计 算 机 学 报
CHINESE JOURNAL OF COMPUTERS

泛化界正则项：理解权重衰减正则形式的统一视角

李翔　陈硕　杨健

(南京理工大学 计算机科学与工程学院 PCALab，南京　210094)

摘　要　经验风险最小化(Empirical Risk Minimization, ERM)旨在学习一组模型参数来尽可能地拟合已观测到的样本，使得模型具有基础的识别能力。除了 ERM，权重衰减(Weight Decay, WD)对于进一步提升模型的泛化能力，即对未观测样本的精准识别也非常重要。然而，WD 的具体形式仅仅是在优化过程中不断缩小所学习的模型参数，这很难与提升泛化能力这个概念直接地联系起来，尤其是对于多层深度网络而言。本文首先从计算学习理论(learning theory)中的鲁棒性(robustness)与泛化性(generalization)之间的量化关系出发，推导出了一个统一的泛化界正则项(Generalization Bound Regularizer, GBR)来理解 WD 的作用。本文证明了优化 WD 项(作为损失目标函数的一部分)本质上是在优化 GBR 的上界，而 GBR 则与模型的泛化能力有着理论上的直接关联。对于单层线性系统，本文可以直接推导出该上界；对于多层深度神经网络，该上界可以通过几个不等式的松弛来获得。本文通过引入均等范数约束(Equivalent Norm Constraint, ENC)即保证上述不等式的取等条件来进一步压缩 GBR 与其上界之间的距离，从而获得具有更好泛化能力的网络模型，该模型的识别性能在大型 ImageNet 数据集上得到了全面的验证。

关键词　泛化界正则项；经验风险最小化；权重衰减；均等范数约束；深度神经网络

中图法分类号　TP391

Generalization Bound Regularizer: A Unified Perspective for Understanding Weight Decay

Li　Xiang　　Chen Shuo　　Yang Jian

(PCALab, Department of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094)

Abstract　Empirical Risk Minimization (ERM) aims to learn parameters of a model that can perfectly, at least, master a set of observed examples. Beyond ERM, the Weight Decay (WD) regularization term is also necessary to ensure the trained models with generalization ability on unseen objects. However, the form of WD targets at making the learning parameters small during optimization, which naturally lacks smooth connection to the concept of generalization, especially for multi-layer deep networks. This paper first aims to bridge this gap through a proposed unified framework, namely Generalization Bound Regularizer (GBR), which is theoretically deduced from the robustness and generalization theory. Specifically, we demonstrate that optimizing WD term, as a part of the loss objective, is actually optimizing an upper bound of the underlying GBR, which is directly related to the generalization ability of models. For a single-layer linear system, this upper bound can be derived directly; for a multi-layer deep network, this upper bound is obtained via additional relaxations of several inequalities. By introducing Equivalent Norm Constraint (ENC) and further equalizing the GBR and its corresponding upper bound, it is easy to get a more generalized model with improved recognition performance, which is comprehensively validated on the large-scale ImageNet dataset.

Key words　generalization bound regularizer; empirical risk minimization; weight decay; equivalent norm constraint; deep neural network

---

本课题得到国家自然科学基金(No.U1713208)、111项目(AH92005)资助. 李翔，博士，主要研究领域为深度学习、计算机视觉. E-mail: xiang.li.implus@njust.edu.cn. 陈硕，博士，主要研究领域为机器学习、度量学习. E-mail: shuochen@njust.edu.cn. 杨健（**通信作者**），博士，教授，主要研究领域为机器学习、模式识别. E-mail: csjyang@njust.edu.cn.

1 引言

机器学习系统经过训练能够在已观测数据上达到尽可能低的平均误差，这通常被称为经验风险最小化(Empirical Risk Minimization, ERM)原理[1]。ERM 保障了学习系统针对已观测样本所具有的基础识别能力，但却容易带来过拟合的风险，从而无法将该识别能力更好地泛化到未观测样本中去。

随着权重衰减(Weight Decay, WD)正则化技术的提出，模型过拟合的影响也随之降低，无论是在单层线性模型[2]还是在多层深度神经网络[3-7]中。WD 可以追溯到文献[8]，它证明了 WD 为线性网络泛化能力的提升带来了诸多益处[8,9]。对于大多数优化器而言[10-12]，WD 在数学上完全等价于 L_2 正则化[13,14]，其表达式为：

$$\frac{1}{2}\lambda\left(\sum_{i=1}^{L}\left\|\mathbf{W}^i\right\|_F^2\right), \qquad (1)$$

其中 λ 是其系数，$\mathbf{W}^i$ 表示总计 L 层中的第 i 层的权重，而 $\|\cdot\|_F$ 表示 Frobenius 范数。

将 WD 最小化作为损失函数目标的一部分可以视为找到模型参数权重处于较小数量级的一个解，这在实践中被广泛地证明能够提升模型的泛化性能。但是，“减小权重”似乎在直觉上与“更好的泛化能力”的联系并不那么紧密。相关文献已经给出了一些简单的解释，例如在文献[8]中，对于非线性网络，作者指出由 WD 优化得到的解具有潜在的最小复杂度，从奥卡姆剃刀的角度来看很有可能成为最好的策略。这种解释缺乏对深度神经网络的进一步探索。尽管文献[9]证明可以通过减小模型的权重范数来缩小传统简单线性模型的 Rademacher 复杂度上界，但是最新的研究[15]表明 Rademacher 复杂度对于深度神经网络而言其参考意义受限，原因在于它们本身的数值都非常接近于 1，并且深度神经网络能够很好地拟合随机标签。因此，传统的 Rademacher 复杂度无法深入并定量地解释 WD 为何能在深度网络中提高泛化能力。

本文首先希望为 WD 提供理论上的更深入且统一的理解。具体来说，基于鲁棒和泛化的理论分析[16]基础，本文从学习算法的泛化上界入手，经过进一步推导，该上界可以通过降低模型最终的预测相对输入的敏感度来同步地被降低。通常，这样的敏感度可以表示为在整个输入空间上，输出对输入梯度的大小(本文将其称为“泛化界正则项”(Generalization Bound Regularizer, GBR))，即

$$\int_{x\sim\mu}\left\|\frac{\partial f\left(x;\left\{\mathbf{W}^1,\ldots,\mathbf{W}^L\right\}\right)}{\partial x}\right\|_F^2 P(x)\mathrm{d}x\,, \qquad (2)$$

其中输入 $x\in\mathbb{R}^{m_0}$ 来自数据分布 μ (概率密度函数为 $P(\cdot)$)的样本，本文记 $f\left(x;\left\{\mathbf{W}^1,\ldots,\mathbf{W}^L\right\}\right)$ 为包含 L 层参数 $\left\{\mathbf{W}^1,\ldots,\mathbf{W}^L\right\}$ 的学习系统的输出，其中 $\mathbf{W}^i\in\mathbb{R}^{m_i\times m_{i-1}}, i\in\{1,\ldots,L\}$， m_i 表示每一层中的特征维度。因此，公式(2)可以视为与泛化能力直接相关的数学表达式。本文可以证明对于单层线性模型，公式(2)等价于 WD 的定义；同时，对于多层深度神经网络，本文也可以通过多个不等式将公式(2)放大到 WD 的形式。这表明了 WD 正则本质上是深度网络 GBR 数值的上界。当网络的训练过程努力最小化这个上界 WD 时，其所期望相应实际的 GBR 值也将隐式地变小，从而提升网络的泛化性能。

在揭示 GBR 和 WD 之间联系的统一视角的基础上，本文探索了进一步增强它们之间的关联，使得在优化过程中能够进一步降低 GBR。具体来说，本文通过让所有层权重的范数相同，从而使得均值不等式取到等号，并将其称为均等范数约束(Equivalent Norm Constraint, ENC)条件。在大型 ImageNet 数据集上进行的实验表明，尽管 ENC 条件非常简单，但它仍然可以持续改善网络的泛化性能，并使深层模型对对抗样本更加鲁棒。

