理解深度学习

许志钦

2023 年 9 月 13 日

目录

Chapter 1

相图分析

1.1 动机

在神经网络的应用中，过参数化是一种常见的情况。在过参数化下，参数的数目大于训练样本的数目，这使得全局最小点的数目并不唯一。一般使用的损失函数只考虑训练样本上的误差，而对于那些未被用来训练模型的样本，不同全局最小点的效果差异会非常大，也就是有非常不一样的泛化。在实际训练中，我们通过给定损失函数、优化算法和一系列超参，包括神经网络的结构、参数的初始化、学习率等，每一次训练基本都能找到一个解。在不同训练中，只要我们的设定基本一致，尽管找到的解不一样，但其泛化性差异也不会很大，否则，神经网络的结果就几乎不可复现。

前面给定的这些条件可以理解为隐式地给全局最小解设定了一些条件，使得我们最终从所有的可行解中找了一个满足这些隐式条件的解。这些条件并没有显式地给出，因此，我们称之为隐式正则化或者隐式偏好。这和传统的优化问题有明显的区别。比如对于一个优化问题 $\min ||Ax-y||_2^2$, 其中 $A \in \mathbb{R}^{m\text{xn}}$, $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, 对于 $m < n$ 的情况，这个问题是有多解的。为了获得唯一的解，一般会施加一些显式的正则项，比如考虑问题 $\min ||Ax-y||_2^2 + \lambda||X||_p$。这里需要注意的一点是，正则项的形式对于我们理解优化问题的解很有帮助，比如当 $p = 1$ 时，获得的解 x 会倾向于具有稀疏性，也就是尽量多维度为 0。但在神经网络中，我们通常没有加显式正则项，也能获得还不错的解。要理解这些解的性质，我们可以通过找出神经网络的隐式正则项。

在前面的章节中，我们已经谈到神经网络具有低频偏好的频率原则，这是从函数空间的角度发现的一种隐式正则化。这类发现，不仅可以帮助我们理解神经网络的训练过程，也可以帮助我们认识到神经网络的不足和指导设计性能更优的神经网络。但不得不说，频率原则给出的

是一种相对粗粒化的认识，它对具体的参数演化和神经网络的结构并不能给出太多的信息。

大量的实验可以发现，不同的参数初始化下，神经网络学习到的解具有很大的差异。所以当我们谈神经网络具有某种特定的性质的时候，往往不能脱离它是在什么样的初始化下才具有的性质。比如，我们说神经网络能力很强，但如果神经网络的初始化特别大的情况下，它就几乎没有任何预测能力，完全展现不出很强的能力。因此，研究神经网络的性质如何依赖初始化是非常重要的。

在这一章节中，对于全连接网络，我们关注不同超参数下，特别关注影响参数初始化的超参，神经网络表现出的不同隐式正则化，也就是神经网络关于初始化超参数的相图分析。主要的内容来源于参考文献Luo et al. (2021)。

提起相图分析，最容易想到的例子就是关于水的相图。通过给定压强和体积，可以判定水的相态，比如固态、液态或者气态。为了达到明确的相变，需要有热力学极限的假设，也就是系统的粒子数目趋于无穷大，同时，粒子数目和体积的比例趋于常数。在统计力学相图的启发下，我们研究神经网络宽度趋于无穷的情况，这样可以得到非常干净的相变。对于实际使用的神经网络，由于宽度是有限的，因此，相变的分界会逐渐模糊，但极限情况的相图分析仍然可以给出定性的理解。

1.2 模型介绍

我们以两层无限宽 ReLU 神经网络为切入点，

$$f_{\boldsymbol{\theta}}^{\alpha}(\boldsymbol{x}) = \frac{1}{\alpha}\sum_{k=1}^{m} a_k \sigma(\boldsymbol{w}_k^{\intercal}\boldsymbol{x}), \tag{1.1}$$

其中 $\boldsymbol{x} \in \mathbb{R}^d$, α 是尺度因子, 参数集为 $\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{\theta}_a, \boldsymbol{\theta}_{\boldsymbol{w}})$，其中 $\boldsymbol{\theta}_a = \mathrm{vec}(\{a_k\}_{k=1}^m)$，$\boldsymbol{\theta}_{\boldsymbol{w}} = \mathrm{vec}(\{\boldsymbol{w}_k\}_{k=1}^m)$ 是通过 $a_k^0 \sim N(0, \beta_1^2)$ 和 $\boldsymbol{w}_k^0 \sim N(0, \beta_2^2 \boldsymbol{I}_d)$ 进行初始化的。偏置项 b_k 可以通过将 $\boldsymbol{x}$ 和 $\boldsymbol{w}_k$ 扩展为 $(\boldsymbol{x}^{\intercal}, 1)^{\intercal}$ 和 $(\boldsymbol{w}_k^{\intercal}, b_k)^{\intercal}$ 的形式来加入。

