理解深度学习

2023 年 9 月 13 日

目录

Chapter 1

凝聚现象

频率原则从函数空间的角度刻画了训练过程的规律。由于输出函数是众多参数共同作用形成的结果，因此，我们也可以说频率原则是一种宏观的描述。对于微观，也就是具体每个参数，在实际训练中是否真的非常复杂呢？过度参数化的神经网络通常通过在没有显式正则化Breiman (1995), Zhang et al. (2021) 的情况下最小化损失函数来在实际问题上表现出良好的泛化性能。对于过度参数化的神经网络，有无限可能的全局最小值点可以达到令人满意的训练损失。但是，它们的泛化性能非常不同。重要的是要研究在训练期间对损失函数施加了哪些隐式正则化，从而将神经网络引导到特定类型的全局最小值点。比如表面上，神经网络有大量的参数，但实际训练过程或者训练完成后，寻找到的全局最小值点是否真的有非常多的有效参数呢？这些问题促使我们去研究参数的演化。

1.1 理想的凝聚现象及凝聚的意义

我们首先给出一个理想的凝聚现象的示意图。如图1.1所示，虽然给定神经网络在初始化时，各个神经元的输入权重差异很大，但是在经过一段时间训练后，中间隐藏神经元分成了两类，前两个神经元是一类，后三个神经元是另一类。在每一类中，不同神经元的输入权重是完全一样的，因此，它们的输出也是一样的，但它们仍然有不一样的输出权重。我们把出现不同神经元具有完全相同的输入权重的现象称为凝聚现象。

显然，一旦同层的神经元发生聚类后，这个网络就可以等效成一个小网络。如1.1所示，该网络的隐藏层聚成两类神经元，因此，其等效成具有两个隐藏神经元的网络。每个等效的隐藏神经元的输入权重和该类里的神经元的输入权重一样，但其输出权重是该类中所有神经元的输出权重的求和。

现在让我们重新回顾一下过参数化带来的问题。传统的学习理论说，一个模型的复杂度太大容易造成过拟合。一种简单的看模型复杂度的方法就是数模型的可训练参数。对于神经网络，大量的经验告诉我们，尽管神经网络的可训练参数远大于样本的数目，但我们却观察到神经网络不容易过拟合。那到底是什么地方出了问题呢?

一种可能就是单纯地数模型的参数在这里是行不通的。我们更关心一个神经网络在实际问题中它的有效参数是多少。当神经网络的同层神经元发生凝聚后，其有效的神经元数目会远小于它实际拥有的神经元数目。因此，它的有效复杂度也变得很小。这种有效复杂度才有衡量其泛化能力合理的量。所以，凝聚对一个大网络获得好的泛化能力可能是非常重要的。

既然凝聚的目的是让大网络等效成一个小网络，那读者自然要问，为什么我们不直接训练一个小的网络呢？为什么要花这么算力去把一个大网络变成一个小网络？这是一个有趣的问题，我们将在研究损失景观的嵌入原则中进一步探讨这个问题。一个快速的回答是，尽管大网络可以在表达能力上等效成小网络，但其在优化上可能会比小网络更容易。

我们还关心为什么会发生凝聚现象。这里涉及到神经网络模型非常本质的一个结构，也就层结构。在每一层中，不同神经元的差别在于它们的序号。在任一优化算法下，不同神经元的动力学形式是完全一样的。考虑一种简单的情况，假如这个动力学只有两个稳定点，而有一万个神经元，那么这些就只会分成两类，跑到这两个稳定点。实际上，神经网络模型，特别是过参数的模型特别复杂，稳定点或者全局最小点太多，要那怎么保证大家尽量跑到相同的稳定点呢？最简单的办法是让所有神经元的初始值都一样。但这样又限制了模型的表达能力。一种折衷的方法就是让所有神经元的初始值非常靠近，但并不相同。为达到这种情况，我们可以让初始化尽量小。这就是为什么我们要小初始化。在初始凝聚的部分，我们会把这些直观解释形成严格的数学推理。

但小初始化并不是必要的。如果我们加一些正则化技巧，比如后面我们会谈到的 Dropout，它也会让不同神经元趋于相同，同样可以达到凝聚的效果。

1.2 三个简单的例子

在相图分析中 (Luo et al. 2021)，我们介绍了两层无穷宽 ReLU 网络可以分成三个区域，一个是线性区域，一个是非线性区域，也称为凝聚区域，以及分割这两个区域的临界区域。首先我们来看一下这三个区域下参数的深化行为。

我们使用相图分析中的一维实验来说明相图中各种典型情况，神经网络拟合在 4 个一维的训练点。图 1.2的第一行显示了不同 γ 值下的典型学习结果, 从相对锯齿形的插值 (NTK 尺度) 到平滑的类似三次样条式的插值 (平均场尺度)，再到线性样条插值。

