Final Project

162050128 杨铭 162050127 颜劭铭

2022 年 5 月 29 日

1 背景

在深度神经网络的训练中，如果数据分布在某一层网络上发生了一定的偏移，这会改变下一层的输入的分布，并且由于网络的层数通常非常多，随着网络训练层数的加深，数据分布的偏移会更加严重，进一步地导致模型优化的难度增加，甚至不能优化，这被称为内部协变量偏移（Internal Covariate Shift）。我们可以用一个形象的例子来理解，比如说高层大厦底部发生了微小偏移，楼层越高，偏移会越来越严重，并且后期会越来越难重铸回原来直上直下的样子。

内部协变量偏移会导致两个问题：在训练过程中，当前一层网络还没有收敛时，后面的网络的参数会随着前面网络的训练有很大的变化；激活函数在处理数据的时候，如果输入的数据特征比较大（例如取到 1000）或者比较小（例如取到 10^{-3} 的数量级甚至更小）的时候，这个数据在经过激活后（例如 sigmoid 函数），他会是比较接近于 1 或 0 的，这个时候梯度非常小，运用批梯度下降方法进行优化的时候非常缓慢。

而为了解决内部协变量偏移的问题，我们要使神经网络中每一层的参数保持独立同分布，比如说归一化到标准正态分布。因为在保证独立同分布的情况下，我们采样的数据会来自于同一空间，并且会散布在整个样本空间中，对于整个空间来说具有一定的代表性。

因此，谷歌研究人员在 2015 年的时候提出了 Batch Normalization [1]。该方法使深度神经网络训练过程中前期的准确率要高，收敛更快；并且减少对 learning rate 的敏感度。

2 相关工作

2015 年提出的 Batch Normalization，又被称为批量归一化。我们在很多任务中，经常会将输入进行归一化，比如说模型中一个特征是 $0 \sim 1$，另一个特征是 $0 \sim 1000$，可想而知，第二个特征对于模型参数会造成很大的影响，而通过将第二个特征归一化到 $0 \sim 1$ 上，可以尽可能地避免过大的特征对于模型后续训练的影响。而这里的批量指的是批量数据，也就是每一次优化时的样本数目。因此，谷歌的研究人员希望在深度神经网络中的每个卷积层后都进行一次批量归一化，这样就可以保证数据在每一层都是同分布的，尽可能地减少 ICS。

Batch Normalization 首先对于某一层中输入的所有样本 $\mathcal{B} = x_1, x_2, \cdots, x_m$ 求取均值和方差，并对于每个元素进行归一化成标准正态分布，而为了保证这层网络中学习到的特征不被破坏，需要对于归一化后的数据进行尺度缩放和偏移操作，实现恒等变换，恢复本层网络所学习的特征。公式如下：

$$\mu_{\mathcal{B}} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\sigma_{\mathcal{B}}^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2$$

$$\hat{x_i} = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$$

$$y_i = \gamma \hat{x_i} + \beta$$

其中 γ 和 β 是两个可学习参数。

我们可以从实验中观察 Batch Normalization 的效果，如图 1所示。作者在有和没有 BatchNorm 的 CIFAR-10 数据集 [2] 上训练标准 VGG 架构。正如预期的那样，图 1(a) 和 (b) 展示了使用 BatchNorm 层训练的网络在优化和泛化性能方面有着非常显著的改进。

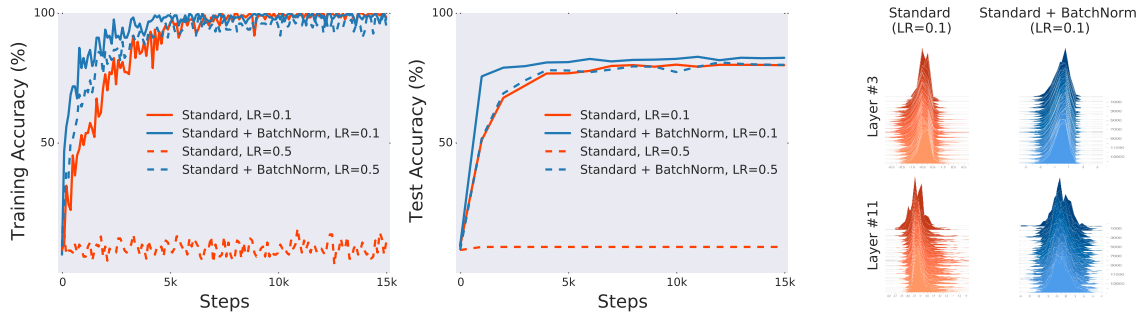


图 1: 比较使用和不使用 BatchNorm 在 $CIFAR-10$ 上训练的标准 VGG 网络的 (a) 训练（优化）和 (b) 测试（泛化）性能，以及输入在网络各层上的分布可视化 (c)

如今已经出现了许多归一化方法可以替代 BatchNorm，包括 Layer Normalization [3]，Group Normalization [4]，对于单张图片的 Instance Normalization [5]，Weight Normalization [6] 等等。这些方法提出了对于标准 BatchNorm 的改进，但是并没有尝试解释 BatchNorm 为什么成功。在 2015 年提出 BatchNorm 的论文 [1] 中，研究人员认为该方法的成功与网络中各层标准化后使得分布相同有关，而本文 How Does Batch Normalization Help Optimization? [7] 对于该解释提出了质疑，并证明了该方法的一个更根本的影响：它使优化环境变得更加平滑。

3 论文中作者解决的问题

在前文中，我们了解到谷歌的研究人员认为 BatchNorm 的成功与网络中各层标准化后使得分布相同有关，因此作者从该思路出发去验证正确性，绘制了输入的数据在网络各层中的可视化效果图 1.(c)，却发现在有和没有 BatchNorm 的网络的各层输入分布之前的差异是不太明显的，虽然有 BatchNorm 的网络的输入层分布会更加贴近于标准正态分布，但是这个差异似乎不能够导致图 1.(a) 和 (b) 中神经网络效果提高如此之多。

3.1 BatchNorm 的性能是否源于控制 ICS

从前文我们可以了解到减少 ICS 即尽可能地保持输入的数据独立同分布，因此，要研究 BatchNorm 的性能是否源于控制 ICS，作者通过三个对比实验说明：不使用 BatchNorm 的 VGG 网络，加入 BatchNorm 的 VGG 网络和加入带有噪声的 BatchNorm 的 VGG 网络，实验效果如图 2所示。

由于 BatchNorm 是让输入数据归一化到标准正态分布，因此作者在使用 BatchNorm 的各层网络上引入了非标准正态分布的噪声，使得各层网络发生严重的 ICS。

在图 2中我们可以发现，添加了噪声的 BatchNorm 的性能比没有添加噪声有着微小的差距，但是相比起标准的 VGG 网络的性能依旧有着非常大的提升，而根据图 2中的网络数据分布可视化图我们可以发现加入噪声的 BatchNorm 已经是不服从标准正态分布的，相比起标准的 VGG 网络和 BatchNorm 产生了非常严重的协变量偏移。因此，控制协变量偏移使得数据分布相同并不是 BatchNorm 性能提升的主要原因。