我们先解释一下模型的线性和非线性的含义。模型关于输入 $\boldsymbol{x}$ 是非线性的，否则该模型只能拟合线性函数。在本书中，当我们说一个模型是线性的时候，我们指的是，对任意一个参数，模型对该参数的导数不依赖于该参数本身的值。比如，对于模型 (1.1), 如果只有 a 是可训练的参数，而所有的 $\boldsymbol{w}$ 均不可训练的，那么该模型是线性的，简单的证明如下，对于任意的 a_k, $f_{\boldsymbol{\theta}}^{\alpha}(\boldsymbol{x})$ 对于 a_k 的导数为 $\frac{1}{\alpha}\sigma(\boldsymbol{w}_k^{\intercal}\boldsymbol{x})$，与 a_k 无关，所以模型是线性的。当激活函数是非线性函数时，模型对于任意 $\boldsymbol{w}_k$ 的导数为 $\frac{1}{\alpha}\sum_{k=1}^m a_k \sigma'(\boldsymbol{w}_k^{\intercal}\boldsymbol{x})\boldsymbol{x}$，这是与 $\boldsymbol{w}_k$ 有关的函数，因此，若任意 $\boldsymbol{w}_k$ 是可训练的参数，则模型是非线性的。

但考虑一种特殊的情况，如果在训练过程中尽管 $\boldsymbol{w}_k$ 是可训练的参数，但 $\boldsymbol{w}_k$ 的变化量极

其的小，在任意训练时刻 t，神经网络模型 $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ 可以被如下线性模型非常好的近似，

$$f_{\boldsymbol{\theta}}^{\text{lin}}(\boldsymbol{x}) = f_{\boldsymbol{\theta}(0)}(\boldsymbol{x}) + \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}(0)}(\boldsymbol{x}) \cdot (\boldsymbol{\theta}(t) - \boldsymbol{\theta}(0)). \tag{1.2}$$

一般来说, 这种线性行为只有当 $\boldsymbol{\theta}(t)$ 始终保持在 $\boldsymbol{\theta}(0)$ 的一个小邻域内时才会发生, 这样一阶 Taylor 展开才是一个好的逼近。对于两层神经网络, 因为其输出层总是线性的, 这个小邻域的要求简化为对输入权重的要求, 即 $\boldsymbol{\theta}_{\boldsymbol{w}}(t)$ 始终保持在 $\boldsymbol{\theta}_{\boldsymbol{w}}(0)$ 的一个邻域内。因此，我们把这个线性近似成立的区域称为线性区域。

具体地，我们主要关注超参数 α, β_1 和 β_2 是如何影响模型的线性与非线性的。正如前一段分析的，分析的关键在于 $\boldsymbol{w}_k$ 的变化。极端情况下，$\boldsymbol{w}_k$ 在训练过程中不会发生变化，那这个模型就是线性的，也称为随机特征模型 (Random feature model)。当神经网络宽度 $m \to \infty$, 并且我们给予 $\beta_1, \beta_2 \sim O(1)$, 对于 $\alpha = \sqrt{m}$，在梯度流的训练中，随着宽度趋于无穷大，$\boldsymbol{w}_k$ 在训练过程中的变化量趋于 0，因此，模型趋于线性，这也是常说的神经正切核（NTK）的线性动力学模型 (Jacot et al. 2018, Arora et al. 2019, Zhang et al. 2020)，而当 $\alpha = m$ 的平均场设定下，随着宽度 m 的变化，$\boldsymbol{w}_k$ 的变化量一直是量级为 $O(1)$ 的常数，神经网络的梯度流展现出高度非线性的平均场特征 (Mei et al. 2018, Rotskoff & Vanden-Eijnden 2018, Chizat & Bach 2018, Sirignano & Spiliopoulos 2020)。

简单地说，我们可以把关于初始化超参的相图分为线性区别和非线性区域。在后面的章节中，我们将展示不同区域对应着不同的隐式正则化。如图 (1.1) 所示，在线性区域，神经网络会找距离初始化距离最近的解，在非线性区域，神经网络会找参数凝聚的解。

判断一个两层神经网络模型是线性还是非线性的核心是刻画 $\boldsymbol{w}_k$ 在训练过程中的变化程度。在下面的篇幅中，我们首先寻找哪些量可以用来作为有效地刻画相图的坐标，以及相图如何划分。

1.3 两层无穷宽 ReLU 网络的无量纲化分析

在统计力学中，为了看到相变的发生，需要让粒子的数目趋于无穷。类似地，我们也考虑神经元数目趋于无穷的情况。为了描述一个热力学系统的相态，必须找到合适的状态量。比如当水处于标准大气压下时，温度为室温 23 度时，我们便知道它处于液态。如果温度是零下 50 度，那应该是固态。类似地，我们能否针对神经网络定义一些统计量，通过统计量判断神经网络处于的状态，比如线性或者非线性。在 $m \to \infty$ 时求取相图坐标位置有以下几条指导原则:

1. 它们应该是独立的，一个坐标的值不能推出另一个坐标的值。

2. 对于相图中的一个具体坐标, 所有对应的神经网络的学习动力学在统计上应该具有相似性, 除了时间尺度上的区别。

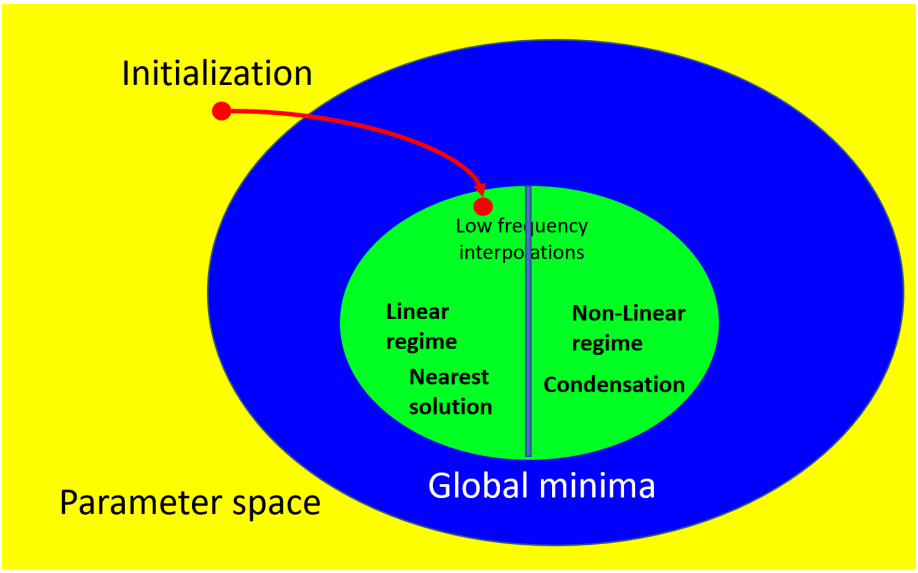


图 1.1: 不同区域的隐式正则化。

3. 除了时间尺度上的差异外, 它们应该能很好地区分动力学上的差异。

以水为例，温度和压强是两个合适的状态量，它们可以独立变化，比如固定温度，压强可以通过改变体积自由变化；不同体积的系统，只要它们的温度和压强一样，它们的状态应该是类似的；通过温度和压强可以很好的区分各种相态。

根据以上原则, 在本节中, 我们执行以下的尺度放缩过程, 以便在不同超参数选择之间进行公平比较, 并获得一个规范化的模型, 其中有两个与梯度流动力学的时间尺度无关的独立的量,。我们从原始模型 (1.1)开始，其定义于给定的样本集 $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$，其中 $\boldsymbol{x}_i \in \mathbb{R}^d$,$i \in [n]$，网络宽度为 m，缩放参数为 $1/\alpha$ 以及 $\sigma = \mathrm{ReLU}$。参数初始化为

$$a_k^0 \sim N(0, \beta_1^2), \quad \boldsymbol{w}_k^0 \sim N(0, \beta_2^2 \boldsymbol{I}_d), \tag{1.3}$$

其中 a_k 和 $\boldsymbol{w}_k$ 分别按不同的尺度 β_1 和 β_2 进行了初始化。损失函数为

$$R_S(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^n (f_{\boldsymbol{\theta}}^\alpha(\boldsymbol{x}_i) - y_i)^2. \tag{1.4}$$

那么基于梯度下降，在连续时间的极限下，其动力学遵守如下关于 $\boldsymbol{\theta}$ 的梯度流，

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = -\nabla_{\boldsymbol{\theta}} R_S(\boldsymbol{\theta}). \tag{1.5}$$

更详细地说, 令 $\boldsymbol{\theta}=\mathrm{vec}(\{\boldsymbol{q}_k\}_{k=1}^m)$, 其中 $\boldsymbol{q}_k=(a_k,\boldsymbol{w}_k^\intercal)^\intercal$,$k\in[m]$, 则 $\boldsymbol{\theta}$ 满足

$$\begin{aligned}\frac{\mathrm{d}a_k}{\mathrm{d}t}&=-\frac{1}{n}\sum_{i=1}^n\frac{1}{\alpha}\sigma(\boldsymbol{w}_k^\intercal\boldsymbol{x}_i)\left(\frac{1}{\alpha}\sum_{k'=1}^m a_{k'}\sigma(\boldsymbol{w}_{k'}^\intercal\boldsymbol{x}_i)-y_i\right)\\ \frac{\mathrm{d}\boldsymbol{w}_k}{\mathrm{d}t}&=-\frac{1}{n}\sum_{i=1}^n\frac{1}{\alpha}a_k\sigma'(\boldsymbol{w}_k^\intercal\boldsymbol{x}_i)\boldsymbol{x}_i\left(\frac{1}{\alpha}\sum_{k'=1}^m a_{k'}\sigma(\boldsymbol{w}_{k'}^\intercal\boldsymbol{x}_i)-y_i\right).\end{aligned}$$

