Generalization in RL

Sham M. Kakade (and Wen Sun)

Outline

0. Announcements

1. Recap

2. Today: SL vs. RL

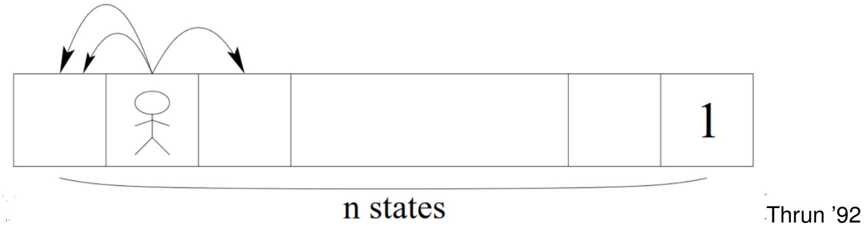
3. Supervised Learning (SL) : Let's review

4. RL and generalization
   - Is Agnostic Learning Possible?
   - Lower bounds

5. Interpretation: How should we study RL ?

(It's nice to see you all!)

The need for strategic exploration



- agent starts at s_0
- length of chain is H
- chance of hitting goal state in H steps is $(1/3)^H$ with a random policy

UCBVI: Optimistic Model-based Learning

strategic exploration

Inside iteration n :

Use all previous data to estimate transitions $\widehat{P}^n_1, \ldots, \widehat{P}^n_{H-1}$

Design reward bonus $b^n_h(s,a), \forall s, a, h$

Optimistic planning with learned model: $\pi^n = \text{Value-Iter}\left(\{\widehat{P}^n_h, r_h + b^n_h\}_{h=1}^{H-1}\right)$

Collect a new trajectory by executing π^n in the real world $\{P_h\}_{h=0}^{H-1}$ starting from s_0

UCBVI: Put All Together

For $n = 1 \to N$:

1. Set $N_h^n(s, a) = \sum_{i=1}^{n-1} \mathbf{1}\{(s_h^i, a_h^i) = (s, a)\}, \forall s, a, h$
2. Set $N_h^n(s, a, s') = \sum_{i=1}^{n-1} \mathbf{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s, a, s')\}, \forall s, a, a', h$
3. Estimate $\widehat{P}^n$: $\widehat{P}_h^n(s' | s, a) = \frac{N_h^n(s, a, s')}{N_h^n(s, a)}, \forall s, a, s', h$
4. Plan: $\pi^n = VI\left(\{\widehat{P}_h^n, r_h + b_h^n\}_h\right)$, with $b_h^n(s, a) = cH\sqrt{\frac{\ln(SAHN/\delta)}{N_h^n(s, a)}}$
5. Execute π^n : $\{s_0^n, a_0^n, r_0^n, \ldots, s_{H-1}^n, a_{H-1}^n, r_{H-1}^n, s_H^n\}$

need $N \gtrsim O(S^2 H) / \epsilon^2$

Theorem: UCBVI Regret Bound

$$\mathbb{E}\left[\text{Regret}_N\right] := \mathbb{E}\left[\sum_{n=1}^{N}\left(V^\star - V^{\pi^n}\right)\right] \leq \widetilde{O}\left(H^2\sqrt{S^2AN}\right)$$

N

Remarks:

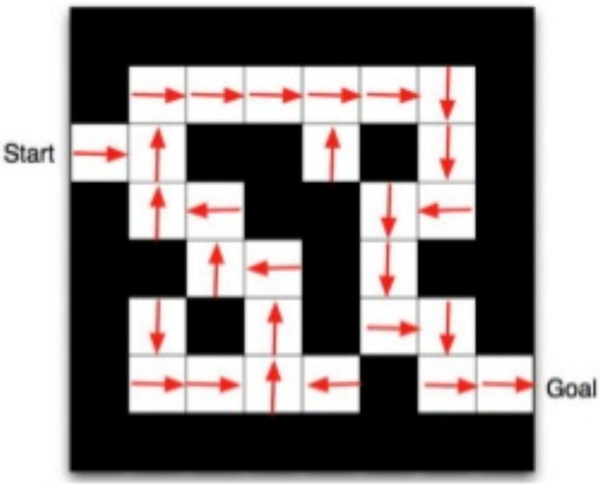
Note that we consider expected regret here (policy π^n is a random quantity).
High probability version is not hard to get (need to do a martingale argument)

