Different Designs For LLM KD Loss

- KLD
- Reversed KLD(RKLD)
- JSD
- Wasserstein Distance

KLD

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

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

最小化KLD(p,q)等价于最小化CE(p,q)等价于最大化似然函数

RKLD

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

最小化RKLD(p,q)等价于最小化CE(q,p)-H(q)

FKLD: Mean-Seeking Behaviour

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

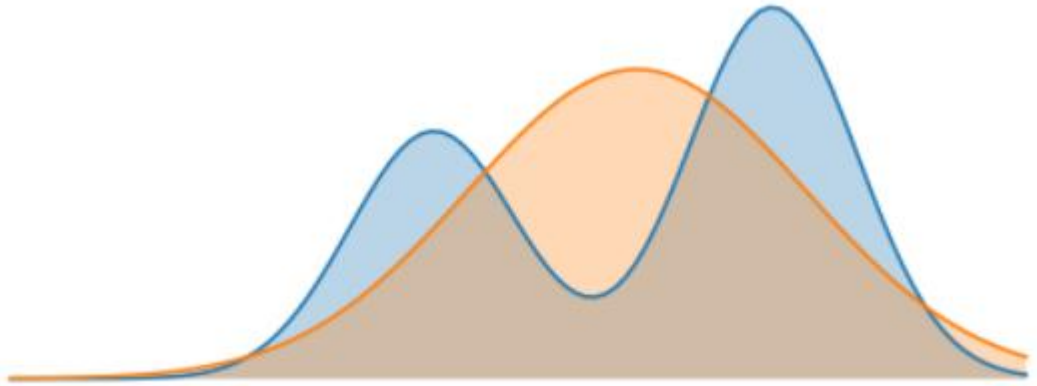
Zero avoiding

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

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

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

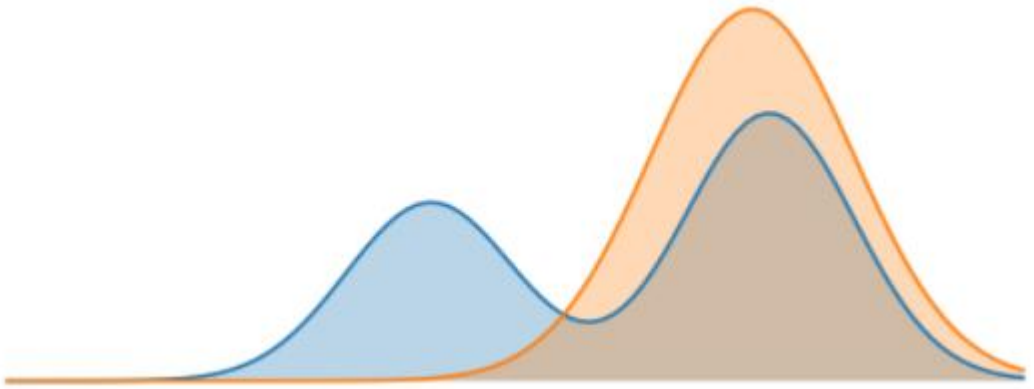
$$q_\theta(y|x) \gg 0, p(y|x) \approx 0 \rightarrow KL(q_\theta, p) = inf$$

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

entropy(q)防止其退化到极小区域上

RKLD倾向于拟合一个峰



MiniLLM

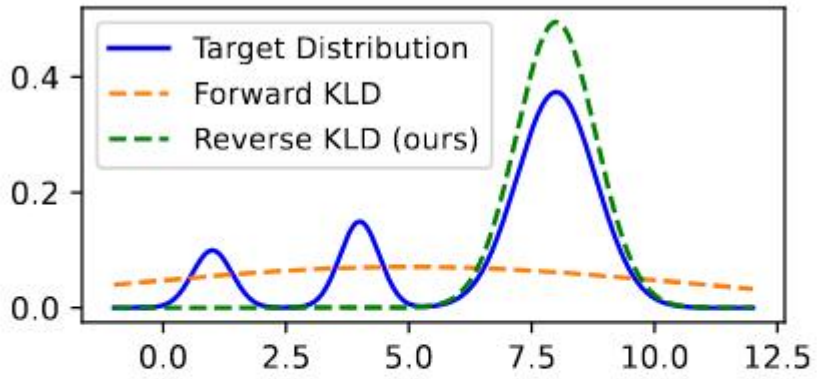


Figure 2: We fit a Gaussian mixture distribution with a single Gaussian distribution using *forward* KLD and *reverse* KLD.