根据上述分析，基于鲁棒和泛化性的基础理论，本文推导了与模型的泛化能力直接相关的泛化界正则项(GBR)框架，并从理论上解释了优化权重衰减(WD)本质上是优化了 GBR 的上界，从而为 WD 提供了一个统一的理解。与此同时，根据所提出的统一框架，对于多层神经网络，本文尝试通过进一步加强 GBR 和 WD 之间的联系，并提出均等范数约束(ENC)条件进一步限制 GBR 的大小，它被证明对提高深度神经网络的泛化能力非常有效。

2 相关工作

本文将从理解权重衰减、泛化能力、对抗样本

和正则技术这四大方面依次展开对相关工作的简要概述。

理解权重衰减(WD)：对于非线性深度神经网络，在早期的工作[8]中，WD 从哲学的视角来看被认为是一种最佳的策略，因为它具有最小的复杂度。它可以被视为一种启发式的且最为简单的惩罚项，用来防止模型具有较大的参数值（或者说较大的复杂性）。此外，具有较小参数数值的函数往往会改善自身的 Lipschitzness[17,18]。同时，WD 可以确保一些特定函数[19]的平滑性。大量仿真实验[20,21]表明 WD 能够提高模型的容错能力。基于 RBF 模型[22]，文献[23]证明了在温和条件下可以将显式的正则化模式[24]简化为 WD。文献[25]确定了 WD 正则化的三种不同机制，包括（1）提高有效学习率[13,26]，（2）正则化输入输出的 Jacobian 范数，以及（3）降低二阶优化的有效阻尼系数。然而，这些相关工作都是从 WD 的作用的角度进行阐述的，而并没有深入探究 WD 形式的根源。而本文的工作希望基于计算学习理论(learning theory)中的鲁棒性(robustness)与泛化性(generalization)之间的量化关系[16]，为理解 WD 的形式提供一个统一的理论框架。

泛化能力：文献[15]证明了 Rademacher 复杂度和 VC 维的经典概念不足以理解深度网络的泛化性。为了解释泛化性，传统方法一般是利用局部最小值[27,28]的平坦度或清晰度的概念，该概念考虑了损失函数对模型参数摄动的敏感性。在本文的框架中，预测对输入信号的敏感性也可以直接与泛化界相关。

对抗样本：GBR 的推导形式表示了预测对于输入的敏感性，其在一些研究[29,30]中是直接被作为先验假设的。因此，它也与对抗样本的概念有关。对抗样本又被称为可以最大程度地增加损失函数从而带来干扰的训练数据[31]。在文献[32]中，作者展示了使用对抗样本进行训练可以提高模型的泛化能力。因此，降低 GBR 很有可能会使模型对干扰噪声和更多未知数据具有更好的鲁棒性，其将在后文中通过实验进行佐证。

正则技术：与 WD 类似，本文所提出的均等范数约束条件(ENC)主要对权重进行正则化。还有许多其他类型的正则化技术，它们着重于对特征[33,34]，数据[35–37]或标签[36–38]进行正则化。Dropout[33]和 DropBlock[34]在训练过程中以元素或元素块的方式引入了随机特征丢弃。Cutout[35]，Mixup[36]和 CutMix[37]分别提出了区域蒙版、成对线性组合以及区域复制和粘贴操作，通过扩展训练数据来提高模型泛化能力。其中，Mixup 和 CutMix 会通过组合系数进一步调整相应的标签。LabelSmooth[38]试图通过将训练标签修改为软分布[39]来防止分类器对某个类别过于自信。本文将在实验部分展示所提出的 ENC 条件与这些正则化方法之间的比较。

3 理解权重衰减形式的统一视角

3.1 基础概念

本节首先介绍基本的概念以及必要的定义和定理，这些定义和定理是从算法的鲁棒和泛化理论[16]中引入的。

本文考虑以下一般的学习模型：给定一组训练样本，本文的目标是从假设集中选择一个假设。除非另有说明，否则本节中训练集的大小固定为 n。Z 和 H 分别表示整体样本分布集合和假设集合。本文使用 $\mathbf{s}$ 表示由 n 个训练样本 $(s_1,\ldots,s_n)$ 组成的训练集。本文旨在学习从 Z^n 到 H 的映射,即算法 A。本文还使用 $\mathrm{A}_s(x)\in\mathrm{Y}$ 来表示 $x\in\mathrm{X}$ 的预测，并使用 $|\,x$ 和 $|\,y$ 来分别表示点的 x 分量和 y 分量。例如，$s_{i|\,x}$ 是 s_i 的 x 分量。在这里，本文将 X 称为输入空间，将 Y 称为输出空间。给定训练集 $\mathbf{s}$，A_s 代表其学习到的假设。本文为每个假设 $h\in\mathrm{H}$ 和一个数据点 $z\in\mathrm{Z}$ 定义一个关联两者的损失函数 $l(h,z)$，并假设 $l(h,z)$ 为非负的，其上界为标量 M。本文根据[16]中的理论，可以定义变量 $\grave{o}(\mathbf{s})$，并给出如下描述：

定义 1. 算法 A 被定义为 $(K,\grave{o}(\mathbf{s}))$ 鲁棒的条件是 Z 能够被划分为 K 个不相交的集合 $\{C_i\}_{i=1}^{K}$，并且对于 $\forall s\in\mathbf{s}$，满足：

$$s,z\in C_i,\Rightarrow\left|l(\mathrm{A}_s,s)-l(\mathrm{A}_s,z)\right|\leq\,?(\mathbf{s}). \tag{3}$$

注意到能够在同一集合 C_i 中的两个数据点意味着它们的欧氏距离相对较小，并以 ε 为界，也就是说，$\|s_{|\,x}-z_{|\,x}\|\leq\varepsilon$ 和 $|s_{|\,y}-z_{|\,y}|\leq\varepsilon$ 同时成立.

接下来，本文考虑标准的学习设置，即样本集合 $\mathbf{s}$ 由 n 个独立同分布的，从未知分布 μ 中抽取的样本构成，本文的学习目标是最小化经验损失。本文令 $\hat{l}(\cdot)$ 和 $l_{\mathrm{emp}}(\cdot)$ 分别表示期望误差和经验误差：

$$\hat{l}(\mathrm{A}_s)\triangleq\mathrm{E}_{z\sim\mu}l(\mathrm{A}_s,z);l_{\mathrm{emp}}(\mathrm{A}_s)\triangleq\frac{1}{n}\sum_{s_i\in s}l(\mathrm{A}_s,s_i) \tag{4}$$

然后，本文可以得到如下定理，该定理显示了学习算法 A_s 的期望误差与经验误差之间差距的上界，如文献[16]所示：

定理 1. 如果 **s** 包含 n 个独立同分布的样本，并且 A 是 $\left(K,\grave{o}(\mathbf{s})\right)$ 鲁棒的,那么对于任意 $\delta>0$，有至少 $1-\delta$ 的概率使得下式成立：

$$\left|\hat{l}\left(\mathrm{A}_s\right)-l_{\mathrm{emp}}\left(\mathrm{A}_s\right)\right|\leq ?(\mathbf{s})+M\sqrt{\frac{2K\ln 2+2\ln(1/\delta)}{n}}. \tag{5}$$

3.2 泛化界正则项

本文首先对分类问题进行分析。公式(5)证明了期望误差与经验误差之间的差异可以通过两项来界定：$\grave{o}(\mathbf{s})$ 和 $M\sqrt{\frac{2K\ln 2+2\ln(1/\delta)}{n}}$。因为训练集是固定的，所以第二项中的所有常数都是固定，无法更改。本文尝试深入到第一项，即 $\grave{o}(\mathbf{s})$，目标是将其重新表述为与训练参数有关的能够参与优化的具体表达式。根据公式(5)可知，$\grave{o}(\mathbf{s})$ 的定义是任意训练样本 s 的损失与来自 Z 的相似样本 z 的损失之差的上界。为了简化推导，本文将 $\mathrm{A}_s\left(s_{|x}\right)$ 表示为其对应标签 $S_{|y}$ 的概率值，而不是整个 Softmax 向量。本文还将 $\mathrm{A}_s'\left(z_{|x}\right)$ 表示为预测相对于输入 $z_{|x}$ 的梯度，从而可以证明如下定理：

