arXiv:2603.01842v1 [cs.NE] 2 Mar 2026

通过运输不等式实现两层神经网络的时间一致集中

Arnaud Guillin*[1]、Boris Nectoux†[1] 和 Paul Stos‡[1]

[1]克莱蒙奥弗涅大学，法国国家科学研究中心 UMR 6620，*LMBP*，克莱蒙费朗，法国 ce

摘要

对于采用二次损失和岭正则化、通过随机梯度下降（SGD）训练的两层神经网络，我们以高概率量化其预测与平均场极限之间的差异度，并且这一结果在时间上保持一致。作为关键步骤，我们为SGD参数的分布建立了 T_p 运输不等式（$p \in \{1,2\}$），其常数明确且与迭代次数无关。随后，我们证明经验参数测度在Wasserstein距离 W_1，下时间一致地集中于其平均场极限周围，并将这些界转化为相对于固定测试函数 φ的预测误差估计。我们还推导出切片Wasserstein距离 SW_1，下的类似集中界，从而得到与维数无关的速率。

1 引言

神经网络已成为解决各种高维预测问题的标准工具 [21, 24, 39]，，其在实践中的成功也激发了人们对理解其行为背后理论机制的浓厚兴趣[44, 40, 1]。

本文聚焦于回归任务的经典监督学习框架。我们观测到一系列独立数据点（x_k, y_k）$\in \mathbb{R}^d \times \mathbb{R}$，它们来自同一个（未知）分布 π。目标是学习一个预测器$\hat{}y$，用于预测标签 y，并能够泛化到此前未见过的特征 x。对于宽度为 N的单隐藏层神经网络，该预测器由 $\theta = (\theta^1, \dots, \theta^N) \in \mathcal{E} = (\mathbb{R}^D)^N$ 参数化，其形式为

$$\hat{y}_\theta(x) = \frac{1}{N}\sum_{i=1}^{N} \sigma_*(\theta^i, x),$$

其中 $\sigma_*\colon \mathbb{R}^D \times \mathbb{R}^d \to \mathbb{R}$ 是一个激活函数。通常，$\theta^i = (a_i, b_i, w_i) \in \mathbb{R}^{d+2}$ 且$\sigma_*(\theta^i, x) = a_i\sigma($ $w_i \cdot x + b_i)$，其中 $\sigma\colon \mathbb{R} \to \mathbb{R}$（例如，Sigmoid函数、双曲正切函数、ReLU）。因子 $1/N$ 是网络输出[15]的标准平均场归一化；引入它是为了方便，也可以通过简单地重新定义激活函数将其吸收。

原则上，我们希望选择 θ，使总体风险最小化

$$\mathcal{R}(\theta) = \mathbb{E}_\pi\left[\ell\big(y, \hat{y}_\theta(x)\big)\right],$$

其中 $\ell\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ 是一个损失函数。对于回归问题，为了简单起见，我们使用平方损失 ℓ $(y, \hat{}y) = \frac{1}{2}(y - \hat{}y)^2$，不过也可以考虑其他损失函数。

在实践中，由于我们无法直接获取数据分布 π，一种标准方法$\hat{}$是通过优化算法（如随机梯度下降（SGD）或其变体 [7, 8]）最小化由观测样本构建的经验代理 $\mathcal{R}$。

*arnaud.guillin@uca.fr
†boris.nectoux@uca.fr
‡paul.stos@uca.fr

在当前设定下，参数从 $\mu_0 \in \mathcal{P}(\mathbb{R}^D)$独立同分布地初始化，并按照固定学习率 $\alpha > 0$的在线随机梯度下降进行更新：对每个 $i \in \{1, \dots, N\}$,

$$\theta_{k+1}^i = (1 - \gamma\lambda)\theta_k^i + \gamma\big(y_{k+1} - \hat{y}_{\theta_k}(x_{k+1})\big)\nabla_{\theta^i}\sigma_*(\theta_k^i, x_{k+1}), \tag{1}$$

其中 $\gamma = \alpha/N$。项 $-\gamma\lambda$ 对应于对经验风险进行岭正则化

$$\widehat{\mathcal{R}}_\lambda(\theta, x, y) = \frac{1}{2}\big(y - \hat{y}_\theta(x)\big)^2 + \frac{\lambda}{2}\|\theta\|_2^2$$

这将有助于我们下面的分析。

在在线设定中，每个数据点只使用一次，不会再次使用。这种理想化设定是学习文献中的标准做法；当数据量充足或以流式方式到达时，这种设定尤为适用，此时反复使用一个固定的有限数据集并非主要现象 [7, 37]。

即使是在通过随机梯度下降训练的基础两层架构中，为神经网络的训练动力学提供定量保证仍然很有挑战性。这种困难不仅源于高维性，也源于学习的反馈结构：在每次迭代中，每个参数都会通过一个取决于当前网络输出的梯度进行更新，而该输出汇总了所有 N 神经元的贡献，并在随机数据样本上进行评估。因此，随机梯度下降会在参数空间中产生强耦合的随机演化 $\mathcal{E}$。

对于宽两层网络，一种尤为有效的方法是将训练建模为大型相互作用粒子系统，并通过神经元参数的经验分布 [36], 或其经时间重标度的版本 [15, 16], 来追踪该系统，该版本对所有 $t \geq 0$ 均有定义，其定义如下：

$$\mu_t^N = \frac{1}{N}\sum_{i=1}^N \delta_{\theta^i_{\lfloor Nt \rfloor}} \in \mathcal{P}(\mathbb{R}^D).$$

随着神经元数量 $N \to \infty$增加，该经验过程会依概率（在适当的拓扑下）收敛到确定性演化 $(\bar{\mu}_t)_{t\geq 0}$ ，其特征是一个非线性测度值方程的解 [27, 42, 16]（见（13））。这一渐近描述给出了一个更易处理的学习动力学宏观模型，并已被证明有助于研究涨落 [41, 16] 以及分析稳定性和长时间行为 [27, 10]。

现有的定量有限宽度界通常是在固定时间范围 $[0, T]$内推导出来的，且随着 T 增大，常数会变差（通常呈指数级）；例如，参见 [27, 定理 3]。这在实践中并不理想，因为学习算法通常需要运行许多次迭代。

在本工作中，我们通过在对 $t \geq 0$ 一致的意义下，量化经验参数测度 μ_t^N与其平均场极限 $\bar{\mu}_t$ 之间的偏差，来弥补这一空缺；这里假设存在由岭正则化诱导的显式收缩性条件。更具体地说，我们研究三种互补的差异度量：

- $\Delta^{N,\ t}{}^{\varphi} = \big|\langle\varphi,\ \mu_t^N\rangle - \langle\varphi,\ \bar{\mu}_t\rangle\big|$ 对于一个固定的1-利普希茨测试函数 φ： $\mathbb{R}^D \to \mathbb{R}$。特别地，假设激活函数满足利普希茨条件（假设(A1)如下），对于任意固定输入 $x \in \mathbb{R}^d$，取 $\varphi = \sigma_*(\cdot, x)$，即可将关于 $\Delta^{N,\ t}{}^{\varphi}$ 的任意界转化为网络输出的界。
- $\Delta_t^N = \mathsf{W}_1(\mu_t^N,\ \bar{\mu}_t)$，其中$\mathsf{W}_1$ 是经典的$1-Wasserstein$ 距离，定义在 $\mathcal{P}_1(\mathbb{R}^D)$ 上（这是定义在 $\mathbb{R}^D$ 上的、具有有限一阶矩的博雷尔概率测度集合）。回顾W_1 的 Kantorovich–Rubinstein 对偶公式（例如，参见[46, 注 6。5]）：

$$\mathsf{W}_1(\mu, \nu) = \sup_{\|\psi\|_{\mathrm{Lip}} \leq 1} \big|\langle\psi, \mu\rangle - \langle\psi, \nu\rangle\big|. \tag{2}$$

  因此， Δ_t^N同时控制 $\Delta^{N,\ t}{}^{\varphi}$ 对于所有1-利普希茨 φ。
- $\Delta^{N,\ \text{其中}\ t} = \mathsf{SW}_1(\mu_t^N, \bar{\mu}_t)$，其中 SW_1 是 1–切片 $Wasserstein$ 距离，定义为

$$\mathsf{SW}_1(\mu, \nu) = \int_{\mathbb{S}^{D-1}} \mathsf{W}_1\big(P_{u\#}\mu, P_{u\#}\nu\big)\,\zeta(du), \qquad P_u(w) = u \cdot w,$$

其中 ζ 是单位球面上的均匀概率测度 $\mathbb{S}^{D-1}$。切片类比 SW_1在高维中得到广泛应用，且计算上易于处理，[6, 34, 29] 因为它将比较简化为一维投影，同时仍能探测丰富的可观测量族（所有线性投影）。因此， $\Delta^{N,\mathsf{s}}_t$ 在针对性很强的差异度 $\Delta^{N,\varphi}_t$ 与全局度量 Δ^N_t之间实现了有益的折衷。

当具体选择无关紧要时，我们记作

$$\Delta^{N,\circ}_t \in \{\Delta^{N,\varphi}_t, \Delta^N_t, \Delta^{N,\mathsf{s}}_t\}.$$

关于 Wasserstein 距离和最优传输的背景知识，我们参见 [38, 46, 34]（另见 [33]，其中介绍更侧重于机器学习）；关于切片 Wasserstein 距离及相关切片散度，我们参见 [6, 28, 29]。

我们的分析分为两个步骤。首先，在初始化分布μ_0 和数据分布 π满足正则性条件的前提下，我们为第 k 次迭代时参数向量的分布 $\nu_k = \mathcal{L}(\theta_k)$ 建立一个 T_p 传输不等式 [22] ($p \in \{1, 2\}$)，其常数具有显式形式且与 k 无关。这正是命题 1 的内容。对于 $p = 1$ 和 $p = 2$，这意味着参数的中心化利普希茨可观测量（因而也包括 μ^N_t）具有次高斯尾部 (7)，且相应常数不会随时间恶化。将这一步应用于 μ^N_t 的适当利普希茨泛函，可得到我们关注的三种差异相对于均值的高概率偏差界（见推论1）。随后，我们通过同步耦合比较相互作用的 SGD 粒子与独立同分布的平均场系统，从而对 $t \geq 0$ 上的剩余偏差项 $\mathbb{E}[\Delta^{N,\circ}_t]$ 建立一致界。由此可将误差分解为两部分：沿轨迹累积的动态误差，通过标准的混沌传播论证处理 [45]（命题 3 和推论 2）；以及经验测度的静态独立同分布抽样误差，通过 Wasserstein 距离下经典的经验收敛率控制 [20]（命题 4 和命题 5）。结合这两个步骤，我们得到一个围绕平均场极限的非渐近时间一致集中界。这就是我们的主要结果，见定理 1。这些结果是在激活函数 σ_* 满足标准有界性和 Lipschitz 假设的条件下建立的。最后一节将说明如何放宽有界性要求：只需允许激活函数在数据支撑集上关于参数变量至多线性增长，即可通过局部化涵盖更一般的 ReLU型激活函数。

相关工作。宽两层神经网络在线训练的平均场近似在 [27], 中被引入，其中导出了极限动力学，并给出了固定时间范围上的有限宽度界 $[0, T]$。经验参数测度在任意固定时间区间上的轨迹大数定律与中心极限定理在 [42, 41],中得到证明，随后在小批量处理与含噪情形下被推广到完整轨迹层面，见 [16]。关于平均场动力学的补充分析，包括各种结构性假设下的稳定性与长时间行为，我们参见 [36, 10, 27]。我们还提及在 [11]中发展的（连续时间）随机梯度下降动态的定量混沌传播分析。

除固定时间范围上的界之外，要获得时间一致控制，通常需要某种形式的耗散性（例如在适当度量下的收缩性），以防止误差累积；例如，参见弱相互作用扩散设定中的 [18, 23, 14]。相关粒子系统（包括异质平均场相互作用）的 Wasserstein 波动的时间一致指数界已在 [2], 中建立，从而推广了经验测度的基于传输的集中估计 [5]；另见 [13]的一般平均场粒子模型框架。

在另一方向上，一些研究将平均场思想拓展至深度架构和其他无限宽度区域。在 [43, 30, 31, 19] 中，研究者为多层网络建立了严格的平均场形式体系。针对 Transformer 架构，也有人提出了平均场视角，其中词元动力学可以被解释为相互作用粒子系统 [35, 9]。

从更具几何性的视角来看，Fournier 和 Guillin 的经典独立同分布 Wasserstein 采样误差界 [20] 近期在 [12], 中得到了进一步改进：该研究以支撑集的覆盖维数取代环境维数，从而获得了类似的最优收敛率。这

