arXiv:2304.00320v1 [cs.LG] 1 Apr 2023

随机标签噪声下的随机梯度下降：双重随机模型与推理稳定器

Haoyi Xiong[1], Xuhong Li[1*]、BoyangYu[1],
Dongrui Wu[2], Zhanxing Zhu[3] 和 Dejing Dou[4]

[1]百度公司，北京，中国。[2]华中科技大学，中国武汉。[3]北京大学，中国北京。[4]BCG Greater China，中国北京。

*通讯作者。电子邮箱：lixuhong@baidu.com；贡献作者：xionghaoyi@baidu.com；boyangjlu1030@163.com；drwu@hust.edu.cn；zhanxing.zhu@gmail.com；dejingdou@gmail.com；

摘要

随机标签噪声（或观测噪声）广泛存在于实际的机器学习场景中。以往研究主要关注标签噪声对学习性能的影响，而我们的工作旨在研究在随机梯度下降（SGD）的小批量采样设置下、假设标签噪声无偏时，标签噪声所产生的隐式正则化效应。具体而言，我们分析了带有无偏标签噪声的二次损失下SGD的学习动力学，并将SGD的动力学建模为具有两个扩散项的随机微分方程（SDE）（即一个 *Doubly StochasticModel*）。尽管第一个扩散项由对（无标签噪声的）损失梯度进行小批量采样所引起，正如许多其他关于SGD的研究 [1, 2], 我们的模型研究了SGD动力学中的第二个噪声项，并将其视为一种隐式正则化器；该噪声项由对标签噪声进行小批量采样所引起。我们的理论分析发现，这种隐式正则化器会偏好某些收敛点，从而使模型输出能够抵抗参数扰动（即 *inference stability*）。尽管类似现象已在 [3], 中得到研究，但我们的工作

不假设随机梯度下降（SGD）服从类奥恩斯坦–乌伦贝克过程，并得出了更具普适性的结果，同时证明了近似收敛。为验证我们的分析，我们设计了两组实证研究，分析在深度神经网络训练和线性回归中使用无偏随机标签噪声时，随机梯度下降的隐式正则化器。第一项实验研究了深度学习中的 *the noisy self-distillation tricks*，其中学生网络使用经过充分训练的教师网络的输出进行训练，并加入无偏随机标签噪声。实验表明，标签噪声引起的隐式正则化器倾向于选择推理稳定性更高的模型。我们还对含无偏标签噪声的随机梯度下降线性回归进行了实验，绘制了每一步所学参数的轨迹，并直观展示了隐式正则化的影响。结果支持了我们的理论发现。

1 引言

随机梯度下降（SGD）已被广泛用作训练大规模数据集上的深度神经网络的有效方法 [4]。尽管小批量采样策略最初是为了降低每次迭代的计算成本而提出的，但人们认为它引入了隐式正则化器，可防止学习过程收敛到泛化性能较差的局部极小值 [1, 5–8]。为解释这种隐式正则化，可以将随机梯度下降建模为带有由小批量采样引起的梯度噪声的梯度下降（GD） [9]。理论研究 [10–13]和实证研究 [1, 7, 8]均已证明，这种隐式正则化或梯度噪声有望提升学习的泛化性能 [1, 12, 13] 。总之，梯度噪声使随机梯度下降避免收敛到泛化能力较差的尖锐局部极小值 [1, 12, 13] ，并会使学习结果趋向于平坦极小值 [14]。

本文旨在研究 SGD 小批量采样下，随机标签噪声对隐式正则化的影响。为简化研究，我们将训练数据集设为向量集合 $\mathcal{D} = \{x_1, x_2, x_3, \ldots, x_N\}$。每个向量 $x_i \in \mathcal{D}$ 对应的标签 $\tilde{y}_i$ 是由真实神经网络 $f^*(x)$ 产生的带噪响应，使得

$$\tilde{y}_i = y_i + \varepsilon_i,\ y_i = f^*(x_i),\ \text{and}\ \mathbb{E}[\varepsilon_i] = 0,\ \text{var}[\varepsilon_i] = \sigma^2\ , \tag{1}$$

其中，标签噪声 ε_i 被假设为独立的零均值**随机变量**。在我们的工作中，随机标签噪声可以是以下两种情况之一：(1) 在训练步骤之前从概率分布中抽取（但会通过 SGD 的小批量采样重新采样），或 (2) 由每次训练迭代中的随机变量实现 [15]。因此，学习的目标是在 $f(x, \theta)$ 中估计 θ，以近似 $f^*(x)$，

符号	定义和方程	
x_i，以及 $y_i = f^*(x_i)$	该 i^{th} 数据点和真实标签	(1)
$\tilde{y}_i$，以及 ε_i	噪声标签 i^{th} 和标签噪声。	(1)
$f(x,\ \theta)$	参数为 θ 、输入为 x 的神经网络输出。	(2)
$\widehat{\theta}$	神经网络的参数估计器。	(2)
L_i^*	$L_i(\theta) = \frac{1}{2}(f(x_i,\ \theta) - y_i)^2$ 基于无噪声样本的损失。	(2)
$\tilde{L}_i$	$\tilde{L}_i(\theta) = \frac{1}{2}(f(x_i,\ \theta) - \tilde{y}_i)^2$ 基于含噪声样本的损失。	(2)
	不对标签噪声作任何假设的随机梯度下降	
B_k	随机梯度下降第 k^{th} 步抽取的样本小批量	(3)
$b = \lvert B_k \rvert$	固定批量大小为 B_k	(3)
θ_k	随机梯度下降的第 k^{th} 步。	(3)
V_k	由损失梯度的小批量采样引起的随机梯度下降噪声。	(5)
η	随机梯度下降的学习率。	(3)
$\Theta(t)$	随机梯度下降的连续时间动力学。	(6)
z_k	标准高斯分布的随机向量。	(9)
$W(t)$	随时间演变的布朗运动。	(6)
$\bar{\theta}_k$	离散时间近似的第 k^{th} 步： $\Theta(t)$。	(7)
	带无偏标签噪声的随机梯度下降（ULN）	
θ_k^{ULN}	带无偏标签噪声的随机梯度下降的第 k^{th} 步。	(8)
ξ_k^*	通过小批量采样真实损失梯度产生的随机梯度下降噪声。	(8)
ξ_k^{ULN}	通过小批量采样无偏标签噪声产生的随机梯度下降噪声。	(8)
Σ_N^{SGD}	真实损失梯度的协方差矩阵。	(15),(16),(17)
Σ_N^{ULN}	基于无偏标签噪声的协方差矩阵。	(15),(16),(17)
$\Theta^{\mathrm{ULN}}(t)$	连续时间双重随机模型。	(16)
$\bar{\theta}_k^{\mathrm{ULN}}$	离散时间双随机模型的第 k^{th} 步。	(17)
$W_1(t)$，以及 $W_2(t)$	随时间演化的两个相互独立的布朗运动。	(16)
z_k，以及 z_k'	两个相互独立的标准高斯随机向量。	(9),(17).
$\Theta^{\mathrm{LNL}}(t)$	Label-NoiseLess 设置下的连续时间动力学。	(20)

表 1：关键符号及其定义

使得

$$\widehat{\theta} \leftarrow \underset{\forall\theta\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{N}\sum_{i=1}^{N} \tilde{L}_i(\theta) := \frac{1}{2N}\sum_{i=1}^{N} (f(x_i,\theta) - \tilde{y}_i)^2 \right\}. \tag{2}$$

注意，本文将 $L_i^*(\theta) = \frac{1}{2}(f(x_i,\theta) - y_i)^2$ 表示基于无噪声样本的损失。受 [1, 14], 启发，我们研究无偏标签噪声 ε_i $(1 \le i \le N)$ 如何从学习动力学的视角影响随机梯度下降在小批量采样下对 θ 可能解的“选择” [16] [12, 17, 18]。本文所用符号请参见表1。

1.1 背景：随机梯度下降动力学与隐式正则化

为分析求解公式（2）中问题的随机梯度下降算法，我们遵循 [17] 中的设置，并将随机梯度下降视为一种算法：在第 k^{th} 次迭代中，给定估计值 θ_k，它会从训练数据集中随机选取一个长度为 b的样本子集，即 $B_k \subset \mathcal{D}$，并估计小批量随机梯度

4 双重随机模型

$\frac{1}{b}$ -$\sum_{\forall x_i \in B_k} \nabla L_i$（$\theta_k$），然后根据 θ_k更新 θ_{k+1} 的估计值，具体如下 w

$$\theta_{k+1} \leftarrow \left(\theta_k - \frac{\eta}{|B_k|} \sum_{\forall x_i \in B_k} \nabla \tilde{L}_i(\theta_k) \right) , \tag{3}$$

其中， η 表示随机梯度下降的步长。此外，我们可以将小批量采样得到的损失梯度分解为全批量损失梯度与噪声之和，使得

$$\begin{aligned} & \frac{\eta}{|B_k|} \sum_{\forall x_i \in B_k} \nabla \tilde{L}_i(\theta_k) \\ =& \frac{\eta}{N} \sum_{\forall x_i \in \mathcal{D}} \nabla \tilde{L}_i(\theta_k) - \left(\frac{\eta}{N} \sum_{\forall x_i \in \mathcal{D}} \nabla \tilde{L}_i(\theta_k) - \frac{\eta}{|B_k|} \sum_{\forall x_i \in B_k} \nabla \tilde{L}_i(\theta_k) \right) \\ =& \frac{\eta}{N} \sum_{\forall x_i \in \mathcal{D}} \nabla \tilde{L}_i(\theta_k) - \sqrt{\eta} V_k(\theta_k) , \end{aligned} \tag{4}$$

其中， η 表示随机梯度下降的步长， $V_k(\theta_k)$表示由小批量采样引起的随机梯度噪声项。随着批量大小增加，噪声将趋于零，如下所示：

$$V_k(\theta_k) = \sqrt{\eta} \left(\frac{1}{N} \sum_{\forall x_i \in \mathcal{D}} \nabla \tilde{L}_i(\theta_k) - \frac{1}{|B_k|} \sum_{\forall x_i \in B_k} \nabla \tilde{L}_i(\theta_k) \right) \to \mathbf{0}_d, \quad \text{as} \quad B \to N . \tag{5}$$

当 $\mathbf{d}t = \eta \to 0$ 且批量大小 $b = |B_k|$为常数时，随机梯度下降算法将扩散为连续时间动力学 $\Theta(t)$，其随机微分方程（随机微分方程）满足弱收敛 [17, 19], ，如下所示：

$$\mathbf{d}\Theta = -\frac{1}{N} \sum_{i=1}^{N} \nabla \tilde{L}_i(\Theta) \mathbf{d}t + \left(\frac{\eta}{b} \tilde{\Sigma}_N^{\mathrm{SGD}}(\Theta) \right)^{\frac{1}{2}} \mathbf{d}W(t) \tag{6}$$

其中， $W(t)$ 是 $\mathbb{R}^d$ 中的标准布朗运动，并将 $\Sigma_N^{\mathrm{SGD}}(\Theta)$ 定义为 $1 \le i \le N$ 的损失梯度 $\nabla L_i(\Theta)$ 的样本协方差矩阵。请注意，关于上述连续时间近似的详细推导及相关假设，请参阅 [17]。我们遵循 [17] ，不对 $\tilde{\Sigma}_N^{\mathrm{SGD}}(\Theta)$ 作低秩假设。通过欧拉离散化， [11, 17], 可以将随机梯度下降近似为 θ_k ，使得

$$\begin{aligned} & \bar{\theta}_{k+1} \leftarrow \bar{\theta}_k - \frac{\eta}{N} \sum_{\forall x_i \in \mathcal{D}} \nabla \tilde{L}_i(\bar{\theta}_k) + \sqrt{\eta} \xi_k(\bar{\theta}_k), \text{ and} \\ & \xi_k(\bar{\theta}_k) = \left(\frac{\eta}{b} \tilde{\Sigma}_N^{\mathrm{SGD}}(\bar{\theta}_k) \right)^{\frac{1}{2}} z_k, \ z_k \sim \mathcal{N}(0, \mathbf{I}_d) . \end{aligned} \tag{7}$$