定理 2. 对于一个分类任务的算法 A_s，本文假设其损失函数 $l\left(\mathrm{A}_s,z\right)$ 是 L-*Lipschitz* 的[1]。如果存在 $\left\|\mathrm{A}_s'\left(z_{|x}\right)\right\|\leq\frac{\grave{o}(\mathbf{s})}{L\varepsilon}$，那么 A 满足**定义 1**，即 A 是 $\left(K,\grave{o}(\mathbf{s})\right)$ 鲁棒的。

Proof. 首先从分类任务的情形入手。由于 $l\left(\mathrm{A}_s,z\right)$ 是 L-*Lipschitz* 的，并且 $\mathrm{A}_s\left(s_{|x}\right)$ 是其对应标签 $s_{|y}$ 的概率数值,注意到在分类任务中通常使用交叉熵损失函数[40,41]即 $-\ln(\cdot)$，于是可以选择 $s,z\in C_i$，即 $\left\|s_{|x}-z_{|x}\right\|\leq\varepsilon$，使得有：

$$\left|l\left(\mathrm{A}_s,s\right)-l\left(\mathrm{A}_s,z\right)\right|\leq L\left|\mathrm{A}_s\left(s_{|x}\right)-\mathrm{A}_s\left(z_{|x}\right)\right|. \tag{6}$$

根据泰勒展开[42]以及 $\left\|s_{|x}-z_{|x}\right\|\leq\varepsilon$，通过忽略高阶无穷小，本文可以进一步得到：

$$\begin{aligned}\left|l\left(\mathrm{A}_s,s\right)-l\left(\mathrm{A}_s,z\right)\right|&\leq L\left|\mathrm{A}_s\left(s_{|x}\right)-\mathrm{A}_s\left(z_{|x}\right)\right|\\&=L\left|\left(\mathrm{A}_s'\left(z_{|x}\right)\right)^{\cdot}\left(s_{|x}-z_{|x}\right)\right|\\&\leq L\left\|\mathrm{A}_s'\left(z_{|x}\right)\right\|\left\|s_{|x}-z_{|x}\right\|\\&\leq L\frac{\grave{o}(\mathbf{s})}{L\varepsilon}\varepsilon=\grave{o}(\mathbf{s}).\end{aligned} \tag{7}$$

现在有了 $\left|l\left(\mathrm{A}_s,s\right)-l\left(\mathrm{A}_s,z\right)\right|\leq ?(\mathbf{s})$，这意味着 A_s 满足了**定义 1**，也就完成了在分类任务下的证明。

本文进一步考虑回归任务的情况。这里需要将条件调整为 $\left\|\mathrm{A}_s'\left(z_{|x}\right)\right\|\leq\frac{\grave{o}(\mathbf{s})-\varepsilon}{L\varepsilon}$。考虑到 $\left|s_{|y}-z_{|y}\right|\leq\varepsilon$ 和 $\mathrm{A}_s\left(s_{|x}\right)$ 是用于作为预测的回归数值，本文也可以类似地得到：

$$\begin{aligned}\left|l\left(\mathrm{A}_s,s\right)-l\left(\mathrm{A}_s,z\right)\right|&=\left|\left|\mathrm{A}_s\left(s_{|x}\right)-s_{|y}\right|-\left|\mathrm{A}_s\left(z_{|x}\right)-z_{|y}\right|\right|\\&?\left|\mathrm{A}_s\left(s_{|x}\right)-s_{|y}-\mathrm{A}_s\left(z_{|x}\right)+z_{|y}\right|\\&?\left|\mathrm{A}_s\left(s_{|x}\right)-\mathrm{A}_s\left(z_{|x}\right)\right|+\left|s_{|y}-z_{|y}\right|\\&?L\left(\frac{\grave{o}(\mathbf{s})-\varepsilon}{L\varepsilon}\right)\varepsilon+\varepsilon=\grave{o}(\mathbf{s}).\end{aligned} \tag{8}$$

因此 $\left|l\left(\mathrm{A}_s,s\right)-l\left(\mathrm{A}_s,z\right)\right|\leq ?(\mathbf{s})$ 依然成立，结论得证。注意到调整的条件 $\left\|\mathrm{A}_s'\left(z_{|x}\right)\right\|\leq\frac{\grave{o}(\mathbf{s})-\varepsilon}{L\varepsilon}$ 与分类的情况虽然有区别，但是对于后续的推导及分析所得出的结论是一致的。

证毕.

由**定理 2** 可知，如果网络预测对输入的梯度的模(即，$\left\|\mathrm{A}_s'\left(z_{|x}\right)\right\|$)能够在学习的过程中被主动地压缩，就能够有一个更小的 $\grave{o}(\mathbf{s})$ 来满足

[1] 对于分类任务而言，这个假设很平常，因为其一般的损失函数为 $l(\cdot)=-\ln(\cdot)$，并且本文此前就已经假设 $l(\cdot)$ 的上限为 M。

$\left\|\mathrm{A}_s'\left(z_{|x}\right)\right\|_{\mathrm{,}}$ $\frac{\grave{o}(\mathbf{s})}{L\varepsilon}$ 这个条件，从而使得 A_s 达到 $\left(K,\grave{o}(\mathbf{s})\right)$ 鲁棒。根据**定理 1**，如果 $\grave{o}(\mathbf{s})$ 较小，即可以直接减小经验误差和期望误差之间的差距，从而有效地提高学习模型的泛化能力。

本文从这一项：$\left\|\mathrm{A}_s'\left(z_{|x}\right)\right\|$ 入手，探索它是否有可能被显示地优化。依据前文表述的符号系统，在整个数据分布中最小化 $\left\|\mathrm{A}_s'\left(z_{|x}\right)\right\|$ 等价于最小化公式(2)：

$$\begin{aligned}\mathrm{L}(f)&=\mathrm{E}_{x\sim\mu}\left\|f'(x)\right\|_F^2\\&=\int_{x\sim\mu}\left\|\frac{\partial f\left(x;\left\{\mathbf{W}^1,\ldots,\mathbf{W}^L\right\}\right)}{\partial x}\right\|_F^2 P(x)\mathrm{d}x\,,\end{aligned}$$

其中 x 是从真实未知分布 μ 中采样的。为了便于参考，本文将其称为模型 f 的“泛化界正则项”(Generalization Bound Regularizer, GBR)，即 $\mathrm{L}(f)$，它源自计算学习理论中的鲁棒性与泛化性之间的量化关系[16]以及本文的进一步分析推导。然而，直接优化公式(2)会遇到两个严重的问题：

- 遍历未知分布 μ 中的所有样本来计算公式(2)是无法实现的。
- 即使本文使用已知的训练样本来近似公式(2)，它仍然需要对整个网络进行额外的反向传播处理，并需要额外的内存来计算梯度，这会产生大量的时间和资源开销。

基于这些问题，本文考虑简化公式(2)并尝试找到其上界，该上界可能会更易于被优化。

3.3 从泛化界正则项到权重衰减

本文的分析主要从两个方面进行：单层线性模型和多层感知机，其中多层感知机以 ReLU[4,43]作为激活函数。在整个分析过程中，本文证明了泛化界正则项(GBR)可以自然地简化为 WD 的精确(或几乎精确)表达式，从而为 WD 提供了统一的理解。

3.3.1 单层线性模型

继承上述的符号系统，本文首先忽略偏差项，并将单层线性模型表示为：

$$y=f\left(x;\left\{\mathbf{W}^1\right\}\right)=\mathbf{W}^1x\,,\tag{9}$$

其中 $y\in\mathbb{R}^{m_1}$.根据公式(2),本文可以推导出公式(9)的 GBR：

$$\begin{aligned}\mathrm{L}(f)&=\int_{x\sim\mu}\left\|\frac{\partial f\left(x;\left\{\mathbf{W}^1\right\}\right)}{\partial x}\right\|_F^2 P(x)\mathrm{d}x\\&=\int_{x\sim\mu}\left\|\frac{\partial\left(\mathbf{W}^1x\right)}{\partial x}\right\|_F^2 P(x)\mathrm{d}x\\&=\left\|\left(\mathbf{W}^1\right)^{\cdot}\right\|_F^2\int_{x\sim\mu}P(x)\mathrm{d}x=\left\|\mathbf{W}^1\right\|_F^2,\end{aligned}\tag{10}$$

