第七章　网络优化与正则化

虽然神经网络具有非常强的表达能力，但是当应用神经网络模型到机器学习时依然存在一些难点。主要分为两大类：（1）优化问题：神经网络模型是一个非凸函数，再加上在深度网络中的梯度消失问题，很难进行优化；另外，深度神经网络模型一般参数比较多，训练数据也比较大，会导致训练的效率比较低。（2）泛化问题：因为神经网络的拟合能力强，反而容易在训练集上产生过拟合。因此，在训练深度神经网络时，同时也需要掌握一定的技巧。目前，人们在大量的实践中总结了一些经验技巧，从优化和正则化两个方面来提高学习效率并得到一个好的网络模型。

7.1　网络优化

深度神经网络是一个高度非线性的模型，其风险函数也是一个非凸问题。在非凸问题中，一个会存在一些局部最优点。

有效地学习深度神经网络的参数是一个具有挑战性的问题，其主要原因有以下几个方面。

网络结构多样性　神经网络的种类非常多，比如卷积网络、循环网络等，其结构也非常不同。有些比较深，有些比较宽。不同参数在网络中的作用也有很大的差异，比如连接权重和偏置的不同，以及循环网络中循环连接上的权重和其它权重的不同。

网络结构的多样性导致了很难找到一种通用的优化方法。不同的优化方法在不同网络结构上的差异也都比较大。

此外，网络的超参数一般也比较多，这也给优化带来很大的挑战。

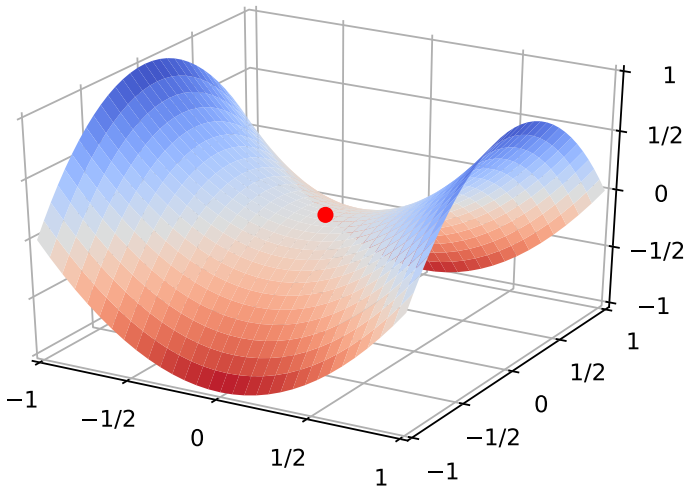


图 7.1: 鞍点示例。

高维变量的非凸优化 深度神经网络的参数非常多，其参数学习是在非常高维空间中的非凸优化问题，其挑战和在低维空间的非凸优化问题有所不同。低维空间的非凸优化问题主要是存在一些局部最优点。采用梯度下降方法时，不合适的参数初始化会导致陷入局部最优点，因此主要的难点是如何选择初始化参数和逃离局部最优点。

Dauphin et al. [2014] 指出在高维空间中，非凸优化的难点并不在于如果逃离局部最优点，而是如何逃离鞍点。鞍点（saddle point）是梯度为0，但是在一些维度上是最高点，在另一些维度上是最低点，如图7.1所示。梯度下降方法同样很难从这些鞍点中逃离。

鞍点的叫法是因为其形状像马鞍。

目前，深度神经网络的参数学习主要是通过梯度下降方法来寻找一组可以最小化结构风险的参数。在具体实现中，梯度下降法可以分为：批量梯度下降、随机梯度下降以及小批量梯度下降三种形式。根据不同的数据量和参数量，可以选择一种具体的实现形式。除了在收敛效果和效率上的差异，这三种方法都存在一些共同的问题，比如1）如何初始化参数；2）预处理数据；3）如何选择合适的学习率，避免陷入局部最优等。

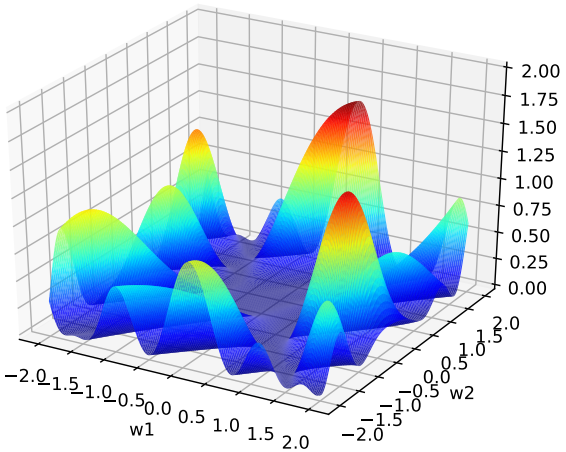


图 7.2: 神经网络中的非凸优化问题。

7.1.1 参数初始化

神经网络的训练过程中的参数学习是基于梯度下降法进行优化的。梯度下降法需要在开始训练时给每一个参数赋一个初始值。这个初始值的选取十分关键。在感知器和 logistic 回归的训练中，我们一般将参数全部初始化为 0。但是这在神经网络的训练中会存在一些问题。因为如果参数都为 0，在第一遍前向计算时，所有的隐层神经元的激活值都相同。这样会导致深层神经元没有区分性。这种现象也称为**对称权重**现象。