令

$$\bar{a}_k=\beta_1^{-1}a_k,\quad \bar{\boldsymbol{w}}_k=\beta_2^{-1}\boldsymbol{w}_k,\quad \bar{t}=\frac{1}{\beta_1\beta_2}t, \tag{1.6}$$

那么

$$\begin{aligned}\frac{\mathrm{d}\bar{a}_k}{\mathrm{d}\bar{t}}&=-\frac{\beta_2}{\beta_1}\frac{1}{n}\sum_{i=1}^n\frac{\beta_1\beta_2}{\alpha}\sigma(\bar{\boldsymbol{w}}_k^\intercal\boldsymbol{x}_i)\left(\frac{\beta_1\beta_2}{\alpha}\sum_{k'=1}^m\bar{a}_{k'}\sigma(\bar{\boldsymbol{w}}_{k'}^\intercal\boldsymbol{x}_i)-y_i\right),\\ \frac{\mathrm{d}\bar{\boldsymbol{w}}_k}{\mathrm{d}\bar{t}}&=-\frac{\beta_1}{\beta_2}\frac{1}{n}\sum_{i=1}^n\frac{\beta_1\beta_2}{\alpha}\bar{a}_k\sigma'(\bar{\boldsymbol{w}}_k^\intercal\boldsymbol{x}_i)\boldsymbol{x}_i\left(\frac{\beta_1\beta_2}{\alpha}\sum_{k'=1}^m\bar{a}_{k'}\sigma(\bar{\boldsymbol{w}}_{k'}^\intercal\boldsymbol{x}_i)-y_i\right).\end{aligned}$$

我们引入两个缩放参数

$$\kappa:=\frac{\beta_1\beta_2}{\alpha},\quad \kappa':=\frac{\beta_1}{\beta_2}, \tag{1.7}$$

其中 κ 和 κ' 分别称为能量缩放参数和动力学缩放参数。则上述动力学可以写为

$$\frac{\mathrm{d}\bar{a}_k}{\mathrm{d}\bar{t}}=-\frac{1}{\kappa'}\frac{1}{n}\sum_{i=1}^n\kappa\sigma(\bar{\boldsymbol{w}}_k^\intercal\boldsymbol{x}_i)\left(\kappa\sum_{k'=1}^m\bar{a}_{k'}\sigma(\bar{\boldsymbol{w}}_{k'}^\intercal\boldsymbol{x}_i)-y_i\right), \tag{1.8}$$

$$\frac{\mathrm{d}\bar{\boldsymbol{w}}_k}{\mathrm{d}\bar{t}}=-\kappa'\frac{1}{n}\sum_{i=1}^n\kappa\bar{a}_k\sigma'(\bar{\boldsymbol{w}}_k^\intercal\boldsymbol{x}_i)\boldsymbol{x}_i\left(\kappa\sum_{k'=1}^m\bar{a}_{k'}\sigma(\bar{\boldsymbol{w}}_{k'}^\intercal\boldsymbol{x}_i)-y_i\right). \tag{1.9}$$

在本文余下的讨论中, 我们将指的是这个初始无量纲化后的模型 (1.9), 称为无量纲化模型, 并为简洁起见, 省略所有的 κ 上标以及 a_k、$\boldsymbol{w}_k$、t 中的 "$\bar{\cdot}$"。注意,κ 和 κ' 在宽度无限的极限下, 它们的值都趋于常数或者无穷, 因此不遵循上述原则 (2) 和 (3)。例如, 对 NTK 和平均场模型来说, 它们都有 $\kappa=0$ 且 $\kappa'=1$, 但是它们的训练行为存在明显差异。为了解释这种动力学差异, 我们考虑 κ 和 κ' 和 m 如下的极限关系

$$\gamma=\lim_{m\to\infty}-\frac{\log\kappa}{\log m},\quad \gamma'=\lim_{m\to\infty}-\frac{\log\kappa'}{\log m}, \tag{1.10}$$

