

لَمْ يَرِيْدُ حِلَّاً

٤٠٢٣٥٢١٩

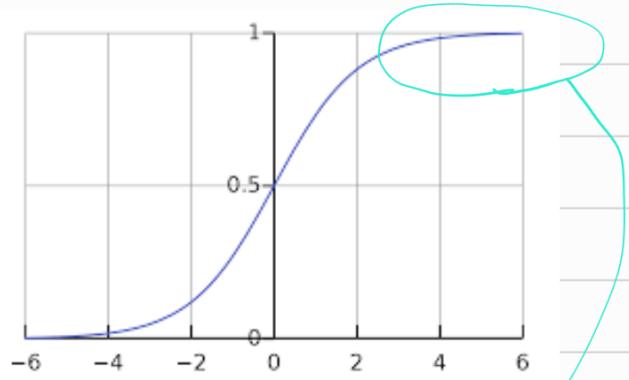
الله بِكَ

۱۱ سوال

درین تابع خوب است از تابع Sigmoid

بین امتیاز را باید

امنیتی داشت  $\approx$  Sigmoid

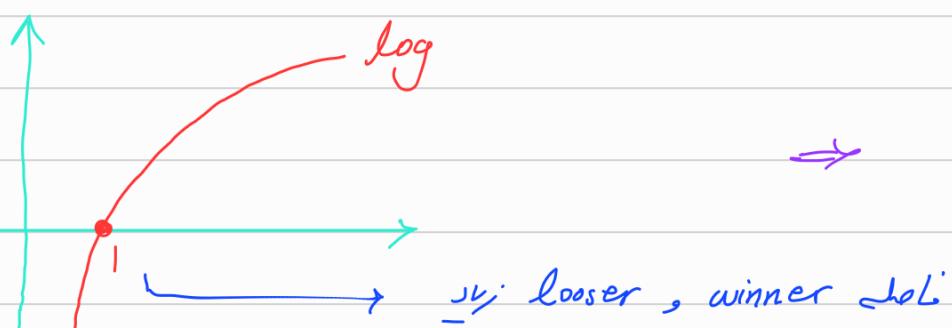


Gradient Vanishing درین توابع دهنده اند که نزدیک به صفر می‌باشند و درین نتیجه loss بـ صفر می‌رسند

از طرفه دوچرخه  $\log$  این مقدار حساب می‌شود و همچنان که این

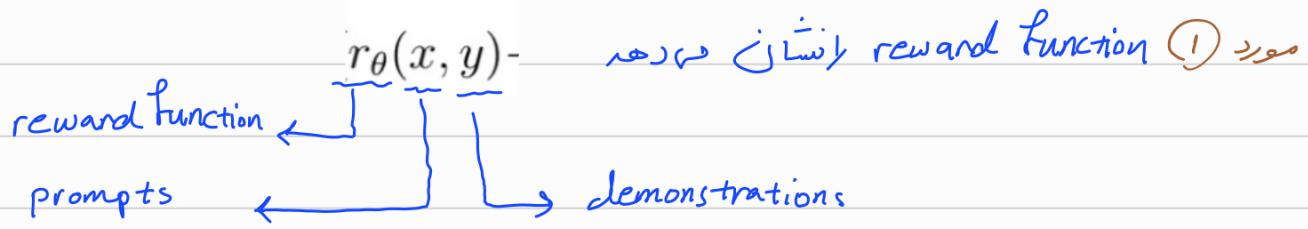
blessing است درین های این مورخوب است و باعث مهشود می‌شوند نتیجه extreme نزدیک

که در آن نتیجه reward بـ خوب می‌شود.



$$\text{objective}(\phi) = \mathbf{E}_{(x,y) \sim D_{\pi_{\phi}^{RL}}} [r_{\theta}(x, y) - \beta \log(\pi_{\phi}^{RL}(y|x)/\pi^{SFT}(y|x))] + \gamma \mathbf{E}_{x \sim D_{\text{pretrain}}} [\log(\pi_{\phi}^{RL}(x))]$$

(١)



برای آنکه می‌توانیم  $r_{\theta}(x, y)$  را بهینه کنیم باید reward func  $\rightarrow$   $\hat{r}_{\theta}(x, y)$  را تخمین زنیم (Supervised fine-tuning) (٢)

برای آنکه می‌توانیم  $\hat{r}_{\theta}(x, y)$  را تخمین زنیم باید  $r_{\theta}(x, y)$  را با  $\hat{r}_{\theta}(x, y)$  مطابقت داشت. برای اینکه  $\hat{r}_{\theta}(x, y)$  را تخمین زنیم باید  $\hat{r}_{\theta}(x, y)$  را با  $r_{\theta}(x, y)$  مطابقت داشته باشد (Supervised fine-tuning) SFT

برای آنکه  $\hat{r}_{\theta}(x, y)$  را تخمین زنیم باید  $\log \pi_{\phi}^{RL}(y|x)$  را بهینه کنیم که این کار با KL divergence می‌باشد.