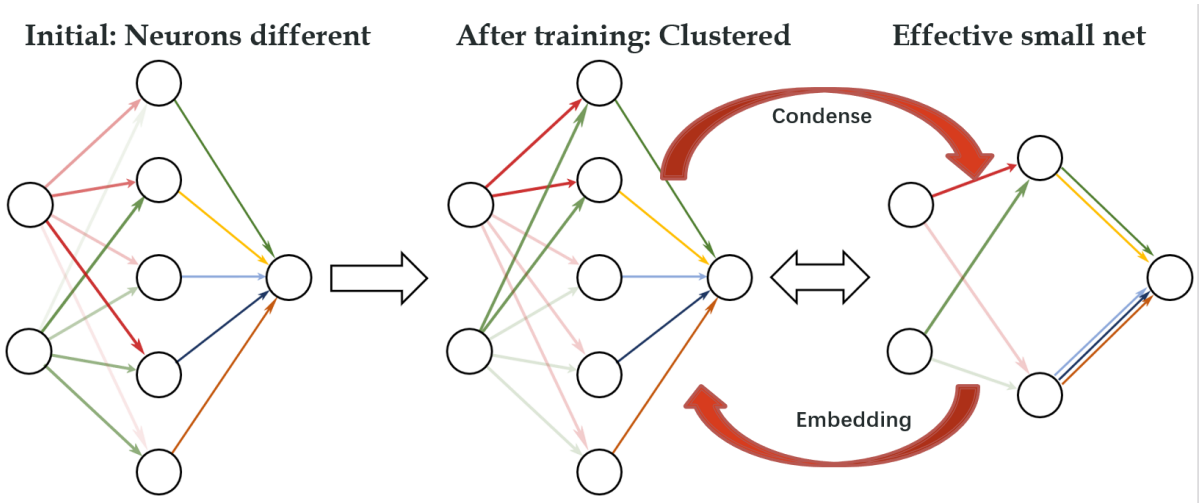


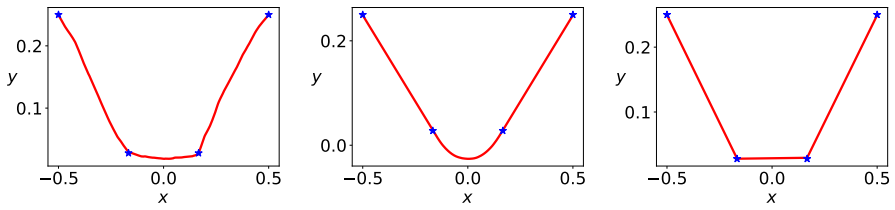
图 1.1: 理想的凝聚现象。

对于一维输入的两层神经网络

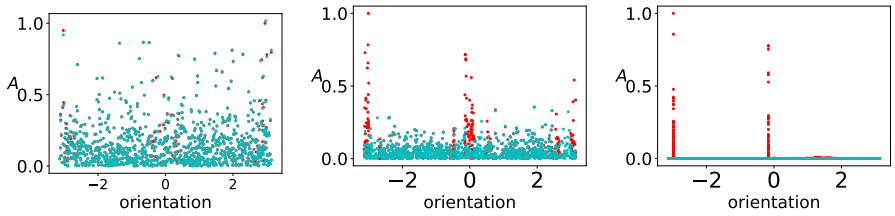
$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{m} a_j \sigma(\boldsymbol{w}_j \cdot \boldsymbol{x}) = \sum_{j=1}^{m} a_j \sigma(w_j x + b_j), \tag{1.1}$$

为了深入探究它们的参数空间表示的细节，我们注意到对于 ReLU 激活函数，每个神经元的参数对 $(a_k, \boldsymbol{w}_k)$ 可以分离为一个单位方向特征 $\hat{\boldsymbol{w}} = \boldsymbol{w}/|\boldsymbol{w}|_2$ 和一个表示其对输出贡献的振幅 $A = |a||\boldsymbol{w}|_2$, 即 $(A, \hat{\boldsymbol{w}})$。对于一维输入，由于加入了偏置项，$\boldsymbol{w}$ 是二维的。因此，我们使用每个 $\hat{\boldsymbol{w}}$ 相对于 x 轴的 $[-\pi, \pi)$ 内的角度来表示其方向。$(A_k, \hat{\boldsymbol{w}}_k)_{k=1}^m$ 的散点图如图 1.2的第二行所示。很明显，图 1.2第一行中的例子的参数演化是不同的。对于线性区域 $\gamma = 0.5$, 训练后的散点图与初始散点图非常接近。因为参数距离很近，所以末态的神经网络函数可以由在初始值的一阶泰勒展开较好地近似。然而, 对于非线性区域 $\gamma = 1.75$, 活跃神经元 (即具有显著振幅 A 的神经元) 被聚集到了几个方向上, 这与初始散点图有很大的偏差。对于平均场模型下的初始化，其也产生了一定的凝聚，但并没有非线性区域这么剧烈。

