Different Designs For LLM KD Loss

Alephia 25/11/20

BACKGROUND

模型q_θ依据前缀w^{t-1}生成文本的时候，loss可以表示为

$$l(q_\theta, w^{t-1}; o) = \mathop{\mathbb{E}}_{w_t \sim o(\cdot|w^{t-1})} \log \frac{o(w_t|w^{t-1})}{q_\theta(w_t|w^{t-1})}$$

由分布o采样下一个token w_t，再进行KLD对齐。

可以展开得到

$$L(q_\theta; o) \approx \sum_{t=1}^{T} \mathop{\mathbb{E}}_{w^{t-1} \sim d_o^{t-1},\, w_t \sim o(\cdot|w^{t-1})} \log \frac{o(w_t|w^{t-1})}{q_\theta(w_t|w^{t-1})} = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{w^{t-1} \sim d_o^{t-1}} KL(o||q_\theta)$$

BACKGROUND

$$KL(p(X), q_\theta(X)) = E_{x\sim p(X)}\left[\log \frac{p(x)}{q_\theta(x)}\right] = E_{x\sim p(X)}[-\log q_\theta(x)] - H(p(x))$$

$$\begin{aligned} argmin_\theta\ KL(p(X), q_\theta(X)) &= argmin_\theta\ E_{x\sim p(X)}\left[-\log q_\theta(x)\right] \\ &= argmax_\theta\ E_{x\sim p(X)}\left[\log q_\theta(x)\right] \\ &\approx argmax_\theta \sum_x \log q_\theta(x) \\ &= argmax_\theta \prod_x q_\theta(x) \end{aligned}$$

最小化$KL(p, q_\theta)$等价于最小化$CE(p, q_\theta)$等价于最大化似然函数

BACKGROUND

$$
\begin{aligned}
RKL(p(X), q_\theta(X)) &= KL(q_\theta(X), p(X)) \\
&= E_{x\sim q_\theta(X)}\left[\log \frac{q_\theta(x)}{p(x)}\right] \\
&= E_{x\sim q_\theta(X)}\left[-\log p(x)\right] - H(q_\theta(x))
\end{aligned}
$$

最小化RKLD(p,q)等价于最小化CE(q,p)-H(q)

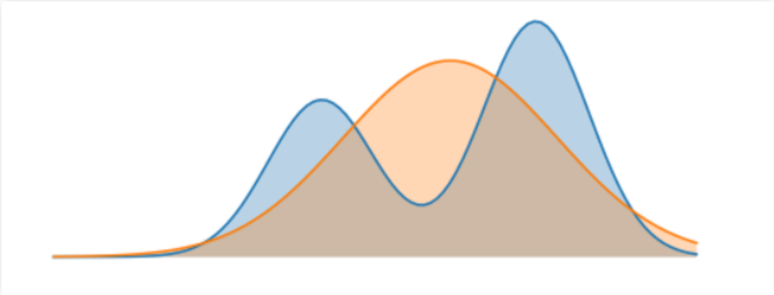
FKLD: MEAN-SEEKING BEHAVIOUR

$$KL(p(X), q_\theta(X)) = E_{x \sim p(X)}[-\log q_\theta(x)] - H(p(x))$$

Zero Avoiding

$$\exists (x, y)\ s.t.\ p(y|x) \gg 0, q_\theta(y|x) \approx 0 \rightarrow KL(p, q_\theta) = inf$$

- p中高概率的地方，q也必须高，需要涵盖所有高概率区域
- q中高概率的地方，p不必高
- FKLD倾向于拟合多个峰

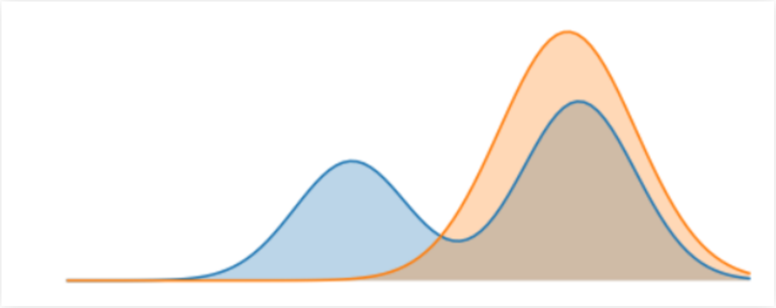


RKLD: MODE-SEEKING BEHAVIOUR

$$RKL(p(X), q_\theta(X)) = E_{x\sim q_\theta(X)}\left[-\log p(x)\right] - H(q_\theta(x))$$

$$\exists(x,y)\ s.t.\ q_\theta(y|x) \gg 0, p(y|x) \approx 0 \rightarrow KL(q_\theta, p) = inf$$