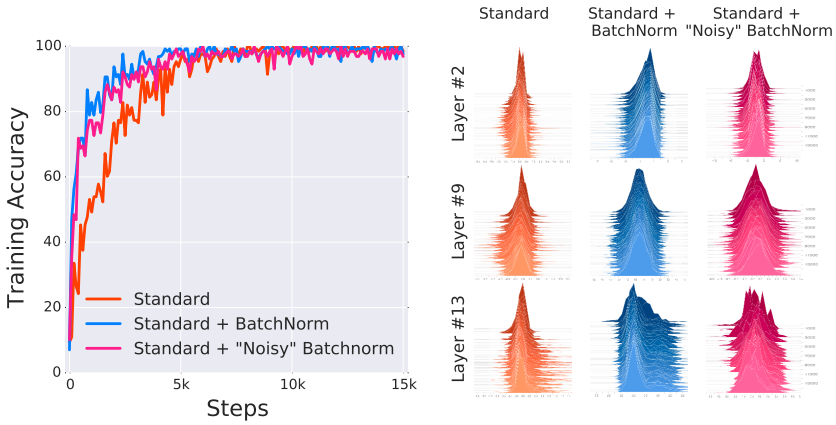


图 2: ICS 与 BatchNorm 性能的联系

3.2 BatchNorm 是否减少了 ICS

在 3.1 中我们可以发现如果将 ICS 与输入数据分布的稳定性联系在一起，BatchNorm 的性能与 ICS 没有直接的联系。但是在提出 BatchNorm 的论文中 [2] 给出了一个对于 ICS 的更广泛的定义：网络中参数的变化引起输入分布的变化。

即将每一层网络都视为一个优化任务：如何使得经验风险最小化，而每一层的输入分布对于损失函数进行优化，此时前面所有层的参数更新都会改变输入分布，使得这个经验风险最小化问题发生改变。因此作者将目光聚焦到了参数变化的问题上，因为在优化过程中，前一层的参数更新都会改变后续的网络层输入。

由于深度神经网络的训练过程中常用的是一阶优化方法，所以作者主要关注损失函数的梯度，并测量了每一层网络在前面所有层参数更新前后的梯度差异，通过这个差异衡量了前面参数的更新而引起当前网络层参数需要调整的程度。

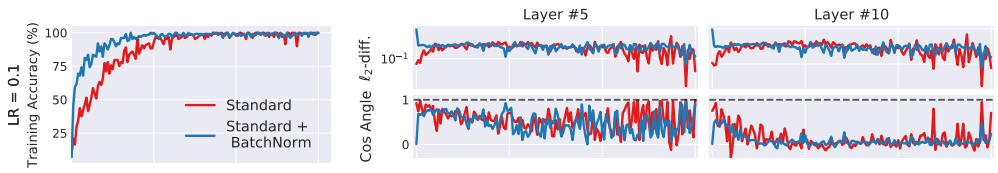
作者定义了以下三个公式进行计算：

$$G_{t,i} = \nabla_{W_i^{(t)}} \mathcal{L}\left(W_1^{(t)}, \ldots, W_k^{(t)}; x^{(t)}, y^{(t)}\right)$$

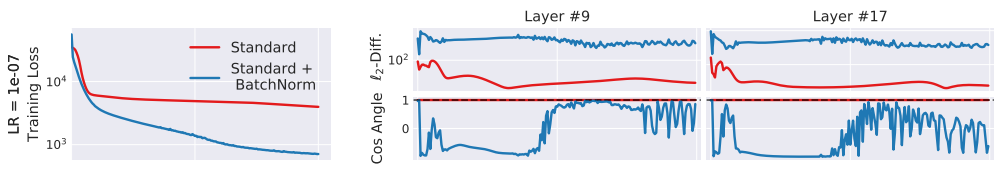
$$G'_{t,i} = \nabla_{W_i^{(t)}} \mathcal{L}\left(W_1^{(t+1)}, \ldots, W_{i-1}^{(t+1)}, W_i^{(t)}, W_{i+1}^{(t)}, \ldots, W_k^{(t)}; x^{(t)}, y^{(t)}\right)$$

梯度差异：$\|G_{t,i} - G'_{t,i}\|_2$

其中 $G_{t,i}$ 对应着当前层参数在其他层参数更新前的梯度，$G'_{t,i}$ 对应着当前层参数在其他层参数更新后的梯度。



(a) VGG



(b) DLN

图 3: 测量应用 BatchNorm 前后的 ICS（3.2 中的定义），训练准确度和损失

根据传统的 ICS 定义，BatchNorm 应该会减少 $G_{t,i}$ 和 $G'_{t,i}$ 之间的差异，增加两者间的相关性，从而减少 ICS。然而在图 3的实验中，作者对于同一层测量了更新参数前后的梯度差异（理想情况为 0）和余弦相似度（理想情况为 1），却惊讶地发现，在应用了 BatchNorm 以后，梯度差异反而更加明显。从左右两张图的对比我们可以发现虽然 BatchNorm 在训练准确度和损失方面表现很好，但是从优化角度考虑，BatchNorm 甚至没有减少 ICS。

4 核心思想与创新点

前面，我们在实验现象中可以看到，在引入噪声后，尽管 ICS 发生了偏移，但性能却几乎不受任何影响，这说明 Batch Norm 起到的作用并不是降低 ICS 来稳定分布. 换句话说，Batch Norm 提高模型性能的原因与人们普遍的观点背道而驰.

实际上，Batch Norm 不仅能降低 ICS，他还有别的好处 [1]. 例如可以防止梯度爆炸、对于初始化超参数的选择更具鲁棒性，远离平滑区域. 这些好处确实会有助于提高模型泛化能力. 但这都不是根本原因. 作者在实验佐证了这一反常现象后，又利用通过某些实验机制初步验证了自己的想法.

4.1 BatchNorm 对光滑性的影响

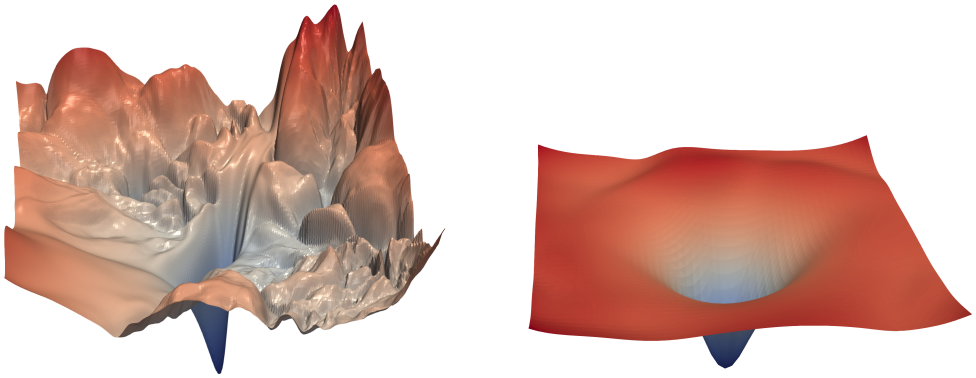
作者在 VGG 网络的优化环境下进行了实验 从图 4.1中明显可以看出，同一迭代步数中，带有 Batch-

(a) loss landscape (b) gradient predictiveness (c) "effective" β-smoothness

图 4: 在 VGG 网络中的性能对比

Norm 层的网络整体趋势不变，但在迭代过程中少了许多剧烈、尖锐的变化波动. 换句话说，BatchNorm 使损失函数变得更加平滑. 对比一般的网络，损失函数不仅有可能是非凸的，还有可能有很多非常平坦的区域、尖锐的顶点，而这些点的存在都会导致梯度消失或梯度爆炸. 这种不平滑会导致损失函数对初始点的选择、学习率的大小十分敏感.BatchNorm 带来的光滑性使优化过程收敛更迅速，更精确. 在下降过程中，梯度的变化十分缓和，不会出现突然变大的情况，因而当我们在取一个步长下降时，我们可以很自信地说，我们正在朝着最优点前进. 总得来说，BatchNorm 的使用正是在重新参数化过程中使得模型在梯度下降过程中更加可靠且更具预测性. 梯度下降的过程不会突变，就像一个小球沿着缓坡慢慢的溜到了地面.