为了打破这个平衡，比较好的方式是对每个参数都随机初始化，这样使得不同神经元之间的区分性更好。

但是一个问题是如何选取随机初始化的区间呢？如果参数太小，会导致神经元的输入过小。经过多层之后信号就慢慢消失了。参数过小还会使得 sigmoid 型激活函数丢失非线性的能力。以 logistic 函数为例，在 0 附近基本上是近似线性的。这样多层神经网络的优势也就不存在了。如果参数取得太大，会导致输入状态过大。对于 sigmoid 型激活函数来说，激活值变得饱和，从而导致梯度接近于 0。

因此，如果要高质量地训练一个网络，给参数选取一个合适的初始化区间是非常重要的。一般而言，参数初始化的区间应该不用神经元的性质进行差异化的设置。如果一个神经元的输入连接很多，它的每个输入连接上的权重就应该小一些，以避免神经元的输出过大（当激活函数为ReLU时）或过饱和（当激活函数为sigmoid函数时）。

经常使用的初始化方法有以下几种：

Gaussian 分布初始化

Gaussian初始化方法是最简单的初始化方法，参数从一个固定均值（比如0）和固定方差（比如0.01）的Gaussian分布进行随机初始化。

初始化一个深度网络时，一个比较好的初始化方案是保持每个神经元输入的方差为一个常量。当一个神经元的输入连接数量为n_{in}时，可以设置其输入连接权重以$\mathcal{N}(0, \sqrt{\frac{1}{n_{in}}})$的Gaussian分布进行初始化。如果同时考虑输出连接的数量n_{out}，则可以按$\mathcal{N}(0, \sqrt{\frac{2}{n_{in}+n_{out}}})$的Gaussian分布进行初始化。

均匀分布初始化

均匀分布初始化是在一个给定的区间$[-r, r]$内采用均匀分布来初始化参数。超参数r的设置也可以按神经元的连接数量进行自适应的调整。

Xavier初始化方法中，Xavier是发明者Xavier Glorot的名字。

***Xavier*初始化方法** Glorot and Bengio [2010]提出一个自动计算超参数r的方法，参数可以在$[-r, r]$内采用均匀分布进行初始化。

如果神经元激活函数为logistic函数，对于第$l-1$到l层的权重参数区间r可以设置为

$$r = \sqrt{\frac{6}{n^{l-1}+n^1}}, \tag{7.1}$$

这里n^l是第l层神经元个数，n^{l-1}是第$l-1$层神经元个数。

对于tanh函数，r可以设置为

$$r = 4\sqrt{\frac{6}{n^{l-1}+n^l}}. \tag{7.2}$$

假设第 l 层的一个隐藏层神经元 z^l，其接受前一层的 n^{l-1} 个神经元的输出 $a_i^{(l-1)}$，$i \in [1, n^{(l-1)}]$，

$$z^l = \sum_{i=1}^{n} w_i^l a_i^{(l-1)} \tag{7.3}$$

为了避免初始化参数使得激活值变得饱和，我们需要尽量使得 z^l 处于激活函数的线性区间，也就是其绝对值比较小的值。这时该神经元的激活值为 $a^l = f(z^l) \approx z^l$。

假设 w_i^l 和 $a_i^{(l-1)}$ 都是相互独立，并且均值都为 0，则 a 的均值为

$$\mathbb{E}[a^l] = \mathbb{E}[\sum_{i=1}^{n} w_i^l a_i^{(l-1)}] = \sum_{i=1}^{d} \mathbb{E}[\mathbf{w}_i]\mathbb{E}[a_i^{(l-1)}] = 0. \tag{7.4}$$

a^l 的方差为

$$\begin{aligned} \mathrm{Var}[a^l] &= \mathrm{Var}[\sum_{i=1}^{n^{(l-1)}} w_i^l a_i^{(l-1)}] && (7.5)\\ &= \sum_{i=1}^{n^{(l-1)}} \mathrm{Var}[w_i^l]\,\mathrm{Var}[a_i^{(l-1)}] && (7.6)\\ &= n^{(l-1)}\,\mathrm{Var}[w_i^l]\,\mathrm{Var}[a_i^{(l-1)}]. && (7.7) \end{aligned}$$

也就是说，输入信号的方差在经过该神经元后被放大或缩小了 $n^{(l-1)}\,\mathrm{Var}[w_i^l]$ 倍。为了使得在经过多层网络后，信号不被过分放大或过分减弱，我们尽可能保持每个神经元的输入和输出的方差一致。这样 $n^{(l-1)}\,\mathrm{Var}[w_i^l]$ 设为 1 比较合理，即

