理解深度学习

许志钦

2023 年 9 月 13 日

目录

Chapter 1

从频率原则理解神经网络

任意算法，如果考虑任意一个数据集，那么它们的性能都是等价的。这就是我们一直提到的 **No Free Lunch Theorem**。通俗来讲，抛开数据集谈算法，就和抛开剂量谈毒性一样，是完全没有意义的。某个算法在数据集 A 上表现出了很好的泛化性，但是很有可能就在数据集 B 上的表现比较差劲，我们可以用频率原则解释这一现象。在这一章中，我们将用频率原则来理解具有低频偏好的神经网络在不同任务中的泛化，以及一些神经网络的现象。

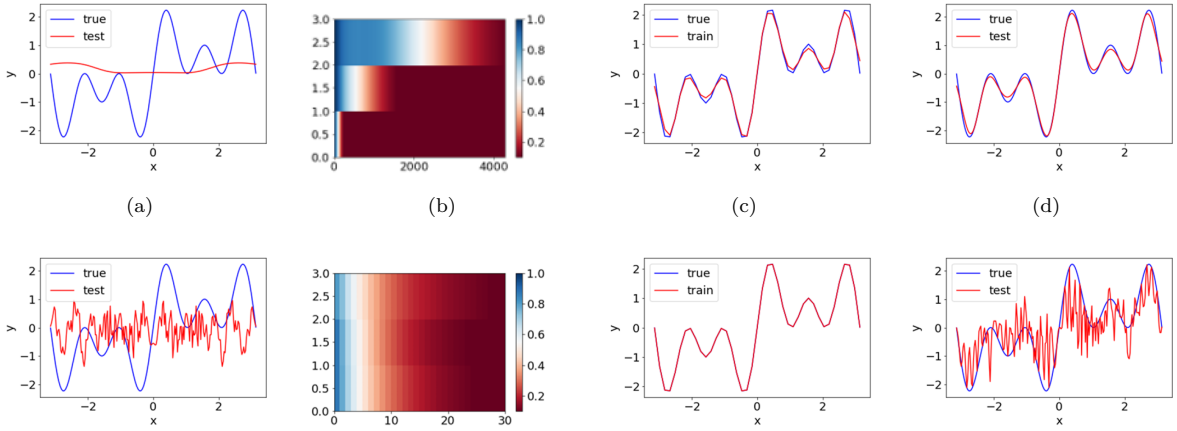
1.1 从一维例子理解为什么需要频率原则

我们用一个全连接网络来拟合 $\sin(x)+\sin(3x)+\sin(5x)$。用带参数 a 的 Ricker 函数作为激活函数：

$$\frac{1}{15a}\pi^{1/4}\left(1-\left(\frac{x}{a}\right)^2\right)\exp\left(-\frac{1}{2}\left(\frac{x}{a}\right)^2\right). \tag{1.1}$$

当 a 比较小时，神经网络拟合高频速度更快。正如图**??**所展示的，在第一行，我们选取一个较大的 a，使得低频收敛得更快，此时，训练点虽然没有完全拟合好，但在测试点上，神经网络的输出和真实值很接近。在第二行图片中，我们选择一个较小的 a，使得高低频收敛速度相近，此时频率原则消失了。神经网络可以完美地拟合训练点，但在测试点上，神经网络的输出有明显振荡，与真实值差异很大。不同的 a 会导致不同结果的原因，请见下一章的分析。

这些实验让我们认识到频率原则是非常重要的，没有频率原则，拟合得到的模型输出将会非常振荡，也就是高频成分很多，而这类拟合通常泛化性能会比较差。



(a) (b) (c) (d)