为我们所研究的情形指明了一个潜在的未来研究方向：如果平均场极限集中在参数空间的某个低维子集上，那么识别一个有效支撑集，并估计（或给出其覆盖维数的上界），就可能得到严格更优的收敛率。

在准备本稿件期间，我们了解到 [47]，该工作为一般的 McKean 型非线性马尔可夫链建立了定量的混沌传播界，并讨论了附加结构性假设下的长时间行为。尽管两项工作的设置和技术有所不同，但它们都支持这样一种观点：一致时间的混沌传播从根本上与潜在的压缩结构相关。

本文其余部分的组织如下。我们在第 2 节陈述基本假设和主要结果。第 3 节包含证明。最后，第 4 节将前面的结果推广到线性增长的激活函数。

记号。 所有随机变量都定义在同一个概率空间 $(\Omega\ \mathcal{F}, \mathbb{P})$ 上。对于随机变量 Z，$\mathcal{L}(Z)$ 表示 Z 的分布。我们用 $|\cdot|$ 既表示 $\mathbb{R}$ 上的绝对值，也表示 $\mathbb{R}^d$ 或 $\mathbb{R}^D$ 上的欧几里得范数。$\mathbb{R}^D$ 上的数量积记作 $v \cdot w$。我们将 $|\cdot|_\infty$ 记作 $\mathcal{Z} = \mathbb{R}^d \times \mathbb{R}$ 上的上确界范数，并将 $\|\cdot\|_p (p \in \{1, 2\})$ 记作 $\mathcal{E} = (\mathbb{R}^D)^N$ 上的通常 ℓ^p 乘积范数：

$$\|\theta\|_p = \Big(\sum_{i=1}^N |\theta^i|^p\Big)^{1/p}, \qquad \theta = (\theta^1, \dots, \theta^N) \in \mathcal{E}.$$

给定度量空间 $(\mathcal{X}, \mathsf{d})$ 和函数 $f\colon \mathcal{X} \to \mathbb{R}$，我们定义（只要其有限）

$$\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|, \qquad \|f\|_{\mathrm{Lip}} = \sup_{x \neq y} \frac{|f(x) - f(y)|}{\mathsf{d}(x, y)}.$$

对于可测的 $\varphi\colon \mathbb{R}^D \to \mathbb{R}$ 和 $\mu \in \mathcal{P}(\mathbb{R}^D)$，我们使用括号记号

$$\langle \varphi, \mu \rangle = \int_{\mathbb{R}^D} \varphi \, d\mu.$$

特别地，对于任意输入 $x \in \mathbb{R}^d$ 和 $t \geq 0$，有

$$\hat{y}_{\theta_{\lfloor Nt \rfloor}}(x) = \langle \sigma_*(\cdot, x), \mu_t^N \rangle.$$

在本文中，$C > 0$ 表示一个通用常数（与 N、t 和 p无关），其值可能随行而变化。

2 主要结果

2.1 假设

我们定义以下一组假设。

(A1) 激活函数 $\sigma_* \in \mathcal{C}^1(\mathbb{R}^D \times \mathbb{R}^d)$; 此外，$\sigma_*$ 一致有界，且混合梯度有界，即

$$\|\sigma_*\|_\infty \leq B, \qquad \|\nabla_{(\theta^i, x)} \sigma_*\|_\infty \leq M.$$

(A2) 标签几乎必然有界：$|y| \leq A$ π-a.s.

(A3) 对每个 $\theta^i \in \mathbb{R}^D$，映射 $x \in \mathbb{R}^d \mapsto \nabla_{\theta^i} \sigma_*(\theta^i, x)$ 关于 θ^i 一致地是 L_x-利普希茨的；对每个 $x \in \mathbb{R}^d$，映射 $\theta^i \in \mathbb{R}^D \mapsto \nabla_{\theta^i} \sigma_*(\theta^i, x)$ 关于 x 一致地是 L_θ-利普希茨的。

假设 (A1)–(A3) 在学习文献中是标准的：例如， [27] 假设有界激活函数和标签以及有界利普希茨漂移项，并对 $\nabla_{\theta^i}\sigma_*(\theta^i,\cdot)$ 采用次高斯控制，以处理高维输入。我们的条件略为更直接（一致有界性和利普希茨常数），从而使稳定性常数得以显式给出，并保持后续集中论证的透明性。主要局限在于 (A1) 排除了非光滑或无界激活函数，例如 ReLU。该条件将在最后的第 4 节中放宽。

在这些假设下，我们汇总以下常数，供下文反复使用。

$$K=\alpha\big((A+B)L_x+M^2+M\big),\qquad \lambda_\star=(A+B)L_\theta+M^2,\qquad L_N=|1-\gamma\lambda|+\gamma\lambda_\star,$$
$$C_\star=8M^4+4(\lambda+(A+B)L_\theta)^2,\qquad N_\star=\left\lceil\frac{4\alpha C_\star}{\lambda-\lambda_\star}\right\rceil. \tag{3}$$

2.2 随机梯度下降动态的运输不等式

运输不等式将相对熵与最优传输联系起来，并为证明 Lipschitz 可观测量的高斯型集中提供了直接途径。在这里，它们构成了我们分析的第一步：我们证明参数向量在第 $\nu_k=\mathcal{L}(\theta_k)$ 次迭代时的分布k 满足一个 T_p 传输不等式（针对 $p\in\{1,2\}$），其常数与 k无关（命题1），进而得到随机梯度下降马尔可夫链上的一致次高斯集中（推论1）。下面我们简要回顾 T_p 的定义以及我们将用到的几个性质；关于这类不等式及其测度集中推论的一般背景，可参见 [25, 26, 22] 及其中的参考文献。

给定一个波兰度量空间 $(\mathcal{X},\mathsf{d})$ 和 $p\geq 1$，若概率测度 $\mu\in\mathcal{P}(\mathcal{X})$ 在 $(\mathcal{X},\mathsf{d})$ 上满足常数为 $C>0$ 的 L^p–传输 不等式，我们记作 $\mu\in T_p(C)$；若对所有 $\nu\in\mathcal{P}(\mathcal{X})$,

$$\mathsf{W}_p(\nu,\mu)\leq\sqrt{2C\,H(\nu\,|\,\mu)}, \tag{4}$$

其中，W_p 表示 $\mathcal{P}_p(\mathcal{X})$上的通常 p-*Wasserstein* 距离（即具有有限 p阶矩的博雷尔概率测度集合），并且

$$H(\nu\,|\,\mu)=\begin{cases}\displaystyle\int_{\mathcal{X}}\log\frac{d\nu}{d\mu}\,d\nu & \text{if } \nu\ll\mu\\ +\infty & \text{otherwise}\end{cases}$$

是 ν 相对于 μ的相对熵。

在 $p\geq 1$的所有取值中， $p=1$ 和 $p=2$ 两种情形最受关注。对于T_1，Bobkov–Götze定理 [4] 给出了（4）与高斯集中之间的等价关系：$\mu\in T_1(C)$ 在 $(\mathcal{X},\mathsf{d})$上成立，当且仅当对每个利普希茨 f：$(\mathcal{X},\mathsf{d})\to\mathbb{R}$， f是 μ–可积的，并且

$$\forall\xi\in\mathbb{R},\quad \mathbb{E}_\mu\big[e^{\xi(f-\mathbb{E}_\mu[f])}\big]\leq\exp\left(\frac{C\xi^2}{2}\|f\|_{\mathrm{Lip}}^2\right). \tag{5}$$

特别地，若 μ的支集是紧的，例如$\operatorname{supp}\mu\subset B(x_0,R)$对某个 $x_0\in\mathcal{X}$成立，且$R>0$，则（5）可直接由Hoeffding引理推出，从而有 $\mu\in T_1(R^2/4)$。更一般地， T_1 也等价于关于$\mathsf{d}(x,x_0)^2$, 的指数可积条件，参见 [17,定理23]。根据Chernoff方法，（5）进而给出次高斯尾界

$$\forall r>0,\quad \mathbb{P}\big(|f(X)-\mathbb{E}[f(X)]|\geq r\big)\leq 2\exp\left(-\frac{r^2}{2C\|f\|_{\mathrm{Lip}}^2}\right),\qquad X\sim\mu. \tag{6}$$

注意， $T_2(C)$ 比 $T_1(C)$ 更强（因为 $\mathsf{W}_1\leq\mathsf{W}_2$），因此也蕴含利普希茨泛函的高斯集中。与 T_1 不同，仅有紧支撑通常并不能保证 T_2；证明 T_2 通常需要额外的结构，例如对数 Sobolev 不等式 [32]。

我们将反复使用 T_p 在利普希茨推前 [17, 引理 2.1] 下的稳定性：若$\mu \in T_p(C)$ 在 $(\mathcal{X}, \mathsf{d})$ 上成立，且 $\Psi : (\mathcal{X}, \mathsf{d}) \to (\mathcal{Y}, \mathsf{d}')$ 是 L-利普希茨的，则 $\Psi_{\#}\mu \in T_p(L^2 C)$ 在 $(\mathcal{Y}, \mathsf{d}')$ 上成立。

我们还将用到以下独立同分布张量化恒等式。对于 $n \geq 1$，在 $\mathcal{X}^n$ 上赋予乘积度量

$$\mathsf{d}_1(x, y) = \sum_{i=1}^{n} \mathsf{d}(x_i, y_i), \qquad \mathsf{d}_2(x, y) = \Big(\sum_{i=1}^{n} \mathsf{d}(x_i, y_i)^2 \Big)^{1/2}.$$

若 $(\mathcal{X}, \mathsf{d})$ 上的 $\mu \in T_1(C)$，则 $(\mathcal{X}^n, \mathsf{d}_1)$ 上的 $\mu^{\otimes n} \in T_1(nC)$；若 $(\mathcal{X}, \mathsf{d})$ 上的 $\mu \in T_2(C)$，则 $(\mathcal{X}^n, \mathsf{d}_2)$ 上的 $\mu^{\otimes n} \in T_2(C)$；参见命题 2.3] 以及命题 1.3] 之后的注。

以下所有陈述对 $p \in \{1, 2\}$ 的任一选择均成立。为了沿 SGD 马尔可夫链传播 T_p 不等式，自然需要假设初始化分布和数据分布满足 T_p 不等式：

($\mathbf{A4}_p$) 初始化分布 μ_0 在 $(\mathbb{R}^D, |\cdot|)$ 上满足 $\mu_0 \in T_p\big(C_0^{(p)}\big)$，数据分布 π 在 $(\mathbb{R}^d \times \mathbb{R}, |\cdot|_\infty)$ 上满足 $\pi \in T_p\big(C_\pi^{(p)}\big)$。

在假设 ($A4_p$) 和满足单步压缩条件 $L_N < 1$ （见（3））的情况下，我们可以仿照 T_p 定理 2[17, 和5]定理 1 [3, ，对随机梯度下降迭代一致地传递 2]不等式。

命题 1。 固定 $p \in \{1, 2\}$。假设 *(A1)–(A3)*、*($A4_p$)*以及 $L_N < 1$。那么对所有 $k \in \mathbb{N}$，$\nu_k \in T_p\big(C_N^{(p)}\big)$ 在 $(\mathcal{E}, \ \|\cdot\|_p)$上，取如下显式常数

$$C_N^{(1)} = N C_0^{(1)} + \frac{K^2}{(1 - L_N^2)} C_\pi^{(1)}, \qquad C_N^{(2)} = \frac{C_0^{(2)} \vee K^2 \, C_\pi^{(2)} N^{-1}}{(1 - L_N)^2}.$$

注。在我们的缩放 $\gamma = \alpha/N$下，通常当 $\gamma\lambda_\star < 1$ 足够大时，N 条件 $L_N < 1$ 对应于正则化系数的可行区间 $\lambda_\star < \lambda < 2\gamma^{-1} - \lambda_\star$。下界 $\lambda_\star < \lambda$ 保证正则化项压过相互作用项，使随机梯度下降更新在参数变量上具有 收缩性 （见下文引理 1 (ii)）。上界 $\lambda < 2\gamma^{-1} - \lambda_\star$ 是（1）的离散化伪影，对于宽网络而言基本不构成约束。