$$\mathrm{Var}[w_i^l] = \frac{1}{n^{(l-1)}}. \tag{7.8}$$

同理，为了使得在反向传播中，误差信号也不被放大或缩小，需要将 w_i^l 的方差保持为

$$\mathrm{Var}[w_i^l] = \frac{1}{n^{(l)}}. \tag{7.9}$$

作为折中，同时考虑信号在前向和反向传播中都不被放大或缩小，可以设置

$$\mathrm{Var}[w_i^l] = \frac{2}{n^{(l-1)} + n^{(l)}}. \tag{7.10}$$

假设随机变量 x 在区间 $[a,b]$ 内均匀分布，则其方差为：

$$\text{Var}[x] = \frac{(b-a)^2}{12}. \tag{7.11}$$

因此，若让 $w_i^l \in [-r,r]$，并且 $\text{Var}[w_i^l] = 1$，则 r 的取值为

$$r = \sqrt{\frac{6}{n^{l-1}+n^1}}. \tag{7.12}$$

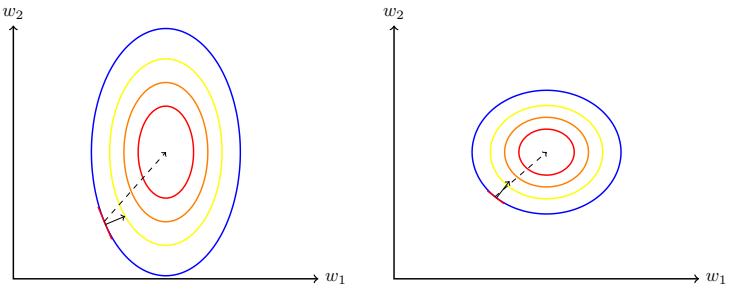
7.1.2 数据预处理

一般而言，原始的训练数据中，每一维特征的来源以及度量单位不同，会造成这些特征值的分布范围往往差异很大。当我们计算不同样本之间的欧式距离时，取值范围大的特征会起到主导作用。这样，对于基于相似度比较的机器学习方法（比如最近邻分类器），必须先对样本进行预处理，将各个维度的特征归一化到同一个取值区间，并且消除不同特征之间的相关性，才能获得比较理想的结果。虽然神经网络可以通过参数的调整来适应不同特征的取值范围，但是会导致训练效率比较低。

假设一个只有一层的网络 $y = \tanh(w_1x_1 + w_2x_2 + b)$，其中 $x_1 \in [0,10]$，$x_2 \in [0,1]$。之前我们提到 tanh 函数的导数在区间 $[-2,2]$ 上是敏感的，其余的导数接近于 0。因此，如果 $w_1x_1 + w_2x_2 + b$ 过大或过小，都会导致梯度过小，难以训练。为了提高训练效率，我们需要使 $w_1x_1 + w_2x_2 + b$ 在 $[-2,2]$ 区间，我们需要将 w_1 设得小一点，比如在 $[-0.1,0.1]$ 之间。可以想象，如果数据维数很多时，我们很难这样精心去选择每一个参数。因此，如果每一个特征的取值范围都在相似的区间，比如 $[0,1]$ 或者 $[-1,1]$，我们就不太需要区别对待每一个参数，减少人工干预。

除了参数初始化之外，不同特征取值范围差异比较大时还会梯度下降法的搜索效率。图7.3给出了数据归一化对梯度的影响。其中，图7.3a为未归一化数据的等高线图。取值范围不同会造成在大多数位置上的梯度方向并不是最优的搜索方向。当使用梯度下降法寻求最优解时，会导致需要很多次迭代才能收敛。如果我们把数据归一化为取值范围相同，如图7.3b所示，大部分位置的梯度方向近似于最优搜索方向。这样，在梯度下降求解时，每一步梯度的方向都基本指向最小值，训练效率会大大提高。

归一化的方法有很多种，比如之前我们介绍的 sigmoid 型函数等都可以将不同取值范围的特征挤压到一个比较受限的区间。这里，我们介绍几种在神经



(a) 未归一化数据的梯度 (b) 归一化数据的梯度

图 7.3: 数据归一化对梯度的影响。

网络中经常使用的归一化方法。

标准归一化 也叫 z-score 归一化，来源于统计上的标准分数。将每一个维特征都处理为符合标准正态分布（均值为0，标准差为1）。假设有 N 个样本 $\{\mathbf{x}^{(i)}\}, i = 1, \cdots, N$，对于每一维特征 x，我们先计算它的均值和标准差：

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}, \tag{7.13}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x^{(i)} - \mu)^2. \tag{7.14}$$

然后，将特征 $x^{(i)}$ 减去均值，并除以标准差，得到新的特征值 $\hat{x}^{(i)}$。

$$\hat{x}^{(i)} = \frac{x^{(i)} - \mu}{\sigma}, \tag{7.15}$$

这里，σ 不能为0。如果标准差为0，说明这一维特征没有任务区分性，可以直接删掉。

在标准归一化之后，每一维特征都服从标准正态分布。

缩放归一化 另外一种非常简单的归一化是通过缩放将特征取值范围归一到 $[0, 1]$ 或 $[-1, 1]$ 之间：

$$\hat{x}^{(i)} = \frac{x^{(i)} - \min(x)}{\max(x) - \min(x)}, \tag{7.16}$$

其中，$\min(x)$ 和 $\max(x)$ 分别为这一维特征在所有样本上的最小值和最大值。

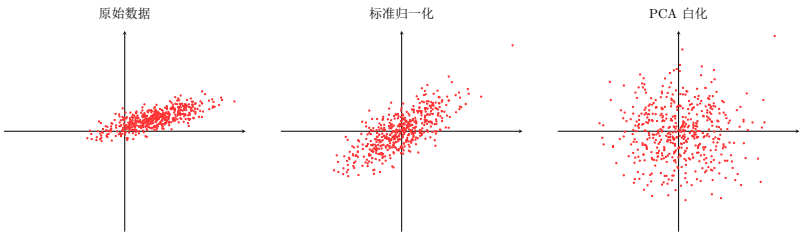


图 7.4: 数据归一化示例

白化（whitening） 白化是一种重要的预处理方法，用来降低输入数据特征之间的冗余性。输入数据经过白化处理后，特征之间相关性较低，并且所有特征具有相同的方差。