KLD下，学生在教师分布的viod 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

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

Rethinking Kullback-Leibler Divergence in Knowledge Distillation for Large Language Models

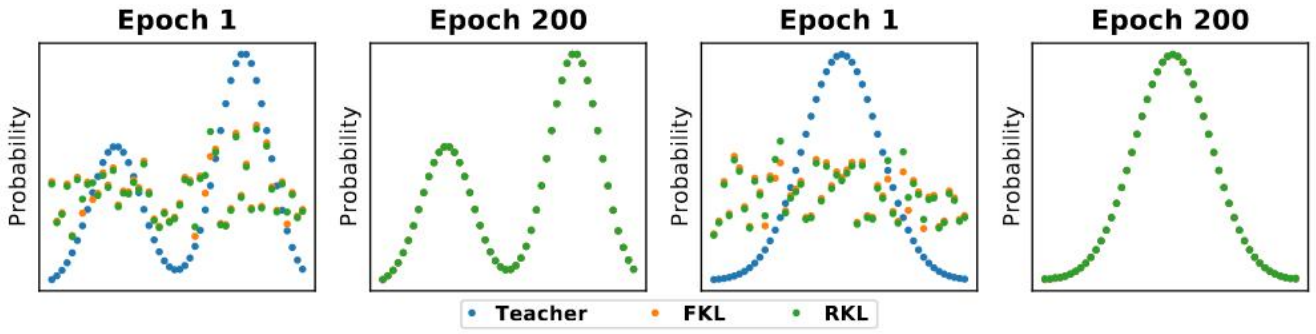


Figure 2: The convergence of FKL and RKL on toy data under epoch 1 and epoch 200. The initial distribution q is the same for FKL and RKL. After 200 epochs, both FKL and RKL can converge to the target distribution well regardless of the shape of p.

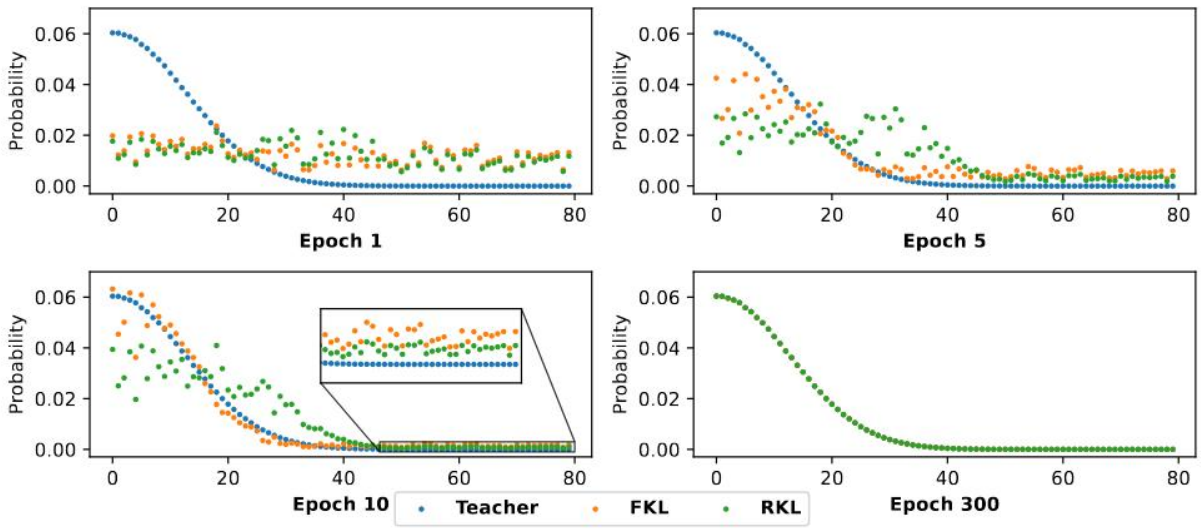
1 教师，学生输出经过softmax之后不一定满足Gaussion分布
2 logits分布是离散的