(e) 初始化 (f) 频率收敛 (g) 训练点结果 (h) 测试点结果

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

1.2 实际训练中的一些困难

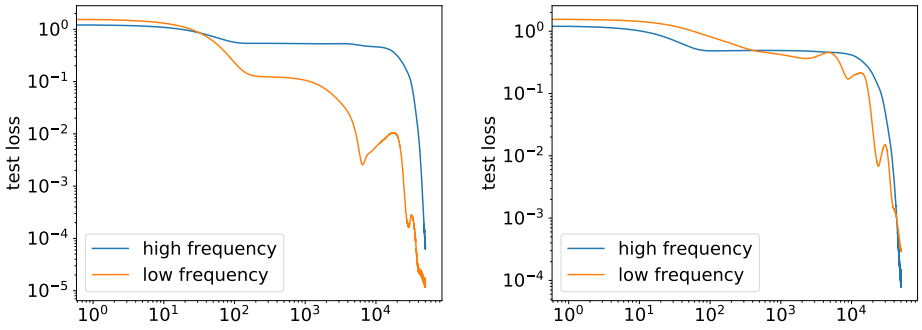
在实际训练中，常常会碰到一些误差特别大或者特别难学习的区域，该区域内大概率有明显的高频。可以通过增加样本或者该区域的训练权重来缓解困难。例如，我们考察目标函数

$$f(x) = \begin{cases} sin(5x) & x \leq 0 \\ sin(x) & x > 0 \end{cases}$$

显然原点左侧为高频部分，右侧为低频部分。我们考察在拟合过程中，原点两侧数据的拟合速度。正如图??(a) 中所示，低频数据点损失值先下降，而后高频部分数据点的损失值再下降。而假如我们对高频部分数据点损失分配更高的权重，如图??(b) 中所示，则高低频数据以几乎一致的速度拟合。

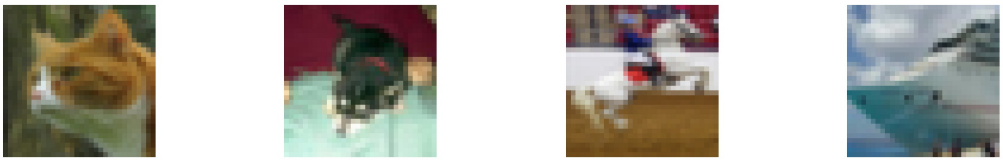
在分类问题中，有一些样本是容易区分的，有一些样本是难以区分的，并且用不同的神经网络结构训练，样本的困难程度常常类似。从频率的角度来看，简单样本应该处于低频区域，而困难样本处于高频区域。例如，我们用 VGG-9，VGG-16，VGG-19 分类 cifar-10 数据集，三个网络测试准确率均为 90% 左右 (共 10000 个测试样本)，而三个网络同时分类错误的测试样本共 494 个，这便是处于高频区域的困难样本。如图??所示，我们给出高频区域样本的几个示例。

用神经网络拟合图片时，模型会自动带有低频倾向，导致一定程度的磨光，同时也使得高频细节很难学习。如图??所示，我们使用神经网络拟合左侧图片，网络往往会学到一定程度磨光后的图片。



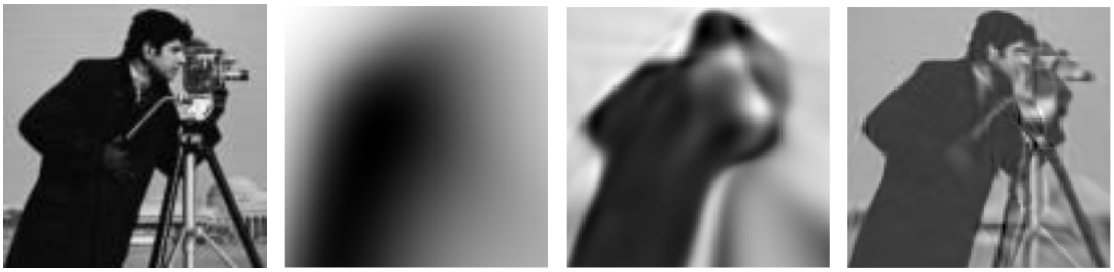
(a) (b)

图 1.2: 高频数据的与低频数据点拟合的速度。(a) 所有数据点损失权重相同。(b) 高频数据点损失权重为低频数据点的 20 倍。



(a) (b) (c) (d)

