最小二乘是注意力机制的一个特例

摘要

最小二乘法在信号处理领域源远流长，具有重要的历史地位；而 Transformer 一经提出，就在机器翻译等序列建模领域取得巨大成功，其中的 Attention 机制（ $Attn(X)=softmax\left((W_qX)^T\cdot(W_kX)\right)\cdot(W_vX)^T$ ），更是与最小二乘（ $LS(X)=(X^T\cdot X)^{-1}\cdot X^T$ ）有着千丝万缕的联系。

本文首先从二者表面上的相似入手，通过鉴别二者X矩阵定义的不同，分析二者在内积和归一化方式的多方面差异，进而得出$LS(X)$和$Attn(X)$在运算模式、几何意义和物理意义上存在本质区别的结论。

在第二部分，本文从信号系统的角度对最小二乘法和注意力机制进行描述和建模，将二者所面向的参数估计和序列建模问题统一为信号系统的辨识问题，为定性衡量二者的相似性提供可行途径。

在第三部分，基于信号系统的角度，对最小二乘系统进行分析和转化，最终从理论上证明，在选取归一化函数为单位映射的条件下、最小二乘系统是注意力系统的一个特例。

实验部分证实上述论断，并展示有/无归一化的注意力系统的收敛情况。

目录

1 引入

Transformer 在自然语言处理领域发挥了无与伦比的威力，它的运算模式天生就适合于处理序列化建模的问题。

Vaswani 等人在“Attention Is All You Need”中提出 Transformer 模型，通过采用注意力机制实现的 Encoder-Decoder 架构，完全移除 RNN 模块和卷积层，不仅具有增强并行性、缩短训练时间，还分别在 WMT2014 英德和英法翻译任务中取得最好成绩。然而，尽管 Vaswani 等在原文中对 Transformer 架构和核心的注意力机制进行了清晰的描述，但缺乏对注意力机制有效性的原因剖析，从而限制我们对于“注意力机制适用于何种问题？如何选取输入输出以使注意力机制更好地学习序列特征？”等问题的认知，影响 Transformer 在其它问题的应用。

Charton 在“Linear algebra with transformers”中应用 Transformer 解决九类线性代数问题，并提出 P10、P1000 等四种矩阵编码方式，是对 Transformer 建模能力和适用性的一次有益探索。然而，其探索局限于对实验现象的观察和总结，缺乏有效的规律总结或更进一步的理论分析。

Garg 等人在“What can transformers learn in-context? a case study of simple function classes”中聚焦于 In-context learning 能力，将 Transformer 应用于函数拟合问题，在线性函数族、双层神经网络等函数族上取得良好效果。尽管在衡量 Transfomer 建模能力时，最小二乘估计被用作对照组，但研究并未深入到分析为何无噪声情况下 Transformer 逼近最小二乘性能，也没有深入到对 Transformer 核心的注意力机制和最小二乘的关系的分析。

受 Vaswani 等人启发，我们从 Transformer 中抽离出（单头）注意力机制作为研究对象。Charton 的不同编码方式启发了关于选取注意力机制输入格式的思考。Garg 等人的实验现象促使形成注意力机制和最小二乘存在某种联系的信念，另外，其对于$tokens$的选取方式也强化对注意力机制输入的重视，引发关于单位制和物理意义的思考，间接引导形成信号系统的描述方式。

2 模型

d维随机变量x和1维随机变量y满足$y = x^T\beta + \varepsilon$，其中$\varepsilon$是高斯白噪声。

称x为该线性带噪系统的激励，称y为激励x下的观测。

激励x取值的向量空间称激励空间，记为S；观测y取值的向量空间称观测

空间，记为T。（注：由于激励x和观测y常常是不同的物理量，因此，S和T不能简单地认为是$\mathbb{R}^d$和$\mathbb{R}$）

2.1 最小二乘问题

首先，我们简要回顾一下最小二乘法要解决的问题，我们希望通过n个激励$(x_1,\dots,x_n)$以及对应的观测$(y_1,\dots,y_n)$，得到参数向量β在最小二乘准则下的“最优”估计 $\hat{\beta}_{LS}=\underset{\beta}{\operatorname{argmin}}\sum_{i=1}^{n}(y_i-x_i^T\beta)^2$ 。

通过求目标函数的导数的零点，解得（向量形式）的最小二乘估计为

$$\hat{\beta}_{LS}=(X^TX)^{-1}X^TY \tag{1}$$

其中，$X=(x_1,\dots,x_n)^T\in\mathbb{R}^{n\times d}$，$Y=(y_1,\dots,y_n)^T\in\mathbb{R}^{n\times 1}$

同时可以证明，最小二乘估计是参数向量β的一致最小方差无偏估计（UMVUE）。我们通过验证$\hat{\beta}_{LS}$达到 Rao-Cramer 下界来说明这一点：

Step1: 求Y的联合概率密度函数f_Y

给定外部激励$\{x_i\}_{i=1}^n$，由$\varepsilon\sim\mathcal{N}(0,\sigma^2I_n)$，知$Y\sim\mathcal{N}(X^T\beta,\sigma^2I_n)$。故有

$$f_Y(y_1,\dots,y_n)=\frac{1}{\sqrt{(2\pi\sigma^2)^d}}\exp\left\{-\frac{\sigma^2}{2}(y-X\beta)^T(y-X\beta)\right\}$$

Step2: 求$\nabla_\beta lnf_Y$和$\nabla_\beta^2 lnf_Y$

$$\begin{cases}\nabla_\beta \ln f_Y=\dfrac{1}{\sigma^2}X^T(y-X\beta)\\ \nabla_\beta{}^2 \ln f_Y=-\dfrac{1}{\sigma^2}X^TX\end{cases}$$

Step3: 求 fisher 信息量$I(\beta)$与 Rao-Cramer 下界 CRLB：

$$\begin{aligned}I(\beta)&=\mathbb{E}[-\nabla_\beta{}^2\ln f_Y]\\&=\frac{1}{\sigma^2}X^TX\end{aligned}$$

$$\begin{aligned}\mathrm{CRLB}&=I(\beta)^{-1}\\&=\sigma^2(X^TX)^{-1}\end{aligned}$$