我们来理解一下方向的函数。对于每个神经网络 $\sigma(wx+b)$, 把它的方向记为 ϕ, 则 $\tan\phi = b/w$。对于 ReLU，其转折点的位置在 $wx + b = 0$，也就是 $x = -b/w$，因此每个转折点就对应了一个角度。这个一维的例子中，最少刚好是两个转折点就可以拟合，恰好神经网络刚好学到两个方向。



(a) $\gamma = 0.5$ (b) $\gamma = 1$ (c) $\gamma = 1.75$



(d) $\gamma = 0.5$ (e) $\gamma = 1$ (f) $\gamma = 1.75$

图 1.2: 第一行所示为用不同 γ 的两层 ReLU 神经网络初始化学习四个数据点的结果。第二行所示为对应的初始 (青色) 和末态 (红色)$\{(A_k, \hat{\boldsymbol{w}}_k)\}_{k=1}^m$ 的散点图。$\gamma' = 0(\beta_1 = \beta_2 = 1)$, 隐层神经元数量 $m = 1000$。

1.3 临界和凝聚区域

通过前节的图 1.2 (d-f) 可以观察到, 与初始化相比, 神经网络在特征空间中的表示 $((A_k, \hat{\boldsymbol{w}}_k)_{k=1}^m)$ 的凝聚是神经网络非线性训练动力学的一个显著特征。具体地, 我们关注 $m \to \infty$ 时的这种凝聚, 即当 $\boldsymbol{\theta}_w$ 的相对变化趋于 $+\infty$。使用与图 1.2中相同的 1 维数据, 我们观察 $(A_k, \hat{\boldsymbol{w}}_k)$ 分布。我们用三种宽度的两层网络 $m = 10^3, 10^4, 10^6$。结果如图 1.3所示, 很容易观察到, 在蓝框指示的边界的右侧, 在非线性区域, 随 $m \to \infty$, 凝聚变得更强。这与我们的直觉一致, 即 $\boldsymbol{\theta}_{\boldsymbol{w}}$ 偏离初始化的距离越远, 这里表现为凝聚的神经网络的非线性越强。因此, 如前文所述, 我们将这个区域称为凝聚区域。在线性区域和凝聚区域之间的临界区域中, 随着 $m \to \infty$, 凝聚的程度几乎保持不变, 这类似于平均场行为。一般来说, 凝聚的机制以及它的隐式正则化效应还不为人所理解, 这仍然是未来研究的重要问题。

我们还观察了 MNIST 数据集上的神经网络的凝聚。对于这样的高维数据, 不可能像上述一维情况那样直接可视化高维特征空间中的分布。因此, 我们考虑一个投影方法, 即将每个 $\hat{\boldsymbol{w}}$ 投影到一个参考方向 $\boldsymbol{p}$, 并画出 A_k 与 $I_{\hat{\boldsymbol{w}}} = \hat{\boldsymbol{w}} \cdot \boldsymbol{p}$ 的关系图。注意参考方向可以任意选择, 不影响我们的结论。在不失一般性的情况下, 我们取 $\boldsymbol{p} = \mathbf{1}/\sqrt{n}$。很明显, 如果神经元确实在高维特征空间中的几个方向上凝聚, 那么它们的一维投影也应在几点上凝聚。如图 1.4所示, 与一维

情况类似, 可以观察到在上文确定的凝聚态中存在凝聚行为。随着参数与边界距离的增加, 凝聚变得更加明显。

1.4 凝聚的过程

接下来，我们用一个一维的例子来说明，神经网络在演化的过程中凝聚方向会逐渐增加。我们选取目标函数为

$$f(x) = -\mathrm{ReLU}(x) + \mathrm{ReLU}(2\times(x+0.3)) - \mathrm{ReLU}(1.5\times(x-0.4)) + \mathrm{ReLU}(0.5\times(x-0.8)), \quad (1.2)$$

并采用两层 ReLU 神经网络进行拟合。如图1.6所示，我们可以看到，随着神经网络的演化，凝聚的方向逐渐增多，每个方向上的神经元的模长也逐渐增加。与此同时，如图1.7，图中损失函数的五个标记的点对应图1.6的五个中间态，我们可以观察到发生凝聚时损失函数处于训练的不同阶段。