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

但是实际不允许蒸馏这么多轮，在训练前期KLD和RKLD依然存在区别

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

Rethinking Kullback-Leibler Divergence in Knowledge Distillation for Large Language Models



In the KD for LLMs, the mean-seeking and mode-seeking behaviors do not hold for forward KL (FKL) and reverse KL (RKL), respectively. Instead, they share the same optimization objective. Meanwhile, FKL focuses on the head part and RKL focuses on the tail part at the beginning.

与之前的结论并不冲突，之前的结论假设更强

Reason

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

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

Solution

$$AKL(p, q_\theta) = \frac{g_{head}}{g_{head}+g_{tail}} FKL(p, q_\theta) + \frac{g_{tail}}{g_{head}+g_{tail}} RKL(p, q_\theta).$$

FKL VS RKL

AKL

Method	GPT2 1.5B → GPT2 120M			LLaMA 6.7B → TinyLLaMA 1.1B		
	Dolly (500)	S-NI (1694)	UnNI (10000)	Dolly (500)	S-NI (1694)	UnNI (10000)
Teacher	26.98±0.27	27.25±0.16	31.61±0.16	26.73±0.65	32.75±0.24	34.61±0.08
SFT	23.01±0.18	16.48±0.28	18.43±0.09	22.05±0.38	27.79±0.20	25.96±0.13
SeqKD	23.30±0.38	16.35±0.17	18.51±0.04	22.67±0.57	26.97±0.32	27.35±0.06
FKL	23.46±0.56	16.63±0.48	19.27±0.06	22.24±0.38	28.07±0.41	26.93±0.09
RKL	22.62±0.34	17.88±0.17	19.35±0.06	23.95±0.53	28.90±0.41	27.89±0.08
SKL	23.47±0.49	16.51±0.29	18.46±0.08	23.29±0.54	29.89±0.18	29.15±0.20
SRKL	23.25±0.22	17.54±0.30	19.31±0.10	22.09±0.22	29.60±0.38	28.81±0.16
FKL+RKL	23.36±0.53	17.83±0.23	20.37±0.08	24.08±0.51	30.98±0.31	30.48±0.10
AKL (Ours)	**23.88±0.46**	**19.15±0.21**	**21.97±0.13**	**24.40±0.42**	**31.37±0.23**	**31.05±0.17**

MiniLLM没有FKL与RKL的对比实验

LLM KD场景下，SKL，RKL各有优劣，RKL并不占据绝对优势，但是组合后效果较好

DistillLLM

Loss Function	Dolly Eval (↑)	Self-Instruct (↑)	Vicuna Eval (↑)	Super-Natural (↑)	Unnatural (↑)
KLD	23.52 (0.22)	11.23 (0.46)	15.92 (0.41)	20.68 (0.16)	23.38 (0.13)
RKLD	23.82 (0.34)	10.90 (0.58)	16.11 (0.46)	22.47 (0.21)	23.03 (0.11)
Generalized JSD	24.34 (0.35)	12.01 (0.54)	15.21 (0.61)	25.08 (0.36)	27.54 (0.07)
SKL	24.80 (0.12) ●	12.86 (0.34) ●	**16.20 (0.57)** ●	**26.26 (0.41)** ●	28.06 (0.08) ●
SRKL	**25.21 (0.27)** ●	**12.98 (0.24)** ●	15.77 (0.39) ●	25.83 (0.15) ●	**28.62 (0.10)** ●