- q中高概率的地方，p也必须高， q中低概率的地方，p也应该较小
- p中高概率的地方，q不必高
- RKLD倾向于拟合一个峰



RKLD IN LLM KD

KLD下，学生在教师分布的void region会高估，进而带来麻烦。这一问题在RKLD下有所缓解

条件：

1. 教师服从混合Gaussion分布，学生服从Gaussion分布
2. 两个分布都是连续的

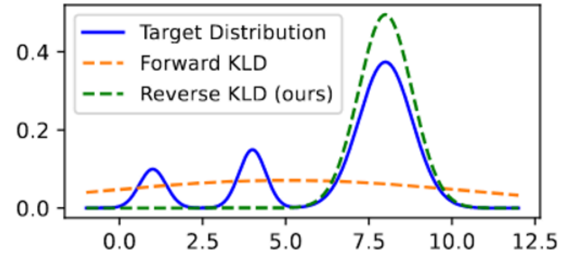


Figure 2: We fit a Gaussian mixture distribution with a single Gaussian distribution using *forward* KLD and *reverse* KLD.

Gu, Y., Dong, L., Wei, MiniLLM: Knowledge Distillation of Large Language Models. In ICLR,24

DOES RKLD REALLY HELPS IN LLM KD?

1. 教师，学生输出经过softmax之后不一定满足Gaussion分布
2. logits分布是离散的

事实上非Gauission+离散情况下，充分训练后，两种loss训练下都会得到同一个拟合结果

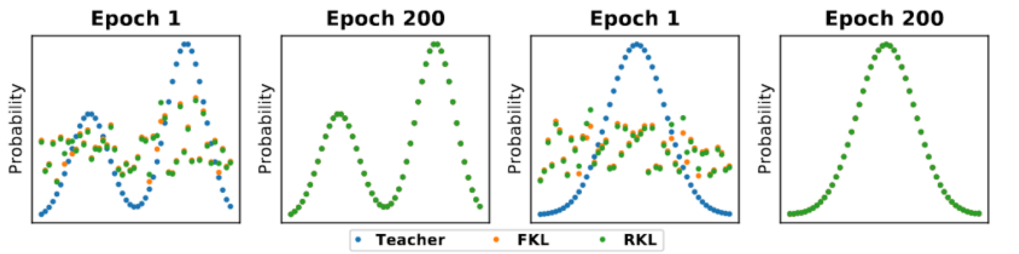
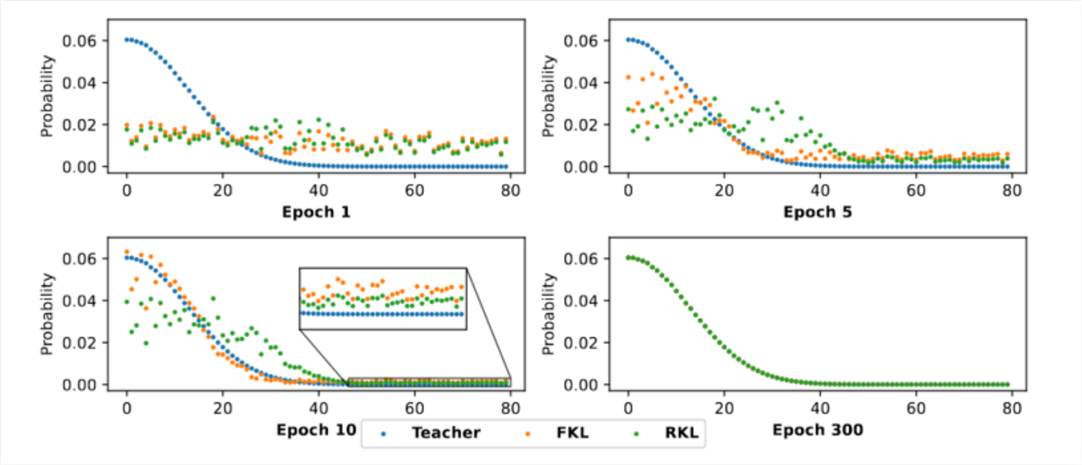