作为命题1 （确实可参见（6））的直接推论，对于任意利普希茨 f：（$\mathcal{E}$，$\|\cdot\|_p$）$\to \mathbb{R}$和 $p \in \{1, 2\}$，对每个 $k \in \mathbb{N}$，均有

$$\forall r > 0, \quad \mathbb{P}\big(|f(\theta_k) - \mathbb{E}[f(\theta_k)]| \geq r\big) \leq 2 \exp\Big(- \frac{r^2}{2 C_N^{(p)} \|f\|_{\mathrm{Lip},p}^2} \Big) \tag{7}$$

其中

$$\|f\|_{\mathrm{Lip},p} = \sup_{\theta \neq \theta'} \frac{|f(\theta) - f(\theta')|}{\|\theta - \theta'\|_p}.$$

对于 $t \geq 0$和 $\theta = (\theta^1, \ldots, \theta^N) \in \mathcal{E}$，令

$$\mu_\theta = \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta^i} \in \mathcal{P}(\mathbb{R}^D),$$

并对任意 1-利普希茨 φ: $\mathbb{R}^D \to \mathbb{R}$，定义

$$f_t^\varphi(\theta) = \big|\langle \varphi, \mu_\theta \rangle - \langle \varphi, \bar{\mu}_t \rangle\big|, \qquad f_t(\theta) = \mathsf{W}_1(\mu_\theta, \bar{\mu}_t), \qquad f_t^{\mathsf{s}}(\theta) = \mathsf{SW}_1(\mu_\theta, \bar{\mu}_t).$$

通过协构造后，在第 $k = \lfloor Nt \rfloor$次迭代处沿随机梯度下降迭代进行评估，我们由此得到

$$f_t^\varphi(\theta_{\lfloor Nt \rfloor}) = \Delta_t^{N,\varphi}, \qquad f_t(\theta_{\lfloor Nt \rfloor}) = \Delta_t^N, \qquad f_t^{\mathsf{s}}(\theta_{\lfloor Nt \rfloor}) = \Delta_t^{N,\mathsf{s}}. \tag{8}$$

使用典范耦合（将第 i 个原子配对），对于任意 θ、$\theta' \in \mathcal{E}$,

$$\mathsf{W}_1(\mu_\theta, \mu_{\theta'}) \le \frac{1}{N}\sum_{i=1}^{N}|\theta^i - \theta'^i| = \frac{1}{N}\|\theta - \theta'\|_1 \le \frac{1}{\sqrt{N}}\|\theta - \theta'\|_2.$$

根据康托罗维奇–鲁宾斯坦对偶（2），任意 1-利普希茨 φ 都满足 $\big|\langle\varphi,\mu\rangle - \langle\varphi,\nu\rangle\big| \le \mathsf{W}_1(\mu,\nu)$，并且由于每个投影 P_u 都是 1-利普希茨的，我们也有 $\mathsf{SW}_1(\mu,\nu) \le \mathsf{W}_1(\mu,\nu)$。因此，根据三角不等式，$f_t^\varphi$, f_t和 f_{st}在 $(\mathcal{E}, \|\cdot\|_p)$ 上都是 $N^{-1/p}$–利普希茨的，其中 $p \in \{1,2\}$。

尽管命题 $C_N^{(1)}$ 1 中的常数可能按 N的量级增长，但请注意，这不会影响（7），因为对于 $p=1$ 和 $p=2$，我们都有 $C_N^{(p)}N^{-2/p} = O(N^{-1})$。因此，结合（7）和（8），可得以下一致偏差不等式。

推论1. 固定 $p \in \{1,2\}$ 并假设 *(A1)–(A3)*、*(A4$_p$)* 和 $L_N < 1$。那么存在$C > 0$（与 N和 t无关），使得对任意 $t \ge 0$ 和 $\delta \in (0,1)$，以概率 $1-\delta$,

$$\big|\Delta_t^{N,\circ} - \mathbb{E}[\Delta_t^{N,\circ}]\big| \le C\sqrt{\frac{\log(2/\delta)}{N}}.$$

2.3 一致偏差衰减与围绕平均场的集中

上述推论1 为 Δ_t^N（同样也适用于 $\Delta^{N,\,t}{}^{\varphi}$ 和 $\Delta^{N,\,t}{}^{\mathsf{s}}$）围绕其均值的集中提供了一个集中不等式。因此，为了得到 μ_t^N 围绕 $\bar{\mu}_t$ 的集中不等式，接下来只需对该偏差进行一致的时间控制。

上文引入的假设 (A4$_p$) 仅用于推导命题 1 和推论 1。对于下文的偏差估计，只需假设初始化时具有有限的 $(2+\varepsilon)$ 阶矩：

(A5) 存在 $q > 2$ 使得 $\mathbb{E}\big[|\theta_0^1|^q\big] < \infty$。

特别地，对于 $p \in \{1,2\}$ 的任一选择，假设 (A4$_p$) 均蕴含 (A5)（因为 T_1 和T_2 都蕴含指数可积性，因而具有任意阶矩）。

命题 2。 假设 *(A1)–(A3)*、*(A5)* 以及 $L_N < 1$, $N \ge N_\star$（见 (3)）。那么存在$C > 0$（与 N 和 t 无关），使得

$$\sup_{t\ge 0}\mathbb{E}[\Delta_t^{N,\circ}] \le C\,\kappa_N,$$

其中

$$\kappa_N = \begin{cases} N^{-1/2} & \text{for } \Delta_t^{N,\varphi} \text{ and } \Delta_t^{N,\mathsf{s}}, \\ N^{-1/(1\vee D)} + N^{-1/2} & \text{for } \Delta_t^N. \end{cases}$$

注。 中唯一依赖于维数的项 κ_N 出现在沃瑟斯坦差异度$\Delta_t^N = \mathsf{W}_1(\mu_t^N, \bar{\mu}_t)$中，其速率为 $N^{-1/(1\vee D)}$。该项源于经验测度在高维空间 W_1 上的 $\mathbb{R}^D$ [20, 定理 1], 所述最优独立同分布抽样误差，体现了高维最优匹配的内在复杂性，也称为维数灾难。相比之下，$\Delta^{N,\,t}{}^{\varphi}$ 和 $\Delta^{N,\,t}{}^{\mathsf{s}}$ 的速率 $N^{-1/2}$ 与维数无关，因为它们可化为一维（或标量）平均。在低维情形下，可以得到更精细的维数特定速率以及可能的对数修正（参见 [20]中定理 1 后的讨论），但上述界已足以满足我们的需要。

最后，结合推论 1 与命题 2，我们得到以下随机梯度下降参数的经验测度与其平均场极限之间的一致时间集中不等式。

定理 1。 假设 *(A1)–(A3)*, *(A4$_p$)* ($p=1$ 或 $p=2$均可），且 $L_N < 1$, $N \ge N_\star$(见 (3))。那么存在 $C > 0$ (与 N 和 t无关），使得对任意 $t \ge 0$ 和$\delta \in (0,1)$，以概率 $1-\delta$,

$$\Delta_t^{N,\circ} \le C\,\kappa_N + C\sqrt{\frac{\log(2/\delta)}{N}}.$$

其中 $\kappa_N \to 0$ 如 $N \to \infty$ 命题 2中所给出。

应用于网络输出。 固定 $x \in \mathbb{R}^d$ 并令 $\varphi_x = \sigma_*(\cdot, x)$。在 (A1) 条件下，$\varphi_x$ 关于 x 一致地满足 M-利普希茨条件。将定理 1 应用于 $\Delta^{N,\varphi}_{\cdot,t}$（其中 $t \geq 0$），并令 $\varphi = \varphi_x/M$，可得对于任意 $\delta > 0$，以概率 $1-\delta$,

$$\left|\hat{y}_{\theta_{\lfloor Nt \rfloor}}(x) - \langle \sigma_*(\cdot, x), \bar{\mu}_t \rangle\right| \leq C\sqrt{\frac{1+\log(2/\delta)}{N}}.$$

3 证明

3.1 SGD 动力学

我们首先将随机梯度下降递推关系改写为一种形式，使其马尔可夫结构显式呈现，因为命题 1 归根结底是关于所得马尔可夫链上传输不等式的结论。证明结合了单步更新的利普希茨控制（引理 1）以及关于 T_p 不等式在压缩马尔可夫核下稳定性的结果。对于 $p=1$，通过基于拉普拉斯变换刻画（T_1）的初等归纳法，可得到5的传播。对于 $p=2$，我们则依赖压缩马尔可夫链的依赖张量化结果 [3,定理 1。2]，在完整轨迹层面应用该结果，再投影到最后一个坐标，从而得到一个 T_2 对迭代指标一致的不等式。

对于 $\theta = (\theta^1, \ldots, \theta^N) \in \mathcal{E} = (\mathbb{R}^D)^N$ 和 $z = (x, y) \in \mathcal{Z} = \mathbb{R}^d \times \mathbb{R}$，定义单步漂移

$$\begin{aligned} F^i_\lambda(\theta, z) &= \big(y - \hat{y}_\theta(x)\big)\, \nabla_{\theta^i} \sigma_*(\theta^i, x) - \lambda\, \theta^i, \qquad i \in \{1, \ldots, N\}, \\ F_\lambda(\theta, z) &= \big(F^1_\lambda(\theta, z), \ldots, F^N_\lambda(\theta, z)\big) \in \mathcal{E}, \end{aligned} \tag{9}$$

及其相应的更新映射 $\Phi(\theta,\ z) = \theta + \gamma F_\lambda(\theta,\ z)$。那么（1）等价于

$$\theta_{k+1} = \theta_k + \gamma F_\lambda(\theta_k, z_{k+1}) = \Phi(\theta_k, z_{k+1}) \tag{10}$$

其中 $z_{k+1} = (x_{k+1}, y_{k+1}) \in \mathcal{Z}$。由于 $(z_k)_{k \geq 1}$ 是独立同分布的，参数序列 $(\theta_k)_{k \in \mathbb{N}}$ 因此构成定义在 $\mathcal{E}$ 上的时间齐次马尔可夫链，其转移核为

$$\forall \theta \in \mathcal{E}, \quad P(\theta, d\theta') = \big[\Phi(\theta, \cdot)_{\#}\pi\big](d\theta'). \tag{11}$$

引理 1. 假设 *(A1)–(A3)*。那么，对于任意 θ，$\theta' \in \mathcal{E}$ 和 z，$z' \in$ 支撑集 π,

(i) $\|\Phi(\theta, z) - \Phi(\theta, z')\|_2 \leq \dfrac{K}{\sqrt{N}} |z - z'|_\infty$，且 $\|\Phi(\theta, z) - \Phi(\theta, z')\|_1 \leq K\, |z - z'|_\infty$,

(ii) $\|\Phi(\theta,\ z) - \Phi(\theta',\ z)\|_p \leq L_N \|\theta - \theta'\|_p$ 对于任意 $p \in \{1, 2\}$,

其中常数 K，L_N 定义于 (3)。

证明。 (i) 令 θ，$\theta' \in \mathcal{E}$ 和 $z = (x,\ y)$ $z' = (x',\ y') \in \operatorname{supp} \pi \subset \mathbb{R}^d \times [-A, A]$（根据 (A2)）。对于每个 $i \in \{1, \ldots, N\}$，由三角不等式可得，

$$\begin{aligned} |\Phi^i(\theta, z) - \Phi^i(\theta, z')| &= \gamma \big|(y - \hat{y}_\theta(x))\nabla_{\theta^i}\sigma_*(\theta^i, x) - (y' - \hat{y}_\theta(x'))\nabla_{\theta^i}\sigma_*(\theta^i, x') \\ &\qquad \pm (y - \hat{y}_\theta(x))\nabla_{\theta^i}\sigma_*(\theta^i, x')\big| \\ &\leq \gamma |y - y'||\nabla_{\theta^i}\sigma_*(\theta^i, x')| + \gamma |y - \hat{y}_\theta(x)||\nabla_{\theta^i}\sigma_*(\theta^i, x) - \nabla_{\theta^i}\sigma_*(\theta^i, x')| \\ &\quad + \gamma |\hat{y}_\theta(x') - \hat{y}_\theta(x)||\nabla_{\theta^i}\sigma_*(\theta^i, x')|. \end{aligned}$$

由 (A1)可知，映射 $x \in \mathbb{R}^d \mapsto \sigma_*(\theta^i,\ x)$ 是 M-利普希茨（关于 i一致），因此 $|\hat{}y_\theta(x') - \hat{}y_\theta(x)| \leq M|x - x'|$。再结合 (A3)，得到

$$|\Phi^i(\theta, z) - \Phi^i(\theta, z')| \leq \gamma M\, |y - y'| + \gamma\big((A+B)L_x + M^2\big)|x - x'|.$$