白化的一个主要实现方式是使用主成分分析（Principal Component Analysis，PCA）方法去除掉各个成分之间的相关性。

7.1.3 批量归一化

在传统机器学习中，一个常见的问题的**协变量偏移**（Covariate Shift）。协变量是一个统计学概念，是可能影响预测结果的变量。在机器学习中，协变量可以看作是输入变量。一般的机器学习算法都要求输入变量在训练集和测试集上的分布是相似的。如果不满足这个要求，这些学习算法在测试集的表现会比较差。

在多层的神经网络中，中间某一层的输入是其前面网络的输出。因为前面网络的参数在每次迭代时也会被更新，而一旦这些参数变化会造成这一个中间层的输入也发生变化，其分布往往会和参数更新之前差异比较大。换句话说，从这一个中间层开始，之后的网络参数白学了，需要重新学习。这个中间层的深度很大时，这种现象就越明显。这种现象叫做内部协变量偏移（internal covariate shift）。

为了解决这个问题，通过对每一层的输入进行归一化使其分布保存稳定，这种方法称为批量归一化（batch normalization）方法 [Ioffe and Szegedy, 2015]。

归一化方法可以采用第**??**节中介绍的几种归一化方法。因为每一层都要进行归一化，所以要求归一化方法的速度要快。这样，PCA 白化的归一化方法就

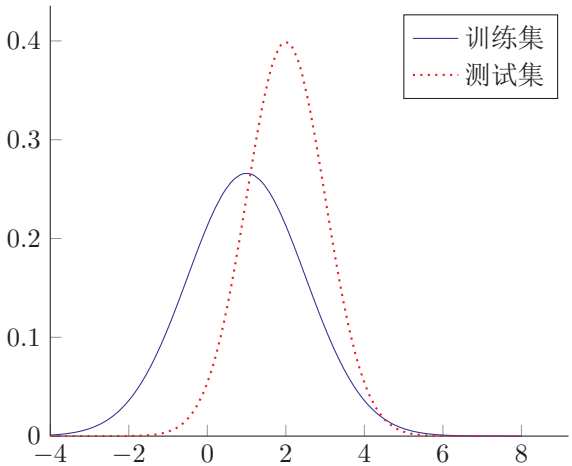


图 7.5: 协变量偏移。

不太合适。为了提高归一化效率，这里使用标准归一化，对每一维特征都归一到标准正态分布。相当于每一层都进行一次数据预处理，从而加速收敛速度。

因为标准归一化会使得输入的取值集中的 0 附近，如果使用 sigmoid 型激活函数时，这个取值区间刚好是接近线性变换的区间，减弱了神经网络的非线性性质。因此，为了使得归一化不对网络的表示能力造成负面影响，我们可以通过一个附加的缩放和平移变换改变取值区间。从最保守的角度考虑，可以通过来标准归一化的逆变换来使得归一化后的变量可以被还原为原来的值。

$$y_k = \gamma_k \hat{x}_k + \beta_k, \tag{7.17}$$

这里 γ_k 和 β_k 分别代表缩放和平移的参数。当 $\gamma_k = \sqrt{\sigma[x_k]}$，$\beta_k = \mu[x_k]$ 时，y_k 即为原始的 x_k。

当训练完成时，用整个数据集上的均值 μ_k 和方差 σ_k^2 来分别代替 $\mu_{\mathcal{B},k}$ 和方差 $\sigma_{\mathcal{B},k}^2$。

通过每一层的归一化，从而减少前面网络参数更新对后面网络输入带来的内部协变量偏移问题，提高训练效率。

算法 7.1: 批量归一化

输入: 一次 mini-batch 的样本集合：$\mathcal{B} = \{\mathbf{x}^{(i)}\}, i = 1, \cdots, m$;

参数：γ, β;

for $k = 1 \cdots K$ **do**

$$\mu_{\mathcal{B},k} = \frac{1}{m}\sum_{i=1}^{m} x_k^{(i)}, \qquad \text{// mini-batch 均值}$$

$$\sigma^2_{\mathcal{B},k} = \frac{1}{m}\sum_{i=1}^{m}(x_k^{(i)} - \mu_{\mathcal{B},k})^2. \qquad \text{//mini-batch 方差}$$

$$\hat{x}_k^{(i)} = \frac{x_k^{(i)} - \mu_{\mathcal{B},k}}{\sqrt{\sigma^2_{\mathcal{B},k} + \epsilon}}, \forall i \qquad \text{//归一化}$$

$$y_k^{(i)} = \gamma \hat{x}_k^{(i)} + \beta \equiv \mathbf{BN}_{\gamma,\beta}(\mathbf{x}^{(i)}), \forall i \qquad \text{//缩放和平移}$$

end

输出: $\{y^{(i)} = \mathbf{BN}_{\gamma,\beta}(\mathbf{x}^{(i)})\}$

7.1.4 梯度下降方法的改进

由于深度学习经常用来处理比较大规模的数据，参数也非常多，因此如果每次迭代都计算整个数据集上的梯度需要计算资源比较多。并且在大规模的数据集中，数据也非常冗余，也没有必要在整个数据集上批量计算梯度。因此，在训练深度模型时，经常使用小批量梯度下降算法。