梯度角度的分析及改进

$$\nabla_\theta D_{KL}(p, q_\theta) = \mathbb{E}_{p(x,y)}\left[\nabla_\theta \log \frac{p(y|x)}{q_\theta(y|x)}\right] = -\mathbb{E}_{p(x,y)}\left[q_\theta(y|x)^{-1}\nabla_\theta q_\theta(y|x)\right] \approx -q_\theta(y^*|x^*)^{-1}\nabla_\theta q_\theta(y^*|x^*)$$

梯度的系数与条件概率成反比 → 受低质量token的影响较大，当条件概率接近0，影响尤其大

total variation distance (TVD)

$$\nabla_\theta D_{KL}(p, q_\theta) = \mathbb{E}_{p(x,y)}\left[\nabla_\theta \log \frac{p(y|x)}{q_\theta(y|x)}\right] = -\mathbb{E}_{p(x,y)}\left[q_\theta(y|x)^{-1}\nabla_\theta q_\theta(y|x)\right] \approx -q_\theta(y^*|x^*)^{-1}\nabla_\theta q_\theta(y^*|x^*)$$

梯度的系数与条件概率成反比 → 受低质量token的影响较大，当条件概率接近0，影响尤其大

$$D_{TV}(p, q_\theta) = 1 - \sum_{y\in\gamma} min(p(y|x), q_\theta(y|x)) = \frac{1}{2}\sum_{y\in\gamma}|p(y|x) - q_\theta(y|x)|$$

$$\nabla_\theta D_{TV}(p, q_\theta) \approx \begin{cases} -p(y^*|x^*)^{-1}\nabla_\theta q_\theta(y^*|x^*) & q_\theta(y^*|x^*) < p(y^*|x^*) \\ 0 & q_\theta(y^*|x^*) \geq p(y^*|x^*) \end{cases}$$

更新幅度比KLD更保守，更新相对稳健

高估时不再更新

Ji, H., Ke, P., Tailoring Language Generation Models under Total Variation Distance. In ICLR,23

total variation distance (TVD)

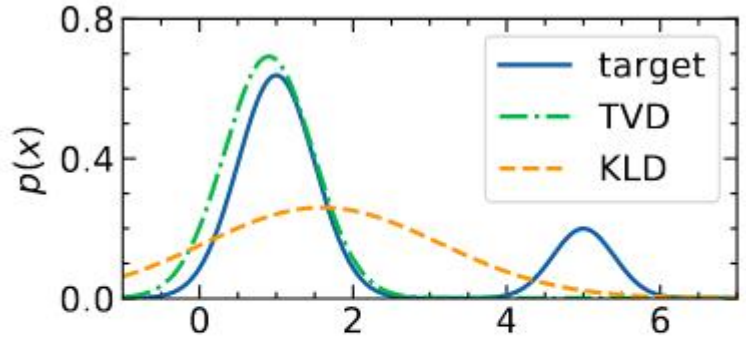


Figure 1: Results of the toy experiment: KLD is sensitive to outliers while TVD is more robust.

Skew KLD(SKLD)

$$\nabla_\theta D_{KL}(p, q_\theta) = \mathbb{E}_{p(x,y)}\left[\nabla_\theta \log \frac{p(y|x)}{q_\theta(y|x)}\right] = -\mathbb{E}_{p(x,y)}\left[q_\theta(y|x)^{-1}\nabla_\theta q_\theta(y|x)\right] \approx -q_\theta(y^*|x^*)^{-1}\nabla_\theta q_\theta(y^*|x^*)$$

梯度的系数与条件概率成反比 → 受低质量token的影响较大，当条件概率接近0，影响尤其大

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

$$\nabla_\theta D_{SKL}^{(\alpha)}(p, q_\theta) = -(1-\alpha)\tilde{q}_\theta(y^*|x^*)^{-1}\nabla_\theta q_\theta(y^*|x^*)$$
$$\tilde{q}_\theta(y^*|x^*) = \alpha p(y^*|x^*) + (1-\alpha) q_\theta(y^*|x^*)$$