随机梯度下降的隐式正则化器为 $\xi_k(\bar{\theta}_k) = \left(\frac{\eta}{b}\tilde{\Sigma}_N^{\mathrm{SGD}}(\bar{\theta}_k)\right)^{\frac{1}{2}} z_k$，它依赖于数据，并受学习率 η 和批量大小 B [20]控制。 [10–12]讨论了用于变分推断的随机梯度下降，并推动了其在采样器中的新型应用 [21, 22]。为理解其对泛化性能的影响， [1, 20]研究了从尖锐局部极小值中逃逸的行为 [8]以及向平坦极小值的收敛。 [23] 从信息论视角探究了随机梯度下降找到平坦局部极小值的方式，并提出了一种新型正则化器以提升性能。最后， [24] 研究了对线性深度神经网络的正则化效应，我们之前的工作 [18] 提出了新的乘性噪声来解释随机梯度下降，并获得了更强的理论性质。

1.2 我们的贡献

在本研究中，我们假设无偏随机标签噪声 $\varepsilon_i\,(1 \le i \le N)$ 与随机梯度下降的小批量采样器相互独立。当随机标签噪声在训练过程之前已从概率分布中抽取后，随机梯度下降会重新采样标签噪声，并通过重新采样带标签噪声的数据以及对随机小批量中的带标签噪声损失梯度求平均，在无标签噪声损失产生的随机梯度噪声之外，生成一种新型的依赖数据的噪声 [25, 26]。

我们的分析表明，在温和条件下，使用带标签噪声的损失梯度时，随机梯度下降可能通过重新采样带标签噪声的样本 [26]，或在迭代过程中动态向标签添加噪声 [15]，引入额外的依赖数据的噪声项，从而补充无标签噪声损失的随机梯度噪声 [17, 18]。我们将此类噪声视为无偏标签噪声所导致的隐式正则化，并将此类噪声的影响解释为学习过程的解选择器。更具体地说，本研究做出了如下独特贡献。

1.2.1 双重随机模型

我们回顾了预备知识 [12, 17, 18, 27]，并在 [17] 中扩展了分析框架，以将无偏标签噪声的影响解释为叠加在随机梯度下降连续时间动力学之上的一种额外隐式正则化器。通过对带标签噪声的随机梯度下降的连续时间动力学进行离散化，我们将学习动力学的离散时间近似记为 θ_k^{ULN}，其中 $k = 1, 2, \ldots$，表示为

$$\theta_{k+1}^{\mathrm{ULN}} \leftarrow \theta_k^{\mathrm{ULN}} - \frac{\eta}{N}\sum_{i=1}^{N}\nabla L_i^*(\theta_k^{\mathrm{ULN}}) + \sqrt{\eta}\xi_k^*(\theta_k^{\mathrm{ULN}}) + \sqrt{\eta}\xi_k^{\mathrm{ULN}}(\theta_k^{\mathrm{ULN}}), \quad (8)$$

其中， $L_i^*(\theta) = (f(x_i, \theta) - f^*(x_i))^2$ 表示带有样本 x_i 及其真实（无噪声）标签 y_i 的无标签噪声损失函数，噪声项 $\xi_k^*(\theta)$表示无标签噪声损失函数 $L_i^*(\theta)$的随机梯度噪声 [17] ，于是我们可以得到由无偏标签噪声引起的新隐式正则化器

（ULN），对于 $\forall\theta \in \mathbb{R}^d$，可近似如下

$$\xi_k^{\mathrm{ULN}}(\theta) \approx \left(\frac{\eta\sigma^2}{bN}\sum_{i=1}^{N}\nabla_\theta f(x_i,\theta)\nabla_\theta f(x_i,\theta)^\top\right)^{\frac{1}{2}} z_k, \text{ and } z_k \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d) , \quad (9)$$

其中， z_k 表示从标准高斯分布中抽取的随机噪声向量， θ_k 表示第 k^{th} 次迭代中网络的参数，$(\cdot)^{1/2}$表示该矩阵的 Chelosky 分解， $\nabla_\theta f(x_i, \theta) = \partial f(x_i, \theta)/\partial\theta$表示样本 x_i 的神经网络输出关于参数 θ_k的梯度， B 和 η 分别定义为随机梯度下降的批量大小和学习率。显然，这种隐式正则化器的强度由 σ^2、B 和 η控制。

结果总结

第 3节将带有无偏随机标签噪声的随机梯度下降算法表述为基于两个噪声项的随机动力学（**命题** 1），从随机梯度下降算法推导出连续时间和离散时间双重随机模型（**定义** 1 和 2），并给出近似误差界（**命题** 2）。两个命题的证明见 A 和B

1.2.2 推理稳定器作为隐式正则化器

无偏随机标签噪声的正则化效应应当

$$\mathbb{E}_{z_k}\left\|\xi_k^{\mathrm{ULN}}(\theta_k)\right\|_2^2 \approx \frac{\eta\sigma^2}{bN}\sum_{i=1}^{N}\|\nabla_\theta f(x_i,\theta_k)\|_2^2 = \frac{\eta\sigma^2}{bN}\sum_{i=1}^{N}\left\|\frac{\partial}{\partial\theta}f(x_i,\theta_k)\right\|_2^2 , \quad (10)$$

其中， $\nabla_\theta f(x, \theta)$表示 f关于 θ 的梯度，其效应受批量大小 B 和标签噪声的方差 σ^2控制。通过假设深度学习算法由类奥恩斯坦–乌伦贝克过程 [3], 驱动，也能得到类似结果；而我们的工作不依赖这一假设，而是完全基于我们提出的双重随机模型。

结果总结

第 4 节分析了无偏随机标签噪声对随机梯度下降的隐式正则化效应，并得出结论：在动态过程中，隐式正则化器是神经网络梯度范数 $\frac{1}{N}\sum_{i=1}^{N}\|\nabla_\theta f(x_i, \theta)\|_2^2$的控制器（**命题** 3）。随后，我们提出 **备注** 2、3和 4，以刻画采用无偏标签噪声时随机梯度下降的行为：（1）由于隐式正则化器驱动的扰动更大，随机梯度下降会逃离梯度范数较高的局部极小值；（2）隐式正则化效应的强度由学习率 η 和批量大小 b控制；（3）由于较低的神经网络梯度范数通常对应较平坦的损失景观，因此可以通过添加并控制无偏标签噪声来调整随机梯度下降的性能。

为验证我们的三点论述，第 5 节介绍了基于 带有无偏标签噪声的自蒸馏 [28, 29]深度神经网络的实验。我们表明，在教师-学生训练设置下，经过充分训练的模型可以通过从自身的噪声输出中学习（即在逻辑值输出中添加无偏标签噪声），逃离局部极小值，并收敛到一个新的点，在该点上神经网络梯度范数更低，泛化性能更好。第 6 节介绍了基于含无偏标签噪声的基于SGD的线性回归的实验，以直观展示隐式正则化效应，以及这些效应与学习率和批量大小之间的联系。可视化结果表明，含无偏标签噪声的基于SGD的线性回归会收敛到一个类似高斯的分布，该分布以线性回归的解为中心，其协方差由数据样本的协方差以及学习率和批量大小共同决定。实验结果支持了我们的理论。

2 相关工作

普通最小二乘法（***OLS***）的随机梯度下降隐式正则化 *)*

该领域最新且最相关的研究是 [27, 30], ，该研究由同一组作者开展，探讨了普通最小二乘法（OLS）中梯度下降和随机梯度下降的隐式正则化。他们研究了一种作用于参数的类似 ℓ_2 范数的隐式正则化器，它将 OLS 正则化为带有递减惩罚项的岭估计量。在此之前，F. Bach 及其团队已在 [31–33]中研究了有噪声和无噪声两种设定下，采用 OLS 和正则化估计量进行线性回归时，基于梯度的解的收敛性。

朗之万动力学与梯度噪声

出于类似的研究目的， [10–12] 从贝叶斯推断和变分推断的角度研究了随机梯度下降（SGD）的极限行为（或动力学的稳态）。他们还推动了随机梯度 MCMC 采样器的新应用 [21, 22]。通过将近收敛区域中的 $\Sigma_N^{\mathrm{SGD}}(\theta)$ 与损失函数海森矩阵 $1/N\nabla^2 L_i(\theta)$ 联系起来， [1] 研究了逃离尖锐局部极小值的行为，而 [8] 则在大批量训练设定下讨论了这一问题。此外， [20] 探讨了学习率和批量大小如何影响泛化性能以及优化结果的平坦性。最后， [24]研究了线性神经网络中的隐式正则化，并且[18] 提出了一种新的乘性噪声模型，用以解释具有更强理论性质的梯度噪声。

自蒸馏与噪声学生

自蒸馏 [28, 29, 34, 35] 已被视为进一步提升训练良好模型泛化性能的有效方法。这类策略以训练良好的模型作为教师模型，从而实现知识蒸馏，并且

可选择在学生模型的训练过程中加入噪声（例如随机失活、随机深度、标签平滑，或潜在的标签噪声）。

相关工作的讨论
与上述工作相比，我们在上述三个类别中仍有所贡献。首先，本文刻画了标签噪声对随机梯度下降动力学的隐式正则化效应。与 [27, 30] 研究线性回归的工作相比，我们提出的双重随机模型可用于解释标签噪声下随机梯度下降在非线性神经网络中的学习动力学。即使从线性回归的角度来看 [27, 30, 33]，我们也精确测量了有无标签噪声时随机梯度下降动力学之间的差异，并通过一个数值模拟的新示例直观展示了隐式正则化效应。

与 [29, 36],相比，我们的分析强调了标签噪声所导致的隐式正则化项在模型选择中的作用，并会选择推断稳定性较高的模型。 [37] 是与我们最相关的工作，作者通过神经切线核（NTK） [38] 近似研究了标签噪声下的梯度下降提前停止。我们的工作分析了随机梯度下降，且不依赖 NTK 等近似假设。

除神经切线核假设外， [3] 还假设深度学习算法由类似奥恩斯坦-乌伦贝克（OU）的过程驱动，并得出了与推理稳定器（我们研究的第三项结果）类似的结果；而我们的工作则通过提出双重随机模型 作出贡献，并以不同的方式得出结论。我们还首次基于常用的深度神经网络架构和基准数据集，提供了实证结果和证据，以展示无偏标签噪声在真实场景中造成的隐式正则化器的影响。请注意，我们此前的一份稿件 [39] 已发布在 OpenReview 上并开放讨论，外部审稿人提出了他们的顾虑——部分结果已在 [3] 中研究，而我们并未以强形式（例如定理或证明）给出这些结果。因此，本工作将主要贡献从标签噪声的隐式正则化转向双重随机模型 及其近似误差界和证明。无需假设存在 OU 过程，即可直接通过双重随机模型估计隐式正则化效应。据我们所知，本工作首次通过处理隐式正则化、普通最小二乘法、自蒸馏、模型选择以及稳定性推断结果等技术问题，理解了无偏标签噪声对随机梯度下降动力学的影响。

3 无偏随机标签噪声下随机梯度下降的双重随机模型

本节将介绍采用无偏随机标签噪声的随机梯度下降，推导连续时间/离散时间双重随机模型，并给出模型之间的近似收敛。

3.1 SGD中的无偏标签噪声建模

在我们的研究中，采用无偏随机标签噪声的随机梯度下降是指一种迭代算法，它从初始值 θ_0^{ULN} 开始，逐步更新估计值。在小批量采样和无偏随机标签噪声的条件下，在第 k^{th} 次迭代中，随机梯度下降算法通过梯度下降规则，利用随机梯度 $\tilde{\mathrm{g}}_k(\theta_k^{\mathrm{ULN}})$ 更新估计值 θ_k^{ULN}，使得

$$\theta_{k+1}^{\mathrm{ULN}} \leftarrow \theta_k^{\mathrm{ULN}} - \eta\tilde{\mathrm{g}}_k(\theta_k^{\mathrm{ULN}}) \ , \tag{11}$$

具体而言，在第 k^{th} 次迭代中，随机梯度下降随机选取一批样本$B_k \subseteq \mathcal{D}$ 来估计随机梯度，如下所示

$$\begin{aligned}
\eta\tilde{\mathrm{g}}_k(\theta_k^{\mathrm{ULN}}) &= \frac{\eta}{|B_k|}\sum_{x_i\in B_k}\nabla\tilde{L}_i(\theta_k^{\mathrm{ULN}}) \\
&= \frac{\eta}{|B_k|}\sum_{x_i\in B_k}\big(\big(f(x_i,\theta_k^{\mathrm{ULN}}) - y_i\big) - \varepsilon_i\big)\cdot\nabla_\theta f(x_i,\theta_k^{\mathrm{ULN}}) \\
&= \frac{\eta}{N}\sum_{i=1}^{N}\nabla L_i^*(\theta_k^{\mathrm{ULN}}) + \sqrt{\eta}\xi_k^*(\theta_k^{\mathrm{ULN}}) + \sqrt{\eta}\xi_k^{\mathrm{ULN}}(\theta_k^{\mathrm{ULN}}),
\end{aligned} \tag{12}$$

