Function Approximation and the NPG

Presented by :
Atmani Hanan

University mohammed vi polytechnic

July 23

Presentation plan

Introduction

- In general, the MDP is infinite, so enumerating over state-action pairs is not possible in terms of computation, space, and statistics. What should we do in this case?
- Generalization via function approximation.
- In our case, we will use linear regression, which is used in supervised learning, to approximate the target functions in both the value iteration and policy iteration algorithms.

Recap on supervised learning: regression

- We have a data distribution $\mathcal{D}, x_i \sim \mathcal{D}, y_i = f^*(x_i) + \epsilon_i$ where noise $\mathbb{E}(\epsilon_i) = 0$
- Unknown target function: $f : X \to Y$
- Set of function hypotheses: $\mathcal{F} = \{f \mid f : X \to Y\}$
- $\hat{f} = \text{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{N} (f(x_i) - y_i)^2$
- Subervised learning theory (e.g VC theory) says that we can indeed genralize i.e we can predict well under the same distribution. Assume $\hat{f} \in \mathcal{F}$ (this is called realizability) we can expect:

$$\mathbb{E}_{x \in \mathcal{D}} (\hat{f}(x) - f^*(x))^2 \leq \delta$$

  Where δ is a small number ($\frac{1}{N} or \frac{1}{\sqrt{N}}$)
- subervised learning can fait if there is train-test distribution mismath,
- However, for som $\mathcal{D}' \neq \mathcal{D}$, $\mathbb{E}_{x \in \mathcal{D}'} (\hat{f}(x) - f^*(x))^2$ might be arbiteary large

- We consider an infinite horizon discounted MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \gamma, P, r, \mu\}$, where μ is the initial state distribution. We assume the reward is bounded, i.e., $\sup_{s,a} r(s,a) \in [0,1]$. For notation simplicity, we denote $V_{\max} := \frac{1}{1-\gamma}$.
- Visitation measure over just the states:

$$d^{\pi}_{s_0}(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \Pr^{\pi}(s_t = s \mid s_0).$$

- We write

$$d^{\pi}_{\mu}(s) = \mathbb{E}_{s_0 \sim \mu}[d^{\pi}_{s_0}(s)].$$

- As we will consider large scale unknown MDP here, we start with a (restricted) function class:

$$\mathcal{F} = \{f \mid f : \mathcal{S} \times \mathcal{A} \to [0, V_{max}]\}$$

Approximate Policy iteration

- Like policy iteration, we iterate between two steep
  1. Policy Evaluation $\hat{Q}^t \approx Q^{\pi^t}$ (where $\hat{Q}^t \in \mathcal{F}$)
  2. Policy Improvement $\pi^{(t+1)}(s) = \text{argmax}_a \hat{Q}^t(s,a)$
- We use supervised learning (regression) to estimat Q^{π^t}
  1. Haw to get training data?
     - Roll-in with policy π to time step t, returning (s_t, a_t). We start with $s_0 \sim \mu$ and $a_0 \sim \pi(.|s_0)$, and follow policy π until time step t, returning the pair (s_t, a_t)
  2. Quality of the learned $\hat{Q}^t$?
     - A roll-out from (s,a) gives an unbaised estimate of $Q^\pi(s,a)$, $\mathbb{E}(y) = Q^\pi(s,a)$
- Summary of the dataset generation process
  1. We roll-in to generate $(s,a) \sim \upsilon$
  2. At (s,a), we roll-out with policy π to generate an unbiased estimate of Q^π $(Q^\pi(s,a) = y)$
  3. In other worlds, one roll-in & roll-out gives us a triple (s,a,y)

Approximate Policy iteration

Estimating the function $Q^\pi(s,a)$ using least square regression:

- Given π, repeat N times of the roll-in & roll-out process, we get a training dataset of N samples:

$$\mathcal{D}^\pi = \left\{s^i, a^i, y^i\right\}_{i=1}^N, \quad s^i, a^i \sim \upsilon, \quad \mathbb{E}(y^i) = Q^\pi(s^i, a^i)$$

