arXiv:2304.06847v1 [math.PR] 13 Apr 2023

最小二乘问题中单遍随机梯度下降的高维极限。

Elizabeth Collins–Woodfin* Elliot Paquette †

摘要

我们描述了最小二乘问题中单遍单批次随机梯度下降（SGD）的高维极限。该极限是在步长不趋于零，且样本数量与问题维数成比例的情况下得到的。我们用高维随机微分方程来描述这一极限，并证明该方程能够近似随机梯度下降的状态演化。作为推论，我们证明，随着维数趋于无穷，统计风险可由一个卷积型沃尔泰拉方程的解近似，且误差趋于零。这里的收敛意义取为能够保证两个过程的统计风险一致的最弱意义。与现有分析相比，本文的区别在于所给出的高维极限类型，以及对样本协方差结构的更一般处理。

关键词： 随机梯度下降；随机矩阵理论；优化；随机微分方程。

MSC2020 学科分类： 60H30。

1 引言

随机优化方法是当今许多大规模计算任务的标准方法，尤其适用于机器学习中出现的任务。对这些算法的分析历史悠久，最早可追溯到 [RM51],的开创性工作，该工作聚焦于固定维度空间中的长期行为。然而，现代随机优化的应用促使研究者采用另一种分析范式：问题维度随算法运行时间成比例增长。

本文推导了维度趋于无穷时，最小二乘问题中随机梯度下降（SGD）在单遍设定下（见下文）的精确标度行为。我们还将结果与近期工作 [Paq+22；Paq+21]进行比较，该工作研究了这一问题的多遍版本。

用于经验风险最小化的随机梯度下降 大多数版本 of
（小批量）随机梯度下降可以在有限和问题的框架下进行表述：

$$\min_{\boldsymbol{x}\in\mathbb{R}^d}\left\{f(\boldsymbol{x}) \coloneqq \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x})\right\}. \tag{1.1}$$

*加拿大麦吉尔大学。电子邮件： elizabeth.collins-woodfin@mail.mcgill.ca https://sites.google.com/view/e-collins-woodfin

†麦吉尔大学，加拿大。电子邮件： elliot.paquette@mcgill.ca http://elliotpaquette.github.io

在此背景下，经验风险最小化假设存在 n 个来自某种数据分布的独立样本，并且每个样本 f_i 表示参数在某个模型中拟合 $\boldsymbol{x}$ 第 i个数据点时的损失。本文将只考虑带有 ℓ^2正则项的线性回归情形。因此，我们假设存在n 个独立同分布样本 $((\boldsymbol{a}_i, \boldsymbol{b}_i) : 1 \le i \le n)$，它们来自某个分布 $\mathcal{D}$，并满足一些将在后文说明的假设。我们将这些数据整理成设计矩阵 $\boldsymbol{A}$ 和标签向量 $\boldsymbol{b}$，其中第 i 行为 $\boldsymbol{a}_i$。最后，我们通过设定以下函数来定义 f_i （1.1）中的函数</style}

$$f_i(\boldsymbol{x}) = \tfrac{1}{2}(\boldsymbol{a}_i \cdot \boldsymbol{x} - \boldsymbol{b}_i)^2 + \tfrac{\delta}{2}\|\boldsymbol{x}\|^2.$$

参数 $\delta \ge 0$ 固定，表示正则化项的强度；全文中，$\|\cdot\|$表示欧几里得范数。

在此背景下，小批量随机梯度下降可描述为

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - \gamma_k \nabla f_{i_k}(\boldsymbol{x}_k) \\ &= \boldsymbol{x}_k - \gamma_k \boldsymbol{A}^T \boldsymbol{e}_{i_k} \boldsymbol{e}_{i_k}^T (\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}) - \gamma_k \delta \boldsymbol{x}_k \end{aligned} \tag{1.2}$$

其中，$\{\gamma_k\}$ 是步长参数，$\boldsymbol{e}_i$ 是第 i 个标准基向量，而 $\{i_k\}$ 是数据选择序列。

本文考虑单遍 情形，此时 $i_k = k$，但算法在执行 n 步后终止。实际上，数据点的顺序可能会在开始前打乱一次；但在我们所设定的独立同分布数据情形下，没有必要加入这项额外的随机化。选择 i_k, 的方式还有其他选择，我们重点介绍其中三种，它们都是多遍 变体。

在有放回随机抽样随机梯度下降中，每个 i_k 都从 $\{1, 2, \dots, n\}$中均匀随机选取。这是 [Paq+22] 所研究的情形，在本文大部分内容中，我们将这种随机梯度下降简称为多遍随机梯度下降。不过，为了提供背景，我们也会提及单次乱序随机梯度下降：其中取 $i_k = k \bmod n$，因此它与单遍随机梯度下降的唯一区别在于，算法每个训练轮次都会执行相同的操作[1] 在随机乱序随机梯度下降中，则通过在每个训练轮次之间随机排列数据来修改上述策略。所有这些策略都在优化文献中得到了广泛研究：人们通常认为，单次乱序和随机乱序策略比随机抽样策略更快 [YSJ21]（另见 [RR12; GOP21; SS20; AYS20]）。

单遍情形是所有这些方法进行比较的基础，既在现象层面更为简单，也代表了随机梯度下降的一种理想化情形：算法的运行时间就是数据量。适当地延长运行时间（即增加 n）只会提升随机梯度下降估计量的统计性能 $\boldsymbol{x}_n$；在这一语境下，这通常被称为 流式随机梯度下降。

随机梯度下降的性能通过总体风险 $\mathcal{P}$ 来衡量，有时还 s
通过带有 ℓ_2正则化的风险 $\mathcal{R}$来衡量，它们分别表示为

$$\mathcal{P}(\boldsymbol{x}) \coloneqq \tfrac{1}{2}\mathbb{E}_{(\boldsymbol{a},\boldsymbol{b})}(\boldsymbol{a} \cdot \boldsymbol{x} - \boldsymbol{b})^2, \qquad (\boldsymbol{a}, \boldsymbol{b}) \sim \mathcal{D}, \tag{1.3}$$

$$\mathcal{R}(\boldsymbol{x}) \coloneqq \tfrac{1}{2}\mathbb{E}_{(\boldsymbol{a},\boldsymbol{b})}(\boldsymbol{a} \cdot \boldsymbol{x} - \boldsymbol{b})^2 + \tfrac{\delta}{2}\|\boldsymbol{x}\|^2, \qquad (\boldsymbol{a}, \boldsymbol{b}) \sim \mathcal{D}. \tag{1.4}$$

这种正则化风险自然表现为单遍随机梯度下降的平均行为，即

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma_k(\nabla \mathcal{R}(\boldsymbol{x}_k) + \xi_{k+1}),$$

对于鞅增量 $(\xi_k : 1 \le k \le n)$。另一个自然的统计场景是分布外回归，在这种情况下，我们将衡量在 $\mathcal{D}$ 上训练的随机梯度下降的性能，但测试时，如（1.3）所示，将 $\mathcal{D}$ 替换为另一个分布 $\mathcal{D}'$。我们不会详细讨论这一情形，但需要指出的是，上述所有例子都是随机梯度下降状态 $\boldsymbol{x}$的某些二次型泛函。

[1]在所有这些情形中（包括有放回情形），我们将 epoch 定义为 n 算法的迭代步数。

数据和步长假设 本分析旨在允许样本数 n 较大，并与问题的维度 d成正比。这意味着必须对数据进行归一化，使其几乎不依赖于维度。此外，我们还需要所涉及的一些随机变量具有良好的尾部性质，因此回顾一下 $p \geq 1$ 的 Orlicz 范数 $\|\cdot\|_{\psi_p}$，其定义为

$$\|X\|_{\psi_p} = \inf\{t : \mathbb{E}e^{|X|^p/t^p} \leq 2\}.$$

我们请读者参阅 [Ver18]，以进一步了解相关阐述、性质和等价公式。

在下文中，我们始终假设，在 $\mathcal{D}$ 条件下，标签由一个底层带噪线性模型给出。形式化地说，我们假设：

假设 1.1。对于从 $\mathcal{D}$ 中采样得到的 $(\boldsymbol{a}, \boldsymbol{b})$，在给定 $\boldsymbol{a}$ 的条件下，$\boldsymbol{b}$ 的分布由 $\boldsymbol{a}\cdot\tilde{\boldsymbol{x}}+w$ 给出，其中 w 的均值为0，方差为 $\eta^2 \geq 0$，并且是次高斯的，满足 $\|w\|_{\psi_2} \leq d^{\varepsilon}$。假设真实参数 $\tilde{\boldsymbol{x}}$ 的范数至多为 d^{ε}。