Figure 2: The convergence of FKL and RKL on toy data under epoch 1 and epoch 200. The initial distribution q is the same for FKL and RKL. After 200 epochs, both FKL and RKL can converge to the target distribution well regardless of the shape of p.

Wu, T., Tao, Rethinking Kullback-Leibler Divergence in Knowledge Distillation for Large Language Models. In COLING,25

COMBINE RKLD WITH FKLD



LLM KD中，所谓mean-seeking和mode-seeking可能并不存在，取而代之的是：FKLD倾向于先拟合分布头部，RKLD倾向于先拟合分布尾部

最终solution：$AKL(p, q_\theta) = \alpha_1 FKL(p, q_\theta) + \alpha_2 RKL(p, q_\theta)$

A BETTER FORMAT

$$D_{AB}^{(\alpha,\beta)}(p,q) = -\frac{1}{\alpha\beta}\sum_k \left[p(k)^a q(k)^\beta - \frac{\alpha}{\alpha+\beta}p(k)^{\alpha+\beta} - \frac{\beta}{\alpha+\beta}q(k)^{\alpha+\beta}\right]$$

α controls `Hardness Concentration`, while β controls `Confidence-Concentration`

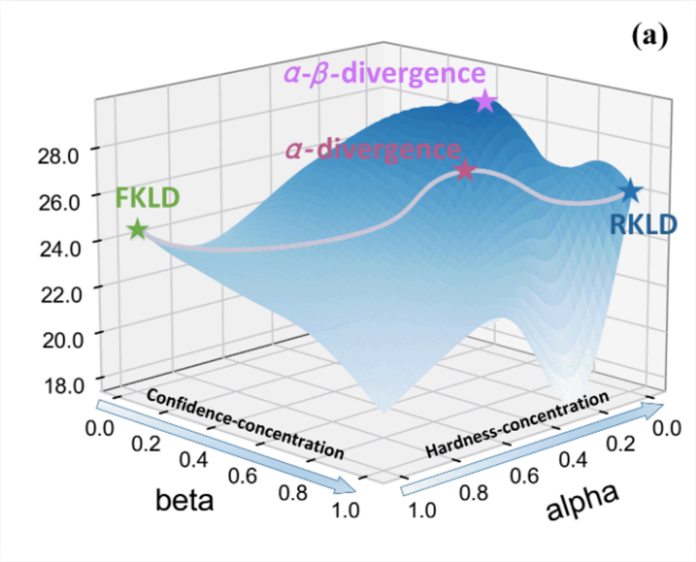
FKLD 表现出弱`Hardness Concentration` 和弱`Confidence-Concentration` , $\alpha = 1, \beta = 0$

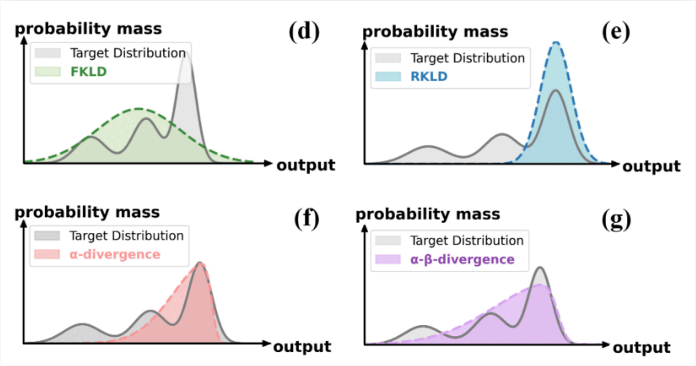
RKLD 表现出强`Hardness Concentration` 和强`Confidence-Concentration`, $\alpha = 0, \beta = 1$