假设 $f(\mathbf{x}^{(i)}, \theta)$ 代表神经网络，θ 为网络参数，以小批量梯度下降为例，在第 t 次迭代（epoch）时，选取 m 个训练样本 $\mathcal{I}_t = \{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^m$，首先计算梯度 $\mathbf{g}_t$

$$\mathbf{g}_t = \frac{1}{m}\sum_{i\in I_t} \frac{\partial \mathcal{L}\big(\mathbf{y}^{(i)}, f(\mathbf{x}^{(i)}, \theta)\big)}{\partial \theta} + \lambda \|\theta\|^2, \tag{7.18}$$

其中，$\mathcal{L}(\cdot,\cdot)$ 为可微分的损失函数，λ 为正则化系数。

然后使用梯度下降来更新参数，

$$\theta_t \leftarrow \theta_{t-1} - \alpha \mathbf{g}_t, \tag{7.19}$$

其中 $\alpha > 0$ 为学习率。

我们定义每次迭代时参数更新的差值 $\Delta\theta_t$ 为

$$\Delta\theta_t = \theta_t - \theta_{t-1}. \tag{7.20}$$

$\Delta\theta_t$ 为每次迭代时参数的实际更新差值，即 $\theta_t = \theta_{t-1} + \Delta\theta_t$。$\Delta\theta_t$ 和梯度 $\mathbf{g}_t$ 并不需要完全一致。在标准的小批量梯度下降中，$\Delta\theta_t = -\alpha\mathbf{g}_t$。

为了更有效地进行训练深度神经网络，在标准的小批量梯度下降方法的基础上，也经常使用一些改进方法以加快优化速度。常见的改进方法主要从以下两个方面进行改进：学习率衰减和动量法。

这些改进的优化方法也同样可以应用在批量或随机梯度下降方法上。

学习率衰减

在梯度下降中，学习率 α 的取值非常关键，如果过大就不会收敛，如果过小则收敛速度太慢。从经验上看，学习率在一开始要保持大些来保证收敛速度，在收敛到最优点附近时要小些以避免来回震荡。因此，比较简单直接的学习率调整可以通过学习率衰减（learning rate decay）的方式来实现。

假设初始化学习率为 α_0，在第 t 次迭代时的学习率 α_t。常用的衰减方式为可以设置为按迭代次数进行衰减。比如反时衰减（inverse time decay）

$$\alpha_t = \alpha_0 \frac{1}{1 + \beta \times t}, \tag{7.21}$$

或指数衰减（exponential decay）

$$\alpha_t = \alpha_0 \beta^t, \tag{7.22}$$

或自然指数衰减（natural exponential decay）

$$\alpha_t = \alpha_0 \exp(-\beta \times t), \tag{7.23}$$

其中 β 为衰减率，一般取值为 0.96。

除了这些固定衰减率的调整学习率方法外，还有些自适应地调整学习率的方法，比如 AdaGrad、RMSprop、AdaDelta 等。这些方法都对每个参数设置不同的学习率。

AdaGrad 在标准的梯度下降方法中，每个参数在每次迭代时都使用相同的学习率。由于每个参数的维度上收敛速度都不相同，因此根据不同参数的收敛情况分别设置学习率。

AdaGrad（Adaptive Gradient）算法 [Duchi et al., 2011] 是借鉴 L2 正则化的思想，每次迭代时自适应地调整每个参数的学习率。在第 t 迭代时，先计算每个参数梯度平方的累计值

$$G_t = \sum_{\tau=1}^{t} \mathbf{g}_\tau \odot \mathbf{g}_\tau, \tag{7.24}$$

其中 $\odot$ 为按元素乘积，$\mathbf{g}_\tau \in \mathbb{R}^{|\theta|}$ 是第 τ 次迭代时的梯度。

AdaGrad 算法的参数更新差值为

$$\Delta\theta_t = -\frac{\alpha}{\sqrt{G_t + \epsilon}} \odot \mathbf{g}_t, \tag{7.25}$$

其中 α 是初始的学习率，ϵ 是为了保持数值稳定性而设置的非常小的常数，一般取值 e^{-7} 到 e^{-10}。此外，这里的开平方、除、加运算都是按元素进行的操作。

在 Adagrad 算法中，如果某个参数的偏导数累积比较大，其学习率相对较小；相反，如果其偏导数累积较小，其学习率相对较大。但整体是随着迭代次数的增加，学习率逐渐缩小。

Adagrad 算法的缺点是在经过一定次数的迭代依然没有找到最优点时，由于这时的学习率已经非常小，很难再继续找到最优点。

RMSprop *RMSprop*算法是 Geoff Hinton 提出的一种自适应学习率的方法，可以在有些情况下避免 AdaGrad 算法中学习率不断单调下降以至于过早衰减的缺点。