常数 ε 将始终取为一个足够小的固定值。小于1 18 的任意值都可以。 —

假设数据协方差经过归一化，使其几乎 st

范数有界，也就是说：

假设 1.2。协方差矩阵 $\boldsymbol{K} := \mathbb{E}[\boldsymbol{a}\boldsymbol{a}^T]$ 的算子范数有界，且上界与 d无关

请注意，虽然我们没有明确假设 $\boldsymbol{a}$ 是中心化的，但若要满足上述假设，其均值在某种意义上必须足够小。

最后，我们假设 $\boldsymbol{a}$具有良好的尾部性质，即

假设 1.3。数据向量 $\boldsymbol{a}$ 满足：对于任意范数小于 1的确定性 $\boldsymbol{x}$，$\|\boldsymbol{a}\cdot\boldsymbol{x}\|_{\psi_2} \leq d^{\varepsilon}$；并且数据向量 $\boldsymbol{a}$ 满足 *Hanson-Wright* 不等式：对于所有 $t \geq 0$ 以及任意确定性矩阵 $\boldsymbol{B}$

$$\mathbb{P}\left(\left|\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a} - \mathbb{E}\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a}\right| \geq t\right) \leq 2\exp\left(-\min\left\{\frac{t^2}{d^{4\varepsilon}||\boldsymbol{B}||_{HS}^2}, \frac{t}{d^{2\varepsilon}||\boldsymbol{B}||}\right\}\right).$$

我们指出，这些假设适用于两种重要情形：

(a) 当 $\boldsymbol{a} = \sqrt{\boldsymbol{K}}\boldsymbol{u}$其中 $\boldsymbol{K}$ 是一个算子范数有界的确定性矩阵，且 $\boldsymbol{u}$ 是一个由独立同分布次高斯随机变量构成的向量；或者

(b) 当 $\boldsymbol{a}$ 是一个具有凸集中性质的向量，详见 [Ada15]。

第二种情形有一些自然的例子，例如具有利普希茨激活函数的随机特征模型 [RR08]（另见 [Paq+22, 命题6。2] 有关随机特征情形的具体说明）。我们指出，通过截断，也可以在上述情形(a)下仅假设某个足够高但有限阶矩一致有界来进行研究，不过本文不作探讨。

最后，步长参数 γ_k 必须进行适当归一化：

假设1.4。对所有 k 以及固定的 $\gamma > 0$，步长 $\gamma_k = \frac{\gamma}{d}$。

我们指出，也可以对有界连续函数 $\gamma : [0,\infty) \to [0,\infty)$选取 $\gamma_k = \gamma(k/d)/d$，这样不会对论证的任何部分造成影响。

鉴于以上所有假设，我们注意到，随机梯度下降最终可化为以下随机递推

$$\boldsymbol{x}_k - \tilde{\boldsymbol{x}} = (\boldsymbol{I}(1-\tfrac{\gamma\delta}{d}) - \gamma\boldsymbol{m}_k\boldsymbol{m}_k^T)(\boldsymbol{x}_{k-1} - \tilde{\boldsymbol{x}}) - \tfrac{\gamma\delta}{d}\tilde{\boldsymbol{x}} + \gamma\boldsymbol{m}_k\eta_k, \tag{1.5}$$

其中 $\eta_k = w_k/\sqrt{d}$ 和 $\boldsymbol{m}_k = \boldsymbol{a}_k/\sqrt{d}$。 ¯

均质化随机梯度下降 我们的定理最容易表述为：随机梯度下降的状态可以与高维空间中的某个扩散模型进行比较。均质化随机梯度下降定义为一个初始条件为 $\boldsymbol{X}_0 = \boldsymbol{x}_0$ 的连续时间过程，它满足以下随机微分方程

$$\mathrm{d}\boldsymbol{X}_t = -\gamma\nabla\mathcal{R}(\boldsymbol{X}_t)\mathrm{d}t + \gamma\sqrt{\tfrac{2}{d}\mathcal{P}(\boldsymbol{X}_t)\boldsymbol{K}}\mathrm{d}B_t \tag{1.6}$$

其中 B_t 是 d维标准布朗运动，且 $\mathcal{R}$ 和 $\mathcal{P}$ 分别是正则化风险和未正则化风险（回顾（1.3）和（1.4））。

我们的主要定理表明，对于二次统计量（特别是风险（1.3）和（1.4）），均质化随机梯度下降与随机梯度下降在首阶近似下可以互换。我们使用概率修饰语以压倒性概率表示某个陈述成立，除非发生概率至多为 $e^{-\omega(\log d)}$ 的事件，其中 ω（$\log\ d$）趋于 ∞，且速度快于 d的对数，当 $d \to \infty$. 时。我们还在二次函数 q： $\mathrm{R}^d \to \mathrm{C}$上引入一个范数 $\|\cdot\|_{C^2}$

$$\|q\|_{C^2} \coloneqq \|\nabla^2 q\| + \|\nabla q(0)\| + |q(0)|,$$

右侧的范数分别是算子范数和欧几里得范数。

定理 1.5。对于任意二次函数 q： $\mathrm{R}^d \to \mathrm{R}$，以及任意满足 $\boldsymbol{x}_0$ 且 $\|\boldsymbol{x}_0\| \leq 1$的确定性初始化，存在常数 $C(\|\boldsymbol{K}\|)$ 使得过程 $\{\boldsymbol{x}_k\}_{k=0}^n$ 和 $\{\boldsymbol{X}_t\}_{t=0}^{n/d}$ 对于任意满足 n 的 $n \leq d\log\ d/C(\|\boldsymbol{K}\|)$

$$\sup_{0\leq k\leq n}\left|q(\boldsymbol{x}_k) - q(\boldsymbol{X}_{k/d})\right| < \|q\|_{C^2}\cdot e^{C(\|\boldsymbol{K}\|)\frac{n}{d}}\cdot d^{-\frac{1}{2}+9\varepsilon} \tag{1.7}$$

以压倒性概率成立。

过程 $\boldsymbol{x}_k$ 和 $\boldsymbol{X}_t$ 相互独立，因此这也是关于集中的陈述。特别地，将 q（$\boldsymbol{X}_{k/d}$）替换为$\mathbb{E}q$（$\boldsymbol{X}_{k/d}$）后，该陈述仍然成立。

显式风险曲线 利用现有理论（见 [Paq+22，定理 1.1]），$\mathcal{P}$（$\boldsymbol{X}_{k/d}$）和$\mathcal{R}$（$\boldsymbol{X}_{k/d}$）可以看出会集中于其均值附近，而这些均值满足一个卷积沃尔泰拉方程。具体而言，$\mathbb{E}\mathcal{P}$（$\boldsymbol{X}_t$）$= \Psi$（t）以及 $\mathbb{E}\mathcal{R}$（$\boldsymbol{X}_t$）$= \Omega$（t），其中

$$\begin{pmatrix}\Psi(t)\\ \Omega(t)\end{pmatrix} = \begin{pmatrix}\mathcal{P}(\mathcal{X}_{\gamma t})\\ \mathcal{R}(\mathcal{X}_{\gamma t})\end{pmatrix} + \int_0^t \begin{pmatrix}K(t-s;\nabla^2\mathcal{R})\Psi(s)\\ K(t-s;\nabla^2\mathcal{P})\Psi(s)\end{pmatrix}\mathrm{d}s, \quad \begin{cases}K(t;\boldsymbol{M}) \coloneqq \frac{\gamma^2}{d}\operatorname{tr}(\boldsymbol{K}\boldsymbol{M}e^{-2\gamma t(\boldsymbol{K}+\delta\boldsymbol{I})}),\\ \mathrm{d}\mathcal{X}_{\gamma t} \coloneqq -\gamma\nabla\mathcal{R}(\mathcal{X}_{\gamma t}), \quad \mathcal{X}_0 = \boldsymbol{X}_0.\end{cases}$$

过程 $\mathcal{X}_t$ 是梯度流，对于这里的问题，它可以显式求解。注意，关于 Ψ 的方程是自治的，而 Ω 的解则可以用它表示并求解。由这个沃尔泰拉方程很容易推导出收敛速率、收敛阈值，以及极限风险的表达式（在双重标度极限中，先取 $d \to \infty$，再取 $t \to \infty$）。参见 [Paq+21] 中的讨论，并参见图 1 中的示例。