系数变成加权的条件概率的逆，相对KLD更小，更新更加稳健

Ko, J., Kim, S., DistiLLM: Towards Streamlined Distillation for Large Language Models. In ICML,24

Skew KLD(SKLD)

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

$$\begin{aligned}\nabla_\theta D_{KL}(q_\theta, p) &= \nabla_\theta \mathbb{E}_{q_\theta(x,y)}\left[\log \frac{q_\theta(y|x)}{p(y|x)}\right] \\ &\approx \frac{1}{p(y^*|x^*)}(\log \frac{q_\theta(y^*|x^*)}{p(y^*|x^*)} - 1)\nabla_\theta q_\theta(y^*|x^*)\end{aligned}$$

$$\begin{aligned}\nabla_\theta D_{SKL}^{\alpha}(q_\theta, p) &= \nabla_\theta \mathbb{E}_{q_\theta(x,y)}\left[\log \frac{q_\theta(y|x)}{\alpha q_\theta(y|x) + (1-\alpha)p(y|x)}\right] \\ &\approx \frac{1}{p(y^*|x^*)}\left(\log \frac{q_\theta(y^*|x^*)}{\tilde{p}(y^*|x^*)} + 1 - \frac{\alpha}{\tilde{p}(y^*|x^*)}\right)\nabla_\theta q_\theta(y^*|x^*)\end{aligned}$$

相对TVD，其可以在KLD和RKLD上迁移

Skew KLD(SKLD)

Loss Function	Dolly Eval (↑)	Self-Instruct (↑)	Vicuna Eval (↑)	Super-Natural (↑)	Unnatural (↑)
KLD	23.52 (0.22)	11.23 (0.46)	15.92 (0.41)	20.68 (0.16)	23.38 (0.13)
RKLD	23.82 (0.34)	10.90 (0.58)	16.11 (0.46)	22.47 (0.21)	23.03 (0.11)
Generalized JSD	24.34 (0.35)	12.01 (0.54)	15.21 (0.61)	25.08 (0.36)	27.54 (0.07)
SKL	24.80 (0.12) ●	12.86 (0.34) ●	**16.20 (0.57)** ●	**26.26 (0.41)** ●	28.06 (0.08) ●
SRKL	**25.21 (0.27)** ●	**12.98 (0.24)** ●	15.77 (0.39) ●	25.83 (0.15) ●	**28.62 (0.10)** ●

1 SKL，SRKL相对KLD，RKLD改进明显
2 SKL，SRKL与之前的结论类似，互相没有明显优势

Jensen-Shannon divergence

$$D_{JS}(p, q_\theta) = \frac{1}{2} D_{KL}\left(p, \frac{p+q_\theta}{2}\right) + \frac{1}{2} D_{KL}\left(q_\theta, \frac{p+q_\theta}{2}\right)$$

相对KLD来说是对称的

$$p(x)q_\theta(x) = 0, p(x) + q_\theta(x) \neq 0 \rightarrow JS(p, q_\theta) = log_2$$ 此时两个分布不重叠

$$p(x)q_\theta(x) \neq 0, p(x) + q_\theta(x) \neq 0 \rightarrow 0 < JS(p, q_\theta) < log_2$$ 此时两个分布存在重叠部分