由此可得

$$\|\Phi(\theta, z) - \Phi(\theta, z')\|_2 \leq \gamma\sqrt{N}\big(M\, |y - y'| + \big((A+B)L_x + M^2\big)|x - x'|\big) \leq \frac{K}{\sqrt{N}}\, |z - z'|_\infty,$$

因此

$$\|\Phi(\theta,z)-\Phi(\theta,z')\|_1 \le \sqrt{N}\|\Phi(\theta,z)-\Phi(\theta,z')\|_2 \le K\,|z-z'|_\infty.$$

(ii) 再次，根据三角不等式，对任意 $i\in\{1,\dots,N\}$,

$$\begin{aligned}|\Phi^i(\theta,z)-\Phi^i(\theta',z)| &= \big|(1-\gamma\lambda)(\theta^i-\theta'^i)+\gamma(y-\hat{y}_\theta(x))\nabla_{\theta^i}\sigma_*(\theta^i,x)\\ &\quad -\gamma\big(y-\hat{y}_{\theta'}(x)\big)\nabla_{\theta^i}\sigma_*(\theta'^i,x)\pm\gamma(y-\hat{y}_\theta(x))\nabla_{\theta^i}\sigma_*(\theta'^i,x)\big|\\ &\le |1-\gamma\lambda||\theta^i-\theta'^i|+\gamma|y-\hat{y}_\theta(x)||\nabla_{\theta^i}\sigma_*(\theta^i,x)-\nabla_{\theta^i}\sigma_*(\theta'^i,x)|\\ &\quad +\gamma\big|\hat{y}_{\theta'}(x)-\hat{y}_\theta(x)\big||\nabla_{\theta^i}\sigma_*(\theta'^i,x)|.\end{aligned}$$

根据 (A1)，映射 $\theta^i\in\mathbb{R}^D\mapsto\sigma_*(\theta^i,\ x)$ 是 M-利普希茨的（关于 $x\in\mathbb{R}^D$一致），因此

$$|\hat{y}_{\theta'}(x)-\hat{y}_\theta(x)| \le \frac{1}{N}\sum_{i=1}^N|\sigma_*(\theta'^i,x)-\sigma_*(\theta^i,x)| \le \frac{M}{N^{1/p}}\|\theta-\theta'\|_p.$$

再结合 (A2) 和 (A3)，可得

$$|\Phi^i(\theta,z)-\Phi^i(\theta',z)| \le \big(|1-\gamma\lambda|+\gamma(A+B)L_\theta\big)\,|\theta^i-\theta'^i|+\gamma\frac{M^2}{N^{1/p}}\,\|\theta-\theta'\|_p.$$

由此可由闵可夫斯基不等式可得 $\|\Phi(\theta,\cdot)-\Phi(\theta',\cdot)\|_p \le L_N\,\|\theta-\theta'\|_p$对所有 $p\in\{1,2\}$ 均成立。 □

有了引理 1 ，我们现在可以从随机梯度下降迭代的单步稳定性过渡到其随时间一致的集中性。

命题的证明 *1*。情形 $p=1$。证明对 $k\in\mathbb{N}$进行归纳。由于 $\mu_0\in T_1(C_0^{(1)})$，由张量化可得 $\nu_0=\mu_0^{\otimes N}\in T_1\big(NC_0^{(1)}\big)$ 在 $(\mathcal{E},\|\cdot\|_1)$上。现在，假设$\nu_k\in T_1(c_k)$ 对某个 $c_k>0$ 和 $k\in\mathbb{N}$成立。下面证明 $\nu_{k+1}\in T_1(c_{k+1})$ 其中

$$c_{k+1}=L_N^2c_k+K^2C_\pi^{(1)}, \tag{12}$$

其中 K 和 L_N 是引理 1 中的利普希茨常数。这里，我们使用了T_1 与高斯集中 (5) 之间的等价性。令 $f:(\mathcal{E},\|\cdot\|_1)\to\mathbb{R}$ 为 1-利普希茨函数。取 $Z\sim\pi$，并令对 $\theta\in\mathcal{E}$, $g(\theta)=\mathbb{E}[f(\Phi(\theta,Z))]$。由引理 1 可知， g是 L_N-利普希茨的，并且对每个固定的 $\theta\in\mathcal{E}$，映射 $z\mapsto f(\Phi(\theta,z))$ 是 K-利普希茨的。由于 $\pi\in T_1\big(C_\pi^{(1)}\big)$，集中界 (5) 对所有 $\theta\in\mathcal{E}$ 给出:

$$\forall\xi\in\mathbb{R},\quad \mathbb{E}\big[e^{\xi(f(\Phi(\theta,Z))-g(\theta))}\big]\le\exp\left(\frac{C_\pi^{(1)}\xi^2}{2}K^2\right).$$

现在，令 $\theta=\theta_k\sim\nu_k$ 为随机变量，且与 Z相互独立，使得 $\Phi(\theta_k,Z)$ 与 θ_{k+1} (参见（10）) 同分布。对 θ_k进行条件化后，对于任意 $\xi\in\mathbb{R}$,

$$\begin{aligned}\mathbb{E}\big[e^{\xi(f(\theta_{k+1})-\mathbb{E}[f(\theta_{k+1})])}\big] &= \mathbb{E}\Big[e^{\xi(g(\theta_k)-\mathbb{E}[g(\theta_k)])}\,\mathbb{E}\big[e^{\xi(f(\theta_{k+1})-g(\theta_k))}\,\big|\,\theta_k\big]\Big]\\ &\le \exp\left(\frac{C_\pi^{(1)}\xi^2}{2}K^2\right)\mathbb{E}\big[e^{\xi(g(\theta_k)-\mathbb{E}[g(\theta_k)])}\big].\end{aligned}$$

由于 $\nu_k\in T_1(c_k)$ 且 g 是 L_N-利普希茨函数，因此再次由（5）得

$$\forall\xi\in\mathbb{R},\quad \mathbb{E}\big[e^{\xi(g(\theta_k)-\mathbb{E}[g(\theta_k)])}\big]\le\exp\left(\frac{c_k\xi^2}{2}L_N^2\right).$$

结合最后两个不等式，便可得到

$$\forall\xi\in\mathbb{R},\quad \mathbb{E}\big[e^{\xi(f(\theta_{k+1})-\mathbb{E}[f(\theta_{k+1})])}\big]\le\exp\left(\frac{\xi^2}{2}\big(L_N^2c_k+K^2C_\pi^{(1)}\big)\right)$$

这又等价于 $\nu_{k+1} \in T_1(c_{k+1})$ 其中 c_{k+1} 如（12）所示。在 $L_N < 1$的条件下，可知对所有 $k \geq 1$,

$$c_k = L_N^{2k} N C_0^{(1)} + K^2 C_\pi^{(1)} \sum_{j=0}^{k-1} L_N^{2j} \leq N C_0^{(1)} + \frac{K^2}{(1 - L_N^2)} C_\pi^{(1)} = C_N^{(1)},$$

这就是针对 $p = 1$ 所声称的界。

情形 $p = 2$。这里我们采用一种不同的方法：首先利用压缩马尔可夫链的依赖张量化定理，为整个轨迹 $(\theta_0,\dots,\theta_k)$ 的联合分布建立一个 T_2 不等式，然后投影到最后一个坐标，从而得到针对 ν_k 的 T_2 不等式，其常数与 k 无关。

固定 $k \in \mathbb{N}$，记 $\nu^{(k)}$ 为轨迹 $(\theta_0, \dots, \theta_k)$ 在 $\mathcal{E}^{k+1}$ 上的分布，并赋予其 ℓ^2 乘积距离

$$\mathsf{d}_2\big((\theta_0, \dots, \theta_k), (\theta'_0, \dots, \theta'_k)\big) = \left(\sum_{j=0}^{k} \|\theta_j - \theta'_j\|_2^2 \right)^{1/2}.$$

我们应用 [3, 定理 1。2] 将其应用于 $s = 2$ 上的测度 $\nu^{(k)}$，该测度位于 $(\mathcal{E}^{k+1},\mathsf{d}_2)$上，并验证其中的假设 (i) 和 (ii) 均得到满足。

首先，由于 $\mu_0 \in T_2\big(C_0^{(2)}\big)$，根据张量化可得 $\nu_0 = \mu_0^{\otimes N} \in T_2\big(C_0^{(2)}\big)$。此外，利用Lipschitz推前下的稳定性以及 $\pi \in T_2\big(C_\pi^{(2)}\big)$，由引理 1 (i) 和 (11) 可得 $P(\theta, \cdot) \in T_2(C_P)$，其中 $C_P = C_\pi^{(2)} K^2/N$与 θ无关。因此，[3,定理 1。2] 中的条件 (i) 以常数 $C_0^{(2)} \vee C_P$成立。

其次，令 $Z \sim \pi$并考虑 $\big(\Phi(\theta、 Z)\ \Phi(\theta'、 Z)\big)$ 的耦合，其中 $P(\theta、 \cdot)$和 $P(\theta'、 \cdot)$。定理 1 中的收缩条件 (ii) [3, 定理 1。2] 随后由引理 1 (ii) 直接得到，因为

$$\mathsf{W}_2^2\big(P(\theta, \cdot), P(\theta', \cdot)\big) \leq \mathbb{E}\big[\|\Phi(\theta, Z) - \Phi(\theta', Z)\|_2^2\big] \leq L_N^2 \|\theta - \theta'\|_2^2.$$

根据 $L_N < 1$，[3, 定理 1。2] 在 $(\mathcal{E}^{k+1}, \mathsf{d}_2)$ 上给出 $\nu^{(k)} \in T_2\big(C_N^{(2)}\big)$，其常数为 $C_N^{(2)} =(1 - L_N)^{-2}\big(C_0^{(2)} \vee C_P\big)$ 且与 k 无关。最后，投影 $(\theta_0, \dots, \theta_k) \mapsto \theta_k$从 $(\mathcal{E}^{k+1}, \mathsf{d}_2)$ 到 $(\mathcal{E}, \|\cdot\|_2)$ 是 1-利普希茨的，因此 $\nu_k \in T_2\big(C_N^{(2)}\big)$，这就是关于 $p = 2$ 的所声称界。 □

3.2 一致偏差衰减

现在，我们推导一个关于偏差项的时间一致上界 $\mathbb{E}[\Delta_t^{N,\circ}]$，该偏差项出现在推论1中。为此，我们将随机梯度下降迭代与一个由平均场动力学驱动的辅助独立同分布粒子系统进行比较，该系统定义如下。

首先，我们通过引入支配无限宽度情形下典型粒子分布的演化方程，将平均场极限形式化。我们将平均场极限 $(\bar{\mu}_t)_{t\geq 0}$定义为连续性方程在 $\mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^D))$ 中的唯一（确定性）解

$$\partial_t \bar{\mu}_t + \alpha \nabla \cdot \big(G_\lambda(\cdot, \bar{\mu}_t)\, \bar{\mu}_t\big) = 0, \qquad \bar{\mu}_0 = \mu_0, \tag{13}$$

其中，G_λ 表示平均场漂移：对于任意 $(\theta^i, \mu) \in \mathbb{R}^D \times \mathcal{P}(\mathbb{R}^D)$,

$$G_\lambda(\theta^i, \mu) = \mathbb{E}_\pi\big[\big(y - \langle \sigma_*(\cdot, x), \mu\rangle\big) \nabla_{\theta^i} \sigma_*(\theta^i, x)\big] - \lambda \theta^i. \tag{14}$$

方程（13）应在弱意义下理解（例如参见 [38, 定义 4。1]），即对于任意光滑且紧支撑的测试函数 $\varphi \in \mathcal{C}_c^1(\mathbb{R}^D)$，

$$\frac{d}{dt}\langle \varphi, \bar{\mu}_t\rangle = \alpha \left\langle \nabla \varphi \cdot G_\lambda(\cdot, \bar{\mu}_t), \bar{\mu}_t \right\rangle.$$

在(A1)–(A3)条件下，惩罚情形中的方程（13）的存在性和唯一性可通过直接调整[16, §2.31] 中的论证得出。（正则化项增加了一个线性耗散漂移，而其中使用的W_1 上的 Lipschitz 估计保持不变。）

分析（13）的一种有用概率工具由相应的非线性动力学给出：按照 [45, 定理 1。1], 存在唯一的（按分布意义）轨迹¯解$(\theta_t)_{t\geq 0}$ 满足

$$\dot{\bar{\theta}}_t = \alpha\, G_\lambda(\bar{\theta}_t, \rho_t), \qquad \bar{\theta}_0 \sim \mu_0, \qquad \rho_t = \mathcal{L}(\bar{\theta}_t), \tag{15}$$