我们可以用图 5来表示 BatchNorm 的结果 [8]。应用 BatchNorm 后，损失函数从左图的原始情况逐渐趋向于右图，使得优化更易于达到最优点。

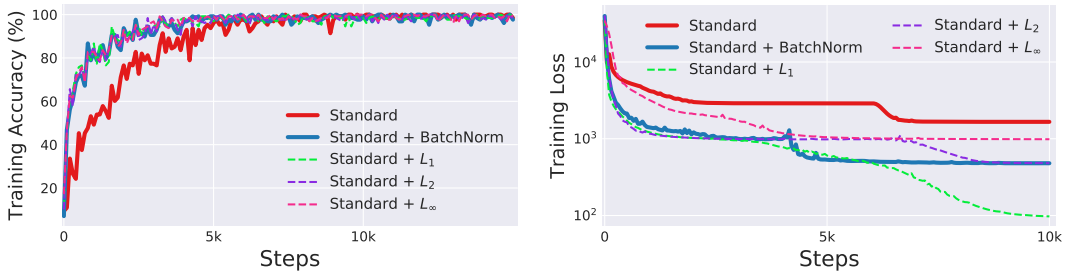


(a) without skip connections (b) with skip connections

图 5: 损失函数的变化

4.2 最优环境的发现

为了说明 Batch Norm 在连续上的稳定性，作者计算了每一步下降的梯度以及损失函数的变化量 从



(a) VGG (b) Deep Linear Model

图 4.2中可以看出，BatchNorm 有效的提高了性能表现，在不同的范数定义中，带有 BatchNorm 的表现更加平稳，没有明显的波动，但却沿着指定目标继续前进. 同时，在某些范数中其性能甚至比 BathcNorm 还要出色，这也有力佐证了 BatchNorm 带来的性能提升与分布的稳定性并没有太大关联. 最后做了多组对比实验，发现按照其他的梯度下降方向，损失函数的变化情况都是类似的.

4.3 不同环境下都可以平滑吗？

为了证明这一观点，作者做了多组不同范数条件下的 BatchNorm 实验，发现在 l_1 范数下性能甚至表现更好，而在 l_∞ 下发生了严重的偏移，这说明 BatchNorm 对光滑性的影响或许要在满足某种准则时才可以起到作用，甚至能比 BathcNorm 的性能更加出众.

5 公式推导部分

5.1 一些有用的结论

5.1.1 C.1

Fact C.1 BN 层的梯度. 在激活函数 $f(C)$ 下 $C = \gamma \cdot B + \beta$, 其中 B 经过 BN 层的标准化 $B = \mathrm{Bn}_{0,1}(A) := \frac{A-\mu}{\sigma}$, 其中 A 是在 m 为批量个数，方差为 σ^2, 我们有如下结论

$$\frac{\partial f}{\partial A^{(b)}} = \frac{\gamma}{m\sigma}\left(m\frac{\partial f}{\partial C^{(b)}} - \sum_{k=1}^{m}\frac{\partial f}{\partial C^{(k)}} - B^{(b)}\sum_{k=1}^{m}\frac{\partial f}{\partial C^{(k)}}B^{(k)}\right)$$

5.1.2 C.2

Fact C.2 正则化输出的梯度有如下公式:

$$\frac{\partial \hat{y}^{(b)}}{\partial y^{(k)}} = \frac{1}{\sigma}\left(\mathbf{1}[b=k] - \frac{1}{m} - \frac{1}{m}\hat{y}^{(b)}\hat{y}^{(k)}\right) \tag{1}$$

也就是:

$$\frac{\partial z_j^{(b)}}{\partial y^{(k)}} = \frac{\gamma}{\sigma}\left(\mathbf{1}[b=k] - \frac{1}{m} - \frac{1}{m}\hat{y}^{(b)}\hat{y}^{(k)}\right) \tag{2}$$

5.2 理论推导

5.2.1 理论 Theprem 4.1

Theorem 4.1(BN 在损失函数利普希茨连续上的影响) 考虑带有 BN 层的网络损失函数 $\hat{\mathcal{L}}$ 与相同的但不带 BN 层的网络损失函数 $\mathcal{L}$,

$$\left\|\nabla_{y_j}\hat{\mathcal{L}}\right\|^2 \le \frac{\gamma^2}{\sigma_j^2}\left(\left\|\nabla_{y_j}\mathcal{L}\right\|^2 - \frac{1}{m}\left\langle \mathbf{1}, \nabla_{y_j}\mathcal{L}\right\rangle^2 - \frac{1}{\sqrt{m}}\left\langle \nabla_{y_j}\mathcal{L}, \hat{y}_j\right\rangle^2\right).$$

证明: 损失函数 $\hat{\mathcal{L}}$ 对 y_j^b 求导有:

$$\frac{\partial \hat{\mathcal{L}}}{\partial y_j^b} = \left(\frac{\gamma}{m\sigma_j}\right)\left(m\frac{\partial \hat{\mathcal{L}}}{\partial z_j^b} - \sum_{k=1}^{m}\frac{\partial \hat{\mathcal{L}}}{\partial z_j^{(k)}} - \hat{y}_j^{(b)}\sum_{k=1}^{m}\frac{\partial \hat{\mathcal{L}}}{\partial z_j^{(k)}}\hat{y}_j^{(k)}\right), \tag{3}$$

我们可以将其写成向量形式:

$$\frac{\partial \hat{\mathcal{L}}}{y_j} = \left(\frac{\gamma}{\sigma_j}\right)\left(m\frac{\partial \hat{\mathcal{L}}}{\partial z_j} - \mathbf{1}\left\langle \mathbf{1}, \frac{\partial \hat{\mathcal{L}}}{\partial z_j}\right\rangle - \hat{y}_j\left\langle \frac{\partial \hat{\mathcal{L}}}{\partial z_j}, \hat{y}_j\right\rangle\right) \tag{4}$$

为了方便，记 $\mu_g = \frac{1}{m}\left\langle \mathbf{1}, \partial\hat{\mathcal{L}}/\partial z_j\right\rangle$, 其为梯度向量的均值，这方便我们使用 $\hat{y}_j$ 的范数和零均值性质，同时也便于我们使用上面的一些结论

$$\frac{\partial \hat{\mathcal{L}}}{\partial y_j} = \left(\frac{\gamma}{\sigma_j}\right)\left(\left(\frac{\partial \hat{\mathcal{L}}}{z_j} - \mathbf{1}\mu_g\right) - \frac{1}{m}\hat{y}_j\left\langle\left(\frac{\partial \hat{\mathcal{L}}}{z_j} - \mathbf{1}\mu_g\right), \hat{y}_j\right\rangle\right) \tag{5}$$

由于 $\|\hat{y}_j\| = \sqrt{m}$

$$= \frac{\gamma}{\sigma_j}\left(\left(\frac{\partial \hat{\mathcal{L}}}{z_j} - \mathbf{1}\mu_g\right) - \frac{\hat{y}_j}{\|\hat{y}_j\|}\left\langle\left(\frac{\partial \hat{\mathcal{L}}}{z_j} - \mathbf{1}\mu_g\right), \hat{y}_j\right\rangle\right) \tag{6}$$

$$\left\|\frac{\partial \hat{\mathcal{L}}}{\partial y_j}\right\|^2 = \frac{\gamma^2}{\sigma_j^2}\left\|\left(\frac{\partial \hat{\mathcal{L}}}{z_j} - \mathbf{1}\mu_g\right) - \frac{\hat{y}_j}{\|\hat{y}_j\|}\left\langle\left(\frac{\partial \hat{\mathcal{L}}}{z_j} - \mathbf{1}\mu_g\right), \hat{y}_j\right\rangle\right\|^2 \tag{7}$$

$$= \frac{\gamma^2}{\sigma_j^2}\left(\left\|\frac{\partial \hat{\mathcal{L}}}{z_j} - \mathbf{1}\mu_g\right\|^2 - \left\langle\left(\frac{\partial \hat{\mathcal{L}}}{z_j} - \mathbf{1}\mu_g\right), \hat{y}_j\right\rangle^2\right) \tag{8}$$

$$= \frac{\gamma^2}{\sigma^2}\left(\left\|\frac{\partial \hat{\mathcal{L}}}{\partial z_j}\right\|^2 - \frac{1}{m}\left\langle \mathbf{1}, \frac{\partial \hat{\mathcal{L}}}{\partial z_j}\right\rangle^2 - \frac{1}{\sqrt{m}}\left\langle \frac{\partial \hat{\mathcal{L}}}{\partial z_j}, \hat{y}_j\right\rangle^2\right) \tag{9}$$

