理解深度学习

许志钦

2023 年 9 月 13 日

目录

Chapter 1

频率原则的简单理论

1.1 简单分析

1.1.1 激活函数的影响

考虑如下的网络结构：使用 $\sigma(x) = \tanh(x)$（选择 tanh 作为激活函数的原因是因为它的傅里叶变换很容易做）作为激活函数的两层 DNN，用来拟合一维目标函数 $f(x)$, 即网络输入输出均为一维：

$$h(x) = \sum_{j=1}^{m} a_j \sigma(w_j x + b_j) \tag{1.1}$$

由傅里叶变换：

$$\begin{aligned} \hat{\sigma}(w_j x + b_j)(k) &= \frac{2\pi i}{|w_j|} e^{\frac{ib_j k}{w_j}} \frac{1}{\exp(-\frac{\pi k}{2w_j}) - \exp(\frac{\pi k}{2w_j})} \\ &= \frac{2\pi i}{|w_j|} \exp\left(\frac{ib_j k}{w_j}\right) \exp\left(-\frac{\pi k}{2w_j}\right) \end{aligned} \tag{1.2}$$

可以发现随着频率的增加，tanh 在频率空间是指数衰减的。它在傅里叶空间快速衰减的主要原因是因为它在时域空间是一个光滑、无穷阶可导的函数。对 $h(x)$ 做连续傅里叶变换，自然就可以得到：

$$\hat{h}(k) \approx \sum_{j=1}^{m} a_j \exp\left(\frac{ib_j}{w_j}\right) \exp\left(-\left|\frac{\pi k}{2w_i}\right|\right) \tag{1.3}$$

定义某个频率上的损失函数：

$$L(k) = \frac{1}{2}|\hat{h}(k) - \hat{f}(k)|^2 \tag{1.4}$$

那么总的损失函数就是所有频率相加，也就是求积分。根据 Parseval' s theorem，频域下的损失函数应该和时域上的损失函数具有相同的能量，我们有：

$$L = \int L(k)dk = \int \frac{1}{2}|f(x) - h(x)|^2 dx \tag{1.5}$$

对网络使用梯度下降法训练，我们有：

$$\theta^{(n+1)} \leftarrow \theta^{(n)} - \eta \sum \frac{\partial L(k)}{\partial \theta} \tag{1.6}$$

其中 η 是学习率，下面考察 $\frac{\partial L(k)}{\partial \theta}$：

$$\left|\frac{\partial L(k)}{\partial \theta}\right| \approx |\hat{h}(k) - \hat{f}(k)| \exp\left(-\left|\frac{\pi k}{2w_i}\right|\right) G(\theta, k) \tag{1.7}$$

其中 $G(\theta, k)$ 是一个 $O(1)$ 的函数，从上面的式子可以看出，如果低频没有收敛，也就是 $A(k) = |\hat{h}(k) - \hat{f}(k)| > 0$，而且参数 w_j 比较小时，$\exp\left(-|\pi k/2w_j|\right)$ 占主导，低频对梯度造成的贡献远大于高频，也就是低频成分主导了梯度流，梯度下降的方向实际上就是更接近低频收敛的方向。所以从上面的理想模型可以获得一个重要理解：深度学习对于低频的偏向性的来源是由于激活函数的光滑性与正则性，以及基于梯度下降的训练方法造成的。类似地对于 ReLU，也可以证明是多项式衰减的。

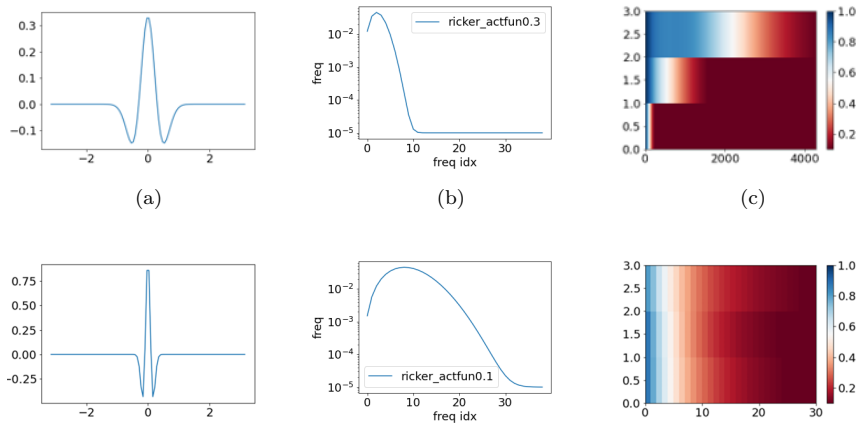
上面的分析展示了激活函数在频率原则中的重要性。对于大部分的激活函数，例如 Tanh 和 ReLU，他们在频率空间单调下降。因此，我们可以容易地观察到频率原则。我们可以设计一类在频率空间不是单调下降的函数，这类函数在频率空间先是单调上升直到某个高频，然后再单调下降。在这种情况下，我们可以观察到相反的频率收敛顺序。我们用带参数 a 的 Ricker 函数作为激活函数：

$$\frac{1}{15a}\pi^{1/4}\left(1 - \left(\frac{x}{a}\right)^2\right)\exp\left(-\frac{1}{2}\left(\frac{x}{a}\right)^2\right). \tag{1.8}$$

当 a 比较小时, Ricker 函数从一个更高的频率开始误差。我们用一个全连接网络来拟合 $\sin(x) + \sin(3x) + \sin(5x)$。正如图**??**所展示的，在第一行，当激活函数从一个比较低的频率开始衰减时，我们可以明显地观察到低频收敛得更快；然而，在第二行，我们调节 a，使激活函数从一个比较高的频率开始衰减，这时，我们无法观察到哪个频率可以明显地收敛更快。这些现象和我们前面的分析是一致，也就是激活函数在频率空间的行为影响了频率的收敛。

若我们将上面的理想模型写成严格的数学表述 (下面的定理，一般读者可以忽略)，就有

Theorem 1. *考虑一个以 $\sigma(x) = \tanh(x)$ 作为激活函数的两层的 DNN, 任取两个频率 $\forall k_1, k_2$, 使得 $\left|\hat{f}(k_1)\right| > 0, \left|\hat{f}(k_2)\right| > 0$，$|k_2| > |k_1| > 0$，一定存在正的常数 c 和 C，使得对足够小的*