Dependency on H and S are suboptimal; but the **same** algorithm can achieve $H^2\sqrt{SAN}$ in the leading term [Azar et.al 17 ICML]

Outline

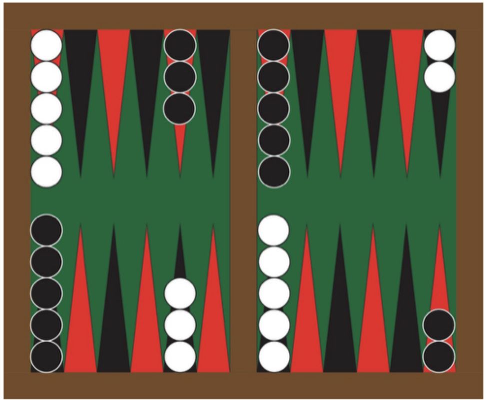
What we can solve:
(the small state space case)

Maze example: r = -1 per time-step and policy



[David Silver. Advanced Topics: RL]

What we want to solve:
(the large state space case)



TD GAMMON [Tesauro 95]

[AlphaZero, Silver et.al, 17]

[OpenAI Five, 18]

Generalization: RL vs Supervised Learning (SL)

To what extent is generalization in RL similar to (or different from) that in supervised learning?

Generalization: RL vs Supervised Learning (SL)

To what extent is generalization in RL similar to (or different from) that in supervised learning?

- Up to now, we have focussed on “tabular” MDPs (theoretically important)
  - We ultimately seek learnability results where number of states is large (or $|\mathcal{S}| = \infty$).
  - This is a question of generalization.

Generalization: RL vs Supervised Learning (SL)

To what extent is generalization in RL similar to (or different from) that in supervised learning?

- Up to now, we have focussed on "tabular" MDPs (theoretically important)
  - We ultimately seek learnability results where number of states is large (or $|\mathcal{S}| = \infty$).
  - This is a question of generalization.
- Supervised Learning: two lines of thinking
  - Optimal learning: try to learn the Bayes optimal classifier. need very strong assumptions. "No free lunch" thms
  - Agnostic learning: try to do as well best classifier in some (restricted) class $\mathcal{H}$.

Generalization: RL vs Supervised Learning (SL)

To what extent is generalization in RL similar to (or different from) that in supervised learning?

- Up to now, we have focussed on "tabular" MDPs (theoretically important)
  - We ultimately seek learnability results where number of states is large (or $|\mathcal{S}| = \infty$).
  - This is a question of generalization.
- Supervised Learning: two lines of thinking
  - Optimal learning: try to learn the Bayes optimal classifier. need very strong assumptions.
  - Agnostic learning: try to do as well best classifier in some (restricted) class $\mathcal{H}$.
- If rather than trying to be 'optimal' in RL, does trying to do agnostic learning make our task easier?

Outline

Binary Classification

- n labeled examples: $(x_i, y_i)_{i=1}^{n}$, with $x_i \in \mathcal{X}$ and $y_i \in \{0, 1\}$.
  A set $\mathcal{H}$ of binary classifiers, where for $h \in \mathcal{H}$, $h : \mathcal{X} \to \{0, 1\}$.
  Define the empirical error and the true error as:

$$\widehat{\text{err}}(h) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(h(x_i) \neq y_i), \quad \text{err}(h) = \mathbb{E}_{(X,Y)\sim D} \mathbf{1}(h(X) \neq Y).$$

  where $\mathbf{1}(h(x) \neq y)$ is 0 if $h(x) = y$ and 1 otherwise.