注意到最小二乘估计$\hat{\beta}_{LS}$的方差为：

$$\begin{aligned}Cov(\hat{\beta}_{LS},\hat{\beta}_{LS})&=(X^TX)^{-1}X^TCov(\varepsilon,\varepsilon)X(X^TX)^{-1}\\&=\sigma^2(X^TX)^{-1}\end{aligned}$$

达到了 CRLB。因此，最小二乘估计是一致最小方差无偏估计量，是统计意义上的最优估计。

2.2 注意力机制

Transfomer 由 Encoder 和 Decoder 组成，Encoder 和 Decoder 都有六层，但 Encoder 每层的主体是自注意力层，Q、K、V是同一组向量，而 Decoder 每层是注意力层，Q是前一层输出，而K、V是 Encoder 编码结果。

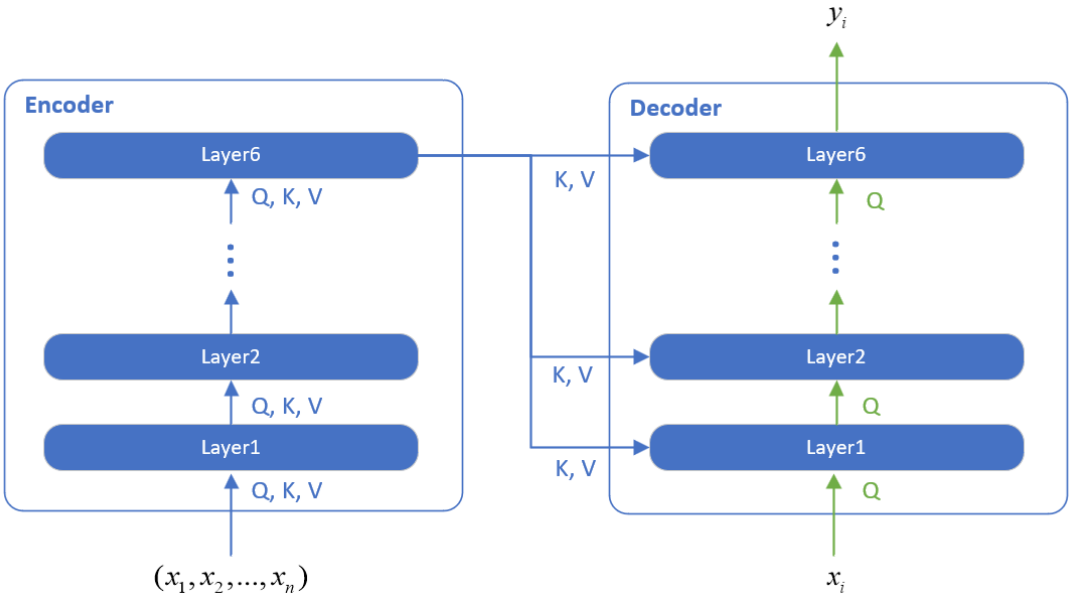


图 1 Transformer 的 Encoder-Decoder 架构

Transformer 在 Seq2Seq 问题中取得优异成绩，一大原因就是它无需对序列长度做出假设和限制。相较于基于n阶马尔可夫性假设（即自然语言处理的n元文法模型）的前馈神经网络，它能够处理任意长度的输入序列；相较于将输入的$token$不断叠加到定长的 hidden state 的循环神经网络，它避开不断叠加$token$带来的信息瓶颈问题。而达成这一效果的，正是内积-归一化机制，也即“Attention Is All You Need”中提到的“Scaled Dot Product[1]”。

对于长度为m的d维序列$K = (k_1, \dots, k_m)$、$V = (v_1, \dots, v_m)$，和长度为n的d维序列$Q = (q_1, \dots, q_n)$，Scaled Dot Product 的运算为：

$$y = h(Q^T \cdot K) \cdot V^T \tag{2}$$

其中，h是归一化函数，在 Transformer 中为以行为单位的$softmax$；输出结果y

[1] 正如“Attention Is All You Need”中 4 号脚注所言，$\sqrt{d}$用于初始化网络参数时归一化方差，多见于早期手写神经网络时初始化矩阵，对于注意力机制的表示能力无影响（因为W_q和$W_q/\sqrt{d}$的表示范围均为$\mathbb{R}^{d\times d}$）

也为长度为n的d维序列。Encoder 中采用自注意力机制，满足：$\begin{cases} Q = W_q X \\ K = W_k X \\ V = W_v X \end{cases}$，其中，$X = (x_1, ..., x_n) \in \mathbb{R}^{d \times n}$。**注意：为了保持与最小二乘的表面相似性，这里$X$的定义与最小二乘中恰好差一个转置。**

注意力机制有明确的几何含义：

1. 由于W_q、W_k、W_v仅与特征维度有关而与序列长度无关，因此注意力层的参数是固定的，而不是像前馈神经网络通过限制序列长度来固定网络参数大小；同时，$W_q(x_1, \dots, x_n) = (W_q x_1, \dots, W_q x_n)$，即参数矩阵$W_q$、$W_k$、$W_v$的几何意义是对输入的 tokens 中的每一个 token 进行线性变换，即映射到特征空间。

2. $Q^T K = \begin{pmatrix} q_1^T K \\ \vdots \\ q_n^T K \end{pmatrix}$，其中$q_i^T K = (q_i^T k_1, \dots, q_i^T k_n)$，是特征向量$q_i$与一组特征向量$(k_1, \dots, k_n)$的内积，而内积的几何意义是相似度。

3. 归一化函数h将n组相似度分别归一化为加权系数，并对特征$V = (v_1, \dots, v_n)$进行加权得到输出y，y也是由n个向量组成的序列，每个向量都属于v所在的向量空间。事实上，当选取$h = softmax$时，加权系数均属于$[0, 1]$，每个向量都在$(v_1, \dots, v_n)$所构成的闭包中。

