# Title:

Recent advances in deep learning theory

# Abstract:

深度学习通常被批评缺乏理论的依据，是实验驱动的领域。这篇文章就是回顾和组织了深度学习的理论。

本文分成六类：

1.基于复杂性和容量的方法，用于分析深度学习的泛化能力；

2.受贝叶斯推理部分启发的随机微分方程及其动态系统，用于模拟随机梯度下降及其变体，这些模型描述了深度学习的优化和泛化；

3.驱动动态系统轨迹的损失景观的几何结构；

4.从正反两方面探讨深度神经网络的过参数化问题；

5.网络架构中几种特殊结构的理论基础；

6.伦理和安全方面日益严重的关注以及它们与泛化能力之间的关系。

# Intro：

深度学习可以广义地定义为使用人工神经网络从经验中发现知识以进行预测或决策的算法家族。深度学习中的典型人工神经网络通常连接着一系列权重矩阵和一个非线性激活函数，这种网络具有相当大的参数规模。

深度学习缺乏理论依据这一黑盒特性给应用带来了未知的风险。这种无法意识的存在严重削弱了我们识别、管理和预防算法引发的灾难的能力，进一步严重影响了将最新进展应用于许多行业的信心，尤其是在安全关键领域，如自动驾驶、医疗诊断和药物发现等领域。这也对创新的深度学习算法设计的未来开发不利。

理论基础的一大部分是泛化性，这对模型处理未知情况尤为重要

人们需要找到一个足够小的模型来防止过拟合，只要该模型能够适应训练样本。然而，深度学习模型通常具有非常大的模型大小，有时甚至会使泛化边界比潜在损失函数的最大值还要大。此外，奥卡姆剃刀原则表明，泛化能力和模型大小之间存在正相关关系，但在深度学习中不再观察到这种关系。相反，更深更宽的网络通常具有更好的性能。深度学习出色的泛化能力和其极端的过参数化之间的不协调对于传统的基于复杂性学习理论来说就像一片“乌云”。

# A brief overview of the status quo

本文将相关的研究分为六类：

## Complexity and capacity-based method for analyzing the generalizability of deep learning.

传统的统计学习建立了基于假设空间复杂性的泛化误差上限，例如VC维度、Rademacher复杂度和覆盖数，这些界限依靠模型大小。然而，深度学习模型的巨大模型规模也使得泛化边界变得无效。因此，如果我们能够开发与规模无关的假设复杂性度量和泛化边界，那将是非常理想的。

解决办法是有效模型复杂度

## Stochastic differential equations (SDE) for modelling stochastic gradient descent (SGD) and its variants, which dominate the optimization algorithms in deep learning.

这些随机微分方程（SDEs）的动态系统决定了训练神经网络中权重的轨迹，而它们的稳定分布则表示已学习到网络的“有效”假设空间 “可以通过SGD找到”。通过SGD研究深度学习的泛化能力是简单直接的。此外，这一方法家族部分受到贝叶斯推理的启发。随机梯度方法和贝叶斯推理之间的相互作用将有助于推动这两个领域的发展。

## The geometrical structures of the highly complex empirical risk landscape that drive the trajectories of the dynamic systems.

理解损失表面将是建立深度学习理论基础的关键步骤。

高度复杂的经验风险景观的几何结构驱动了动态系统的轨迹。损失景观的几何形状在驱动随机微分方程的轨迹方面起着重要作用：（1）损失的导数是随机微分方程的组成部分；（2）损失将作为随机微分方程的边界条件。

通常，在“正则化”问题（使病态问题可解的方法）中，学习能力和优化能力通常得到保证。“正则化”可以用各种术语来描述，包括凸性、Lipschitz连续性和可微性。然而，这些因素在深度学习中不再得到保证，至少不明显。神经网络通常由大量非线性激活函数组成。激活函数中的非线性使得损失表面非常不平滑和非凸。已建立的凸优化保证变得无效。损失表面的复杂性长期以来一直是触及损失景观几何形状甚至深度学习理论的难题。然而，损失表面的复杂几何形状恰好描述了深度学习的行为。通过其损失表面理解深度学习正是一条“高速公路”。

## The roles of over-parameterization of deep neural networks.

过参数化通常被视为通过基于复杂度的方法为深度学习建立有意义的泛化边界的主要障碍。然而，最近的研究表明，过参数化将在塑造深度学习的损失表面方面发挥重要作用 - 使损失表面更加平滑甚至“类似”凸性。此外，许多工作已经证明，极端过参数化的神经网络等效于一些更简单的模型，例如高斯核（softmax回归）。