Binary Classification

- n labeled examples: $(x_i, y_i)_{i=1}^n$, with $x_i \in \mathcal{X}$ and $y_i \in \{0, 1\}$.
  A set $\mathcal{H}$ of binary classifiers, where for $h \in \mathcal{H}$, $h : \mathcal{X} \to \{0, 1\}$.
  Define the empirical error and the true error as:

$$\widehat{\text{err}}(h) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(h(x_i) \neq y_i), \quad \text{err}(h) = \mathbb{E}_{(X,Y)\sim D} \mathbf{1}(h(X) \neq Y).$$

  where $\mathbf{1}(h(x) \neq y)$ is 0 if $h(x) = y$ and 1 otherwise.
- If the samples are drawn i.i.d. according to a joint distribution D over (x, y), then, by Hoeffding's inequality, for a fixed $h \in \mathcal{H}$, with probability at least $1 - \delta$:

$$|\text{err}(h) - \widehat{\text{err}}(h)| \leq \sqrt{\frac{1}{2N} \log \frac{2}{\delta}}.$$

SL is an RL problem with $\gamma = 0$

- Binary classification is special case of RL.
  Consider learning in an MDP, with two actions where the effective horizon is 1.
- $|\mathcal{A}| = 2$, $\gamma = 0$, and the reward function is $r(s, a) = \mathbf{1}(\text{label}(s) = a)$.
- Note in SL, we rarely make restrictions that $\mathcal{X}$ (i.e. $\mathcal{S}$) is finite.
- Note that $\mu(s_0) \leftrightarrow D(x)$ (D is the distribution of our data)

$s \longleftrightarrow x$ $\qquad$ $\mathcal{S} \longleftrightarrow \mathcal{X}$

Occams Razor and Generalization

Data - Reuse

Your HW0: This and the union bound give rise to what is often referred to as the "Occam's razor" bound:

Proposition

(The "Occam's razor" bound) Suppose $\mathcal{H}$ *is finite. Let* $\widehat{h} = \arg\min_{h\in\mathcal{H}} \widehat{err}(h)$ *and* $h^\star = \arg\min_{h\in\mathcal{H}} err(h)$*. With probability at least* $1-\delta$*:*

$$err(\widehat{h}) - err(h^\star) \le \sqrt{\frac{2}{N}\log\frac{2|\mathcal{H}|}{\delta}}.$$

Occams Razor and Generalization

Your HW0: This and the union bound give rise to what is often referred to as the "Occam's razor" bound:

Proposition

(The "Occam's razor" bound) Suppose $\mathcal{H}$ is finite. Let $\widehat{h} = \arg\min_{h\in\mathcal{H}} \widehat{err}(h)$ and $h^\star = \arg\min_{h\in\mathcal{H}} err(h)$. With probability at least $1-\delta$:

$$err(\widehat{h}) - err(h^\star) \le \sqrt{\frac{2}{N}\log\frac{2|\mathcal{H}|}{\delta}}.$$

(The logarithmic dependence is the most naive complexity measure of $\mathcal{H}$, yet the bound is strong.)

The VC Dimension

- $|\mathcal{H}|$: a set of Boolean functions on $\mathcal{X}$.
  Even though $|\mathcal{H}|$ may be infinite, the number of possible behaviors of on a finite set of states is not necessarily exhaustive.

The VC Dimension

- $|\mathcal{H}|$: a set of Boolean functions on $\mathcal{X}$.
  Even though $|\mathcal{H}|$ may be infinite, the number of possible behaviors of on a finite set of states is not necessarily exhaustive.
- We say that the set $\{x_1, x_2, \ldots x_d\}$ is shattered if there exists an $h \in \mathcal{H}$ that can realize any of the possible 2^d labellings.