3 最小二乘与注意力机制的关系分析

对比最小二乘估计$\beta_{LS} = (X^T \cdot X)^{-1} \cdot X^T Y$和 Transformer 中 Scaled Dot Product 的$softmax(Q^T \cdot K) \cdot V$，可以发现，二者的运算有以下相似之处，一是内部均为内积的形式，可以视为相似度的计算；二是矩阵求逆运算和$softmax$运算均可视为某种归一化运算的方式。但通过进一步的分析，我们发现这两处表面上的相似，本质上都存在较为明显的不合理之处。但通过重新审视二者所尝试解决的问题，我们能够断言，最小二乘问题可以被转化为 Transformer 所面向的序列建模问题，而最小二乘法就是注意力机制的一种特例。

我们将在第一小节深入分析$(X^T X)^{-1} X^T$与自注意力层的 scaled dot product 这两类运算的差异，从多个角度说明为何$(X^T X)^{-1} X^T$和$softmax\left((W_q X)^T (W_k X)\right)(W_v X)^T$这两个看似极为相似的运算存在本质上的差异；

在第二小节重新审视注意力机制本身，从注意力层输入输出的选取入手，

分析研究其他学者用注意力机制解决最小二乘问题的方式和不合理之处，并从点积和$softmax$的几何意义出发，将注意力层在 high-level 上抽象为一个由激励空间S到观测空间T的一个映射；

基于映射的观点，在第三小节我们将注意力机制用信号系统的方式建模，进而将最小二乘问题和注意力机制序列建模问题都转化为信号系统的辨识问题，在此基础上，分析“最小二乘系统”和“注意力系统”的相似性成为可能。

3.1 注意力机制和最小二乘的内积和归一化的异同

正如 2.1 和 2.2 所提到的，在最小二乘中，

$$LS(X)=(X^TX)^{-1}X^T \tag{3}$$

其中$X=(x_1,\dots,x_n)^T\in\mathbb{R}^{n\times d}$；

而在（自）注意力机制中，

$$Attn(X)=softmax\left((W_qX)^TW_kX\right)(W_vX)^T \tag{4}$$

其中，$X=(x_1,\dots,x_n)\in\mathbb{R}^{d\times n}$。

尽管式(3)和式(4)具有极为相似的形式，但遗憾的是，通过进一步的分析，我们发现，式(3)和式(4)的相似性可能是表面的，差异则是本质的。

二者分歧的关键，在于X矩阵的定义相差一个转置，这一差异看似无关紧要，实则可能导致几何意义和物理意义的截然不同，主要体现在内积和归一化两方面。

在注意力机制中W_q、W_k和W_v的主要作用是对每个x_i进行线性变换，映射到新的特征空间进行运算，不失一般性，我们对$W_q=W_k=W_v=I_d$的情况进行分析，其他情况是类似的。

3.1.1 内积的差异——相似与相关

由于X矩阵定义的差异，在注意力机制中，$X^TX=\begin{pmatrix}x_1^T\\ \vdots\\ x_n^T\end{pmatrix}\cdot\begin{pmatrix}x_1 & \cdots & x_n\end{pmatrix}=\begin{pmatrix}x_1^Tx_1 & \cdots & x_1^Tx_n\\ \vdots & \ddots & \vdots\\ x_n^Tx_1 & \cdots & x_n^Tx_n\end{pmatrix}=\left(x_i^Tx_j\right)_{n\times n}$，即$X^TX$矩阵的第$i$行第$j$列的元素是第$i$和$j$个激励的点积。因此，从几何意义上来说，$X^TX$矩阵是历史激励两两之间的相似度矩阵。自然地，$X^TX$的长和宽均为样本容量$n$，在实际的自然语言处理问题中，$X^TX$的维数随句子长度（即$tokens$的长度，也即这里的样本容量$n$）的增加而增加。

而在最小二乘中，$X^TX=\begin{pmatrix}x_1 & \cdots & x_n\end{pmatrix}\cdot\begin{pmatrix}x_1^T\\ \vdots \\ x_n^T\end{pmatrix}=\sum_{i=1}^{n}x_ix_i^T$，注意到$x$各分量（也即，$d$个特征）之间的相关矩阵的无偏估计是$\frac{1}{n}\sum_{i=1}^{n}x_ix_i^T$。因此，如果不考虑常数项，最小二乘的$X^TX$可以看作是对$d$个特征的相关矩阵的估计。$X^TX$的长和宽均为特征维数$d$，不仅维数不随样本容量的变化而变化，而且由大数定律，当样本容量较大时，$\frac{1}{n}X^TX$收敛于真实的相关矩阵。

3.1.2 归一化的差异——单位制

尽管LS中矩阵的取逆和$Attn$中的$softmax$都可视作某种意义上的归一化操作，但二者的差异不仅体现在求逆是对矩阵整体的操作而$softmax$是对行的操作，同时体现在单位的不同（进而体现在物理意义的不同）上。

这里，我们不妨假设x是物理量电流，单位是安培，记作A。

对于最小二乘，X的单位是A，$(X^TX)^{-1}$的单位是$1/A^2$，而$LS(X)=(X^TX)^{-1}X^T$的单位为$1/A$。因此，从物理意义上考虑，$LS(X)$将电流映射为电流的倒数，LS整体上是一个归一化因子。

对于注意力机制，X的单位也为A，但通过内积操作，X^TX成为无单位矩阵，每个元素代表相似度，经过$softmax$归一化后，每行是一组和为 1 的加权系数。再右乘X后，得到的$Attn(X)$是单位为A的矩阵。

综上，最小二乘中的$(X^TX)^{-1}$是单位为$1/A^2$的$d\times d$维矩阵，代表相关矩阵的逆；而注意力机制中的$softmax(X^TX)$是无单位的$n\times n$维矩阵，代表归一化的加权系数。无论从物理意义、几何意义和矩阵参数上，二者都存在明显差异。除此之外，最小二乘解决的是参数估计问题，而注意力机制面向的是序列建模问题，二者适用场景的差异也是分析二者关系中一个不可忽视的问题。

这迫使我们重新审视二者的适用场景，只有将二者的应用场景统一起来，找到正确的切入角度，分析二者的相似性才不是无稽之谈。为达成这一目的，我们先分析注意力机制的输入输出特性。

