Lecture 8
Planning with global access and uniform realizability

Administrivia: Projects

- What makes a good project?

- Choose a topic
- If you choose a paper, improve on it
  - How does it fit into the big picture?
  - What is left? Can you add anything?
  - Can you make the proof nicer?
- Move on if the paper is not “good”
- Start early
- Ask for help

Global access

Can get all the features at all states, can preprocess it

ζ −uniform action-value realizability

$$\sup_{\pi \in \mathrm{DETML}} \inf_{\theta} \|q^{\pi} - \Phi\theta\|_{\infty} \leq \zeta$$

Note

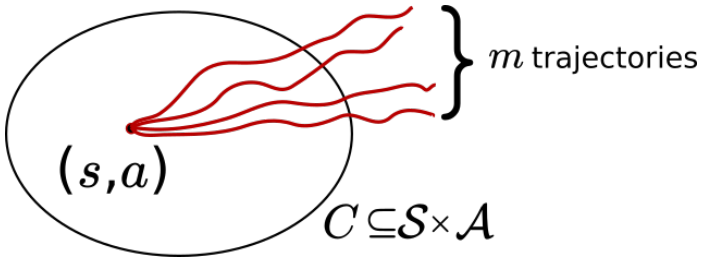
For finite MDPs, $\forall \pi \in \mathrm{ML}, \exists m > 0, (\pi_i)_{i \in [m]}$ st. $q^{\pi} \in \Sigma_i \alpha_i q^{\pi_i}$ with some $\alpha_i \geq 0, \sum_i \alpha_i = 1$

(Dadashi et al.)

least-squares policy evaluation

Least-squares Policy Evaluation

1. Rollouts from a set $\mathcal{C}$ of well-chosen state-action pairs
2. Average over those
3. Least-squares fit



$$\left(\mathbb{E}\left[\hat{R}_m(z)\right] = q^{\pi}(z)\right)$$

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \sum_{z \in \mathcal{C}} \varrho(z) \Big(\langle \theta, \varphi(z) \rangle - \hat{R}_m(z) \Big)^2$$

$$\hat{R}_m(z) = \frac{1}{m} \sum_{j=1}^{m} \sum_{t=0}^{H^{(j)}-1} r_{A_t^{(j)}}(S_t^{(j)})$$

$$\mathbb{P}(H^{(j)} \geq t+1) = \gamma^t$$

First head with $X_1, X_2, \ldots$:
$\mathbb{P}(X_i = \text{Head}) = 1 - \gamma$

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \sum_{z \in \mathcal{C}} \varrho(z) \Big(\langle \theta, \varphi(z) \rangle - \hat{R}_m(z) \Big)^2$$

Alternative: Choose H large enough and let

$$\hat{R}_m(z) = \frac{1}{m} \sum_{j=1}^{m} \sum_{t=0}^{H-1} \gamma^t r_{A_t^{(j)}}(S_t^{(j)}).$$

Lemma will be for this choice.
Homework: Think about the pros and cons of switching to the trajectories with random length

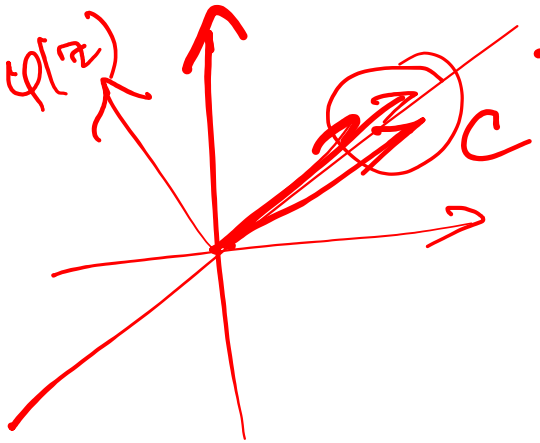
$$G_\varrho = \sum_{z \in C} \varrho(z)\varphi(z)\varphi(z)^\top$$

Lemma (extrapolation error control in least-squares): Fix any $\theta \in \mathbb{R}^d$, $\varepsilon : \mathcal{Z} \to \mathbb{R}$, $C \subset \mathcal{Z}$ and $\varrho \in \Delta_1(C)$ such that the moment matrix G_ϱ is nonsingular. Define

$$\hat{\theta} = G_\varrho^{-1} \sum_{z' \in C} \varrho(z') \Big(\varphi(z')^\top \theta + \varepsilon(z') \Big) \varphi(z') \,.$$