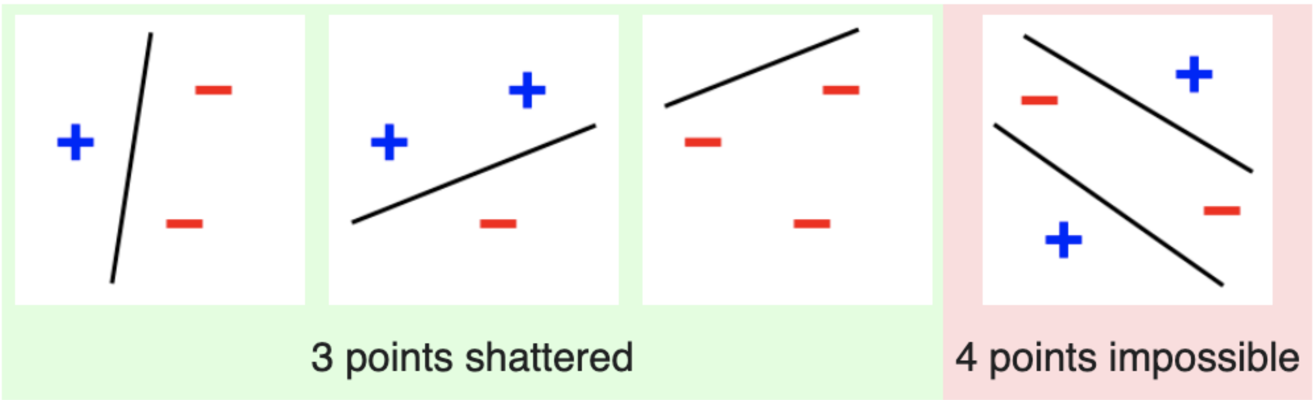
The VC Dimension

- $|\mathcal{H}|$: a set of Boolean functions on $\mathcal{X}$.
  Even though $|\mathcal{H}|$ may be infinite, the number of possible behaviors of on a finite set of states is not necessarily exhaustive.
- We say that the set $\{x_1, x_2, \ldots x_d\}$ is shattered if there exists an $h \in \mathcal{H}$ that can realize any of the possible 2^d labellings.
- The Vapnik–Chervonenkis (VC) dimension is the size of the largest shattered set.

The VC Dimension

- $|\mathcal{H}|$: a set of Boolean functions on $\mathcal{X}$.
  Even though $|\mathcal{H}|$ may be infinite, the number of possible behaviors of on a finite set of states is not necessarily exhaustive.
- We say that the set $\{x_1, x_2, \ldots x_d\}$ is shattered if there exists an $h \in \mathcal{H}$ that can realize any of the possible 2^d labellings.
- The Vapnik–Chervonenkis (VC) dimension is the size of the largest shattered set.
- Let $d = VC(\mathcal{H})$. The Sauer–Shelah lemma: the number of possible labellings on a set of n points by functions in $\mathcal{H}$ is at most $\left(\frac{en}{d}\right)^d$. For $d << m$, this is much less than 2^n.

Review: Half Spaces



- Let $\mathcal{H}_{\text{half}}$ be the set of halspaces on $\mathcal{X} = \mathbb{R}^d$.
- The VC dimension is $VC(\mathcal{H}_{\text{half}}) = d + 1$

Infinite Hypothesis Classes

The following classical bound highlights how generalization is possible on infinite hypothesis classes with bounded complexity.

Proposition

(VC dimension and generalization) Let $\widehat{h} = \arg\min_{h\in\mathcal{H}} \sum_{i=1}^{n} \mathbf{1}(h(x_i) \neq y_i)$ *and* $h^\star = \arg\min_{h\in\mathcal{H}} err(h)$*. Suppose* $\mathcal{H}$ *has a bounded VC dimension. For* $m \geq VC(\mathcal{H})$*, we have that with probability at least* $1-\delta$*:*

$$err(\widehat{h}) - err(h^\star) \leq \sqrt{\frac{c}{n}\left(VC(\mathcal{H}) \log \frac{2n}{VC(\mathcal{H})} + \log \frac{2}{\delta}\right)},$$

regret

where c is an absolute constant

Outline

1. Recap
2. Today: SL vs. RL
3. Supervised Learning (SL) : Let's review
4. RL and generalization
   - Is Agnostic Learning Possible?
   - Lower bounds
5. Interpretation: How should we study RL

Can we entirely avoid the dependence on S?