将结论 1 代入该式即可得到 **Theorem 4.1**. 现在可以利用这个等式来推导 **Theorem 4.4**

5.2.2 Theorem 4.4

在权重空间最小化利普希茨连续的上界，其中 $\hat{\mathcal{L}}$ 是 BN 的 loss,$\mathcal{L}$ 是不带 BN 的 loss（其余网络的结构相同）.

$$g_j = \max_{\|X\|\le\lambda}\left\|\nabla_W\mathcal{L}\right\|^2, \quad \hat{g}_j = \max_{\|X\|\le\lambda}\left\|\nabla_W\widehat{\mathcal{L}}\right\|^2 \Longrightarrow \hat{g}_j \le \frac{\gamma^2}{\sigma_j^2}\left(g_j^2 - m\mu_{g_j}^2 - \lambda^2\left\langle\nabla_{y_j}\mathcal{L}, \hat{\boldsymbol{y}}_j\right\rangle^2\right)$$

为了证明此式，对于最大的特征向量 λ_0 有以下结论:

$$\lambda_0 = \max_{x \in \mathcal{R}^d; \|x\|_2 = 1} x^T M x, \tag{10}$$

这意味着对于一个矩阵 X 有着约束 $\|X\|_2 \leq \lambda$, 则 $\forall v \in \mathcal{R}^d, v^T X v \leq \lambda \|v\|^2$，这让矩阵 X 有着一个较紧密的上界，因为当且仅当 $\lambda = \lambda_0$ 时等号才成立

再通过链式法则求解权重的梯度:

$$\frac{\partial \widehat{\mathcal{L}}}{\partial W_{ij}} = \sum_{b=1}^{m} \frac{\partial \widehat{\mathcal{L}}}{\partial y_j{}^{(b)}} \frac{\partial y_j{}^{(b)}}{\partial W_{ij}} \tag{11}$$

$$\frac{\partial \widehat{\mathcal{L}}}{\partial W_{ij}} = \sum_{b=1}^{m} \frac{\partial \widehat{\mathcal{L}}}{\partial y_j{}^{(b)}} x_i{}^{(b)} \tag{12}$$

$$= \left\langle \frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j}, \boldsymbol{x_i} \right\rangle \tag{13}$$

$$\frac{\partial \widehat{\mathcal{L}}}{\partial W_{\cdot j}} = \boldsymbol{X}^\top \left(\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j} \right) \tag{14}$$

再有 $X \in \mathcal{R}^{m \times d}$ 是输入矩阵且 $X_{bi} = x_i^{(b)}$，由此可得:

$$\left\| \frac{\partial \widehat{\mathcal{L}}}{\partial W_{\cdot j}} \right\|^2 = \left(\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y_j}} \right)^\top \boldsymbol{X} \boldsymbol{X}^\top \left(\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y_j}} \right), \tag{15}$$

我们将前面的结论 $\|X\|_2 \leq \lambda$ 变形，$\left\|X^T X\right\|_2 \leq \lambda^2$, 故可以得到:

$$\max_{\|X\|_2 < \lambda} \left\| \frac{\partial \widehat{\mathcal{L}}}{\partial W_{\cdot j}} \right\|^2 \leq \lambda^2 \left(\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y_j}} \right)^\top \left(\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y_j}} \right) = \lambda^2 \left\| \frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y_j}} \right\|^2 \tag{16}$$

代入 **Theorem 4.1** 的结论有:

$$\hat{g_j} := \max_{\|X\|_2 < \lambda} \left\| \frac{\partial \widehat{\mathcal{L}}}{\partial W_{\cdot j}} \right\|^2 \leq \frac{\lambda^2 \gamma^2}{\sigma^2} \left(\left\| \frac{\partial \mathcal{L}}{\partial \boldsymbol{y}_j} \right\|^2 - \frac{1}{m} \left\langle 1, \frac{\partial \mathcal{L}}{\partial \boldsymbol{y}_j} \right\rangle^2 - \frac{1}{\sqrt{m}} \left\langle \frac{\partial \mathcal{L}}{\partial \boldsymbol{y}_j}, \hat{\boldsymbol{y}}_j \right\rangle^2 \right) \tag{17}$$

再应用一次公式 (10)

$$g_j := \max_{\|X\|_2 < \lambda} \left\| \frac{\partial \widehat{\mathcal{L}}}{\partial W_{\cdot j}} \right\|^2 = \lambda^2 \left\| \frac{\partial \mathcal{L}}{\partial \boldsymbol{y}_j} \right\|^2 \tag{18}$$

得到了最终的形式:

$$\hat{g}_j \leq \frac{\gamma^2}{\sigma^2} \left(g_j^2 - m \mu_{g_j}^2 - \lambda^2 \left\langle \frac{\partial \mathcal{L}}{\partial \boldsymbol{y}_j}, \hat{\boldsymbol{y}}_j \right\rangle^2 \right)$$

Theorem 4.2 (BN 对光滑性的影响). 令 $\hat{\boldsymbol{g}}_j = \nabla_{\boldsymbol{y}_j} \mathcal{L}$ ， $\boldsymbol{H}_{jj} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{y}_j \partial \boldsymbol{y}_j}$ 分别为 loss 输出层的梯度和 Hessian 矩阵，则

$$\left(\nabla_{y_j} \widehat{\mathcal{L}} \right)^\top \frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j \partial \boldsymbol{y}_j} \left(\nabla_{y_j} \widehat{\mathcal{L}} \right) \leq \frac{\gamma^2}{\sigma^2} \left(\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j} \right)^\top \boldsymbol{H}_{jj} \left(\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j} \right) - \frac{\gamma}{m \sigma^2} \left\langle \hat{\boldsymbol{g}}_j, \hat{\boldsymbol{y}}_j \right\rangle \left\| \frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j} \right\|^2$$

如果 $\boldsymbol{H}_{jj}$ 保留了 $\hat{\boldsymbol{g}}_j$ 和 $\nabla_{\boldsymbol{y}_j} \widehat{\mathcal{L}}$ 的范数性质，则有:

$$\left(\nabla_{y_j} \widehat{\mathcal{L}} \right)^\top \frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j \partial \boldsymbol{y}_j} \left(\nabla_{y_j} \widehat{\mathcal{L}} \right) \leq \frac{\gamma^2}{\sigma^2} \left(\hat{\boldsymbol{g}}_j^\top \boldsymbol{H}_{jj} \hat{\boldsymbol{g}}_j - \frac{1}{m \gamma} \left\langle \hat{\boldsymbol{g}}_j, \hat{\boldsymbol{y}}_j \right\rangle \left\| \frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j} \right\|^2 \right)$$