Then, for any $z \in \mathcal{Z}$ we have

$$\left| \varphi(z)^\top \hat{\theta} - \varphi(z)^\top \theta \right| \le \|\varphi(z)\|_{G_\varrho^{-1}} \max_{z' \in C} |\varepsilon(z')| \,.$$



optimal design

Theorem (Kiefer-Wolfowitz): Let $\mathcal{Z}$ be finite. Let $\varphi : \mathcal{Z} \to \mathbb{R}^d$ be such that the underlying feature matrix Φ is rank d. There exists a set $\mathcal{C} \subseteq \mathcal{Z}$ and a distribution $\varrho : C \to [0, 1]$ over this set, i.e. $\sum_{z' \in \mathcal{C}} \varrho(z') = 1$, such that

1. $|\mathcal{C}| \leq d(d+1)/2$;
2. $\sup_{z \in \mathcal{Z}} \|\varphi(z)\|_{G_\varrho^{-1}} \leq \sqrt{d}$;
3. In the previous line, the inequality is achieved with equality and the value of $\sqrt{d}$ is best possible under all possible choices of $\mathcal{C}$ and ρ.

extrapolation error control with LS

Corollary (extrapolation error control in least-squares via optimal design): Fix any $\varphi : \mathcal{Z} \to \mathbb{R}^d$ full rank. Then, there exists a set $\mathcal{C} \subset \mathcal{Z}$ with at most $d(d+1)/2$ elements and a weighting function $\varrho \in \Delta_1(\mathcal{C})$ such that for any $\theta \in \mathbb{R}^d$ and any $\varepsilon : \mathcal{C} \to \mathbb{R}$,

$$\max_{z \in \mathcal{Z}} \left| \varphi(z)^\top \hat{\theta} - \varphi(z)^\top \theta \right| \leq \sqrt{d} \max_{z' \in C} \left| \varepsilon(z') \right| .$$

where $\hat{\theta}$ is given by

$$\hat{\theta} = G_\varrho^{-1} \sum_{z' \in C} \varrho(z') \Big(\varphi(z')^\top \theta + \varepsilon(z') \Big) \varphi(z') .$$

C and ϱ are chosen independently of θ and ϵ!

Lemma (LSPE-G extrapolation error control): Fix any full-rank feature-map $\varphi : \mathcal{Z} \to \mathbb{R}^d$ and take the set $\mathcal{C} \subset \mathcal{Z}$ and the weighting function $\varrho \in \Delta_1(\mathcal{C})$ as in the Kiefer-Wolfowitz theorem. Fix an arbitrary policy π and let θ and ε_π such that $q^\pi = \Phi\theta + \varepsilon_\pi$ and assume that immediate rewards belong to the interval $[0, 1]$. Let $\hat{\theta}$ be as in Eq. (6). Then, for any $0 \le \delta \le 1$, with probability $1 - \delta$,

$$\left\| q^\pi - \Phi\hat{\theta} \right\|_\infty \le \|\varepsilon_\pi\|_\infty (1 + \sqrt{d}) + \sqrt{d} \left(\frac{\gamma^H}{1-\gamma} + \frac{1}{1-\gamma} \sqrt{\frac{\log(2|C|/\delta)}{2m}} \right). \tag{7}$$

2ε

KW: $|\mathcal{C}| \le d(d+1)/2$, hence to make second term $\le 2\varepsilon$, enough if

$$H \ge H_{\gamma, \varepsilon/\sqrt{d}} \quad \text{and} \quad m \ge \frac{d}{(1-\gamma)^2 \varepsilon^2} \log \frac{d(d+1)}{\delta}$$

Total # samples: $|\mathcal{C}|Hm \approx \dfrac{d^3 \log(d/\varepsilon) \log(d/\delta)}{(1-\gamma)^3 \varepsilon^2}$

Questions from slack

Ehsan Imani 5 hours ago

In the proof of extrapolation error control lemma we bound the absolute value of (weighted) sum of

$$\sum_{z'} \varepsilon(z')\varphi(z')^\top G_\varrho^{-1}\varphi(z')$$

by the (weighted) sum of

$$\sum_{z'} |\varepsilon(z')|\,\left|\varphi(z')^\top G_\varrho^{-1}\varphi(z')\right|$$

Generally, for two uncorrelated random variables X and Y, the magnitude of the sum of XY over some samples could be much smaller than the sum of $|X|\ |Y|$. Now in the proof if ε is mostly made up of a "variance" component that is not correlated with the features, can we reduce its effect on the bound in this way? Would the improvement in the bound worth the extra hassle?