RL and Agnostic Learning

- We have a set of policies Π (either finite or infinite).
  - Π could be a parametric set. (neural policies)
  - Π could be greedy policys on a a set of parametric value functions $\mathcal{V} = \{f_\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R} | \theta \in \mathbb{R}^d\}$.
  - Π may not contain $\pi^\star$.

either $|\Pi| < |A|^S$

or complexity(Π) is restricted in some way

RL and Agnostic Learning

- We have a set of policies Π (either finite or infinite).
  - Π could be a parametric set.
  - Π could be greedy policys on a a set of parametric value functions $\mathcal{V} = \{f_\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R} | \theta \in \mathbb{R}^d\}$.
  - Π may not contain $\pi^\star$.
- in agnostic learning, we have the optimization problem:

$$\max_{\pi \in \Pi} \mathbb{E}_{s_0 \sim \mu} V^\pi(s_0)$$

We want to (approx) solve this with a small number of sample trajectories.

RL and Agnostic Learning

- We have a set of policies Π (either finite or infinite).
  - Π could be a parametric set.
  - Π could be greedy policys on a a set of parametric value functions $\mathcal{V} = \{f_\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R} | \theta \in \mathbb{R}^d\}$.
  - Π may not contain $\pi^\star$.
- in agnostic learning, we have the optimization problem:

$$\max_{\pi \in \Pi} \mathbb{E}_{s_0 \sim \mu} V^\pi(s_0)$$

  We want to (approx) solve this with a small number of sample trajectories.
- analogous to agnostic learning in SL
  - binary classification: $|\mathcal{A}| = 2$, $\gamma = 0$, $r(\cdot)$ being the labeling reward.
  - relevant dependencies for RL:

$$\text{Complexity}(\Pi), |\mathcal{S}|, |\mathcal{A}|, N \quad , \quad H \approx \frac{1}{1-\gamma}$$

RL Sampling Model

- Assume sampling access to the MDP in a μ-reset model:
  - start at a state $s_0 \sim \mu$
  - we can rollout a policy π of our choosing
  - we can terminate the trajectory at will.

  (weaker model than generative model)

RL Sampling Model

- Assume sampling access to the MDP in a μ-reset model:
  - start at a state $s_0 \sim \mu$
  - we can rollout a policy π of our choosing
  - we can terminate the trajectory at will.

  (weaker model than generative model)

Lemma

(Effective Horizon and Truncation) We have that:

$$\left| V^\pi(s_0) - \mathbb{E}_\pi \left[\sum_{t=0}^{H} \gamma^t r(s_t, a_t) \mid s_0 \right] \right| \leq \gamma^H / (1 - \gamma),$$

For $H = \frac{\log\left(1/\left(\epsilon(1-\gamma)\right)\right)}{1-\gamma}$ *we will have an* ϵ *approximation to* $V^\pi(s_0)$.

Importance Sampling

Lemma

(Near unbiased estimation of $V^\pi(s_0)$) Let π_{uar} denote the policy which chooses actions uniformly at random at every state. We have that:

$$|\mathcal{A}|^H \, \mathbb{E}_{\pi_{\text{uar}}} \left[\mathbf{1}\Big(\pi(s_0) = a_0, \ \ldots, \pi(s_H) = a_H \Big) \sum_{t=0}^{H} \gamma^t r(s_t, a_t) \right]$$

$$= \mathbb{E}_{\pi} \left[\sum_{t=0}^{H} \gamma^t r(s_t, a_t) \right].$$

π_{UAR} – random policy.

RHS is what we want

Importance Sampling

Lemma

(Near unbiased estimation of $V^\pi(s_0)$) Let π_{uar} denote the policy which chooses actions uniformly at random at every state. We have that:

$$|\mathcal{A}|^H \, \mathbb{E}_{\pi_{\text{uar}}} \left[\mathbf{1}\Big(\pi(s_0) = a_0, \ \ldots, \pi(s_H) = a_H\Big) \sum_{t=0}^{H} \gamma^t r(s_t, a_t) \right]$$
$$= \mathbb{E}_{\pi} \left[\sum_{t=0}^{H} \gamma^t r(s_t, a_t) \right].$$

- the estimated the reward of π on a trajectory is nonzero only when π takes precisely the same actions as the π_{uar} on the trajectory then estimated reward of $|\mathcal{A}|^H$ is equal to that of π_{uar}.

