(1)

a) ۱. استفاده از $\log\sigma\left(r_\theta(x,y_w) - r_\theta(x,y_l)\right)$ باعث می‌شود به هنگام زیاد شدن اختلاف امتیاز $r_\theta(x,y_w)$ و $r_\theta(x,y_l)$ گرادیان به سمت صفر شدن میرود زیرا $\log\sigma(\cdot)$ در مقادیر زیاد گرادیانش صفر می‌شود و به همین علت از زیاد شدن اختلاف امتیاز این دو جلوگیری می‌شود.

b)

$$\text{objective}(\phi) = E_{(x,y)\sim D_{\pi_\phi^{RL}}}\left[r_\theta(x,y) - \beta\log\left(\frac{\pi_\phi^{RL}(y|x)}{\pi^{SFT}(y|x)}\right)\right] + \gamma E_{x\sim D_{pretrain}}\left[\log\left(\pi_\phi^{RL}(x)\right)\right]$$

$r_\theta(x,y) \longrightarrow$ جهت حداکثر تابع reward برای جفت (x,y) به نوعی alignment مورد نظر.

$\log\left(\frac{\pi_\phi^{RL}(y|x)}{\pi^{SFT}(y|x)}\right) \longrightarrow$ برای جلوگیری از فراموشی زبان توسط مدل و تولید توکن‌های بی‌معنی جهت حداکثر کردن $\pi^{RL}(y|x)$. به نوعی باعث می‌شود خیلی از توزیع اولیه $\pi^{SFT}(y|x)$ optimization فاصله نگیرد.

$\log\left(\pi_\phi^{RL}(x)\right) \longrightarrow$ برای جلوگیری از خراب شدن کارایی مدل در تسک های NLP که در pretraining مدل وجود دارند.

c) با توجه به اینکه $(x,y)\sim D_{\pi_\phi^{RL}}$ می‌آید و به نوعی دیتا static نیست و وابسته به π_ϕ^{RL} دارد باعث $r_\theta(x,y)$ به هنگام مشتق‌گیری برحسب ϕ صفر نشود.

d)

$$L_\theta = \mathbb{E}_{\pi_\theta}[G_t]$$

$$\nabla_\theta L_\theta = \nabla_\theta \mathbb{E}_{\pi_\theta(\tau),\tau}[G_t] = \nabla_\theta \int \pi_\theta(\tau) G_\tau \, d\tau$$

$$= \int \nabla_\theta \pi_\theta(\tau) G_\tau \, d\tau = \int \pi_\theta(\tau) \frac{1}{\pi_\theta(\tau)} \nabla_\theta \pi_\theta(\tau) G_\tau \, d\tau$$

$$= \int \pi_\theta(\tau) \nabla_\theta \log \pi_\theta(\tau) G_\tau \, d\tau$$

$$= \mathbb{E}_{\pi_\theta,\tau}[\nabla_\theta \log \pi_\theta(\tau) G_\tau]$$

$$= \mathbb{E}_{\pi_\theta,\tau}\Big[\Big(\sum_t \nabla_\theta \log \pi_\theta(a_t | s_t) G_t\Big)\Big]$$

e)

برای محاسبه KL نیاز هست که توزیع $\pi_\phi^{RL}(y|x)$, $\pi^{SFT}(y|x)$ را به ازای همه نقاط x بشناسیم اما ما این مقادیر را به ازای همه نقاط نداریم به همین علت باید با نمونه گیری این KLD را تخمین زد. (نمونه گیری $\pi^{RL}(y|x)$).

$$D_{KL}(q\|p) \approx \frac{1}{N} \sum_{i=1}^{N} \log \frac{q(x_i)}{p(x_i)} \quad , x_i \sim q(x)$$

این تخمینگر unbiased هست اما variance زیادی دارد.

$$r = \frac{p(x)}{q(x)} \Rightarrow k_1 = -\log r$$

به صورت شهودی می‌توان گفت که در این تخمین به طور متوسط نصف نمونه‌ها منفی می‌شوند در حالی که KL منفی نمی‌شود. پس variance بالایی دارد. اما تخمینگر زیر مقادیر منفی ندارد پس variance کمتری دارد.

$$D_{KL}(q\|p) \approx \frac{1}{N}\sum \frac{1}{2}\left[\log q(x_i) - \log p(x_i)\right]^2, \quad x_i \sim q(x)$$

این تخمینگر biased هست اما دارای variance کمتری است.