3.2 注意力机制的输入输出特性

注意力机制利用内积和归一化计算权重向量并对一组特征向量进行线性组合，那么，在利用注意力机制建模最小二乘问题时，如何选取查询向量、键向量和值向量是无法回避的问题。

Garg 等将最小二乘的样本和观测[2]组合为$(x_1, y_1, \dots, x_n, y_n, x)$，并将其作为$tokens$输入 Transformer，对于其中 Encoder 的自注意力层，$tokens$同时作为查询向量、键向量和值向量。其中，x_i和y_i维数不一致的问题是通过向y_i填充零解决的。

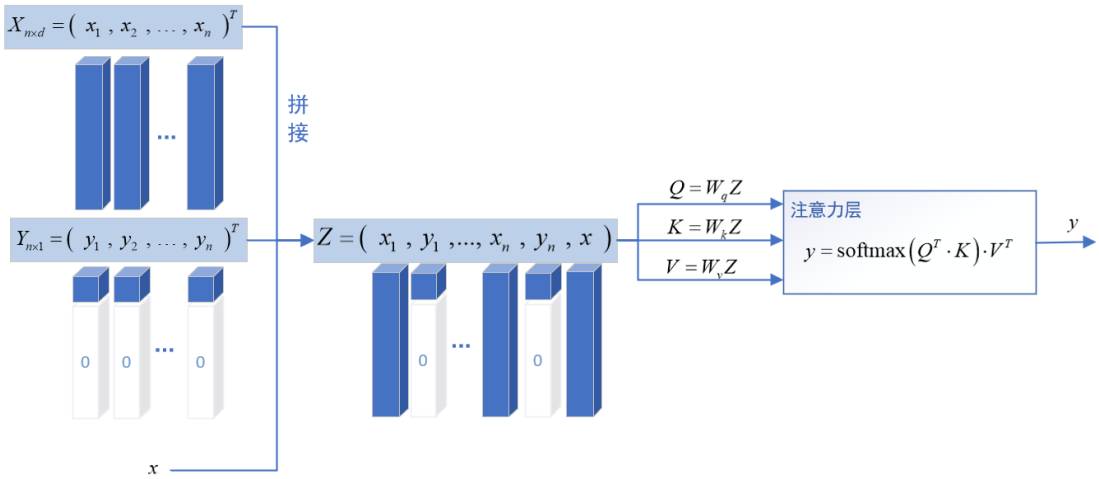


图 2 X和Y拼接输入的自注意力层

这种方式虽然也充分利用训练数据$\{x_i\}_{i=1}^n$和$\{y_i\}_{i=1}^n$，但显然不甚优雅，原因在于$\{x_i\}_{i=1}^n$和$\{y_i\}_{i=1}^n$处于不同的向量空间S和T，从问题的实际背景来看，x和y在尺度（例如，x以mA为单位，而y以A为单位）甚至单位制上（例如，x是电流而y是电压）都是不同的，而自注意力机制的出发点在于利用矩阵乘法得到同一组 tokens 在三个向量空间下的表示Q、K、V，再利用点积计算Q和K之间的相似度，通过$softmax$归一化后，对V进行线性组合。直接将$\{y_i\}_{i=1}^n$填充后和$\{x_i\}_{i=1}^n$进行拼接尽管在计算格式上符合自注意力机制的要求，但其物理意义不甚明晰。

然而，既然x_i和y_i处于不同向量空间中，且天然地存在一一对应的关系；又考虑到Q、K、V三者中，K和V长度相同，且K_i和V_i也存在一一对应关系。那么，一个自然的想法是用$(x_1, \dots, x_n)$构成K而用$(y_1, \dots, y_n)$来构成V。

[2] Garg 原文中使用记号$f(x_i)$，而本文使用y_i

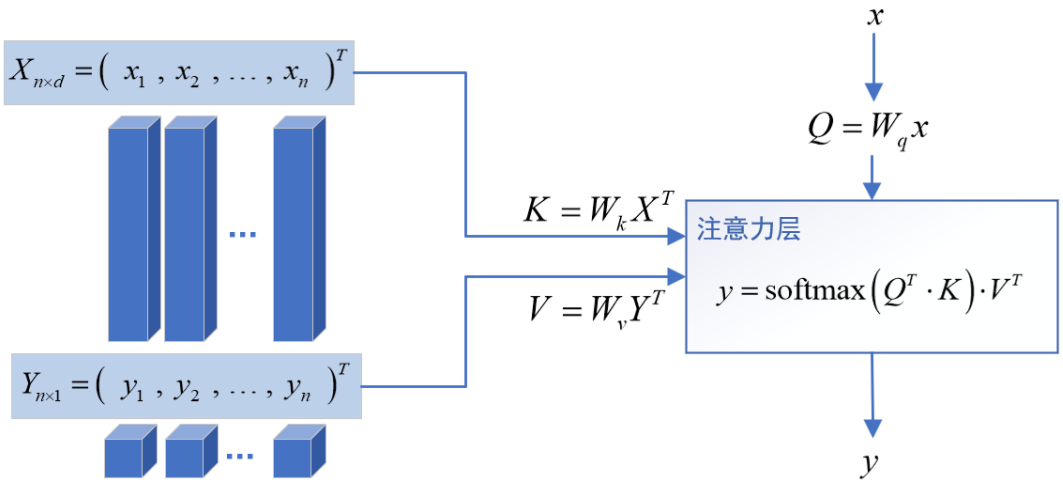


图 3 X和Y分别输入的注意力层

同时，注意到注意力机制中的$softmax$运算将相似度向量/矩阵$Q^T\cdot K$归一化为权重向量/矩阵，因此注意力层的输出其实是观测$(y_1,\ldots,y_n)$的线性组合。

在固定历史激励$\{x_i\}_{i=1}^n$和历史观测$\{y_i\}_{i=1}^n$的条件下，注意力层的输入仅为d维向量x，而输出为观测y。从 high-level 的角度来看，注意力层其实是激励空间S到观测空间T的一个映射！