其中 $\nabla L_i^*(\theta)$ 对于 $\forall\theta \in \mathbb{R}^d$ 指基于无标签噪声样本（x_i, y_i）的损失梯度，而 $y_i = f^*(x_i)$、$\xi_k^*(\theta)$ 指的是通过对无标签噪声样本的梯度进行小批量采样而产生的随机梯度噪声，[17] 而 $\xi_k^{\mathrm{ULN}}(\theta)$ 是由小批量采样和无偏随机标签噪声引起的附加噪声项，使得

$$\begin{aligned}
\nabla L_i^*(\theta) &= \frac{\partial}{\partial\theta}\frac{(f(x_i,\theta) - f^*(x_i))^2}{2} = (f(x_i,\theta) - f^*(x_i))\cdot\nabla f(x_i,\theta) \ , \\
\xi_k^*(\theta) &= \frac{\sqrt{\eta}}{|B_k|}\sum_{x_j\in B_k}\left(\nabla L_j^*(\theta) - \frac{1}{N}\sum_{i=1}^{N}\nabla L_i^*(\theta)\right), \\
\xi_k^{\mathrm{ULN}}(\theta) &= -\frac{\sqrt{\eta}}{|B_k|}\sum_{x_j\in B_k}\varepsilon_j\cdot\nabla_\theta f(x_j,\theta) \ .
\end{aligned} \tag{13}$$

命题1（两个噪声项的均值和方差）。噪声项 $\xi_k^*(\theta)$ 和 $\xi_k^{\mathrm{ULN}}(\theta)$ 的均值和方差应为如下向量值函数

$$\begin{aligned}
&\mathbb{E}_{B_k}[\xi_k^*(\theta)] = \mathbf{0}_d, \qquad\qquad and\ \mathrm{Var}_{B_k}[\xi_k^*(\theta)] = \frac{\eta}{|B_k|}\Sigma_N^{\mathrm{SGD}}(\theta) \\
&\mathbb{E}_{B_k,\varepsilon_i}[\xi_k^{\mathrm{ULN}}(\theta)] = \mathbf{0}_d, \quad and\ \mathrm{Var}_{B_k\varepsilon_i}[\xi_k^{\mathrm{ULN}}(\theta)] = \frac{\eta}{|B_k|}\Sigma_N^{\mathrm{ULN}}(\theta) \ .
\end{aligned} \tag{14}$$

定义在 $\theta \in \mathbb{R}^d$ 上的两个矩阵值函数 $\Sigma_N^{\mathrm{SGD}}(\theta)$ 和 $\Sigma_N^{\mathrm{ULN}}(\theta)$ 刻画了噪声向量的方差。当我们假设标签噪声与

小批量采样相互独立时，

$$\Sigma_N^{\mathrm{SGD}}(\theta) = \frac{1}{N}\sum_{j=1}^{N}\left(\nabla L_j^*(\theta) - \frac{1}{N}\sum_{i=1}^{N} L_i^*(\theta)\right)\left(\nabla L_j^*(\theta) - \frac{1}{N}\sum_{i=1}^{N} L_i^*(\theta)\right)^{\top}$$
$$\Sigma_N^{\mathrm{ULN}}(\theta) = \frac{\sigma^2}{N}\sum_{j=1}^{N}\nabla_\theta f(x_j,\theta)\nabla_\theta f(x_j,\theta)^{\top} \quad as \quad \mathrm{var}[\varepsilon_j] = \sigma^2 \ . \tag{15}$$

两个噪声项 $\xi_k^*(\theta)$ 和 $\xi_k^{\mathrm{ULN}}(\theta)$ 受学习率和批量大小控制，将在很大程度上影响随机梯度下降动力学。证明请参见 *A*。

根据两个噪声项的均值和方差，我们可以轻松地将含无偏标签噪声的随机梯度下降学习动力学表述如下。

3.2 双重随机模型与近似

我们考虑一种含无偏随机标签噪声的随机梯度下降算法，其形式为带加性依赖数据的噪声的梯度下降。，_使得 $\theta_{k+1} = \theta_k - \frac{\eta}{N}\sum_{i=1}^{N}\nabla\tilde{L}_i(\theta_k) + \sqrt{\eta}\tilde{V}_k(\theta_k)$。当 $\eta \rightarrow 0$ 时，我们假设噪声项 $\xi_k^*(\theta_k)$ 和 $\xi_k^{\mathrm{ULN}}(\theta_k)$ 相互独立，随后可以沿用 [19] 中的分析，推导含无偏随机标签噪声的随机梯度下降的扩散过程，记为 $\Theta^{\mathrm{ULN}}(t)$，其定义在连续时间 $t \geq 0$ 上。由此，我们定义双重随机模型，用于刻画含无偏标签噪声的随机梯度下降的连续时间动力学，如下所示。

定义 1 (连续时间双重随机模型)。 给定第 *3.1* 节中定义并指定的随机梯度下降算法 θ_k^{ULN}，其中 $\eta = \mathbf{dt}$，我们假设$B_k = B$对于 $k = 1$、2、3.....并将其连续时间动力学表述为

$$\mathbf{d}\Theta^{\mathrm{ULN}} = -\frac{1}{N}\sum_{i=1}^{N}\nabla L_i^*(\Theta^{\mathrm{ULN}})\mathbf{d}t + \left(\frac{\eta}{b}\Sigma_N^{\mathrm{SGD}}(\Theta^{\mathrm{ULN}})\right)^{\frac{1}{2}}\mathbf{d}W_1(t) + \left(\frac{\eta}{b}\Sigma_N^{\mathrm{ULN}}(\Theta^{\mathrm{ULN}})\right)^{\frac{1}{2}}\mathbf{d}W_2(t) \ , \tag{16}$$

其中 $W_1(t)$ 和 $W_2(t)$ 表示随时间演化的两个相互独立的布朗运动，$\mathbf{d}t = \eta$ 和 $\Theta^{\mathrm{ULN}}(0) = \theta_0^{\mathrm{ULN}}$。

显然，我们可以得到随机梯度下降动力学的离散时间近似 [11, 17] ，如下所示。

定义 2 （离散时间双重随机模型）。 我们将 θ_k^{ULN}在 $k = 1$，2，..... 记作以下模型的离散时间近似：用于具有无偏标签噪声的随机梯度下降的双重随机模型，该模型在 k^{th} 次迭代中表现为

as

$$\bar{\theta}_{k+1}^{\mathrm{ULN}} \leftarrow \bar{\theta}_k^{\mathrm{ULN}} - \frac{\eta}{N}\sum_{i=1}^{N}\nabla L_i^*(\bar{\theta}_k^{\mathrm{ULN}}) + \sqrt{\eta}\left(\frac{\eta}{b}\Sigma_N^{\mathrm{SGD}}(\bar{\theta}_k^{\mathrm{ULN}})\right)^{\frac{1}{2}} z_k + \sqrt{\eta}\left(\frac{\eta}{b}\Sigma_N^{\mathrm{ULN}}(\bar{\theta}_k^{\mathrm{ULN}})\right)^{\frac{1}{2}} z_k', \tag{17}$$

其中， z_k 和 z_k' 是两个相互独立的 d 维随机向量，每次迭代均独立地从标准 d 维高斯分布 $\mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ 中抽取，且 $\bar{\theta}_0^{\mathrm{ULN}} = \Theta^{\mathrm{ULN}}(0)$。

当 $t = k\eta$ 时， $\bar{\theta}_k^{\mathrm{ULN}}$与 $\Theta^{\mathrm{ULN}}(t)$ 之间的收敛是紧的，且收敛界如下。

命题 2（近似收敛）。设 $T \geq 0$。设 $\Sigma_N^{\mathrm{SGD}}(\theta)$ 和 $\Sigma_N^{\mathrm{ULN}}(\theta)$ 为方程 *15* 中定义的两个扩散矩阵。假设 ***A1*** 存在某个 $M > 0$，使得 $\max_{i=1,\,2,\,\ldots,N}\{(\|\nabla L_i^*(\theta)\|_2)\} \leq M$，并且

$$\max_{i}\{(\|\nabla_\theta f(x_i;\,\theta)\|_2)\} \leq M;$$

A2 存在某个 $L > 0$，使得 $\nabla L_i(\theta)$ 和 $\nabla_\theta f(x,\,\theta)$ 对于 $\forall x \in \mathcal{D}$ 都是利普希茨连续的，并且利普希茨常数 $L > 0$ 有界，对所有 $i = 1,\,2,\,\ldots,\,N$ 一致成立。

带无偏标签噪声的随机梯度下降（*ULN*）的连续时间动力学，记为 $\Theta^{\mathrm{ULN}}(t)$ 见方程（16），是对离散时间随机梯度下降动力学 $\bar{\theta}_k^{\mathrm{ULN}}$（见方程（17））的一阶 1 强近似。也就是说，存在常数 C，其与 η 无关但取决于 σ^2、L 和 M，使得

$$\mathbb{E}\|\Theta^{\mathrm{ULN}}(k\eta) - \bar{\theta}_k^{\mathrm{ULN}}\|^2 \leq C\eta^2, \quad \textit{for all } 0 \leq k \leq \lfloor T/\eta \rfloor. \tag{18}$$

证明请参见 *B* 。

备注 1. 基于上述近似的强收敛界，我们可以将 $\bar{\theta}_k^{\mathrm{ULN}}$– 即方程 (17) 的解 – 视为基于相同初始化、含无偏标签噪声的随机梯度下降算法的紧致近似。噪声项 $\xi_k^{\mathrm{ULN}}(\theta)$ *(*定义见方程 (27)*)* 的紧致近似可写为

$$\xi_k^{\mathrm{ULN}}(\theta) \approx \left(\frac{\eta}{b}\Sigma_N^{\mathrm{ULN}}(\theta)\right)^{\frac{1}{2}} z_k', \ \textbf{\textit{and}}\ z_k' \sim \mathcal{N}(\mathbf{0_d}, \mathbf{I_d}) \ . \tag{19}$$

我们使用这样的离散时间迭代和近似来解释随机梯度下降与无偏标签噪声算法的隐式正则化行为，其中噪声项为$\xi_k^{\mathrm{ULN}}(\theta)$ θ_k^{ULN}。

4 神经网络的隐式正则化效应

在本节中，我们使用我们的模型来解释带有无偏标签噪声的随机梯度下降对一般神经网络的正则化效应，而不对神经网络的结构作任何假设。

4.1 受无偏随机标签噪声影响的隐式正则化器

将含无偏随机标签噪声的随机梯度 $\tilde{g}_k(\theta)$ 与基于无标签噪声损失的随机梯度进行比较，我们发现了一个附加噪声项 $\xi_k^{\mathrm{ULN}}(\theta)$，它充当**隐式正则化器**。

为了阐释 $\xi_k^{\mathrm{ULN}}(\theta)$，我们首先定义基于无标签噪声损失的随机梯度下降（SGD）扩散过程，即 $L_i^*(\theta)$，对于 $1 \leq i \leq N$，其定义为

$$\mathbf{d}\Theta^{\mathrm{LNL}} = -\frac{1}{N}\sum_{i=1}^{N}\nabla L_i^*(\Theta^{\mathrm{LNL}})\mathbf{d}t + \left(\frac{\eta}{b}\Sigma_N^{\mathrm{SGD}}(\Theta^{\mathrm{LNL}})\right)^{\frac{1}{2}}\mathbf{W}(t)\ . \tag{20}$$

通过比较 $\Theta^{\mathrm{ULN}}(t)$ 与 $\Theta^{\mathrm{LNL}}(t)$，$\xi_k^{\mathrm{ULN}}(\Theta)$ 对连续时间形式的影响应为 $\sqrt{\eta/B}(\Sigma_N^{\mathrm{ULN}}(\Theta))^{1/2}\mathrm{d}W(t)$。然后，在离散时间下，我们可以得到如下结果。

命题3 (隐式正则化器 $\xi_k^{\mathrm{ULN}}(\theta)$)**.** 带有无偏随机标签噪声的随机梯度下降的隐式正则化器可近似如下：

$$\xi_k^{\mathrm{ULN}}(\theta) \approx \left(\frac{\sigma^2\eta}{bN}\sum_{i=1}^{N}\nabla_\theta f(x_i,\theta)\nabla_\theta f(x_i,\theta)^\top\right)^{\frac{1}{2}} z_k,\ \textit{and}\ z_k \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)\ . \tag{21}$$