- Least square regression:

$$\hat{Q} = argmin_{Q\in\mathcal{F}} \sum_{i=1}^{N} (Q(s^i, a^i) - y^i)^2$$

Assume success ful supervise learning we have,

$$\mathbb{E}_{s,a\sim\upsilon}(\hat{Q}^\pi(s,a) - Q^\pi(s,a))^2 \leq \delta$$

Approximate Policy iteration

- Initialize $Q^0 \in \mathcal{F}$, $\pi^0 = \text{argmax}_a \hat{Q}^0(s,a)$
- Data generalization process (roll-in & roll-out $(\mathcal{D}^{\pi^t} = \{s^i, a^i, y^i\}_{i=1}^N, \quad s^i, a^i \sim \upsilon, \quad \mathbb{E}(y^i) = Q^{\pi^t}(s^i, a^i))$
- Square Regression oracle

$$\hat{Q}^t = argmin_{Q \in \mathcal{F}} \sum_{i=1}^{N} (Q(s^i, a^i) - y^i)^2$$

- Policy Improvement $\pi^{(t+1)}(s) = \text{argmax}_a \hat{Q}^t(s,a)$
- Remark

$$\text{argmax}_a A^\pi(s,a) = \text{argmax}_a Q^\pi(s,a)$$

Therefore, we can use the advantage function A in the algorithm (API) instead of using Q.

- Given the current policy π^t, let's find a new policy that has large adv over π^t, under $d_\mu^{\pi^t}$ (i.e let's aim to solve the following program $\text{argmax}_{\pi\in\Pi}\mathbb{E}_{s\sim d_\mu^{\pi^t}} A^{\pi^t}(s,\pi(s))$ Greedy policy selector)
- How to impliment such greedy policy selector?
- Implementing Approximate Greedy policy Selector via regression
- We can do a reduction to regression via advantage function approximation

$$\mathcal{F} = \{f \mid f : \mathcal{S}\times\mathcal{A}\to\mathbb{R}\}, \Pi = \{\pi(s) = \text{argmax}_a f(s,a) \mid f\in\mathcal{F}\}$$

- $\{s^i, a^i, y^i\}_{i=1}^N, \quad s^i \sim d_\mu^{\pi^t}, a^i \sim \mathcal{U}(\mathcal{A}), \quad \mathbb{E}(y^i) = A^{\pi^t}(s^i, a^i)$
- Regression oracle:

$$\hat{A}^t = \text{argmin}_{f\in\mathcal{F}} \sum_{i=1}^N (f(s^i,a^i) - y^i)^2$$

- At greedy with the estimator $\hat{A}^t$ (as we hope $\hat{A}^t \approx A^{\pi^t}$): $\hat{\pi}(s) = \text{argmax}_a \hat{A}^t(s,a), \forall s$
- Assume this regression is successful i.e

$$\mathbb{E}_{s\sim d_\mu^{\pi^t}, a^i \sim \mathcal{U}} (\hat{A}^t(s,a) - A^{\pi^t}(s,a))^2 \leq \epsilon$$

- Then, $\hat{\pi}(s) = \text{argmax}_a \hat{A}^t(s,a)$ is the approximate greedy policy:

$$\mathbb{E}_{s\sim d_\mu^{\pi^t}} (A^{\pi^t}(s, \hat{\pi}(s)) \geq \max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}} A^{\pi^t}(s, \pi(s)) - \epsilon,$$

  Where $\epsilon = \frac{1}{N}$ or $\frac{1}{\sqrt{N}}$, denote $\hat{\pi} = \mathcal{G}_\varepsilon(\pi^t, \Pi, \mu)$
- The failure case of the API: An abrupt distribution change means the API cannot guarantee success. To ensure monotonic improvement of the API, we need a strong concentrability ratio assumption $\frac{d^\pi(s,a)}{\upsilon(s,a)}$

Conservative Policy Iteration

- Key idea of CPI: Incremental update-no abrupt distribution change,
- Making small incremental update to the policy by forcing that the new policy's state action distribution is not too far away from the current policy's
- Achieves that by forming a new polcy is mixture of the current policy and a local greedy policy