通过将系数 $\frac{1}{2}\lambda$ 引入到公式(10)中，本文可以直接得到针对单层线性模型的 GBR 的公式为：

$$\frac{1}{2}\lambda\left\|\mathbf{W}^1\right\|_F^2.\tag{11}$$

这恰好是 WD 正则项的表达式。此外，其他传统的非深度方法的正则项也可以用类似的方式导出。接下来，本文将偏差项引入到公式(9)中，得到：

$$y=f^b\left(x;\left\{\mathbf{W}^1,b^1\right\}\right)=\mathbf{W}^1x+b^1.\tag{12}$$

其中 $b^1\in\mathbb{R}^{m_1}$。再一次推导其 GBR，能够得到：

$$\begin{aligned}\mathrm{L}\left(f^b\right)&=\int_{x\sim\mu}\left\|\frac{\partial f^b\left(x;\left\{\mathbf{W}^1,b^1\right\}\right)}{\partial x}\right\|_F^2 P(x)\mathrm{d}x\\&=\int_{x\sim\mu}\left\|\frac{\partial\left(\mathbf{W}^1x+b^1\right)}{\partial x}\right\|_F^2 P(x)\mathrm{d}x=\left\|\mathbf{W}^1\right\|_F^2,\end{aligned}\tag{13}$$

从而再一次得到了相同的表示。同时，本文也在理论上证明，约束偏差项对于泛化能力的提升是不必要的，因为根据上述推导，在 GBR(即 WD)的表达式中实际上不存在偏差项。

3.3.2 多层感知机

本节考虑多层感知机的情况。根据上一节的推导，本文仅考虑没有偏差项的情况，因为上文已经表明偏差项不会出现在最终的表达式中。另外，卷积神经网络可以看成是卷积参数在空间上共享的复杂多层感知机，也可以推导得到类似的结论，在这里为了简便起见，仅对多层感知机进行推导展开。具体地，本文通过简单地堆叠每一层将单层模型扩展为多层模型：

$$y=\hat{f}\left(x;\left\{\mathbf{W}^1,\ldots,\mathbf{W}^L\right\}\right)=\mathbf{W}^L\left(\cdots\delta^2\left(\mathbf{W}^2\delta^1\left(\mathbf{W}^1x\right)\right)\right),\tag{14}$$

其中 $\delta^i(\cdot)$ 表示 ReLU 函数，$\nabla\delta^i$ 表示为其对应的梯度。请注意，本文在这里不考虑“批量归一化”[44](Batch Normalization, BN)层，因为在测试阶段，它会变成纯线性变换，并且可以融合到卷积权重[45]中。同样地，根据公式(2)可以推导出其 GBR：

$$\begin{aligned}\mathrm{L}(\hat{f}) &= \int_{x\sim\mu}\left\|\frac{\partial\hat{f}\left(x;\left\{\mathbf{W}^1,\ldots,\mathbf{W}^L\right\}\right)}{\partial x}\right\|_F^2 P(x)\mathrm{d}x \\ &= \int_{x\sim\mu}\left\|\frac{\partial\left(\mathbf{W}^L\left(\cdots\delta^2\left(\mathbf{W}^2\delta^1\left(\mathbf{W}^1x\right)\right)\right)\right)}{\partial x}\right\|_F^2 P(x)\mathrm{d}x \\ &= \int_{x\sim\mu}\left\|\left(\mathbf{W}^1\right)^{\top}\left(\nabla\delta^1\odot\left(\mathbf{W}^2\right)^{\top}\right)\cdots\right. \\ &\quad\left.\left(\nabla\delta^{L-2}\odot\left(\mathbf{W}^{L-1}\right)^{\top}\right)\left(\nabla\delta^{L-1}\odot\left(\mathbf{W}^L\right)^{\top}\right)\right\|_F^2 P(x)\mathrm{d}x,\end{aligned} \tag{15}$$

其中 $\odot$ 表示 Hadamard 乘积[46]，它沿 $\mathbf{W}^i$，$i\in\{2,\mathrm{L},L\}$ 与 $\nabla\delta^{i-1}$ 所匹配的维度进行逐元素乘法。由于 $\delta(\cdot)$ 是 ReLU 激活函数，因此 $\nabla\delta(\cdot)$ 数值为 0 或 1，这意味着 $\nabla\delta^i$ 的第 j 个元素满足 $\nabla\delta_j^i\le 1$，即：

$$\left\|\nabla\delta^{i-1}\odot\left(\mathbf{W}^i\right)^{\top}\right\|_F^2 \le \left\|\left(\mathbf{W}^i\right)^{\top}\right\|_F^2 = \left\|\mathbf{W}^i\right\|_F^2. \tag{16}$$

基于矩阵不等式[47]和均值不等式[48]，容易得到：

$$\begin{aligned}\mathrm{L}(\hat{f}) &\le \int_{x\sim\mu}\left\|\mathbf{W}^1\right\|_F^2\left\|\nabla\delta^1\odot\left(\mathbf{W}^2\right)^T\right\|_F^2\cdots \\ &\left\|\nabla\delta^{L-2}\odot\left(\mathbf{W}^{L-1}\right)^T\right\|_F^2\left\|\nabla\delta^{L-1}\odot\left(\mathbf{W}^L\right)^T\right\|_F^2 P(x)\mathrm{d}x \\ &\le \int_{x\sim\mu}\prod_{i=1}^{L}\left\|\mathbf{W}^i\right\|_F^2 P(x)\mathrm{d}x \\ &= \prod_{i=1}^{L}\left\|\mathbf{W}^i\right\|_F^2 \le \left(\frac{\sum_{i=1}^{L}\left\|\mathbf{W}^i\right\|_F^2}{L}\right)^L.\end{aligned} \tag{17}$$

从而再次得到与 WD 非常相似的定义。它们唯一的区别是，公式(17)中还有一个幂次项 $L(L>1)$。值得一提的是，如果考察其它的激活函数，例如 Sigmoid 函数，根据其梯度的性质能够得到 $\nabla\delta_j^i\le\frac{1}{4}$，可以得出公式(16)(17)依然成立。

3.4 统一理解

基于计算学习理论中的鲁棒性与泛化性之间的量化关系[16]，本文介绍了统一视角的泛化界正则项(GBR)框架，揭示了如何通过隐式优化 GBR 来提高学习模型的泛化能力。本文发现它与广泛使用的 WD 项(从单层线性模型到多层感知机)具有非常紧密和统一的联系。本文通过该框架得出如下重点结论：

- 对于单层线性模型，GBR 等价于 WD 项。
- 对于多层非线性网络，优化 WD 项实际上是在优化该模型 GBR 的上界，这是提高模型泛化能力的主要原因。
- 偏差项没有必要进行衰减，它们对泛化性能没有任何影响，因为根据本文的理论推导，它们实际上不会出现在优化目标中。注意到仅在最近几年，研究员们才开始逐渐经验性地在深度神经网络中对偏差项不采用权重衰减[39,49]，本文为这个实践提供了具体的理论依据。

4 方法

4.1 权重衰减与泛化界正则项的统计关系

在多层深度神经网络的情况下，优化 WD 项本质上是优化 GBR 的上限。因此，本文期望通过减少 WD 项间接减少 GBR，从而提升模型的泛化性能。为了进一步分析，本文首先对 WD 和 GBR 进行了一些数值统计，以观察两者的关联。具体来说，本文选择 ResNet-50[5]作为主干网络，并在集合{0，10^{-8}，10^{-6}，10^{-4}，10^{-3}，10^{-2}}中遍历 WD 的系数 λ 进行考察。在训练这些模型的过程中，本文基于训练集 $\mathbf{s}$ 中相同的训练数据来计算公式(18)，以近似 GBR 数值。尽管 GBR 是在实际数据分布 μ 中定义的，但是本文利用已知的训练样本对其进行的统计估计也能够一定程度上反映出它的统计特性，即如下式所示：

$$\frac{1}{n}\sum_{x\in\mathbf{s}}\left\|\frac{\partial f\left(x;\left\{\mathbf{W}^1,\ldots,\mathbf{W}^L\right\}\right)}{\partial x}\right\|_F^2. \tag{18}$$