这样，我们可以估计隐式正则化器 $\|\xi_k^{\mathrm{ULN}}(\theta)\|_2$ 的期望正则化效应，如下所示：

$$\mathbb{E}_{z_k}\|\xi_k^{\mathrm{ULN}}(\theta)\|_2^2 = \frac{\eta\sigma^2}{bN}\sum_{i=1}^{N}\|\nabla_\theta f(x_i,\theta)\|_2^2\,. \tag{22}$$

证明请参见 C 。

因此，我们得出结论：由无偏随机标签噪声引起的随机梯度下降隐式正则化效应，与神经网络在样本上的 $\frac{1}{N}\sum_{i=1}^{N}\|\nabla_\theta f(x_i,\theta)\|_2^2$ 平均梯度范数 $f(x,\theta)$成正比。

4.2 将无偏标签噪声理解为推理稳定器

在此，我们扩展了关于随机梯度下降的现有结果 [1, 40]，以理解**命题3**，并得出如下评述。

备注 2（推理稳定性）。 在偏导数形式下，梯度范数可以写成

$$\frac{1}{N}\sum_{i=1}^{N}\|\nabla_\theta f(x_i,\theta)\|_2^2=\frac{1}{N}\sum_{i=1}^{N}\|\frac{\partial}{\partial\theta}f(x_i,\theta)\|_2^2 \tag{23}$$

刻画了神经网络输出 f（x、θ）基于样本x_i（对于 $1\leq i\leq N$）在点 θ 附近的参数插值上的变化。较低的 $\frac{1}{N}\sum_{i=1}^{N}\|\nabla_\theta f$（$x_i$、$\theta$）$\|_2^2$ 会带来神经网络 f（x、$\bar{\theta}$）输出对参数上的（随机）扰动具有更高的稳定性。

备注 3（逃逸与收敛）。 当噪声 $\xi_k^{\mathrm{ULN}}(\theta)$ 依赖于 θ 时（第 *4* 节将介绍一种特殊情况，即 $\xi_k^{\mathrm{ULN}}(\theta)$ 与 *OLS* 无关），我们遵循 *[1]* 的观点，并认为隐式正则化器有助于 *SGD* 从点 θ 逃逸，该点具有较大的神经网络梯度范数 $\frac{1}{N}\sum_{i=1}^{N}\|\nabla_\theta f$（$x_i$、$\tilde{\theta}$）$\|_2^2$，因为此时噪声 $\xi_k^{\mathrm{ULN}}(\tilde{\theta})$ 的尺度较大。反过来，我们遵循 *[40]*的观点，并认为当带有无偏随机标签噪声的 *SGD* 收敛时，该算法将收敛到点 θ^*，其 $\frac{1}{N}\sum_{i=1}^{N}\|\nabla_\theta f$（$x_i$、$\theta^*$）$\|_2^2$较小。在假设深度学习算法由奥恩斯坦–乌伦贝克过程驱动时，*[3]* 中也获得了类似的结果。

注 4 (性能调优)。 将 $\eta\sigma^2/B$ 视为平衡隐式正则化器与原始随机梯度下降的系数，可以在固定 η 和 B 的情况下，通过增大σ^2来更强地对随机梯度下降学习过程进行正则化/惩罚。更具体地说，随着 σ^2 增大，正则化效应会增强，因此我们可以期待获得具有更低 $\frac{1}{N}\sum_{i=1}^{N}\|\nabla_\theta f(x_i,\theta)\|_2^2$ 的正则化解，或获得推理稳定性更高的神经网络。

5 无偏标签噪声下的自蒸馏实验

本实验旨在理解**命题3** 和备注*3*和*4*，具体考察：(1) 无偏标签噪声是否会降低神经网络的梯度范数；(2) 这类无偏标签噪声是否会提升神经网络的性能；(3) 在真实世界深度学习场景中，是否可以通过控制无偏标签噪声的方差来进行性能调优。

5.1 实验设计

为了评估带有无偏标签噪声的随机梯度下降，我们基于带有无偏标签噪声的自蒸馏设计了一组新颖的实验。除了直接从噪声标签中学习外，我们的实验还旨在通过二次回归损失，利用教师网络的噪声输出训练一个（学生）网络；学生网络的初始权重取自教师网络，并随机向教师网络的软输出添加无偏标签噪声。

我们旨在每个训练轮次后直接测量神经网络的梯度范数 $\frac{1}{N}\sum_{i=1}^{N}\|\nabla_\theta f(x_i,\theta)\|_2^2$，以验证无偏标签噪声对随机梯度下降隐式正则化的影响（即**命题3**）。教师网络、（使用无偏标签噪声训练的）学生网络和（无噪声训练的）学生网络之间的性能比较，展示了无偏标签噪声在随机梯度下降用于回归任务时的优势（即备注 *3* 和 *4*）。

具体而言，我们基于含无偏标签噪声的自蒸馏 设计了一组新颖的实验，并详细阐述所提出的含无偏标签噪声的随机梯度下降 以何种方式适用于含无偏标签噪声的自蒸馏设置。此外，我们介绍实证实验的目标，并列出预期证据，随后说明实证评估的实验设置。最后，我们展示实验结果，并提供有力证据来验证本文提出的方法。

5.2 含噪自蒸馏

给定一个训练良好的模型，自蒸馏算法 [28, 29, 34, 35]旨在通过学习模型（作为教师）输出的“软标签”（即逻辑值）来进一步提升模型性能。此外，一些实践发现，在训练过程中引入一定的随机性和随机扰动，即采用含噪自蒸馏，可以进一步提升自蒸馏效果，从而获得更好的泛化性能 [29, 34]。在本文中，我们研究以下两种广为人知的加性噪声策略。

1. 高斯噪声。 给定一个具有 $\mathbb{L}$维逻辑值输出的预训练模型，在每次自蒸馏迭代中，这个简单方法首先从 $\mathbb{L}$维高斯分布 $\mathcal{N}(\mathbf{0}_\mathbb{L},\ \sigma^2\mathbf{I}_\mathbb{L})$中抽取随机向量，然后将这些向量加到模型的Logit输出上。这样，学生模型就能从含噪输出中学习。需要注意的是，在我们的分析中，我们假设模型的输出为单维，而在自蒸馏中，逻辑值标签是多维的。因此，对角矩阵 $\sigma^2\mathbf{I}_\mathbb{L}$现在指方差的完整形式，而 σ^2 控制噪声方差的尺度。

2. 对称噪声。。基本上，该策略源自 [15] 通过在 $\mathbb{L}$ 个维度之间随机交换逻辑值输出的值来生成噪声的方法。具体而言，在每次自蒸馏迭代中，给定

交换概率为 p，预训练模型的每个 Logit输出（此处记为 y ），以及记为 y_l的逻辑值输出的每个维度，概率为 p 的策略会将与 y_l 对应维度中的逻辑值与任意其他维度 $y_{m\neq l}$ 以相同的先验概率进行交换（即以 $(\mathbb{L}-1)^{-1}$ 的概率）。在其余 $1-p$ 的概率下，该策略保持该处原有的逻辑值输出。通过这种方式，新的噪声标签 $\tilde{y}$ 的期望 $\mathbb{E}[\tilde{y}]$如下:

$$\mathbb{E}[\tilde{y}_l] = (1-p)\cdot y_l + \frac{p\cdot\sum_{\forall m\neq l} y_m}{\mathbb{L}-1} \tag{24}$$

这种策略会给原始 Logit输出引入显式偏差。然而，当我们将期望 $\mathbb{E}[\tilde{y}]$ 视为创新的软标签时，新软标签周围的随机噪声在所有维度上仍然是无偏的$\mathbb{E}[\tilde{y}-\mathbb{E}[\tilde{y}]]=0$。请注意，这种噪声并非鲁棒学习研究的对称噪声 [41]。

因此，从字面上看，我们提出的无偏标签噪声下的随机梯度下降设置非常符合噪声自蒸馏的实践。

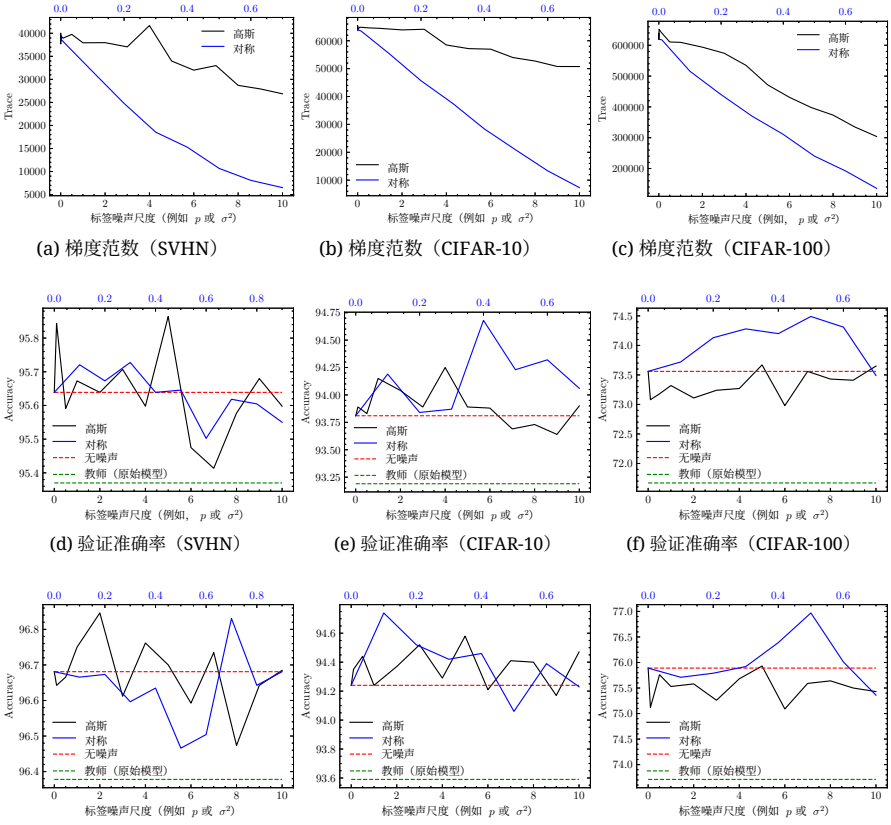
5.3 数据集和深度神经网络模型

我们选择 ResNet-56 [42]，，这是一种最实用的深度模型，用于在三个数据集上进行实验：SVHN [43]，CIFAR-10 和 CIFAR-100 [44]。我们遵循标准训练过程 [42] 来训练教师模型（原始模型）。具体而言，我们从头开始训练模型 200 个训练轮次，并采用批量大小为 64、动量为 0.9 的随机梯度下降优化器。训练开始时，学习率设为 0.1，并在第 100^{th} 训练轮次和第 150^{th} 训练轮次时除以 10。我们采用标准权重衰减，并使用较小的正则化参数（10^{-4}）。对于无噪声自蒸馏，我们遵循标准过程 [45] ，将教师模型的知识蒸馏给网络结构相同的学生模型。基本训练设置与训练教师模型时相同。

对于带噪声的自蒸馏设置，我们将原始训练集划分为新的训练集（80%）和验证集（20%）。为清晰起见，我们还展示了在三个数据集上采用不同程度的无偏标签噪声时的结果，其中原始训练集用于训练。

5.4 实验结果

图1 展示了上述两种方法在噪声规模逐渐增大时的结果，即高斯噪声下 σ^2 逐渐增大，对称噪声下 p 逐渐增大。在图1(a)–(c)中，我们展示了神经网络 $\frac{1}{N}\sum_{i=1}^{N}\|\nabla_\theta f(x_i,\ \theta)\|_2^2$的梯度范数会随着 σ^2和 p 的增大而在两种策略下均有所下降。这些结果印证了我们的理论分析，意味着正则化降低了神经网络输出在参数可能受到随机扰动时的变化幅度，从而使模型具有较高的推断稳定性