And does the "helpful averaging" in the notes refer to this?!

+8

Discussion

What is compute cost?

Example 1:

Theorem (Runtime Bound for Policy Iteration): Consider a finite, discounted MDP with rewards in $[0, 1]$. Let k^* be as in the progress lemma, $\{\pi_k\}_{k \geq 0}$ the sequence of policies obtained by policy iteration starting from an arbitrary initial policy π_0. Then, after at most $k = k^*(\mathrm{SA} - \mathrm{S}) = \tilde{O}\left(\frac{\mathrm{SA}-\mathrm{S}}{1-\gamma}\right)$ iterations, the policy π_k produced by policy iteration is optimal: $v^{\pi_k} = v^*$. In particular, policy iteration computes an optimal policy with at most $\tilde{O}\left(\frac{\mathrm{S}^4\mathrm{A}+\mathrm{S}^3\mathrm{A}^2}{1-\gamma}\right)$ arithmetic and logic operations.

Is this satisfactory? Does this mean that if PI is coded up, it will give the optimal policy?

Problems

No infinite precision arithmetic on computers

Floating point is funky:

```
>>> .1+.1+.1==.3
False
```

Errors can propagate, get large, overwhelm, ..

Order of operations matter

sum([1]+[1.0/1 billion]*1 billion) == 2?

Can't invert "ill conditioned" matrices. Do we have those?

$(-1)^{\text{sign}}(1.b_{51}b_{50}\ldots b_0)_2 \times 2^{e-1023}$

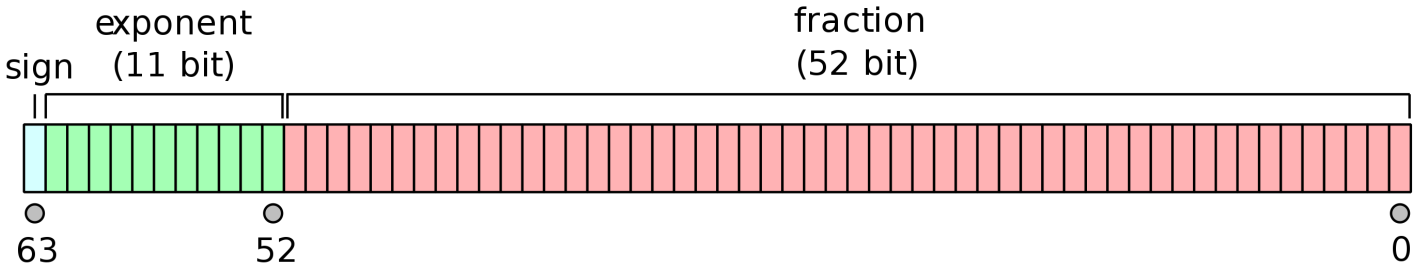
The goal

To know whether some calculations are ”safe”

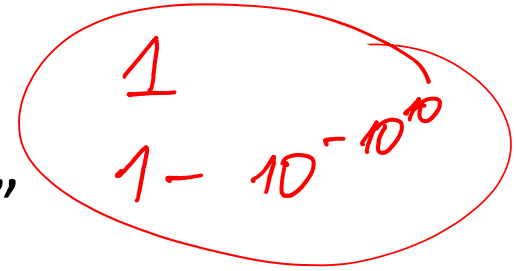
Proposition:

With floating point using $e + f$ bits and target accuracy $\varepsilon > 0$, effective horizon $H = 1/(1-\gamma)$, provided $SA \leq u(e, f, 1/\varepsilon)$, for any MDP M of this size, policy iteration returns an ε-optimal policy.

What is u? How do we get u?



Approaches: Computation over reals



Want computation over the "reals"

E.g. cost of $A \mapsto A^{-1}b$, or $\mathrm{x} \mapsto \sqrt{x}$

- Model 1: Turing machines
- Model 2: #operations in program that uses infinite precision arithmetic (BSS model)
- Model 3: Bit model. Program runs on Turing machine, gets as input target accuracy, computes desired precision of input, gets arbitrarily rounded inputs

What are the strengths and weaknesses?