图 1.3: cifar-10 数据集中，高频区域样本的几个示例

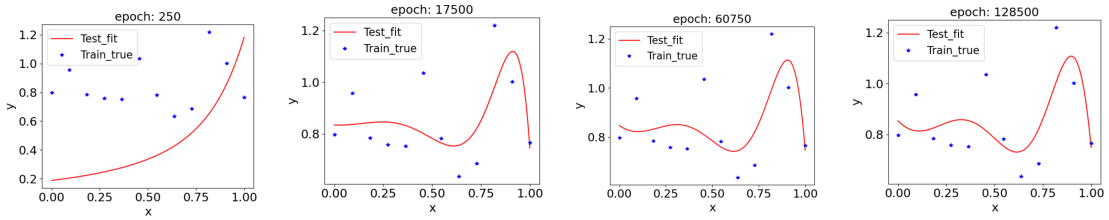


(a) target (b) epoch=80 (c) epoch=2000 (d) epoch=58000

图 1.4: 图片拟合问题中各个阶段网络输出。

用高阶多项式拟合数据时，在完全拟合好时，往往会有龙格现象，使拟合函数非常振荡。但若在梯度下降过程中使用提前停止的技巧，就可以大大缓解过拟合的出现。如图**??**所示，随

着训练的进行，模型先拟合低频信息，后拟合高频信息。对高频信息 (噪声) 的过度拟合，使得模型出现龙格现象。



(a) (b) (c) (d)

图 1.5: 使用均方损失和梯度下降来找到具有 11 阶和 12 阶等间距点的多项式插值的解。

如果满足损失函数的解有的是低频占优，有的是高频占优，那神经网络会倾向于选低频。这在神经网络解微分方程时，当解只能是弱解的时候会出现与真实解不一致的情况，原因就是选择了低频解。如图??所示，真实解相对于 PINN 得到的解更加陡峭。

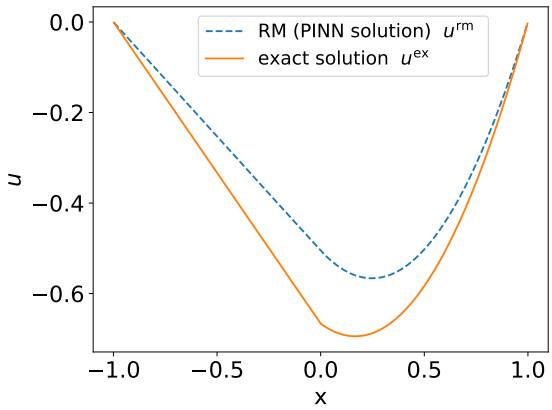


图 1.6: 通过 PINN 训练得到的解 (蓝色虚线) 及 PDE 的真实解 (橙色实线)。

1.3 Early-stopping 为什么有效的一种解释

在实际使用神经网络的过程中，人们经常会发现并不是把训练误差降到最低是最好的，而在训练过程中，训练误差会不断下降，但测试误差会先下降然后上升。我们来看一个例子，在图??，我们用一个神经网络去拟合带有噪音的红色数据，中间的图的绿色曲线说明随着训练持续，训练误差持续下降，但红色的曲线说明测试误差却在绿色虚线后开始上升。显然在这个例

子中，我们需要在绿色虚线处提前停止训练。为了理解，在这个例子中，为什么提前停止会有效，我们首先看一下，训练到绿色虚线时，神经网络的输出是什么样的。从左图的蓝色曲线可以看出，此时神经网络的输出是很光滑的 sin 曲线。实际上，我们的数据正是从 sin(x) 加上噪音中采样得到的。然后我们在频率空间检验一下此时神经网络的拟合效果。

在图**??**的右图中，红色和黑色的两条线分另是采样数据中的训练集和测试集的频谱图，可以看到他们在低频的位置重合得很好，但在高频的位置，它们的幅度都比较小，且不一致。这主要是因为噪音是幅度较小的白噪音，也就是在频率空间，噪音的各个频率的幅度是一样的且较小。因为数据是低频占优的，所以高频更容易受到噪音影响。我们再来看一下神经网络在训练集和测试集上的频谱。我们选择中间图绿色虚线的时刻，可以看到此时，神经网络在训练集和测试集上都与真实数据在低频上符合得很好，而高频此时仍然比较小。这是因为神经网络具有先学低频的频率原则，所以此时尚未拟合被噪音严重影响的高频，因此没有尚未发生过似合。

噪音对高频影响更大在很多真实数据中是很常见的。对于这类数据，频率原则使得神经网络先学习重要且信噪比高的低频信号。使用提前停止的技巧就可以防止神经网络拟合带大量噪音的高频，而不产生明显过拟合。