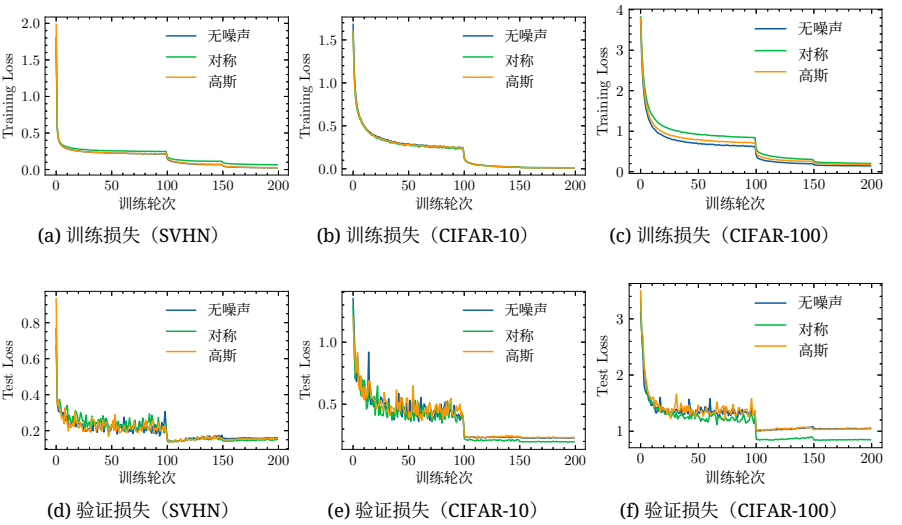
(a) 梯度范数（SVHN） (b) 梯度范数（CIFAR-10） (c) 梯度范数（CIFAR-100）

(d) 验证准确率（SVHN） (e) 验证准确率（CIFAR-10） (f) 验证准确率（CIFAR-100）

(g) 测试准确率（SVHN） (h) 测试准确率（CIFAR-10） (i) 测试准确率（CIFAR-100）

图 1：使用 ResNet-56 进行含噪自蒸馏时，基于 SVNH、CIFAR-10 和 CIFAR-100 数据集，在不同标签噪声尺度（例如， p 和 σ^2）下的梯度范数（对应推理稳定性）、验证准确率和测试准确率。

。在图1(d)–(f)和(g)–(i)中，我们绘制了通过含噪自蒸馏获得的模型的验证准确率和测试准确率。结果表明：(1) 与教师模型相比，学生模型的神经网络 $\frac{1}{N}\sum_{i=1}^{N}\|\nabla_\theta f(x_i,\ \theta)\|_2^2$ 梯度范数更低，并且梯度范数会随着噪声规模（即 σ^2 和 P）的增大而进一步降低；(2) 与教师模型相比，部分模型经过含噪自蒸馏后有所改进，而且含噪自蒸馏的表现优于无噪声自蒸馏；(3) 可以根据验证准确率选择含噪自蒸馏模型，从而获得更好的整体泛化性能（在测试数据集上）。此处所有结果均基于进行200个训练轮次的含噪自蒸馏。



(a) 训练损失（SVHN） (b) 训练损失（CIFAR-10） (c) 训练损失（CIFAR-100）

(d) 验证损失（SVHN） (e) 验证损失（CIFAR-10） (f) 验证损失（CIFAR-100）

Fig. 2：训练过程中每个训练轮次的训练损失和验证损失 e

我们展示了整个训练过程中训练损失和测试损失的变化，并比较了在含噪自蒸馏中不添加标签噪声、添加对称噪声和高斯噪声的设置。图 2 展示了在验证集上采用最优标签噪声尺度时，三个数据集（即 SVHN、CIFAR-10 和 CIFAR-100 ）上的结果。结果表明，所有算法最终都会收敛到训练损失接近于零的局部极小值，而随机梯度下降结合对称噪声所找到的局部极小值更加平坦，泛化性能也更好（尤其是在 CIFAR-100 数据集上）。

6 无偏标签噪声下的线性回归实验

为了验证我们在线性回归设置下的发现，我们使用合成数据进行数值评估，以直观展示带标签噪声的普通最小二乘法和无标签噪声的普通最小二乘法下，随机梯度下降算法随迭代的动态变化。

6.1 无偏标签噪声下的线性回归

我们希望了解无偏标签噪声会如何影响普通最小二乘法（OLS）的随机梯度下降迭代过程，其中 OLS 采用一个简单的二次损失函数，即

$$\widehat{\beta}_{\mathrm{OLS}} \leftarrow \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \tilde{L}_i(\beta) := \frac{1}{2N} \sum_{i=1}^{N} \left(x_i^\top \beta - \tilde{y}_i \right)^2 \right\} , \tag{25}$$

其中，样本通过 $\tilde{y}_i = x_i^\top \beta^* + \varepsilon_i$、 $\mathbb{E}[\varepsilon_i] = 0$ 生成，且 $\mathbf{var}[\varepsilon_i] = \sigma^2$。请注意，在本节中，我们将记号 θ 替换为 β，以表示线性回归模型的参数。

将式（25）和式（11）合并，我们将带有无偏标签噪声的普通最小二乘法的随机梯度下降写成如下迭代形式 β_k^{ULN}其中 $k = 1,\ 2,\ 3$.....如下

$$\beta_{k+1}^{\mathrm{ULN}} \leftarrow \beta_k^{\mathrm{ULN}} - \frac{\eta}{N}\sum_{i=1}^{N} \nabla L_i^*(\beta_k^{\mathrm{ULN}}) + \sqrt{\eta}\xi_k^*(\beta_k^{\mathrm{ULN}}) + \sqrt{\eta}\xi_k^{\mathrm{ULN}}(\beta_k^{\mathrm{ULN}}), \quad (26)$$

其中，对于 $\forall \theta \in \mathbb{R}^d$， $\nabla L_i^*(\beta)$ 表示基于无标签噪声样本（x_i， y_i）的损失梯度； $y_i = x_i^\top \beta^*$， $\xi_k^*(\beta)$ 表示对无标签噪声样本的梯度进行小批量采样所产生的随机梯度噪声；而 $\xi_k^{\mathrm{ULN}}(\beta)$ 是由小批量采样和无偏标签噪声引起的附加噪声项，使得

$$\begin{aligned} \nabla L_i^*(\beta) &= \frac{\partial}{\partial \beta}\frac{(x_i^\top \beta - x_i^\top \beta^*)^2}{2} = x_i x_i^\top (\beta - \beta^*), \\ \xi_k^*(\beta) &= \frac{\sqrt{\eta}}{|B_k|}\sum_{x_j \in B_k}\left(\nabla L_j^*(\beta) - \frac{1}{N}\sum_{i=1}^{N} L_i^*(\beta)\right), \\ \xi_k^{\mathrm{ULN}}(\beta) &= -\frac{\sqrt{\eta}}{|B_k|}\sum_{x_j \in B_k} \varepsilon_j \cdot x_j \ . \end{aligned} \quad (27)$$

我们将 N 个样本的样本协方差矩阵 记为

$$\bar{\Sigma}_N = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^\top \ . \quad (28)$$

根据一般形式下关于隐式正则化的备注 *1*，我们可以将带有随机标签噪声的普通最小二乘法的随机梯度下降的隐式正则化器写为:

$$\xi_k^{\mathrm{ULN}}(\beta) \approx \left(\frac{\eta\sigma^2}{B}\bar{\Sigma}_N\right)^{\frac{1}{2}} z_k' = \left(\frac{\eta\sigma^2}{bN}\sum_{i=1}^{N} x_i x_i^\top\right)^{\frac{1}{2}} z_k', \text{ and } z_k' \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d), \quad (29)$$

该方法是无偏的 $\mathbb{E}[\xi_k^{\mathrm{ULN}}(\beta)] = \mathbf{0_d}$，具有不变协方差结构，并且与 β （位置）和 k （时间）相互独立。

将**命题 2** 与线性回归的设定相结合，我们可得到具有无偏标签噪声的线性回归的连续时间动力学，记为 $\beta^{\mathrm{ULN}}(t)$。根据 [33, 46], ，我们可以看出，无噪声线性回归的随机梯度下降及其连续时间动力学（分别记为 β_k^{LNL}和 $\beta^{\mathrm{LNL}}(t)$）均会渐近收敛到最优解 β^*。由于附加噪声项 ξ^{ULN} 是无偏的，且具有不变协方差结构，当

$t \to \infty$时，我们可以直接得出 $\lim_{t\to\infty} \mathbb{E}\ \beta^{\mathrm{ULN}}(t) = \lim_{t\to\infty} \mathbb{E}\ \beta^{\mathrm{LNL}}(t) = \beta^*$，$\lim_{t\to\infty} \mathrm{d}\beta^{\mathrm{ULN}}(t) = (\frac{\eta\sigma^2}{B}\bar{\Sigma}_N)^{1/2}\mathrm{d}W(t)$。根据随机过程的分布定义，我们可以得出，当 $t \to \infty$时，$\beta^{\mathrm{ULN}}(t)$ 收敛到一个平稳分布，使得 $\beta^{\mathrm{ULN}}(t) \sim \mathcal{N}(\beta^*, \frac{\eta\sigma^2}{B}\bar{\Sigma}_N)$。

注 5. 因此，在 $k \to \infty$ 的情况下，采用无偏标签噪声的普通最小二乘法随机梯度下降算法将收敛到如下类似高斯分布的分布

$$\lim_{k\to\infty} \mathbb{E}\ [\beta_k^{\mathrm{ULN}}] = \beta^*, and\ \lim_{k\to\infty} \mathrm{Var}\ [\beta_k^{\mathrm{ULN}}] = \frac{\eta\sigma^2}{B}\bar{\Sigma}_N, \tag{30}$$

当 η 和 B 为常数时，该分布的跨度和形状由 σ^2 和 $\bar{\Sigma}_N$ 控制。

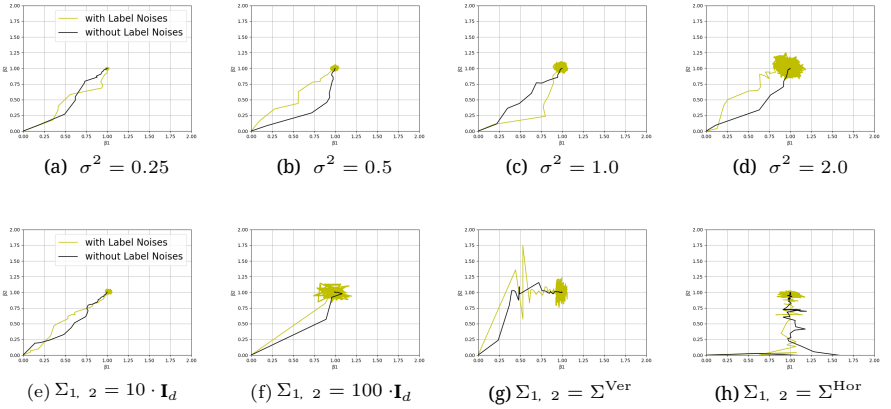
在本实验中，我们希望通过数值模拟评估上述说明，以验证：(1) β_k^{ULN} 的轨迹是否收敛到 $\mathcal{N}(\beta^*, \frac{\eta\sigma^2}{B}\bar{\Sigma}_N)$ 的分布；(2) 收敛区域的形状是否可由数据 Σ_N 的样本协方差矩阵控制；(3) 收敛区域的大小是否可由标签噪声的方差 σ^2 控制。

6.2 实验设置

在我们的实验中，我们使用从二维高斯分布 $X_i \sim \mathcal{N}(\mathbf{0}, \Sigma_{1,2})$ 中生成的 100 个随机样本，其中 $\Sigma_{1,2}$ 是一个对称协方差矩阵，用于控制随机样本的生成。为了给标签添加噪声，我们首先从方差为 $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ 的正态分布中抽取 100 个随机噪声副本，然后使用 $\tilde{y}_i = X_i^\top \beta^* + \varepsilon_i$ 和 $\beta^* = [1,1]^\top$以及 σ^2 和 $\Sigma_{1,2}$的各种设置，通过 $(X_i, \tilde{y}_i)$ 样本对建立普通最小二乘法问题。我们将随机梯度下降算法的学习率固定为 $\eta = 0.01$，批量大小设为 $B = 5$，迭代总数设为 $K = 1,000,000$，以展示完整的路径。

6.3 实验结果

图3 展示了数值验证的结果。在图3(a)–(d)中，我们逐渐将标签噪声的方差 σ^2 从 0.25 增大到 2.0，可以观察到：(1) 对无标签噪声的普通最小二乘法进行随机梯度下降，会快速收敛到最优解 $\beta^* = [1.0, 1.0]^\top$；(2) 对含有无偏随机标签噪声的普通最小二乘法进行随机梯度下降，渐近地收敛到以最优点为中心的分布；(3) 随着 σ^2 增大，收敛分布的范围也会变大。在图3(e)–(h)中，我们采用了四种 $\Sigma_{1,2}$ 设置，可以看到，无论如何设置普通最小二乘法问题的 $\Sigma_{1,2}$，(4) 含有无偏随机标签噪声的随机梯度下降都会渐近地收敛到以最优点为中心的分布。