(a) (b) (c)

(d) activation in spatial domain (e) activation in frequency domain (f) convergence order

图 1.1: Ricker 激活函数. 第一行，$a = 0.3$ ； 第二行，$a = 0.1$。

δ，有

$$\frac{\mu\left(\left\{W : \left|\frac{\partial L(k_1)}{\partial \theta_{lj}}\right| > \left|\frac{\partial L(k_2)}{\partial \theta_{lj}}\right| \quad for\ all \quad l, j\right\} \cap B_\delta\right)}{\mu\left(B_\delta\right)} \geq 1 - C\exp(-c/\delta) \tag{1.9}$$

其中 $B_\delta \subset \mathbb{R}^m$ 是一个球心在原点，半径为 σ 的圆球，$\mu(\cdot)$ 是勒贝格测度。

也就是说对于任意两个都没收敛的 k_1,k_2，可以证明只要我们的参数在某个球面内，当球的半径很小，也就是参数比较小的情况下，低频的梯度大于高频的梯度的测度比上整个空间的测度是以指数收敛的形式趋近于 1 的。换句话说，当 $\delta \to 0$ 的时候，低频对应的梯度一定是大于高频对应的梯度的。

证明. 不失一般性，假定 $k_1, k_2 > 0, b_j > 0, w_j \neq 0, j = 1, \cdots, m$, 由于频域中损失函数的梯度:

$$\frac{\partial L(k)}{\partial a_j} = \frac{2\pi}{w_j} \sin\left(\frac{b_j k}{w_j} - \phi(k)\right) E_0 \tag{1.10}$$

$\left|\frac{\partial L(k_1)}{\partial a_j}\right| \leq \left|\frac{\partial L(k_2)}{\partial a_j}\right|$ 可以被等价地表示为

$$\frac{A\left(k_2\right)}{A\left(k_1\right)} \left|\frac{\exp\left(\frac{\pi k_1}{2w_j}\right) - \exp\left(-\frac{\pi k_1}{2w_j}\right)}{\exp\left(\frac{\pi k_2}{2w_j}\right) - \exp\left(-\frac{\pi k_2}{2w_j}\right)}\right| \cdot \left|\sin\left(\frac{b_j k_2}{w_j} - \phi\left(k_2\right)\right)\right| \geq \left|\sin\left(\frac{b_j k_1}{w_j} - \phi\left(k_1\right)\right)\right| \tag{1.11}$$

又因为 $k > 0$ 时有 $|\hat{h}(k)| \leq C\sum_{j=1}^{m} \frac{|a_j|}{|w_j|} \exp\left(-\frac{\pi k}{2|w_j|}\right) (k > 0)$，所以

$$\lim_{W \to 0} \hat{h}(k) = 0 \quad , \lim_{W \to 0} D(k) = -\hat{f}(k) \tag{1.12}$$

所以

$$\lim_{W\to 0} A(k) = |\hat{f}(k)|, \quad \lim_{W\to 0} \phi(k) = \pi + \arg(\hat{f}(k)) \tag{1.13}$$

当　足够小时，对于球中任意一点 $W \in B_\delta$, 有 $A(k_1) > \frac{1}{2}\left|\hat{f}(k_1)\right| > 0$,$A(k_2) < 2\left|\hat{f}(k_2)\right|$，又 $\left|\sin\left(\frac{b_j k_2}{w_j} - \phi(k_2)\right)\right| \le 1$，所以对于足够小的 ，有

$$\left|\frac{\exp\left(\frac{\pi k_1}{2w_j}\right) - \exp\left(-\frac{\pi k_1}{2w_j}\right)}{\exp\left(\frac{\pi k_2}{2w_j}\right) - \exp\left(-\frac{\pi k_2}{2w_j}\right)}\right| \le 2\exp\left(\frac{-\pi(k_2-k_1)}{2|w_j|}\right) \tag{1.14}$$

又由式 **??**，可以得到

$$\left|\sin\left(\frac{b_j k_1}{w_j} - \phi(k_1)\right)\right| \le \frac{8\left|\hat{f}(k_2)\right|}{\left|\hat{f}(k_1)\right|}\exp\left(-\frac{\pi(k_2-k_1)}{2|w_j|}\right) \tag{1.15}$$

由于 $\frac{2}{\pi}|x| \le |\sin x|\left(|x| \le \frac{\pi}{2}\right)$，结合式 **??**, 有

$$\left|\frac{b_i k_1}{w_i} - \arg\left(\hat{f}(k_1)\right) - q\pi\right| \le \frac{8\pi\left|\hat{f}(k_2)\right|}{\left|\hat{f}(k_1)\right|}\exp\left(-\frac{\pi(k_2-k_1)}{2\delta}\right) \tag{1.16}$$

$$\Rightarrow -c_1\exp(-c_2/\delta) + q\pi + \arg\left(\hat{f}(k_1)\right) \le \frac{b_i k_1}{w_i} \le c_1\exp(-c_2/\delta) + q\pi + \arg\left(\hat{f}(k_1)\right)$$

其中 $c_1 = \frac{8\pi\left|\hat{f}(k_2)\right|}{\left|\hat{f}(k_1)\right|}$，$c_2 = \pi(k_2 - k_1)$, 定义 $I := I^+ \cup I^-$ 其中

$$I^+ := \{w_j > 0 : W \in S_{1j,\delta}\}, \quad I^- := \{w_j < 0 : W \in S_{1j,\delta}\} \tag{1.17}$$

对于 $w_j > 0$，有

$$0 < \frac{b_j k_1}{c_1\exp(-c_2/\delta) + q\pi + \arg\left(\hat{f}(k_1)\right)} \le w_j \le \frac{b_j k_1}{-c_1\exp(-c_2/\delta) + q\pi + \arg\left(\hat{f}(k_1)\right)} \tag{1.18}$$

