Tuesday, April 14

- Midterm next Thursday (4/23)
- No class next Tuesday (4/21)
- Office hours:
  - ↳ Canceled next Thursday (4/23)
  - ↳ Virtual Monday (4/20)
  - ↳ Extra tomorrow (4/15)

If second midterm grade is higher, I will replace first midterm grade.

Plan

Reinforcement learning!

Up to now, we assumed data is static and learning is off line but...

↳ playing games
↳ autonomous movement
↳ self-driving
↳ trading
↳ chatting

Challenge: Predictions influence future data

⇒ Reinforcement learning
(AlphaGo, DOTA 2 AI, ChatGPT)

Motivation: Natural intelligence

↳ infant
↳ 2-year old
↳ middle school
↳ college

In state, take action, get reward, update state

	Temple Run	Stock Market	Chat
State	Pixels		
Action	Left/right/jump/slide		
Reward	Alive + coins		
Update	Next pixels		

Thursday, April 16

Midterm Thursday, April 23

↳ virtual OH Monday

↳ no class Tuesday

↳ no OH Thursday

Project

Proposal due April 27

↳ explore topic/algorithm of your choice from class

Probability Review

X random variable

$\Pr(X=x)$ = "probability X takes value x"

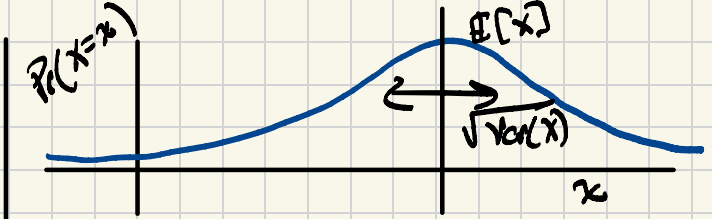
$$\mathbb{E}[X] = \sum_x x \Pr(X=x)$$

Linearity of Expectation: X, Y r.v.

$$\mathbb{E}[X+Y] = \sum_x \sum_y (x+y) \Pr(X=y \text{ and } Y=y)$$

$$= \sum_x x \sum_y \Pr(X=x \text{ and } Y=y)$$
$$+ \sum_y y \sum_x \Pr(X=x \text{ and } Y=y)$$

$$= \sum_x x \Pr(X=x) + \sum_y y \Pr(Y=y)$$
$$= \mathbb{E}[X] + \mathbb{E}[Y]$$



$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

$$= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2]$$

$$= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2$$

$$= \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Linearity of Variance: If X, Y independent i.e., for all x, y

$\Pr(X=x \text{ and } Y=y) = \Pr(X=x)\Pr(Y=y)$,

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

Mathematical Formulation

Time steps $t = 1, 2, \ldots, T$

At time t,

- State s_t
- Take action a_t
- Get reward $r(s_t, a_t)$
- Update $s_{t+1} = f(s_t, a_t)$

} Can all be stochastic!

Trajectory $\tau = s_1, a_1, s_2, a_2, \ldots, s_T, a_T$

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

↖ discount future reward $0 < \gamma < 1$

Model

Policy π_θ: State → action dist.

$$a_t \overset{*}{\sim} \pi_\theta(s_t)$$

Loss

$$\mathcal{L}(\theta) = \mathop{\mathbb{E}}_{\tau \sim \pi_\theta} R(\tau)$$

$$= \sum_{\tau} \pi_\theta(\tau)\, R(\tau)$$

$\pi_\theta(\tau)$: prob. of τ

In practice, sample τ and approximate $\mathcal{L}$.

* If deterministic, could memorize e.g., Atari-style games

Policy Gradient

$$\mathcal{L}(\theta) = \mathop{\mathbb{E}}_{\tau \sim \pi_\theta} R(\tau) = \sum_\tau \pi_\theta(\tau)\, R(\tau) \approx \frac{1}{m} \sum_{\ell=1}^{m} \pi_\theta(\tau_\ell)\, R(\tau_\ell)$$

estimate

$$\nabla_\theta \mathcal{L}(\theta) = \sum_\tau R(\tau)\, \nabla_\theta \pi_\theta(\tau)$$

can compute using back prop but where's the expectation?

Goal: We want $\nabla_\theta \mathop{\mathbb{E}}_{\tau \sim \pi_\theta} R(\tau)$ to be expectation so we can estimate

Hint: Compute $\nabla_\theta \log \pi_\theta(\tau) = \ldots$

Then $\nabla_\theta \mathop{\mathbb{E}}_{\tau \sim \pi_\theta} R(\tau) = \ldots$

REINFORCE Algorithm

For $l = 1, \ldots, m$:

Sample $\tau = (s_1, a_1, \ldots, s_T, a_T)$ from π_θ

Compute $R(\tau) = \sum_{t=1}^{T} -r(s_t, a_t)\, \gamma^{t-1}$

$\theta \leftarrow \theta - \alpha\, R(\tau)\, \nabla_\theta \log \pi(\tau)$

$\tau_1 = (s_1, a_1, \ldots, s_T, a_T)$

$\tau_2 = (s_2, a_2, \ldots, s_T, a_T)$

$\vdots$

Note that:

↳ reward can be non-differentiable

↳ we can extract more trajectories

↳ Right in expectation but variance can be high... subtract baseline

(Also view as "centering" neutral reward)