1.5 卷积网络的凝聚现象

对于卷积网络，我们把每一个卷积核当作一个神经元。在这个设定下，卷积网络的凝聚现象就是不同卷积核的权重在学习后呈现出一样的数值。为了刻画卷积网络的凝聚现象，我们把卷积核展成向量，然后算向量之间的内积。为此，我们定义余弦相似度。**余弦相似度:** 两个向量 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 的余弦相似度定义为

$$D(\boldsymbol{u},\boldsymbol{v}) = \frac{\boldsymbol{u}^\intercal \boldsymbol{v}}{(\boldsymbol{u}^\intercal \boldsymbol{u})^{1/2}(\boldsymbol{v}^\intercal \boldsymbol{v})^{1/2}}. \quad (1.3)$$

如图1.8所示, 我们用不同卷积核之间的余弦相似度来度量参数的凝聚程度。我们用交叉熵损失训练了一个三层 $\tanh(x)$ CNN 拟合数据集 CIFAR10。如图1.8(a)-(c)，当训练准确率达到 100% 时，神经网络发生了比较明显的参数凝聚现象。

1.6 实际网络的凝聚现象

通过对在 timm 中模型的预训练的参数模型的观察后，我们发现在实际训练的过程中也存在参数凝聚的现象。如图1.9，我们可以观察到 Resnet-18 预训练的参数下，该卷积层的参数有一定的凝聚现象，而随机正态初始化的参数没有发生这一现象。

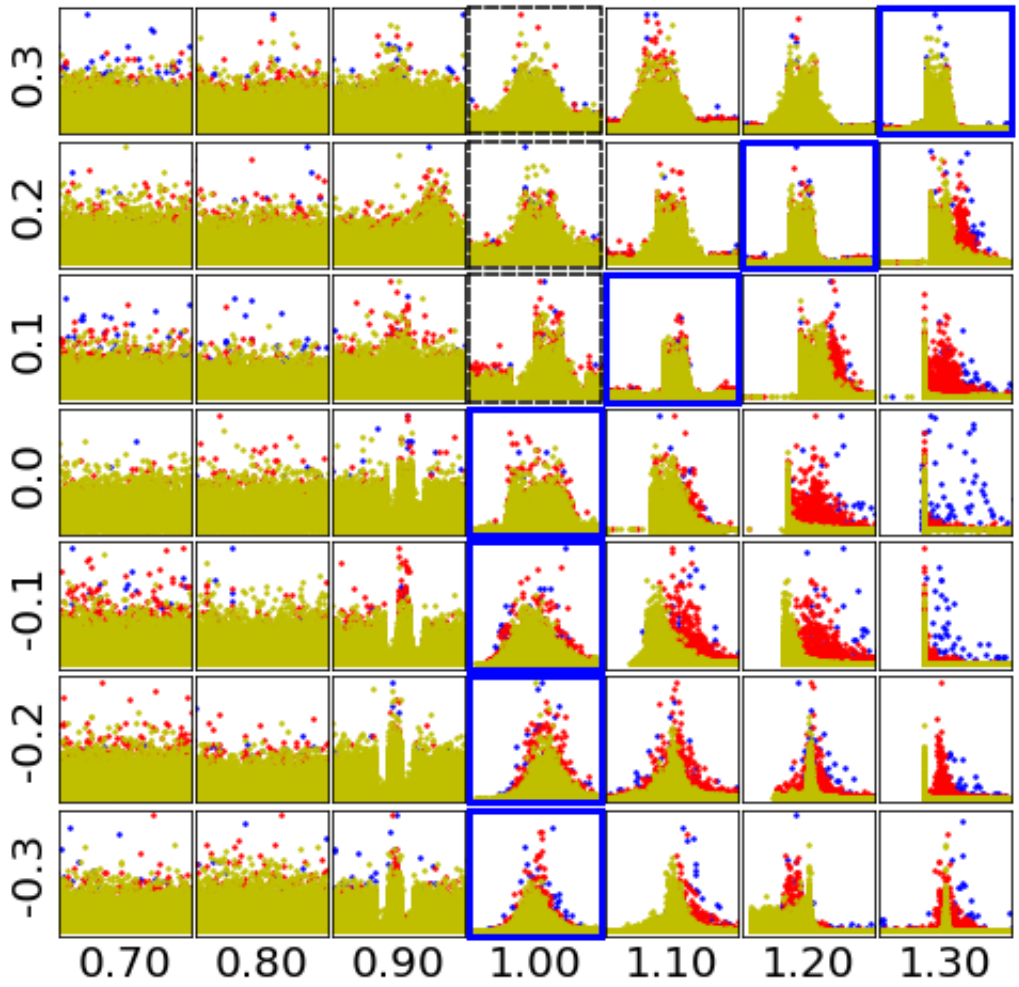


图 1.3: 如图为 1 维合成数据的凝聚图。每个框中的每个颜色表示对应 γ 和 γ' 的神经网络上的 $\{(A_k, \hat{\boldsymbol{w}}k)\}k = 1^m$ 的散点分布。隐层神经元数量为:$m = 10^3$(蓝色),10^4(红色),10^6(黄色)。横坐标为 γ, 纵坐标为 γ'。