3.3 注意力机制和最小二乘的信号系统描述

从信号系统的角度来分析，若将$y=x^T\beta+\varepsilon$作为原系统，那么，无论是最小二乘法还是注意力机制，都可以看作基于历史激励$\{x_i\}_{i=1}^n$和历史观测$\{y_i\}_{i=1}^n$对原系统的一种建模。

最小二乘法所描述的信号系统（简称为最小二乘系统）将系统建模为线性，基于最小二乘准则得到参数向量β的估计$\hat{\beta}_{LS}=(X^TX)^{-1}X^TY$，然后将新的激励$x$映成$\hat{y}_{LS}$。也即，最小二乘系统的输入输出特性：

$$\hat{y}_{LS}=x^T(X^TX)^{-1}X^TY \tag{5}$$

注意力机制所描述的信号系统（简称为注意力系统）对新的激励x进行线性变换后，与线性变换后的历史激励分别内积得到相似度，然后利用归一化函数h作用得到加权系数，对线性变换后的历史观测进行线性组合得到输出$\hat{y}_{attn}$。也即：注意力系统的输入输出特性：

$$\begin{aligned}\hat{y}_{attn}&=h\left((W_qx)^TW_kX^T\right)(W_vY^T)^T\\&=h(x^TW_q{}^TW_kX^T)YW_v^T\end{aligned} \tag{6}$$

其中W_q、W_k是$d\times d$维的矩阵，而W_v是标量。

原系统、最小二乘系统和注意力系统的运算可由下图表示：

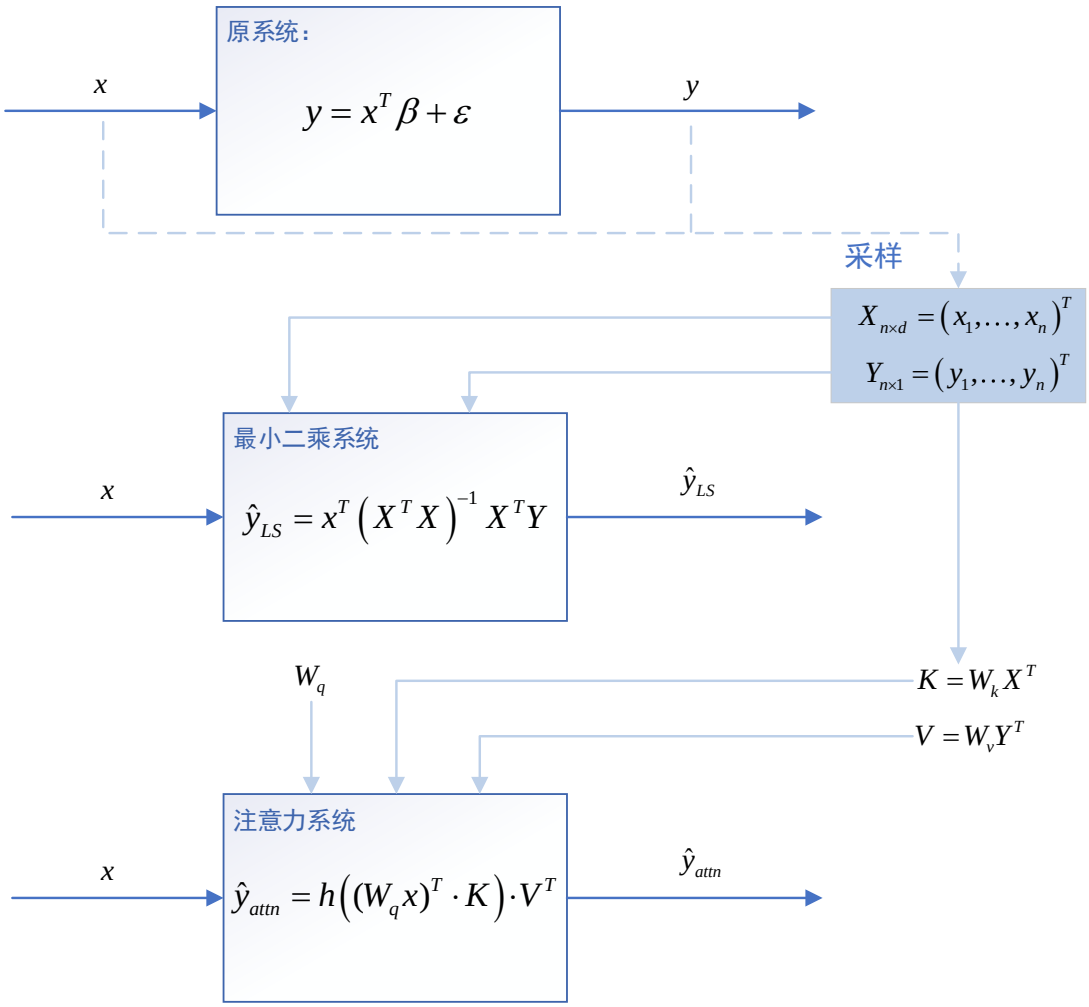


图 4 三种信号系统的运算图

在将二者所要解决的问题都转化为信号系统的形式之后，我们终于可以对二者的相似性进行分析。这里需要注意的是，最小二乘系统是一个无学习参数的、确定性的系统，给定激励$\{x_i\}_{i=1}^n$和观测$\{y_i\}_{i=1}^n$，其系统的内部参数随即确定；而注意力系统是一个有学习参数的系统（W_q、W_k、W_v），学习参数需要在数据集上训练。不同的学习参数对应不同的输入输出特性，对应不同的系统。