本小节的主要参考文献是**?**。

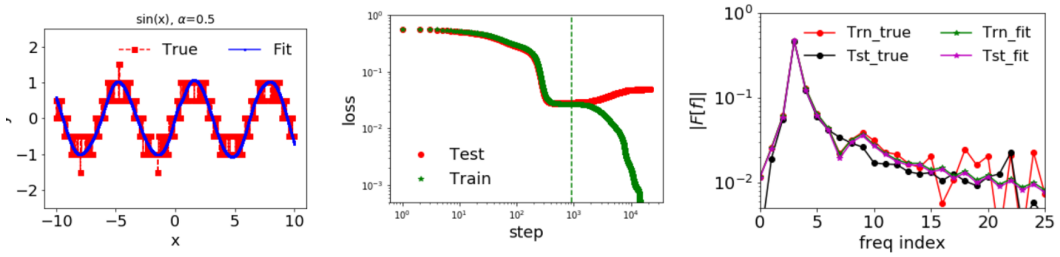


图 1.7: DNN 拟合带噪声的数据启示我们：训练的步数并不是越多越好。对真实数据的采样总会带来噪声，如果训练的步数太多，那么神经网络一定会学到数据中的噪声信息，反而降低了泛化性能。图片来源**?**。

1.4 神经网络的优势与局限

我们在介绍泛化问题的部分谈到神经网络有一个著名的泛化迷团，也就是为什么过参数化的神经网络和不容易发生过拟合。事实上，我们已经提到在谈泛化的时候离不开没有免费的午餐这个定理，也就是这个泛化迷团是针对某类特定的数据，及图像分类识别等任务。我们从频

率原则已经知道神经网络具有低频偏好，这些可以被泛化好的数据是不是本身也有低频主导的特性呢？以及我们是不是就可以找到一些高频主导的数据，预测神经网络在这类数据中是没有办法泛化好的？在这一小节中，我们通过对几类数据的频域分析来肯定这些猜想。

如图**??**所示，我们之前在 MNIST 和 CIFAR10 的主成分方向上进行了傅里叶变换发现它们是低频占优的，红色的点是由训练集计算得到的，绿色的点是测试集加上训练集的结果，蓝色的虚线是 DNN 输出在训练点和测试点上做傅里叶变换的结果，这里都是选真实数据集中的主成分方向做为研究方向。我们可以发现，在低频占有的数据集上，神经网络拟合的结果对低频的信息拟合的很好，而高频成分的幅值很小，所以即便高频没有拟合好也不会造成特别大的误差。

而考虑一个高频占优的数据集：定义一个 d 维空间中的函数，对于每个维度只取 1 或者 -1，也就是 $\boldsymbol{x}=(x_1,\cdots,x_d)\in\{-1,1\}^d$，函数为 $f(\boldsymbol{x})=\prod_{j=1}^d x_j$。当所有维度中取了偶数个 -1 时，$f(\boldsymbol{x})$ 的值就是 1，如果取了奇数个 -1，那么函数的值就是 -1，因此这个函数叫做奇偶函数。这种题目经常被用来做智商测试，一旦发现输出只依赖于 -1 的数目，那就很容易被用来预测其它点的值。比如对于 $d=10$，这个函数只定义在 $2^{10}=1024$ 个点上，若我们只给一部分的数据点做为训练集，然后预测那些没有给标签的点的输出。若我们用神经网络来学习这些训练集，神经网络可以在训练集准确地预测每一个点的输出，或者说能够记住所有训练点对应的输出，但在测试点上，它的准确率永远在 50% 左右。可以看出这里的人工智能是多么地不智能！同样地，我们用频率空间来分析一下这个例子。

对这个函数做离散傅里叶变换：

$$\frac{1}{2^d}\sum_{x\in\Omega}\prod_{j=1}^d x_j \mathrm{e}^{-\mathrm{i}2\pi\boldsymbol{k}\cdot\boldsymbol{x}}=(-\mathrm{i})^d\prod_{j=1}^d\sin 2\pi k_j \tag{1.2}$$