如后续的理论和实验结果所示, 这些量满足了上述的所有原则, 并且能够把一些常用的初始化方法都包括进来。

注记 1. 这里我们列出了一些常用的初始化方法及相关工作, 以及它们对应的缩放参数, 如表 *1.1*所示。

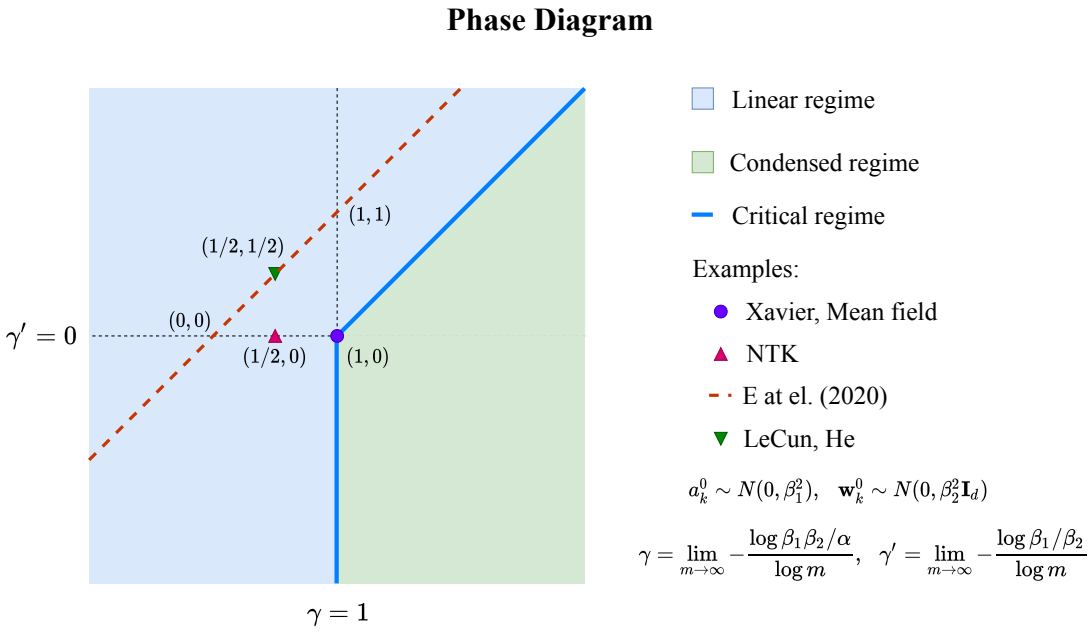


图 1.2: 两层无穷宽 ReLU 神经网络的相图。符号标出的例子可以在表格 (1.1) 找到对应。

图 (1.2) 展示了两层无穷宽 ReLU 神经网络的相图结果，这些结果将在后面的篇幅中具体论证。请读者注意一个细节，我们把非线性区域称为凝聚区域 (condensed regime)，因为我们在实验上发现在非线性区域，参数都会发生一种称为凝聚的现象，这是后面几章的重点。

1.4 典型例子

使用 γ 和 γ' 作为坐标, 在本节中, 我们通过实验来说明相图中各种典型情况在简单的 4 个训练点的 1 维问题上的行为, 这个例子能够进行易于理解的可视化。图 1.3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。

1.5 划分相图的关键量

刻画模型的线性与否关于在于其能否被线性模型 (1.2) 准确地逼近，而这个逼近效果的好坏取决于参数 $\boldsymbol{w}$ 在训练后的变化程度。因此我们使用以下相对距离作为 $\boldsymbol{\theta}_{\boldsymbol{w}}(t)$ 在训练期间偏

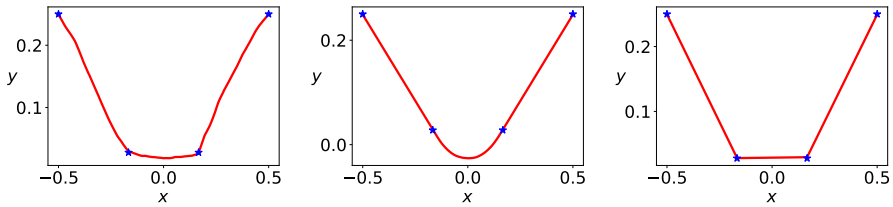
Name (相关工作)	α	β_1	β_2	κ $\left(\frac{\beta_1\beta_2}{\alpha}\right)$	κ' $\left(\frac{\beta_1}{\beta_2}\right)$	γ $\left(\lim_{m\to\infty}\frac{\log 1/\kappa}{\log m}\right)$	γ' $\left(\lim_{m\to\infty}\frac{\log 1/\kappa'}{\log m}\right)$
LeCun (LeCun et al. 2012)	1	$\sqrt{\frac{1}{m}}$	$\sqrt{\frac{1}{d}}$	$\sqrt{\frac{1}{md}}$	$\sqrt{\frac{d}{m}}$	$\frac{1}{2}$	$\frac{1}{2}$
He (He et al. 2015)	1	$\sqrt{\frac{2}{m}}$	$\sqrt{\frac{2}{d}}$	$\sqrt{\frac{4}{md}}$	$\sqrt{\frac{d}{m}}$	$\frac{1}{2}$	$\frac{1}{2}$
Xavier (Glorot & Bengio 2010)	1	$\sqrt{\frac{2}{m+1}}$	$\sqrt{\frac{2}{m+d}}$	$\sqrt{\frac{4}{(m+1)(m+d)}}$	$\sqrt{\frac{m+d}{m+1}}$	1	0
NTK (Jacot et al. 2018)	$\sqrt{m}$	1	1	$\sqrt{\frac{1}{m}}$	1	$\frac{1}{2}$	0
Mean-field (Mei et al. 2018) (Sirignano & Spiliopoulos 2020) (Rotskoff & Vanden-Eijnden 2018)	m	1	1	$\frac{1}{m}$	1	1	0
E et al. (Weinan et al. 2019)	1	β	1	β	β	$\lim_{m\to\infty}\frac{\log 1/\beta}{\log m}$	$\lim_{m\to\infty}\frac{\log 1/\beta}{\log m}$