又因为 $W \in B_\delta$,$c_1\exp(-c_2/\delta) + \arg\left(\hat{f}(k_1)\right) \le 2\pi$，所以 $\frac{b_j k_1}{2\pi + q\pi} \le w_j \le \delta$,**??** 只对比较大的 q 成立，所以对 I^+ 的勒贝格测度进行估计，有:

$$\mu\left(I^+\right) \le \sum_{q=q_0}^{\infty} \left|\frac{b_j k_1}{-c_1\exp(-c_2/\delta) + q\pi + \arg\left(\hat{f}(k_1)\right)} - \frac{b_j k_1}{c_1\exp(-c_2/\delta) + q\pi + \arg\left(\hat{f}(k_1)\right)}\right| \tag{1.19}$$

$$\le 2|b_j| k_1 c_1\exp(-c_2/\delta)\sum_{q=q_0}^{\infty} \frac{1}{\left(q\pi + \arg\left(\hat{f}(k_1)\right)\right)^2 - \left(c_1\exp(-c_2/\delta)\right)^2} \tag{1.20}$$

$$\le C\exp(-c/\delta) \tag{1.21}$$

□

这部分理论不仅给我们呈现了为什么会有频率原则，同时也能告诉我们在哪些情况下频率原则会不明显，甚至不成立。下面讲两种情况，一种是初始的权重很大的情况，另一种是当损失函数中含有神经网络的输出关于输入的导数的情况。

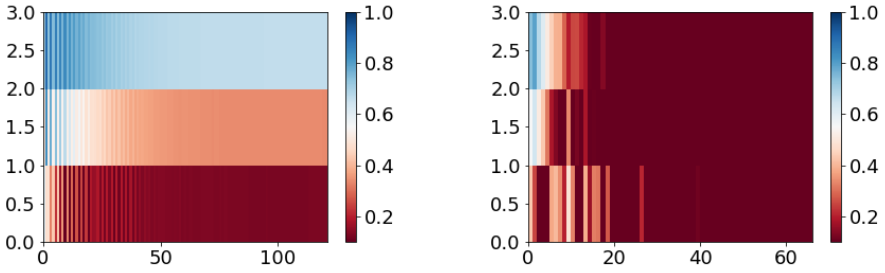
1.1.2 损失函数中包含的频率权重

损失函数的形式可以显著地影响频率的收敛。例如，我们可以显式地在损失函数中给一些特定的频率加很大的权重，这可以会使这个频率的收敛加快。我们可以考虑两种损失函数，一种是普通的均方损失 L_{nongrad}，另一种是在 L_{nongrad} 上额外添加一个导数的损失，

$$L_{\text{nongrad}} = \sum_{i=1}^{n} (f_\theta(x_i) - f^*(x_i))^2/n, \tag{1.22}$$

$$L_{\text{grad}} = L_{\text{nongrad}} + \sum_{i=1}^{n} (\nabla_x f_\theta(x_i) - \nabla_x f^*(x_i))^2/n. \tag{1.23}$$

同样，我们用神经网络拟合 $\sin(x) + \sin(3x) + \sin(5x)$。正如图**??**所展示的，当添加导数的损失后，高频的收敛更快。关键的原因是在频率空间，导函数的傅里叶变换等于原函数的傅里叶变换乘以频率，这等价于对高频施加更大的权重。



(a) L_{nongrad} (b) L_{grad}

图 1.2: 在损失函数 L_{nongrad} 和 L_{grad} 下，各个主要频率的收敛速度。

1.1.3 激活函数和损失函数的联合作用

前面的分析表明频率的收敛行为是受到激活函数和损失函数的联合作用的。对于一般的激活函数，频率原则可以被容易地观察到。但在一些特殊设定的任务中，例如解微分方程，损失函数经常包含梯度信息，频率原则可能被弱化或者消失。

1.2 初始化权重很大时

我们前面讨论的范畴都是在初始化权重比较小的范畴下，那么初始化权重比较大的时候呢？考虑对于 tanh 来说，从前面的理论来看，当权重值很大的时候，指数控制那一项 $\exp\left(-\frac{\pi k}{2w_j}\right)$ 几乎不随频率发生变化，此时低频的优势自然就少了。从实域来看，以图 **??**为例，激活函数在 $x=0$ 处关于输入是一个更快变化的函数，高频成分增多。当神经网络由这样的激活函数组合而成时，其初始化的输出就有更多高频成分。而当训练集的高频不足以抑制神经网络在初始化时所具有的高频时，在训练完成后，神经网络仍然有很多高频成分，这就会对网络泛化性能产生影响：

