【题目】最小二乘 vs 大模型

自 2023 年起，大模型毋庸置疑地占据着 AI 研究的 C 位。但是，迄今为止，关于大模型工作机理的研究仍然处于起步阶段，它为什么拥有如此强大的能力，人们还未窥门径。与之相对应，就我们的课程而言，最小二乘方法也同样居于核心地位，从线性估计到 Kalman 滤波，从 LMS 到 RLS，诸多信号处理算法的本质，都是最小二乘。两种热门方法，一个已有 300 年历史，一个是火箭蹿升的新贵，它们相遇在一起，会碰撞出什么样的火花？

如所周知，大模型采用 Transformer 架构，其核心组件是 Attention Head，也就是所谓的“注意力头”。大模型的“神奇”能力，主要来自于“注意力头”因此，透彻理解“注意力头”，对于深入揭示大模型的工作机理至关重要。有一个有趣的观察，通常意义下的“注意力头”运算如下：假设注意力头的输入是一组矢量（这些矢量在 AI 里称为 Token）

$$X = (X_1, X_2, \cdots, X_n) \in R^{d\times n}$$

那么“注意力头”对其进行的处理为

$$AH(X) = softmax\left(\frac{(W^Q X)^T (W^K X)}{\sqrt{d}}\right)(W^V X)^T$$

其中，W^Q, W^K, W^V是三个参数矩阵（分别称为 Query, Key, Value）。与此同时，对于同样的输入X，最小二乘的处理为

$$LS(X) = (X^T X)^{-1} X^T$$

你注意到了吗？两者之间非常类似，（表面上）简直是像极了！这难道是偶然的吗？其中有没有什么值得研究的有趣联系？

我们的大作业，就从这里开始。我们的目标是，从熟悉的最小二乘出发，对“注意力头”中的各种运算进行分析，建立对“注意力头”的直观与理性认识，从而深入剖析 Transformer 乃至于 LLM 的工作机理。

我们这里重点考虑如下问题：

1. 参数矩阵W^Q, W^K, W^V的作用，究竟是什么？标准的最小二乘中，是没有参数矩阵的。如果在最小二乘中加上参数矩阵，会出现什么情况？
2. 最小二乘中的矩阵求逆，是处理的关键。这个矩阵求逆，起到的是什么作用？

与之相对，“注意力头”使用的，是 Softmax。这个操作的本质是什么？和求

逆相比，其差异在什么地方？

3. “注意力头”里的 Softmax 函数中，还有一个尺度因子$\sqrt{d}$。最小二乘中并没有这一项。这一项起什么作用？如果加入到最小二乘里，会有什么效果？
4. Transformer 的强大能力，还来源于其所谓的“多头注意力”机制。简单地说，就是把每一个 Token 的d维矢量，拆分为H份，每一份的维度变为d/H，每一份用一个“注意力头”来处理，然后将处理结果组合在一起。最小二乘里没有这样做。如果我们在最小二乘里这样做，会有什么效果？是否对于高维数据有特别的增益？
5. 同学们在阅读文献的时候，如果有其他的想法和思路，也很欢迎。

为了帮助大家尽快进入研究，我们提供部分文献供参考。【1】是 Transformer 最早也是最经典的论文，文中创立了 Transformer 架构，值得仔细阅读。事实上，后续的很多文章，从描述的清晰程度上看，都不如这一篇。【2】使用 Transformer 来解决简单的数学问题，把目标和最小二乘进行了对齐。【3】使用 Transformer 解决线性代数问题，进一步推进对于 Transformer 的认识。抱歉！由于我们所研究的问题比较前沿，参考文献不是很多。

我们此次大作业的任务如下：

1. 阅读文献，形成对 Transformer 的初步认识（最小二乘不用读文献了吧？应该小脑反应了），并根据自身经验/兴趣/实验室方向，参考上面列出的问题，确定研究方向。
2. 收集并仔细阅读选定研究方向的相关文献，切实把握其基本概念和基本方法，并找到相应突破点。
3. 通过相应的理论计算与数值仿真（工具任选），得到有意义的结论，同步获得对于最小二乘以及 Transformer 的深入认知。
4. 整理所得到的结果（理论公式/曲线/表格），完成大作业报告。

【参考文献】

【1】 Attention Is All You Need

【2】 What Can Transformers Learn In-Context

【3】 Linear algebra with transformers