این کار را Align کردن می‌گویند.

برای آنکه  $\hat{r}_{\theta}(x, y)$  را تخمین زنیم باید  $\hat{r}_{\theta}(x, y)$  را با  $r_{\theta}(x, y)$  مطابقت داشته باشد (Align) (٣)

برای آنکه  $\hat{r}_{\theta}(x, y)$  را تخمین زنیم باید  $\hat{r}_{\theta}(x, y)$  را با  $r_{\theta}(x, y)$  مطابقت داشته باشد (Align) (٤)

NLP می‌تواند  $\hat{r}_{\theta}(x, y)$  را تخمین زنیم.

آنچه در آنکه  $\hat{r}_{\theta}(x, y)$  را تخمین زنیم می‌تواند  $\hat{r}_{\theta}(x, y)$  را تخمین زنیم و  $r_{\theta}(x, y)$  را تخمین زنیم.

Dynamic reward می‌تواند  $\hat{r}_{\theta}(x, y)$  را تخمین زنیم.

(>)

$$L_\theta = E_{\pi_\theta}[G_t]$$

$$\nabla_\theta L_\theta = \nabla_\theta E_{\pi_\theta(\tau), z}[G_t] = \nabla_\theta \int \pi_\theta(z) G_z dz$$

لـ  $\nabla_\theta L_\theta$  مسقـ نـيـرـوـ لـمـسـيـهـ تـسـدـ مـوـكـلـ حـجـمـ دـوـ وـ لـ عـوـضـنـ كـرـدـ

وـ جـعـنـ اـنـ تـحـعـ بـعـدـ سـبـبـ تـمـ نـاسـنـ دـادـهـ شـعـورـ سـوـيـنـيـ وـ مـسـقـ نـيـرـيـ رـاـدـارـ. سـاـمـانـ دـارـمـ :

$$\int \nabla_\theta \pi_\theta(z) G_z dz$$

اـنـ اـسـكـلـ حـيـرـنـ دـرـخـنـيـ تـبـعـ مـوـسـودـ قـاـلـ مـاـمـبـيـتـ.

اـنـ اـسـكـلـ حـيـرـنـ دـرـخـنـيـ تـبـعـ مـوـسـودـ قـاـلـ مـاـمـبـيـتـ.

$$= \int \pi_\theta(z) \frac{1}{\pi_\theta(z)} \nabla_\theta \pi_\theta(z) G_z dz$$

$$= \int \pi_\theta(z) \nabla_\theta \log \pi_\theta(z) G_z dz = E_{\pi_\theta, z} [\nabla_\theta \log \pi_\theta(z) G_z]$$

اـنـ اـسـكـلـ حـيـرـنـ دـرـخـنـيـ تـبـعـ مـوـسـودـ قـاـلـ مـاـمـبـيـتـ.

$$= E_{\pi_\theta, z} \left[ \left( \sum_t \nabla_\theta \log \pi_\theta(a_t | s_t) G_t \right) \right]$$

$$\bullet D(q \parallel p) = \sum_x q(x) \log q(x)/p(x).$$

باوچاره به آنکه محاسبه KL سخت است

برای محاسبه دقتی از نیاز است توزیع دقیق  $p$  و  $q$  را در هم مقایری خواهند داشت باشیم و این سخت است. و فرم نسبت می‌شود محاسبه ندارد. محاسبات و حافظه بالا نیاز است. و فرم نسبت ندارد.

لذا محاسبه از توزیع  $q$  sample از توزیع  $p$  سخت است.

$$\bullet D(q \parallel p) \approx \frac{1}{N} \sum_{i=1}^N \log q(x_i)/p(x_i), \quad x_i \sim q(x)$$

KL divergence  $\approx$  نمونه از  $p$  و براساس قابل اعتماد است این همچنان Sample نیست  $\leftarrow$

این تخمین  $\leftarrow$  نام واریانس نباید کرد. زیرا اگر هر بار نامسالم

محاذیت برآورده واریانس را زیاد می‌کند. از  $q$  نمونه می‌شوند.

این تخمین  $\leftarrow$  نام واریانس  $\approx$  در در خاطر نیست لطفاً مفهوم دخالت است  $\leftarrow$  دخالت

نمی‌دانیم. تعریف می‌شود

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

تخمین  $r$  در  $\leftarrow$  واریانس  $\approx$  در این نام واریانس است برای اینست  $\leftarrow$

$$k_r = \frac{1}{r} \left( \log \frac{p(x)}{q(x)} \right)^2 = \frac{1}{r} (\log r)^2$$

خوبی در  $\leftarrow$  محاسبه کن می‌خواهد  $k_r$  و  $k_1$  را مقایسه کنید. و ممکن است  $k_r$  باشد

$\rightarrow$   $k_1$   $\leftarrow$   $k_r$  Emperically True?