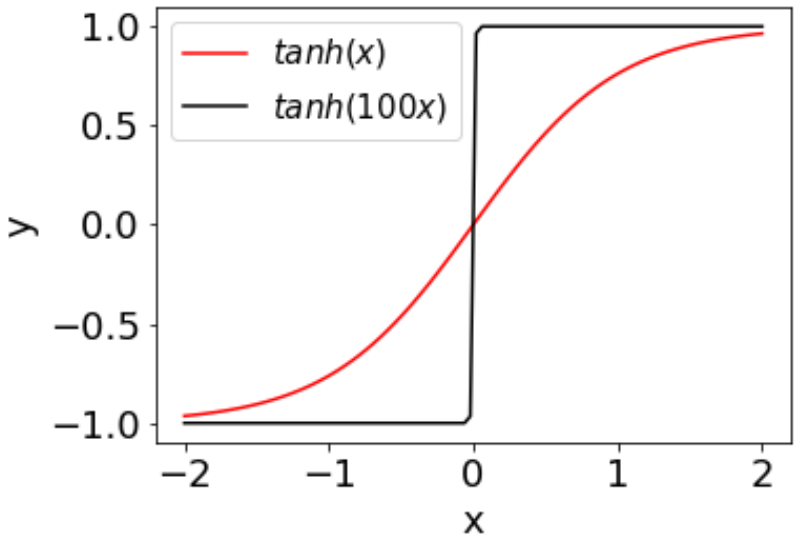


图 1.3: 初始化权重较大与初始化权重较小的情况下拟合一个函数的结果

1.2.1 一维实验

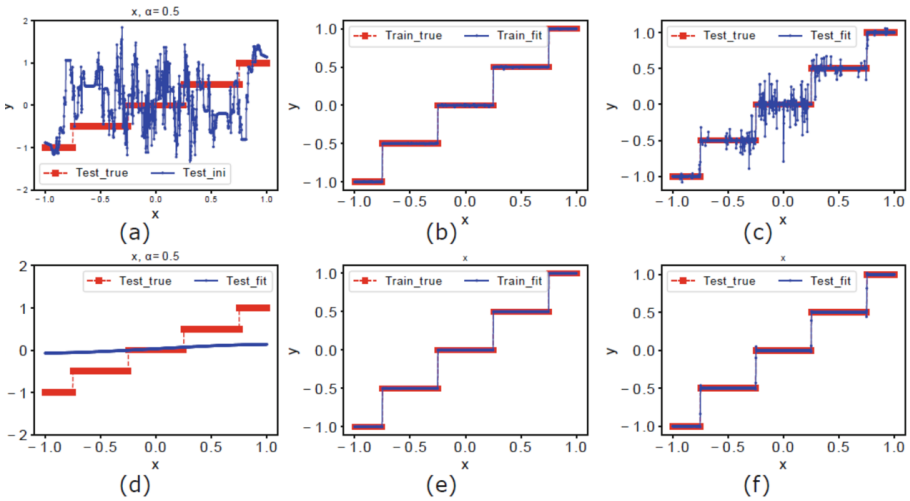


图 1.4: 初始化权重较大与初始化权重较小的情况下拟合一个函数的结果。

如图??所示，如果初始化权重设的太大，训练过后依然有比较不明显的振荡，而在测试集上可以看出泛化性很差。当初始化权重比较小时，由于 DNN 不会拟合超过数据集最高频率的信息，所以最终的泛化误差比较理想。但是如果初始化权重设置得太大，就会在初始状态下引入很多高频的信息，有些甚至比数据集中的高频信息还要高。又因为 DNN 不会学习比数据集频率更高的信息，也就意味着训练的过程没有能力抵消初始化带来的高频。在这种情况下，虽然在训练集上仍有高度的准确性，但是这些高频信息对泛化性能产生了很大的影响，其输出中就夹杂了很多超过数据集最高频率的高频成分。

1.2.2 二维实验

对于二维、真实的数据，同样也有这样的性质。如图??所示，我们只用一奇数列的像素点，采用大的初始化值来进行训练，把所有像素点放进去测试的时候，发现结果已经是面目全非。将 x 方向的结果绘制出来，可以发现训练的结果好像并没有什么不对，但是测试的时候却输出了高频振荡。产生这个现象的原因和上面的一维实验差不多，测试时输出的高频成分来源于初始化时引入的高频，在训练过程中 DNN 并没有学到比数据集最高频率更高的信息，所以初始化引入的高频就被一直保留到最后，从而对泛化能力产生不利的影响。

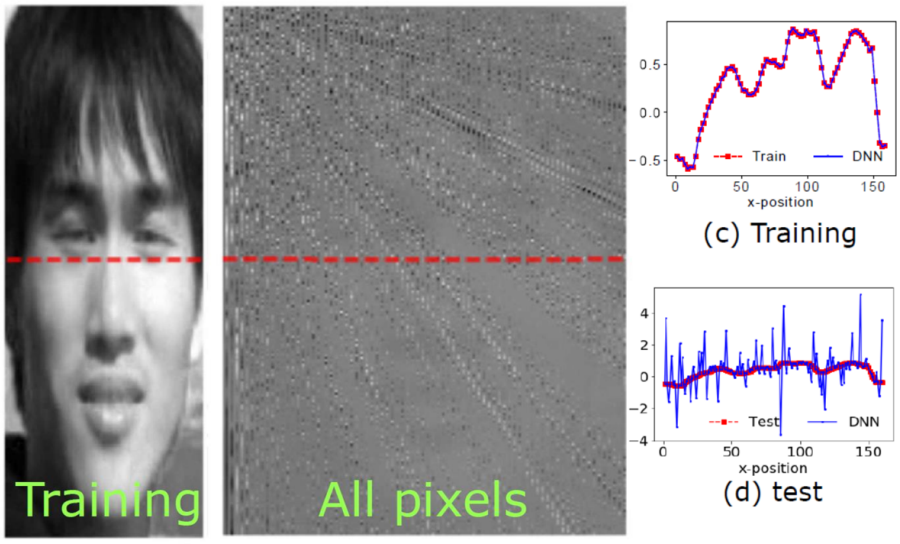


图 1.5: 初始化权重较大的二维图片拟合实验，取奇数行列的数据进行训练，并取剩下的数据作为测试，可以发现产生了糟糕的结果，这主要是由初始化的高频成分带来的振荡。

1.3 线性频率原则

1.3.1 基本 LFP 模型

受频率原则的启发，我们提出了一个有效的模型来对以 ReLU 为激活函数的两层宽神经网络的动力学行为进行建模，将这一模型称之为线性频率原则（LFP）。

相关工作?? 表明，如果神经网络每层神经元个数足够多，给定合适的参数初始化，使神经网络处于线性区域?，比如 NTK 区域，那么对任何 $t \geq 0$，有 $\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(0)\| \ll 1$，这里 $\boldsymbol{\theta}(t)$ 表示 t 时刻神经网络的参数。因为参数变化相当小，神经网络 t 时刻的输出 $h(\boldsymbol{x}, \boldsymbol{\theta}(t))$ 可以被 $\theta(0)$ 处的一阶泰勒展开逼近，这意味着以下表达式将是 DNN 的输出 $h(\boldsymbol{x}, \boldsymbol{\theta}(t))$ 的一个很好的逼近（$\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$）：

$$h_{\text{lin}}(\boldsymbol{x}, \boldsymbol{\theta}) = h(\boldsymbol{x}, \boldsymbol{\theta}_0) + \nabla_{\boldsymbol{\theta}} h\left(\boldsymbol{x}, \boldsymbol{\theta}_0\right) \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0), \tag{1.24}$$