与多遍情形的比较 在 [Paq+22] 中，定理1.5 已针对（随机样本）多遍情形得到证明。在该情形下，扩散过程有所不同。定义经验风险 $\mathcal{L}(\boldsymbol{x}) = \frac{1}{2n}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2$ 和正则化经验风险$f(\boldsymbol{x})$ $= \mathcal{L}(\boldsymbol{x})$ $+ \delta\|\boldsymbol{x}\|^2$。那么，多遍情形下的均质化随机梯度下降变为

$$\mathrm{d}\boldsymbol{X}_t = -\gamma\nabla f(\boldsymbol{X}_t)\mathrm{d}t + \gamma\sqrt{\tfrac{2}{d}\mathcal{L}(\boldsymbol{X}_t)(\tfrac{1}{n}\boldsymbol{A}^T\boldsymbol{A})}\mathrm{d}B_t \tag{1.8}$$

因此，多遍情形与单遍情形之间的差异在于，总体风险被经验风险所替代，数据协方差矩阵 $\boldsymbol{K}$ 被经验数据协方差矩阵所替代 ($\frac{1}{n}\boldsymbol{A}^T\boldsymbol{A}$)。请注意，如果以数据 ($\boldsymbol{A}$， $\boldsymbol{b}$)为条件，那么多遍随机梯度下降实际上 就是 流式随机梯度下降，只不过其数据分布是有限支撑的（具体而言，是数据的经验分布）；然而，数据的经验分布与满足假设1.3相距甚远。

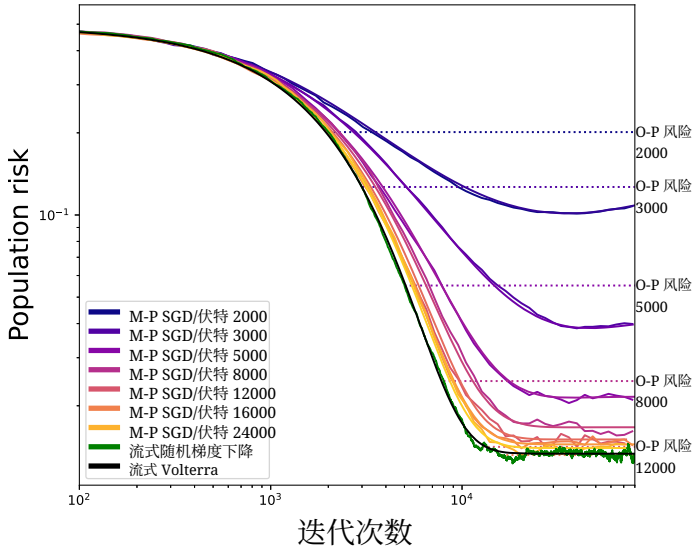


图 1: $d = 2000$中一个简单线性回归问题的风险曲线。图中绘出了多遍随机梯度下降、其高维对应形式（均质化随机梯度下降下的期望风险，即“Volterra”）、流式随机梯度下降（即数据集大小不断变化的单遍随机梯度下降），以及均质化随机梯度下降的期望风险（“流式 Volterra”）。同时绘出了流式随机梯度下降在不同水平 n 下的风险，以便与对应的多遍版本进行比较。请注意，在数据集较小时，多遍随机梯度下降的表现远优于单遍随机梯度下降。而在数据集较大时，两者表现相近；事实上，多遍随机梯度下降始终表现较差。

讨论 我们提出了一种方法，用于研究最小二乘问题上单遍随机梯度下降的高维极限，其中步数与问题的维度成正比。该极限对象可用朗之万型扩散来描述，并且可以直接与多遍情形中的同一对象进行比较。

关于单遍随机梯度下降训练标度极限的文献很多，因此这里只提及一些最相关的研究。 [AGJ22]或许是最接近的高维扩散近似，它适用于存在隐藏的有限维结构的情况；当 $\boldsymbol{K} = \boldsymbol{I}$ 时，本文研究的情形也属于其涵盖范围，此外它还涵盖 $\boldsymbol{K}$ 具有有限多个特征值的情形。另请参阅 [AGJ21]。

还有其他标度极限采用了不同于本文的表述方式。 [WL19；WML17] 针对具有单位协方差的广义线性模型，给出了状态的偏微分方程极限。 [BGH23]在协方差满足光滑性假设的情况下，给出了随机梯度下降的标度极限 $\boldsymbol{K}$，并从几何角度对其进行了解释；他们还以某种方式描述了随机梯度下降的涨落。需要注意的是，定理1.5 本质上给出了风险的大数定律，而非涨落。

[Ger+22] 使用动力学场论，为采用大批量小批量随机梯度下降的浅层神经网络得出了密切相关的结果；动力学平均场理论对小批量噪声的自相关及其他几个过程提供了隐式刻画。另请参见 [CCM21]的相关工作。我们指出，在批量大小成比例的情况下， [Lee+22]中还给出了离散 Volterra 描述。

组织在第2 节中，我们概述主要证明，并将其归结为核心技术问题。在第3节中，我们界定随机误差项，表示

本文的主要技术贡献。

2 证明的主要论证

为了比较随机梯度下降与均质化随机梯度下降，我们采用扩散近似中的一种鞅方法（参见 [EK 09]）。实际上，我们证明了 $q(\boldsymbol{x}_k)$几乎满足伊藤引理的结论。此外，我们证明两个Doob分解中的鞅项都很小，因此只需证明 $q(\boldsymbol{x}_k)$ 和 $q(\boldsymbol{X}_t)$的可预测部分彼此接近。

为推进讨论，我们计算这个Doob分解。为了利用移除 $\tilde{\boldsymbol{x}}$后得到的更简单结构，引入

$$\boldsymbol{v}_k \coloneqq \boldsymbol{x}_k - \tilde{\boldsymbol{x}} \quad \text{and} \quad \boldsymbol{V}_t \coloneqq \boldsymbol{X}_t - \tilde{\boldsymbol{x}}. \tag{2.1}$$

我们将第一个以整数为索引的函数扩展到实数索引，定义如下$\boldsymbol{v}_t = \boldsymbol{v}_{\lfloor t \rfloor}$。我们还令$(\mathcal{F}_t : \quad t \geq 0)$ 表示由 $(\boldsymbol{v}_t : \quad t \geq 0)$ 和 $(\boldsymbol{V}_{t/d} : \quad t \geq 0)$ 生成的滤过。因此，对于所有 $k \in \mathbb{N}$， $\boldsymbol{v}_k$ 都关于 $\mathcal{F}_k$ 可测。回顾二次型的递推关系（1.5） q

$$\begin{aligned} q(\boldsymbol{v}_k) - q(\boldsymbol{v}_{k-1}) &= -\gamma(\nabla q(\boldsymbol{v}_{k-1}))^T(\boldsymbol{u}_{k-1} + \Delta_k) + \tfrac{\gamma^2}{2}(\boldsymbol{u}_{k-1} + \Delta_k)^T(\nabla^2 q)(\boldsymbol{u}_{k-1} + \Delta_k), \quad \text{where} \\ \Delta_k &= \boldsymbol{m}_k(\boldsymbol{m}_k^T \boldsymbol{v}_{k-1} - \eta_k) \quad \text{and} \quad \boldsymbol{u}_{k-1} = \tfrac{\delta}{d}(\boldsymbol{v}_{k-1} + \tilde{\boldsymbol{x}}). \end{aligned} \tag{2.2}$$

上述方程都可以分解为一个可预测部分和两个鞅增量

$$\begin{aligned} q(\boldsymbol{v}_k) - q(\boldsymbol{v}_{k-1}) =& -\gamma(\nabla q(\boldsymbol{v}_{k-1}))^T\big((\tfrac{\delta}{d}\boldsymbol{I} + \tfrac{1}{d}\boldsymbol{K})\boldsymbol{v}_{k-1} + \tfrac{\delta\tilde{\boldsymbol{x}}}{d}\big) + \Delta\mathcal{M}_k^{\text{lin}} \\ &+ \tfrac{\gamma^2}{2}\operatorname{tr}(\tfrac{1}{d}\boldsymbol{K}(\nabla^2 q))\left(\tfrac{1}{d}\boldsymbol{v}_{k-1}^T\boldsymbol{K}\boldsymbol{v}_{k-1} + \mathbb{E}[\eta_k^2]\right) + \Delta\mathcal{E}_k^{\text{quad}} + \Delta\mathcal{M}_k^{\text{quad}}, \\ \text{where} \quad \Delta\mathcal{M}_k^{\text{quad}} :=& \Delta_k^T\nabla^2 q\Delta_k - \mathbb{E}[\Delta_k^T\nabla^2 q\Delta_k \mid \mathcal{F}_{k-1}]. \end{aligned} \tag{2.3}$$

