last time

- SFT / instruction tuning
  - full model tuning, $\frac{dL}{d\theta}$
  - LoRA
- distillation
  - target response is generated by a "teacher" model

---

policy: the model we are currently training π_{ref}, π

rollout: a complete response sampled from the policy model

reward: a score of how good a particular rollout is

↳ dense reward: a score for each token

↳ sparse reward: one score for the entire rollout

off-policy training

↳ target response comes from humans, teacher models, etc

↳ SFT, vanilla distillation, pretraining

on-policy training

↳ responses come from the model that's being trained

↳ OPD, RL

exposure bias (issue w/ off-policy algos)

↳ during training, model always is trained to continue a ground-truth prefix

↳ however, at test time, the model continues its own prefixes

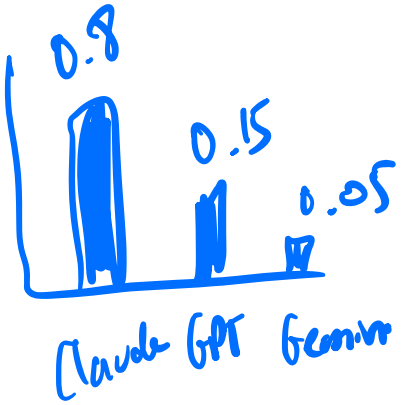
↳ off-policy training doesn't contain prefixes with the model's own mistakes

---

distillation

↳ so far, we covered SFT on teacher rollouts

teacher model: my favorite LLM is



cross-entropy loss:

↳ p : target / teacher model

q : current / student model

$$CE : \sum_{w \in V} p(w)(-\log q(w))$$

↳ we can compute this CE loss on soft targets (not one-hot) from the teacher at each timestep

↳ only works w/ access to teacher logits

↳ does not work if you have just blackbox API access

How do we measure how similar p and q are?

↳ p: teacher ↳ q: student

KL divergence

$$KL(p \| q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$p(x)$: weight each outcome by teacher prob

$\frac{p(x)}{q(x)}$: how far off q is from p on a particular word

↳ if $q(x) = p(x)$, no penalty

$q(x) < p(x)$, positive penalty

$q(x) > p(x)$, negative term

KL always ≥ 0,

not symmetric, $KL(p \| q) \neq KL(q \| p)$

$$KL(p \| q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$$= \underbrace{\sum_x p(x)(-\log q(x))}_{CE(p,q)} - \underbrace{\sum_x p(x)(-\log p(x))}_{\text{entropy of the target/teacher}}$$

↳ w/ one-hot p (pretraining, SFT, Vanilla distillation)

entropy $= 0$

$$KL(p \| q) = -\log q(w)$$

Expectation

↳ probability-weighted average

$$\mathop{E}_{x \sim p}\left[r(x)\right] = \sum_{x} p(x) r(x)$$

↳ usually intractable to enumerate all rollouts w/ nonzero $p(x)$

↳ we can estimate it by sampling

$$\mathop{E}_{x \sim p}\left[r(x)\right] \approx \frac{1}{N} \sum_{i}^{N} r(x_i)$$

forward KL:

$$KL(p \| q) = E_{x \sim p}\left[\log p(x) - \log q(x)\right]$$

↳ rollouts | targets come from teacher

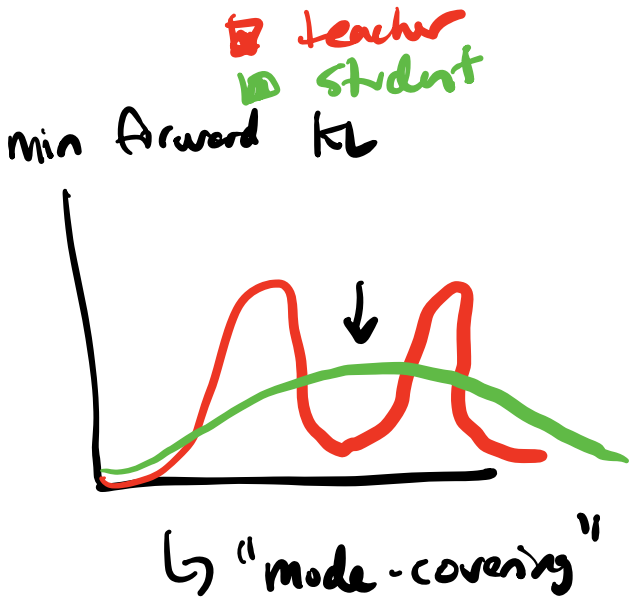
reverse KL:

$$KL(q \| p) = E_{x \sim q}\left[\log q(x) - \log p(x)\right]$$

↳ rollouts from student

↳ need access to the teacher probabilities

forward vs. reverse KL



↳ "mode-covering"



↳ "mode seeking"