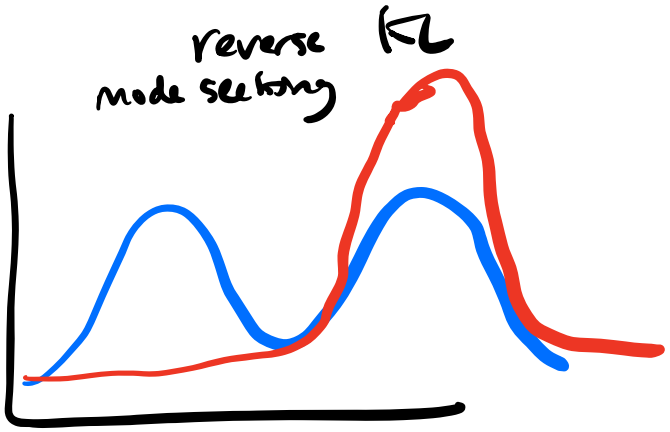




notation

p : teacher dist., also π_T

q : student dist., π_S

x : prompts

y : response / rollout

forward KL:

$$KL(\pi_T \| \pi_S) = \mathop{E}_{y \sim \pi_T}\left[\log \pi_T(y) - \log \pi_S(y)\right]$$

reverse:

$$KL(\pi_S \| \pi_T) = \mathop{E}_{y \sim \pi_S}\left[\log \pi_S(y) - \log \pi_T(y)\right]$$

On-policy distillation:

1. decode a response y from the student, given a prompt

$$y \sim \pi_S(\cdot \mid x)$$

2. a powerful teacher model scores every token of the generated response

↳ scores $\Rightarrow$ conditional prob. dist of next token given prompt and partial response from student

$$\log \pi_T(y_i \mid x, y_{<i})$$

↳ all of the scores computed in parallel (prefill)

3. at each token

$$r_i = \log \pi_T(y_i \mid x, y_{<i}) - \log \pi_S(y_i \mid x, y_{<i})$$

if $r_i > 0$: teacher likes the token more than student

$r_i < 0$: student is putting more prob. on the token than the teacher

4. perform backprop / gradient on the student model

---

x: what is 15% of 80?

Student y:

15% of 80 = 0.15 × 80 = 1.2

teacher: $r_i \approx 0$

$r_i << 0$

↳ when student + teacher agree, not much to change

↳ the "bad" token gets a large

gradient update

↳ what if we only knew whether the response was correct or not?

↳ standard RL setting

↳ we don't know which token(s) were responsible for the error

↳ "credit assignment"

↳ not an issue w/ OPD because we have scores from the teacher at every token

---

Thinking Machines blog on OPD:

↳ Qwen 3-8B-Base model + SFT on 400k math problems

↳ 60% acc on held-out problems

↳ then, they did OPD using Qwen 3-8B Instruct as the teacher
  ↳ after 150 steps, reached 70%

↳ in comparison to SFT:
  ↳ SFT would need 2M examples to reach 70% acc
  ↳ OPD can yield similar perf but 10-30x cheaper