(a) $\sigma^2 = 0.25$ (b) $\sigma^2 = 0.5$ (c) $\sigma^2 = 1.0$ (d) $\sigma^2 = 2.0$

(e) $\Sigma_{1,\ 2} = 10 \cdot \mathbf{I}_d$ (f) $\Sigma_{1,\ 2} = 100 \cdot \mathbf{I}_d$ (g) $\Sigma_{1,\ 2} = \Sigma^{\mathrm{Ver}}$ (h) $\Sigma_{1,\ 2} = \Sigma^{\mathrm{Hor}}$

图 3：使用不同的 $\tilde{\sigma}^2$ 和 $\Sigma_{1,\ 2}$ 设置进行（含噪）随机数据生成时，采用和不采用无偏随机标签噪声的 OLS 上 SGD 的轨迹。对于图 (a)–(d)，实验设置为固定$\Sigma_{1,\ 2} = [[20, 0]^\top, [0, 20]^\top]$并改变 $\tilde{\sigma}^2$。对于图 (e)–(h)，实验设置为固定 $\tilde{\sigma}^2 = 0.5$并改变$\Sigma_{1,\ 2}$，其中我们设置 $\Sigma^{\mathrm{Ver}} = [[10, 0]^\top, [0, 100]^\top]$ 和 $\Sigma^{\mathrm{Hor}} = [[100, 0]^\top, [0, 10]^\top]$ 以塑造收敛分布。

比较 (e) 与 (f) 中的结果可以发现，当$\Sigma_{1,2}$ 的迹增大时，收敛分布的范围也会增大。此外，(5) 收敛分布的形状取决于$\Sigma_{1,2}$。在图3(g) 中，当我们将$\Sigma_{1,\ 2}$ 的主成分置于垂直轴上（即$\Sigma^{\mathrm{Ver}} = [[10, 0]^\top, [0, 100]^\top]$）时，分布主要沿垂直轴展开。图3(h) 展示了分布相反的布局，此时我们将 $\Sigma^{\mathrm{Hor}} = [[100, 0]^\top, [0, 10]^\top]$ 设为$\Sigma_{1,2}$。收敛分布的尺度和形状支持了我们在式（29）中的理论分析。

请注意，无偏随机标签噪声是在学习过程开始前添加到标签中的。在这种设置下，是SGD的小批量采样器重新采样噪声，并通过形成隐式正则化器影响SGD的动态过程。

7 讨论与结论

以往研究主要关注标签噪声或受损标签导致的性能下降 [37, 47], ，而我们研究随机标签噪声在随机梯度下降（SGD）的小批量采样设置下产生的隐式正则化效应。具体而言，我们采用 SGD 的动力系统解释，分析基于带有无偏随机标签噪声的二次损失的学习过程。我们将基于含标签噪声损失的小批量随机梯度分解为式（11）中的三部分：(i) $\nabla L^*(\theta)$——无标签噪声损失的真实梯度，(ii) $\xi_k^*(\theta)$

– 由对无标签噪声损失进行小批量采样而产生的随机梯度噪声，以及 (iii) $\xi_k^{\mathrm{ULN}}(\theta)$ – 受随机标签噪声和小批量采样共同影响的噪声项。我们的研究将 $\xi_k^{\mathrm{ULN}}(\theta)$ 视为隐式正则化器，并发现这种隐式正则化器的作用是在学习过程中降低神经网络的梯度范数 $\frac{1}{N}\sum_{i=1}^{N}\|\nabla_\theta f(x_i,\theta)\|_2^2$。这里,_神经网络的梯度范数刻画了神经网络输出相对于参数周围随机扰动的变化／稳定性。总之，新的隐式正则化器 $\xi_k^{\mathrm{ULN}}(\theta)$ 有助于随机梯度下降选择一个推理稳定性更高的收敛点。

我们进行了大量实验，以验证我们的理论研究。基于深度神经网络的评估表明，在自蒸馏设置下，通过迭代地向教师模型的输出添加噪声，可以降低神经网络的梯度范数、提高网络的推理稳定性，并获得更好的解。线性回归的数值研究清晰展示了有无无偏随机标签噪声时基于随机梯度下降的线性回归轨迹；这一观察与我们理论推导出的随机梯度下降动力学一致（即收敛形态受数据、学习率和批量大小控制）。实证观察充分支持了我们的理论发现。

声明

- 经费支持 - Haoyi Xiong 和 Xuhong Li 获得中国国家重点研发计划资助（项目编号：2021ZD0110303）。
- 利益冲突/竞争性利益 - 不适用
- 伦理审批 - 不适用（所有研究均基于公开可用的开源数据集）
- 参与知情同意 - 不适用（所有研究均基于公开可用的开源数据集）
- 发表知情同意 - 不适用（所有研究均基于公开可用的开源数据集）
- 数据和材料可用性——所有实验均基于公开可用的开源数据集。
- 代码可用性——所有代码将在论文被接收后发布。
- 作者贡献——Haoyi Xiong 博士提出了最初的研究构想，明确了研究问题，提出了部分算法，撰写了论文并参与讨论。Xuhong Li 博士提出了最初的研究构想，完成了部分实验，并撰写了部分论文。Boyang Li 先生协助完成了部分数学证明。Dongrui Wu 和 Zhanxing Zhu 参与了讨论。Dejing Dou 教授监督了研究进展。

参考文献

[1] Z. Zhu、J. Wu、B. Yu、L. Wu、J. Ma，《随机梯度下降中的各向异性噪声：其逃离尖锐极小值的行为及正则化效应》，载于：国际机器学习会议，2019年，第7654–7663页。[2] J. Wu、W. Hu、H. Xiong、J. Huan、V. Braverman、Z. Zhu，《论能够实现如随机梯度下降般泛化的含噪梯度下降》，载于：国际机器学习会议，机器学习研究论文集，2020年，第10367–10376页。[3] G. Blanc、N. Gupta、G. Valiant、P. Valiant，《由类奥恩斯坦–乌伦贝克过程驱动的深度神经网络的隐式正则化》，载于：学习理论会议，机器学习研究论文集，2020年，第483–513页。[4] L. Bottou，《神经网络中的随机梯度学习》，Neuro-Nımes 91会议论文集（8）。[5] C. Zhang、S. Bengio、M. Hardt、B. Recht、O. Vinyals，《理解深度学习需要重新思考泛化》，载于：国际学习表征会议，2017年。[6]S. Jastrzebski、Z. Kenton、D. Arpit、N. Ballas、A. Fischer、Y. Bengio、A. Storkey，《影响随机梯度下降中极小值的三个因素》，arXiv预印本 arXiv:1711.04623。[7] E. Hoffer、I. Hubara、D. Soudry，《训练更久，泛化更好：弥合神经网络大批量训练中的泛化差距》，载于：I. Guyon、U. V. Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan、R. Garnett（编），《神经信息处理系统进展第30卷》，Curran Associates, Inc.，2017年，第1731–1741页。[8] N. S. Keskar、D. Mudigere、J. Nocedal、M. Smelyanskiy、P. T. P. Tang，《深度学习的大批量训练：泛化差距与尖锐极小值》，载于：国际学习表征会议（ICLR），2017年。[9] L. Bottou、F. E. Curtis、J. Nocedal，《大规模机器学习的优化方法》，SIAM评论 60（2）（2018年），第223–311页。[10]S. Mandt、M. D. Hoffman、D. M. Blei，《作为近似贝叶斯推断的随机梯度下降》，《机器学习研究期刊》18（1）（2017年），第4873–4907页。[11] P. Chaudhari、S. Soatto，《随机梯度下降执行变分推断，并在深度网络中收敛至极限环》，载于：国际学习表征会议，2018年。

[12] W. Hu、Z. Zhu、H. Xiong、J. Huan，将准势作为随机梯度下降中损失函数的隐式正则化器，arXiv预印本 arXiv:1901.06054。[13]U. Simsekli、L. Sagun、M. Gurbuzbalaban，深度神经网络中随机梯度噪声的尾指数分析，arXiv预印本 arXiv:1901.06053。[14] S. Hochreiter、J. Schmidhuber，平坦极小值，《神经计算》9（1）（1997年）1–42。[15] B. Han、Q. Yao、X. Yu、G. Niu、M. Xu、W. Hu、I. Tsang、M. Sugiyama，协同教学：针对噪声标签极多的深度神经网络的鲁棒训练，载于：《神经信息处理系统进展》，2018年，第8527–8537页。[16] A. M. Saxe、J. L. McClelland、S. Ganguli，深度线性神经网络学习非线性动力学的精确解，载于：国际学习表征会议（ICLR），2014年。[17] Q. Li、C. Tai、E. Weinan，随机修正方程与自适应随机梯度算法，载于：国际机器学习会议，2017年，第2101–2110页。[18] J. Wu、W. Hu、H. Xiong、J. Huan、V. Braverman、Z. Zhu，关于能够像随机梯度下降一样实现泛化的含噪梯度下降，载于：国际机器学习会议（ICML），2020年。[19] W. Hu、C. J. Li、L. Li、J.-G. Liu，非凸随机梯度下降的扩散近似，《数学科学与应用年刊》4（1）（2019年）3–32。[20] S. L.Smith、P.-J. Kindermans、C. Ying、Q. V. Le，不要降低学习率，而应增大批量大小，载于：国际学习表征会议，2018年。[21] R. Zhang、C. Li、J. Zhang、C. Chen、A. G. Wilson，面向贝叶斯深度学习的周期性随机梯度马尔可夫链蒙特卡洛，载于：国际学习表征会议，2019年。[22]H. Xiong、K. Wang、J. Bian、Z. Zhu、C.-Z. Xu、Z. Guo、J. Huan，谱哈密顿蒙特卡洛，载于：《AAAI人工智能会议论文集》第33卷，2019年，第5516–5524页。[23] Z. Jia、H. Su，基于信息论的局部极小值刻画与正则化，载于：国际机器学习会议，机器学习研究论文集，2020年，第4773–4783页。

[24] G. Gidel、F. Bach、S. Lacoste-Julien，线性神经网络中离散梯度动力学的隐式正则化，载于：《神经信息处理系统进展》，2019年，第3202–3211页。[25] Y. Fang、J. Xu、L. Yang，随机梯度下降估计量的在线自助法置信区间，《机器学习研究期刊》19（1）（2018年）3053–3073。[26] T. Li、L. Liu、A. Kyrillidis、C. Caramanis，使用随机梯度下降进行统计推断，载于：人工智能促进协会，2018年。[27] A. Ali、J. Z. Kolter、R. J. Tibshirani，最小二乘回归提前停止的连续时间视角，载于：人工智能与统计学会议。[28] L. Zhang、J. Song、A. Gao、J. Chen、C. Bao、K. Ma，成为自己的教师：通过自蒸馏提升卷积神经网络的性能，载于：IEEE 国际计算机视觉会议论文集，2019年，第3713–3722页。[29] K. Kim、B. Ji、D. Yoon、S. Hwang，自知识蒸馏：实现更佳泛化的简单方法，arXiv预印本 arXiv:2006.12000。[30]A. Ali、E. Dobriban、R. J. Tibshirani，最小二乘问题中随机梯度流的隐式正则化，载于：国际机器学习会议，2020年。[31] A. Dieuleveut、N. Flammarion、F. Bach，更难、更好、更快、更强的最小二乘回归收敛速率，《机器学习研究期刊》18（1）（2017年）3520–3570。[32]U. Marteau-Ferey、D. Ostrovskii、F. Bach、A. Rudi，超越最小二乘：通过自协调性实现正则化经验风险最小化的快速速率，载于：学习理论会议，2019年，第2294–2340页。[33] R. Berthier、F. Bach、P. Gaillard，无噪声线性模型下随机梯度下降的紧非参数收敛速率，arXiv预印本 arXiv:2006.08212。[34] Q. Xie、M.-T. Luong、E. Hovy、Q. V. Le，使用含噪学生模型进行自训练可提升ImageNet分类性能，载于：IEEE/CVF计算机视觉与模式识别会议，2020年，第10687–10698页。[35]G. Xu、Z. Liu、X. Li、C. C. Loy，知识蒸馏遇上自监督，载于：欧洲计算机视觉会议，2020年。