这里我们假设 h 有一些比较好的性质：对任何 $\boldsymbol{\theta} \in \mathbb{R}^m$，函数 $h(\cdot, \boldsymbol{\theta}_0)$ 均具有弱导数 $\nabla_{\boldsymbol{\theta}} h(\cdot, \boldsymbol{\theta}_0) \in L^2(\mathbb{R}^d)$。这个性质对于一般的 DNN 而言都是很好满足的。

考察两层神经网络，我们可以把两层神经网络表示为

$$h(\boldsymbol{x}, \boldsymbol{\theta}(t)) = \sum_{j=1}^{m} a_j \sigma(\boldsymbol{w}_j \cdot \boldsymbol{x} + b_j), \tag{1.25}$$

其中 σ 为激活函数，$\boldsymbol{w}_j$ 是输入权重，a_j 是输出权重，b_j 是偏置项。

根据之前提到的线性假设以及偏置项的初始化方差足够大的情况下，?? 推导出了线性频率原则（LFP）来刻画两层宽神经网络的动力学行为，其中我们假设 DNN 的损失函数为均方损失。由 LFP 表征的 DNN 动力学模型如下：

$$\partial_t \mathcal{F}[u](\boldsymbol{\xi}, t) = -(\gamma(\boldsymbol{\xi}))^2 \mathcal{F}[u_\rho](\boldsymbol{\xi}), \tag{1.26}$$

其中 $u(\boldsymbol{x}, t) = h(\boldsymbol{x}, \boldsymbol{\theta}(t)) - f^*(\boldsymbol{x})$, $u_\rho(\boldsymbol{x}, t) = u(\boldsymbol{x}, t)\rho(\boldsymbol{x})$, $\rho(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n \delta(\boldsymbol{x} - \boldsymbol{x}_i)$，这里我们假设训练集为 n 个数据点 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$。权重项 $\gamma(\boldsymbol{\xi})$ 依赖于激活函数、参数初始化值和频率值，对于以 ReLU 为激活函数的 DNN，$\gamma(\boldsymbol{\xi})$ 有以下相对简单的形式：

$$(\gamma(\boldsymbol{\xi}))^2 = \mathbb{E}_{a(0), r(0)} \left[\frac{r(0)^3}{16\pi^4 \|\boldsymbol{\xi}\|^{d+3}} + \frac{a(0)^2 r(0)}{4\pi^2 \|\boldsymbol{\xi}\|^{d+1}} \right],$$

其中 $r(0) = |\boldsymbol{w}(0)|$，并且初始化参数 $a(0)$ 和 $\boldsymbol{w}(0)$ 是按照给定的分布随机生成的。$\boldsymbol{\xi}$ 有两个指数，分别是 $d+3$ 和 $d+1$. 第一项的是因为对 ReLU（对 a 求导留下的部分）做傅里叶变换，第二项是因为对 ReLU 的导数（对 w, b 求导留下的部分）做傅里叶变换。

公式??中所示的动力学模型表明，低频部分要比高频部分收敛速度要快。因此 DNN 更加倾向于获得低频系数较大的函数。

这个动力学方程的左右两端的 u 差一个采样密度。左端的 u 是整个神经网络函数，而右端因为乘以了采样密度，使得其只受训练点的影响。这符合神经网络的训练过程，即训练目标着眼于将训练点拟合好。

?? 推导了与 LFP 等价的变分模型：

$$\min_{h - h_{\text{ini}} \in F_\gamma} \int_{\mathbb{R}^d} (\gamma(\boldsymbol{\xi}))^{-2} |\mathcal{F}[h](\boldsymbol{\xi}) - \mathcal{F}[h_{\text{ini}}](\boldsymbol{\xi})|^2 \, \mathrm{d}\boldsymbol{\xi},$$

并且 $h(\boldsymbol{x})$ 需要额外满足限制条件 $h(\boldsymbol{x}_i) = y_i$, $i = 1, \ldots, n$。当频率 $\boldsymbol{\xi}$ 增大时，权重项 $(\gamma(\boldsymbol{\xi}))^{-2}$ 也会随之增大，这意味着对于 $h(\boldsymbol{x}) - h_{\text{ini}}(\boldsymbol{x})$ 中的高频项施加了更大的惩罚。根据这个分析也可以得出 $h(x) - h_{ini}(x)$ 将会是一个低频函数。但是我们注意到如果 $h_{ini}(x)$ 不是 0 的话，那么 $h(x)$ 将会含有 $h_{ini}(x)$ 中的高频成分。在论文 (?) 中，提出一种反对称初始化的技术，可以使 DNN 的初始输出 $h_{ini}(x)$ 为 0。这样，DNN 最后的输出将是一个低频函数，从另一个角度得出了频率原则的结论。

在数值上，我们通过解决以下问题来解决 LFP 模型 [1]

$$\min_{a_n, b_n} \sum_{i=1}^{M} \left(\sum_{j \in I} \left[a_j \sin\left(2\pi \frac{j}{L'} x_i\right) + b_j \cos\left(2\pi \frac{j}{L'} x_i\right) \right] - y_i \right)^2 + \varepsilon \sum_{j \in I} \gamma\left(2\pi \frac{j}{L'}\right)^{-2} \left(a_j^2 + b_j^2\right), \tag{1.27}$$

[1] 代码可见 https://github.com/xuzhiqin1990/LFP

其中，我们令 $I=\left\{0,\cdots,\frac{L'}{L}K-1\right\}$, $L'=10L$, L 是训练输入的范围, $K=200$ 远远大于训练样本的数量, $\varepsilon=10^{-6}$. 记 $M_I=\frac{L'}{L}K-1$. 我们可以将上述问题改写为向量形式：

$$\min_{\boldsymbol{a}}\left(\boldsymbol{E}\boldsymbol{a}-\boldsymbol{Y}\right)^{\intercal}\left(\boldsymbol{E}\boldsymbol{a}-\boldsymbol{Y}\right)+\varepsilon\boldsymbol{a}^{\intercal}\boldsymbol{W}^{-1}\boldsymbol{a}, \tag{1.28}$$

