LLM Alignment Week 8 Progression

Yuanqi Zhao

University of Oxford

2.Oct.2024

本周进度总结：领域知识总结

蒸馏阶段通常采用的损失函数为 $L = a \cdot L_{\text{soft}} + (1 - a) \cdot L_{\text{hard}}$，
其中 L_{soft} 通常是学生模型与教师模型之间的 Kullback-Leibler 散度，
而 L_{hard} 指的是学生模型在进行预测时的交叉熵损失。

但这也导致了以下几个问题：

- 最小化 KL 散度可能会导致学生模型所学习的概率分布出现偏移，对教师模型中的低概率区域产生了更大的关注，而对教师模型中的高概率区域出现低于预期的拟合。
- 当两个分布没有重合区域时，KL 散度会输出无穷大，导致不稳定的梯度（梯度消失）。

本周进度总结：领域知识总结

最新的研究如 MiniLLM(清华 COAI & 微软，2024 ICLR) 中首次在 LLM-distillation 中采用了 reverse-KL（逆 KL 散度）方法来缓解上述问题。即让学生模型专注于教师模型的高概率区域，而放弃对低概率区域的关注。

但是，这也并不是一种很好的方法。在 Wasserstein GAN 论文中，作者讨论了 Reverse-KL Divergence，Forward-KL Divergence 与 Jensen-Shannon (JS) divergence 在学习直线问题（Learning a Parallel Line）中无法提供一个可用的梯度。这也说明了如若使用 RKLD/FKLD/JSD，则无法在低维流形（Manifolds）中学习概率分布。

当然，这一切建立在流形假说上 (Manifolds Assumption)：即我们的语言分布只是高维空间中的一个流形。

本周进度总结：领域知识总结-学习直线问题

Example 1 (Learning parallel lines)**.** Let $Z \sim U[0,1]$ the uniform distribution on the unit interval. Let $\mathbb{P}_0$ be the distribution of $(0, Z) \in \mathbb{R}^2$ (a 0 on the x-axis and the random variable Z on the y-axis), uniform on a straight vertical line passing through the origin. Now let $g_\theta(z) = (\theta, z)$ with θ a single real parameter. It is easy to see that in this case,

- $W(\mathbb{P}_0, \mathbb{P}_\theta) = |\theta|$,
- $JS(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} \log 2 & \text{if } \theta \neq 0\ , \\ 0 & \text{if } \theta = 0\ , \end{cases}$
- $KL(\mathbb{P}_\theta \| \mathbb{P}_0) = KL(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} +\infty & \text{if } \theta \neq 0\ , \\ 0 & \text{if } \theta = 0\ , \end{cases}$
- and $\delta(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} 1 & \text{if } \theta \neq 0\ , \\ 0 & \text{if } \theta = 0\ . \end{cases}$

When $\theta_t \to 0$, the sequence $(\mathbb{P}_{\theta_t})_{t \in \mathbb{N}}$ converges to $\mathbb{P}_0$ under the EM distance, but does not converge at all under either the JS, KL, reverse KL, or TV divergences. Figure 1 illustrates this for the case of the EM and JS distances.