## Theoretical foundations for several special structures in network architectures.

回顾了CNN，RNN，满足置换不变性（permutation-invariant）与置换同变性（permutation-equivariant）的通用网络结构（Deep Sets：https://zhuanlan.zhihu.com/p/612736617）

## The intensive concerns in ethics and security and their relationships with deep learning theory

深度学习已经在越来越多的应用领域得到部署。其中一些涉及高度私人的个人数据，例如手机上的图像和视频、健康数据和最终记录。其他一些场景可能需要深度学习来做出高度敏感的决策，例如抵押贷款批准、大学录取和信用评估。此外，深度学习模型已经显示出容易受到对抗性示例的攻击。如何保护深度学习系统免受与隐私保护、公平性和对抗性攻击相关的漏洞的侵害具有重要意义。

# Stochastic gradient descent: An implicit regularization：

随机梯度方法（SGM），包括随机梯度下降（SGD）及其变体，是深度学习中的主要优化方法。它们只探索了假设空间的一小部分。这在算法和数据相关的方式下对深度学习模型进行了隐式正则化，从而控制了有效模型容量。因此，通过SGM接近算法和数据相关的界限可能克服基于VC维数或Rademacher复杂度的传统统计学习理论。

## Stochastic gradient methods

Gradient Descent（GD）是优化预期风险的工具，预期风险的梯度可表示如下：

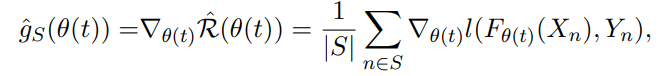


-空心E为期望，l为loss函数，F(X)为神经网络的输出，倒三角为Nabla算子，值为梯度

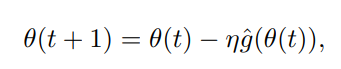
更新参数方程式为：

η为学习率，t为迭代次数

SGD：S为mini batch的索引，所有的索引都是独立同分布的（i.i.d），N为训练样本大小，在每个mini-batch下，SGD的迭代定义如下：



相当于将每个小批量的梯度取平均，接着更新参数：



本文后面将用以下式子简要表示：



SGD是将初始状态Q0经过i步（状态Qi）收敛到Q。

## Generalization bounds on convex loss surfaces

深度神经网络的loss函数表面是高度非凸的，尽管对于凸函数的梯度下降这方面有一些研究，但是不可用

## Generalization bounds on non-convex loss surface

相关工作包括通过算法稳定性和PAC-贝叶斯理论获得泛化界限。

Bousquet和Elisseeff 建议使用算法稳定性来测量当训练样本集受到干扰时输出假设的稳定性。

算法稳定性：

**定义4：对于相邻的样本组（S，S’），有以下不等式：**



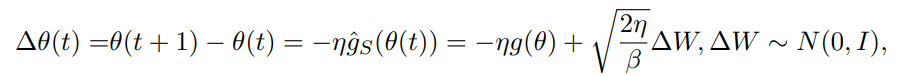
-Z是任意示例，A是机器学习算法，常数β称为算法A的uniform stability（一致稳定性）。

一个算法对训练数据中的干扰不敏感，那它自然具有良好的泛化能力。依据这种直觉，基于算法稳定性已经证明一些**泛化界限**。例如，基于均匀稳定性，Bousquet和Elisseeff 证明了以下关于期望泛化的定理。

**定理5：算法A是β-一致稳定，那么：**

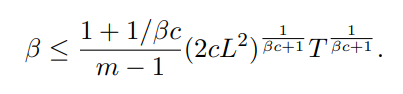


SGM通常由等式（3）和（4）来描述。随机梯度在权重更新中引入了梯度噪声。当梯度噪声由高斯分布建模时，SGM可以简化为随机梯度Langevin动力学（SGLD），如下所示：



-反温度常数β>0

定理6：假设损失函数是-smmoth（epslion-smmoth）的（标签平滑：一种减少独热编码缺陷的正则化手段），满足L-Lipschitz（利普希茨连续条件），并且每个数据不大于1，SGM以逐步递增的步长运行，则SGM为均匀分布带以下式子：