RMSprop 算法首先计算每次迭代梯度 $\mathbf{g}_t$ 平方的指数衰减移动平均，

$$G_t = \beta G_{t-1} + (1-\beta)\mathbf{g}_t \odot \mathbf{g}_t, \tag{7.26}$$

其中 β 为衰减率，一般取值为 0.9。

RMSprop 算法的参数更新差值为

$$\Delta\theta_t = -\frac{\alpha}{\sqrt{G_t + \epsilon}} \odot \mathbf{g}_t, \tag{7.27}$$

其中 α 是初始的学习率，比如 0.001。

从上式可以看出，RMSProp 算法和 Adagrad 算法的区别在于 G_t 的计算有累积方式变成了指数衰减移动平均。在迭代过程中，每个参数的学习率并不是呈衰减趋势，既可以变小也可以变大。

AdaDelta *AdaDelta*算法 [Zeiler, 2012] 也是 Adagrad 算法的一个改进。和 RMSprop 算法类似，AdaDelta 算法通过梯度平方的指数衰减移动平均来调整学习率。此外，AdaDelta 算法还引入了每个平方的指数衰减权移动平均。

第 t 次迭代时，每次参数更新差 $\Delta\theta_\tau, 1 \leq \tau \leq t-1$ 的指数衰减权移动平均为

此时 $\Delta\theta_t$ 未未知，因此只能计算到 ΔX_{t-1}。

$$\Delta X_{t-1}^2 = \beta \Delta X_{t-2}^2 + (1-\beta)\Delta\theta_{t-1} \odot \Delta\theta_{t-1}. \tag{7.28}$$

其中 β 为衰减率。

AdaDelta 算法的参数更新差值为

$$\Delta\theta_t = -\frac{\sqrt{\Delta X_{t-1}^2 + \epsilon}}{\sqrt{G_t + \epsilon}}\mathbf{g}_t \tag{7.29}$$

其中 G_t 的计算方式和 RMSprop 算法一样（公式 (7.26)），ΔX_{t-1}^2 为参数更新差 $\Delta\theta$ 的指数衰减权移动平均。

从上式可以看出，AdaDelta 算法将 RMSprop 算法中的初始学习率 α 改为动态计算的 $\sqrt{\Delta X_{t-1}^2}$，在一定程度上平抑了学习率的波动。

动量法

除了调整学习率之外，还可以通过使用最近一段时间内的平均梯度来代替当前时刻的梯度来作为参数更新的方向。这就是*动量法*。*动量*是模拟物理中的概念。一般而言，一个物体的动量指的是这个物体在它运动方向上保持运动的趋势，是物体的质量和速度的乘积。

动量法（Momentum Method）[Rumelhart et al., 1988] 是用之前积累动量来替代真正的梯度。每次迭代的梯度可以看作是加速度。

在第 t 次迭代时，计算*负梯度*的“加权移动平均”作为参数的更新方向，

实际上是相当于对 $-\frac{\alpha}{1-\rho}\mathbf{g}_t$ 做指数衰减移动平均。

$$\Delta\theta_t = \rho\Delta\theta_{t-1} - \alpha\mathbf{g}_t, \tag{7.30}$$

其中 ρ 为动量因子，通常设为0.9；α 为学习率。

这样每个参数的实际更新差值取决于最近一段时间内梯度的加权平均值。当某个参数在最近一段时间内的梯度方向不一致时，其真实的参数更新幅度变小；相反，当在最近一段时间内的梯度方向都一致时，其真实的参数更新幅度变大，起到加速作用。一般而言，在迭代初期，梯度方法都比较一致，动量法会起到加速作用，可以更快地到达最优点。在迭代后期，梯度方法会取决不一致，在收敛值附近震荡，动量法会起到减速作用，增加稳定性。从某种角度来说，当前梯度叠加上部分的上次梯度，一定程度上可以近似看作二阶梯度。

AdaM *Adam*（Adaptive Moment Estimation）算法 [Kingma and Ba, 2015] 可以看作是动量法和RMSprop的结合，不但使用动量作为参数更新方向，而且可以自适应调整学习率。

Adam算法一方面计算梯度平方 $\mathbf{g}_t^2$ 的指数加权平均（和RMSprop类似），另一方面计算梯度 $\mathbf{g}_t$ 的指数加权平均（和动量法类似）。

$$M_t = \beta_1 M_{t-1} + (1-\beta_1)\mathbf{g}_t, \tag{7.31}$$

$$G_t = \beta_1 G_{t-1} + (1-\beta_1)\mathbf{g}_t \odot \mathbf{g}_t, \tag{7.32}$$

其中 β_1 和 β_2 分别为两个移动平均的衰减率，通常取值为 $\beta_1 = 0.9, \beta_2 = 0.99$。

M_t 可以看作是梯度的均值（一阶矩），G_t 可以看作是梯度的未减去均值的方差（二阶矩）。

假设 $M_0 = 0, G_0 = 0$，那么在迭代初期 M_t 和 G_t 的值会比真实的均值和方差要小。特别是当 β_1 和 β_2 都接近于1时，偏差会很大。因此，需要对偏差进行修正。

$$\hat{M}_t = \frac{M_t}{1-\beta_1^t}, \tag{7.33}$$

$$\hat{G}_t = \frac{G_t]_{t-1}}{1-\beta_2^t}. \tag{7.34}$$

Adam算法的参数更新差值为

$$\Delta\theta_t = -\frac{\alpha}{\sqrt{G_t}+\epsilon}\hat{M}_t, \tag{7.35}$$

其中学习率 α 通常设为0.001，并且也可以进行衰减，比如 $\alpha_t = \alpha_0/\sqrt{t}$。

梯度截断

在深层神经网络或循环神经网络中，梯度消失或爆炸是影响学习效率的主要因素。除了上面介绍优化算法之外，还可以使用一种比较简单的启发式方法：*梯度截断*（gradient clipping）[Pascanu et al., 2013]。

