Week 11

↳ Attendance required (unless prior communication)

↳ Anonymous eval:

☺ white board practice, office hours, lecture notes, practice exams

(me) homework difficulty (require talking), grade curve, more math, practice problems, less math, programming in class, quiz questions taken advantage of

(you) office hours, structure time,

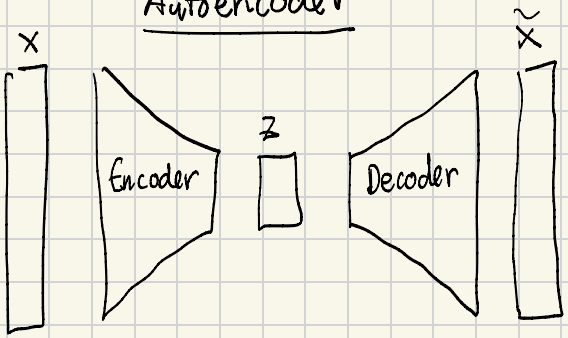
difficult class,

I've been thinking about how to say this: I feel like at CMC, you're the only professor in the Math/CS department that believes that we are capable of mastering material at the level that you are presenting it. I appreciate this external belief in my capabilities because it pushes me to spend that extra hour thinking about the problem set, and it encourages me to ask questions in class. Although I'm sure by now you realize that the students aren't as comfortable with linear algebra, notation, this "sitting on the rock" idea (without external support/without personal effort), I'd advocate that you continue to push for this level of content/understanding from us. Compared to day one, many of us are more comfortable with notation on the board, following a proof, and thinking through complex problems on the problem set.

↳ Slower pace this week: RL (no q-learning)

Unsupervised Learning

Autoencoder



Why? • efficient
• meaningful

Variational Autoencoder

Latent space nicely distributed

Principal Component Analysis

Optimal linear autoencoder ≈ SVD

Motivation for Semisupervised learning

In supervised and unsupervised, data is static and learning is offline

Often, data depends on predictions:

↳ Games (video games, board games)

↳ Autonomous movement

↳ Driving

↳ Stock market

⇒ Reinforcement Learning

↳ Alpha GO ↳ DOTA 2

↳ LLMs

Reinforcement Learning

Q: How do humans learn?

A: Try things in chaotic environment,
repeat actions that lead to good outcomes

In a state, take some action to max reward

	Temple Run	Stock Market	LLM
State	Pixel world		
Action	left/right/slide/jump		
Reward	- Perish, coins		

Mathematical Formulation

Time step $t = 1, 2, \cdots, L$ (horizon length)

state: $s_t = f(s_{t-1}, a_{t-1})$

action: a_t comes from policy $\pi_\theta(s_t)$ (This is what we learn! (in terms of θ))

reward: $r(s_t, a_t)$

Trajectory $s_0, a_0, s_1, a_1, \cdots, s_{L-1}, a_{L-1}$

$$\underset{\pi}{\text{minimize}} \sum_{t=0}^{L-1} -r(s_t, a_t)\, \gamma^t$$

(γ^t: discount future reward via $\gamma \in (0, 1]$)

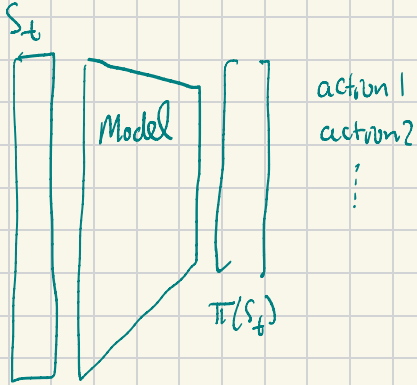
where $s_{t+1} = f(s_t, a_t)$

$a_t \sim \pi_\theta(s_t)$

Random state $f(s_t, a_t)$

Random reward $r(s_t, a_t)$

Random policy $\pi(s_t)$ is a distribution



$$\underset{\pi}{\text{minimize}}\ \mathbb{E} \sum_{t=0}^{L-1} -r(s_t, a_t)\, \gamma^t$$

where $s_{t+1} = f(s_t, a_t)$

$a_t \sim \pi_\theta(s_t)$

Logistics

↳ OH Monday and Thursday 12:30-2

↳ Midterm 11/25

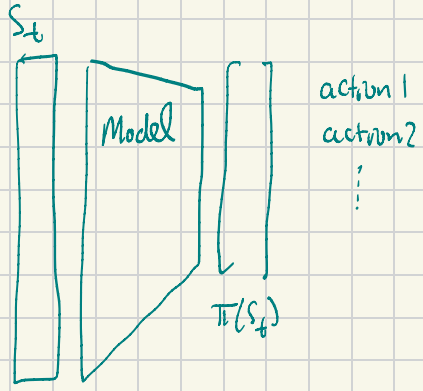
↳ Project after Thanksgiving

Review

$$\min_{\pi} \mathbb{E}\left[\sum_{t=0}^{L-1} -r(s_t, a_t)\ \gamma^t\right]$$

where $s_{t+1} = f(s_t, a_t)$

$a_t \sim \pi_\theta(s_t)$



Policy Gradients

Idea: estimate expectation via roll outs!

$$\tau = s_0, a_0, r_0, s_1, a_1, r_1, \ldots$$

$$R(\tau) = \sum_{t=0}^{L-1} \gamma^t \cdot r(s_t, a_t)$$

$\pi_\theta(\tau)$ = probability of τ from policy π_θ

$$\min_{\pi} \; \mathbb{E}_{\tau \sim \pi_\theta}[R(\tau)]$$

Goal: Compute $\nabla_\theta \mathbb{E}_{\tau \sim \pi_\theta}[R(\tau)]$

$$\mathbb{E}_{\tau \sim \pi_\theta}[R(\tau)] = \sum_\tau R(\tau)\, \pi_\theta(\tau)$$

↑ how do we differentiate?

log-derivative: $\frac{\partial}{\partial \theta} \pi(\tau) = \pi(\tau) \frac{\partial}{\partial \theta} \log \pi(\tau)$

$$\frac{\partial}{\partial \theta} \mathbb{E}_{\tau \sim \pi_\theta} R(\tau) = \sum_\tau R(\tau) \frac{\partial}{\partial \theta} \pi(\tau)$$

$$= \sum_t R(\tau)\, \pi(\tau) \frac{\partial}{\partial \theta} \log \pi(\tau)$$

$$= \mathbb{E}_{\tau \sim \pi_\theta}\left[R(\tau) \frac{\partial}{\partial \theta} \log \pi(\tau) \right]$$

REINFORCE Algorithm

1. Sample $\tau = (s_0, a_0, r_0, \ldots, s_{L-1}, a_{L-1}, r_{L-1})$

   for $\ell = 0, \ldots, L$:

2. Compute $R_\ell(\tau) = \sum_{t=\ell}^{L-1} r(s_t, a_t)$

3. $\theta \leftarrow \theta - \alpha \; R_\ell(\tau) \frac{\partial}{\partial \theta} \log \pi(s_t)[a_t]$

Fundamentally, we are estimating $\mathbb{E}_{\tau \sim \pi_\theta}\left[R(\tau) \frac{\partial}{\partial \theta} \log \pi(\tau) \right]$

Discussion

↳ Reduce variance by subtracting baseline b

↳ Reward can be non-differentiable!!

↳ Need differentiable policy, can avoid with evolutionary search

↳ A kind of random search

1. sample direction v
2. Search for η to minimize loss on $\theta - \eta v$
3. Update $\theta \leftarrow \theta - \eta v$