上式证明了对于任意样本来说，损失函数l的权重是L-Lipschitz的常数，例：

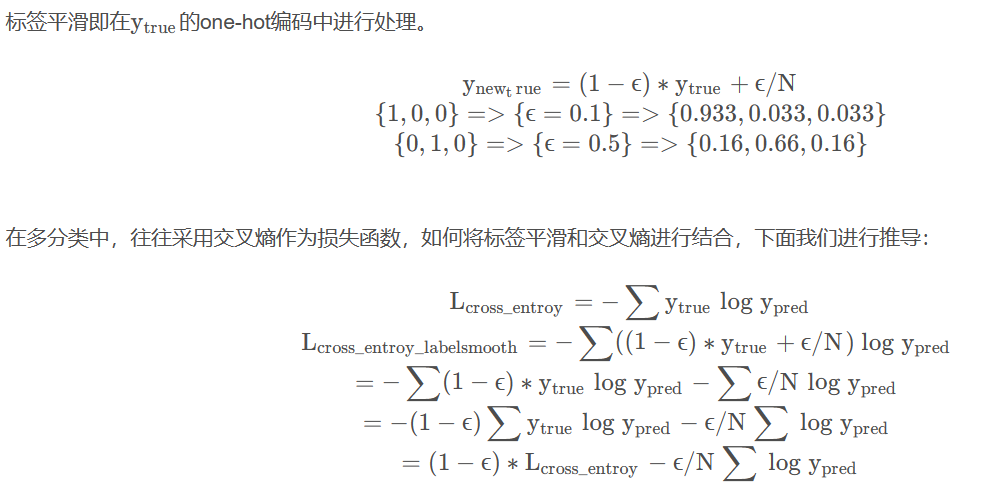


这里|| ||为向量二范数（欧几里得范数），为向量长度

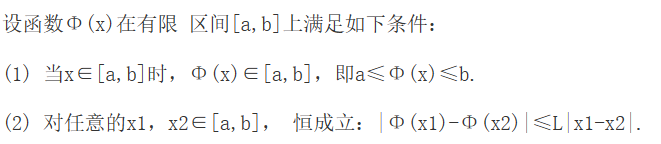
即稳定性可以通过权重的稳定性来衡量。因此，人们可以只分析随着时间t的增加，权重如何发散。作者们然后根据步长和迭代次数给出了泛化界限。