Wang, G., Yang, Z., Wang, Z., Wang, S., Xu, Q., & Huang, Q. (2025). ABKD: Pursuing a Proper Allocation of the Probability Mass in Knowledge Distillation via α-β-Divergence. In ICML, 25

A BETTER FORMAT

$$D_\alpha(p,q) = -\frac{1}{\alpha(1-\alpha)} \sum_k \left[p(k)^a q(k)^{1-\alpha} - 1 \right]$$

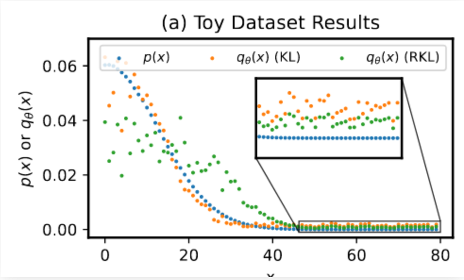




APPLY ASSISTANT DISTRIBUTION

$$SKL^{\alpha}(p, q_\theta) = KL(p, \alpha p + (1-\alpha) q_\theta)$$

$$SRKL^{\alpha}(p, q_\theta) = RKL(q_\theta, \alpha q_\theta + (1-\alpha) p)$$



pulling-up effect for SKL

pulling-down effect for SRKL

$$L_{CALD} = \frac{1}{2|D|} \sum_{(x, y_t, y_s) \sim D} SKL^{\alpha}(x, y_t) + SRKL^{\alpha}(x, y_s)$$

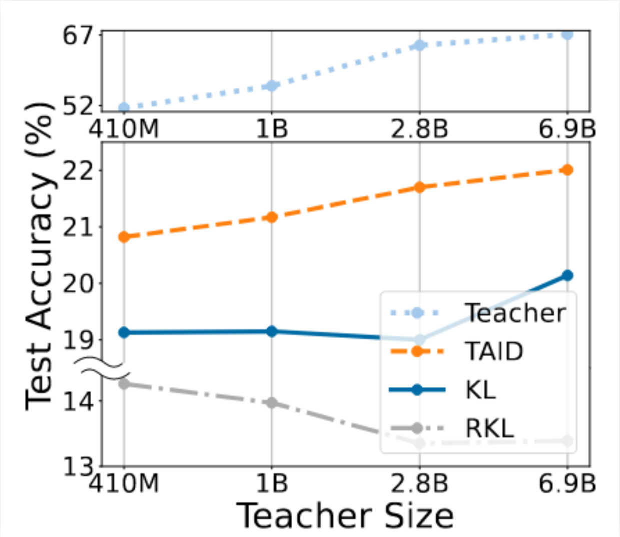
Ko, J., Kim, S., Chen, T., & Yun, S. (2024). DistiLLM: Towards Streamlined Distillation for Large Language Models. In ICML, 24

Ko, J., Chen, T., Kim, S., Ding, T., Liang, L., Zharkov, I., & Yun, S. (2025). DistiLLM-2: A Contrastive Approach Boosts the Distillation of LLMs. In ICML, 25

APPLY ASSISTANT DISTRIBUTION

$$r = softmax((1-\alpha)\cdot \mathrm{logit}(q_\theta(y_s|y_{<s})) + \alpha \cdot \mathrm{logit}p((y_s|y_{<s})))$$

$$r = p^\alpha q^{1-\alpha}$$

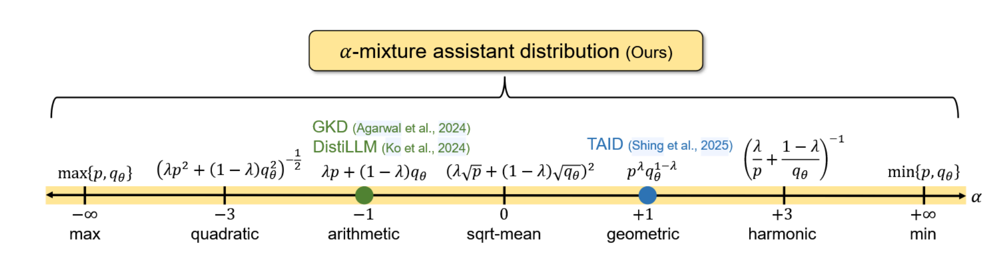


TAID: Temporally Adaptive Interpolated Distillation for Efficient Knowledge Transfer in Language Models. (2025). In ICLR, 25