其余鞅增量由 $\Delta\mathcal{M}_k^{\text{lin}}$ 给出，且均关于 Δ_k 呈线性。可预测部分又进一步分解为主导阶项和误差项 $\Delta\mathcal{E}_k^{\text{quad}}$ 。

这些可预测部分又取决于不同的统计量 $q_1(\boldsymbol{v}_{k-1})$。在有限维情形下，我们可以通过一个封闭递推系统，将这些统计量（或某个足够大的有限汇总统计量集合 $q, q_1, \ldots, q_r$）与其自身联系起来。但在当前情形下，这是不可能的。另一方面，对于这里的问题，我们证明存在一个近似封闭的函数流形。具体而言，我们令

$$\begin{aligned} Q_n(q) \coloneqq Q_n(q, \boldsymbol{K}) = \Big\{ & q(\boldsymbol{x}), \quad (\nabla q(\boldsymbol{x}))^T R(z;\boldsymbol{K})\boldsymbol{x}, \quad \boldsymbol{x}^T R(y;\boldsymbol{K})(\nabla^2 q)R(z;\boldsymbol{K})\boldsymbol{x}, \\ & (\nabla q(\boldsymbol{x}))^T R(z;\boldsymbol{K})\tilde{\boldsymbol{x}}, \quad \boldsymbol{x}^T R(y;\boldsymbol{K})(\nabla^2 q)R(z;\boldsymbol{K})\tilde{\boldsymbol{x}}, \quad \forall\, z, y \in \Gamma \Big\}. \end{aligned} \tag{2.4}$$

这里， $R(z;\boldsymbol{K}) = (\boldsymbol{K} - z\boldsymbol{I})^{-1}$ 是预解矩阵，而 Γ 是半径为 $\max\{1, 3\|\boldsymbol{K}\|\}$ 的圆。为了控制鞅，方便起见，我们引入一个停时

$$\tau \coloneqq \inf\left\{k : ||\boldsymbol{v}_k|| > d^{\varepsilon}\right\} \cup \left\{td : ||\boldsymbol{V}_t|| > d^{\varepsilon}\right\}, \tag{2.5}$$

并引入相应的停时过程

$$\boldsymbol{v}_k^{\tau} = \boldsymbol{v}_{k\wedge\tau}, \quad \boldsymbol{V}_t^{\tau} = \boldsymbol{V}_{t\wedge(\tau/d)}. \tag{2.6}$$

我们证明了适用于停时过程的定理版本，随后 gnɥg出停时大于 n 的概率为压倒性概率。

我们用于比较 $\boldsymbol{v}_{td}$ 和 $\boldsymbol{V}_t$ 的关键工具是以下引理。

引理 2.1。给定一个满足 $\|q\|_{C^2} \leq 1$ 的二次型 q，其中 $Q = Q_n(q) \cup Q_n(\mathcal{P}) \cup Q_n(\|\cdot\|^2)$ 如上，

$$\begin{aligned}\max_{0\leq t\leq \frac{n}{d}} |q(\boldsymbol{v}_{td}^\tau) - q(\boldsymbol{V}_t^\tau)| \leq \sup_{0\leq t\leq n/d} \left(|\mathcal{M}_{\lfloor td\rfloor}^{\mathrm{lin},\tau}| + |\mathcal{M}_{\lfloor td\rfloor}^{\mathrm{quad},\tau}| + |\mathcal{E}_{\lfloor td\rfloor}^{\mathrm{quad},\tau}| + |\mathcal{M}_t^{HSGD,\tau}|\right) \\ + C(\|\boldsymbol{K}\|) \cdot \sup_{g\in Q} \int_0^{n/d} |g(\boldsymbol{v}_{sd}^\tau) - g(\boldsymbol{V}_s^\tau)| ds.\end{aligned} \tag{2.7}$$

这里，$\mathcal{M}^{HSGD}{}_{,t}{}^{\tau}$ 是 $q(\boldsymbol{V}_t^\tau)$的半鞅分解中的鞅部分。

证明概要。由于该结论与 [Paq+22，命题 4 的证明1]，十分相似，我们仅说明其主要思想。其思路是，取一个 $g \in Q$，并应用（2.3），那么在 $g(\boldsymbol{v}_t)$的可预测部分中，我们有

$$I_1 := \int_0^t \nabla g(\boldsymbol{v}_{sd})^T (\delta \boldsymbol{I} + \boldsymbol{K}) \boldsymbol{v}_{sd} \mathrm{d}s, \quad I_2 := \int_0^t \nabla g(\boldsymbol{v}_{sd})^T \tilde{\boldsymbol{x}} \mathrm{d}s, \quad I_3 := \int_0^t \boldsymbol{v}_{sd}^T \boldsymbol{K} \boldsymbol{v}_{sd} \mathrm{d}s.$$

这些项也会以系数的形式出现，而这些系数仅可利用 $\|g\|_{C^2}$ 和 $\|\boldsymbol{K}\|$进行界定。对 $g(\boldsymbol{V}_t)$ 应用伊藤引理，同样可以得到这一结果，不过需要将 $\boldsymbol{v}_t \to \boldsymbol{V}_t$替换为。我们希望界定例如 $I_1(\boldsymbol{v}_t) - I_1(\boldsymbol{V}_t)$。为此，我们将其被积函数表示为$p(\boldsymbol{v}_t) - p(\boldsymbol{V}_t)$，其中 p 是多项式。若 g 是线性的（即（2.4）的最后一行），那么 p 仍然是线性的。例如，若其为 $g(\boldsymbol{x}) = \nabla q(\boldsymbol{x})^T R(z; \boldsymbol{K})\tilde{\boldsymbol{x}}$，那么 p仍然是线性的，且表示为

$$p(\boldsymbol{x}) = \boldsymbol{x}^T (\delta \boldsymbol{I} + \boldsymbol{K}) R(z; \boldsymbol{K}) \tilde{\boldsymbol{x}} = \delta \boldsymbol{x}^T R(z; \boldsymbol{K}) \tilde{\boldsymbol{x}} + \boldsymbol{x}^T R(z; \boldsymbol{K}) \tilde{\boldsymbol{x}} - z \boldsymbol{x}^T \tilde{\boldsymbol{x}},$$

其中我们使用了预解式恒等式 $(\boldsymbol{K} - z)R(z; \boldsymbol{K}) = \boldsymbol{I}$。注意，由于函数$\boldsymbol{x}^T R(z; \boldsymbol{K})\tilde{\boldsymbol{x}}$ 属于 $Q_n(\|\cdot\|^2)$，因此它包含在 Q 中。此外，根据柯西积分公式，我们可以通过对 $y \in \Gamma$上的 $\frac{-1}{2\pi i} \boldsymbol{x}^T R(y; \boldsymbol{K})\tilde{\boldsymbol{x}}$取平均来表示 $\boldsymbol{x}^T \tilde{\boldsymbol{x}}$。因此 —

$$|p(\boldsymbol{v}_{td}) - p(\boldsymbol{V}_t)| \leq (\delta + 1 + 3\|\boldsymbol{K}\|) \max_{g\in Q} |g(\boldsymbol{v}_{td}) - g(\boldsymbol{V}_t)|.$$

通过相同的推导，最终可以证明，Q 中包含的每一项都可以用类似的方式控制，所用的是类 Q中的其他元素。 □

第二个重要思路是将集合 Q离散化。

引理 2.2. 存在 $\bar{Q} \subseteq Q$，满足 $|\bar{Q}| \leq C(\|\boldsymbol{K}\|)d^{4m}$，使得对每个 $q \in Q$，都存在某个 $\bar{q} \in \bar{Q}$ 满足 $||q - \bar{q}||_{C^2} \leq d^{-2m}$。

证明。在谱曲线 Γ 上，我们可以界定预解式的范数。由于

$$\frac{\mathrm{d}}{\mathrm{d}z} R(z; \boldsymbol{K}) = (\boldsymbol{K} - z\boldsymbol{I})^{-2},$$