梯度截断是限定一个梯度的模限定一个区间，当梯度的模小于或大于这个区间时就进行截断。一般截断的方式有以下几种：

按值截断 在第 t 次迭代时，梯度为 $\mathbf{g}_t$，给定一个区间 $[a, b]$，如果一个参数的梯度小于 a 时，就将其设为 a；如果小于 b 时，就将其设为 b。

$$\mathbf{g}_t = \max(\min(\mathbf{g}_t, b), a). \tag{7.36}$$

按模截断 *按模截断*是将梯度的模截断到一个给定的截断阈值 b。

如果 $\|\mathbf{g}_t\|^2 \leq b$，保持 $\mathbf{g}_t$ 不变。如果 $\|\mathbf{g}_t\|^2 > b$，令

$$\mathbf{g}_t = \frac{b}{\|\mathbf{g}_t\|}\mathbf{g}_t. \tag{7.37}$$

截断阈值 b 是一个超参数，也可以根据一段时间内的平均梯度来自动调整。Pascanu et al. [2013] 在实验中发现训练过程对阈值 a 并不十分敏感，通常一个小的阈值就可以得到很好的结果。

在训练循环神经网络时，按模截断是避免梯度爆炸问题的有效方法。图7.6给出了一个循环神经网络的损失函数关于参数的曲面。在使用梯度下降方法来进行参数学习的过程中，有时梯度会突然增大，如果用大的梯度进行更新参数，反而会导致其远离最优点。为了避免这种情况，当梯度的模大于一定阈值时，对梯度进行截断。

7.1.5 超参数优化

除了可学习的参数之外，神经网络模型中还存在很多超参数。这些超参数对模型的性能也十分关键。常见的超参数有

- 网络层数
- 每层的神经元数量

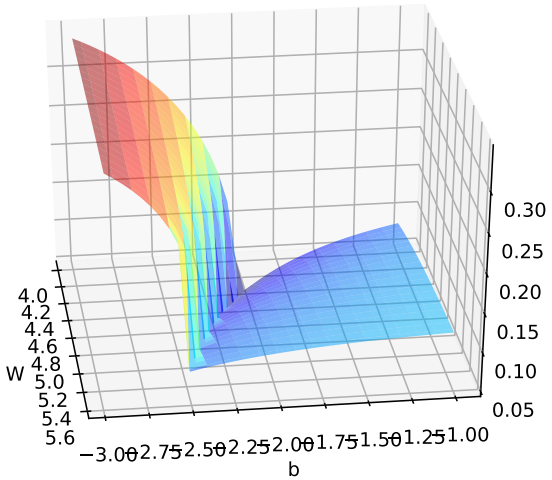


图 7.6: 梯度爆炸问题示例。图中的曲面为只有一个隐藏神经元的循环神经网络 $h_t = \sigma(wh_{t1} + b)$ 的损失函数，其中 w 和 b 为参数。假如 h_0 初始值为 0.3，损失函数为 $\mathcal{L} = (h_{100} - 0.65)^2$。

- 激活函数的类型
- 卷积核的大小以及数量
- 梯度下降方法中的学习率，以及选择哪种优化方法
- 正则化系数
- 小批量梯度下降中的每次选取样本的数量

超参数的优化问题是一个组合优化问题，没有通用的优化方法。对于超参数的设置，一般用**网格搜索**（Grid Search）或者人工搜索的方法来进行。假设总共有 K 个超参数，第 k 个超参数的可以取 m_k 个值。如果参数是连续的，可以将参数离散化，选择几个“经验”值。比如学习率 α，我们可以设置

$$\alpha \in \{0.01, 0.1, 0.5, 1.0\}.$$

这样，这些超参数可以有 $m_1 \times m_2 \times \cdots \times m_K$ 个取值组合。所谓网格搜索就是根据这些超参数的不同组合分别训练一个模型，然后评价这些模型在检验数据集上的性能，选取一组性能最好的组合。

如果每个超参数对模型性能的影响有很大差异，有些超参数对模型性能的影响有限，而有些则非常大。在这种情况下，网格搜索可能会有些浪费。采用随机搜索会比网格搜索更加有效，而且在实践中也更容易实现 [Bergstra and Bengio, 2012]。

7.2 网络正则化

机器学习模型的关键是泛化问题，即在样本真实分布上的期望风险最小化。对于同样，最小化神经网络模型在训练数据集上的经验风险并不是唯一目标。由于神经网络的拟合能力非常强，其在训练数据上的错误往往都可以降到非常低（比如错误率为0），因此如果提高神经网络的泛化能力反而成为影响模型能力的最关键因素。