证明: 首先回顾一下 Hessian 矩阵的形式, 同时我们作出假设

$$\boldsymbol{H}_{jk} \in \mathbb{R}^{m\times m}; H_{jk} := \frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{z}_j \partial \boldsymbol{z_k}} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{y}_j \partial \boldsymbol{y_k}}$$

为了方便起见，定义一个函数 $\mu_{(\cdot)}$ 用来计算向量的平均值. 最后，我们得到了梯度的形式:

$$\hat{g}_j = \frac{\partial \widehat{\mathcal{L}}}{\partial z_j} = \frac{\partial \mathcal{L}}{\partial y_j}$$

同样，这里的第二个等式是基于假设. 现在，在 Hessian 矩阵对应的 loss 在 BN 层前的激活函数 y_j 使用扩展的梯度:

$$\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j \partial \boldsymbol{y}_j} = \frac{\partial}{\partial \boldsymbol{y}_j}\left(\left(\frac{\gamma}{m\sigma_j}\right)\left[m\hat{\boldsymbol{g}}_j - m\mu_{(\hat{\boldsymbol{g}}_j)} - \hat{y}_j^{(b)}\left\langle \hat{\boldsymbol{g}}_j, \hat{\boldsymbol{y}}_j\right\rangle\right]\right) \tag{19}$$

利用了乘积法则和链式法则:

$$= \frac{\gamma}{m\sigma}\left(\frac{\partial}{\partial z_q}\left[m\hat{\boldsymbol{g}}_j - m\mu_{(\hat{\boldsymbol{g}}_j)} - \hat{\boldsymbol{y}}_j\left\langle \hat{\boldsymbol{g}}_j, \hat{\boldsymbol{y}}_j\right\rangle\right]\right)\cdot\frac{\partial z_{\boldsymbol{q}}}{\partial \boldsymbol{y}_j} \tag{20}$$

$$+\left(\frac{\partial}{\partial \boldsymbol{y}_j}\left(\frac{\gamma}{m\sigma_j}\right)\right)\cdot\left(m\hat{\boldsymbol{g}}_j - m\mu_{(\hat{\boldsymbol{g}}_j)} - \hat{\boldsymbol{y}}_j\left\langle \hat{\boldsymbol{g}}_j, \hat{\boldsymbol{y}}_j\right\rangle\right) \tag{21}$$

利用减法来拆分等式:

$$= \left(\frac{\gamma}{\sigma_j}\right)\left(\boldsymbol{H}_{jj} - \frac{\partial \mu_{(\hat{g}_j)}}{\partial z_j} - \frac{\partial}{\partial z_j}\left(\frac{1}{m}\hat{\boldsymbol{y}}_j\left\langle \hat{\boldsymbol{g}}_j, \hat{\boldsymbol{y}}_j\right\rangle\right)\right)\cdot\frac{\partial \boldsymbol{z}_j}{\partial \boldsymbol{y}_j} \tag{22}$$

$$+\left(\hat{\boldsymbol{g}}_j - \boldsymbol{\mu}_{(\hat{\boldsymbol{g}}_j)} - \frac{1}{m}\hat{\boldsymbol{y}}_j\left\langle \hat{\boldsymbol{g}}_j, \hat{\boldsymbol{y}}_j\right\rangle\right)\left(\frac{\partial}{\partial \boldsymbol{y}_j}\left(\frac{\gamma}{\sigma_j}\right)\right) \tag{23}$$

我们对上面的式子一个一个代入此式子:

$$\frac{\partial \mu_{(\hat{g}_j)}}{\partial z_j} = \frac{1}{m}\frac{\partial \mathbf{1}^\top \hat{g}_j}{\partial z_j} = \frac{1}{m}1\cdot\mathbf{1}^\top \boldsymbol{H}_{jj} \tag{24}$$

$$\frac{\partial}{\partial z_j}\left(\hat{\boldsymbol{y}}_j\left\langle \hat{\boldsymbol{y}}_j, \hat{\boldsymbol{g}}_{\boldsymbol{j}}\right\rangle\right) = \frac{1}{\gamma}\frac{\partial}{\partial \hat{\boldsymbol{y}}_j}\left(\hat{\boldsymbol{y}}_j\left\langle \hat{\boldsymbol{y}}_j, \hat{\boldsymbol{g}}_j\right\rangle\right) \tag{25}$$

$$= \frac{1}{\gamma}\frac{\partial \hat{\boldsymbol{y}}_j}{\partial \hat{\boldsymbol{y}}_j}\left\langle \hat{\boldsymbol{g}}_{\boldsymbol{j}}, \hat{\boldsymbol{y}}_{\boldsymbol{j}}\right\rangle + \hat{\boldsymbol{y}}_j\hat{\boldsymbol{y}}_j^\top \boldsymbol{H}_{jj} + \frac{1}{\gamma}\hat{\boldsymbol{y}}_j\hat{\boldsymbol{g}}_j^\top\frac{\partial \hat{\boldsymbol{y}}_j}{\partial \hat{\boldsymbol{y}}_j} \tag{26}$$

$$= \frac{1}{\gamma}\boldsymbol{I}\left\langle \hat{\boldsymbol{g}}_{\boldsymbol{j}}, \hat{\boldsymbol{y}}_{\boldsymbol{j}}\right\rangle + \hat{\boldsymbol{y}}_j\hat{\boldsymbol{y}}_j^\top \boldsymbol{H}_{jj} + \frac{1}{\gamma}\hat{\boldsymbol{y}}_j\hat{\boldsymbol{g}}_j^\top\boldsymbol{I} \tag{27}$$

$$\frac{\partial}{\partial \boldsymbol{y}_j}\left(\frac{\gamma}{\sigma_j}\right) = \gamma\sqrt{m}\frac{\partial\left(\left(\boldsymbol{y}_{\boldsymbol{j}} - \boldsymbol{\mu}_{(\boldsymbol{y}_j)}\right)^\top\left(\boldsymbol{y}_{\boldsymbol{j}} - \boldsymbol{\mu}_{(\boldsymbol{y}_j)}\right)\right)^{-\frac{1}{2}}}{\partial \boldsymbol{y}_j} \tag{28}$$

$$= \frac{-1}{2}\gamma\sqrt{m}\left(\left(\boldsymbol{y}_j - \boldsymbol{\mu}_{(\boldsymbol{y}_j)}\right)^\top\left(\boldsymbol{y}_{\boldsymbol{j}} - \boldsymbol{\mu}_{(\boldsymbol{y}_j)}\right)\right)^{-\frac{3}{2}}\left(2\left(\boldsymbol{y}_j - \boldsymbol{\mu}_{(\boldsymbol{y}_j)}\right)\right) \tag{29}$$

$$= -\frac{\gamma}{m\sigma^2}\hat{\boldsymbol{y}}_j \tag{30}$$

用之前的来重写 Hessian 矩阵:

$$\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j \partial \boldsymbol{y}_j} = \left(\frac{\gamma}{m\sigma_j}\right)\left(m\boldsymbol{H}_{jj} - 1\cdot\mathbf{1}^\top \boldsymbol{H}_{jj} - \frac{1}{\gamma}\boldsymbol{I}\left\langle \hat{\boldsymbol{g}}_j, \hat{\boldsymbol{y}}_j\right\rangle - \hat{\boldsymbol{y}}_j\hat{\boldsymbol{y}}_j^\top \boldsymbol{H}_{jj} - \frac{1}{\gamma}\left(\hat{\boldsymbol{y}}_j\hat{\boldsymbol{g}}_j^\top\right)\right)\cdot\frac{\partial \boldsymbol{z}_j}{\partial \boldsymbol{y}_j} \tag{31}$$

$$-\frac{\gamma}{m\sigma^2}\left(\hat{\boldsymbol{g}}_j - \boldsymbol{\mu}_{(\hat{\boldsymbol{g}}_j)} - \frac{1}{m}\hat{\boldsymbol{y}}_j\left\langle \hat{\boldsymbol{g}}_j, \hat{\boldsymbol{y}}_j\right\rangle\right)\hat{\boldsymbol{y}}_j^\top \tag{32}$$