且 $(\rho_t)_{t\geq 0}$ 满足（13）。此外，在 (A1)、(A2) 和 (A5) 的条件下，容易证明 $(\rho_t)_{t\geq 0} \in \mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^D))$，并且 $t \geq 0 \mapsto \rho_t$ 实际上关于 W_1 是利普希茨连续的。由平均场偏微分方程（13）解的唯一性可知，对所有 $t \geq 0$，均有 $\rho_t = \bar{\mu}_t$。因此，我们按惯例引入独立同分布粒子系统 $(\bar{\theta}_t^i)_{t\geq 0}$, $i \in \{1, \ldots, N\}$，其初始点与（1）相同，并根据非线性动力学（15）独立演化 $t \geq 0$：

$$\forall t \geq 0, \quad \bar{\theta}_t^i = \theta_0^i + \alpha \int_0^t G_\lambda(\bar{\theta}_s^i, \bar{\mu}_s)\, ds, \tag{16}$$

使得 $\mathcal{L}(\bar{\theta}_t^i) = \bar{\mu}_t$ 对所有 $t \geq 0$均成立。相应的经验测度记为

$$\bar{\mu}_t^N = \frac{1}{N} \sum_{i=1}^N \delta_{\bar{\theta}_t^i}.$$

根据三角不等式，我们将偏差分解为

$$\mathbb{E}[\Delta_t^N] \leq \mathbb{E}\big[\mathsf{W}_1(\mu_t^N, \bar{\mu}_t^N)\big] + \mathbb{E}\big[\mathsf{W}_1(\bar{\mu}_t^N, \bar{\mu}_t)\big], \tag{17}$$

对于 $\mathbb{E}[\Delta_t^{N,\varphi}]$ 和 $\mathbb{E}[\Delta_t^{N,\mathsf{s}}]$，情况类似。式（17）中的第一项刻画了随机梯度下降粒子与其平均场对应粒子之间的动态误差。利用基于同步耦合的标准 $t \geq 0$ 混沌传播论证 [45], ，可对所有 $t \geq 0$ 一致地控制该误差。

命题 3。假设 *(A1)–(A3)*、*(A5)*，以及 $\lambda > \lambda_\star$， $N \geq N_\star$ （见(3)）。则存在$C > 0$ （与 N 和 t无关），使得

$$\sup_{t\geq 0} \mathbb{E}\big[\mathsf{W}_1(\mu_t^N, \bar{\mu}_t^N)\big] \leq \frac{C}{\sqrt{N}}.$$

回顾 $|\langle \varphi、\ \mu\rangle - \langle \varphi、\ \nu\rangle| \leq \mathsf{W}_1(\mu, \nu)$，对于任意1-利普希茨 φ和$\mathsf{SW}_1 \leq \mathsf{W}_1$,，命题 3直接推出下面的推论 2。

推论 2. 在命题 *3*的假设下，存在 $C > 0$ *(*与 N和 t无关*)*，使得

$$\sup_{t\geq 0} \mathbb{E}\big[|\langle \varphi, \mu_t^N\rangle - \langle \varphi, \bar{\mu}_t^N\rangle|\big] \leq \frac{C}{\sqrt{N}}, \qquad \sup_{t\geq 0} \mathbb{E}\big[\mathsf{SW}_1(\mu_t^N, \bar{\mu}_t^N)\big] \leq \frac{C}{\sqrt{N}}.$$

(17)中的第二项是纯粹的 静态 独立同分布经验测度抽样误差，可由最优 W_1 速率控制，参见 [20, 定理 1], ，我们将在下面的命题中回顾这一结果。

命题 4. 假设 *(A1)*、*(A2)* 和*(A5)*成立。那么存在 $C > 0$ *(*与 N和 t无关*)*，使得

$$\sup_{t\geq 0} \mathbb{E}\big[\mathsf{W}_1(\bar{\mu}_t^N, \bar{\mu}_t)\big] \leq C N^{-1/(1\vee D)}.$$

我们还要求经验测度 $\bar{\mu}_t^N$在针对测试函数 φ 进行检验时以及经由SW_1时具有静态的独立同分布采样速率。与W_1, 不同，我们预期这两个量具有与参数维数 D无关的一致 $N^{-1/2}$ 速率，因为前者只是标量样本均值，而后者通过投影归结为一维 Wasserstein 距离。这一直观认识在下一命题中得到形式化。

命题 5. 假设*(A5)*。则存在 $C > 0$ （与 N 和 t无关），使得

$$\sup_{t\geq 0}\mathbb{E}\big[\big|\langle\varphi,\bar{\mu}_t^N\rangle - \langle\varphi,\bar{\mu}_t\rangle\big|\big] \;\leq\; \frac{C}{\sqrt{N}}, \qquad \sup_{t\geq 0}\mathbb{E}\big[\mathsf{SW}_1(\bar{\mu}_t^N,\bar{\mu}_t)\big] \;\leq\; \frac{C}{\sqrt{N}}.$$

结合（17）、命题 3–5 以及推论 2 可得命题 2 （注意，命题 $L_N < 1$ 中的条件 2 蕴含了 $\lambda > \lambda_\star$ 命题 3中的条件）。命题 3、4和 5的证明将在接下来的两节中给出。

3.2.1 命题 4-5的证明

我们首先为独立同分布的平均场粒子 $\bar{\theta}_t^i$建立一致时间矩界。该界将在命题 3、4和 5的证明过程中反复使用。

引理 2。 假设 *(A1)–(A2)*，并且 $\mathbb{E}\big[|\theta_0^1|^q\big] < \infty$ 对某个 $q > 1$成立。则存在 $C_q > 0$（与 N 和 t无关），使得

$$\sup_{t\geq 0}\mathbb{E}\big[|\bar{\theta}_t^1|^q\big] \leq C_q < \infty.$$

证明。 固定 $q > 1$ 使得 $\mathbb{E}\big[|\theta_0^1|^q\big] < \infty$，并令 $f_q(t) = \mathbb{E}\big[|\bar{\theta}_t^1|^q\big]$， $t \geq 0$。对于 $q > 1$，映射 N_q： $w \mapsto |w|^q$ 在 $\mathbb{R}^D$ 上连续可微，且 $\nabla N_q(w) = q|w|^{q-2}w$ 对于 $w \neq 0$ 和 $\nabla N_q(0) = 0$成立。因此，利用（15）和链式法则，可得

$$\begin{aligned} f_q'(t) &= \mathbb{E}\big[\nabla N_q(\bar{\theta}_t^1)\cdot\dot{\bar{\theta}}_t^1\big] \\ &= q\alpha\,\mathbb{E}\big[|\bar{\theta}_t^1|^{q-2}\bar{\theta}_t^1\cdot G_\lambda(\bar{\theta}_t^1,\bar{\mu}_t)\big] \\ &\leq -q\alpha\lambda\,\mathbb{E}\big[|\bar{\theta}_t^1|^q\big] + q\alpha C\,\mathbb{E}\big[|\bar{\theta}_t^1|^{q-1}\big] \end{aligned}$$

在 $C = (A+B)M$中，最后一个不等式由 (A1)–(A2)得到。由指数为q 和 $q' = q/(q-1)$的杨氏不等式，对于任意 $\varepsilon > 0$,

$$C|\bar{\theta}_t^1|^{q-1} \leq \frac{\varepsilon^{q'}}{q'}|\bar{\theta}_t^1|^q + \frac{C^q}{q\varepsilon^q}.$$

取 $\varepsilon^{q'} = \lambda$ 并代入前一个微分不等式，得到

$$f_q'(t) \leq -\alpha\lambda f_q(t) + \alpha\frac{C^q}{\lambda^{q-1}}.$$

由格朗沃尔估计可得，对所有 $t \geq 0$,

$$f_q(t) \leq e^{-\alpha\lambda t}f_q(0) + \frac{C^q}{\lambda^q}(1-e^{-\alpha\lambda t}) \leq \mathbb{E}\big[|\theta_0^1|^q\big] + \frac{C^q}{\lambda^q} = C_q < \infty,$$

至此证明完毕。 □

命题 *4* 的证明。 设 $q > 2$ 由 (A5)给出。由引理 2，存在 $C_q > 0$ (与 t 和 N无关）使得

$$\sup_{t\geq 0}\mathbb{E}\big[|\bar{\theta}_t^1|^q\big] \leq C_q.$$

因此，将 [20, 定理 1] 应用于 $p = 1$ ，并取指数 $q > 2$，对分布为$(\bar{\theta}_t^1,\dots,\bar{\theta}_t^N)$ 的独立同分布样本， $\bar{\mu}_t$可得，对所有 $t \geq 0$,

$$\mathbb{E}\big[\mathsf{W}_1(\bar{\mu}_t^N,\bar{\mu}_t)\big] \leq C\left(\mathbb{E}[|\bar{\theta}_t^1|^q]\right)^{1/q} N^{-1/(1\vee D)},$$

其中 C 仅依赖于 D 和 q。通过对 $t \geq 0$取上确界，即可得到该结果。 □

命题5的证明。设 $t \geq 0$，并令 $U_t^i = \varphi(\bar{\theta}_t^i)$，它们与 $\mathbb{E}[U_t^1] = \langle \varphi, \bar{\mu}_t \rangle$独立同分布。由柯西–施瓦茨不等式，

$$\mathbb{E}\big[\big|\langle \varphi, \bar{\mu}_t^N \rangle - \langle \varphi, \bar{\mu}_t \rangle\big|\big] = \mathbb{E}\Big[\Big|\frac{1}{N}\sum_{i=1}^{N}(U_t^i - \mathbb{E}[U_t^1])\Big|\Big] \leq \frac{\sqrt{\mathrm{Var}(U_t^1)}}{\sqrt{N}}.$$

由于 φ是 1-利普希茨的，$\mathrm{Var}(U_t^1) \leq \mathbb{E}\big[|U_t^1 - \varphi(0)|^2\big] \leq \mathbb{E}\big[|\bar{\theta}_t^1|^2\big]$。由 (A5)， $\mathbb{E}[|\theta_0^1|^2] < \infty$。因此，我们可以对引理 2 应用一致矩界，其中 $q = 2$ ，并对前一显示式中的$t \geq 0$ 取上确界，得到

$$\sup_{t \geq 0} \mathbb{E}\big[\big|\langle \varphi, \bar{\mu}_t^N \rangle - \langle \varphi, \bar{\mu}_t \rangle\big|\big] \leq \frac{C}{\sqrt{N}}.$$

现在我们转向切片Wasserstein距离的类似界。由富比尼定理，

$$\mathbb{E}\big[\mathsf{SW}_1(\bar{\mu}_t^N, \bar{\mu}_t)\big] = \int_{\mathbb{S}^{D-1}} \mathbb{E}\big[\mathsf{W}_1\big(P_{u\#}\bar{\mu}_t^N, P_{u\#}\bar{\mu}_t\big)\big]\, \zeta(du). \tag{18}$$

对于每个 $u \in \mathbb{S}^{D-1}$， $P_{u\#}\bar{\mu}_t^N$是由 N 个同分布于$P_{u\#}\bar{\mu}_t$ 上、服从 $\mathbb{R}$的独立同分布样本所构成的经验测度。由 (A5) 和引理 2，存在 $q > 2$ 和 $C_q > 0$ （与 D、 N 和t无关），使得

$$\sup_{t \geq 0} \sup_{u \in \mathbb{S}^{D-1}} \int_{\mathbb{R}} |w|^q \, P_{u\#}\bar{\mu}_t(dw) = \sup_{t \geq 0} \sup_{u \in \mathbb{S}^{D-1}} \mathbb{E}\big[|\langle u, \bar{\theta}_t^1 \rangle|^q\big] \leq \sup_{t \geq 0} \mathbb{E}\big[|\bar{\theta}_t^1|^q\big] \leq C_q.$$

因此，应用[20, 定理 1] 于维度 $D = 1$ 并取 $p = 1$ 可得常数 $C > 0$（仅依赖于 q），使得对所有 $t \geq 0$ 以及所有 $u \in \mathbb{S}^{D-1}$，

$$\mathbb{E}\big[\mathsf{W}_1\big(P_{u\#}\bar{\mu}_t^N, P_{u\#}\bar{\mu}_t\big)\big] \leq \frac{C}{\sqrt{N}}.$$