由于最小二乘系统的输入输出关系是一个函数，而注意力系统的输入输出关系是一个函数族，因此，研究这两个系统的关系，本质上是在研究一个函数和一个函数族之间的关系。

事实上，只需移除注意力系统的 softmax 运算，即可看出最小二乘系统的函数实际上属于注意力系统所代表的函数族。我们在下一节进行具体的公式推导以说明这一点。

4 对最小二乘系统的分析与转化

这里我们首先要指出一点：当我们将最小二乘问题和注意力机制用信号系统的方式建模之后（如图 4 所示），最小二乘系统已经表现出和注意力系统相似的形式。比如$(X^TX)^{-1}$是$d\times d$维的矩阵，可以看作对键向量空间到新特征空间的线性变换；而通过对注意力系统进行如下选取：

$$\begin{cases} W_q = I_d \\ W_k = (X^TX)^{-1} \\ W_v = 1 \end{cases}$$

以及将归一化函数h选为单位映射id，立得

$$\begin{aligned} \hat{y}_{attn} &= id\left(\left(I_d x\right)^T\cdot(X^TX)^{-1}X^T\right)\cdot\left(1\cdot Y^T\right)^T \\ &= x^T(X^TX)^{-1}X^TY \\ &= \hat{y}_{LS} \end{aligned}$$

然而，将归一化函数h直接选取为单位映射显得不太自然，也不能体现出不同的激励如何影响权重的大小。但从单位制的角度来分析，$(X^TX)^{-1}$为负幂次，直觉上类似于一个平方项的归一化因子，恰好能够抵消Q^TK的平方项的单位，使得整体成为一个无单位的加权系数。

因此，为更进一步分析$(X^TX)^{-1}$，我们采用奇异值分解的手段，将X分解为物理意义更明确的正交矩阵和对角矩阵。

4.1 公式推导

首先对X进行奇异值分解

$$X = U\Sigma V^T$$

其中，正交阵$U\in\mathbb{R}^{n\times n}$，分块对角阵$\Sigma\in\mathbb{R}^{n\times d}$，正交阵$V\in\mathbb{R}^{d\times d}$。

在最小二乘问题中，我们通常假定X列满秩。由于样本个数通常远大于特征维数，因此，列满秩的要求通常是成立的。在此条件下，$\exists\Lambda\in\mathbb{R}^{d\times d}\quad s.t.\quad X=\begin{pmatrix}\Lambda\\0\end{pmatrix}$，其中$\Lambda=\begin{pmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_d\end{pmatrix}$且$\lambda_i>0,\ i=1,2,...$

代入X^TX中，得到

$$\begin{aligned} X^TX &= (U\Sigma V^T)^T(U\Sigma V^T) \\ &= V\Sigma^T\Sigma V^T \\ &= V\begin{pmatrix}\Lambda^T & 0\end{pmatrix}\begin{pmatrix}\Lambda \\ 0\end{pmatrix}V^T \\ &= V\Lambda^T\Lambda V^T \end{aligned}$$

取逆之后，得到

$$\begin{aligned} (X^TX)^{-1} &= (V^T)^{-1}\Lambda^{-1}(\Lambda^{-1})^TV^{-1} \\ &= (V\Lambda^{-1})(V\Lambda^{-1})^T \end{aligned}$$

记$L=(V\Lambda^{-1})^T$，则参数向量β的最小二乘估计：

$$\begin{aligned} \hat{\beta}_{LS} &= (X^TX)^{-1}X^TY \\ &= L^TL(x_1,x_2,\ldots,x_n)Y \end{aligned}$$

假设第$n+1$时刻，外部施加的输入为x，则第 n+1 时刻预测输出$\hat{y}_{LS}$满足

$$\begin{aligned} \hat{y}_{LS} &= x^T\hat{\beta}_{LS} \\ &= \left((Lx)^T\cdot L(x_1,\ldots,x_n)\right)\cdot\left(1\cdot(y_1,\ldots,y_n)\right)^T \end{aligned} \tag{7}$$

4.2 小结

4.2.1 结论 1-最小二乘系统是注意力系统的一种特例

通过对注意力系统作如下选取：

$$\begin{cases} W_q = L \\ W_k = L \\ W_v = 1 \end{cases}$$

和$h=id$，能够得到：

$$\hat{y}_{attn} = id\left((Lx)^T\cdot L(x_1,x_2,\ldots,x_n)\right)\left(1\cdot(y_1,\ldots,y_n)\right)^T$$

和公式(7)比较，立得该种选取方式下，$\hat{y}_{LS}=\hat{y}_{attn}$，也即最小二乘系统是注意力系统在一种选取下的特例。这里需注意选取方式不止上述一种，事实上，当选取$h=id$时，任意满足$W_q^TW_k\cdot W_v$（$W_v\in\mathbb{R}$，是标量）的选取都能使这两个系统的输入输出特性相同。

4.2.2 结论 2-L沿特征维度对激励进行归一化和去单位化

根据L的定义式，$L=(V\Lambda^{-1})^T=\Lambda^{-1}V^T$。由于$X^TX=V\Sigma^T\Sigma V^T$，故$V$是$X^TX$的单位特征向量构成的矩阵。

注意到$X^T = (x_1, x_2, \dots, x_n)$，故有：

$$X^T X = (x_1, \dots, x_n) \cdot \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}$$

$$= \sum_{i=1}^{n} x_i x_i^T$$

由于$x_i \in S = \mathbb{R}^d (i = 1,2, \dots, n)$是$d$维随机变量，因此$X^T X$实际上是激励$x$的$d$个分量（AKA，$d$个特征）的相关矩阵。

因此，V实际上是特征的相关矩阵在空间中延伸的d个正交单位方向向量(orthonormal)，是白化后的d个特征。$V^T x$代表激励x在这d个正交方向上的投影。

而 $Lx = \Lambda^{-1} V^T x = \begin{pmatrix} \lambda_1^{-1} & & 0 \\ & \ddots & \\ 0 & & \lambda_d^{-1} \end{pmatrix} V^T x$ 中，d个投影分别除以相关矩阵的d个奇异值，也即以白化后的d个特征的标准差进行归一化。由于标准差是有单位的，因此L在对x进行白化和归一化的同时，也进行了去单位化。

4.2.3 结论 3-注意力系统能收敛到最小二乘系统

在 2.1 我们提到最小二乘估计满足：$\hat{\beta}_{LS} = \underset{\beta}{\operatorname{argmin}} \left\| y - x^T \beta \right\|_2^2$。而当$h = id$时，注意力系统是线性系统，且最小二乘系统是注意力系统的一个取值，因此，若将注意力机制后的损失函数设为最小二乘损失，则必然能够也必将收敛到最小二乘系统。