[36] D. Lopez-Paz,L. Bottou, B. Sch¨olkopf, V. Vapnik，统一蒸馏与特权信息，载于：国际学习表征会议（ICLR），2016年。[37] M. Li, M. Soltanolkotabi, S. Oymak，带提前停止的梯度下降对于过参数化神经网络中的标签噪声具有可证明的鲁棒性，载于：国际人工智能与统计会议，机器学习研究论文集，2020年，第4313–4324页。[38] A. Jacot, F. Gabriel, C. Hongler，神经切线核：神经网络中的收敛与泛化，载于：神经信息处理系统进展，2018年，第8571–8580页。[39] H. Xiong, X. Li, B. Yu, D. Dou, D. Wu, Z. Zhu，https://openreview.net/forum?id=g4szfsQUdy3 （2021年）。网址：https://openreview.net/forum?id=g4szfsQUdy3[40] L. Wu, C. Ma, E. Weinan，过参数化学习中 SGD 如何选择全局极小值：动力稳定性视角，载于：神经信息处理系统进展，2018年，第8279–8288页。[41] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, J. Bailey，面向噪声标签鲁棒学习的对称交叉熵，载于：IEEE 国际计算机视觉会议论文集，2019年，第322–330页。[42] K. He, X. Zhang, S. Ren, J. Sun，面向图像识别的深度残差学习，载于：IEEE 计算机视觉与模式识别会议论文集，2016年，第770–778页。[43] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng，利用无监督特征学习识别自然图像中的数字，载于：NeurIPS 深度学习与无监督特征学习研讨会，2011年。[44] A. Krizhevsky, G. Hinton 等，学习小尺寸图像的多层特征，技术报告（2009）。[45]G. Hinton, O. Vinyals, J. Dean, http://arxiv.org/abs/1503.02531，载于：NeurIPS 深度学习与表征学习研讨会，2015年。网址：http://arxiv.org/abs/1503.02531[46]J. Latz，连续时间下随机梯度下降的分析，《统计与计算》31（4）（2021年）1–25。[47] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, L. Fei-Fei，Mentornet：在受损标签上为极深神经网络学习数据驱动课程，载于：国际机器学习会议，2018年，第2304–2313页。

V. S. Borkar, S. K. Mitter，随机递归算法的强近似定理，《优化理论与应用杂志》100（3）（1999）499–513。

[49] B. Øksendal，《随机微分方程》，载于：《随机微分方程》，施普林格，2003年，第65–84页。

命题1的证明

Proof 由于小批量 B_k 是从完整样本集 $\mathcal{D}$中随机、独立且均匀抽取的，因此对于 $\forall\theta\in\mathbb{R}^d$ 和 $\forall x_j\in B_k$，有

$$\mathbb{E}_{x_j\in B_k}\left[\nabla L_j^*(\theta)\right]=\frac{1}{N}\sum_{i=1}^{N}\nabla L_i^*(\theta)\Rightarrow\mathbb{E}_{B_k}\left[\frac{1}{|B_k|}\sum_{x_j\in B_k}\nabla L_j^*(\theta)\right]=\frac{1}{N}\sum_{i=1}^{N}\nabla L_i^*(\theta)\,. \tag{31}$$

由此，我们可以推导出

$$\begin{aligned}\mathbb{E}_{B_k}[\xi_k^*(\theta)]&=\mathbb{E}_{B_k}\sqrt{\eta}\left(\frac{1}{|B_k|}\sum_{i=1}^{k}\nabla_j^*(\theta)-\frac{1}{N}\sum_{i=1}^{N}L_i^*(\theta)\right)\\&=\sqrt{\eta}\left(\mathbb{E}_{B_k}\left[\frac{1}{|B_k|}\sum_{i=1}^{k}\nabla_j^*(\theta)\right]-\frac{1}{N}\sum_{i=1}^{N}L_i^*(\theta)\right)=\mathbf{0_d}\,.\end{aligned} \tag{32}$$

此外，对于一个 $\theta\in\mathbb{R}^d$，有

$$\begin{aligned}\mathrm{Var}[\xi_k^*(\theta)]&=\mathbb{E}_{B_k}\left(\xi_k^*(\theta)-\mathbb{E}_{B_k}\xi_k^*(\theta)\right)\left(\xi_k^*(\theta)-\mathbb{E}_{B_k}\xi_k^*(\theta)\right)^\top\\&=\mathbb{E}_{B_k}\left[\xi_k^*(\theta)\xi_k^*(\theta)^\top\right]\\&=\frac{\eta}{|B_k|^2}\sum_{x_j\in B_k}\mathrm{Var}[\nabla L_j^*(\theta)]\\&=\frac{\eta}{|B_k|N}\sum_{j=1}^{N}\left(\nabla L_j^*(\theta)-\frac{1}{N}\sum_{i=1}^{N}L_i^*(\theta)\right)\left(\nabla L_j^*(\theta)-\frac{1}{N}\sum_{i=1}^{N}L_i^*(\theta)\right)^\top\\&=\frac{\eta}{|B_k|}\Sigma_N^{\mathrm{SGD}}(\theta),\end{aligned} \tag{33}$$

其中 $\Sigma_N^{\mathrm{SGD}}(\theta)$ 定义为式（15）。

类似地，由于小批量 B_k 是从完整样本集 $\mathcal{D}$中随机、独立且均匀抽取的，因此对于 $\forall\theta\in\mathbb{R}^d$，有

$$\mathbb{E}_{x_j\in B_k}\left[\nabla_\theta f(x_j,\theta)\right]=\frac{1}{N}\sum_{i=1}^{N}\nabla_\theta f(x_i,\theta)\,. \tag{34}$$

因此有

$$\mathbb{E}_{B_k,\varepsilon_i}[\xi_k^{\mathrm{ULN}}(\theta)]=-\frac{\sqrt{\eta}}{|B_k|}\sum_{x_i\in B_k}\left\{\mathbb{E}_{x_i\in B_k}\left[\nabla f(x_i,\theta)\right]\cdot\mathbb{E}_{\varepsilon_i}[\varepsilon_i]\right\}=\mathbf{0_d}\,. \tag{35}$$

Again, by the definition, there has

$$\begin{aligned}\mathrm{Var}_{B_k,\varepsilon_i}[\xi_k^{\mathrm{ULN}}(\theta)]&=\mathbb{E}_{B_k,\varepsilon_i}\left[\xi_k^{\mathrm{ULN}}(\theta)\xi_k^{\mathrm{ULN}}(\theta)^\top\right]\\&\text{Let us assume } B_k \text{ and } \varepsilon_i \text{ for } 1\le i\le N \text{ are independent.}\\&=\frac{\eta}{|B_k|^2}\sum_{x_j\in B_k}\mathbb{E}_{B_k\varepsilon_i}\left[(\varepsilon_i\cdot\nabla_\theta f(x_i,\theta)-\mathbb{E}_{B_k\varepsilon_i}(\varepsilon_i\cdot\nabla_\theta f(x_i,\theta))^2\right]\\&\text{As } \mathrm{Var}[\varepsilon_i]=\sigma^2 \text{ for } 1\le i\le N, \text{ there has}\\&=\frac{\eta\sigma^2}{|B_k|N}\sum_{i=1}^{N}\nabla_\theta f(x_i,\theta)\nabla_\theta f(x_i,\theta)^\top=\frac{\eta\sigma^2}{|B_k|}\Sigma_N^{\mathrm{ULN}}(\theta)\,.\end{aligned} \tag{36}$$

其中 $\Sigma_N^{\mathrm{ULN}}(\theta)$ 定义为公式 (15)。 □

B 命题 2 的证明

Proof 我们证明，当 $\eta \to 0$ 时，公式 ($\bar{\theta}_k$ 17) 的离散迭代在强范数下、有限时间区间内接近随机微分方程的解。(16)。主要技术沿用 [48], , 但 [48] 仅考虑了以下情况: $\Sigma_N^{\mathrm{SGD}}(\theta)$ 和 $\Sigma_N^{\mathrm{ULN}}(\theta)$ 均为常数。

令 $C_1(\theta) = \sqrt{\frac{1}{B}\Sigma_N^{\mathrm{SGD}}(\theta)}$、$C_2(\theta) = \sqrt{\frac{1}{B}\Sigma_N^{\mathrm{ULN}}(\theta)}$ 以及 $L^*(\theta) = \frac{1}{N}\sum_{i=1}^N L_i^*(\theta)$。

令 $\widehat{\Theta}(t)$ 表示由随机微分方程的积分形式定义的过程

$$\begin{aligned}\widehat{\Theta}(t) - \widehat{\Theta}(0) = &-\int_0^t \nabla L^*(\widehat{\Theta}_{\lfloor \frac{s}{\eta} \rfloor \eta})\mathbf{d}s + \sqrt{\eta}\int_0^t C_1(\widehat{\Theta}(\lfloor \frac{s}{\eta} \rfloor \eta))\mathbf{d}W_1(s) \\ &+ \sqrt{\eta}\int_0^t C_2(\widehat{\Theta}(\lfloor \frac{s}{\eta} \rfloor \eta))\mathbf{d}W_2(s)\ ,\ \widehat{\Theta}(0) = \theta_0\ .\end{aligned} \tag{37}$$

对于正实数 $a > 0$，我们定义 $\lfloor a \rfloor$ =最大值 $\{k \in \mathbb{N}_+, k < a\}$。由式（37）可知，对于 $k = 0,\ 1,\ 2,$ .....

$$\begin{aligned}\widehat{\Theta}((k+1)\eta) - \widehat{\Theta}(k\eta) = &-\eta\nabla L^*(\widehat{\Theta}(k\eta)) - \sqrt{\eta}C_1(\widehat{\Theta}(k\eta))(W_1((k+1)\eta) - W_1(k\eta)) \\ &- \sqrt{\eta}C_2(\widehat{\Theta}(k\eta))(W_2((k+1)\eta) - W_2(k\eta)).\end{aligned} \tag{38}$$

由于 $\sqrt{\eta}(W_1((k+1)\eta) - W_1(k\eta)) \sim \mathcal{N}(0,\ \eta^2 I)$ 且 $\sqrt{\eta}(W_2((k+1)\eta) - W_2(k\eta)) \sim \mathcal{N}\ 0\ \eta^2 I$ $\eta z_{k+1} = \sqrt{\eta}\ W_1\ k+1\ \eta\ - W_1\ k\eta$ $z'_{k+1} = (,)$，我们可以令 $\sqrt{\eta}\ W_2\ k+1\ \eta\ - W_2\ k\eta$ z_{k+1} z'_{k+1} 和 $(((\ )))$ 和 $(((\ )))$，其中 和 是式（16）中的独立同分布高斯序列。由此可见,

$$\widehat{\Theta}(k\eta) = \bar{\theta}_k^{\mathrm{ULN}}\ , \tag{39}$$

其中 $\bar{\theta}_k^{\mathrm{ULN}}$是方程（17）的解。

接下来，我们尝试对式（ $\widehat{\Theta}_t$ ）中的 37）以及式（ $\Theta^{\mathrm{ULB}}(t)$ 中的（16）进行界定。最后，只需令 $\bar{\theta}_k^{\mathrm{ULN}} = \widehat{\Theta}(k\eta)$ ，即可得到 $\Theta^{\mathrm{ULN}}(k\eta)$ 和 $t = k\eta$）的误差估计。由于我们假设 $\nabla L_i^*(\theta)$ 和 $\nabla_\theta f(x,\theta)$ 是 L–Lipschitz 连续的，因此可得

$$\|C_1(\theta_1) - C_1(\theta_2)\|_2 = \sqrt{\frac{1}{bN}\sum_{i=1}^N \|\nabla L_i^*(\theta_1) - \nabla L_i^*(\theta_2)\|_2^2} \le L\|\theta_1 - \theta_2\|_2 \tag{40}$$

由于批量大小 $b \ge 1$。同理,

$$\|C_2(\theta_1) - C_2(\theta_2)\|_2 = \sqrt{\frac{\sigma^2}{bN}\sum_{i=1}^N \|\nabla_\theta f(x_i,\theta_1) - \nabla_\theta f(x_i,\theta_2)\|_2^2} \le \sigma L\|\theta_1 - \theta_2\|_2 \tag{41}$$

由于批量大小 $B \ge 1$。因此 $C_1(\theta)$ 和 C_2 均为 L–利普希茨连续。对 (37) 和 (16) 作差可得

$$\begin{aligned}\widehat{\Theta}(t) - \Theta^{\mathrm{ULN}}(t) = &-\int_0^t [\nabla L^*(\widehat{\Theta}(\lfloor \frac{s}{\eta} \rfloor \eta)) - \nabla L^*(\Theta^{\mathrm{ULN}}(s))]\mathbf{d}s \\ &+ \sqrt{\eta}\int_0^t [C_1(\widehat{\Theta}(\lfloor \frac{s}{\eta} \rfloor)) - C_1(\Theta^{\mathrm{ULN}}(s))]\mathbf{d}W_1(s) \\ &+ \sqrt{\eta}\int_0^t [C_2(\widehat{\Theta}(\lfloor \frac{s}{\eta} \rfloor)) - C_2(\Theta^{\mathrm{ULN}}(s))]\mathbf{d}W_2(s).\end{aligned} \tag{42}$$