表 1.1: Initialization methods with their scaling parameters

离 $\boldsymbol{\theta}_{\boldsymbol{w}}(0)$ 的程度的指标

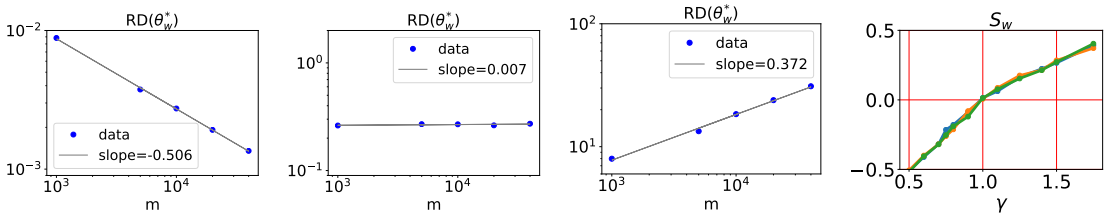
$$\mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) = \frac{\|\theta_{\boldsymbol{w}}(t) - \theta_{\boldsymbol{w}}(0)\|_2}{\|\theta_{\boldsymbol{w}}(0)\|_2}. \tag{1.11}$$

具体地, 我们关注 $\sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t))$, 它表示在训练过程中 $\boldsymbol{\theta}_{\boldsymbol{w}}(t)$ 从初始化的最大偏移距离。当 $m \to \infty$ 时, 如果 $\sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) \to 0$, 则神经网络训练动力学属于线性域。否则, 如果它接近 $O(1)$ 或 $+\infty$, 则神经网络训练动力学是非线性的。注意, 对于后者, 即 $\boldsymbol{\theta}_{\boldsymbol{w}}$ 与初始化偏离到无穷远的情况, 可以观察到特征空间中非常强的非线性凝聚动力学行为, 因此我们



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

图 1.3: 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。



(a) $\gamma = 0.5$ (b) $\gamma = 1$ (c) $\gamma = 1.75$ (d) S_w vs. γ

图 1.4: 随着 $m \to \infty$, $\mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}^*)$ 与 m 的关系。对于 (a-c), $\beta_1 = 1$, $\beta_2 = 1$, 且五个蓝点分别表示隐层神经元数为 $1000, 5000, 10000, 20000, 40000$ 的神经网络上的 $\mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}^*)$ 关于 m 的曲线。灰线是数据点的线性拟合并在标签中给出斜率标注。对于 (d), 曲线表示 $S_{\boldsymbol{w}}$ 对 γ 的关系, 其中 $\gamma' = 0$。每条线对应一对 β_1 和 β_2 的值: 蓝色:$\beta_1 = 1$,$\beta_2 = 1$; 橙色:$\beta_1 = m^{-1/2}$,$\beta_2 = m^{-1/2}$; 绿色:$\beta_1 = m^{-1}$,$\beta_2 = m^{-1}$。注意 α 由 $\alpha = \beta_1\beta_2 m^{\gamma}$ 确定。

将 $\sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) \to +\infty$ 的区域称为凝聚区域, 这将在后续的详细实验中得到证实。对于 $\sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) \to O(1)$, 神经网络表现出中间水平的非线性行为。我们将这个区域称为临界区域。

在下文中, 我们将通过实验和理论方法准确地在相图中分离线性区域和凝聚区域。

1.6 实验上划分相图

为了实验区分线性区域和非线性区域, 我们需要估计

$$\sup_{t\in[0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)),$$

经验上, 这可以用 $\mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}^*)(\boldsymbol{\theta}_{\boldsymbol{w}}^* := \boldsymbol{\theta}_{\boldsymbol{w}}(\infty))$ 来逼近。不失一般性, 因为我们不可能在 $m \to \infty$ 下运行实验, 所以我们用 $\mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}^*)$ 随 $m \to \infty$ 的增长来量化。由图 (1.4(a-c)) 可见, 它们近似满足幂律关系。因此, 我们定义

$$S_{\boldsymbol{w}} = \lim_{m\to\infty} \frac{\log \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}^*)}{\log m}, \tag{1.12}$$