其中

$$\boldsymbol{a}=[a_0,\cdots,a_{M_I},b_0,\cdots,b_{M_I}]^{\intercal},$$

$$\boldsymbol{E}=\left[\sin\left(2\pi\frac{0}{L'}\boldsymbol{X}\right),\cdots,\sin\left(\frac{2\pi}{L'}M_I\boldsymbol{X}\right),\cos\left(2\pi\frac{0}{L'}\boldsymbol{X}\right),\cdots,\cos\left(\frac{2\pi}{L'}M_I\boldsymbol{X}\right)\right],$$

$$\boldsymbol{X}=[x_1,\cdots,x_M]^{\intercal},\quad \boldsymbol{Y}=[y_1,\cdots,y_M]^{\intercal},$$

$$\boldsymbol{W}^{-1}=\operatorname{diag}\left\{\gamma\left(2\pi\frac{0}{L'}\right)^{-2},\cdots,\gamma\left(2\pi\frac{1}{L'}\left(\frac{L'}{L}K-1\right)\right)^{-2}\right\}.$$

上述问题的解满足

$$\boldsymbol{E}^{\intercal}\left(\boldsymbol{E}\boldsymbol{a}-\boldsymbol{Y}\right)+\varepsilon\boldsymbol{W}^{-1}\boldsymbol{a}=0. \tag{1.29}$$

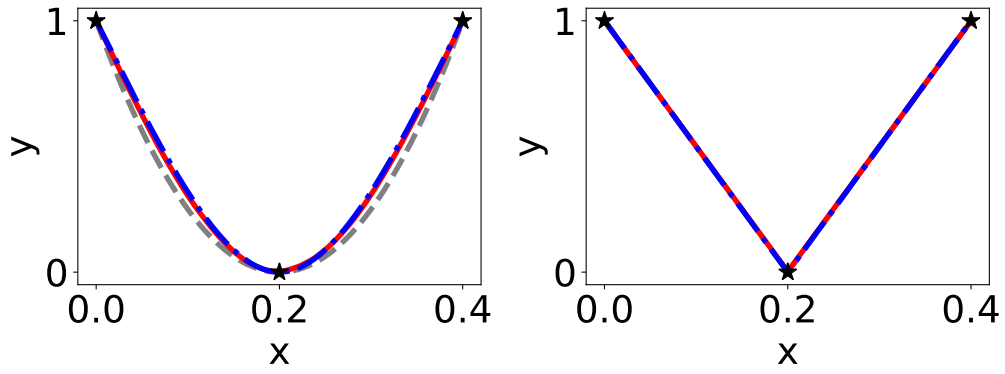
求解 $\boldsymbol{a}$ 我们有，

$$\boldsymbol{a}=\left[\boldsymbol{E}^{\intercal}\boldsymbol{E}+\varepsilon\boldsymbol{W}^{-1}\right]^{-1}\boldsymbol{E}^{\intercal}\boldsymbol{Y}. \tag{1.30}$$

对于一维输入示例，我们使用具有 10000 隐藏神经元和不同初始化的 ReLU 神经网络。如图 **??**(a) 所示，我们取三个数据点用作训练数据集，所有参数的初始化均从均值为零的均匀分布中采样，用 $[-U,U]$ 表示采样范围。为了使 $1/\xi^4$ 项占主导地位，我们设置 $\boldsymbol{w}$ 初始化分布 $U=3$，a 初始化分布 $U=0.01$，偏置项初始化分布 $U=3$。值得注意的是，偏差项初始化值不会很大。如图 **??**(a) 所示，神经网络通过平滑函数（用 f_{NN} 表示，红色实线）对训练数据进行插值，这与 LFP 模型的预测（用 f_{LFP} 表示，蓝色虚线）和三次样条插值（灰色虚线）几乎重叠。相反，为了使 $1/\xi^2$ 项占主导地位，我们设置 $\boldsymbol{w}$ 初始化分布 $U=0.1$，a 初始化分布 $U=2$，偏置项初始化分布 $U=2$。如图 **??**(b) 所示，NN 通过函数对训练数据进行插值，这与 LFP 模型的预测和线性样条插值几乎重叠。

1.3.2 高频的收敛极限

LFP 模型中，每个频率收敛的速度依赖于激活函数，具体在模型中，体现在 $\|\xi\|$ 的指数会不一样。这些指数决定了各个频率成分收敛的快慢。于是，从 LFP 模型中我们可以抽象出



(a) (b)

图 1.6: f_{NN} (红色实线) vs. f_{LFP} (蓝色点线) vs. 样条 (灰色虚线, 其中 (a) 为三次样条, (b) 为线性样条), 每张图中所有的曲线都几乎重合。我们考察宽度为 10000 的两层 ReLU 神经网络, 初始化为: (a) $\langle r^2\rangle_r \gg \langle a^2\rangle_a$, (b) $\langle r^2\rangle_r \ll \langle a^2\rangle_a$。图中黑星表示训练点。

一个简单的监督学习的模型,

$$\min_{h\in\mathcal{H}} Q_\alpha[h] = \int_{\mathbb{R}^d} \langle\boldsymbol{\xi}\rangle^\alpha \left|\mathcal{F}[h](\boldsymbol{\xi})\right|^2 \mathrm{d}\boldsymbol{\xi}, \tag{1.31}$$

$$\text{s.t.} \quad h(\boldsymbol{x}_i) = y_i, \quad i = 1, \cdots, n, \tag{1.32}$$