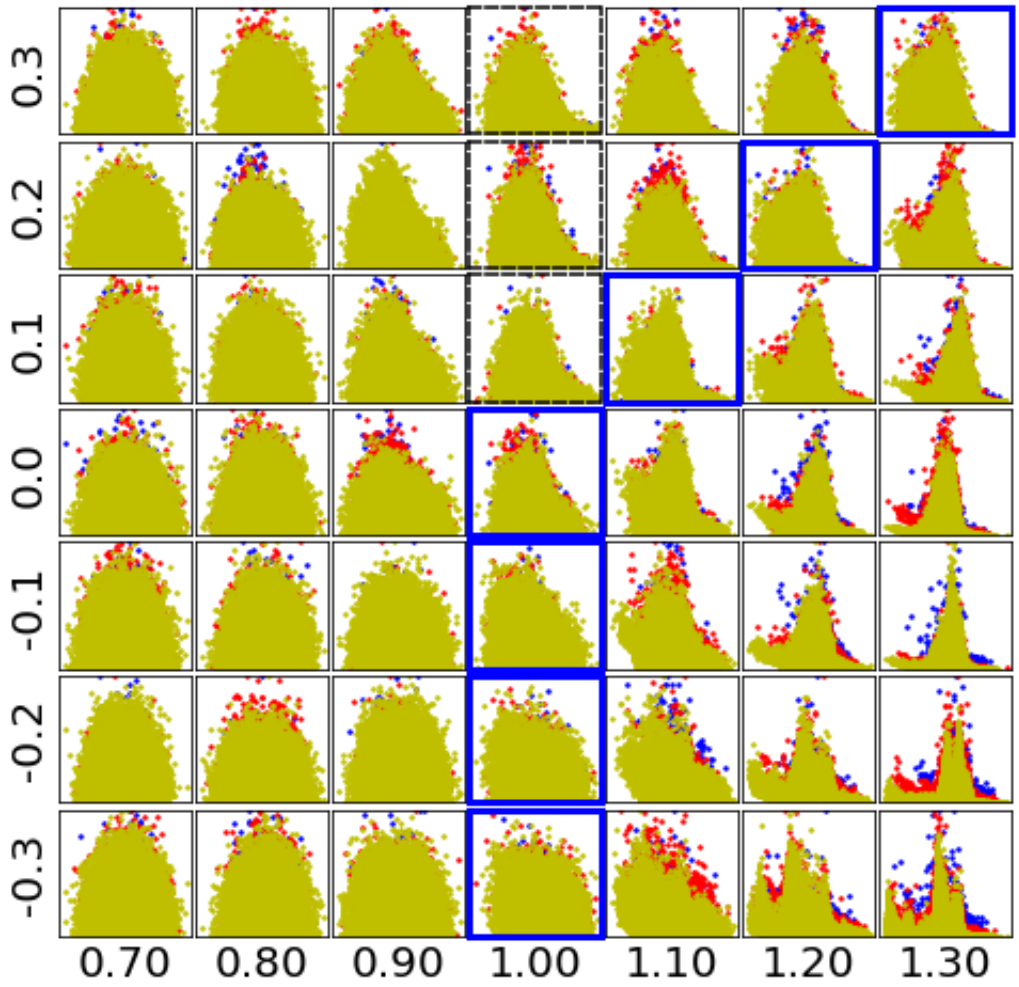


图 1.4: 如图为 MNIST 数据集的凝聚图。每个框中的每个颜色表示对应 γ 和 γ' 的神经网络上的 $\{(A_k, I_{\hat{\boldsymbol{w}}})\}_{k=1}^m$ 的散点分布。隐层神经元数量为:$m = 10^3$(蓝色),10^4(红色),2.5×10^5(黄色)。横坐标为 γ, 纵坐标为 γ'。