统计结果呈现在图 1 中。随着 WD 系数的增大，近

似的 GBR 值逐渐减小，这完全符合本文的预期。从图 1 得到的另一个有趣的观察结果是，当没有 WD 时(即 $\lambda=0$)，GBR 值在训练期间将变得非常不稳定(请参见最上方曲线)。而仅仅只是使用非常小的 WD(例如 $\lambda=10^{-8}$)，它的稳定性也会大大提高。尽管这与本文的主题无关，但本文认为这个有趣的现象值得在未来的工作中进一步探索。

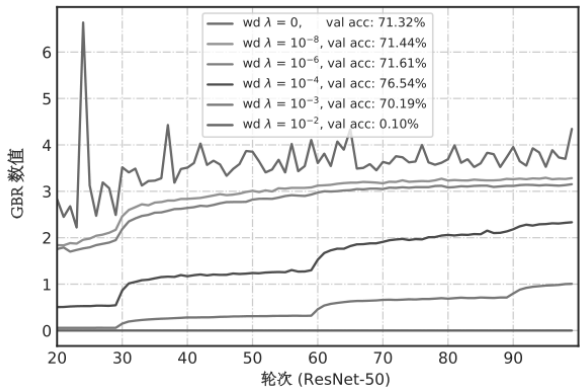


图 1　不同 WD 系数 λ 在训练过程中的近似 GBR 数值。本文同时还给出了最终模型在验证集上的 Top-1 精度

4.2 均等范数约束

图 1 表明增加 WD 系数 λ 实际上能够同时降低 GBR。但是，当 $\lambda>10^{-4}$ 时，模型性能会严重下降，因为它破坏了众所周知的“偏差方差平衡”[50]，带来了严重的欠拟合问题。因此，这促使本文去寻找新的方式(而不是简单地增加 λ)来进一步限制 GBR 的增长，并希望能不破坏偏差方差的平衡(即，至少保持最佳的 WD 系数 $\lambda=10^{-4}$)。为了实现此目标，本文考虑进一步加强它们的连接(即，公式(17))：是否可以进一步缩小它们的差距，以便新的优化目标可以在训练过程中进一步限制 GBR 的增加？

一个极端的情况是使公式(17)中的所有的不等式取等，这样就使得在深度非线性模型的情况下，最小化 WD 等效于直接最小化 GBR，这可能会进一步降低 GBR。但是，公式(17)中不是所有的不等式都能到等号(例如，$\nabla\delta_j^i=1$ 不可能在深层非线性网络中始终成立，否则它将退化为线性模型)。幸运的是，均值不等式可以有机会让其保持等号，而这唯一的附加条件是让每一层的参数权重的范数值相同(即等于 α)：

$$\left\|\mathbf{W}^1\right\|_F^2=\left\|\mathbf{W}^2\right\|_F^2=\cdots=\left\|\mathbf{W}^L\right\|_F^2=\alpha. \tag{19}$$

公式(19)即为本文所提出的均等范数约束(ENC)，其目的是进一步连接 GBR 和 WD。本文期望通过执行 ENC 条件，直接优化的上限(即 WD)更接近 GBR，从而可以进一步隐式地减少 GBR，以提升模型的泛化性能。

4.3 均等范数约束的实现

本节描述 ENC 的可能实现。通过回顾一系列权重重参数化方法，例如权重归一化[51](WN)和权重标准化[18](WS)，本文可以通过重参数化整个权重得到一个简单的解决方案，记 $\hat{\mathbf{W}}^i$ 为 $\mathbf{W}^i\in\mathbb{R}^{m_i\times m_{i-1}}$ 重参数化后的变量：

$$\hat{\mathbf{W}}^i=\sqrt{\alpha}\frac{\mathbf{W}^i}{\left\|\mathbf{W}^i\right\|_F},\forall i\in\{1,\ldots,L\}, \tag{20}$$

归一化的权重 $\hat{\mathbf{W}}^i$ 将直接参与神经网络的正向和反向的传播过程，这正好符合 ENC 中 $\left\|\hat{\mathbf{W}}^i\right\|_F^2=\alpha$ 的条件。请注意，它与 WN[51]不同：公式(20)通过计算 $\mathbf{W}^i$ (即 $m_i\times m_{i-1}$ 个元素)中所有元素的 Frobenius 范数来标准化整个权重。而 WN 仅沿着矩阵第一维，对第二维(即 m_{i-1} 个元素)的所有元素的权重进行归一化，这意味着 $\left\|\hat{\mathbf{W}}^i\right\|_F=\sqrt{m_i}$，而 m_i 在不同层之间是不相同的。因此，作为实现 ENC 的替代方法，本文可以部分地仿照 WN 并在每个层上添加一个额外的缩放常数，给定 $\forall i\in\{1,\ldots,L\},\forall j\in\{1,\ldots,m_i\}$：

$$\hat{\mathbf{W}}_j^i=\frac{\sqrt{\alpha}}{\sqrt{m_i}}\frac{\mathbf{W}_j^i}{\left\|\mathbf{W}_j^i\right\|}, \tag{21}$$

其中 ENC 条件 $\left\|\hat{\mathbf{W}}^i\right\|_F^2=\alpha$ 也随即成立。

类似地，本文也定义了基于权重标准化(WS)的 ENC 条件，给定 $\forall i\in\{1,\ldots,L\},\forall j\in\{1,\ldots,m_i\}$：

$$\hat{\mathbf{W}}_j^i=\frac{\sqrt{\alpha}}{\sqrt{m_i m_{i-1}}}\frac{\mathbf{W}_j^i-\overline{\mathbf{W}_j^i}}{\sqrt{\frac{\left\|\mathbf{W}_j^i-\overline{\mathbf{W}_j^i}\right\|^2}{m_{i-1}}}},\overline{\mathbf{W}_j^i}=\frac{\sum_{k=1}^{m_{i-1}}\mathbf{W}_{j,k}^i}{m_{i-1}}, \tag{22}$$

通过将零均值运算扩展到公式(20)中，本文可

以进一步得到 ENC 的一个新变体，即当 $\forall i \in \{1,\ldots, L\}$：

$$\hat{\mathbf{W}}^i = \sqrt{\alpha}\frac{\mathbf{W}^i - \overline{\mathbf{W}^i}}{\left\|\mathbf{W}^i - \overline{\mathbf{W}^i}\right\|_F}, \overline{\mathbf{W}^i} = \frac{\sum_{j=1}^{m_i}\sum_{k=1}^{m_{i-1}}\mathbf{W}^i_{j,k}}{m_i m_{i-1}}, \tag{23}$$

综上所述，公式(20)-(23)都符合 ENC 条件，其中，公式(21)是基于 WN 的实现，公式(22)则是基于 WS 实现的，而公式(20)和(23)则是基于 WN 或 WS 的变体实现的。它们都是训练过程中可能的选择，本文将在后续实验部分对其进行验证。

由于本文假定每一层参数 $\mathbf{W}^i$ 之后都有一个 BN 层，因此 $\mathbf{W}^i$ 的缩放比率将通过 BN 的特征归一化来消除。因此，除非另有说明，否则本文简单地令 $\alpha = 1$，即在训练过程中只需要确保 $\left\|\hat{\mathbf{W}}^i\right\|_F^2 = 1$，$\forall i \in \{1,\ldots,L\}$。除了 $\alpha = 1$ 外，在后续的实验中，本文还通过探索其他常数或在优化过程中使 α 变得可学习来研究 α 值的影响。通过严格遵循本文在多层网络中对 GBR 的推导，本文将 ENC 条件应用到所有具有可学习权重的网络层(BN 层除外)，并根据上述假设，本文确保每个层后面都有一个 BN 层。