对 $u \sim \zeta$ 积分并对 $t \geq 0$ 取上确界，即可得到所需的界。 □

3.2.2 命题3的证明

尽管网络参数是以独立同分布的方式初始化的，但在随机梯度下降动态下，这种独立性通常会消失。然而，不难看出，系统的可交换性始终得以保持，即参数的分布在其坐标的任意置换下保持不变。事实上，这是更新映射 Φ 的置换等变性直接导致的：对于 $\{1, \ldots, N\}$ 的任意置换 σ，

$$\forall \theta \in \mathcal{E}, \quad \Phi(\theta^\sigma, \cdot) = \Phi(\theta, \cdot)^\sigma \tag{19}$$

其中 $\theta^\sigma = (\theta^{\sigma(1)}, \ldots, \theta^{\sigma(N)})$。此外，尽管随机梯度下降系统和平均场系统通过相同的初始化耦合在一起（因此先验上并不独立），由此得到的成对粒子族仍然是可交换的。以下引理将对此作出精确说明。

引理 3. 对于每个 $t \geq 0$， $\big(\theta_{\lfloor Nt \rfloor}, \bar{\theta}_t\big)$ 的分布都是可交换的。

证明。 固定一个置换 σ，属于 $\{1, \ldots, N\}$，并令 $t \geq 0$和 $k = \lfloor Nt \rfloor$。根据定义（16），存在一个确定性映射 $\Psi_t \colon \mathbb{R}^D \to \mathbb{R}^D$ （与 i无关），使得对于每个 $i \in \{1, \ldots, N\}$， $\bar{\theta}_t^i = \Psi_t\ (\theta_0^i)$ 。特别地，

$$\bar{\theta}_t^\sigma = \big(\bar{\theta}_t^{\sigma(1)}, \ldots, \bar{\theta}_t^{\sigma(N)}\big) = \big(\Psi_t(\theta_0^{\sigma(1)}), \ldots, \Psi_t(\theta_0^{\sigma(N)})\big) = \Psi_t(\theta_0^\sigma),$$

其中，我们通过记号滥用将 $\Psi_t : \mathcal{E} \to \mathcal{E}$ 逐分量扩展。

另一方面，通过迭代（10），我们可以写成 $\theta_k = \Phi^{(k)}\ (\theta_0,\ z_1,\ \ldots,\ z_k)$ ，其中 $\Phi^{(k)} : \mathcal{E} \times \mathcal{Z}^k \to \mathcal{E}$ 递归定义为 $\Phi^{(0)}\ (\theta)\ = \theta$ ，且对于 $j \in \mathbb{N}$，

$$\Phi^{(j+1)}(\theta, z_1, \ldots, z_{j+1}) = \Phi(\Phi^{(j)}(\theta, z_1, \ldots, z_j), z_{j+1}).$$

因此，通过反复应用（19），可得对所有 $\theta \in \mathcal{E}$ 和 $(z_1, \ldots, z_k) \in \mathcal{Z}^k$，

$$\Phi^{(k)}(\theta^\sigma, z_1, \ldots, z_k) = \Phi^{(k)}(\theta, z_1, \ldots, z_k)^\sigma.$$

因此，映射 Ψ_t 与 $\Phi^{(k)}$ 也是置换等变的。利用初始参数向量 θ_0 的可交换性，以及 θ_0 与数据点 $(z_1, \ldots, z_k)$ 之间的独立性，我们于是有

$$(\theta_k, \bar{\theta}_t)^\sigma = \big(\Phi^{(k)}(\theta_0^\sigma, z_1, \ldots, z_k), \Psi_t(\theta_0^\sigma)\big) \overset{d}{=} \big(\Phi^{(k)}(\theta_0, z_1, \ldots, z_k), \Psi_t(\theta_0)\big) = (\theta_k, \bar{\theta}_t)$$

这就证明了同步耦合粒子的可交换性。 □

利用可交换性，并采用与引理 3 中相同的同步耦合，我们有 e

$$\mathbb{E}\big[\mathsf{W}_1(\mu_t^N, \bar{\mu}_t^N)\big] \le \frac{1}{N}\sum_{i=1}^N \mathbb{E}\big[|\theta^i_{\lfloor Nt \rfloor} - \bar{\theta}^i_t|\big] = \mathbb{E}\big[|\theta^1_{\lfloor Nt \rfloor} - \bar{\theta}^1_t|\big]. \tag{20}$$

因此，可交换性将 W_1 差异度归结为单个标记粒子的期望偏差。

命题 3 的证明。根据上述论证，只需对 $\mathbb{E}\big[|\theta^1_{\lfloor Nt \rfloor} - \bar{\theta}^1_t|\big]$ 关于 $t \ge 0$ 给出一致的界。对于 $t \ge 0$，令 $k = \lfloor Nt \rfloor$ 与 $t_k = k/N$（从而 $t \in [t_k, t_{k+1})$）。对 $i \in \{1, \ldots, N\}$ 定义，

$$\delta_k^i = \theta_k^i - \bar{\theta}^i_{t_k}, \qquad \delta_k = \delta_k^1, \qquad D_t = \bar{\theta}^1_{t_k} - \bar{\theta}^1_t,$$

从而

$$|\theta^1_{\lfloor Nt \rfloor} - \bar{\theta}^1_t| \le |\delta_k| + |D_t|. \tag{21}$$

根据 (16)，(A1) 和 (A2)，对所有 $t \ge 0$，

$$\mathbb{E}\big[|D_t|\big] \le \alpha \int_{t_k}^t \mathbb{E}\big[|G_\lambda(\bar{\theta}^1_s, \bar{\mu}_s)|\big]\, ds \le \gamma\big((A+B)M + \lambda \sup_{t \ge 0} \mathbb{E}\big[|\bar{\theta}^1_t|\big]\big)$$

其中在最后一个不等式中，我们使用了 $t - t_k \le 1/N$ 和 $\gamma = \alpha/N$。根据引理 2，在假设 (A5) 下，我们有 $\sup_{t\ge 0}\mathbb{E}\big[|\bar{\theta}^1_t|\big] \le 1 + \sup_{t\ge 0}\mathbb{E}\big[|\bar{\theta}^1_t|^2\big] < \infty$。因此，离散化误差 D_t 一致有界：

$$\sup_{t>0} \mathbb{E}\big[|D_t|\big] \le \frac{C}{N}. \tag{22}$$

剩下的工作是对 $\mathbb{E}\big[|\delta_k|\big]$ 在 $k \in \mathbb{N}$ 上的一致性进行控制。根据 (16)，对所有 $k \in \mathbb{N}$，

$$\bar{\theta}^1_{t_{k+1}} = \bar{\theta}^1_{t_k} + \alpha \int_{t_k}^{t_{k+1}} G_\lambda(\bar{\theta}^1_s, \bar{\mu}_s)\, ds,$$

因此，加上并减去 $\gamma G_\lambda(\bar{\theta}^1_{t_k}, \bar{\mu}_{t_k})$，

$$\delta_{k+1} = \delta_k + \gamma\left(F^1_\lambda(\theta_k, z_{k+1}) - G_\lambda(\bar{\theta}^1_{t_k}, \bar{\mu}_{t_k})\right) + E_{k+1} \tag{23}$$

其中

$$E_{k+1} = \gamma G_\lambda(\bar{\theta}^1_{t_k}, \bar{\mu}_{t_k}) - \alpha \int_{t_k}^{t_{k+1}} G_\lambda(\bar{\theta}^1_s, \bar{\mu}_s)\, ds.$$

我们按照以下方式进行分解：[27, 引理 7。2]

$$\begin{aligned} F^1_\lambda(\theta_k, z_{k+1}) - G_\lambda(\bar{\theta}^1_{t_k}, \bar{\mu}_{t_k}) &= F^1_\lambda(\theta_k, z_{k+1}) - G_\lambda(\theta^1_k, \mu^N_{t_k}) \\ &\quad + G_\lambda(\theta^1_k, \mu^N_{t_k}) - G_\lambda(\theta^1_k, \bar{\mu}_{t_k}) \\ &\quad + G_\lambda(\theta^1_k, \bar{\mu}_{t_k}) - G_\lambda(\bar{\theta}^1_{t_k}, \bar{\mu}_{t_k}) \\ &\equiv A_k + B_k + C_k. \end{aligned}$$

令 $m_k = \mathbb{E}\big[|\delta_k|^2\big]$。注意到 $m_0 = 0$ 由于 $\delta_0 = 0$ 在同步耦合下。利用 δ_{k+1} (23) 中的递推关系以及上述分解，展开 $|\delta_{k+1}|^2$ 并取期望可得

$$\begin{aligned} m_{k+1} \le m_k + \gamma^2\,\mathbb{E}\big[|A_k + B_k + C_k|^2\big] + 2\gamma\,\mathbb{E}\big[\delta_k \cdot (A_k + B_k + C_k)\big] \\ + \mathbb{E}\big[|E_{k+1}|^2\big] + 2\,\mathbb{E}\big[\delta_k \cdot E_{k+1}\big] + 2\gamma\,\mathbb{E}\big[(A_k + B_k + C_k) \cdot E_{k+1}\big]. \end{aligned}$$

我们研究以下技术引理中等式右侧的每一项。

引理 4. 假设 *(A1)–(A3)* 和 *(A5)*。则对所有 $k \in \mathbb{N}$:

(i) 令$(\mathcal{F}_k)_{k\in\mathbb{N}}$ 为由 $\mathcal{F}_k = \sigma(\theta_0,\ z_1,\ \ldots,\ z_k)$给出的滤过。则 $\mathbb{E}[A_k \mid \mathcal{F}_k] = 0$ 且 $\mathbb{E}\big[|A_k|^2\big] \le C$。

(ii) $\mathbb{E}\big[\delta_k \cdot B_k\big] \le M^2\, m_k + CN^{-1}$，且 $\mathbb{E}\big[|B_k|^2\big] \le 2M^4\, m_k + 2M^2B^2N^{-1}$。

(iii) $\mathbb{E}\big[\delta_k \cdot C_k\big] \le -\big(\lambda - (A+B)L_\theta\big)\, m_k$，且 $\mathbb{E}\big[|C_k|^2\big] \le \big(\lambda + (A+B)L_\theta\big)^2\, m_k$.(iv) $\mathbb{E}\big[|E_{k+1}|\big] \le CN^{-2}$，$\mathbb{E}\big[|E_{k+1}|^2\big] \le CN^{-4}$，且 $\mathbb{E}\big[\delta_k \cdot E_{k+1}\big] \le \frac{1}{4}\gamma(\lambda - \lambda_\star)\, m_k + C\gamma^2$。

引理4 的证明推迟到本节末尾。我们现在给出结论。

由于 δ_k、B_k、C_k、E_{k+1} 是 $\mathcal{F}_k$-可测的，并且 $\mathbb{E}\big[A_k \mid \mathcal{F}_k\big] = 0$ 由引理4 (i)，所有涉及 A_k 的交叉项均为零:

$$\mathbb{E}\big[\delta_k \cdot A_k\big] = \mathbb{E}\big[A_k \cdot B_k\big] = \mathbb{E}\big[A_k \cdot C_k\big] = \mathbb{E}\big[A_k \cdot E_{k+1}\big] = 0.$$

因此，利用 $2\gamma\,(B_k + C_k) \cdot E_{k+1} \le \gamma^2|B_k + C_k|^2 + |E_{k+1}|^2$ 和 $|B_k + C_k|^2 \le 2|B_k|^2 + 2|C_k|^2$，我们得到

$$\begin{aligned} m_{k+1} \le m_k + \gamma^2\,\mathbb{E}\big[|A_k|^2\big] + 4\gamma^2\,\mathbb{E}\big[|B_k|^2 + |C_k|^2\big] + 2\gamma\,\mathbb{E}\big[\delta_k \cdot B_k\big] + 2\gamma\,\mathbb{E}\big[\delta_k \cdot C_k\big] \\ + 2\,\mathbb{E}\big[|E_{k+1}|^2\big] + 2\,\mathbb{E}\big[\delta_k \cdot E_{k+1}\big]. \end{aligned}$$

利用引理4的界，在重新排列各项后，我们得到

$$m_{k+1} \le m_k - \frac{3}{2}\gamma(\lambda - \lambda_\star)\, m_k + C_\star\gamma^2\, m_k + C\gamma^2 + \frac{C}{N^3}$$

其中 $C_\star = 8M^4 + 4\big(\lambda + (A+B)L_\theta\big)^2$（见（3））。由于 $N \ge N_\star$，我们有$C_\star\gamma^2 \le \frac{1}{4}\gamma(\lambda - \lambda_\star)$，因此

$$m_{k+1} \le \big(1 - \gamma(\lambda - \lambda_\star)\big)m_k + C\gamma^2 + \frac{C}{N^3}.$$