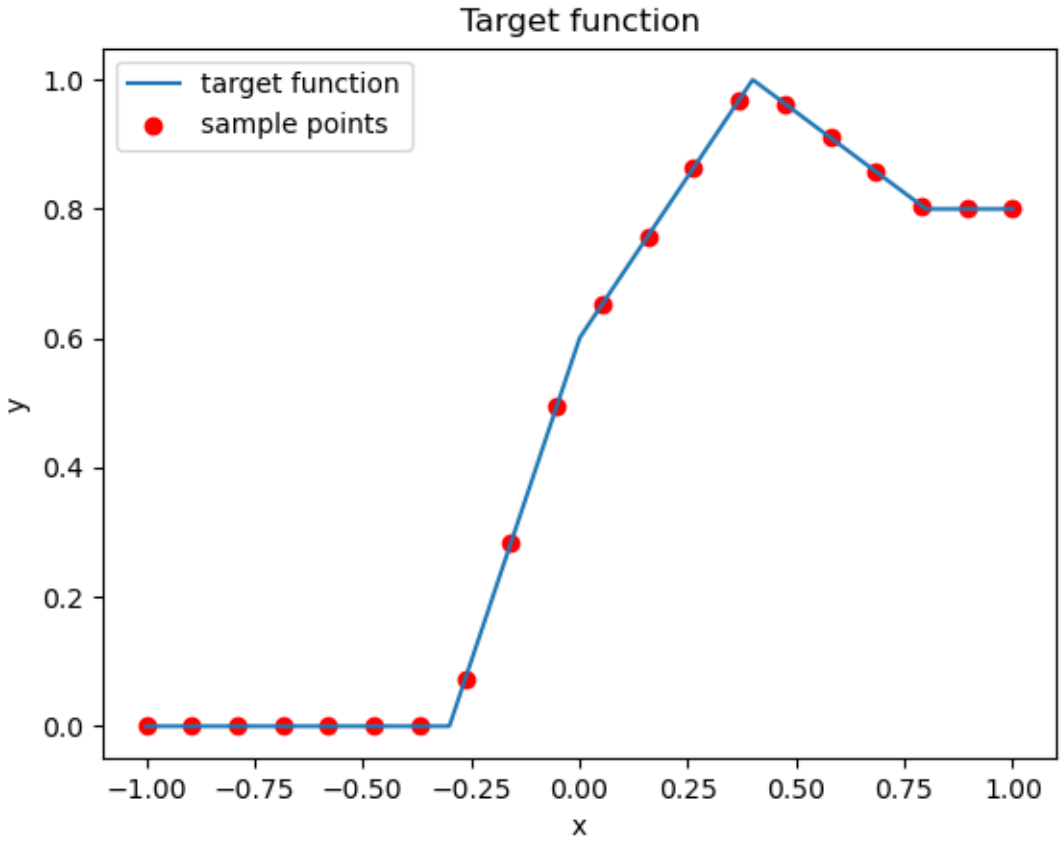

Target function
target function
sample points
1.0
0.8
0.6
0.4
0.2
0.0
y
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
x

