb{E}_{s \sim \nu, a \sim \pi}\left[\mathbb{E}_{s' \sim P(s,a)}\left[\left(V_{f_{k-1}}(s') - V^*(s') \right)^2 \right] \right].$$

only func of s'.

$$= \gamma^2 \|V_{f_{k-1}} - V^*\|_{2, P(\nu \times \pi)}^2.$$

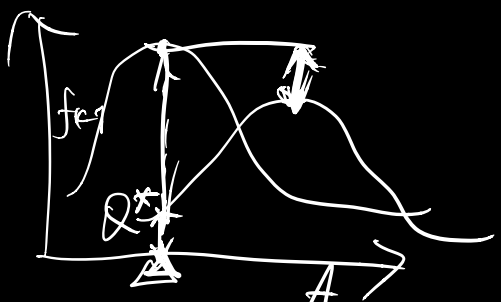
$$= \gamma^2 \sum_{s \in S} P(\nu \times \pi)(s) \cdot \left(\max_a f_{k-1}(s,a) - \max_a Q^*(s,a) \right)^2.$$

$$\le \gamma^2 \|f_{k-1} - Q^*\|_{P(\nu \times \pi) \times \pi_{f_{k-1}, Q^*}}^2. \quad \boxed{\swarrow}$$

"contraction-like".

To recap: $J(\pi^*) - J(\hat{\pi}) \le \cdots \le \|f_K - Q^*\|_{2, \nu \times \pi}$

where ν is $d_h^{\hat{\pi}}$.



$$\begin{cases} \|f_K - Q^*\|_{2,\nu\times\pi} \le \|f_K - \mathcal{T} f_{K-1}\|_{2,\nu\times\pi} \\ \quad + \gamma \cdot \|f_{K-1} - Q^*\|_{2, P(\nu\times\pi)\times \pi_{f_{K-1}, Q^*}} \end{cases}$$

$a^* = \operatorname{argmax}_a \left(\max f_{k-1}(s,a), \max Q^*(s,a) \right)$

$\pi_{f_{k-1}, Q^*} : s \mapsto a^*$.

Note: this holds for all ν that takes form of d_h^{π} for (non-stationary π).

Claim: if ν is such a dist. then $P(\nu\times\pi)\times \pi_{f_{k-1}, Q^*}$ is also such a dist. ← "admissible"

$$\max_{\nu' : \text{admissible}} \|f_K - Q^*\|_{2,\nu'} \le \max_{\nu' : \text{ad}} \|f_K - \mathcal{T} f_{K-1}\|_{2,\nu'} + \gamma \max_{\nu' : \text{ad}} \|f_{K-1} - Q^*\|_{2, P(\nu')\times \pi_{f_{K-1}, Q^*}}$$

$$\le \max_{\nu' : \text{ad}} \|f_{K-1} - Q^*\|_{2,\nu'}$$

Remains to bound $\max_{\nu' : \text{ad}} \|f_K - \mathcal{T} f_{K-1}\|_{2,\nu'}$.

Recall: $\forall$ admissible ν, $\frac{\nu(s)}{\mu(s)} \le C$. $\frac{\nu'(s\times a)}{(\mu\times U)(s,a)} \le C \cdot |A|$.

$$\forall \nu', \quad \|f_K - \mathcal{T} f_{K-1}\|_{2,\nu'} \le \sqrt{C \cdot |A|} \cdot \|f_K - \mathcal{T} f_{K-1}\|_{2,\mu\times U}$$

Data distribution

$$\|f_K - \mathcal{T} f_{K-1}\|^2_{2,\mu\times U} = \mathcal{L}_{\mu\times U}(f_K; f_{K-1}) - \mathcal{L}_{\mu\times U}(\mathcal{T} f_{K-1}; f_{K-1})$$

crucial. "inherent noise"

$\forall f, f' \in \mathcal{F}$. $|\mathcal{L}_D(f; f') - \mathcal{L}_{\mu\times U}(f; f')| \le \varepsilon$.

$$\le \mathcal{L}_D(f_K; f_{K-1}) - \mathcal{L}_D(\mathcal{T} f_{K-1}; f_{K-1}) + 2\varepsilon$$

(relies on $\mathcal{T} f_{K-1} \in \mathcal{F}$).

$\leq \boxed{2\varepsilon}$ (b/c: f_k minimizing $L_D(\cdot\,; f_{k-1})$ + $\exists f_{k-1} \in \mathcal{F}$).

$$J(\pi^*) - J(\hat{\pi}) \leq \frac{2}{1-\gamma}\left(\frac{1-\gamma^k}{1-\gamma}\sqrt{2|A|C\varepsilon} + \gamma^k \frac{R_{max}}{1-\gamma}\right).$$

$k \to \infty$. $\quad \frac{2}{1-\gamma^2}\sqrt{2|A|C\varepsilon}$.