下面是关于标签平滑和利普西茨连续条件的解释：

**标签平滑：label-smmoth，标签平滑因子为ϵ ，**



**L-Lipschitz（利普希茨连续条件）：**

L称为利普西茨常数

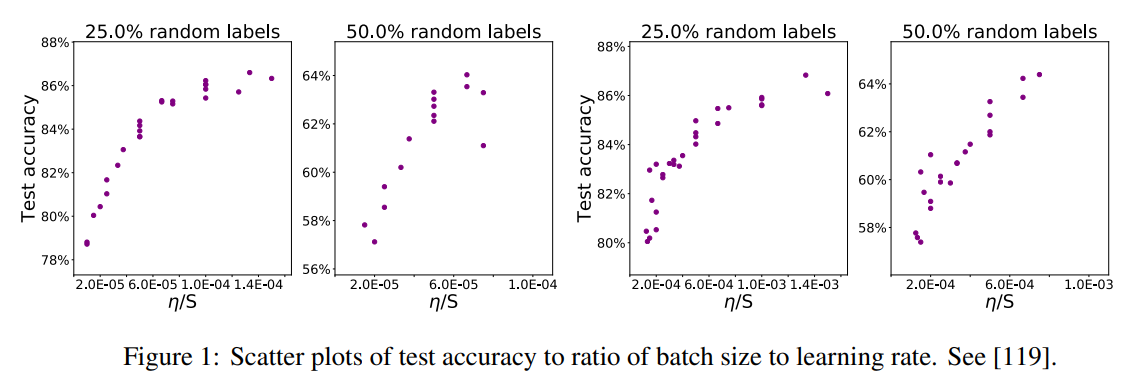
解释：一条曲线上任意两点连线的斜率的绝对值都有小于某一个数

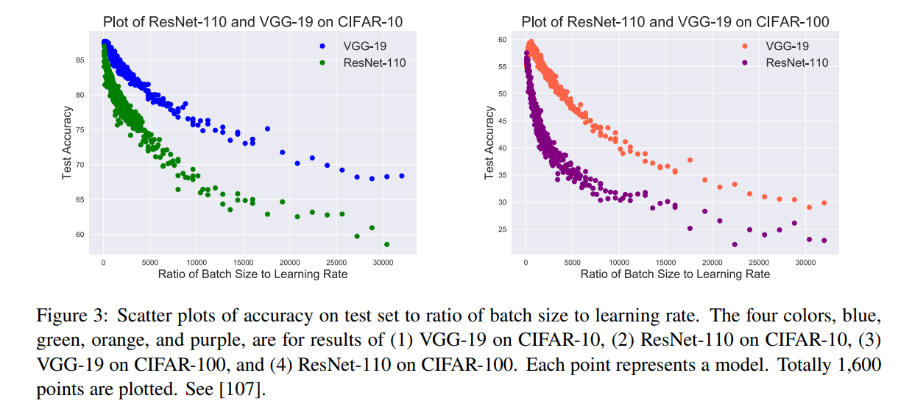
利普希茨连续条件（Lipschitz continuity）是以德国数学家鲁道夫·利普希茨命名，是一个比一致连续更强的光滑性条件。直观上，利普希茨连续函数限制了函数改变的速度，符合利普希茨条件的函数的斜率，必小于一个称为利普希茨常数的实数（该常数依函数而定）。

此外，我们可以轻松地基于这个定理获得泛化界限。（详细看论文）

## The role of learning rate and batch size in shaping generalizability

学习率和批量大小对泛化能力的影响

He等人在一些假设下证明了PAC-Bayesian泛化界限。该泛化界限还表明，泛化能力和学习率与批量大小的比率之间存在正相关关系。然后，He等人进行了实验，结果显示出明显的泛化能力与批量大小与学习率之比之间的相关性



测试集上准确率与批量大小与学习率之比的散点图。每个点代表一个模型。总共绘制了1600个点。

# Taking the high way: Study on loss surface

深度神经网络的损失表面非常非凸甚至非平滑。这种非凸性和非平滑性使得优化和泛化性质的分析变得极其困难。一个自然的想法是绕过几何性质，然后接近理论解释。一些论文认为，这些“令人生畏”的几何性质正是塑造深度神经网络特性的主要因素，也是解释深度学习的关键。

## Does spurious local minimum exist?

如果所有局部极小值都是全局最优的，那么我们就不需要担心深度学习的优化问题。那么，这是真的吗？

1. Linear networks have no spurious local minima

线性网络没有伪局部最小值

线性神经网络指的是其激活函数都是线性函数的神经网络。对于线性神经网络，损失表面没有任何虚假的局部极小值：所有局部极小值都是同样好的；它们都是全局最小值。

2.Nonlinear activations bring infinite spurious local minima

非线性的激活函数带来无限的伪局部最小值

## Geometric structure of loss surface.损失函数的几何结构

Linear partition of the loss surface.

分段线性激活函数中的非线性将损失表面划分为多个平滑且多线性的开放区域。具体来说，激活函数中的每个非线性点都会在“细胞”之间创建一组非可微分边界，而激活函数的线性部分则对应于平滑且多线性的内部。

相当于一个个折线，每个“折”的地方是非线性，不可微的边界，线性的部分像一个个细胞，当激活函数为线性时，细胞收敛于1

# Reflecting the role of over-parameterization: Is it only harmful?

## Neural tangent kernel

无限宽度的浅层神经网络等价于高斯过程。

在梯度下降训练过程中，任意深度的无限宽度神经网络都表现为线性模型

## Influence of over-parameterization on the loss surface.

过度参数化会显著改变神经网络的损失面。

当网络规模增加时，具有较小或正常规模的神经网络中的虚假局部极小值的比例迅速减少。如果神经网络充分过度参数化，则全局极小值相互连接并集中在唯一的山谷中。

## Influence of over-parameterization on the optimization.

过度参数化可能也是深度学习成功的原因之一

研究表明，过度参数化有助于确保基于梯度的优化算法的优化性能。Du等人证明，对于两层神经网络，只要网络宽度足够大，在二次损失下，梯度下降以线性收敛率收敛到全局局部极小值。

对于超过一层隐藏层的情况，Allen-Zhu等人、Du等人和Zou等人同时证明了SGD可以在多项式时间内收敛到过度参数化的深度神经网络的全局最优解，这些研究在不同的网络结构和训练数据假设上略有不同。