我们有其范数被绝对常数所界定。曲线的弧长至多为 $C(\|\boldsymbol{K}\|)$，因此通过选取流形 $\Gamma \times \Gamma$ 的最小网 $d^{-2\varepsilon}$，引理得证。

□

现在，论证的主要技术部分是控制鞅和误差。由于我们处理的是停止过程 $\boldsymbol{v}_k^\tau$，我们引入停止过程 $\mathcal{M}^{\mathrm{lin}、k}{}^{\tau}$、$\mathcal{M}^{\mathrm{quad}、k}{}^{\tau}$、$\mathcal{E}^{\mathrm{quad}、k}{}^{\tau}$，它们的定义类似于（2.6）。

引理 2.3。对于任意二次型 q 满足 $\|q\|_{C^2} \le 1$，各项 $\mathcal{M}_k^{\mathrm{lin},\tau}$、$\mathcal{M}_k^{\mathrm{quad},\tau}$、$\mathcal{E}_k^{\mathrm{quad},\tau}$ 以压倒性概率满足以下界（该界关于 q 是一致的），其中 $n \le d\log d$

(i) $\sup_{1\le k\le n} |\mathcal{M}_k^{\mathrm{lin},\tau}| \le d^{-\frac{1}{2}+5\varepsilon}$,

(ii) $\sup_{1\le k\le n} |\mathcal{M}_k^{\mathrm{quad},\tau}| \le d^{-\frac{1}{2}+9\varepsilon}$,

(iii) $\sup_{1\le k\le n} |\mathcal{E}_k^{\mathrm{quad},\tau}| \le d^{-1+9\varepsilon}$.

结合引理 2.1 和 2.2，再结合上述结果（另见引理 3.2，其中齐次化随机梯度下降鞅有界），我们得出，对于任意 $\bar{q} \in Q$ 满足$\|q\|_{C^2} = 1$,

$$|\bar{q}(\boldsymbol{v}_{td}^\tau) - \bar{q}(\boldsymbol{V}_t^\tau)| \le 4d^{-\frac{1}{2}+9\varepsilon} + C(\|\boldsymbol{K}\|) \max_{g\in Q} \int_0^t |g(\boldsymbol{v}_{sd}^\tau) - g(\boldsymbol{V}_s^\tau)| ds. \tag{2.8}$$

因此，由引理 2.2 并对 $\|g\|_{C^2}$ 所有 Q进行界定,

$$\max_{g\in Q} |q(\boldsymbol{v}_{td}^\tau) - q(\boldsymbol{V}_t^\tau)| \le C(\|\boldsymbol{K}\|)\left(d^{-2} + d^{-\frac{1}{2}+9\varepsilon} + \int_0^t \max_{g\in Q} |g(\boldsymbol{v}_{sd}^\tau) - g(\boldsymbol{V}_s^\tau)| ds\right). \tag{2.9}$$

由格朗沃尔不等式可知，以压倒性概率

$$\max_{g\in Q} \max_{0\le t\le n/d} |g(\boldsymbol{v}_{td}^\tau) - g(\boldsymbol{V}_t^\tau)| \le C(\|\boldsymbol{K}\|)(d^{-2} + 4d^{-\frac{1}{2}+9\varepsilon}) e^{C(\|\boldsymbol{K}\|)n/d}. \tag{2.10}$$

现在注意到，范数函数 $\boldsymbol{x} \mapsto \|\boldsymbol{x}\|^2$ 是 Q. 中包含的二次型之一。因此，若令 $\mathcal{G}$ $\mathcal{E} = \{$ $0\le s\le n/d \|\boldsymbol{V}_s\| \le$ 表示上述显示中的事件，$d^{\varepsilon/2}\}$ 并令最大值，则有

$$\mathcal{G} \cap \mathcal{E} \cap \{\tau \le n/d\} \subseteq \{\|\boldsymbol{v}_\tau\| - \|\boldsymbol{v}_{\tau-1}\| \ge d^{\varepsilon/2}\} \cap \{\tau \le n/d\}.$$

这是因为在事件 $\{\tau \le n/d\} \cap \mathcal{E}$ 中，我们必然有过 $\|\boldsymbol{v}_\tau\| > d^\varepsilon$；但在前一步 τ，我们有 $\boldsymbol{v}_{\tau-1}$，可以将其与 $\boldsymbol{V}_{\tau-1}$ 进行比较（由于 $\mathcal{G}$，且 $\boldsymbol{V}_{\tau-1}$ 的范数很小。现在很容易看出，以压倒性概率，从时刻 0 到 n/d 之间，随机梯度下降的增量都不会使范数增大一个 d的幂次。因此，要完成证明，只需证明 $\mathcal{E}$ 以压倒性概率成立。

因此，证明可由以下引理完成:

引理 2.4. 对于任意 $\delta > 0$ 和任意 $t > 0$，以压倒性概率有

$$\max_{0\le s\le t} \|\boldsymbol{X}_s\|^2 \le e^{C(\|\boldsymbol{K}\|)t} d^\delta.$$

证明。我们对 $\phi(\boldsymbol{X}_t) := \log(1 + \|\boldsymbol{X}_t\|^2)$应用伊藤公式，得到

$$\mathrm{d}\phi(\boldsymbol{X}_t) = -2\gamma \frac{\boldsymbol{X}_t \cdot \nabla \mathcal{R}(\boldsymbol{X}_t)}{1+\|\boldsymbol{X}_t\|^2} \mathrm{d}t + \frac{\boldsymbol{X}_t \cdot \gamma \sqrt{\frac{2}{d}\mathcal{P}(\boldsymbol{X}_t)\boldsymbol{K}}\, \mathrm{d}B_t}{1+\|\boldsymbol{X}_t\|^2} + \Big(\frac{\mathcal{P}(\boldsymbol{X}_t)}{1+\|\boldsymbol{X}_t\|^2} \frac{2\gamma^2}{d} \operatorname{tr}(\boldsymbol{K}) - \frac{2\gamma^2 \mathcal{P}(\boldsymbol{X}_t) \boldsymbol{X}_t^T \boldsymbol{K} \boldsymbol{X}_t}{d}\Big) \mathrm{d}t$$

漂移项和二次变差项可以用某个 $C(\|\boldsymbol{K}\|)$来界定。因此，取此常数，对所有 $r \ge 0$,

$$\mathbb{P}\Big(\max_{0\le s\le t} \phi(\boldsymbol{X}_s) \ge C(\|\boldsymbol{K}\|)(t + r\sqrt{t})\Big) \le 2\exp(-r^2/2).$$

取 $r = \sqrt{\text{对数 } d\text{对数的对数}}$ d，我们得出，以压倒性概率

$$\max_{0\le s\le t} \phi(\boldsymbol{X}_s) \le C(\|\boldsymbol{K}\|)(t + \sqrt{t \log d \log\log d}).$$

□

3 控制误差

本节的主要目标是控制鞅项和误差项；具体而言，我们将证明引理 2.3。为得到这些界，我们需要以下集中引理，这是一个标准结果（参见 [Ver18，定理 2.8。1], 其中证明了非鞅情形的界。对鞅情形的推广只是一个小扩展）：

引理 3.1 （鞅伯恩斯坦不等式）。 如果 $(M_n)_1^N$是滤过概率空间 $(\Omega,\ (\mathcal{F}_n)_1^N,\ \mathbb{P})$ 上的鞅，并定义

$$\sigma_{n,p} := \left\| \inf\{t \geq 0 : \mathbb{E}\left(e^{|M_n - M_{n-1}|^p / t^p} | \mathcal{F}_{n-1}\right) \leq 2\} \right\|_{L^\infty(\mathbb{P})}, \tag{3.1}$$

那么存在一个绝对常数 $C > 0$，使得对所有 $t > 0$,

$$\mathbb{P}\left(\sup_{1\leq n\leq N} |M_n - \mathbb{E}M_0| \geq t\right) \leq 2\exp\left(-\min\left\{\frac{t}{C\max\sigma_{n,1}}, \frac{t^2}{C\sum_1^N \sigma_{n,1}^2}\right\}\right). \tag{3.2}$$

我们还将记录一个关于 ∇q 的估计，以备后用；该估计由 $\|\cdot\|_{C^2}$ 控制推出。

$$||\nabla q(\boldsymbol{x})|| \leq ||\nabla^2 q|| \cdot ||\boldsymbol{x}|| + ||\nabla q(0)|| \leq \|q\|_{C^2} \cdot (||\boldsymbol{x}|| + 1). \tag{3.3}$$