代入 **Fact C.2** 有:

$$\frac{\partial z_j}{\partial y_j} = \left(\frac{\gamma}{\sigma_j}\right)\left(I - \frac{1}{m}1\cdot 1^\top - \frac{1}{m}\hat{y}_j\hat{y}_j^\top\right), \tag{33}$$

代入上式有:

$$\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j \partial \boldsymbol{y}_j} = \frac{\gamma^2}{m\sigma^2}\left(m\boldsymbol{H}_{jj} - M\boldsymbol{H}_{jj} - \frac{1}{\gamma}\boldsymbol{I}\left\langle \hat{\boldsymbol{g}}_{\boldsymbol{j}}, \hat{\boldsymbol{y}}_{\boldsymbol{j}}\right\rangle - \hat{\boldsymbol{y}}_j\hat{\boldsymbol{y}}_j^\top\boldsymbol{H}_{jj} - \frac{1}{\gamma}\left(\hat{\boldsymbol{y}}_j\hat{\boldsymbol{g}}_j^\top\right)\right) \tag{34}$$

$$- \frac{\gamma^2}{m\sigma^2}\left(\boldsymbol{H}_{jj}\boldsymbol{M} - \frac{1}{m}\boldsymbol{M}\boldsymbol{H}_{jj}\boldsymbol{M} - \frac{1}{m\gamma}M\left\langle \hat{\boldsymbol{g}}_j, \hat{\boldsymbol{y}}_{\boldsymbol{j}}\right\rangle - \frac{1}{m}\hat{\boldsymbol{y}}_j\hat{\boldsymbol{y}}_j^\top\boldsymbol{H}_{jj}\boldsymbol{M} - \frac{1}{m\gamma}\left(\hat{\boldsymbol{y}}_j\hat{\boldsymbol{g}}_j^\top\boldsymbol{M}\right)\right) \tag{35}$$

$$- \frac{\gamma^2}{m\sigma^2}\left(\boldsymbol{H}_{jj}\hat{\boldsymbol{y}}_j\hat{\boldsymbol{y}}_j^\top - \frac{1}{m}\boldsymbol{M}\boldsymbol{H}_{jj}\hat{\boldsymbol{y}}_j\hat{\boldsymbol{y}}_j^\top - \frac{1}{m\gamma}\hat{\boldsymbol{y}}_j\hat{\boldsymbol{y}}_j^\top\left\langle \hat{\boldsymbol{y}}_j, \hat{\boldsymbol{y}}_{\boldsymbol{j}}\right\rangle - \frac{1}{m}\hat{\boldsymbol{y}}_j\hat{\boldsymbol{y}}_j^\top\boldsymbol{H}_{jj}\hat{\boldsymbol{y}}_j\hat{\boldsymbol{y}}_j^\top - \frac{1}{m\gamma}\left(\hat{\boldsymbol{y}}_j\hat{\boldsymbol{g}}_j^\top\hat{\boldsymbol{y}}_j\hat{\boldsymbol{y}}_j^\top\right)\right) \tag{36}$$

$$- \frac{\gamma}{m\sigma^2}\left(\hat{\boldsymbol{g}}_j - \boldsymbol{\mu}_{(\hat{\boldsymbol{g}}_j)} - \frac{1}{m}\hat{\boldsymbol{y}}_j\left\langle \hat{\boldsymbol{y}}_j, \hat{\boldsymbol{y}}_j\right\rangle\right)\hat{\hat{y}}_j^\top \tag{37}$$

整合这个式子，并令 $\overline{\hat{\boldsymbol{g}}_j} = \hat{\boldsymbol{g}}_j - \mu_{(\hat{\boldsymbol{g}}_j)}$：

$$\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j \partial \boldsymbol{y}_j} = \frac{\gamma^2}{m\sigma^2}\Big[m\boldsymbol{H}_{jj} - M\boldsymbol{H}_{jj} - \hat{\boldsymbol{y}}_j\hat{y}_j^\top\boldsymbol{H}_{jj} - \boldsymbol{H}_{jj}M + \frac{1}{m}\boldsymbol{M}\boldsymbol{H}_{jj}M \tag{38}$$

$$+\frac{1}{m}\hat{y}_j\hat{y}_j^\top\boldsymbol{H}_{jj}M - \boldsymbol{H}_{jj}\hat{\boldsymbol{y}}_j\hat{\boldsymbol{y}}_j^\top + \frac{1}{m}\boldsymbol{M}\boldsymbol{H}_{jj}\hat{\boldsymbol{y}}_j\hat{\boldsymbol{y}}_j^\top + \frac{1}{m}\hat{\boldsymbol{y}}_j\hat{y}_j^\top\boldsymbol{H}_{jj}\hat{\boldsymbol{y}}_j\hat{\boldsymbol{y}}_j^\top\Big] \tag{39}$$

$$- \frac{\gamma}{m\sigma^2}\left(\hat{g}_j\hat{y}_j^\top - \mu_{(\hat{g}_j)}\hat{y}_j^\top - \frac{3}{m}\hat{y}_j\hat{y}_j^\top\left\langle \hat{g}_j, \hat{y}_j\right\rangle + \left(\left\langle \hat{g}_j, \hat{y}_j\right\rangle I + \hat{y}_j\hat{g}_j^\top\right)\left(I - \frac{1}{m}M\right)\right) \tag{40}$$

$$= \frac{\gamma^2}{\sigma^2}\left[\left(I - \frac{1}{m}\hat{y}_j\hat{y}_j^\top - \frac{1}{m}M\right)\boldsymbol{H}_{jj}\left(I - \frac{1}{m}\hat{\boldsymbol{y}}_j\hat{\boldsymbol{y}}_j^\top - \frac{1}{m}M\right)\right. \tag{41}$$

$$\left. -\frac{1}{m\gamma}\left(\overline{\hat{g}_j}\hat{y}_j^\top + \hat{y}_j\overline{\hat{g}_j}^\top - \frac{3}{m}\hat{\boldsymbol{y}}_j\hat{\boldsymbol{g}}_j^\top\hat{y}\hat{\boldsymbol{y}}_j^\top + \left\langle \hat{\boldsymbol{g}}_j, \hat{\boldsymbol{y}}_j\right\rangle\left(I - \frac{1}{m}M\right)\right)\right] \tag{42}$$

我们希望计算 BN 光滑度利用每一层批量的激活函数，也就是 g^THg:

$$M\overline{\hat{g}_j} = 0 \tag{43}$$

$$\left(I - \frac{1}{m}M - \frac{1}{m}\hat{y}_j\hat{y}_j^\top\right)^2 = \left(I - \frac{1}{m}M - \frac{1}{m}\hat{y}_j\hat{y}_j^\top\right) \tag{44}$$

$$\hat{y}_j^\top\left(I - \frac{1}{m}\hat{y}_j\hat{y}_j^\top\right) = 0 \tag{45}$$

$$\left(I - \frac{1}{m}M\right)\left(I - \frac{1}{m}\hat{y}_j\hat{y}_j^\top\right)\overline{\hat{g}_j} = \left(I - \frac{1}{m}\hat{y}_j\hat{y}_j^\top\right)\overline{\hat{g}_j} \tag{46}$$

再回到前面的公式:

$$\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j} = \frac{\gamma}{\sigma}\overline{\hat{g}_j}^\top\left(I - \frac{1}{m}\hat{y}_j\hat{\boldsymbol{y}}_j^\top\right) \tag{47}$$