我们可以估计

$$\begin{aligned}
&\|\nabla L^*(\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)) - \nabla L^*(\Theta^{\mathrm{ULN}}(s))\|_2^2 \\
\leq& 2\|\nabla L^*(\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)) - \nabla L^*(\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta))\|_2^2 + 2\|\nabla L^*(\Theta(\lfloor\frac{s}{\eta}\rfloor\eta)) - \nabla L^*(\Theta^{\mathrm{ULN}}(s))\|_2^2 \\
\leq& 2L^2\|\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta) - \Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)\|_2^2 + 2L^2\|\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta) - \Theta^{\mathrm{ULN}}(s)\|_2^2 ,
\end{aligned} \tag{43}$$

其中我们使用了由 L-Lipschitz 条件推导出的不等式。同样地，我们估计 e

$$\begin{aligned}
&\|C_1(\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)) - C_1(\Theta^{\mathrm{ULN}}(s))\|_2^2 \\
\leq& 2\|C_1(\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)) - C_1(\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta))\|_2^2 + 2\|C_1(\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)) - \Theta^{\mathrm{ULN}}(s))\|_2^2 \\
\leq& 2L^2\|\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta) - \Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)\|_2^2 + 2L^2\|\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta) - \Theta^{\mathrm{ULN}}(s))\|_2^2 .
\end{aligned} \tag{44}$$

同样地，基于由 L-Lipschitz 条件推导出的不等式，我们还可以得到

$$\begin{aligned}
&\|C_2(\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)) - C_2(\Theta^{\mathrm{ULN}}(s))\|_2^2 \\
\leq& 2\|C_2(\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)) - C_2(\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta))\|_2^2 + 2\|C_2(\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)) - \Theta^{\mathrm{ULN}}(s))\|_2^2 \\
\leq& 2\sigma^2L^2\|\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta) - \Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)\|_2^2 + 2\sigma^2L^2\|\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta) - \Theta^{\mathrm{ULN}}(s))\|_2^2 .
\end{aligned} \tag{45}$$

另一方面，根据式（42），由伊藤等距定理 [49] 和柯西–施瓦茨不等式可得

$$\begin{aligned}
\mathbb{E}|\widehat{\Theta}(t) - \Theta^{\mathrm{ULN}}(t)|^2 \leq& 2\mathbb{E}\left\|\int_0^t [\nabla L(^*(\Theta(\lfloor\frac{s}{\eta}\rfloor\eta)) - \nabla L^*(\Theta^{\mathrm{ULN}}(s))]\mathbf{d}s\right\|_2^2 \\
&+ 2\eta\mathbb{E}\left\|\int_0^t [C_1(\widehat{\Theta}_{\lfloor\frac{s}{\eta}\rfloor}) - C_1(\Theta^{\mathrm{ULN}}(s))]\mathbf{d}W_1(s)\right\|^2 \\
&+ 2\eta\mathbb{E}\left\|\int_0^t [C_2(\widehat{\Theta}_{\lfloor\frac{s}{\eta}\rfloor}) - C_2(\Theta^{\mathrm{ULN}}(s))]\mathbf{d}W_1(s)\right\|^2 \\
\leq& 2\int_0^t \mathbb{E}\left\|\nabla L(^*(\Theta(\lfloor\frac{s}{\eta}\rfloor\eta)) - \nabla L^*(\Theta^{\mathrm{ULN}}(s))\right\|_2^2 \mathbf{d}s \\
&+ 2\eta\int_0^t \mathbb{E}\left\|C_1(\widehat{\Theta}_{\lfloor\frac{s}{\eta}\rfloor}) - C_1(\Theta^{\mathrm{ULN}}(s))\right\|^2 \mathbf{d}s \\
&+ 2\eta\int_0^t \mathbb{E}\left\|C_2(\widehat{\Theta}_{\lfloor\frac{s}{\eta}\rfloor}) - C_2(\Theta^{\mathrm{ULN}}(s))\right\|^2 \mathbf{d}s .
\end{aligned} \tag{46}$$

结合式（43）、（44）、（45）和（46），可得

$$
\begin{aligned}
&\mathbb{E}\|\widehat{\Theta}(t)-\Theta^{\mathrm{ULN}}(t)\|_2^2 \\
&\le 2\int_0^t\left(2L^2\mathbb{E}\|\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)-\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)\|_2^2+2L^2\mathbb{E}\|\Theta_{\lfloor\frac{s}{\eta}\rfloor\eta}-\Theta^{\mathrm{ULN}}(s))\|_2^2\right)\mathbf{d}s \\
&+2\eta\int_0^t\left(2L^2\mathbb{E}\|\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)-\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)\|_2^2+2L^2\mathbb{E}\|\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)-\Theta^{\mathrm{ULN}}(s))\|_2^2\right)\mathbf{d}s \\
&+2\eta\int_0^t\left(2\sigma^2L^2\mathbb{E}\|\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)-\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)\|_2^2+2\sigma^2L^2\mathbb{E}\|\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)-\Theta^{\mathrm{ULN}}(s))\|_2^2\right)\mathbf{d}s\ . \\
=&4(1+\eta+\eta\sigma^2)L^2\cdot \\
&\left(\int_0^t\mathbb{E}\|\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)-\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)\|_2^2\mathbf{d}s+\int_0^t\mathbb{E}\|\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)-\Theta^{\mathrm{ULN}}(s))\|_2^2\mathbf{d}s\right)\ .
\end{aligned}
\tag{47}
$$

由于我们假设存在一个 $M>0$，使得 $\max_{i=1,\ 2.....,\ N}(|\nabla L_i^*(\theta)|)\le M$，因此我们得出 $|\nabla L^*(\theta)|\le\frac{1}{N}\sum_{i=1}^N|\nabla L_i^*(\theta)|\le M$ 且

$$
\|C_1(\theta)\|_2\le\frac{1}{\sqrt{b}}\sqrt{\frac{1}{N}\sum_{i=1}^N\|\nabla L_i^*(\theta)\|_2^2}\le M, \tag{48}
$$

$$
\|C_2(\theta)\|_2\le\frac{1}{\sqrt{b}}\sqrt{\frac{\sigma^2}{N}\sum_{i=1}^N\|\nabla_\theta f(x_i,\theta)\|_2^2}\le\sigma M \tag{49}
$$

因为 $B\ge1$。根据公式（16），伊藤等距定理 [49], 柯西-施瓦茨不等式和 $0\le s-\lfloor\frac{s}{\eta}\rfloor\eta\le\eta$ 我们知道

$$
\begin{aligned}
&\mathbb{E}\|\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)-\Theta^{\mathrm{ULN}}(s))\|_2^2 \\
=&\mathbb{E}\left\|-\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^s\nabla L^*(\Theta_u)\mathbf{d}u+\sqrt{\eta}\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^sC_1(\Theta_u)\mathbf{d}W_1(u)+\sqrt{\eta}\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^sC_2(\Theta_u)\mathbf{d}W_2(u)\right\|_2^2 \\
\le&2\mathbb{E}\left\|\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^s\nabla L^*(\Theta_u)\mathbf{d}u\right\|_2^2+2\eta\mathbb{E}\left\|\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^sC_1(\Theta_u)\mathbf{d}W_1(u)\right\|_2^2+2\eta\mathbb{E}\left\|\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^sC_2(\Theta_u)\mathbf{d}W_2(u)\right\|_2^2 \\
\le&2\mathbb{E}\left(\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^s\left\|\nabla L^*(\Theta_u)\right\|\mathbf{d}u\right)^2+2\eta\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^s\mathbb{E}\|C_1(\Theta_u)\|_2^2\mathbf{d}u+2\eta\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^s\mathbb{E}\|C_2(\Theta_u)\|_2^2\mathbf{d}u \\
\le&2\eta\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^s\mathbb{E}|\nabla L^*(\Theta_u)|^2\mathbf{d}u+2\eta\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^s\mathbb{E}\|C_1(\Theta_u)\|_2^2\mathbf{d}u+2\eta\int_{\lfloor\frac{s}{\eta}\rfloor\eta}^s\mathbb{E}\|C_2(\Theta_u)\|_2^2\mathbf{d}u \\
\le&2\eta^2M^2+2\eta^2M^2+2\eta^2\sigma^2M^2=(4+2\sigma^2)\eta^2M^2\ .
\end{aligned}
\tag{50}
$$

结合式（50）和（47），可得

$$
\begin{aligned}
&\mathbb{E}\|\widehat{\Theta}(t)-\Theta^{\mathrm{ULN}}(t)\|_2^2 \\
\le&4(1+\eta+\eta\sigma^2)L^2\cdot\left(\int_0^t\mathbb{E}\|\widehat{\Theta}(\lfloor\frac{s}{\eta}\rfloor\eta)-\Theta^{\mathrm{ULN}}(\lfloor\frac{s}{\eta}\rfloor\eta)\|_2^2\mathbf{d}s+(4+2\sigma^2)\eta^2M^2t\right)\ .
\end{aligned}
\tag{51}
$$

令 $T>0$ 和 $m(t)=\max_{0\le s\le t}\mathbb{E}|\widehat{\Theta}_s-\Theta^{\mathrm{ULN}}(s))|^2$，注意到 $m(\lfloor\frac{s}{\eta}\rfloor\eta)\le m(s)$（因为 $\lfloor\frac{s}{\eta}\rfloor\eta\le s$），则上述结果对任意 $0\le t\le T$ 均成立，

$$m(t)\le 4(1+\eta+\eta\sigma^2)L^2\cdot\left(\int_0^t m(s)\mathbf{d}s+(4+2\sigma^2)\eta^2M^2T\right). \tag{52}$$

由格朗沃尔不等式可得，对于 $0\le t\le T$,

$$m(t)\le 4(1+\eta+\eta\sigma^2)(4+2\sigma^2)L^2\eta^2M^2Te^{4(1+\eta+\eta\sigma^2)L^2t}. \tag{53}$$

假设 $0<\eta<1$，则存在一个与 η 无关的常数 C，使得

$$\mathbb{E}\|\widehat{\Theta}(t)-\Theta^{\mathrm{ULN}}(t)\|_2^2\le m(t)\le C\eta^2. \tag{54}$$

令 $t=k\eta$ 代入（54）并利用（39），证明完成。 □

C 命题3的证明

Proof 为得到式（22），我们可以使用简单的向量-矩阵-向量乘积变换。对于随机向量 v 和对称矩阵 A，有 $\mathbb{E}_v[v^\top Av]=$ 迹（$A\mathbb{E}[vv^\top]$），使得

$$\begin{aligned}
\mathbb{E}_{z_k}\|\xi_k^{\mathrm{ULN}}(\theta)\|_2^2&=\mathbb{E}_{z_k}\left[\xi_k^{\mathrm{ULN}}(\theta)^\top\xi_k^{\mathrm{ULN}}(\theta)\right]\\
&\approx\frac{\eta\sigma^2}{b}\mathbb{E}_{z_k}\left[z_k^\top\left(\frac{1}{N}\sum_{i=1}^N\nabla_\theta f(x_i,\theta)\nabla_\theta f(x_i,\theta)^\top\right)z_k\right]\\
&=\frac{\eta\sigma^2}{b}\mathrm{trace}\left(\left(\frac{1}{N}\sum_{i=1}^N\nabla_\theta f(x_i,\theta)\nabla_\theta f(x_i,\theta)^\top\right)\mathbb{E}_{z_k}[z_kz_k^\top]\right)\\
&\text{as }\mathbb{E}_{z_k}[z_kz_k^\top]=\mathbf{I}_d\ for\ z_k\sim\mathcal{N}(\mathbf{0}_d,\mathbf{I}_d)\\
&=\frac{\eta\sigma^2}{b}\mathrm{trace}\left(\frac{1}{N}\sum_{i=1}^N\nabla_\theta f(x_i,\theta)\nabla_\theta f(x_i,\theta)^\top\right)\\
&=\frac{\eta\sigma^2}{bN}\sum_{i=1}^N\|\nabla_\theta f(x_i,\theta)\|_2^2.
\end{aligned} \tag{55}$$

□