3.1 递推关系中梯度部分的鞅

引理的证明 2.3 第 (i) 部分。 比较 (2.2) 和 (2.3，可知对于 $k \leq \tau$

$$\begin{aligned} \Delta\mathcal{M}_k^{\text{lin},\tau} &= \left[\left(\boldsymbol{w}_{k-1}^T \boldsymbol{m}_k\right)\left(\boldsymbol{m}_k^T \boldsymbol{v}_{k-1}^\tau - \eta_k\right) - \tfrac{1}{d}\boldsymbol{w}_{k-1}^T \boldsymbol{K} \boldsymbol{v}_{k-1}^\tau\right] =: [\Delta\mathcal{M}_k^{\text{lin}\,1,\tau} - \Delta\mathcal{M}_k^{\text{lin}\,2,\tau}], \\ &\text{where} \quad \boldsymbol{w}_{k-1} := -\gamma\nabla q(\boldsymbol{v}_{k-1}^\tau) + \tfrac{\gamma^2\delta}{d}(\boldsymbol{v}_{k-1}^\tau + \tilde{\boldsymbol{x}}). \end{aligned} \tag{3.4}$$

注意，对于 $k > \tau$，停止后的鞅增量为 0。 利用 (3.3)， $\|\boldsymbol{w}_{k-1}\| \leq C(\gamma,\ \delta)d^\varepsilon$.我们将分别用它们的 Orlicz 范数来界定 $\Delta\mathcal{M}^{\text{lin}\,1}{}_{,\ k}{}^{\ \tau}$ 和 $\Delta\mathcal{M}^{\text{lin}\,2}{}_{,\ k}{}^{\ \tau}$ 的贡献。对于第一部分，任取固定的 k，我们对 $\mathcal{F}_{k-1}$ 进行条件化，并结合假设1.3，可得

$$||\Delta\mathcal{M}_k^{\text{lin}\,1,\tau}||_{\psi_1} \leq \left\|\boldsymbol{w}_{k-1}^T \boldsymbol{m}_k\right\|_{\psi_2} \left\|\boldsymbol{m}_k^T \boldsymbol{v}_{k-1}^\tau - \eta_k\right\|_{\psi_2} \leq Cd^{-\frac{1}{2}+2\varepsilon} \cdot d^{-\frac{1}{2}+2\varepsilon} \tag{3.5}$$

其中 C 是某个绝对常数。对于第二部分，我们有

$$|\Delta\mathcal{M}_k^{\text{lin}\,2,\tau}| = |\tfrac{1}{d}\boldsymbol{w}_{k-1}^T \boldsymbol{K} \boldsymbol{v}_{k-1}^\tau| \leq Cd^{-1+2\varepsilon}. \tag{3.6}$$

综合以上结果可知，对于每个 k,

$$\sigma_{k,1} := \inf\{t > 0 : \mathbb{E}[\exp(|\Delta\mathcal{M}_k^{\text{lin}\,1,\tau} - \Delta\mathcal{M}_k^{\text{lin}\,2,\tau}|/t)|\mathcal{F}_{k-1}] \leq 2\} \leq Cd^{-1+4\varepsilon} \tag{3.7}$$

并且根据鞅伯恩斯坦不等式,

$$\begin{aligned} \mathbb{P}\left(\sup_{1\leq k\leq n} |\mathcal{M}_k^{\text{lin},\tau} - \mathbb{E}\mathcal{M}_0^{\text{lin}}| \geq t\right) &\leq 2\exp\left(-\min\left\{\frac{t}{c\max\sigma_{k,1}}, \frac{t^2}{c\sum_{k=1}^n \sigma_{k,1}}\right\}\right) \\ &\leq 2\exp\left(-\min\left\{Ctd^{1-4\varepsilon}, Ct^2 d^{2-8\varepsilon} n^{-1}\right\}\right). \end{aligned} \tag{3.8}$$

由于我们假设 $n \leq d$对数 d，因此可得

$$\sup_{1\leq k\leq n} |\mathcal{M}_k^{\text{lin},\tau}| \leq d^{-\frac{1}{2}+5\varepsilon} \tag{3.9}$$

以压倒性概率。 □

3.2 递推关系中 Hessian 部分的鞅

引理的证明 2.3 的部分（*ii*）和（*iii*）。接下来，我们考虑递推关系中 Hessian 部分的贡献。我们写作

$$\begin{aligned}&\tfrac{\gamma^2}{2}(\boldsymbol{m}_k\boldsymbol{m}_k^T\boldsymbol{v}_{k-1}^\tau-\boldsymbol{m}_k\eta_k)^T(\nabla^2 q)(\boldsymbol{m}_k\boldsymbol{m}_k^T\boldsymbol{v}_{k-1}^\tau-\boldsymbol{m}_k\eta_k)\\&=\mathbb{E}\left[\tfrac{\gamma^2}{2}(\boldsymbol{m}_k\boldsymbol{m}_k^T\boldsymbol{v}_{k-1}^\tau-\boldsymbol{m}_k\eta_k)^T(\nabla^2 q)(\boldsymbol{m}_k\boldsymbol{m}_k^T\boldsymbol{v}_{k-1}^\tau-\boldsymbol{m}_k\eta_k)|\mathcal{F}_{k-1}\right]+\Delta\mathcal{M}_k^{\text{quad}}.\end{aligned}\tag{3.10}$$

重新排列各项后，得到

$$\Delta\mathcal{M}_k^{\text{quad}}=A_kB_k-\mathbb{E}[A_kB_k|\mathcal{F}_{k-1}]\tag{3.11}$$

其中

$$A_k:=\boldsymbol{m}_k^T(\nabla^2 q)\boldsymbol{m}_k,\quad B_k:=(\boldsymbol{m}_k^T\boldsymbol{v}_{k-1}^\tau-\eta_k)^2.\tag{3.12}$$

这可以展开为

$$\begin{aligned}\Delta\mathcal{M}_k^{\text{quad}}=&(A_k-\mathbb{E}[A_k])(B_k-\mathbb{E}[B_k])+\mathbb{E}[A_k]\mathbb{E}[B_k]-\mathbb{E}[A_kB_k]\\&+(A_k-\mathbb{E}[A_k])\mathbb{E}[B_k]+(B_k-\mathbb{E}[B_k])\mathbb{E}[A_k],\end{aligned}\tag{3.13}$$

因此，我们先使用 Hanson-Wright 不等式，为量 $A_k-\mathbb{E}[A_k]$ 和 $B_k-\mathbb{E}[B_k]$ 推导次指数界。对于 A_k，有

$$\begin{aligned}\mathbb{P}(|A_k-\mathbb{E}A_k|\geq t)&\leq 2\exp\left[-c\min\left(\frac{t^2}{d^{-2+4\varepsilon}||\nabla^2 q||_{HS}^2},\frac{t}{d^{-1+2\varepsilon}||\nabla^2 q||}\right)\right]\\&\leq 2\exp[-c'\min(t^2d^{1-4\varepsilon},td^{1-2\varepsilon})]\leq 2\exp[-c''td^{\frac{1}{2}-2\varepsilon}]\end{aligned}\tag{3.14}$$

因此我们得到次指数界

$$||A_k-\mathbb{E}[A_k]||_{\psi_1}<Cd^{-\frac{1}{2}+2\varepsilon}.\tag{3.15}$$

接下来，我们为 B_k 得到一个次指数界。对于 B_k 中不涉及 η_k 的部分，我们使用 Hanson-Wright 不等式得到

$$\begin{aligned}&\mathbb{P}\left(\left|\boldsymbol{m}_k^T\boldsymbol{v}_{k-1}^\tau(\boldsymbol{v}_{k-1}^\tau)^T\boldsymbol{m}_k-\mathbb{E}\boldsymbol{m}_k^T\boldsymbol{v}_{k-1}^\tau(\boldsymbol{v}_{k-1}^\tau)^T\boldsymbol{m}_k\right|\geq t\right)\\&\leq 2\exp\left[-c\min\left(\frac{t^2}{d^{-2+4\varepsilon}||\boldsymbol{v}_{k-1}^\tau(\boldsymbol{v}_{k-1}^\tau)^T||_{HS}^2},\frac{t}{d^{-1+2\varepsilon}||\boldsymbol{v}_{k-1}^\tau(\boldsymbol{v}_{k-1}^\tau)^T||}\right)\right]\\&\leq 2\exp[-c\min(t^2d^{2-8\varepsilon},td^{1-4\varepsilon})].\end{aligned}\tag{3.16}$$