应用上式:

$$\frac{\partial \widehat{\mathcal{L}}^\top}{\partial \boldsymbol{y}_j}\cdot\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j\partial \boldsymbol{y}_j}\cdot\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j} = \frac{\gamma^4}{\sigma^4}\overline{\hat{g}}_j^\top\left(I - \frac{1}{m}\hat{\boldsymbol{y}}_j\hat{\boldsymbol{y}}_j^\top\right)\boldsymbol{H}_{jj}\left(I - \frac{1}{m}\hat{y}_j\hat{\boldsymbol{y}}_j^\top\right)\underline{\hat{\boldsymbol{g}}_j} \tag{48}$$

$$- \frac{\gamma^3}{m\sigma^4}\overline{\hat{g}}_j^\top\left(I - \frac{1}{m}\hat{\boldsymbol{y}}_j\hat{\boldsymbol{y}}_j^\top\right)\overline{\hat{\boldsymbol{g}}_j}\left\langle \hat{\boldsymbol{g}}_{\boldsymbol{j}}, \hat{\boldsymbol{y}}_j\right\rangle \tag{49}$$

$$= \frac{\gamma^2}{\sigma^2}\left(\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j}\right)^\top\boldsymbol{H}_{jj}\left(\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j}\right) - \frac{\gamma}{m\sigma^2}\left\langle \hat{\boldsymbol{g}}_j, \hat{\boldsymbol{y}}_j\right\rangle\left\|\frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j}\right\|\|^2 \tag{50}$$

这里包括了证明中的第一个部分，注意如果满足我们的假设，这第一个部分可以由 H_{jj} 范数诱导，亦即:

$$\frac{\partial \widehat{\mathcal{L}}^{\top}}{\partial \boldsymbol{y}_j} \cdot \frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j \partial \boldsymbol{y}_j} \cdot \frac{\partial \widehat{\mathcal{L}}}{\partial y_j} \leq \frac{\gamma^2}{\sigma^2} \left[\hat{g}_j^{\top} \boldsymbol{H}_{jj} \hat{g}_j - \frac{1}{m\gamma} \left\langle \hat{g}_j, \hat{\boldsymbol{y}}_j \right\rangle \left\| \frac{\partial \widehat{\mathcal{L}}}{\partial y_j} \right\|^2 \right] \tag{51}$$

这里的最小化上界提供了分离: 、textbfTheorem C.1 (最小化平滑上界). 对应着前面的式子

$$\max_{\|X\| \leq \lambda} \left(\frac{\partial \widehat{\mathcal{L}}}{\partial W_{\cdot j}} \right)^{\top} \frac{\partial \widehat{\mathcal{L}}}{\partial W_{\cdot j} \partial W_{\cdot j}} \left(\frac{\partial \widehat{\mathcal{L}}}{\partial W_{\cdot j}} \right) < \frac{\gamma^2}{\sigma^2} \left[\max_{\|X\| \leq \lambda} \left(\frac{\partial \mathcal{L}}{\partial W_{\cdot j}} \right)^{\top} \frac{\partial \mathcal{L}}{\partial W_{\cdot j} \partial W_{\cdot j}} \left(\frac{\partial \mathcal{L}}{\partial W_{\cdot j}} \right) - \lambda^4 \kappa \right],$$

κ 是之前中给出的分离系数

证明:

$$\frac{\partial \mathcal{L}}{\partial W_{ij} \partial W_{kj}} = \boldsymbol{x}_i^{\top} \frac{\partial \mathcal{L}}{\partial \boldsymbol{y}_j \partial \boldsymbol{y}_j} \boldsymbol{x_k} \tag{52}$$

$$\frac{\partial \widehat{\mathcal{L}}}{\partial W_{ij} \partial W_{kj}} = \boldsymbol{x}_i^{\top} \frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j \partial \boldsymbol{y}_j} \boldsymbol{x_k} \tag{53}$$

$$\frac{\partial \widehat{\mathcal{L}}}{\partial W_{\cdot j} \partial W_{\cdot j}} = \boldsymbol{X}^{\top} \frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j \partial \boldsymbol{y}_j} \boldsymbol{X} \tag{54}$$

我们用在第一部分证明中的梯度来观察一下梯度的预测性:

$$\beta := \left(\frac{\partial \widehat{\mathcal{L}}}{\partial W_{\cdot j}} \right)^{\top} \frac{\partial \widehat{\mathcal{L}}}{\partial W_{\cdot j} \partial W_{\cdot j}} \left(\frac{\partial \widehat{\mathcal{L}}}{\partial W_{\cdot j}} \right) \tag{55}$$

$$= \hat{g}_j^{\top} \left(I - \frac{1}{m} \hat{\boldsymbol{y}}_j \hat{\boldsymbol{y}}_j^{\top} \right) X X^{\top} \frac{\partial \widehat{\mathcal{L}}}{\partial \boldsymbol{y}_j \partial \boldsymbol{y}_j} \boldsymbol{X} \boldsymbol{X}^{\top} \left(I - \frac{1}{m} \hat{\boldsymbol{y}}_j \hat{\boldsymbol{y}}_j^{\top} \right) \hat{g}_j \tag{56}$$

最大化范数:

$$\max_{\|X\| \leq \lambda} \beta = \lambda^4 \hat{g}_j^{\top} \left(I - \frac{1}{m} \hat{y}_j \hat{y}_j^{\top} \right) \frac{\partial \widehat{\mathcal{L}}}{\partial y_j \partial y_j} \left(I - \frac{1}{m} \hat{y}_j \hat{y}_j^{\top} \right) \hat{g}_j \tag{57}$$

在这里过去的证明可以得出结论: Lemma 4.5 (BN 的偏好初始化). 令 $\boldsymbol{W}^*$ 和 $\widehat{\boldsymbol{W}}^*$ 分别是在正常网络和 BN 网络中的局部最优权重. 对于任意的初始化:

$$\left\| W_0 - \widehat{W}^* \right\|^2 \leq \|W_0 - W^*\|^2 - \frac{1}{\|W^*\|^2} \left(\|W^*\|^2 - \langle W^*, W_0 \rangle \right)^2,$$

if $\langle W_0, W^* \rangle > 0$, where $\widehat{W}^*$ and W^* are closest optima for BN and standard network, respectively.

证明: 这是 BN 缩放的结果. 我们有:

$$k = \frac{\langle W^*, W_0 \rangle}{\|W^*\|^2} \tag{58}$$

因此对于任何一个最优点 w^*, 一定有:$\hat{W} := kW^*$ 是一个 BN 网络的最优权重, 在这里局部最优和最优点由这个式子给出:

$$\left\| W_0 - \widehat{W} \right\|^2 - \|W_0 - W^*\|^2 = \|W_0 - kW^*\|^2 - \|W_0 - W^*\|^2 \tag{59}$$

$$= \left(\|W_0\|^2 - k^2 \|W^*\|^2 \right) - \left(\|W_0\|^2 - 2k \|W^*\|^2 + \|W^*\|^2 \right) \tag{60}$$

$$= 2k \|W^*\|^2 - k^2 \|W^*\|^2 - \|W^*\|^2 \tag{61}$$

$$= - \|W^*\|^2 \cdot (1 - k)^2 \tag{62}$$

6 实验

由于本文的目的是修正对于 Batch Normalization 的错误认知，所以实验主要是通过自己提出的梯度可预测性和梯度差异等参数进行计算并绘制图像进行比较。作者实验中展示的数据虽然多但是由于针对性非常强，如数据分布图，或者梯度的预测性变化，我们可以一目了然地看出结果，并且通过实验有力地证明了 BatchNorm 的的成功与网络中各层标准化后使得分布相同无关，而是它使优化环境变得更加平滑。