Conservative Policy Iteration Algorithm

1. Initialize π^0,
2. For $t = 0, \ldots$
   1. Greedy policy selector:
   $$\pi' = \mathcal{G}_\varepsilon(\pi^t, \Pi, \mu)$$
   2. If $\mathbb{E}_{s \sim d_\mu^{\pi^t}} A^{\pi^t}(s, \pi'(s)) \leq \epsilon$: Return π^t
   3. Incremental update :
   $$\pi^{(t+1)}(.|s) = (1 - \alpha)\pi^t(.|s) + \alpha\pi'(.|s), \forall s$$
3. Fin for

- If (2) is true: This means that if I cannot find a policy π' that has a positive advantage over the current policy π^t, then π^t is optimal.
- With Probability $1 - \alpha$, $a \sim \pi^t(.|s)$
- With Probability α, $\pi'(.|s)$

Conservative Policy Iteration

1. Why this is incremental? In what sense?
2. Can we get monotonic poicy improvement?

Lemma

(Similar Policies imply similar state visitations). Consider any t, we have that:

$$\left\|\pi^{t+1}(\cdot \mid s) - \pi^{t}(\cdot \mid s)\right\|_1 \leq 2\alpha, \forall s;$$

Further, we have:

$$\left\| d_{\mu}^{\pi^{t+1}} - d_{\mu}^{\pi^{t}} \right\|_1 \leq \frac{2\alpha\gamma}{1-\gamma}.$$

Conservative Policy Iteration

- Monotonic improvement before termination: Before termination, we have a non-trivial average local advantage: $\mathbb{A} = \mathbb{E}_{s\sim d_\mu^{\pi^t}} A^{\pi^t}(s, \pi'(s)) \geq \epsilon$.
- By the performance difference lemma

$$\begin{aligned} V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t} &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}}\mathbb{E}_{a\sim\pi^{t+1}(.|s)} A^{\pi^t}(s,a) \\ &= \frac{\alpha}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi^{t+1}}} A^{\pi^t}\left(s, \pi'(s)\right) \end{aligned}$$

Monotonic Improvement in CPI

Theorem

Consider any episode t. Denote $\mathbb{A} = \mathbb{E}_{s\sim d_\mu^{\pi^t}} A^{\pi^t}(s, \pi'(s))$. We have:

$$V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t} \geq \frac{\alpha}{1-\gamma}\left(\mathbb{A} - \frac{2\alpha\gamma}{(1-\gamma)^2}\right)$$

Set $\alpha = \frac{\mathbb{A}(1-\gamma)^2}{4\gamma}$, we get:

$$V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t} \geq \frac{\mathbb{A}^2(1-\gamma)}{8\gamma}$$

Proof: Indication $|\mathbb{E}_{x\sim p} f(x) - \mathbb{E}_{x\sim q} f(x)| \leq max_x |f(x)|||p-q||_1$

Local optimality of CPI

Theorem

Algorithn CPI terminates in at most $8\gamma/c^2$ *steps and outputs a policy* π^t *satisfying* $\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}} A^{\pi^t}(s,\pi(s)) \leq 2\varepsilon$.

Global optimality

Theorem

Upon termination, we have a policy π such that:

$$V^\star - V^\pi \leq \frac{2\varepsilon + \epsilon_\Pi}{(1-\gamma)^2} \left\| \frac{d^{\pi^*}}{\mu} \right\|_\infty ,$$

where
$\epsilon_\Pi := \mathbb{E}_{s\sim d_\mu^\pi} [\max_{a\in\mathcal{A}} A^\pi(s,a)] - \max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^\pi} [A^\pi(s,\pi(s))]$. *In other words, if our policy class is rich enough to approximate the policy* $\max_{a\in A} A^\pi(s,a)$ *under* d_μ^π, *i.e.,* ϵ_Π *is small, and* μ *covers* d^{π^*} *in a sense that* $\left\| \frac{d^{d^*}}{\mu} \right\|_\infty \leq \infty$, *CPI guarantees to find a near optimal policy.*