5 实验

5.1 实验设置

为了验证所提出的 ENC 条件的有效性，本文对大型 ImageNet[52]图像分类数据集进行了综合实验。为了公平地进行比较，所有实验都在统一的 pytorch[53]框架下运行，包括每个基线模型的结果。训练设置与文献[54]保持一致，不同之处在于本文将网络[39]中所有偏差部分的权重衰减 λ 设置为 0，这也是本文 GBR 框架推导的结果(详见章节 3.4)。对于基本的数据增广，本文遵循文献[55]中标准做法：随机大小裁剪和随机水平翻转。本文通过 SGD[56]从头开始训练所有网络结构，其中权重衰减为系数为 0.0001，动量为 0.9，共学习 100 个轮次，学习率从 0.1 开始，每 30 个轮次将其减小 10 倍。总批次大小设置为 256，并且使用 8 个 GPU(每个 GPU32 张图像)进行训练。本文采用文献[57]中默认的权重初始化策略。请注意，由于权重归一化家族与 WD 之间存在一定的冲突[58]，因此在应用 ENC 时，本文还采用了 ò 偏移的二范数正则来保障训练的稳定。

5.2 ImageNet分类及消融实验

在本节中，本文通过如下多个问题来验证 ENC 的性能：

ENC 的最佳实现形式是什么？ 本文首先验证上述四个实现(即公式(20)-(23))的性能情况。本文选择 ResNet-50[5]作为主干网络，并在其之上使用 ENC 的四种不同实现。表 1 列出了相应的性能情况。为了更好地对比和参考，本文还列出了 WN-ResNet-50 和 WS-ResNet-50 的识别精度，因为 ENC 的实现是基于 WN 和 WS 的。训练和验证的精度曲线如图 2 中所示。从表 1 和图 2 中，本文观察到在 ENC 条件下，网络的训练可以收敛得更快，泛化性能也得到了较大的提高，精度超过原始基准大约 1 个百分点。同时，公式(22)取得了最佳的验证精度，所以本文将在后续实验中采用公式(22)作为ENC的默认实现。值得强调的是，在测试时，ENC 具有与基准模型相同的复杂度，而在训练阶段，类似于 WN 或 WS，归一化部分的额外计算成本也几乎可以忽略不计。这些特性使得 ENC 变得非常实用。

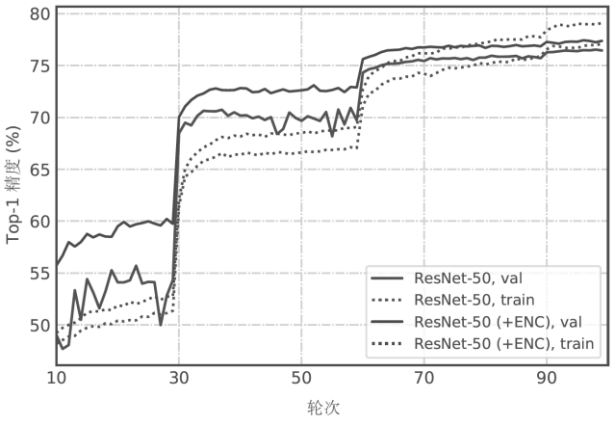


图 2 带有/不带有 ENC 的 ResNet-50 的精度曲线图，“train”代表训练集，“val”代表评估集

表 1 ENC 不同实现的性能以及多个基准模型的性能。括号中的数字代表相对于最原始基准的提升

类型		主干网络	Top-1 精度(%)
基线		ResNet-50[5]	76.54
		WN-ResNet-50[51]	76.44
		WS-ResNet-50[18]	76.74
ENC	公式(20)	ResNet-50[5]	77.09
	公式(21)	ResNet-50[51]	77.25
	公式(22)	ResNet-50[18]	**77.44**(+0.9)
	公式(23)	ResNet-50[5]	77.21

ENC 是否真的通过降低了 GBR 来提升模型泛化能力？本文从增强 GBR 和 WD 之间的联系出发，提出了 ENC 条件，希望 ENC 条件可以进一步限制 GBR 的增加。因此，本文在 ResNet-50[5]上令 WD 系数 $\lambda=10^{-4}$，对带有和不带有 ENC 条件下的 GBR 数值(即，公式(18))进行了绘制。本文还同时绘制了 $\lambda=10^{-3}$ 的曲线做进一步的参考。从图 3 中可以看出，ENC 确实进一步减小了 GBR 数值，而且其最终数值甚至比 $\lambda=10^{-3}$ 的情况还要小，从而验证了 ENC 的确是通过降低了 GBR 来提升模型泛化能力。

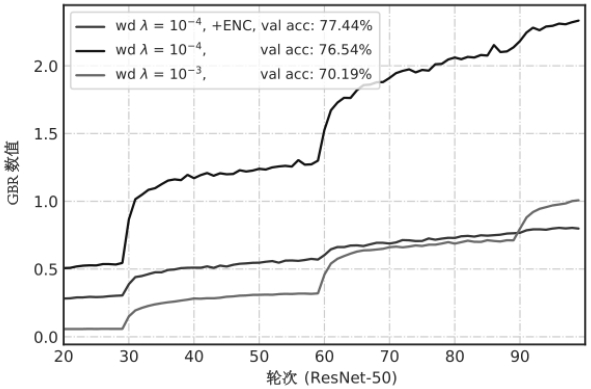


图 3 带有/不带有 ENC 的近似 GBR 数值的比较

表 2 使用最新正则化方法进行比较以及组合比较

类型	模型(ResNet-50)	Top-1 精度(%)
权重	+ L_2 (基线，WD $\lambda=10^{-4}$)	76.54
	+ENC	**77.44**
特征	+Dropout[33]	76.67
	+DropBlock[34]	76.79
数据/标签	+Cutout[35]	76.48
	+Mixup[36]	76.70
	+CutMix[37]	76.57
	+LS[38]	76.75
组合	+ENC+Cutout[35]	77.47
	+ENC+Mixup[36]	77.50
	+ENC+CutMix[37]	**77.94**
	+ENC+LS[38]	77.37
	+ENC+Cutout[35]+LS[38]	77.54
	+ENC+Mixup[36]+LS[38]	77.59
	+ENC+CutMix[37]+LS[38]	**77.96**

ENC 是否比其他提升泛化能力的正则方法有优势？作为一种正则化方法，本文将 ENC 与许多不同类型的最新正则化方法进行了比较。所有实验都是在相同的设置下进行公平地比较。根据表 2 中方法的类型，本文的实验可以分为四组：

- 权重：这组正则化方法着重于网络权重的调整，包括具有 L_2 正则化的基线 ResNet-50 结果和本文所提出的 ENC 条件。注意后续所有方法都会默认使用 L_2 正则化。ENC 条件将 L_2 基线提高了 0.9 个百分点的精度。

- 特征：本文利用 Dropout[33]和 DropBlock[34]作为两种主要作用于特征的正则化方法进行对比。参考文献[59]中的建议，Dropout 仅在最后一个 BN 层之后使用。DropBlock 采用了原文[34]中基于 ResNet-50 的最佳配置。该组正则化方法仅表现出相对于基线的细微改进(0.1%-0.2%)，与 ENC 的提升幅度差距很大(0.9%)。

表 3 在基于 ResNet-50 的 ImageNet 验证集上进行 FGSM 白盒攻击后，其 Top-1 (%) 准确性

类型	Top-1 精度(%)
基准	17.54
ENC	**29.87**$_{(+12.3)}$
Cutout[35]	18.26
Mixup[36]	21.64
CutMix[37]	**32.94**
LabelSmooth[38]	27.11
ENC+Cutout[35]	31.49
ENC+Mixup[36]	26.34
ENC+CutMix[37]	33.12
ENC+LabelSmooth[38]	34.03
ENC+CutMix[37]+LabelSmooth[38]	**34.81**

- 数据/标签：LabelSmooth[38](LS)，Cutout[35]，Mixup[36]和 CutMix[37]等最近流行的正则化方法旨在增强数据或/和标签以改善模型泛化性能。给定相同的 100 个训练轮次，Cutout 甚至会稍微降低性能，而其他方法与 Dropout/DropBlock 类似，仅有细微的改进(0.1%-0.2%)。而 ENC 非常有效地提高了基准模型的性能。