尽管该结论已经从上述的推导中证实，我们在下一节简要验证，并与使用其他归一化函数的情况进行收敛速度和收敛效果上的比较。

5 验证性实验

采用 python 语言编程实现[3]；其中，主要使用 pytorch 搭建注意力层，使用 matplotlib 绘制损失图和误差图。d

算法的流程框图如下，测试一和测试二分别对应 5.1 和 5.2 。

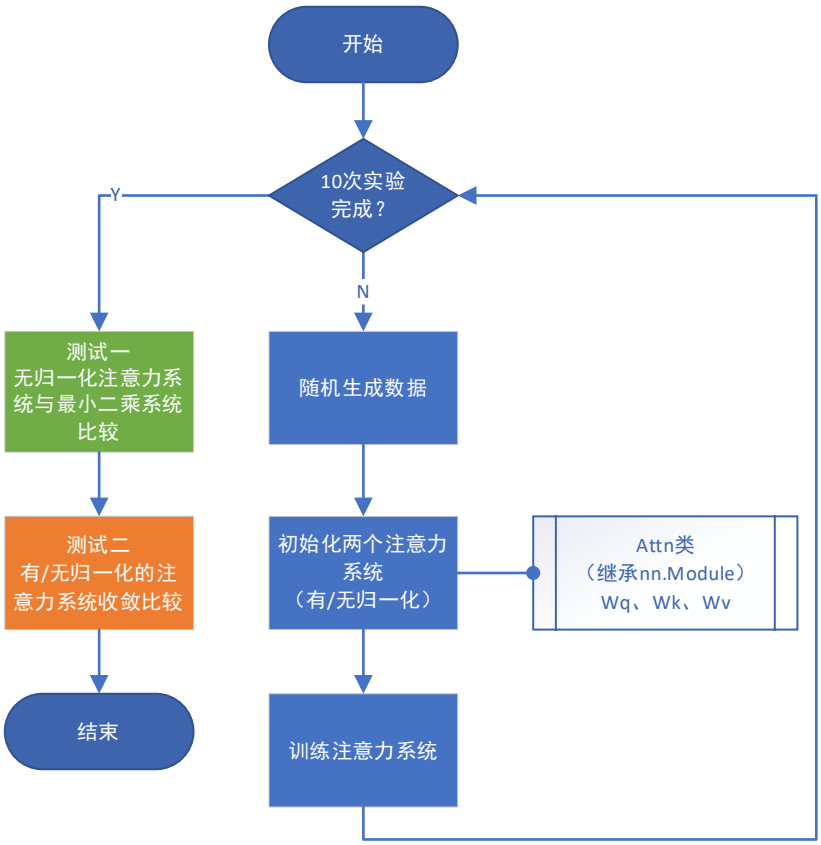
数据生成：

```
torch.manual_seed(0)
```

$n = 100$， $d = 10$

[3] 代码见 code/main.py

$X_{ij} \overset{i.i.d}{\sim} U[-10,10]$， $\varepsilon \sim \mathcal{N}(0, I_n)$， $\beta \sim \mathcal{N}(0, I_d)$



5.1 （无归一化）注意力系统的收敛情况

当$h = id$时（无归一化），注意力系统是输入和输出的线性函数，因此，验证注意力系统收敛到最小二乘系统，只需验证注意力系统等效的参数$\hat{\beta}_{eq}$是否与最小二乘系统的$\hat{\beta}_{LS}$相等。

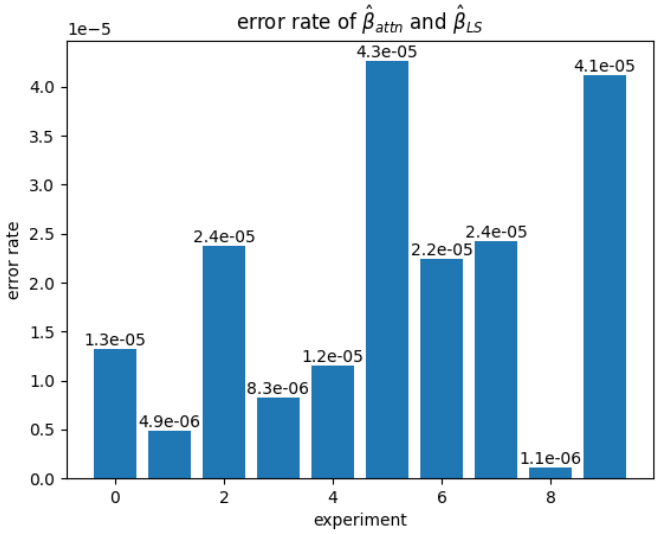
由 3.3 中一般注意力系统的输入输出特性描述（式(6)），无归一化的注意力系统可写为：

$$\hat{y}_{attn} = x^T (W_q{}^T W_k W_v) X^T \cdot Y$$

因此，等效参数$\hat{\beta}_{eq}$为：

$$\hat{\beta}_{eq} = x^T W_q{}^T W_k W_v X^T$$

在实验中，使用误差率$error_rate = \frac{\left\|\hat{\beta}_{eq} - \hat{\beta}_{LS}\right\|_2}{\|\hat{\beta}_{LS}\|_2}$作为判别指标。十次独立实验的误差率均小于万分之一（如图，最大误差率为 0.0043%），验证了无归一化的注意力系统能够收敛到最小二乘系统的论断。