Importance Sampling

Lemma

(Near unbiased estimation of $V^\pi(s_0)$) Let π_{uar} denote the policy which chooses actions uniformly at random at every state. We have that:

$$|\mathcal{A}|^H \, \mathbb{E}_{\pi_{\mathrm{uar}}}\left[\mathbf{1}\Big(\pi(s_0) = a_0, \ \ldots, \pi(s_H) = a_H\Big) \sum_{t=0}^{H} \gamma^t r(s_t, a_t)\right]$$
$$= \mathbb{E}_\pi\left[\sum_{t=0}^{H} \gamma^t r(s_t, a_t)\right].$$

- the estimated the reward of π on a trajectory is nonzero only when π takes precisely the same actions as the π_{uar} on the trajectory then estimated reward of $|\mathcal{A}|^H$ is equal to that of π_{uar}.
- the factor of $|\mathcal{A}|^H$ which is due to this being a high variance estimate. We will return to this point in the next section.

An Occams Razor Bound for RL

Gather N trajectories with π_{UAR}

Denote the n-th sample by $(s_0^n, a_0^n, r_1^n, s_1^n, \ldots, s_H^n)$, where H is a cutoff time where the trajectory ends. A nearly, unbiased estimate of the γ-discounted reward of a given policy π is given by:

$$\widehat{V}^\pi(s_0) = \frac{|\mathcal{A}|^H}{N} \sum_{n=1}^{N} \mathbf{1}\Big(\pi(s_0^n) = a_0^n, \ldots \pi(s_H^n) = a_H^n\Big) \sum_{t=0}^{H} \gamma^t r(s_t^n, a_t^n).$$

An Occams Razor Bound for RL

Denote the n-th sample by $(s_0^n, a_0^n, r_1^n, s_1^n, \ldots, s_H^n)$, where H is a cutoff time where the trajectory ends. A nearly, unbiased estimate of the γ-discounted reward of a given policy π is given by:

$$\widehat{V}^\pi(s_0) = \frac{|\mathcal{A}|^H}{N} \sum_{n=1}^{N} \mathbf{1}\Big(\pi(s_0^n) = a_0^n, \ldots \pi(s_H^n) = a_H^n\Big) \sum_{t=0}^{H} \gamma^t r(s_t^n, a_t^n).$$

if we try every policy

$H \approx \frac{\log 1/\epsilon}{1-\gamma}$

Proposition

(Generalization in RL) Suppose Π *is finite. Let* $\widehat{\pi} = \arg\max_{\pi \in \Pi} \widehat{V}^\pi(s_0)$ *and* ~~$\pi^\star = \arg\max_{\pi \in \Pi} V^\pi(s_0)$~~. *With probability at least* $1 - \delta$*:*

truncation error.

exp(t) dep.

$$V^{\widehat{\pi}}(s_0) \geq \arg\max_{\pi \in \Pi} V^\pi(s_0) - \frac{\epsilon}{2} - |\mathcal{A}|^H \sqrt{\frac{2}{N} \log \frac{2|\Pi|}{\delta}}.$$

A VC Theorem for RL

- Suppose $|\mathcal{A}| = 2$. Each $\pi \in \Pi$ can be viewed as Boolean function.
- $\mathrm{VC}(\Pi)$ is well defined.
- For N trajectories, the Sauer–Shelah lemma bounds the number of possible labellings on a set of N trajectories (of length H) by $\left(\frac{eNH}{d}\right)^d$, where $d = \mathrm{VC}(\Pi)$.

growth rate argument

- this leads to the following proposition:

A VC Theorem for RL

- Suppose $|\mathcal{A}| = 2$. Each $\pi \in \Pi$ can be viewed as Boolean function.
- $\mathrm{VC}(\Pi)$ is well defined.
- For N trajectories, the Sauer–Shelah lemma bounds the number of possible labellings on a set of N trajectories (of length H) by $\left(\frac{eNH}{d}\right)^d$, where $d = \mathrm{VC}(\Pi)$.
- this leads to the following proposition:

Proposition

(A VC Theorem for RL) Suppose $|\mathcal{A}| = 2$. *Let* $\widehat{\pi} = \arg\max_{\pi\in\Pi} \widehat{V}^\pi(s_0)$ *and* $\pi^\star = \arg\max_{\pi\in\Pi} V^\pi(s_0)$. *With probability at least* $1-\delta$*:*

$$V^{\widehat{\pi}}(s_0) \geq \arg\max_{\pi\in\Pi} V^\pi(s_0) - 2^H \sqrt{\frac{c}{n}\left(VC(\Pi)\log\frac{2n}{VC(\Pi)} + \log\frac{2}{\delta}\right)},$$

where c is an absolute constant.

Outline

Can we do “effective” agnostic learning?

- What we want, for our agnostic sample complexity:
  - no dependence on $|\mathcal{S}|$ (or logarithmic)
  - poly H dependence
  - to depend reasonably on a complexity measure of $\mathcal{H}$
    e.g. poly $\log|\mathcal{H}|$ dependence

Can we do "effective" agnostic learning?

- What we want, for our agnostic sample complexity:
  - no dependence on $|\mathcal{S}|$ (or logarithmic)
  - poly H dependence
  - to depend reasonably on a complexity measure of $\mathcal{H}$
    e.g. poly $\log|\mathcal{H}|$ dependence
- Is this possible?

Can we do “effective” agnostic learning?

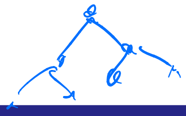
- What we want, for our agnostic sample complexity:
  - no dependence on $|\mathcal{S}|$ (or logarithmic)
  - poly H dependence
  - to depend reasonably on a complexity measure of $\mathcal{H}$
    e.g. poly $\log|\mathcal{H}|$ dependence
- Is this possible?
  No :(

Can we do "effective" agnostic learning?

- What we want, for our agnostic sample complexity:
  - no dependence on $|\mathcal{S}|$ (or logarithmic)
  - poly H dependence
  - to depend reasonably on a complexity measure of $\mathcal{H}$ e.g. poly $\log |\mathcal{H}|$ dependence
- Is this possible?
  No :(
- This is why RL is hard!
  - it is hard in practice...
  - how should we study it?

A "Easy" Lower Bound

$|\Pi| = 2^H$



Proposition

(Lower Bound for The Complete Policy Class) Suppose $|\mathcal{A}| = 2$ *and* $|\mathcal{S}| = 2^H$*, where* $H = \lfloor \frac{\log(2)}{1-\gamma} \rfloor$*. Let* Π *be the set of all* 2^H *policies. There exists a family of MDPs such that if a deterministic algorithm* $\mathcal{A}$ *is guaranteed to find a policy* π *such that:*

$$V^{\widehat{\pi}}(s_0) \geq \arg\max_{\pi \in \Pi} V^{\pi}(s_0) - 1/4.$$

then $\mathcal{A}$ *must use* $N \geq 2^H$ *trajectories.*

A "Easy" Lower Bound

Proposition

(Lower Bound for The Complete Policy Class) Suppose $|\mathcal{A}| = 2$ and $|\mathcal{S}| = 2^H$, where $H = \lfloor \frac{\log(2)}{1-\gamma} \rfloor$. Let Π be the set of all 2^H policies. There exists a family of MDPs such that if a deterministic algorithm $\mathcal{A}$ is guaranteed to find a policy π such that:

$$V^{\widehat{\pi}}(s_0) \geq \arg\max_{\pi \in \Pi} V^{\pi}(s_0) - 1/4.$$

then $\mathcal{A}$ must use $N \geq 2^H$ trajectories.

Observe that $\log|\Pi| = H\log(2)$, so this already rules out the possibility of logarithmic dependence on the size of the policy class, without having an exponential dependence on H.