由于 $\lambda > \lambda_\star$，通过离散 Grönwall 论证和 $m_0 = 0$ 可得

$$\sup_{k\in\mathbb{N}} m_k \le \frac{C\gamma^2}{\gamma(\lambda - \lambda_\star)} \le \frac{C}{N}. \tag{24}$$

最后，将 (22) 与 (24) 代入 (21)，得到

$$\sup_{t\ge 0} \mathbb{E}\big[|\theta^1_{\lfloor Nt\rfloor} - \bar\theta^1_t|\big] \le \sup_{k\in\mathbb{N}} \sqrt{m_k} + \frac{C}{N} \le \frac{C}{\sqrt{N}}.$$

回顾 (20)，命题 3 的证明至此完成。 □

引理 *4* 的证明。 固定 $k \in \mathbb{N}$ 并回顾 $t_k = k/N$，因此 $\mu^N_{t_k}$ 是第 k步的随机梯度下降参数的经验测度。

(i) 根据定义（9）和（14），由于 z_{k+1} 与 $\mathcal{F}_k$ 相互独立，且 θ_k 是 $\mathcal{F}_k$可测的，

$$\mathbb{E}\big[F^1_\lambda(\theta_k, z_{k+1}) \mid \mathcal{F}_k\big] = G_\lambda(\theta^1_k, \mu^N_{t_k}),$$

因此 $\mathbb{E}\big[A_k \mid \mathcal{F}_k\big] = 0$。此外，根据(A1)–(A2)，

$$|A_k| = \big|\big(y_{k+1} - \hat{y}_{\theta_k}(x_{k+1})\big)\nabla_{\theta^1}\sigma_*(\theta_k^1, x_{k+1}) - \mathbb{E}_\pi\big[(y - \hat{y}_{\theta_k}(x))\nabla_{\theta^1}\sigma_*(\theta_k^1, x)\big]\big| \le 2(A+B)M,$$

so that $\mathbb{E}\big[|A_k|^2\big] \le 4(A+B)^2M^2$.

(ii) 令 $B_k = B_{1,k} + B_{2,k}$ 满足

$$B_{1,k} = G_\lambda(\theta_k^1, \mu_{t_k}^N) - G_\lambda(\theta_k^1, \bar{\mu}_{t_k}^N) \qquad B_{2,k} = G_\lambda(\theta_k^1, \bar{\mu}_{t_k}^N) - G_\lambda(\theta_k^1, \bar{\mu}_{t_k}).$$

利用 $\|\nabla_{(\theta^i,x)}\sigma_*\|_\infty \le M$ 以及映射 M 的利普希茨连续性（常数为 $\theta^i \in \mathbb{R}^D \mapsto \sigma_*(\theta^i, x)$），根据 (A1)，可得

$$|B_{1,k}| \le M\,\mathbb{E}_\pi\big[|\hat{y}_{\theta_k}(x) - \langle\sigma_*(\cdot, x), \bar{\mu}_{t_k}^N\rangle|\big] \le \frac{M}{N}\sum_{i=1}^N \mathbb{E}_\pi\big[|\sigma_*(\theta_k^i, x) - \sigma_*(\bar{\theta}_{t_k}^i, x)|\big] \le \frac{M^2}{N}\sum_{i=1}^N |\delta_k^i|.$$

因此，由可交换性（见引理 3），

$$\mathbb{E}\big[|B_{1,k}|^2\big] \le M^4\,\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^N |\delta_k^i|\right)^2\right] \le \frac{M^4}{N}\sum_{i=1}^N \mathbb{E}\big[|\delta_k^i|^2\big] = M^4\,m_k.$$

再次利用 $\|\nabla_{(\theta^i,x)}\sigma_*\|_\infty \le M$ 和詹森不等式，可得

$$|B_{2,k}|^2 \le M^2\,\mathbb{E}_\pi\left[\left|\frac{1}{N}\sum_{i=1}^N \sigma_*(\bar{\theta}_{t_k}^i, x) - \mathbb{E}\big[\sigma_*(\bar{\theta}_{t_k}^1, x)\big]\right|^2\right] = \frac{M^2}{N}\mathbb{E}_\pi\big[\,\mathrm{Var}\left(\sigma_*(\bar{\theta}_{t_k}^1, x)\right)\big] \le \frac{M^2B^2}{N}.$$

结合 $B_{1,k}$ 和 $B_{2,k}$ 的两个界，可得

$$\mathbb{E}\big[|B_k|^2\big] \le 2\,\mathbb{E}\big[|B_{1,k}|^2\big] + 2\,\mathbb{E}\big[|B_{2,k}|^2\big] \le 2M^4\,m_k + \frac{2M^2B^2}{N}.$$

因此，由柯西–施瓦茨不等式，

$$\mathbb{E}\big[\delta_k \cdot B_k\big] \le \sqrt{m_k}\,\sqrt{\mathbb{E}\big[|B_k|^2\big]} \le \sqrt{C}\,m_k + \sqrt{\frac{C}{N}}\sqrt{m_k}.$$

最后，应用杨氏不等式

$$\forall a, b \in \mathbb{R},\ \forall \varepsilon > 0, \quad ab \le \frac{\varepsilon}{2}a^2 + \frac{1}{2\varepsilon}b^2 \tag{25}$$

取 $a = \sqrt{m_k}$、$b = \sqrt{C/N}$ 和 $\varepsilon = 2M^2$，我们得到

$$\mathbb{E}\big[\delta_k \cdot B_k\big] \le M^2\,m_k + \frac{C}{N}.$$

(iii) 根据定义（14），

$$C_k = \mathbb{E}_\pi\big[\big(y - \langle\sigma_*(\cdot, x), \bar{\mu}_{t_k}^N\rangle\big)\big(\nabla_{\theta^1}\sigma_*(\theta_k^1, x) - \nabla_{\theta^1}\sigma_*(\bar{\theta}_{t_k}^1, x)\big)\big] - \lambda\delta_k.$$

利用柯西–施瓦茨不等式和 (A3)，我们有

$$\big|\mathbb{E}_\pi\big[\big(y - \langle\sigma_*(\cdot, x), \bar{\mu}_{t_k}^N\rangle\big)\big(\nabla_{\theta^1}\sigma_*(\theta_k^1, x) - \nabla_{\theta^1}\sigma_*(\bar{\theta}_{t_k}^1, x)\big)\big]\big| \le (A+B)L_\theta\,|\delta_k|.$$

因此

$$\mathbb{E}\big[\delta_k \cdot C_k\big] \le -\big(\lambda - (A+B)L_\theta\big)\,m_k,$$
$$\mathbb{E}\big[|C_k|^2\big] \le \big(\lambda + (A+B)L_\theta\big)^2\,m_k.$$

(iv)将 E_{k+1} 改写为

$$E_{k+1} = \alpha \int_{t_k}^{t_{k+1}} \left(G_\lambda(\bar{\theta}^1_{t_k}, \bar{\mu}_{t_k}) - G_\lambda(\bar{\theta}^1_s, \bar{\mu}_s) \right) ds.$$

利用利普希茨界

$$|G_\lambda(\theta, \mu) - G_\lambda(\theta', \mu)| \le \left(\lambda + (A+B)L_\theta\right) |\theta - \theta'|,$$
$$|G_\lambda(\theta, \mu) - G_\lambda(\theta, \nu)| \le M^2 \, \mathsf{W}_1(\mu, \nu),$$

对所有 θ, $\theta' \in \mathcal{E}$ 以及 μ, $\nu \in \mathcal{P}(\mathbb{R}^D)$，结合以下事实： $s \mapsto \bar{\theta}^1_s$ 关于 L^1 是利普希茨函数，且 $s \mapsto \bar{\mu}_s$ 关于 W_1 是利普希茨函数（其常数由 $(A+B)M + \lambda \sup_{t \ge 0} \mathbb{E}\left[|\bar{\theta}^1_t|\right] < \infty$在 (A5)条件下控制），我们得到

$$\mathbb{E}\left[|G_\lambda(\bar{\theta}^1_{t_k}, \bar{\mu}_{t_k}) - G_\lambda(\bar{\theta}^1_s, \bar{\mu}_s)|\right] \le C \, |s - t_k|.$$

因此，

$$\mathbb{E}\left[|E_{k+1}|\right] \le \alpha \int_{t_k}^{t_{k+1}} C(s - t_k) \, ds \le \frac{C}{N^2},$$

类似地，

$$\mathbb{E}\left[|E_{k+1}|^2\right] \le \frac{C}{N^4}.$$

最后，由杨氏不等式（25）以及 $a = |\delta_k|$， $b = |E_{k+1}|$ 和 $\varepsilon = \frac{1}{2}\gamma(\lambda - \lambda_\star)$，可得

$$\mathbb{E}\left[\delta_k \cdot E_{k+1}\right] \le \frac{1}{4}\gamma(\lambda - \lambda_\star) \, m_k + \frac{C}{\gamma(\lambda - \lambda_\star) N^4} \le \frac{1}{4}\gamma(\lambda - \lambda_\star) \, m_k + C\gamma^2.$$

证明至此完成。 □

4 推广至无界激活函数

在本节中，我们将前面的结果推广到不再一致有界的激活函数。为此，我们去掉假设 (A1)，并用以下局部化假设 (B1)–(B3)取而代之。我们证明，在这些条件下（并且适当选择 λ），随机梯度下降迭代始终一致有界，对所有时刻 均成立，因此始终处于某个确定性紧集内。一旦建立了局部化，分析便可归结为该紧集上的有界情形。

4.1 局部化假设

下文中，我们用以下三个假设替代 (A1)。

(B1) 存在一个紧集 $K_x \subset \mathbb{R}^d$ 和半径 $R_0 > 0$，使得

$$\operatorname{supp} \pi \subset K_x \times [-A, A], \qquad \operatorname{supp} \mu_0 \subset \overline{B}(0, R_0) \subset \mathbb{R}^D.$$

(B2) 存在 b、 $c \ge 0$，使得对所有 $(\theta^i, x) \in \mathbb{R}^D \times K_x$, $|\sigma_*(\theta^i, x)| \le b + c|\theta^i$ |.

(B3) 激活函数 $\sigma_* \in \mathcal{C}^1(\mathbb{R}^D \times \mathbb{R}^d)$ 连续可微，且在 $\mathbb{R}^D \times K_x$ 上的梯度一致有界：

$$\forall (\theta^i, x) \in \mathbb{R}^D \times K_x, \quad |\nabla_{(\theta^i, x)} \sigma_*(\theta^i, x)| \le M.$$

请注意，(B1) 蕴含(A2)。此外，根据霍夫丁不等式，(B1)保证了 Lipschitz 可观测量满足次高斯型拉普拉斯界，这等价于一个 T_1 不等式。然而，单凭这一条件通常不足以推出一个 T_2 不等式。(B3) 中的光滑性条件纯粹是技术性的。具体而言，它排除了 ReLU 等非光滑激活函数；也可以改用平滑逼近（例如 softplus）或分段光滑变体（例如 leaky-ReLU），而不影响下文的局部化论证。

4.2 一致局部化界

我们将使用以下常数：

$$a_\infty = R_0 \vee \frac{M(A+b)}{\lambda - Mc}, \qquad R_\infty = R_0 \vee \frac{M(A+b+ca_\infty)}{\lambda}.$$

引理 5。 假设 *(B1)–(B3)*，且 $\lambda > Mc$，$\gamma\lambda \le 1$。那么，几乎必然有

$$\sup_{k\in\mathbb{N}} \max_{1\le i\le N} |\theta_k^i| \le R_\infty.$$

证明。 由（1）和 (B1)–(B3)，对所有 $k \in \mathbb{N}$ 和 $i \in \{1,\dots,N\}$，有

$$|\theta_{k+1}^i| \le (1-\gamma\lambda)|\theta_k^i| + \gamma M|y_{k+1} - \hat{y}_{\theta_k}(x_{k+1})| \le (1-\gamma\lambda)|\theta_k^i| + \gamma M(A+b+ca_k)$$

其中 $a_k = \dfrac{1}{N}\sum_{i=1}^{N} |\theta_k^i|$。对 i 取平均，得到

$$a_{k+1} \le \big(1-\gamma(\lambda - Mc)\big)\, a_k + \gamma M(A+b). \tag{26}$$

在 $\gamma\lambda \le 1$ 和 $\lambda > Mc$ 的条件下，有 $0 \le 1-\gamma\lambda \le 1$ 和 $\gamma(\lambda - Mc) \le \gamma\lambda \le 1$，因此 $1-\gamma(\lambda - Mc) \in [0,1)$。由（26）的迭代可得

$$\sup_{k\in\mathbb{N}} a_k \le a_0 \vee \frac{M(A+b)}{\lambda - Mc} \le a_\infty$$