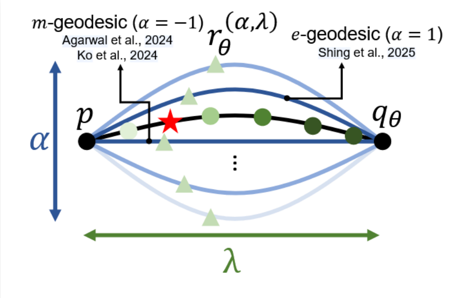
APPLY ASSISTANT DISTRIBUTION

$$\tilde{r}^{(\alpha,\lambda)}(z) = \begin{cases} \left(\lambda p(z)^{\frac{1-\alpha}{2}} + (1-\lambda) q(z)^{\frac{1-\alpha}{2}}\right)^{\frac{2}{1-\alpha}} & \text{if } \alpha \neq 1 \\ p(z)^{\lambda} q(z)^{1-\lambda} & \text{if } \alpha = 1 \end{cases}$$

$$r^{(\alpha,\lambda)}(z) = \frac{\tilde{r}^{(\alpha,\lambda)}(z)}{Z_r}, \ Z_r = \int \tilde{r}^{(\alpha,\lambda)}(x)dx$$



(a) Illustration of α-mixture assistant distributions $r_\theta^{(\alpha,\lambda)}$ with varying α.



(a) Role of α and λ in the distribution space.

Shin, D., Kim, Y., Jo, S., Na, B., & Moon, I. (2025). AMiD: Knowledge Distillation for LLMs with α-mixture Assistant Distribution.

APPLY ASSISTANT DISTRIBUTION

Proposition 3.5. *(Gradient analysis) The gradient of f-divergence $D_f(p||r_\theta^{(\alpha,\lambda)})$ be expressed as:*

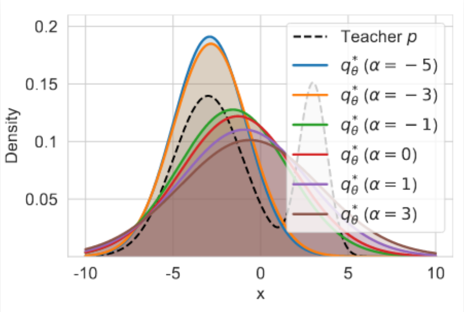
$$\nabla_\theta D_f\left(p||r_\theta^{(\alpha,\lambda)}\right) = \mathbb{E}_{r_\theta^{(\alpha,\lambda)}}\left[w \cdot \left\{\psi_f\left(\frac{p}{r_\theta^{(\alpha,\lambda)}}\right) - \mathbb{E}_{r_\theta^{(\alpha,\lambda)}}\left[\psi_f\left(\frac{p}{r_\theta^{(\alpha,\lambda)}}\right)\right]\right\} \cdot \nabla_\theta \log q_\theta\right] \quad (11)$$

where $w := \frac{(1-\lambda)q_\theta^{\frac{1-\alpha}{2}}}{\lambda p^{\frac{1-\alpha}{2}} + (1-\lambda)q_\theta^{\frac{1-\alpha}{2}}}$ *and* $\psi_f(v) := f(v) - vf'(v)$.

w controls the magnitude of gradient

when $p > q$, a larger α exhibits a mode-seeking behavior

when $p < q$, a smaller α exhibits a mean-seeking behavior



(c) Role of α for q_θ^*.

CONCLUSION

🤔 **Different choices of divergence**

KLD, RKLD, AKD, ABKD...

Focusing on **mode-seeking** ,**mean-seeking** behaviors and conbination of different divergence

🤔 **Apply assistant distribution**

Mitigate the teacher-student gap: AMiD, TAID, DistiLLM

THANKS!