这可以通过估计如图 1.4中的对数坐标图的斜率来经验地得到 $\mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}^*)$ 在 $m \to \infty$ 的情况。如图 1.4(d) 所示, 对于具有相同的 γ 和 γ' 但不同 α、β_1 和 β_2 的神经网络, 其 S_w 非常相似, 这验证了归一化模型的有效性。在下面的实验中, 我们对于每对 γ 和 γ' 只展示一组 α、β_1 和 β_2 的结果。

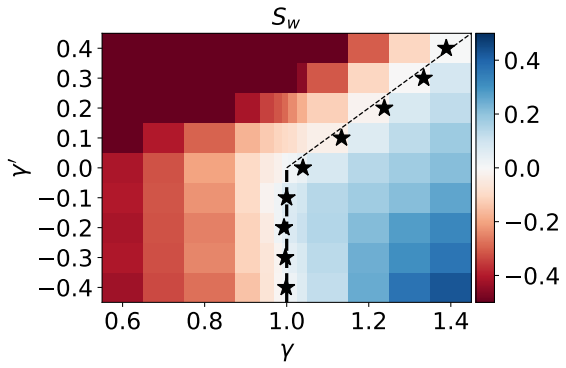


图 1.5: 对于合成数据, $S_{\boldsymbol{w}}$ 在具有 1000, 5000, 10000, 20000, 40000 个隐层神经元的两层 ReLU 神经网络上估计, 横坐标为 γ, 纵坐标为 γ'。星号表示在固定 γ' 下通过线性插值获得的零点。虚线是理论获得的边界。

接下来, 我们通过在相空间上实验扫描 S_w 来可视化相图。与图1.3相同的 1 维问题的结果如图 1.5所示。在红色区域, 其中 $S_{\boldsymbol{w}}$ 小于零,RD($\boldsymbol{\theta}^*_{\boldsymbol{w}}$) 随 $m \to \infty$ 趋于 0, 表示线性区域。相反, 在蓝色区域, 其中 $S_{\boldsymbol{w}}$ 大于零,RD($\boldsymbol{\theta}^*_{\boldsymbol{w}}$) 随 $m \to \infty$ 趋于无穷大, 表示高度非线性行为。这两种区域的边界通过插值可以近似获得, 在图 1.5中用星号表示, 其中 RD($\boldsymbol{\theta}^*_{\boldsymbol{w}}$) $\sim O(1)$。

同样, 我们使用两层 ReLU 神经网络拟合 MNIST 数据集, 采用均方误差损失。在我们的实验中, 输入是 784 维向量, 输出是输入图像的一维标签 ($0 \sim 9$)。如图1.6所示, 通过 1 维数据获得的相图与实际高维数据集得到的结果非常类似。

1.7 两层无穷宽 ReLU 网络的相图的尺度分析

在我们进行详细分析之前, 通过直观的尺度分析, 我们首先说明线性域和凝聚域之间的区分。在初始化附近, 两层 ReLU 神经网络的拟合能力, 即可以拟合的目标函数的量级, 可以粗略估计为

$$C = m\beta_1\beta_2/\alpha = m\kappa.$$

不失一般性, 目标函数总是 $O(1)$。因此, 线性域的必要条件是神经网络有在初始化附近拟合目标的能力, 即 $C \gtrsim O(1)$。因此,

$$\kappa \gtrsim 1/m,$$

这导致在 $m \to \infty$ 时 $\gamma \leq 1$。我们进一步注意到, 输出层总是线性的。因此, 即使输出权重 $\boldsymbol{\theta}_a$ 发生显著变化, 如果输入层权重 $\boldsymbol{\theta}_{\boldsymbol{w}}$ 保持在初始化附近, 动力学仍然可以线性化。如动力学方程 (1.9) 所示, 当 (i) 初始化时 $\kappa' \ll 1$ ($\beta_1 \ll \beta_2$),a 很小，$\boldsymbol{w}$ 很大，这样，在训练中，主要的变