$$k_2 = \frac{1}{2}(\log r)^2$$

$$E_q[k_2] = E_q\left[\frac{1}{2}\log^2 r\right] \xrightarrow[f = \frac{1}{2}\log^2 x]{f\text{-divergence}} \quad f(x) = -\log x \;,\; f''(1) = 1$$

تخمینگر با بایاس کمتر :

$$D_{KL}(q\|p) \approx \frac{1}{N}\sum_{i=1}^{N} \exp\left\{\frac{1}{2}\left[\log q(x_i) - \log p(x_i)\right]^2 - 1 - \left[\log q(x_i) - \log p(x_i)\right]\right.$$

2 (a

$$\max_{\pi} E_{x \sim D, y \sim \pi_\theta(y|x)} \left[r_\varphi(x,y) \right] - \beta D_{KL} \left[\pi_\theta(y|x) \,||\, \pi_{ref}(y|x) \right]$$

Langrangian :

$$\max_{\pi} E_{x \sim D, y \sim \pi_\theta(y|x)} \left[r_\varphi(x,y) - \beta D_{KL}\left(\pi_\theta(y|x) \,||\, \pi_{ref}(y|x) \right) \right] + \lambda \left(1 - \int \pi_\theta(y|x) p(x) \, dx \, dy \right)$$

$$= \int \left(r_\varphi(x,y) - \beta \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} \right) \pi_\theta(y|x) \, p(x) \, dx \, dy + \lambda \left(1 - \int \pi_\theta(y|x) \, p(x) \, dx \, dy \right)$$

$$\frac{\partial L}{\partial \pi_\theta(x|y)} = \left(r_\varphi(x,y) - \beta \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} \right) p(x) - \beta p(x) - \lambda p(x) = 0$$

$$\Rightarrow \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} = e^{\frac{1}{\beta} r_\varphi(x,y) - \frac{\lambda + \beta}{\beta}}$$

$$\Rightarrow \pi_\theta(y|x) = e^{-\frac{\lambda+\beta}{\beta}} \pi_{ref}(y|x) \, e^{\frac{1}{\beta} r_\varphi(x,y)}$$

$$= \frac{1}{Z(x)} \pi_{ref}(y|x) \, e^{\frac{1}{\beta} r_\varphi(x,y)}$$

(b

با توجه به رابطه

$$\pi_\theta(y|x) = \frac{1}{Z(x)}\pi_{ref}(y|x)\, e^{\frac{1}{\beta} r_\phi(x,y)}$$

خواهیم داشت :

$$\log\frac{Z(x)\,\pi_\theta(y|x)}{\pi_{ref}(y|x)} = \frac{1}{\beta} r_\phi(x,y)$$

$$\Rightarrow \quad r_\phi(x,y) = \beta\log Z(x) + \beta\log\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} \qquad (1)$$

$$L_R(r_\phi, D) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log\sigma\left(r_\phi(x,y_w) - r_\phi(x,y_l)\right)\right] \qquad (2)$$

$$(1),(2) \Rightarrow L_{DPO}(\pi;\pi_{ref}) = -\mathbb{E}_{x,y_w,y_l\sim D}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right)\right]$$

(c

$$\nabla_\theta L_{DPO}(\pi_\theta;\pi_{ref}) = \nabla_\theta -\mathbb{E}_{x,y_w,y_l\sim D}\left[\log\sigma\underbrace{\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right)}_{u}\right]$$

$$= -\mathbb{E}_{x,y_l,y_l\sim D}\left[\frac{\sigma'(u)}{\sigma(u)}\,\nabla_\theta\left(\beta\log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)} - \beta\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)}\right)\right]$$

$$= -\mathbb{E}_{x,y_w,y_l\sim D}\left[\beta\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right)\left[\nabla_\theta\log\pi(y_w|x) - \nabla_\theta\log\pi(y_l|x)\right]\right]$$

Reward مربوط به مدل زبانی $\pi_\theta(y|x)$ و مدل زبانی Reference

یعنی $\pi_{ref}(y|x)$ که میتوان به صورت مستقیم بدون نیاز به

Reward function، Policy را optimize کرد.

$$\beta\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right) :$$

$\nabla_\theta \log\pi(y_w|x)$: جهت افزایش likelihood خروجی مناسب y_w

$-\nabla_\theta \log\pi(y_w|x)$: جهت کاهش likelihood خروجی نامناسب y_l