Figure 1: Learning A Parallel Line

本周进度总结：领域知识总结-学习直线问题

以下的图像为学习直线问题中的 JS 距离与 EM 距离。横坐标为 θ, 纵坐标为对应距离。图中我们可以看到 JS 距离无法提供有效的梯度信息。

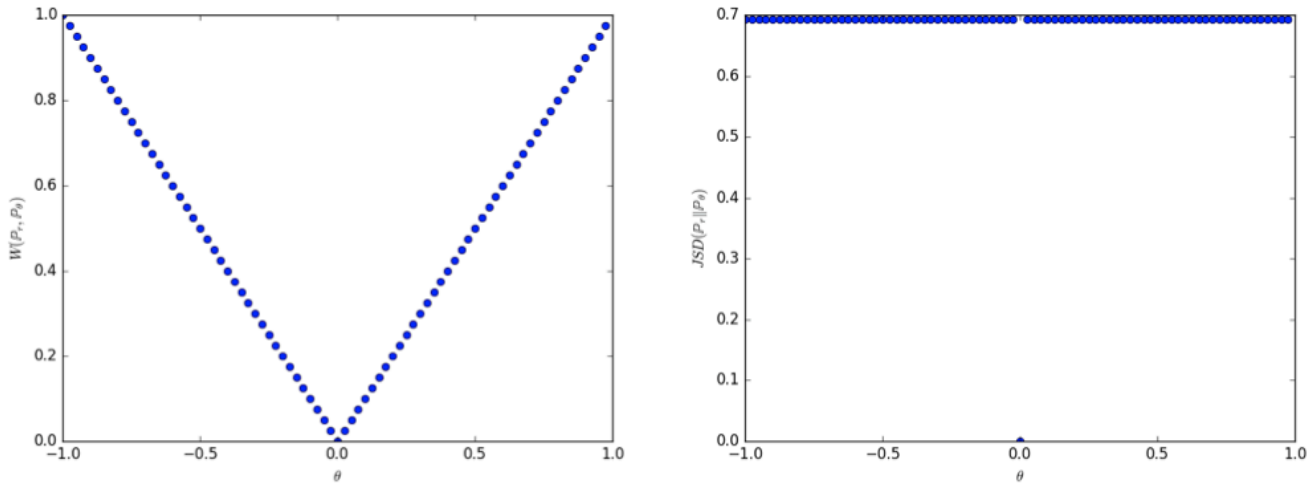

Figure 1: These plots show $\rho(\mathbb{P}_\theta, \mathbb{P}_0)$ as a function of θ when ρ is the EM distance (left plot) or the JS divergence (right plot). The EM plot is continuous and provides a usable gradient everywhere. The JS plot is not continuous and does not provide a usable gradient.

Figure 2: Distance Plot between Wasserstain and KL

本周进度总结：领域知识总结-个人想法延伸

我们是否可以将 Distillation 中的 KL 项替换为 Wasserstein Distance?

定义

Wasserstein Distance(Earth-Move Distance)

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma} \left[\|x - y\| \right], \quad (1)$$

where $\Pi(\mathbb{P}_r, \mathbb{P}_g)$ denotes the set of all joint distributions $\gamma(x, y)$ whose marginals are respectively $\mathbb{P}_r$ and $\mathbb{P}_g$. Intuitively, $\gamma(x, y)$ indicates how much "mass" must be transported from x to y in order to transform the distribution $\mathbb{P}_r$ into the distribution $\mathbb{P}_g$. The EM distance then is the "cost" of the optimal transport plan.

本周进度总结：领域知识总结-个人想法延伸

Wasserstein 距离的优点：

- 更稳定的梯度信息，连续性和可微性。（详细证明在 Wasserstein Gan 论文中）
- 减少模式崩溃：由于 Wasserstein 距离关注整体分布的匹配，它能够减少学生模型只关注教师模型某些高概率区域而忽略其他重要区域的现象（例如对于双峰混合高斯模型，RKL 散度只会关注其中一个峰），从而有效缓解模式崩溃的问题。

Wasserstein 距离的缺点：

- 计算 Wasserstein 距离通常比 KL 散度和 JS 散度更复杂，需要解决一个最优传输问题。（但 spaCy 提供加速算法）

本周进度总结：领域知识总结-个人想法延伸

LLM 的输出符合多项式分布（Multi-Nominal Distribution，MND），也就是 $X|Y$ 属于 MND，可以看到对于长尾分布的 MND（更符合 LLM 的输出概率分布）来说，RKL 散度与 KL 散度并不能对两个分布之间的显著差异有很好的捕获。