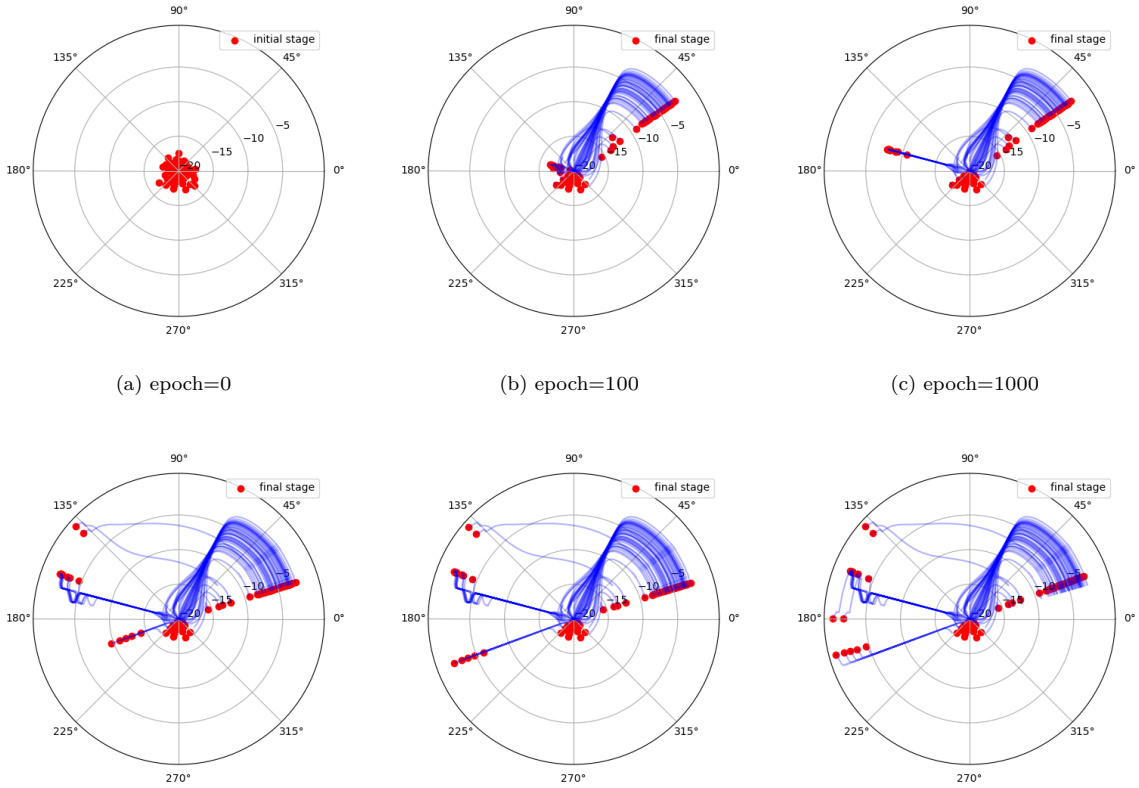
图 1.5: (1.2)的函数图像

表 1.1: Comparison of common (Glorot & Bengio 2010) and condensed Gaussian initializations on resnet18. $\bar{m} = (m_{\text{in}} + m_{\text{out}})/2$. m_{in}: in-layer width. m_{out}: out-layer width.

	common			condensed		
	Glorot_uniform	Glorot_normal	$N(0, \frac{1}{\bar{m}})$	$N(0, \frac{1}{m_{\text{out}}^4})$	$N(0, \frac{1}{m_{\text{out}}^3})$	$N(0, (\frac{1}{\bar{m}})^2)$
Test 1	0.8807	0.8777	0.8816	0.8847	0.8824	0.8826
Test 2	0.8857	0.8849	0.8806	0.8785	0.8813	0.8807
Test 3	0.8809	0.8860	0.8761	0.8824	**0.8861**	0.8800

1.7 更多关于小初始化的讨论

实际上，在小初始化下的神经网络，其泛化性能并不会比一般线性区域和临界区域的性能差。以一个学习 CIFAR10 数据集上的 resnet18-likeHe et al. (2016) 结构的神经网络为例，不难发现，当我们对其卷积部分和全连接的部分分别使用一般初始化和凝聚区域的初始化时，其泛化性能是差不多 (Zhou et al. 2022)。而这，也就体现了凝聚区域的初始化是实际可用有价值的，而不仅仅停留在理论之上。



(a) epoch=0 (b) epoch=100 (c) epoch=1000

(d) epoch=3000 (e) epoch=5000 (f) epoch=100000

图 1.6: 两层 ReLU 神经网络拟合(1.2)时神经元随时间演化的 $(A_k, \hat{\boldsymbol{\alpha}}_k)$ 分布，其中 A_k 如同之前定义，$\hat{\boldsymbol{\alpha}}_k = \operatorname{acrtan}(\frac{b_k}{w_k})$。红色的点表示 $(A_k, \hat{\boldsymbol{\alpha}}_k)$ 在当前时刻的分布，蓝色轨迹为对应红色神经元模长随时间演化的轨迹。隐藏层神经元数目为 $m = 100$。模长的坐标尺度为 log 尺度。

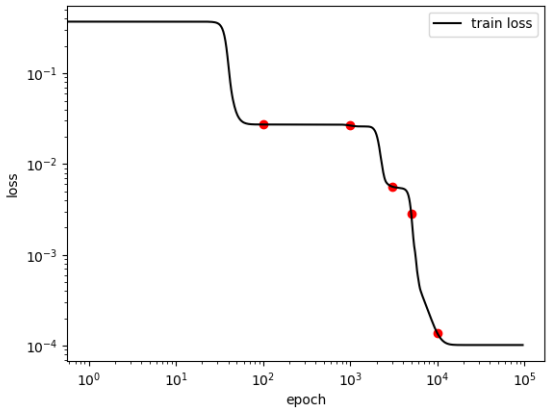
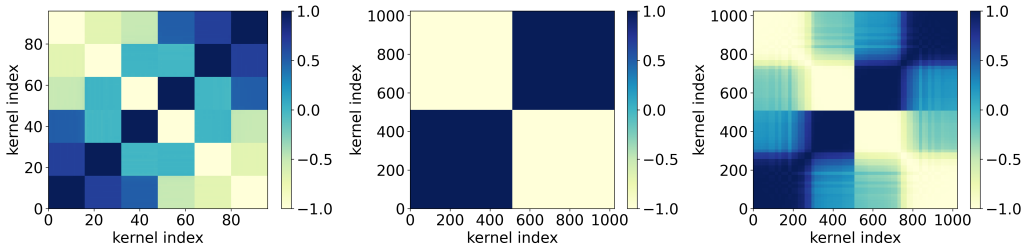
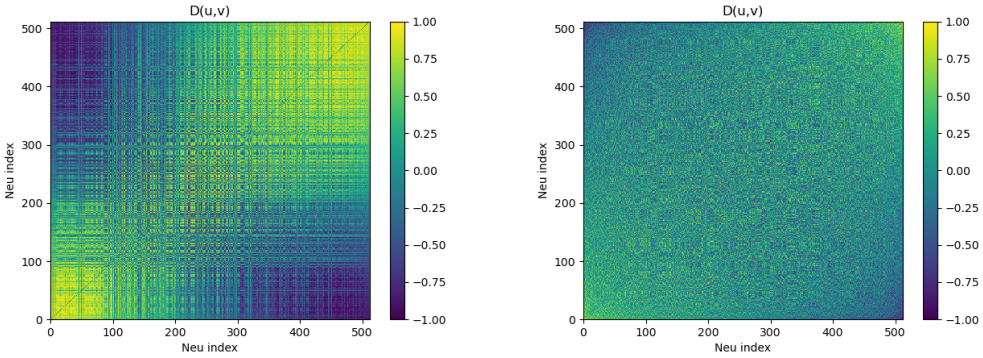