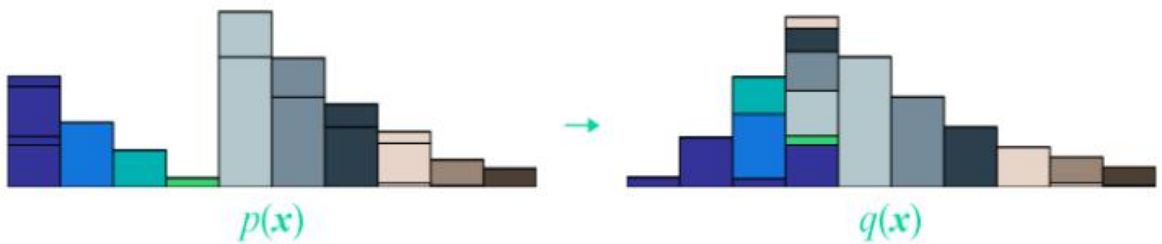
不重叠时梯度消失，无法继续优化，也不再能度量两个分布的相似程度

Wasserstein Distance

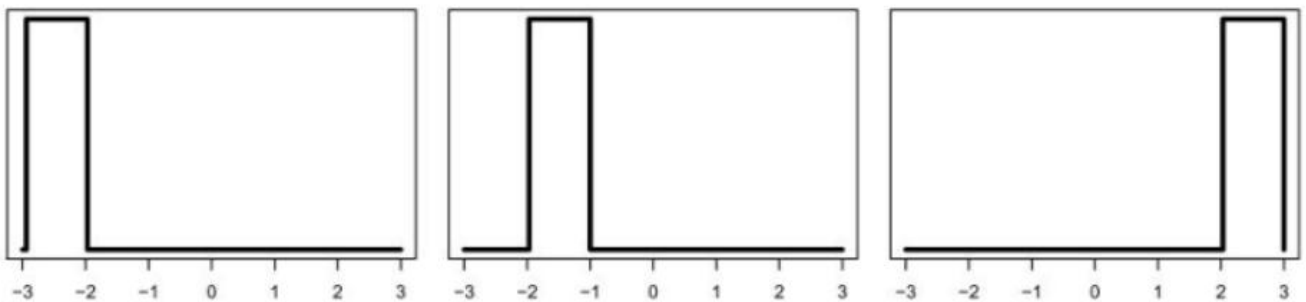
$$W^p(P,Q) = \left(\inf_{\gamma \in \pi(P,Q)} \int \int \gamma(x,y) c(x,y) dx dy \right)^{1/p}$$

$$c(x,y) = ||x-y||^p$$

$$\int \gamma(x,y) dy = p(x) \ , \quad \int \gamma(x,y) dx = q(y)$$



相比KLD，JSD，更加关注整体特征



Wasserstein Distance

线性规划的形式

$$W(P,Q) = min_{\Gamma}\{< \Gamma, C > | A\Gamma = b, \Gamma \geq 0\}$$

$$\Gamma = (\gamma(x_1,y_1), \gamma(x_1,y_2), \dots | \gamma(x_2,y_1), \gamma(x_2,y_2), \dots | \dots | \gamma(x_n,y_1), \gamma(x_n,y_2), \dots)^T$$
$$C = (c(x_1,y_1), c(x_1,y_2), \dots | c(x_2,y_1), c(x_2,y_2), \dots | \dots | c(x_n,y_1), c(x_n,y_2), \dots)^T$$
$$b = (p(x_1), p(x_2), \dots p(x_n) | q(y_1), q(y_2), \dots q(y_n))^T$$

Sinkhorn Algorithm (Sinkhorn distance)

对Wasserstein Distance 引入熵，迭代逼近其数值解

$$W(P,Q) = min_{\Gamma}\{<\Gamma, C> | A\Gamma = b, \Gamma \geq 0\}$$

$$W_{\lambda}(P,Q) = min_{\Gamma}\{<\Gamma, C> -\frac{1}{\lambda}H(\Gamma) | A\Gamma = b, \Gamma \geq 0\}$$ Sinkhorn distance