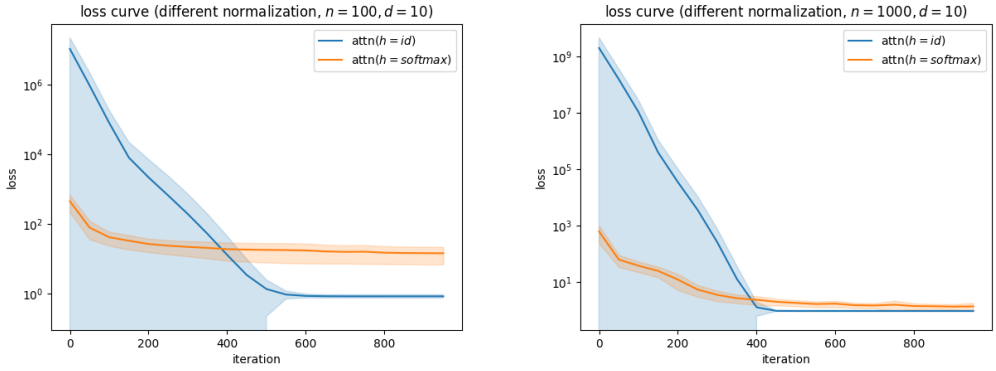


5.2 有/无归一化注意力系统的收敛比较

通过十次独立实验，测试无归一化和$softmax$归一化的注意力系统的收敛情况。损失函数仍采用 MSE Loss，迭代次数为 1000 次。

由于损失曲线的跨度较大，纵轴采用对数坐标系。

曲线为十次实验的损失曲线的均值，阴影部分为一个标准差长度（由于对数坐标，上下不对称）



十次实验中，无归一化注意力系统在迭代前期（迭代次数小于 400 轮）的收敛速度有较大差异，标准差可高达10^6；在迭代次数超过 500 轮后标准差趋于 0，说明已收敛到最小二乘系统。

$softmax$归一化注意力系统的收敛速度则较为稳定，初始损失也集中于10^2~10^3，受实验初始条件的影响较小。

有趣的是，$softmax$归一化注意力系统在$n = 100$时，最终损失比无归一化注意力系统要大 1~2 个数量级；然而，当$n = 1000$时，其最终损失逐渐逼近于归一化的。这说明，**当n充分大时，$softmax$归一化注意力系统可能也具有逼近最小二乘系统的能力，但相对于无归一化注意力系统的无条件逼近，其逼近**

条件可能要复杂得多。

比如，我们可能作出如下猜想：

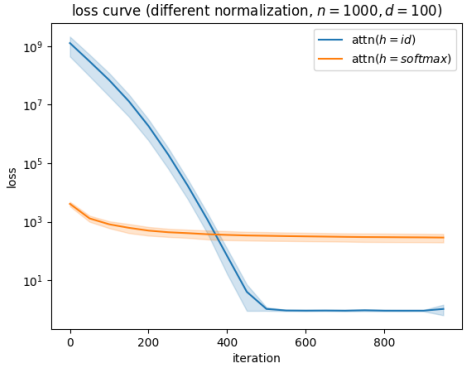
也许与特征维数d有关，

也许与激励在空间中分布的稠密程度有关（注意力机制的闭包特性），

也许与激励的平稳性有关（n较大时，x可看作平稳信号）

……

事实上，对上图中的$n=1000$，$d=10$的情况，当我们增大特征维度d至100时，十次实验得到的收敛曲线图说明 softmax 归一化注意力系统又不能逼近最小二乘系统了，最终损失大概比最优解要差 3 个数量级。这说明d对逼近效果有影响，关于稠密程度的猜想可能正确，但对于真实逼近条件的猜想和验证，仍然是复杂而困难的。



6 结论

1. $LS(X)=(X^TX)^{-1}X^T$ **和** $Attn(X)=softmax\left((W_qX)^TW_kX\right)(W_vX)^T$ **尽管表面上相似，但存在本质的差别。**

二者的X矩阵的定义恰好相差一个转置，这导致最小二乘中X^TX是$d\times d$维，几何含义是各特征之间的相关矩阵，而注意力机制中$\left(W_qX\right)^TW_kX$是$n\times n$维，几何含义是各样本之间的相似度矩阵。

2. 通过将最小二乘法和注意力机制用信号系统的方式描述，能够将二者所面向的最优估计问题和序列建模问题统一起来。

将最小二乘的参数估计和预测过程用一个信号系统来建模（AKA，最小二乘系统，确定性系统，不含网络参数）；受注意力机制的内积和归一化的几何直观和物理意义所启发，将注意力层也用一个信号系统来建模（AKA，注意力系统，含网络参数W_q、W_k、W_v）。

用信号系统的输入输出关系来刻画最小二乘和注意力机制的关系。

3. 最小二乘系统是注意力系统的一个特例，注意力系统能够收敛到最小二乘系统

（无归一化）注意力系统是线性系统，最小二乘系统是注意力系统在W_q、W_k、W_v某种选取下的特例。在训练注意力系统时，如果选取损失函数为 MSE Loss，注意力机制必将收敛于最优解——最小二乘系统。

另外，在保持其他条件不变的前提下，随样本容量n的增大，$softmax$归一化注意力系统也逐渐逼近于最优解。

7 致谢

感谢张颢老师提供这一独特的研究方向并以大作业的形式督促完成，平心而论，仅凭表面上的相似，我个人可能缺乏投身而入的魄力和抽丝剥茧的耐心。正是大作业这一悬在头上的达摩克里斯之剑，我才最终得以克服思维上的“不可能”。从“无从下手”到“努力分析相似性，列举各种相似的方面”，再到“发现由转置导致的各种差异，怀疑二者没关系”，最后到“信号系统角度分析，发现二者的本质等价之处”。这一“山重水复疑无路，柳暗花明又一村”的心路历程，确实是独特而难得的奇妙体验。

感谢我的课题组伙伴和导师，他们有着对大模型和 Transformer 的独到理解，为我理解 Transformer 的组成结构和工作原理提供了宝贵的帮助，和他们的讨论也启发了我的思考。

感谢知乎和博客上对 Transformer 和注意力机制提出个人见解的所有博主，尽管有的可能在专业性和规范性上有所欠缺，但丰富和独特的观点确实对我从多角度理解注意力机制起到了积极作用。

8 参考文献

1. Charton F. Linear algebra with transformers[J]. arXiv preprint arXiv:2112.01898, 2021.

2. Garg S, Tsipras D, Liang P S, et al. What can transformers learn in-context? a case study of simple function classes[J]. Advances in Neural Information Processing Systems, 2022, 35: 30583-30598.

3. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.