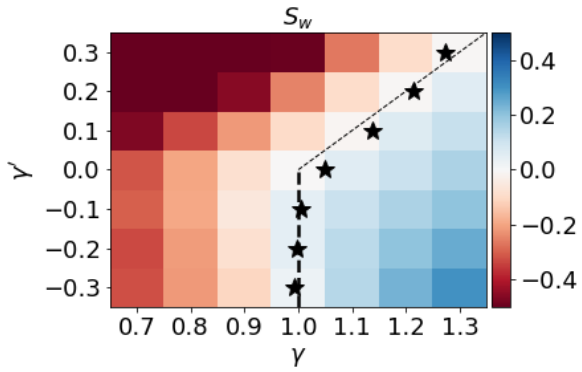


图 1.6: 对于 MNIST 数据, $S_{\boldsymbol{w}}$ 在具有 1000, 10000, 50000, 250000, 40000 个隐层神经元的两层 ReLU 神经网络上估计, 横坐标为 γ, 纵坐标为 γ'。星号表示在固定 γ' 下通过线性插值获得的零点。虚线是理论获得的边界。

化是 a，模型仍然有机会是线性的，且 (ii) a 的量级, 例如用期望 $\mathbb{E}(|a|)$ 表示, 在整个训练过程中满足 $\mathbb{E}(|a|) \ll \beta_2$ 时, 这是可能的。在这种情况下, 在训练结束时,

$$C = m\beta_2\mathbb{E}(|a|)/\alpha \ll m\beta_2^2/\alpha = m\kappa/\kappa'. \tag{1.13}$$

因为 $C \gtrsim O(1)$，我们有

$$1/\kappa' \gg 1/m\kappa, \tag{1.14}$$

在 $m \to \infty$ 时, 对于 $\gamma' > 0$，这导出了条件 $\gamma' > \gamma - 1$。

相反, 如果 $\gamma' < \gamma - 1$ 和 $\gamma > 1$, 即当 $m \to \infty$ 时,$m\kappa \ll 1$ 和 $m\kappa/\kappa' \ll 1$, 则在 $\boldsymbol{\theta}_{\boldsymbol{w}}$ 保持在初始化附近时, 神经网络没有拟合 $O(1)$ 目标的能力。神经网络的拟合能力必须经历巨大的增长才能拟合数据, 这是凝聚态的特征。

上述尺度分析直观地论证了相图中界线 $\gamma = 1$(对于 $\gamma' \leq 0$) 和 $\gamma' = \gamma - 1$(对于 $\gamma' > 0$) 将线性域和凝聚域分开的关键性。

1.8 两层 ReLU 神经网络的理论结果

直观的尺度分析和实验结果展示出一致的边界来区分线性区域和凝聚区域。一个自然的问题是——是否存在一个理论来使直观的尺度分析成为严格的, 并将上述一个维例子的经验相图推广到一般高维数据在两层 ReLU 神经网络上的情况。接下来, 我们通过提供两个非正式陈述的定理来回答这个问题。

Theorem 1*. 如果 $\gamma < 1$ 或 $\gamma' > \gamma - 1$, 那么对于 $\boldsymbol{\theta}^0$ 的选择, 我们有较高的概率

$$\lim_{m \to +\infty} \sup_{t \in [0, +\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) = 0. \tag{1.15}$$

Theorem 2*. 如果 $\gamma > 1$ 并且 $\gamma' < \gamma - 1$, 那么对于 $\boldsymbol{\theta}^0$ 的选择, 我们有较高的概率

$$\lim_{m \to +\infty} \sup_{t \in [0, +\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) = +\infty. \tag{1.16}$$

注记 2. $\lim\limits_{m \to +\infty} \sup\limits_{t \in [0, +\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t))$ 就像统计力学中相变分析的一个序参量, 它对于区域分离至关重要, 并在边界上呈现不连续性。

参考文献

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R. & Wang, R. (2019), 'On exact computation with an infinitely wide neural net', *Advances in neural information processing systems* **32**.

Chizat, L. & Bach, F. (2018), 'On the global convergence of gradient descent for over-parameterized models using optimal transport', *Advances in neural information processing systems* **31**.

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. (2015), Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, *in* 'Proceedings of the IEEE international conference on computer vision', pp. 1026–1034.

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

LeCun, Y. A., Bottou, L., Orr, G. B. & Müller, K.-R. (2012), Efficient backprop, *in* 'Neural networks: Tricks of the trade', Springer, pp. 9–48.

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.

Mei, S., Montanari, A. & Nguyen, P.-M. (2018), 'A mean field view of the landscape of two-layer neural networks', *Proceedings of the National Academy of Sciences* **115**(33), E7665–E7671.

Rotskoff, G. & Vanden-Eijnden, E. (2018), Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks, *in* 'Advances in neural information processing systems', pp. 7146–7155.

Sirignano, J. & Spiliopoulos, K. (2020), 'Mean field analysis of neural networks: A central limit theorem', *Stochastic Processes and their Applications* **130**(3), 1820–1852.

Weinan, E., Ma, C. & Wu, L. (2019), 'A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics', *Sci. China Math* .

Zhang, Y., Xu, Z.-Q. J., Luo, T. & 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.