其中我们使用了 $a_0 \le$ 最大值$_{1\le i\le N} |\theta_0^i| \le R_0$，由 (B1)得到。将此界代回（26）可得

$$|\theta_{k+1}^i| \le (1-\gamma\lambda)|\theta_k^i| + \gamma M(A+b+ca_\infty).$$

最后，再次迭代此不等式并利用 (B1)，我们便可对所有 $k \in \mathbb{N}$ 以及每个$i \in \{1,\dots,N\}$

$$\begin{aligned} |\theta_k^i| &\le (1-\gamma\lambda)^k\, |\theta_0^i| + \frac{M(A+b+ca_\infty)}{\lambda}\big(1-(1-\gamma\lambda)^k\big) \\ &\le R_0 \vee \frac{M(A+b+ca_\infty)}{\lambda} = R_\infty, \end{aligned}$$

这就是所要证明的界。 □

注。 这里，条件 $\lambda > Mc$ 与有界情形中的下界 $\lambda > \lambda_\star$ 起着相同的作用：它确保岭正则化强于激活函数增长所引起的有效相互作用。额外的要求 $\gamma\lambda \le 1$ 是一个方便的离散时间稳定性条件，可确保 $1-\gamma\lambda \ge 0$，并且在缩放 $\gamma = \alpha/N$ 下，当 N 足够大时自动满足。

4.3 化归到有界情形

引理 5 表明，在满足 (B1)–(B3) 且正则化足够强的条件下，随机梯度下降迭代始终位于闭球 $\overline{B}(0, R_\infty)$ 内。于是，我们可以在紧集$\overline{B}(0, R_\infty) \times K_x$上进行分析；在该集合上，激活函数及其梯度均一致有界。为了应用第 3节中的论证，我们还要求 $\nabla_{\theta^i}\sigma_*$ 满足局部 Lipschitz 性质，以替代(A3)：

(B4) 对于每个 $R > 0$，都存在常数 $L_x(R)$ $L_\theta(R) > 0$，使得对于所有 x、$x' \in K_x$ 以及 θ^i、$\theta'^i \in \overline{B}(0,R) \subset \mathbb{R}^D$，

$$\begin{aligned} |\nabla_{\theta^i}\sigma_*(\theta^i, x) - \nabla_{\theta^i}\sigma_*(\theta^i, x')| &\le L_x(R)\,|x - x'|, \\ |\nabla_{\theta^i}\sigma_*(\theta^i, x) - \nabla_{\theta^i}\sigma_*(\theta'^i, x)| &\le L_\theta(R)\,|\theta^i - \theta'^i|. \end{aligned}$$

假设 (B4) 在当前设定下并不苛刻：例如，只要 σ_* 是 $\mathcal{C}^2$ 且在紧集上具有有界的混合二阶导数，该假设便成立；这包括实践中使用的非光滑激活函数的标准光滑近似。因此，第 3 节的论证只需稍作符号上的调整即可沿用，只需将全局常数（B、 M、 L_x、 L_θ）替换为其在 $\overline{B}(0,\ R_\infty) \times K_x$ 上的局部化对应项。

致谢

A.G 获得 ANR-23-CE-40003 和 Conviviality 的资助，并受益于法国国家研究署在法国2030投资计划下管理的政府拨款 ANR-23-EXMA-0001。P.S 获得克莱蒙I-SITE项目 CAP 2025 的资助。

参考文献

[1] Francis Bach。《从第一原理出发的学习理论》。自适应计算与机器学习。麻省理工学院出版社，2024。[2] Erhan Bayraktar 和 Ruoyu Wu。“Graphon粒子系统：一致时间浓度界。”《随机过程及其应用》156（2023年2月），第196–225页。doi: 10.1016/j.spa.2022.11.008。[3] Gordon Blower 和 François Bolley。“乘积空间上的测度集中及其在马尔可夫过程中的应用。”《数学研究》175（2006），第47–72页。doi: 10.4064/sm175-1-3。[4] Sergey G. Bobkov 和 Friedrich Götze。“对数Sobolev不等式相关的指数可积性与运输成本。”《泛函分析杂志》163.1（1999），第1–28页。doi: 10.1006/jfan.1998.3326。[5] François Bolley、Arnaud Guillin 和 Cédric Villani。“非紧空间上经验测度的定量浓度不等式。”《概率论及相关领域》137.3–4（2007），第541–593页。doi: 10.1007/s00440-006-0004-7。[6] Nicolas Bonneel 等。“测度的切片与Radon Wasserstein重心。”《数学成像与视觉杂志》51.1（2015），第22–45页。doi: 10.1007/s10851-014-0506-3。[7] Léon Bottou。“使用随机梯度下降进行大规模机器学习。”COMPSTAT’2010论文集。Physica-Verlag，2010，第177–186页。doi: 10.1007/978-3-7908-2604-3_16。[8] Léon Bottou、Frank E. Curtis 和 Jorge Nocedal。“大规模机器学习的优化方法。”《SIAM评论》60.2（2018），第223–311页。doi: 10.1137/16M1080173。[9] Shi Chen 等。均值场Transformer模型中的定量聚类。2025。arXiv: 2504。14697 [cs.LG]。[10] Lénaïc Chizat 和 Francis Bach。“利用最优传输研究过参数化模型中梯度下降的全局收敛。”第32届国际神经信息处理系统会议论文集（蒙特利尔，加拿大）。Curran Associates Inc.，2018，第3040–3050页。[11] Valentin De Bortoli 等。“宽神经网络中随机梯度下降的定量混沌传播。”《神经信息处理系统进展》。第33卷。Curran Associates, Inc.，2020，第278–288页。[12] Jérôme Dedecker、Aurélie Fischer 和 Bertrand Michel。Wasserstein距离下经验测度的集中：涉及覆盖维数的界。2026。arXiv: 2601。11133 [math.PR]。[13] Pierre Del Moral 和 Emmanuel Rio。“均值场粒子模型的浓度不等式。”《应用概率年刊》21.3（2011），第1017–1052页。doi: 10.1214/10-AAP716。[14] François Delarue 和 Alvin Tse。“环面上的一致时间弱混沌传播。”《亨利·庞加莱研究所年刊：概率与统计》61.2（2025年5月），第1021–1074页。doi: 10.1214/23-AIHP1451。[15] Arnaud Descours。“通过均值场分析，从数学角度理解深度神经网络。”博士论文。克莱蒙奥弗涅大学，2023。url: https://theses.hal.science/tel-045[16] Arnaud Descours 等。“宽两层神经网络的大数定律与中心极限定理：小批量与含噪情形。”机器学习研究杂志25（2024），第1–76页。

[17] Hacène Djellout、Arnaud Guillin 和 Liming Wu。“运输成本-信息不等式及其在随机动力系统与扩散过程中的应用。”《概率年刊》32.3B（2004），第2702–2732页。doi: 10.1214/009117904000000531。[18] Alain Durmus 等。“一致时间混沌传播的初等方法。”《美国数学会会刊》（2018）。[19] Cong Fang 等。“从特征建模：过参数化深度神经网络的平均场框架。”第三十四届学习理论会议（COLT）论文集。第134卷。《机器学习研究论文集》。PMLR，2021，第1887–1936页。[20] Nicolas Fournier 和 Arnaud Guillin。“经验测度在Wasserstein距离下的收敛速率。”《概率论及相关领域》162（2014），第707–738页。doi: 10.1007/s00440-014-0583-7。[21] Ian Goodfellow、Yoshua Bengio 和 Aaron Courville。《深度学习》。麻省理工学院出版社，2016。[22] Nathael Gozlan 和 Christian Léonard。“传输不等式：综述。”《马尔可夫过程及相关领域》16（2010年第4期），第635–736页。[23] Daniel Lacker 和 Luc Le Flem。“尖锐的一致时间混沌传播。”《概率论及相关领域》187.1-2（2023），第443–480页。doi: 10.1007/s00440-023-01192-x。[24] Yann LeCun、Yoshua Bengio 和 Geoffrey Hinton。“深度学习。”《自然》521.7553（2015），第436–444页。doi: 10.1038/nature14539。[25] Michel Ledoux。《测度集中现象》。第89卷。《数学调查与专著》。普罗维登斯，罗得岛州：美国数学会，2001。[26] Michel Ledoux。测度集中、运输成本与泛函不等式。《数学模型中的奇异现象与尺度变换暑期学校》。德国波恩，2003。url: https://lsp.ups-tlse.fr/Ledoux。[27] Song Mei、Andrea Montanari 和 Phan-Minh Nguyen。“两层神经网络景观的平均场视角。”《美国国家科学院院刊》115.33（2018），E7665–E7671。doi: 10.1073/pnas.1806579115。[28] Kimia Nadjahi 等。“切片概率散度的统计与拓扑性质。”《神经信息处理系统进展》33：2020年神经信息处理系统年会，2020年12月6–12日，线上举行。Hugo Larochelle 等编。2020。[29] Kimia Nadjahi 等。“利用随机投影的测度集中快速逼近切片Wasserstein距离。”《神经信息处理系统进展》34：2021年神经信息处理系统年会，2021年12月6–14日，线上举行。Marc'Aurelio Ranzato 等编。2021，第12411–12424页。[30] Phan-Minh Nguyen。多层神经网络学习动力学的平均场极限。2019。arXiv: 1902。02880 [stat.ML]。[31] Phan-Minh Nguyen 和 Huy Tuan Pham。“多层神经网络平均场极限的严谨框架。”《数理统计与学习》6（2023），第201–357页。doi: 10.4171/MSL/42。[32] Felix Otto 和 Cédric Villani。“Talagrand不等式的推广及其与对数Sobolev不等式的联系。”《泛函分析杂志》173.2（2000），第361–400页。doi: 10.1006/jfan.1999.3557。[33] Gabriel Peyré。面向机器学习者的最优传输。2025。arXiv: 2505。06589 [stat.ML]。[34] Gabriel Peyré 和 Marco Cuturi。“计算最优传输。”《机器学习基础与趋势》11.5–6（2019），第355–607页。doi: 10.1561/2200000073。[35] Philippe Rigollet。Transformer的平均场动力学。2025。arXiv: 2512。01868 [cs.LG]。[36] Grant M. Rotskoff 和 Eric Vanden-Eijnden。“神经网络的可训练性与精度：相互作用粒子系统方法。”《纯粹与应用数学通讯》75.9（2022），第1889–1935页。doi: 10.1002/cpa.22074。[37] David Saad 和 Sara A. Solla。“多层神经网络在线梯度下降学习的动力学。”《神经信息处理系统进展》。1995。[38] Filippo Santambrogio。《应用数学家的最优传输》。变分法、偏微分方程与建模。第87卷。《非线性微分方程及其应用进展》。Cham：Birkhäuser，2015。doi: 10.1007/978-3-319-20828-2。

[39] Jürgen Schmidhuber. “神经网络中的深度学习：综述。” 神经网络 61 (2015), 第85–117页. doi: 10.1016/j.neunet.2014.09.003.[40] Shai Shalev-Shwartz与Shai Ben-David. 《理解机器学习：从理论到算法》. 剑桥大学出版社, 2014.[41] Justin Sirignano与Konstantinos Spiliopoulos. “神经网络的均值场分析：中心极限定理。” 随机过程及其应用 130.3 (2020), 第1820–1852页. doi: 10.1016/j.spa.2019.06.003.[42] Justin Sirignano与Konstantinos Spiliopoulos. “神经网络的均值场分析：大数定律。” SIAM应用数学杂志 80.2 (2020), 第725–752页. doi: 10.1137/18M1192184.[43] Justin Sirignano与Konstantinos Spiliopoulos. “深度神经网络的均值场分析。” 运筹数学 47.1 (2022年2月), 第120–152页. doi: 10.1287/moor.2020.1118.[44] Namjoon Suh与Guang Cheng. “深度学习统计理论综述：逼近、训练动力学与生成模型。”《统计学及其应用年评》12 (2025), 第177–207页.[45] Alain-Sol Sznitman. “混沌传播专题。” 圣弗卢尔概率暑期学校 XIX – 1989. 数学讲义. 柏林、海德堡：Springer, 1991, 第165–251页.[46] Cédric Villani. 《最优传输：旧与新》. 数学科学基础丛书. 柏林、海德堡：Springer, 2008. doi: 10.1007/978-3-540-71050-9.[47] James Vuckovic. 《非线性马尔可夫链的混沌传播》. 2026. arXiv: 2602.07537 [math

.PR].