## Influence of over-parameterization on the generalization and learnability.

作者们证明了当使用通用梯度下降算法进行训练并且所有权重都是随机初始化时，足够宽的神经网络可以学习任何低阶多项式函数。Brutzkus等人也在一些假设下给出了两层神经网络的泛化保证。Li和Yuan证明，对于多分类问题，如果过度参数化的单隐藏层ReLU网络使用SGD从随机初始值开始训练，那么它们的泛化误差可以得到保证。

## Influence of the network depth on the generalizability.

网络深度对泛化能力的影响：更深的神经网络有更好的泛化能力。这是神经网络过度参数化与其优秀泛化能力之间矛盾的一个重要原因。

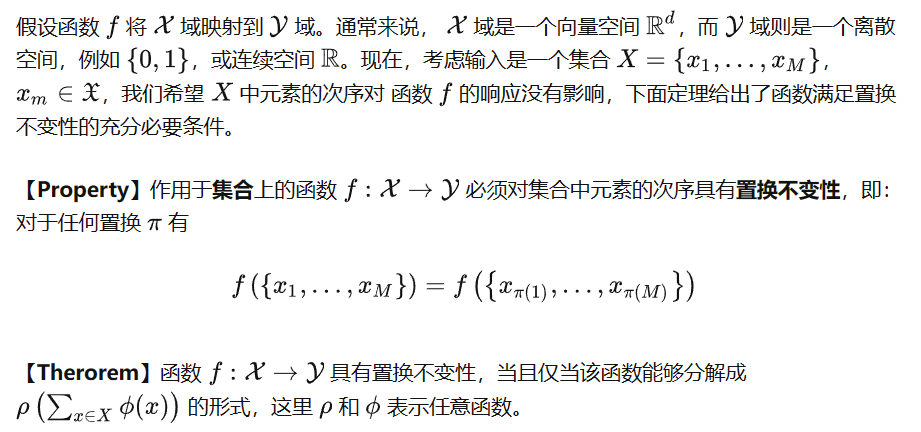
# Theoretical foundations for specific architectures

接下来讨论一些特定网络的理论基础

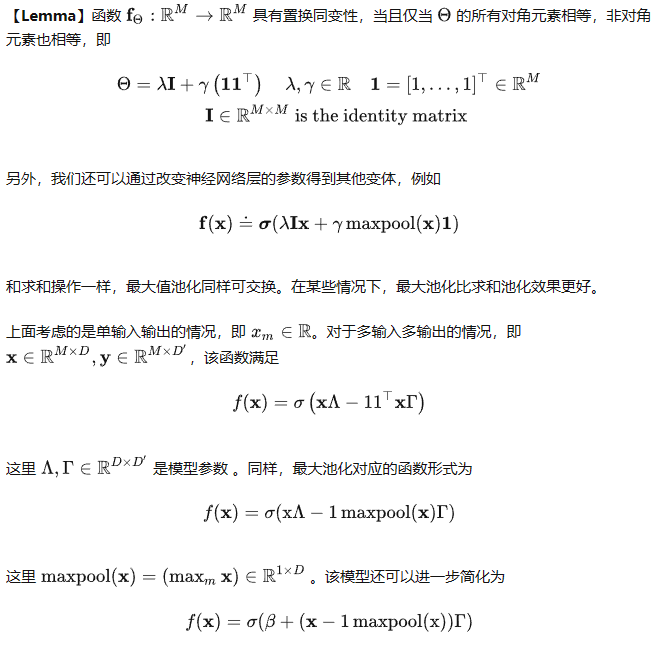
## Networks for permutation invariant/equivariant functions

在某些问题中，问题的输入和输出是满足置换不变性（permutation-invariant）的集合（sets），而非固定维度的向量（vectors）。针对这一特殊的问题设计了一种通用的网络结构——deep sets，该结构具有置换不变性（permutation-invariant）与置换同变性（permutation-equivariant）两大特点，可用于处理输入输出是集合的问题。

1.置换不变性（permutation-invariant）



2.置换同变性（permutation-equivariant）



# Looking beyond: Rising concerns of ethics and security

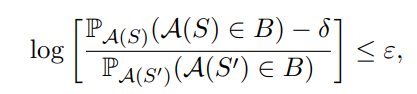
理解各种问题的理论基础也是很重要的

## Privacy preservation

**Differential privacy**

衡量算法隐私保护能力的主要指标是差分隐私。我们称一个学习算法为(ε， δ)-差分私有的，当训练数据受到一些小的干扰时，输出假设的变化受到如下限制

privacy loss.