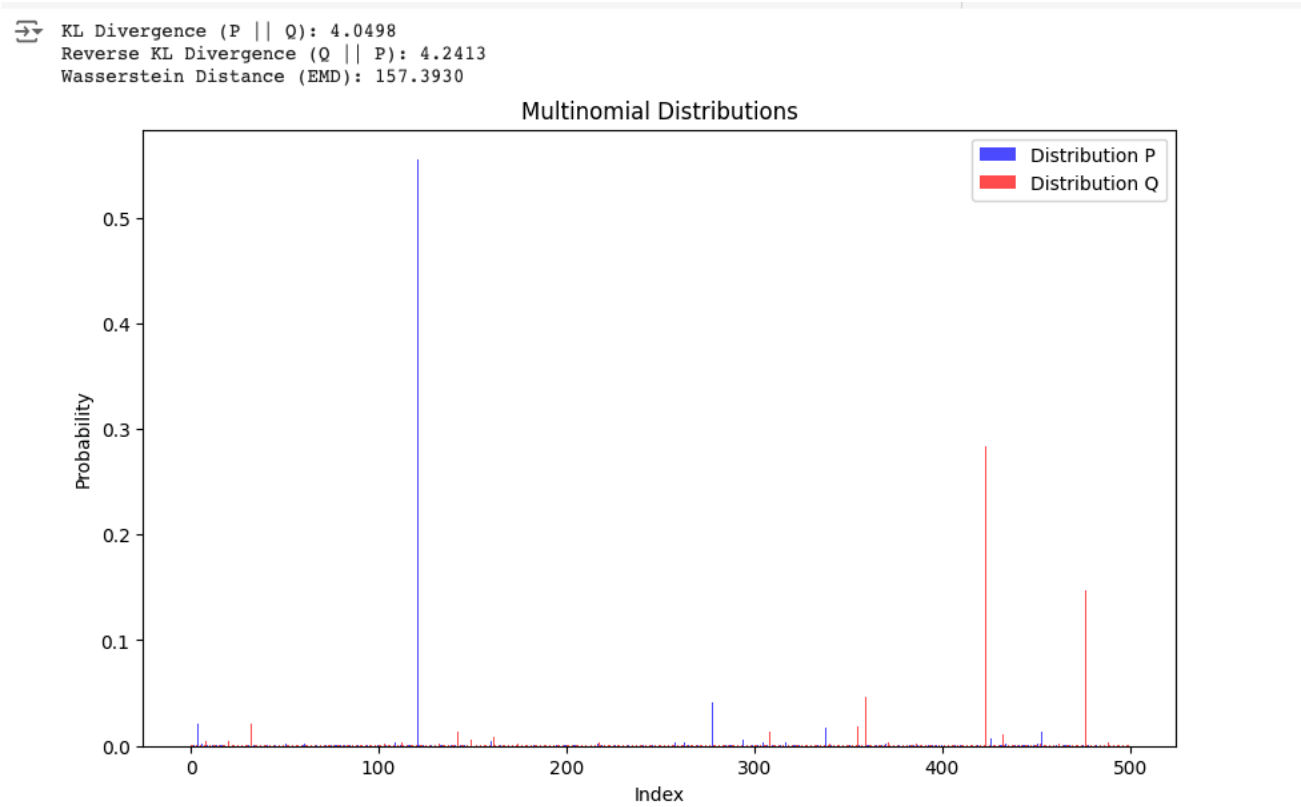

Figure 3: 长尾 MND 比较

下周所做

目前仍待探索的地方是：理论上对于用神经网络去学习多项式分布来说，Wasserstain 距离是否优于 KLD/RKLD？假设词汇表的数量是 20000，即对应的多项式分布是这 2 万维中的 19999 维度，即多项式分布应属于 Manifold。根据前面所讨论的内容来看，wasserstain 应该更合适去作为差异度量。

如果经过理论证明 + 实验，Wasserstain 确实可行。那么下一步应去思考如何做一些 Loss function 上的创新或者训练方法上的创新。