我们把这个变换的曲线展示在图**??**(c) 的绿线，当频率 k 变大时，它的幅度也慢慢增长，所以它是一组高频占优的数据。当我们采样不完整时，训练集做离散傅里叶变换的结果由红线呈现，显然，在幅度较小的低频，它与真实频谱有明显区别，事实上，由于采样不足，有一种频率混叠效应 (aliasing) 使得出现一些假的低频。现在用 DNN 去对它做一个拟合，可以发现 DNN 会产生很多假的低频，也就是对未知标签的数据，它用低频来拟合，这样也就导致 DNN 学习到的高频比真实数据的高频要少很多，因此，DNN 的频谱（蓝色虚线）与真实数据的频谱完全不一样，这样也就不难理解为什么神经网络完全无法在奇偶函数上做好预测了。在前面我们介绍到的实验引用前文的实验?，频率越高的目标函数，神经网络的泛化性也越差，而这里的奇偶函数是一个更极端的情况，完全无法泛化。

本小节的主要参考文献是**?**。

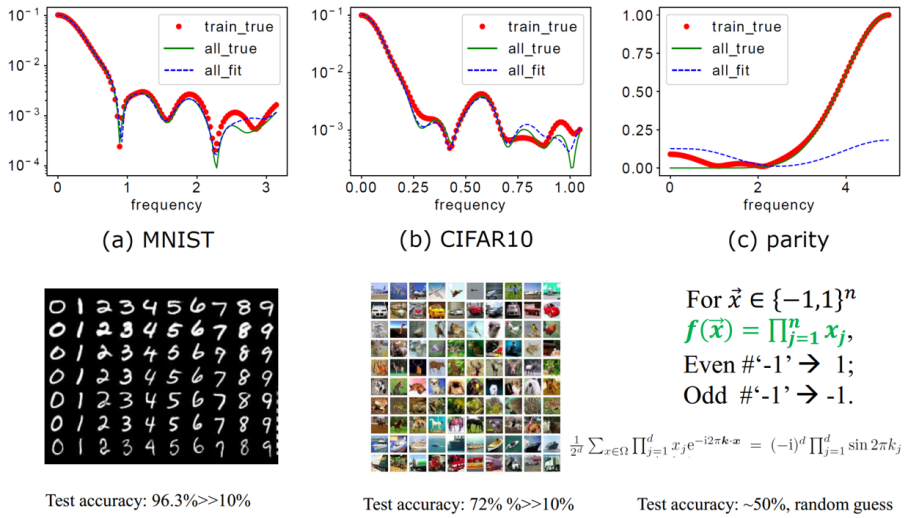


图 1.8: DNN 拟合真实数据，对于 MNIST,CIFAR10 等低频占主导的数据集，DNN 产生了良好的效果，而对于人为创造的高频占优的数据集，DNN 并没有产生良好的泛化性。

1.5 深度频率原则

本小节基于频率原则提出深度频率原则来理解为什么较深的网络能以较快的速度学习，主要参考文献是?。

1.5.1 深度加速训练的效应

近些年来，随着深度学习的发展，其已经在图像、语音、自然语言处理等各个不同的领域展现出了优异的性能。在运用中，人们发现，通过例如 Resnet 残差的手段，更深层的神经网络往往比隐藏层较少的神经网络训练得快，也有更好的泛化性能。

泛化的问题往往还与数据集本身有密切的关系。因此，我们首先关注为什么加深网络可以加快训练。为避免歧义，我们定义训练快慢是通过看网络达到一个固定误差所需要的训练步数。尽管更深的网络每步需要的计算量更大，但这里我们先忽略这个因素。

为了研究这个问题，首先我们用一个简单的实验来重现这个现象。图??是用不同层数但每层大小一致的 DNN 学习目标函数 cos(3x)+cos(5x)，训练到一个固定精度所需要的步数图。我们发现，越深层的神经网络，需要越少的步数，就能够完成任务。