S和S’是相邻的样本对

**Generalization-privacy relationship**

一些工作已经表明，由差分隐私衡量的隐私保护能力几乎等同于泛化能力。论文列举了一系列泛化界限。

## Algorithmic fairness

深度学习已被部署在许多关键决策领域，如抵押贷款批准、信用卡评估、大学录取、员工选择和累犯预测。

然而，在这些领域中，已经观察到长期存在的基于人们的背景，包括种族、性别、国籍、年龄和宗教系统性歧视。此外，少数族裔背景的人在历史数据中机构代表性不足。与此同时，越来越多的报告揭示出，数据中的歧视是继承的，甚至在基于机器学习算法的模型学习中得到强化。由于黑箱性质，公平性问题尤为严重。例如，据报道，面部识别“如果你是一个白人男性，那么就是准确的”；一个用于累犯预测的警察系统，即替代制裁的罪犯管理档案（COMPAS），往往将少数民族标记为再犯罪的可能性显著更高。减轻公平风险是及时且重要的。

为了解决公平性问题，提出了许多方法。它们可以大致分为三个部分：（1）预处理方法，旨在在学习算法处理之前“纠正”数据；（2）后处理方法，调整学习系统的输出以消除潜在的歧视；以及（3）内处理方法在训练过程中适应约束或正则化器来强制实现公平性。

## Adversarial robustness

Szegedy等人发现神经网络容易受到对抗性样本的攻击，这些对抗性样本与真实数据略有不同，但很可能被神经网络错误分类。自那时以来，已经提出了许多方法来通过对抗性样本攻击神经网络或防御对抗性攻击

**1.Generalizability in the presence of adversarial examples.**

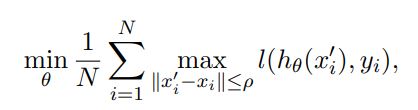
在存在对抗性样本的情况下的泛化能力：

Schmidt等人首次研究了存在对抗性样本时的泛化能力。他们定义了一个新的概念，即对抗性鲁棒泛化，用于衡量当数据中存在对抗性样本时的泛化能力。相比之下，没有对抗性样本时的泛化能力由标准泛化评估。他们证明，当数据来自某些特定分布且对抗性样本遵循l∞范数时，为了获得相同的泛化能力，对抗性鲁棒泛化需要更高的样本量。

**2.Generalizability of adversarial training.**

对抗性训练的泛化能力：

在防御方法中，对抗性训练可以显著提高深度神经网络对对抗性样本的对抗鲁棒性。具体而言，对抗性训练可以表示为解决以下最小最大损失问题：



-hθ是由θ参数化的假设，N是训练样本大小，xi是一个特征，yi是相应的标签，l是损失函数

在提高对抗鲁棒性时，对抗性训练会提高泛化能力。

# Summary and discussion

本文回顾了深度学习理论的文献，从六个方面进行了综述：(1)基于复杂度和容量的方法用于分析深度学习的泛化能力；(2)随机微分方程及其动态系统用于建模随机梯度下降法（SGM），这些方法描述了由有限探索空间引入的隐式正则化；(3)损失景观的几何结构，它支配着动态系统的轨迹；(4)深度神经网络过参数化的积极和消极作用；(5)与前面部分平行，一些特殊网络架构的理论基础；(6)对伦理和安全性的关注以及它们与深度学习理论的关系。

许多工作给出了深度学习泛化误差的上界，结果看论文表1

这六个方向在建立深度学习的理论基础上是有前景的。然而，目前的“最先进技术”仍然远远不够；所有这些方向都面临着重大困难。在本文中，列出了未来工作的三个主要方向：

• 目前对复杂度的度量评估整个假设空间，但很难捕捉到训练算法将输出的“有效”假设空间的复杂性。因此，一个“局部”版本的复杂度将是有帮助的。

• 目前对SGM的模型通常忽略了一些流行的训练技术，如动量、自适应学习率、梯度裁剪、批量归一化等。一些工作已经在研究这些问题，但一个完整的“landscape”仍然缺失。此外，这些工作通常假设梯度噪声是从某些特定分布中抽取的，如高斯分布或α稳定分布。

• 损失表面的几何结构仍然不成熟。神经网络的极端复杂性使得简单的解释不可行。需要努力理解一些复杂的几何性质，如局部极小值的尖锐度/平坦度、根据其性能对局部极小值（谷）进行潜在分类以及来自不同类别的局部极小值的体积。