对于涉及 η_k 的项，我们使用设定中的假设所给出的 Orlicz 界，得到

$$\begin{aligned}||\boldsymbol{m}_k^T\boldsymbol{v}_{k-1}^\tau\eta_k||_{\psi_1}&\leq||\boldsymbol{m}_k^T\boldsymbol{v}_{k-1}^\tau||_{\psi_2}\cdot||\eta_k||_{\psi_2}=d^{-\frac{1}{2}+2\varepsilon}d^{-\frac{1}{2}+\varepsilon}\\&=d^{-1+3\varepsilon}.\end{aligned}\tag{3.17}$$

同样，由于 $||\eta_k^2||_{\psi_1}=d^{-1+2\varepsilon}$ 结合界（3.16）和（3.17），我们有

$$||B_k-\mathbb{E}[B_k]||_{\psi_1}<Cd^{-1+4\varepsilon}.\tag{3.18}$$

此外，我们有

$$\mathbb{E}[A_k]=O(1),\qquad\mathbb{E}[B_k]=O(d^{-1}),\tag{3.19}$$

对所有 k 基于关于 η_k 和 $\boldsymbol{m}_k$的假设。现在我们使用 (3.15)、(3.18)、(3.19) 依次对 (3.13) 中的每一项进行界定。

为界定 $(A_k - \mathbb{E}[A_k])(B_k - \mathbb{E}[B_k])$ 的贡献，我们注意到，对每个k，以压倒性概率，$|A_k - \mathbb{E}[A_k]| < d^{-\frac{1}{2}+3\varepsilon}$ 且 $|B_k - \mathbb{E}[B_k]| < d^{-1+5\varepsilon}$，因此我们可以得出结论：以压倒性概率，

$$\sum_{k=1}^{n} \Big|(A_k - \mathbb{E}[A_k])(B_k - \mathbb{E}[B_k])\Big| < nd^{-\frac{3}{2}+8\varepsilon} < d^{-\frac{1}{2}+9\varepsilon}. \tag{3.20}$$

对于（3.13）中的第二项，我们有

$$\Big|\mathbb{E}[A_k]\mathbb{E}[B_k] - \mathbb{E}[A_k B_k]\Big| = \Big|\mathbb{E}\big[(A_k - \mathbb{E}A_k)(B_k - \mathbb{E}B_k)\big]\Big| \le \mathbb{E}\Big|(A_k - \mathbb{E}A_k)(B_k - \mathbb{E}B_k)\Big|. \tag{3.21}$$

我们可以利用以下结果来界定该量

$$\begin{aligned}\mathbb{P}\left(\Big|(A_k - \mathbb{E}A_k)(B_k - \mathbb{E}B_k)\Big| \ge t\right) &\le \mathbb{P}(|A_k - \mathbb{E}A_k| \ge \sqrt{t}) + \mathbb{P}(|B_k - \mathbb{E}B_k| \ge \sqrt{t}) \\ &\le 4\exp\left[-c\min(td^{1-4\varepsilon}, \sqrt{t}d^{1-4\varepsilon})\right]\end{aligned} \tag{3.22}$$

最后一行中的界由（3.14）和（3.18）合并得到。利用这个界，我们得到

$$\begin{aligned}\Big|\mathbb{E}[A_k]\mathbb{E}[B_k] - \mathbb{E}[A_k B_k]\Big| &\le \int_0^\infty x\mathbb{P}\left(\Big|(A_k - \mathbb{E}A_k)(B_k - \mathbb{E}B_k)\Big| \ge x\right) dx \\ &\le \int_0^1 4x\exp(-cxd^{1-4\varepsilon})dx + \int_1^\infty 4x\exp(-c\sqrt{x}d^{1-4\varepsilon})dx\end{aligned} \tag{3.23}$$

在第一个积分中作变量替换 $y = xd^{1-4\varepsilon}$，并在第二个积分中作变量替换 $z = \sqrt{x}d^{1-4\varepsilon}$，可得

$$4d^{-2+8\varepsilon}\int_0^{d^{1-4\varepsilon}} y\exp(-cy)dy + 4d^{-4+8\varepsilon}\int_{d^{1-4\varepsilon}}^\infty z^2\exp(-cz)dz = O(d^{-2+8\varepsilon}). \tag{3.24}$$

因此，

$$\sum_{k=1}^{n}\Big|\mathbb{E}[A_k]\mathbb{E}[B_k] - \mathbb{E}[A_k B_k]\Big| = O(nd^{-2+8\varepsilon}). \tag{3.25}$$

最后，我们注意到式（3.13）中的其余项，即$(A_k - \mathbb{E}[A_k])\mathbb{E}[B_k]$ 和$(B_k - \mathbb{E}[B_k])\mathbb{E}[A_k]$，是鞅增量，且

$$||(A_k - \mathbb{E}[A_k])\mathbb{E}[B_k]||_{\psi_1} \le Cd^{-\frac{3}{2}+2\varepsilon}, \qquad ||(B_k - \mathbb{E}[B_k])\mathbb{E}[A_k]||_{\psi_1} \le Cd^{-1+4\varepsilon}. \tag{3.26}$$

应用鞅伯恩斯坦不等式，我们得到

$$\begin{aligned}&\mathbb{P}\left(\sup_{1\le k\le n}\left|\sum_{j=1}^{k}(A_j - \mathbb{E}[A_j])\mathbb{E}[B_j] + (B_j - \mathbb{E}[B_j])\mathbb{E}[A_j]\right| \ge t\right) \\ &\le 2\exp\left(-\min\left\{\frac{t}{c\max\sigma_{k,1}}, \frac{t^2}{c\sum_{k=1}^{n}\sigma_{k,1}}\right\}\right) \\ &\le 2\exp\left(-\min\left\{Ctd^{1-4\varepsilon}, Ct^2d^{2-8\varepsilon}n^{-1}\right\}\right).\end{aligned} \tag{3.27}$$

因此，对于 $n \le d$对数 d，我们得到

$$\sup_{1\le k\le n}\left|\sum_{j=1}^{k}(A_j - \mathbb{E}[A_j])\mathbb{E}[B_j] + (B_j - \mathbb{E}[B_j])\mathbb{E}[A_j]\right| \le d^{-\frac{1}{2}+5\varepsilon} \tag{3.28}$$

以压倒性概率成立。最后，结合（3.20）、（3.25）和（3.28）中的界，我们得出，对于 $n \le d$对数 d,

$$\sup_{1\le k\le n} |\mathcal{M}_k^{\mathrm{quad},\tau}| \le d^{-\frac{1}{2}+8\varepsilon} \tag{3.29}$$

以压倒性概率成立。这就完成了引理第（ii）部分的证明。

对于第（iii）部分，我们注意到 $\Delta\mathcal{E}^{\mathrm{quad}}{}_{,k}{}^{\tau} = \mathbb{E}[A_k B_k] - \mathbb{E}[A_k]\mathbb{E}[B_k] + O(d^{-2+4\epsilon})$，由 $\boldsymbol{u}_k$ 交叉项产生的误差项，因此 $\mathcal{E}^{\mathrm{quad}}{}_{,k}{}^{\tau}$ 的界可由（3.25）立即推出。

□

3.3 HSGD 的鞅

回顾一下，HSGD 过程 $\{\boldsymbol{V}_t\}$ 满足微分方程

$$d\boldsymbol{V}_t = -\gamma\nabla\mathcal{R}(\boldsymbol{V}_t+\tilde{\boldsymbol{x}})dt + \gamma\sqrt{\tfrac{2}{d}\mathcal{P}(\boldsymbol{V}_t+\tilde{\boldsymbol{x}})\boldsymbol{K}}dB_t \tag{3.30}$$

其中

$$\mathcal{P}(\boldsymbol{x}) = (\boldsymbol{x}-\tilde{\boldsymbol{x}})^T\boldsymbol{K}^2(\boldsymbol{x}-\tilde{\boldsymbol{x}}) + \eta^2. \tag{3.31}$$

由伊藤引理可得

$$q(\boldsymbol{V}_t^\tau) = q(\boldsymbol{V}_0^\tau) - \gamma\int_0^t (\nabla q(\boldsymbol{V}_s^\tau))^T\nabla\mathcal{R}(\boldsymbol{V}_s^\tau+\tilde{\boldsymbol{x}})ds + \tfrac{\gamma^2}{2}\int_0^t \mathcal{P}(\boldsymbol{V}_s^\tau+\tilde{\boldsymbol{x}})\,\mathrm{tr}(\tfrac{1}{d}\boldsymbol{K}(\nabla^2 q))ds + \mathcal{M}_t^{HSGD\,\tau}, \tag{3.32}$$