图 1.7: 图1.6训练过程中损失函数的图像，红点依次为 epoch= $100, 1000, 3000, 5000, 10000$。



(a) layer 1 (b) layer 2 (c) layer 3

图 1.8: 小初始化下三层 CNN 训练后的最终阶段的凝聚。激活函数为 tanh(x)。如果神经元在同一个米色块中, 则 $D(\boldsymbol{u}, \boldsymbol{v}) \sim 1$(深蓝色块中,$D(\boldsymbol{u}, \boldsymbol{v}) \sim -1$), 表示它们的输入权重方向相同 (相反)。颜色表示两个不同卷积核的 $D(\boldsymbol{u}, \boldsymbol{v})$, 其索引由横轴和纵轴分别表示。训练集为 CIFAR10。线性层使用 ReLU 作为激活函数，输出层使用 softmax，损失函数为交叉熵，优化器为 Adam。卷积核大小 $m = 5$，学习率 $\mathrm{lr} = 2 \times 10^{-6}$。



(a) epoch=100 (b) epoch=1000

图 1.9: Resnet-18 中第十四个卷积层某一输入通道对应的所有输出通道的卷积核向量的余弦相似度。左图为预训练参数，右图为随机正态初始化的参数。

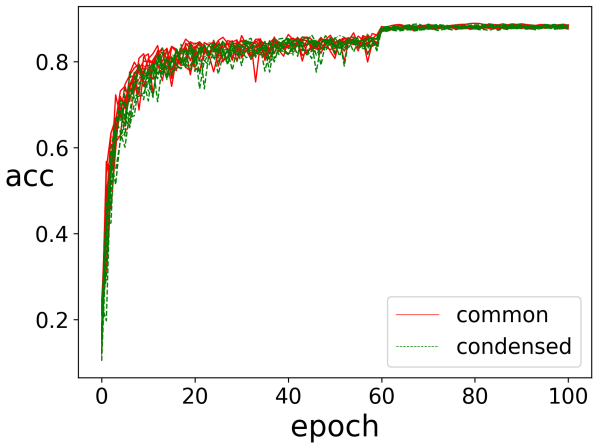


图 1.10: 凝聚初始化下的泛化性与一般常用初始化的泛化性对比

参考文献

Breiman, L. (1995), 'Reflections after refereeing papers for nips', *The Mathematics of Generalization* **XX**, 11–15.

Glorot, X. & Bengio, Y. (2010), Understanding the difficulty of training deep feedforward neural networks, *in* 'Proceedings of the thirteenth international conference on artificial intelligence and statistics', pp. 249–256.

He, K., Zhang, X., Ren, S. & Sun, J. (2016), Deep residual learning for image recognition, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 770–778.

Luo, T., Xu, Z.-Q. J., Ma, Z. & Zhang, Y. (2021), 'Phase diagram for two-layer relu neural networks at infinite-width limit', *Journal of Machine Learning Research* **22**(71), 1–47.

Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. (2021), 'Understanding deep learning (still) requires rethinking generalization', *Communications of the ACM* **64**(3), 107–115.

Zhou, H., Zhou, Q., Luo, T., Zhang, Y. & Xu, Z.-Q. (2022), 'Towards understanding the condensation of neural networks at initial training', *Advances in Neural Information Processing Systems* **35**, 2184–2196.