A General Lower Bound

before: we had a complete policy class.
practice: Π is a restricted (and smaller) class.

$|\Pi|$ – set of all policies.

A General Lower Bound

before: we had a complete policy class.
practice: Π is a restricted (and smaller) class.

Proposition

(Lower Bound for an Arbitrary Policy Class) Suppose $|\mathcal{A}| = 2$ and Π is an arbitrary policy class. There exists a family of MDPs s.t. any algorithm $\mathcal{A}$ that is guaranteed to find a policy $\widehat{\pi}$ s.t.:

$$\mathbb{E}\left[V^{\widehat{\pi}}(s_0)\right] \geq \arg\max_{\pi\in\Pi} V^{\pi}(s_0) - \epsilon.$$

then $\mathcal{A}$ must use an expected number of trajectories N where

$$N \geq c\frac{\min\{2^H, 2^{VC(\Pi)}\}}{\epsilon^2}.$$

VC(π) = d
for halfspaces

A General Lower Bound

before: we had a complete policy class.
practice: Π is a restricted (and smaller) class.

Proposition

(Lower Bound for an Arbitrary Policy Class) Suppose $|\mathcal{A}| = 2$ and Π is an arbitrary policy class. There exists a family of MDPs s.t. any algorithm $\mathcal{A}$ that is guaranteed to find a policy $\widehat{\pi}$ s.t.:

$$\mathbb{E}\left[V^{\widehat{\pi}}(s_0)\right] \geq \arg\max_{\pi \in \Pi} V^{\pi}(s_0) - \epsilon.$$

then $\mathcal{A}$ must use an expected number of trajectories N where

$$N \geq c\frac{\min\{2^H, 2^{VC(\Pi)}\}}{\epsilon^2}.$$

in the worst case, we (nearly) have to do exhaustive search (trying $2^{\mathrm{VC}(\Pi)}$ polices, which is the effective the number of policies in Π)

Outline

How should we study RL?

- The tabular case:
  - we can understand fundamental issues of exploration vs exploitation
  - we can't get at generalization.

How should we study RL?

- The tabular case:
  - we can understand fundamental issues of exploration vs exploitation
  - we can't get at generalization.
- need stronger assumptions/side info.

How should we study RL?

- The tabular case:
  - we can understand fundamental issues of exploration vs exploitation
  - we can't get at generalization.
- need stronger assumptions/side info.
- this course (and the field) take the following approaches:

How should we study RL?

- The tabular case:
  - we can understand fundamental issues of exploration vs exploitation
  - we can't get at generalization.
- need stronger assumptions/side info.
- this course (and the field) take the following approaches:
  - Structural (and Modelling) Assumptions: By making stronger assumptions about the world, we can move away from agnostic learning and escape the curse of dimensionality. We will see examples of this in Part 2.

How should we study RL?

- The tabular case:
  - we can understand fundamental issues of exploration vs exploitation
  - we can't get at generalization.
- need stronger assumptions/side info.
- this course (and the field) take the following approaches:
  - Structural (and Modelling) Assumptions: By making stronger assumptions about the world, we can move away from agnostic learning and escape the curse of dimensionality. We will see examples of this in Part 2.
  - Distribution Dependent Results (and Distribution Shift): When we move to policy gradient methods (in Part 3), we will consider results which depend on given distribution of how we obtain samples. Here, we will make connections to transfer learning.

How should we study RL?

- The tabular case:
  - we can understand fundamental issues of exploration vs exploitation
  - we can't get at generalization.
- need stronger assumptions/side info.
- this course (and the field) take the following approaches:
  - Structural (and Modelling) Assumptions: By making stronger assumptions about the world, we can move away from agnostic learning and escape the curse of dimensionality. We will see examples of this in Part 2.
  - Distribution Dependent Results (and Distribution Shift): When we move to policy gradient methods (in Part 3), we will consider results which depend on given distribution of how we obtain samples. Here, we will make connections to transfer learning.
  - Imitation learning and behavior cloning: here will consider models where the agent has input from, effectively, a teacher, and we will see how this alleviates the problem of curse of dimensionality.