对于图 2中数据分布的可视化，作者分别对于给定网络层中连续步骤之间其输入分布的均值和方差变化进行了绘制，如图 6所示，添加噪声的 BatchNorm 的均值和方差有着非常大的变化，发生了非常严重的 ICS。

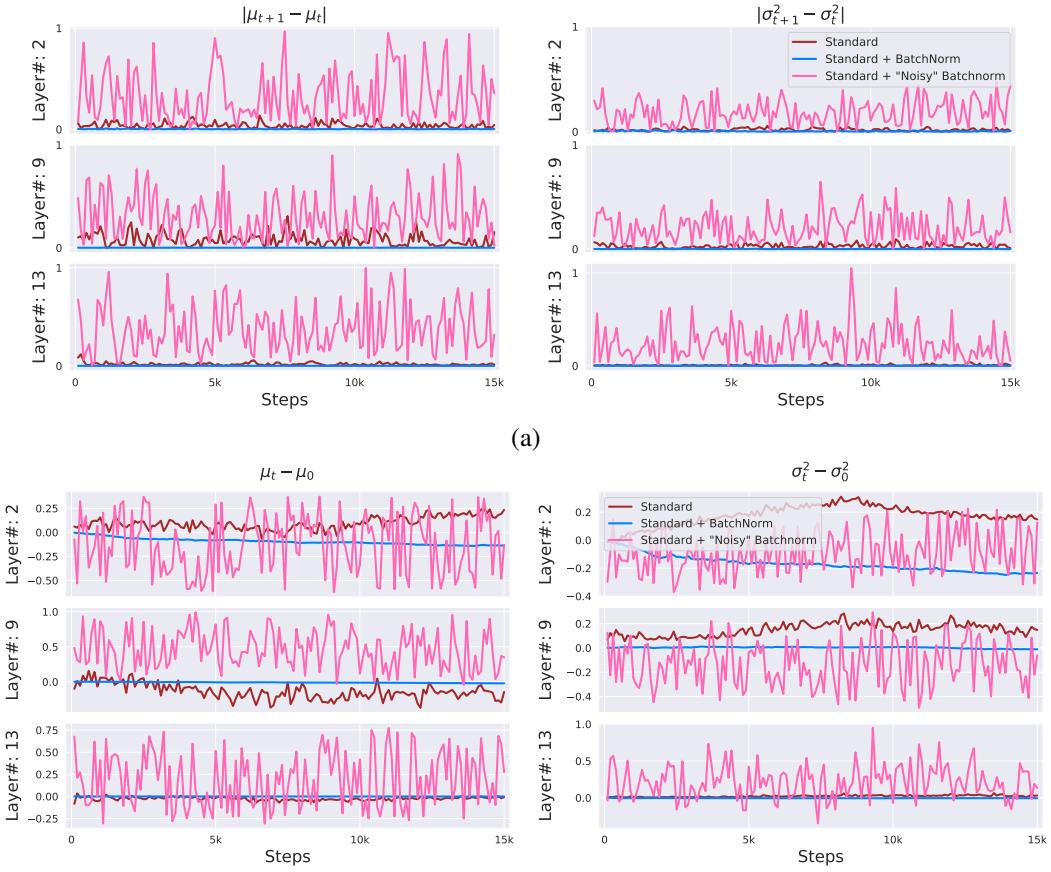
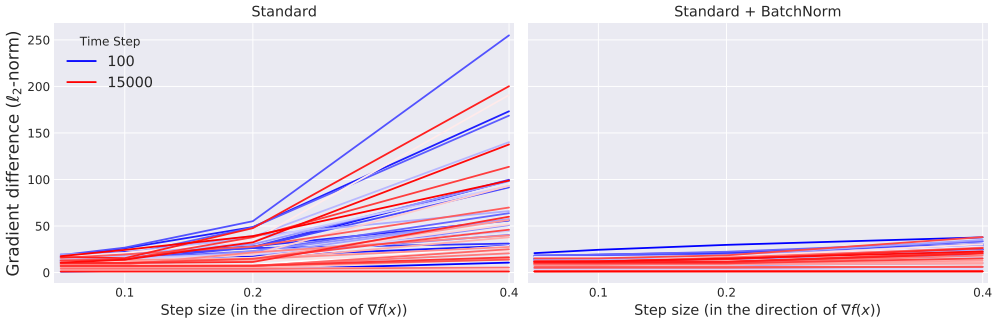
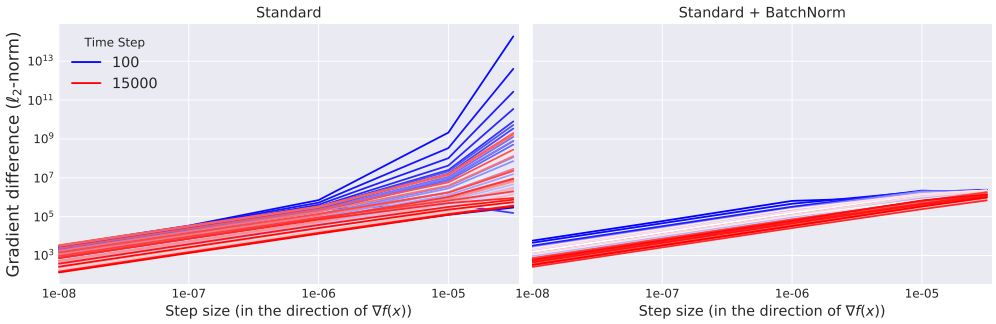


图 6: 确定网络层中连续步骤之间其输入分布的均值和方差变化

除了图 3以外，作者还用折线图重新绘制了在有和没有 BatchNorm 的 VGG 和 DLN 网络上的梯度差异，如图 7所示，可以观察到梯度在具有 BatchNorm 的网络中更具有预测性，并且在邻域中变化缓慢。因此具有 BatchNorm 的网络在很大程度上对广泛的学习率具有鲁棒性。



(a) VGG



(b) DLN

图 7: 是否使用 BatchNorm 的 VGG 和 DLN 网络上的梯度差异

7 启发

这篇文章最大的贡献是两个方面，一是对于普遍认知提出质疑，二是找到了根本原因.BatchNorm 作为深度学习中常用的技术手段，不仅有利于从原理上理解深度学习的可解释性，还可以帮助我们改进技术手段，促进深度学习的发展. BatchNorm 的作用不是在于稳定分布，虽然这是提出这项技术的出发点，但是真正提高性能的原因是让损失函数变得更加平滑和连续. 这告诉我们阅读论文要带着批判的眼光去看，不能写什么就相信什么，要有自己的想法，善于发现，这样才会有科研产出.

BatchNorm 并不是一剂万能药，只有在满足某种机制时才可以提高模型性能. 这是未来研究 BatchNorm 的一个问题之一，深度学习作为当前人工智能最热门的领域之一，尝试去探究他的可解释性，不仅可以解决长时间以来解释性差的问题，还可以促进将深度学习应用到更为广泛，更加符合学科交叉的领域当中. 这篇文章纠正了长期以来深度学习学者的认知错误，也给所有深度学习新技术带来了思考，或许某些问题的出发点并不是根本原因, 以严谨的科学态度才可以发现新的理论成果.

8 分工

论文的核心部分例如公式的推导和问题的解决由两人讨论共同完成，背景、作者要解决的问题、实验由颜劭铭撰写，作者如何解决问题、公式推导、启发由杨铭撰写.

参考文献

[1] 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. PMLR, 2015.

[2] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[4] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[5] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[6] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016.

[7] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? *Advances in neural information processing systems*, 31, 2018.

[8] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.