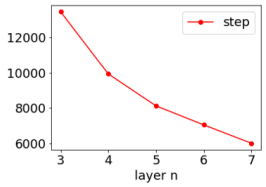


图 1.9: 不同深度神经网络在达到固定误差时的训练时长与隐藏层数关系图。

1.5.2 子网络的等效目标函数

接下来，我们将从频率视角来看深度的影响。对于隐藏层 hi，它的输入是它前一层的输出。在神经网络优化过程中，梯度是反向传播的，也就是说，当我们在更新隐藏层 hi 的参数时，误差的信号是从真实标签和神经网络输出的差异开始向后传播的。因此，对于子网络（从隐藏层 hi 到输出层），它的等效目标函数是由隐藏层 hi 的前一层的输出和真实的标签构成。基于此，若我们考虑从隐藏层 hi 到输出层这个子网络（我们暂且把它叫做感兴趣的学习层，learning component），从输入到隐藏层 hi-1 则可以看成感兴趣的学习层的预处理层（pre-condition component）。我们来研究对于不同深度的预处理层，我们感兴趣的学习层的等效目标函数的频谱。注意，训练时，我们仍然像往常一样，训练所有的参数。

假设两个不同的神经网络有相同的 learning component，即它们的最后若干层是相同的。若其中一个 learning component 的等效目标函数更加低频，那这个神经网络的 learning component 会学得更快。显然，learning component 学得更快，自然整个网络也就学得更快。特别地，当 learning component 学好的时候，整个神经网络也就学好了。因此，这给了我们充分的理由相信，通过研究 learning component 的性质，从这个角度出发，能够对多层神经网络的本质窥探一二。

1.5.3 频率比例密度函数

现在我们需要做的就是找到一个可以刻画高维函数频率分布的量，再利用 F-principle 低频先收敛的特性，我们就可以研究深度带来的效应了。因此，我们定义了频率比例密度函数（Ratio Density Function，RDF）。

本质上，我们首先通过在傅立叶空间画半径为 k_0 的球，定义目标函数在 k_0 球内的能量（L2 积分）占整个函数的能量比（通过高斯滤波获得），即低频能量比（Low frequency ratio，LFR）。具体地，我们利用高维频率原则实验中的滤波方法获得低频部分 $\boldsymbol{y}_i^{\mathrm{low},\delta(k_0)}$。然后，我们定义 LFR

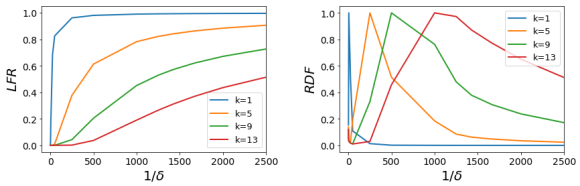
$$\mathrm{LFR}(k_0) = \frac{\sum_i |\boldsymbol{y}_i^{\mathrm{low},\delta(k_0)}|^2}{\sum_i |\boldsymbol{y}_i|^2}. \tag{1.3}$$

这类似于概率的累积分布函数。然后我们对 LFR 在 k_0 上求导数得到 RDF，

$$\mathrm{RDF}(k_0) = \frac{\partial \mathrm{LFR}(k_0)}{\partial k_0}. \tag{1.4}$$

这可以解释为函数在每个频率上的能量密度。令 δ 为用来做卷积的高斯函数的标准差，那 $1/\delta$ 则是频域空间的高斯函数的标准差，也近似为频域空间上取的频率半径 k_0。

为了方便大家理解，我们计算 $\sin(k\pi x)$ 的 LFR 和 RDF。如图 **??**(a)，低频的 LFR 随着截断频率 $1/\delta$ 增加，最快到达 1 的位置。图**??**(b) 中的 RDF 的峰值位置与函数中的频率有一致的关系，频率越低，峰值越靠近零。所以通过 RDF 的峰值可以简单判断该函数的主要成分在什么频率。下面，我们就用 RDF 来刻画高维函数的频谱。



(a) (b)

图 1.10: $\sin(k\pi x)$ vs. $1/\delta$ 的 LFR 和 RDF。注意，为了可视化，我们通过每条曲线的最大值来归一化 (b) 中的 RDF。

1.5.4 深度频率原则的实验

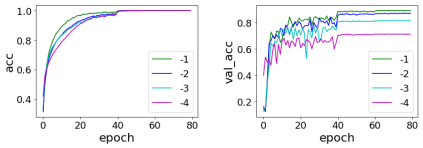
最后，我们需要研究 learning component 的等效目标函数的 RDF。如果 learning component 的等效目标函数的 RDF 趋近于低频，那么通过 F-principle，我们就知道其收敛得会比较快；相反，若其趋近于高频，则其收敛得就会比较慢。