- 组合：本文进一步研究 ENC 带来的增益能否与其他类型的正则化技术正交，以便于同时将这些方法集成在一起，从而获得更高的性能。通过组合实

验，ENC 结合其他正则化方法能够进一步显著提高模型精度，例如“ENC+CutMix+LS”能够最大提升 1.4 个百分点。

ENC 可以抵抗更多的对抗性攻击吗？由于 ENC 旨在隐式降低 GBR，即预测对输入的敏感性，所以它对于对抗样本[32]应该具有更强的鲁棒性。为了验证这一点，本文使用了在 ImageNet 上预训练的 ResNet-50 作为主干网络，并采用了标准工具箱[2]中的快速渐变符号方法(Fast Gradient Sign Method, FGSM)[32]进行对不同正则化方式训练得到的网络进行白盒攻击。表 3 中展示了 ImageNet 验证集遭受攻击后不同方法的 Top-1 精度，其中 ENC 以~12 个百分点的绝对优势优于原始基准。值得一提的是，由于 ENC 不会像 Cutout，Mixup 和 LabelSmooth 那样引入大量的数据/标签的增广，但 ENC 仍以相当大的幅度超越了它们。ENC 没有能够超越最新的 CutMix，本文推测主要原因是因为 CutMix 在数据增广上做的非常灵活，使得网络有很大的机会在训练的阶段就能够接触到很多类似于对抗样本的增广样本。尽管如此，ENC 的性能与 CutMix 也非常接近，并且可以结合使用 LabelSmooth 进一步提高~2 个百分点。

表 4　在最新的卷积神经网络上使用 ENC 在 ImageNet 评估集上的性能提升。R 代表 ResNet，X 代表 ResNeXt，D 代表 DenseNet

模型	参数量	计算量	Top-1 精度(%)
R-50[5]	25.56M	4.122	76.54
R-50[5]+ENC			**77.44**$_{(+0.9)}$
R-101[5]	44.55M	7.850	78.17
R-101[5]+ENC			**78.44**$_{(+0.3)}$
X-50[60]	25.03M	4.273	77.64
X-50[60]+ENC			**78.26**$_{(+0.6)}$
X-101[60]	44.18M	8.033	78.71
X-101[60]+ENC			**78.96**$_{(+0.3)}$
SE-R-50[61]	28.09M	4.130	77.55
SE-R-50[61]+ENC			**78.23**$_{(+0.7)}$
SE-R-101[61]	49.36M	7.863	78.43
SE-R-101[61]+ENC			**78.75**$_{(+0.3)}$
D-201[6]	20.01M	4.367	77.54
D-201[6]+ENC			**77.96**$_{(+0.4)}$

[2]https://github.com/IBM/adversarial-robustness-toolbox

ENC 可以推广到更多深度卷积神经网络结构吗？此外，本文将最佳的 ENC 条件扩展到了其他各种最新卷积神经网络结构中，包括 ResNeXt[60]，SE-ResNet[61]和 DenseNet[6]，如表 4 所示。本文观察到了在模型复杂度没有变化的情况下，随着 ENC 的引入，模型性能都有了一致的提升。

α 的影响。如上所述，由于深度网络的批归一化模块的存在，在所有实验中，本文主要使用 $\alpha=1$ 的设置。本文进一步对 α 的取值做了研究，并尝试回答如下两个问题：(1) α 是否对其他常量敏感？(2) 是否可以直接学习 α？基于 ResNet-50，本文通过更多实验来回答上述问题，如表 5 所示。本文观察到使 α 变得可学习会降低网络的准确性，并且类似的这种现象也在 WS[18]的论文中得到佐证。本文推测引起该现象的主要原因是由于位于卷积层之后的 BN 层会再次对该卷积特征进行归一化，这实际上会抵消掉 α 对特征进行缩放的效果，从而带来学习效率的降低。同时，实验表明 α 取不同的常量不会对最终结果带来较大的影响。

表 5　基于 ResNet-50 对 α 影响的研究

α	0.5	1	2	可学习的
Top-1(%)精度	77.31	77.44	77.40	77.08

6 结论

本文首先基于计算学习理论中的鲁棒性与泛化性的量化关系，介绍了与模型泛化能力直接相关的泛化界正则项(Generalization Bound Regularizer, GBR)，并为理解权重衰减(Weight Decay, WD)提供了统一的视角：优化 WD 实际上是在优化 GBR 的上限，从而在理论上提高了模型的泛化能力。接下来，本文提出了均等范数约束(Equivalent Norm Constraint, ENC)条件，以进一步压缩 GBR 与其上界 WD 之间的距离，使得网络在训练过程中能够有效地抑制 GBR 的增加。通过在大型 ImageNet 图像数据集上进行的实验，本文全面验证了 ENC 的有效性和鲁棒性。

本文的意义分为理论和实践两个层面。首先在理论上，本文对训练神经网络的过程中所使用的权重衰减给出了一个统一的、崭新的理解视角。基于该视角，本文还能够从理论上回答为何近年来研究员们广泛且经验性地对偏差项不采用权重衰减。其

次，在实践上，本文根据理论的推导自然地给出了针对性的改进方案，该方案简单且有效，能够在不引入额外参数和计算复杂度的情况下提升基准网络的性能。

本文所提出的 ENC 方法也存在一定的不足。相比较于一些数据增强的正则化方法（如 CutMix），本文所提出的 ENC 条件在训练轮次增多的情况下，增益会逐渐弱于 CutMix。当然，这也是由方法本身性质所决定的。在不采用较强的数据增广技术（如 CutMix 等）的情况下，将训练轮次由 100 轮增加至 300 轮，ResNet-50 基准结果由 76.54 下降为 76.30，这是由于基准设置中数据增广幅度较弱造成了数据层面的严重的过拟合现象，导致增加训练轮次并不能带来性能的提升。在 300 轮的情况下，CutMix 可以带来接近 2 个点的增益，而 ENC 仅能够维持住此前约 1 个点的增益，其主要原因是由于随着轮次的增加，网络在数据层面的过拟合风险成为了最大的瓶颈，只有激进地增强数据增广力度的方法能够取得更加明显的效果。

参 考 文 献

[1] VAPNIK V. The nature of statistical learning theory. Berlin, Germany: Springer-Verlag, 1995: 1-50.

[2] HANSEL D, SOMPOLINSKY H. Learning from examples in a single layer neural network. EPL (Europhysics Letters), 1990, 11(7):687.

[3] RUCK D W, ROGERS S K, KABRISKY M, et al. The multilayer perceptron as an approximation to a bayes optimal discriminant function. IEEE Transactions on neural networks, 1990, 1(4):296-298.

[4] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks//Advances in neural information processing systems. Lake Tahoe, USA. 2012: 1097-1105.

[5] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, USA. 2016: 770-778.

[6] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, USA. 2017: 4700-4708.

[7] WANG W, LI X, YANG J, et al. Mixed link networks. arXiv preprint arXiv:1802.01808, 2018.

[8] KROGH A, HERTZ J A. A simple weight decay can improve generalization//Advances in neural information processing systems. Colorado, USA. 1992:950-957.

[9] SHALEV-SHWARTZ S, BEN-DAVID S. Understanding machine learning: From theory to algorithms. Cambridge, UK: Cambridge university press, 2014.

[10] BOTTOU L. Large-scale machine learning with stochastic gradient descent//Proceedings of COMPSTAT. New York, USA, 2010: 177-186.

[11] BOTTOU L. Stochastic gradient descent tricks//Neural networks: Tricks of the trade. Berlin, Germany: Springer, 2012: 421-436.

[12] LOSHCHILOV I, HUTTER F. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983, 2016.

[13] VAN LAARHOVEN T. L2 regularization versus batch and weight normalization. arXiv preprint arXiv:1706.05350, 2017.

[14] CORTES C, MOHRI M, ROSTAMIZADEH A. L2 regularization for learning kernels. arXiv preprint arXiv:1205.2653, 2012.

[15] ZHANG C, BENGIO S, HARDT M, et al. Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530, 2016.

[16] XU H, MANNOR S. Robustness and generalization. Machine learning, 2012, 86(3):391-423.

[17] SANTURKAR S, TSIPRAS D, ILYAS A, et al. How does batch normalization help optimization?//Advances in neural information processing systems. Montréal, Canada. 2018: 2483-2493.

[18] QIAO S, WANG H, LIU C, et al. Weight standardization. arXiv preprint arXiv:1903.10520, 2019.

[19] GIROSI F. An equivalence between sparse approximation and support vector machines. Neural computation, 1998, 10(6):1455-1480.