$f$ -divergence  $\rightarrow$  expectation expectation  $\rightarrow$   $K_f$  ( $\rightarrow$ )

$$D_f(p, q) = \mathbb{E}_{q(x)} \left[ f\left(\frac{p(x)}{q(x)}\right)\right] : \text{مقدار طبعي تابع } f \text{ يعطى}$$

KL divergence  $f$  في ترتيب  $p, q$  هو .  $\rightarrow$   $f$ -divergence  $\rightarrow$  KL divergence

: خاصية دالة

$$D_f(p_0, p_\theta) = \frac{f''(1)}{2} \theta^T F \theta + O(\theta^3)$$

F: Fisher information matrix for  $p_\theta = p_0$

$$\mathbb{E}_q[k_2] = \mathbb{E}_q \left[ \frac{1}{r} (\log r)^2 \right] \rightarrow f(x) = \frac{1}{r} (\log x)^2$$

.  $\rightarrow$   $f$ -divergence

Control Variate  $\leftarrow$  مقدار معيون

$\checkmark$   $K_1$  بمعنى متغير دليل على صغرها  $\rightarrow$  افتراضه صبور

$$\frac{P(x)}{q(x)} - 1 = r - 1 \quad : \text{دالمة}$$

unbiased estimator  $\leftarrow -\log r + \lambda(r-1) \rightarrow$   $\lambda$  يعطى

$\log(x) \leq x - 1$  :  $\rightarrow$   $\lambda$  موافق لـ  $r$  .  $\rightarrow$   $\lambda = 1$

$K_F = (r-1) - \log r$  :  $\rightarrow$   $\lambda = 1$  يعطى  $\lambda$  بمعنى

Bregman divergence

$$\text{نیز} \quad E_q[r] = 1 \quad \text{convex} \quad f \text{ موج}$$

$$F(r) - f'(1)(r-1)$$

:  $f'(x) = 1$  ✓  $F(x) = x \log x$  موج ✓ . Convex Minima ✓

$$[r \log r - (r-1)]$$

## سؤال دوم

در این سوال سعی در رسیدن به تابع هدف الگوریتم DPO داریم.

(آ) نشان دهید جواب تابع زیر

$$\max_{\pi_\theta} \mathbb{E}_{x \sim D, y \sim \pi_\theta(y|x)} [r_\phi(x, y)] - \beta \mathbb{D}_{KL} \left[ \frac{\pi_\theta(y|x)}{\text{Policy}} \parallel \frac{\pi_{ref}(y|x)}{\text{reference}} \right]$$

برابر با

$$\pi_r(y|x) = \frac{1}{Z(x)} \pi_{ref}(y|x) \exp\left(\frac{1}{\beta} r(x, y)\right)$$

است.

: متریک Lagrangian معادله آن //

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

$\rightarrow$   $\lambda \left( L - \int \pi_\theta(y|x) p(x) dy dx \right)$

برای هر مقدار  $\lambda$   $\pi_\theta(y|x)$  متناسب با  $\pi_{ref}(y|x)$  باشد

$$\frac{\partial L}{\partial \pi_\theta(y|x)} = \left( r_\phi(x, y) - \beta \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} \right) P(x) - \beta P(x) - 1 = 0$$

$$\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} = \exp \left\{ \frac{1}{\beta} r_\phi(x, y) \right\} \cdot \exp \left\{ - \frac{1+\beta}{\beta} \right\}$$

$$\pi_\theta(y|x) = \frac{\exp \left\{ - \frac{1+\beta}{\beta} \right\} \pi_{ref}(y|x) \exp \left\{ \frac{1}{\beta} r_\phi(x, y) \right\}}{Z(x)}$$

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

rewards

reward objective:

$$L_R(r_\theta, D) = -E_{(x, y_w, y_l) \sim D} \left[ \log \sigma (r_\theta(x, y_w) - r_\theta(x, y_l)) \right]$$

$$L_{DPO}(\pi_\theta, \pi_{\text{ref}}) = -E_{(x, y_w, y_l) \sim D} \left[ \log \sigma \left( \beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right]$$

$$\nabla_\theta L_{DPO}(\pi_\theta, \pi_{\text{ref}}) =$$

$$-\beta E_{(x, y_w, y_l) \sim D} \left[ G \left( \hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w) \right) \left[ \nabla_\theta \log \pi(y_w|x) - \nabla_\theta \log \pi(y_l|x) \right] \right]$$

↑  
↑  
↑

① increase likelihood of  $y_w$

② decrease likelihood of  $y_l$