其中

$$\mathcal{M}_t^{HSGD,\tau} = \gamma\int_0^t (\nabla q(\boldsymbol{V}_s^\tau))^T\sqrt{\tfrac{1}{d}\mathcal{P}(\boldsymbol{V}_s^\tau+\tilde{\boldsymbol{x}})\boldsymbol{K}}dB_s. \tag{3.33}$$

引理 3.2. 对于任意二次型 q，若满足 $\|q\|_{C^2} \le 1$，施加停时 τ 后得到的鞅 $\mathcal{M}^{HSGD}{}_{,t}{}^{\tau}$ 满足

$$\sup_{0\le t\le n/d}\left[\mathcal{M}_t^{HSGD,\tau}\right] \le Cd^{-\frac{1}{2}+3\varepsilon}, \tag{3.34}$$

当 $n \le d\log d$ 时。

证明。 该鞅的二次变差为

$$\left[\mathcal{M}_t^{HSGD,\tau}\right] = \frac{\gamma^2}{d}\int_0^t \mathcal{P}(\boldsymbol{V}_s^\tau+\tilde{\boldsymbol{x}})(\nabla q(\boldsymbol{V}_s^\tau))^T\nabla^2\mathcal{P}(\boldsymbol{V}_s^\tau+\tilde{\boldsymbol{x}})(\nabla q(\boldsymbol{V}_s^\tau))ds \tag{3.35}$$

并且对于所有 s，都有如下界

$$|\mathcal{P}(\boldsymbol{V}_s^\tau+\tilde{\boldsymbol{x}})| \le ||\boldsymbol{V}_s^\tau||^2||\boldsymbol{K}^2|| + \eta^2 \le C_1 d^{2\varepsilon}, \tag{3.36}$$

因此，我们有

$$\sup_{0\le t\le n/d}\left[\mathcal{M}_t^{HSGD,\tau}\right] \le C(\log d)d^{-1+4\varepsilon} \tag{3.37}$$

几乎必然成立。根据有界二次变差连续鞅的高斯尾界，$\mathbb{P}(\sup_t[\mathcal{M}_t^{HSGD,\tau}] > a) < \exp(-a^2/(2C(\log d)d^{-1+4\varepsilon}))$。取 $t = d^{-\frac{1}{2}+3\varepsilon}$，我们得出

$$\sup_{0\le t\le n/d}\left[\mathcal{M}_t^{HSGD,\tau}\right] \le Cd^{-\frac{1}{2}+3\varepsilon} \tag{3.38}$$

以压倒性概率。 □

参考文献

[Ada15] R. Adamczak. “关于具有依赖关系的随机向量的 Hanson-Wright 不等式的注记”。载于：电子通信与概率 20 (2015)，第 72 期，13 页。DOI：10.1214/ECP.v20-3829。[AGJ21] G. B. Arous、R. Gheissari 和 A. Jagannath. “高维推断中非凸损失的在线随机梯度下降”。载于：机器学习研究期刊 22.1 (2021)，第 4788–4838 页。[AGJ22] G. B. Arous、R. Gheissari 和 A. Jagannath. “随机梯度下降的高维极限定理：有效动力学与临界标度”。载于：神经信息处理系统进展。A. H. Oh、A. Agarwal、D. Belgrave 和 K. Cho 编。2022。[AYS20] K. Ahn、C. Yun 和 S. Sra. “随机打乱下的随机梯度下降：无需分量凸性且不需要大量训练轮次即可达到最优速率”。载于：神经信息处理系统进展。H. Larochelle、M. Ranzato、R. Hadsell、M. Balcan 和 H. Lin 编。第 33 卷。Curran Associates, Inc.，2020，第 17526–17535 页。[BGH23] K. Balasubramanian、P. Ghosal 和 Y. He. “具有光滑协方差的在线最小二乘随机梯度下降的高维标度极限与涨落”。载于：arXiv 电子预印本，arXiv:2304.00707 (2023 年 4 月)，arXiv:2304.00707。DOI：10.48550/arXiv.2304.00707。arXiv：2304.00707 [math.PR].[CCM21] M. Celentano、C. Cheng 和 A. Montanari. “随机数据下一阶方法的高维渐近分析”。载于：arXiv 电子预印本，arXiv:2112.07572 (2021 年 12 月)，arXiv:2112.07572。DOI：10.48550/arXiv:2112.07572。arXiv：2112.07572 [math.PR].[EK09] S. N. Ethier 和 T. G. Kurtz. 马尔可夫过程：刻画与收敛。John Wiley & Sons，2009。[Ger+22] C. Gerbelot、E. Troiani、F. Mignacco、F. Krzakala 和 L. Zdeborova. “随机梯度下降方法的严格动力学平均场理论”。载于：arXiv 电子预印本，arXiv:2210.06591 (2022 年 10 月)，arXiv:2210.06591。DOI：10.48550/arXiv:2210.06591。arXiv：2210.06591 [math-ph].[GOP21] M. Gürbüzbalaban、A. Ozdaglar 和 P. A. Parrilo. “随机重排为何胜过随机梯度下降”。载于：数学规划 186 (2021)，第 49–84 页。[Lee+22] K. Lee、A. N. Cheng、C. Paquette 和 E. Paquette. “小批量动量的轨迹：批量大小饱和与高维收敛”。载于：即将在 NeurIPS 2022 发表，arXiv:2206.01029 (2022 年 6 月)，38 页。arXiv：2206.01029 [math.OC].[Paq+21] C. Paquette、K. Lee、F. Pedregosa 和 E. Paquette. “大规模随机梯度下降：平均情形分析、渐近分析与步长临界性”。载于：第三十四届学习理论会议论文集。M. Belkin 和 S. Kpotufe 编。第 134 卷。机器学习研究论文集。PMLR，2021 年 8 月，arXiv:2102.04396，第 3548–3626 页。arXiv：2102.04396 [math.OC].[Paq+22] C. Paquette、E. Paquette、B. Adlam 和 J. Pennington. “高维随机梯度下降的均匀化：精确动力学与泛化性质”。载于：arXiv 电子预印本，arXiv:2205.07069 (2022 年 5 月)，64 页。arXiv：2205.07069[math.ST]。

[RM51] H. Robbins 和 S. Monro。《一种随机逼近方法》。载于：《统计数学年刊》（1951）。DOI：10.1214/aoms/1177729586。[RR08] A. Rahimi 和 B. Recht。《用于大规模核机器的随机特征》。载于：神经信息处理系统进展（NeurIPS）。第 20 卷。2008 年，第 1177–1184 页。[RR12] B. Recht 和 C. Re。《迈向非交换算术-几何平均不等式：猜想、案例研究及其推论》。载于：第 25 届年度学习理论会议论文集。S. Mannor、N. Srebro 和 R. C. Williamson 编。第 23 卷。机器学习研究论文集。苏格兰爱丁堡：PMLR，2012 年 6 月 25–27 日，第 11.1–11.24 页。[SS20] I. Safran 和 O. Shamir。《随机洗牌下的随机梯度下降效果如何？》。载于：第三十三届学习理论会议论文集。J. Abernethy 和 S. Agarwal 编。第 125 卷。机器学习研究论文集。PMLR，2020 年 9 月，第 3250–3284 页。[Ver18] R. Vershynin。《高维概率：导论及其在数据科学中的应用》。第 47 卷。剑桥大学出版社，2018 年。[WL19] C. Wang 和 Y. M. Lu。《高维在线独立成分分析的标度极限*》。载于：《统计力学杂志：理论与实验》2019.12（2019 年 12 月），第 124011 页。DOI：10.1088/1742-5468/ab39d6。[WML17] C. Wang、J. Mattingly 和 Y. M. Lu。《标度极限：在线学习算法的精确且易处理的分析及其在正则化回归和主成分分析中的应用》。载于：arXiv 电子预印本，arXiv:1712.04332（2017 年 12 月），arXiv:1712.04332。DOI：10.48550/arXiv.1712.04332。arXiv：1712.04332[cs.LG]。[YSJ21] C. Yun、S. Sra 和 A. Jadbabaie。《开放问题：单次洗牌随机梯度下降能否优于重新洗牌随机梯度下降和梯度下降？》。载于：学习理论会议。PMLR。2021 年，第 4653–4658 页。