实验上，我们先做了关于 Resnet18 的实验。我们保持全连接层不变，改变 Resnet 卷积模块的个数，并定义最后三层全连接层为 learning component。

整个训练和往常一样，训练所有的参数。在图**??**中，-1、-2、-3、-4 的残差块依次减少，不难发现，拥有更多残差块的网络不仅收敛速度更快，同时泛化性能也更好。

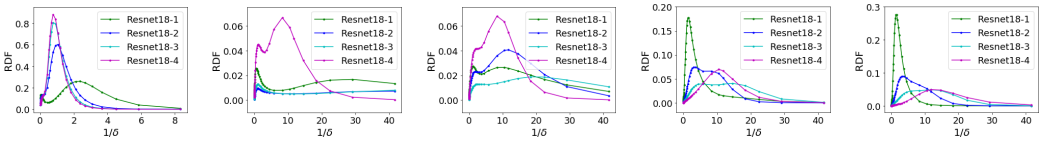
如图**??**，观察其 learning component 的等效目标函数的 RDF，我们发现，拥有更多隐藏层（也就是网络更深）的神经网络其 learning component 相比浅网络会趋于更低频，并最后保持在更加低频处。

我们得到了 Deep Frequency Principle——更深层神经网络的有效目标函数在训练的过程中会更趋近于低频。再基于 F-principle——低频先收敛，我们就能够得到更深层的神经网络



(a) Training (b) Validation

图 1.11: Training accuracy and validation accuracy vs. epoch for variants of Resnet18.

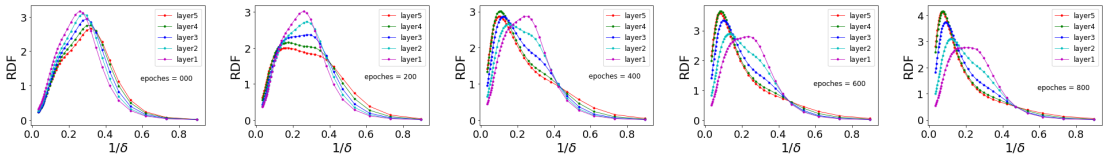


(a) epoch 0 (b) epoch 1 (c) epoch 2 (d) epoch 15 (e) epoch 80

图 1.12: RDF of $S^{[-3]}$ (effective target function of layer "-2") vs. $1/\delta$ at different epochs for variants of Resnet18.

收敛得更快的结果。尽管频率是一个相对可以定量和容易分析的量，但当前实验跨越了多个不同结构的网络，也会给未来理论分析造成困难。因此，我们后面研究单个神经网络中的 Deep Frequency Principle。

于是，我们探究同一个深度神经网络内不同隐藏层的等效目标函数的 RDF，即改变 precondition 和 learning component 的层数（但保持网络的结构和总层数不变）。这个实验是在 MNIST 上的，深度神经网络（DNN），并取了 5 个相同大小的隐藏层。在图**??**中，我们发现，虽然初始时神经网络更深层的等效目标函数的 RDF 聚集于较高频处，但随着训练，更深层的 RDF 会快速地趋于更低频的地方，并保持在低频处。这也是 Deep Frequency Principle——更深层的神经网络的有效目标函数会在训练的过程中会更趋近于低频。



(a) epoch 0 (b) epoch 200 (c) epoch 400 (d) epoch 600 (e) epoch 800

图 1.13: 全连接神经网络在学习 MNIST 数据时，不同隐层的 RDF 与不同阶段 $1/\delta$ 的比较。五条彩色曲线分别对应五个隐含层。带有 "层 i" 图例的曲线是 $S^{[i]}$ 的 RDF。

参考文献

Xu, Z.-Q. J., Zhang, Y., Luo, T., Xiao, Y. and Ma, Z. (2019), 'Frequency principle: Fourier analysis sheds light on deep neural networks', *arXiv preprint arXiv:1901.06523* .

Xu, Z.-Q. J., Zhang, Y. and Xiao, Y. (2018), 'Training behavior of deep neural network in frequency domain', *arXiv preprint arXiv:1807.01251* .

Xu, Z.-Q. J. and Zhou, H. (2021), Deep frequency principle towards understanding why deeper learning is faster, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 35.