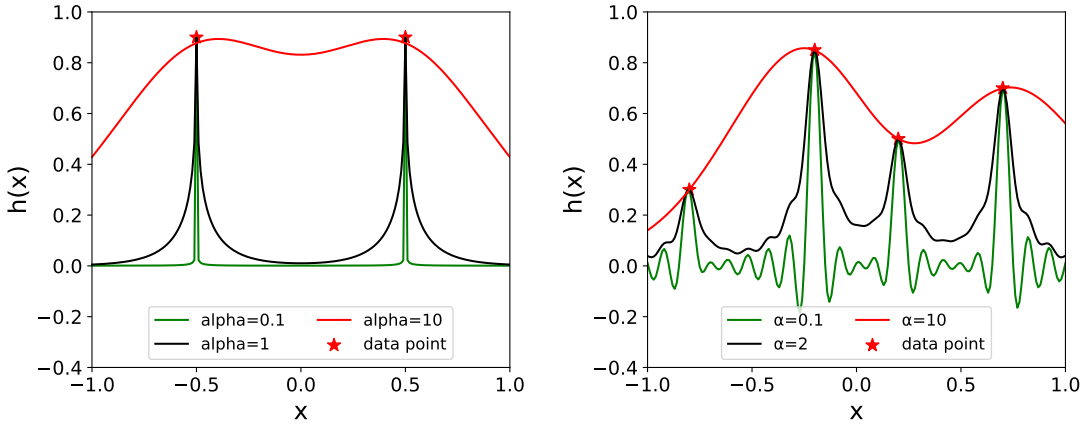
其中 $\mathcal{H} = \{h(x) | \int_{\mathbb{R}^d} \langle\boldsymbol{\xi}\rangle^\alpha \left|\mathcal{F}[h](\boldsymbol{\xi})\right|^2 \mathrm{d}\boldsymbol{\xi} < \infty\}$. 根据 **?** 中证明的等价定理, $-\alpha$ 就是梯度流动力学中关于频率的衰减率。α 越小, 收敛就越快。一个自然的问题是 α 最小可以是多少? 有没有一个临界值? 如果超过这个临界值, 拟合会发生什么情况?

? 研究了这个临界问题。研究发现, 对于数据维度为 d 的情况, 问题 (**??**) 中的 $\alpha = d$ 是一个临界情况。当 $\alpha > d$ 时, 拟合的结果是一个比较连续光滑的解, 而当 $\alpha < d$ 时, 拟合的结果是一种类似 Dirac 函数的形状, 也就是在训练点, 其数值与真实值一样, 而在非训练点, 其值为 0。下面展现两个非常的例子来验证这个临界现象。

假设有 n 个数据点: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, 每个数据点是 d 维的, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^\intercal$。数据对应的标签记为 $(y_1, y_2, \ldots, y_n)^\intercal$。为了符号简单, 我们记 $(j_1, j_2, \cdots, j_d)^\intercal$ 为 $\boldsymbol{J}_{j_1\ldots j_d}$。优化模型可以变为

$$\min_{\phi\in\mathbb{R}^{(2M)^d}} \sum_{j_1,\ldots,j_d=-M}^{M} (1 + \|\boldsymbol{J}_{j_1\ldots j_d}\|^2 \Delta\xi^2)^{\frac{\alpha}{2}} \left|\phi_{j_1\ldots j_d}\right|^2, \tag{1.33}$$

$$\text{s.t.} \sum_{j_1,\ldots,j_d=-M}^{M} \phi_{j_1\ldots j_d} \mathrm{e}^{2\pi\mathrm{i}\Delta\xi \boldsymbol{J}_{j_1\cdots j_d}^\intercal \boldsymbol{x}_k} = y_k, \quad k = 1, 2, \ldots, d \tag{1.34}$$



(a) 2 points in 1 dimension

(b) 20 points in 2 dimension

图 1.7: 用不同的衰减率 α 拟合数据。

这个问题的求解实际上是岭回归。在图**??**中，对一维的例子，我们设置 $M = 1000$，对二维的例子，我们设置 $M = 100$。通过改变 α，我们都能够看到在拟合训练数据点（星星）时，当 $\alpha < d$ 时，插值的解几乎没有意义，而当 $\alpha > d$ 时，插值的解看起来是有意义的解。

这些理论表明，对于一个有意义的插值，α 不能太小，也就是收敛的速度有一个依赖维度的上限。而且这个收敛率大于数据的维度，因此，当维度越来越大时，各个频率的收敛速度也会越来越慢，产生了在收敛时间上关于频率的维数灾难。

1.3.3 线性区域下的理论简述

在我们工作的同时，几项并行工作也分析了 NTK 区域中的频率原则（或谱偏差）。**?** 和 **?** 在假设有足够大的训练数据均匀分布在超球面上的前提下，估计 NTK 区域中两层宽 ReLU 网络的每个频率的收敛速度。**?** 将数据分布的假设放宽为非均匀分布，该假设仅限于二维球面，他们为 NTK 区域中的两层宽 ReLU 网络得出了类似的频率偏差。**?** 研究谱偏差对样本量的依赖性。其他几项工作也侧重于研究 NTK 区域中 Gram 矩阵的谱 **??**。在这项工作中，我们对线性频率原理动力学的精确推导没有对训练数据的分布和大小做出任何假设。在 LFP 模型需要假设偏置项的分布方差比较大，而在我们的其它分析中不需要这个假设

关于文章**?** 中的图 1，在一段完全没有数据点的区间，神经网络也能学出非常漂亮的正弦函数的振荡，正常情况，按我自己的经验，这不应该出现，因为平滑没有振荡地连接两点（不太大的初始化）以及非常混乱地振荡地连接两点（比较大的初始化）才是通常出现的。该文章也表达了神经网络的神奇！我尝试重现很久都无法得到文章中的结果，只好向作者要源代码。

后来发现，作者使用的模型是没有偏置项的网络，并且在输入做了一点小变换。当我把自己的代码调成一样的设定后，确实能出现文章中的现象，但这种现象极容易因为加深网络或者添加偏置项而消失。

1.4 一般神经网络的频率原则理论

??节的理论在推广到高维时，傅里叶变换会有困难，而且对一般的激活函数，其傅里叶变换可能很难精确地写出来。而线性区域下研究频率收敛速度的理论又依赖于线性化，对于一般的中度过参数化的网络并不适用。对于有明显非线性的训练过程，其 Gram 矩阵的特征向量是一直在变的。虽然特征值可以用来刻画神经网络在相应特征向量的成分的收敛速度，但这些特征向量一直在变，使得我们基本不知道到底是哪些成分在收敛。这个问题在线性区域不存在的原因是 Gram 矩阵在线性网络的训练过程中几乎不变，所以其对应的特征值和特征向量均不变，也可以证明它他们频率几乎一一对应。因此，对于一般的神经网络，我们要理解它的训练过程就需要找一组具有明确意义并且不依赖训练过程的特征。大量的实验发现频率是一个很好的量。事实上，在很多科学研究中，频率常常用来刻画被研究的系统，因为它能够被测量，也较容易做理论分析，比如研究波动力学的时候，常常把研究的波分解成不同频率的模态。