迭代更新

$$u^{l+1} = \frac{P}{Kv^{l}},\ v^{l+1} = \frac{Q}{Ku^{l}},\ K = \exp(-\lambda C)$$

$$\Gamma_{i,j} = u_i K_{i,j} v_j$$

$$W_{\lambda}(P,Q) = <\Gamma, C>$$

SinKD

将Sinkhorn Distance用于LLM蒸馏

Method	Complexity	COLA (MCC)	SST-2 (ACC)	MNLI (ACC)	MRPC (F1)	RTE (ACC)	QNLI (ACC)	QQP (ACC)
RKL	$O(bd)$	53.9	91.6	82.9/83.4	90.5	67.1	90.1	91.1
JS	$O(bd)$	54.2	92.2	83.1/83.7	90.7	68.9	90.3	91.2
TVD	$O(bd)$	54.1	92.1	83.3/83.8	90.9	70.0	90.2	91.2
SinKD	$O\left(b^2(d+T)\right)$	60.2	93.1	83.8/84.2	91.3	71.1	90.5	91.3

Table 4: Comparison with distillation methods based on variants of f-divergence on GLUE.

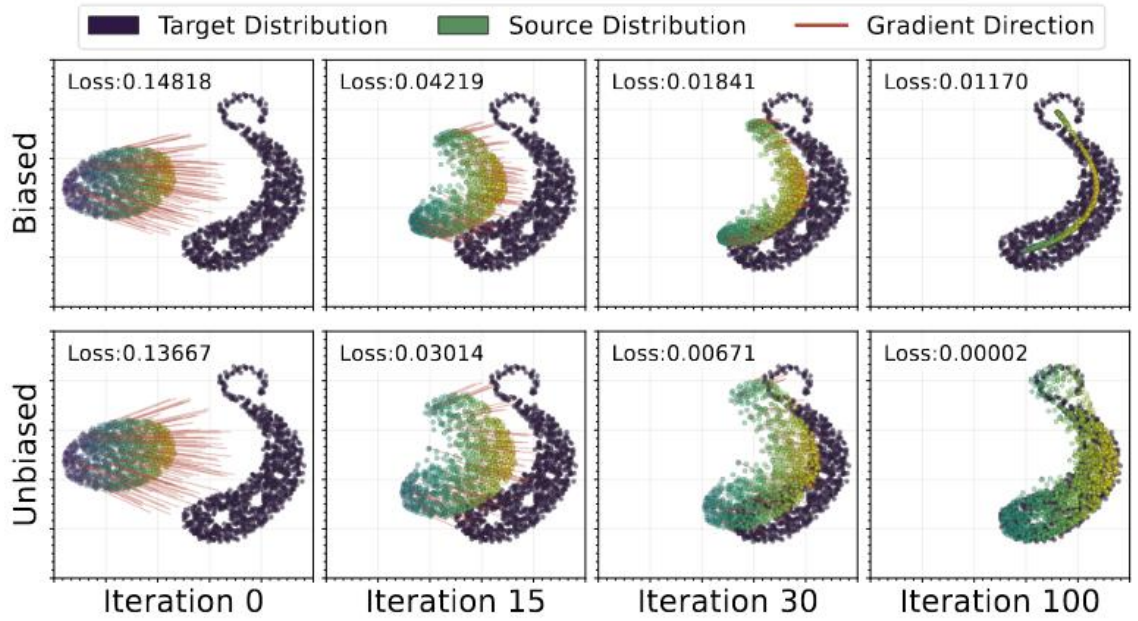
Cui, X., Qin, Y., Sinkhorn Distance Minimization for Knowledge Distillation. In COLING,24

Unbiased Sinkhorn Divergence

$$W_\lambda(P,Q) = < \Gamma, C >$$

引入的熵带来偏差，可能$P = Q$，但是$W_\lambda(P,Q) \neq 0$

$$W_\lambda^{Unbiased}(P,Q) = W_\lambda(P,Q) - \frac{1}{2}W_\lambda(P,P) - \frac{1}{2}W_\lambda(Q,Q)$$



Z,Z., H,Du,H,Bridging the Gap: Teacher-Assisted Wasserstein Knowledge Distillation for Efficient Multi-Modal Recommendation,In WWW,25

Thanks!