在传统的机器学习中，提高泛化能力的方法主要是限制模型复杂度，比如采用权重衰减等方式。而在训练深度神经网络时，特别是在过度参数（over-parameterized）时，权重衰减的效果往往不如浅层机器学习模型中显著。

过度参数是指模型参数的数量远远大于训练数据的数量。

因此训练深度学习模型时，往往还会使用其它的正则化方法，比如数据增强、早期停止、丢弃法、集成法等。

7.2.1 权重递减

深度神经网络很容易产生过拟合现象，因为增加的抽象层使得模型能够对训练数据中较为罕见的依赖关系进行建模。对此，权重递减（ℓ_2 正规化）或者稀疏（ℓ_1 -正规化）等方法可以利用在训练过程中以减小过拟合现象 [Bengio et al., 2013]。

7.2.2 数据增强

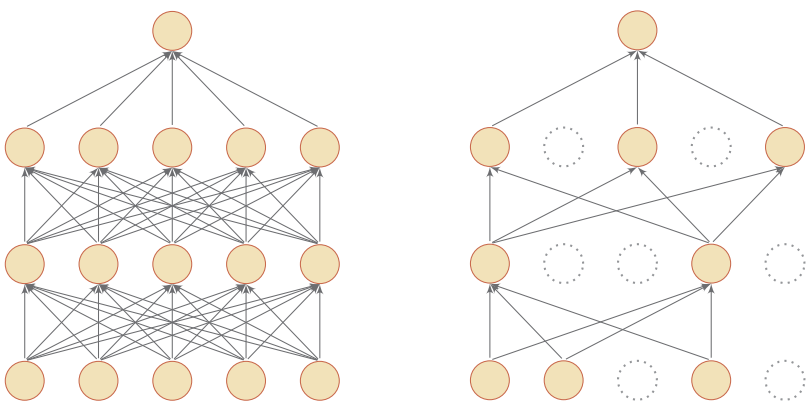
深层神经网络一般都需要大量的训练数据才能获得比较理想的结果。在数据量有限的情况下，可以通过数据增强（Data Augmentation）来增加数据量，

提高模型鲁棒性，避免过拟合。目前，数据增强还主要应用在图像数据上，在文本等其它类型的数据还没有太好的方法。

图像数据的增强主要是通过算法对图像进行转变，引入噪声等方法来增加数据的多样性。增强的方法主要有几种：

- 旋转（Rotation）：将图像按顺时针或逆时针方向随机旋转一定角度；
- 翻转（Flip）：将图像沿水平或垂直方法随机翻转一定角度；
- 缩放（Zoom In/Out）：将图像放大或缩小一定比例；
- 平移（Shift）：将图像沿水平或垂直方法平移一定步长；
- 加噪声（Noise）：加入随机噪声。

7.2.3 Dropout



(a) 标准网络 (b) Dropout 神经网络

图 7.7: Dropout 神经网络模型

训练一个大规模的神经网络经常容易过拟合。过拟合在很多机器学习中都会出现。一般解决过拟合的方法有正则化、早期终止、集成学习以及使用验证集等。Srivastava et al. [2014] 提出了适用于神经网络的避免过拟合的方法，叫**dropout**方法（丢弃法），即在训练中随机丢弃一部分神经元（同时丢弃其对应的连接边）来避免过拟合。图7.7给出了 dropout 网络的示例。

每做一次dropout，相当于从原始的网络中采样得到一个更瘦的网络。如果一个神经网络有n个神经元，那么总共可以采样出2^n个子网络。每次迭代都相当于训练一个不同的子网络，这些子网络都共享原始网络的参数。那么，最终的网络可以近似看作是集成了指数级个不同网络的组合模型。每次选择丢弃的神经元是随机的。最简单的方法是设置一个固定的概率p。对每一个神经元都一个概率p来判定要不要保留。p可以通过验证集来选取一个最优的值。或者，p也可以设为0.5，这对大部分的网络和任务有比较有效。

一般情况下dropout是针对神经元进行随机丢弃，但是也可以扩展到对每条边进行随机丢弃，或每一层进行随机丢弃。

当$p = 0.5$时，在训练时有一半的神经元被丢弃，只剩余一半的神经元是可以激活的。而在测试时，所有的神经元都是可以激活的。因此每个神经元训练时的净输入值平均比测试时小一半左右。这会造成训练和测试时网络的输出不一致。为了缓解这个问题，在测试时需要将每一个神经元的输出都折半，也相当于把不同的神经网络做了平均。

一般来讲，对于隐藏层的神经元，其dropout率等于0.5时效果最好，因为此时通过dropout方法，随机生成的网络结构最具多样性。对于输入层的神经元，其dropout率通常设为更接近1的数，使得输入变化不会太大。当对输入层神经元进行dropout时，相当于给数据增加噪声，以此来提高网络的鲁棒性。

7.3 模型压缩

待写。。。

参考文献

Yoshua Bengio, Nicolas Boulanger-Lewandowski, and Razvan Pascanu. Advances in optimizing recurrent networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8624–8628. IEEE, 2013.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle

point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pages 2933–2941, 2014.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International conference on artificial intelligence and statistics*, pages 249–256, 2010.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of International Conference on Learning Representations*, 2015.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the International Conference on Machine Learning*, pages 1310–1318, 2013.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5:3, 1988.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Matthew D Zeiler. Adadelta: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.