[20] MURRAY A F, EDWARDS P J. Enhanced mlp performance and fault tolerance resulting from synaptic weight noise during training. IEEE Transactions on neural networks, 1994, 5(5):792-802.

[21] LEUNG C S, SUM J P F. A fault-tolerant regularizer for rbf networks . IEEE Transactions on neural networks, 2008, 19(3):493-507.

[22] PARK J, SANDBERG I W. Universal approximation using radial-basis-function networks. Neural computation, 1991, 3(2):246-257.

[23] SUM J, LUO W H, HUANG Y F, et al. Equivalence between weight decay learning and explicit regularization to improve fault tolerance of rbf//International conference on intelligent systems design and applications. Taiwan, China. 2008: 152-157.

[24] BERNIER J L, ORTEGA J, ROJAS I, et al. Obtaining fault tolerant multilayer perceptrons using an explicit regularization. Neural processing letters, 2000, 12(2):107-113.

[25] ZHANG G, WANG C, XU B, et al. Three mechanisms of weight decay regularization. arXiv preprint arXiv:1810.12281, 2018.

[26] HOFFER E, BANNER R, GOLAN I, et al. Norm matters: efficient and accurate normalization schemes in deep networks//Advances in neural information processing systems. Montréal, Canada. 2018: 2160-2170.

[27] HOCHREITER S, SCHMIDHUBER J. Flat minima. Neural computation, 1997, 9(1):1-42.

[28] KESKAR N S, MUDIGERE D, NOCEDAL J, et al. On large-batch training for deep learning: Generalization gap and sharp minima. arXiv preprint arXiv:1609.04836, 2016.

[29] YOSHIDA Y, MIYATO T. Spectral norm regularization for improving

the generalizability of deep learning. arXiv preprint arXiv:1705.10941, 2017.

[30] RIFAI S, VINCENT P, MULLER X, et al. Contractive auto-encoders: Explicit invariance during feature extraction//Proceedings of international conference on machine learning. Washington, USA. 2011: 833-840.

[31] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

[32] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples. arXivpreprintarXiv:1412.6572, 2014.

[33] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 2014, 15(1):1929-1958.

[34] GHIASI G, LIN T Y, LE Q V. Dropblock: A regularization method for convolutional networks//Advances in neural information processing systems. Montréal, Canada. 2018: 10727-10737.

[35] DEVRIES T, TAYLOR G W. Improved regularization of convolutional neural networks withcutout. arXivpreprintarXiv:1708.04552, 2017.

[36] ZHANG H, CISSE M, DAUPHIN Y N, et al. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412, 2017.

[37] YUN S, HAN D, OH S J, et al. Cutmix: Regularization strategy to train strong classifiers with localizable features//Proceedings of the IEEE international conference on computer vision. Seoul, Korea. 2019: 6023-6032.

[38] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, USA. 2016: 2818-2826.

[39] HE T, ZHANG Z, ZHANG H, et al. Bag of tricks for image classification with convolutional neural networks//Proceedings of the IEEE conference on computer vision and pattern recognition. Long Beach, USA. 2019: 558-567.

[40] DE BOER P T, KROESE D P, MANNOR S, et al. A tutorial on the cross-entropy method. Annals of operations research, 2005, 134(1):19-67.

[41] ZHANG Z, SABUNCU M. Generalized cross entropy loss for training deep neural networks with noisy labels//Advances in neural information processing systems. Montréal, Canada. 2018: 8778-8788.

[42] KANWAL R, LIU K. A taylor expansion approach for solving integral equations. International journal of mathematical education in science and technology, 1989, 20(3):411-414.

[43] AGARAPAF. Deep learning using rectified linear units(relu). arXiv preprint arXiv:1803.08375, 2018.

[44] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep net-work training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015.

[45] LIN D, TALATHI S, ANNAPUREDDY S. Fixed point quantization of deep convolutional networks//Proceedings of international conference on machine learning. New York, USA. 2016: 2849-2858.

[46] KIM J H, ON K W, LIM W, et al. Hadamard product for low-rank bilinear pooling. arXiv preprint arXiv:1610.04325, 2016.

[47] BOYD S, BALAKRISHNAN V, FERON E, et al. Control system analysis and synthesis via linear matrix inequalities//IEEE American Control Conference. 1993: 2147-2154.

[48] DIANANDA P. A simple proof of the arithmetic mean-geometric mean inequality. Transactions of the American Mathematical Society. 1960, 67(10):1007.

[49] JIA X, SONG S, HE W, et al. Highly scalable deep learning training system with mixed-precision: Training imagenet in four minutes. arXiv preprint arXiv:1807.11205, 2018.

[50] DOMINGOS P. A unified bias-variance decomposition//Proceedings of international conference on machine learning. 2000:231-238.

[51] SALIMANS T, KINGMA D P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks//Advances in neural information processing systems. Barcelona, Spain. 2016: 901-909.

[52] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database//Proceedings of the IEEE conference on computer vision and pattern recognition. Miami, USA. 2009: 248-255.

[53] PASZKE A, GROSS S, MASSA F, et al. Pytorch: An imperative style, high-performance deep learning library//Advances in neural information processing systems. 2019: 8026-8037.

[54] LI X, HU X, YANG J. Spatial group-wise enhance: Improving semantic feature learning in convolutional networks. arXiv preprint arXiv:1905.09646, 2019.

[55] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions //Proceedings of the IEEE conference on computer vision and pattern recognition. Boston, USA. 2015: 1-9.

[56] SUTSKEVER I, MARTENS J, DAHL G, et al. On the importance of initialization and momentum in deep learning//Proceedings of international conference on machine learning. Atlanta, USA. 2013: 1139-1147.

[57] HE K, ZHANG X, REN S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification//Proceedings of the IEEE international conference on computer vision. Santiago, Chile. 2015: 1026-1034.

[58] LI X, CHEN S, XIA Y, et al. Understanding the disharmony between weight normalization family and weight decay: ϵ−shifted $l2$ regularizer . arXiv preprint, 2019.

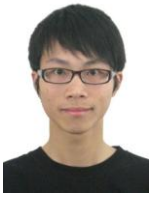
[59] LI X, CHEN S, HU X, et al. Understanding the disharmony between dropout and batch normalization by variance shift//Proceedings of the IEEE conference on computer vision and pattern recognition. Long Beach, USA. 2019:2682-2690.

[60] XIE S, GIRSHICK R, DOLLÁR P, et al. Aggregated residual transformations for deep neural networks//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, USA. 2017: 1492-1500.

[61] HU J, SHEN L, SUN G. Squeeze-and-excitation networks//Proceedings

of the IEEE conference on computer vision and pattern recognition. Salt Lake City, USA. 2018: 7132-7141.

[62] MONTGOMERY D C, PECK E A, VINING G G. Introduction to linear regression analysis: volume 821. Cambridge, UK: John Wiley & Sons, 2012.



Li Xiang, Ph.D. His research interests include computer vision, deep learning, and data mining.

Chen Shuo, Ph.D. His research interests include machine learning, and metric learning.

Yang Jian, Ph.D., professor. His research interests include machine learning, and pattern recognition.

Background

Weight Decay (WD) is a fundamental and important technique for preventing overfitting risks in both linear and deep neural systems. It is recognized as the best strategy which has the smallest complexity from a philosophical point of view. However, the underlying role of the specific form of WD still needs more exploration and understanding.

In this paper, based on the robustness and generalization theory, we derive a framework of the Generalization Bound Regularizer (GBR) which is directly related to the generalization ability of learning models, and theoretically explain that optimizing WD is essentially optimizing the exact or upper bound of the underlying GBR, thus providing a unified view for understanding WD. According to the unified framework, for multi-layer neural networks, we explore to further strengthen the connection between GBR and WD by holding the equal sign of the last scaling inequality, and propose the Equal Norm Constrain (ENC) condition to further constrain the increase of underlying GBR, which is proved very effective to improve the generalization ability of state-of-the-art deep neural networks.

This work was supported by the National Science Fund of China under Grant No. U1713208, Program for Changjiang Scholars. The full name of No. U1713208 is “Active Environment Recognition and Target Behavior Recognition Method of Service Robot”.