对于一般的神经网络的频率原则分析，要精确地做傅里叶变换是困难的，主要有两个原因，一个是输入是高维的，另一个是函数的复合。退而求其次，我们可以定性分析。基本的思路是利用函数的正则性和傅里叶系数的衰减关系。这里的正则性可以理解为函数可以做几阶层数，而傅里叶系数的幅度通过随着频率增加会减小。我们不加演算地来举几个例子。我们要利用函数与其导数在频率空间的关系

$$\mathcal{F}[\partial f/\partial x](\xi) = -2\pi j\xi\mathcal{F}[f](\xi). \tag{1.35}$$

Dirac 函数的定义为

$$\delta(x) = \begin{cases} 0, & x \neq 0 \\ \infty, & x = 0 \end{cases}, \tag{1.36}$$

和

$$\int_{-\infty}^{\infty} \delta(x)dx = 1. \tag{1.37}$$

并且对任意函数

$$f(x) = \int f(y)\delta(y-x)dy. \tag{1.38}$$

考虑函数 $\exp(-2\pi jx\xi)$,

$$1 = \exp(-2\pi jx\xi)|_{x=0} = \int \exp(-2\pi jy\xi)\delta(y)dy. \tag{1.39}$$

上式的右端实际上就是 Dirac 函数的傅里叶变换。因此，Diract 的傅里叶变换是一个常数。这和量子力中学的不确定原理有很深的联系。在实域空间，Dirac 函数只在一个点有值，可以理解它的分辨率是无穷细的，而它在傅里叶空间则是处处一样，完全无法通过函数值来区分其对应的频率，所以没有任何分辨率。总结一下，我们需要的性质是，对于 Dirac 函数，它在频率空间的幅度不随频率衰减。

接下来，我们来看 Heaviside 函数，

$$H(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}, \tag{1.40}$$

Heaviside 函数和 Dirac 函数有导数的关系

$$\delta(x) = H'(x). \tag{1.41}$$

因此

$$\mathcal{F}[\delta(x)](\xi) = -2\pi j\xi \mathcal{F}[H(x)](\xi), \tag{1.42}$$

于是有

$$\|\mathcal{F}[H(x)](\xi)\| = \frac{C}{\|\xi\|}, \tag{1.43}$$

我们用 C 表示常数，为了简便，不同常数，我们用同一个 C 表示。再考虑 $\mathrm{ReLU}(x) = \max\{0, x\}$，满足 $H(x) = \mathrm{ReLU}'(x)$。类似地，我们可以得到

$$\|\mathcal{F}[\mathrm{ReLU}(x)](\xi)\| = \frac{C}{\|\xi\|^2}. \tag{1.44}$$

依此类推，我们可以得到当一个函数可以求导的次数越高（正则性），它在频率空间的幅度随频率增长衰减得更快。如果一个函数无穷可导，那这将是指数衰减。

在定性上，我们可以只关注频率幅度的衰减率，利用梯度下降，这些衰减会对梯度产生影响，从而对损失函数的不同频率成分的收敛产生影响。对于函数的复合，我们只需要估计复合函数的正则性即可。**?** 严格发展这些理论，并且将损失函数的类型推广到 L^p 损失函数 $(p \geq 2)$.

参考文献

Arora, S., Du, S. S., Hu, W., Li, Z. and Wang, R. (2019), 'Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks', *arXiv preprint arXiv:1901.08584* .

Basri, R., Galun, M., Geifman, A., Jacobs, D., Kasten, Y. and Kritchman, S. (2020), 'Frequency bias in neural networks for input of non-uniform density', *arXiv preprint arXiv:2003.04560* .

Basri, R., Jacobs, D., Kasten, Y. and Kritchman, S. (2019), 'The convergence rate of neural networks for learned functions of different frequencies', *arXiv preprint arXiv:1906.00425* .

Bordelon, B., Canatar, A. and Pehlevan, C. (2020), 'Spectrum dependent learning curves in kernel regression and wide neural networks', *arXiv preprint arXiv:2002.02561* .

Cao, Y., Fang, Z., Wu, Y., Zhou, D.-X. and Gu, Q. (2019), 'Towards Understanding the Spectral Bias of Deep Learning', *arXiv:1912.01198 [cs, stat]* .

Jacot, A., Gabriel, F. and Hongler, C. (2018), Neural tangent kernel: Convergence and generalization in neural networks, *in* 'Advances in neural information processing systems', pp. 8571–8580.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Sohl-Dickstein, J. and Pennington, J. (2019), 'Wide neural networks of any depth evolve as linear models under gradient descent', *arXiv preprint arXiv:1902.06720* .

Luo, T., Ma, Z., Wang, Z., Xu, Z.-Q. J. and Zhang, Y. (2021), 'An upper limit of decaying rate with respect to frequency in deep neural network', *arXiv preprint arXiv:2105.11675* .

Luo, T., Ma, Z., Xu, Z.-Q. J. and Zhang, Y. (2019), 'Theory of the frequency principle for general deep neural networks', *arXiv preprint arXiv:1906.09235* .

Luo, T., Ma, Z., Xu, Z.-Q. J. and Zhang, Y. (2020), 'On the exact computation of linear frequency principle dynamics and its generalization', *arXiv preprint arXiv:2010.08153* .

Yang, G. and Salman, H. (2019), 'A fine-grained spectral perspective on neural networks', *arXiv preprint arXiv:1907.10599* .

Zhang, Y., Xu, Z.-Q. J., Luo, T. and Ma, Z. (2019), 'Explicitizing an Implicit Bias of the Frequency Principle in Two-layer Neural Networks', *arXiv:1905.10264 [cs, stat]* . arXiv: 1905.10264.
URL: *http://arxiv.org/abs/1905.10264*

Zhang, Y., Xu, Z.-Q. J., Luo, T. and Ma, Z. (2020), A type of generalization error induced by initialization in deep neural networks, *in* 'Mathematical and Scientific Machine Learning', PMLR, pp. 144–164.