广义样本协方差矩阵最大特征值的大偏差

Jonathan Husson密歇根大学 电子邮箱：*jhusson@umich.edu*

Benjamin McKenna哈佛大学数学科学与应用中心 电子邮箱：*bmckenna@fas.harvard.edu*

摘要

我们为广义样本协方差矩阵的最大特征值建立了一个大偏差原理，即对于一个与 $Z^{\mathrm{T}}\Gamma Z$ 成比例的矩阵，其中 Z 具有独立同分布的实数或复数条目，且 Γ 不一定是单位矩阵。我们研究了经典情形： Z 为高斯分布且 Γ 为正定矩阵；同时也涵盖两种正交扩展：其一， Z 的元素可以改为尖锐次高斯分布，这类分布包括拉ademacher分布和均匀分布，此时我们得到与高斯模型相同的速率函数；其二，如果 Z 仍为高斯分布，则 Γ 可以具有负特征值。后一种情形验证了 Maillard 在物理学文献中给出的公式。

我们还将这些方法应用于实数或复数变形Wigner矩阵的最大特征值，将先前的大偏差估计提升为完整的大偏差原理。最后，我们去除了先前相关研究中存在的若干技术性假设。

日期： 2023年2月3日

关键词和短语： 大偏差、样本协方差矩阵、Wishart 矩阵、变形 Wigner 矩阵*2020*年数学学科分类： 60B20, 60F10, 15B52

arXiv:2302.02847v1 [math.PR] 6 Feb 2023

目录

1 引言 **2**

1.1 我们的结果.2 1.2 随机矩阵大偏差的先前结果.2 1.3 广义样本协方差矩阵的先前结果.3 1.4 文章结构.3 1.5 记号.4 1.6 致谢.5

2 结果 **5**

2.1 设定.5 2.2 戴森方程与经验测度的极限.6 2.3 退化情形.7 2.4 实数情形下的主要结果.8 2.5 复数情形下的主要结果.9

3 无穷 x_c 处的证明 **10**

3.1 Coulomb 气体的预备知识.10 3.2 退化球面积分.12 3.3 测度集中.13 3.4 弱大偏差原理上界的证明.21 3.5 弱大偏差原理下界的证明.22 3.6 退化情形.23 3.7 斯蒂尔切斯变换的第二分支.23

4 有限情形的证明 x_c **26**

5 复数情形 29

附录 A 多维乘积测度的集中 30 附录 B “紧支撑或对数 Sobolev”假设 31 附录 C 变形 Wigner 矩阵 33

参考文献 **36**

1 引言

1.1 我们的结果。

在本文中，我们给出了速度为 N 的最大特征值 $\lambda_{\max}$ 的大偏差原理，该最大特征值对应于一个广义样本协方差矩阵，即形如以下形式的随机矩阵

$$H_N = \frac{1}{M} Z^T \Gamma Z = \frac{1}{M} \sum_{i=1}^{M} d_i z_i z_i^T, \tag{1.1}$$

其中所有 符号将在下文定义。非正式地说，这意味着我们找到一个函数 I，使得

$$\mathbb{P}(\lambda_{\max}(H_N) \approx x) \approx \exp(-NI(x)). \tag{1.2}$$

这里，$Z \in \mathbb{R}^{M\times N}$ 的元素独立同分布，服从某个中心化概率测度 μ，且具有单位方差；其各行是长度为 N 的独立向量，为与文献保持一致，以列形式写作 $\{z_i\}_{i=1}^M$；广义表示确定性矩阵 $\Gamma = \mathrm{diag}(d_1, \ldots, d_M) = \mathrm{diag}(d_1^{(M)}, \ldots, d_M^{(M)})$ 不一定是单位矩阵；而 M 和 N 是两个参数，它们都趋于无穷大，且其比值为常数量级。

我们的结果也适用于复数情形 $H_N = M^{-1} Z^* \Gamma Z$，但为简便起见，本文大部分内容中我们仅使用实数情形的符号表示。

当所有 d_i均为正时，矩阵 H_N 有时称为非齐次样本协方差矩阵（或非齐次威沙特矩阵）。本文允许 d_i为负，但仅限于 μ 服从高斯分布的特殊情形。我们的这一部分结果与 Maillard 最近在物理学文献中发表的一项结果一致，该结果启发了我们的研究 [Mai21]。我们也允许 μ非高斯，但要求它们属于一类特殊的测度，称为尖锐次高斯；其中重要的特殊情形包括拉ademacher分布和均匀分布。在此情形下，我们要求所有 d_i均为正；此时我们证明，速率函数与对应的高斯情形相同。（当某些不为负时，速率函数可能 d_i不会与高斯情形相同；参见注2.17。）

当 Γ 为正定矩阵时，这种情形更为人熟知。在这种情况下，H_N 是由相互独立的观测构成的样本协方差矩阵，这些观测可能包含相关数据；也就是说，它与 $\frac{1}{M}\sum_{i=1}^M x_i x_i^T$具有相同的分布，其中向量 $\overline{x_i}$ 彼此独立，但每个 x_i 的分量可能具有非平凡的协方差矩阵 Γ。在这种情况下，H_N 当然也与其变体$\frac{1}{M}\Gamma^{1/2} Z Z^T \Gamma^{1/2}$具有相同的非零特征值，文献中有时会采用这种表述。但我们要强调，Γ 同时具有正、负特征值的情形也具有实际的统计意义：例如，当观测值 x_i 彼此不再独立时，在对上述情形进行多元方差分析估计时就会出现这种情况。关于这一情形的讨论，请参阅 [FJ22]。

1.2 随机矩阵大偏差的先前结果。

我们的工作属于近期一系列论文，这一系列始于 [GH20] 和 [GM20], ，它们利用球面积分倾斜变换的方法建立了随机矩阵的大偏差原理。这一系列的论文包括数学文献中的[BG20、AGH21、McK21、GH22、Hus22、BGH20] ，以及物理学文献中的 [Mai21、MP22] 。与这些先前工作相比，我们工作的主要技术创新在于，对所谓的 x_c 阈值之外的情形建立了严格的非高斯结果。也就是说，在模型 (1.1) 中，事实证明存在某个 $x_c = x_c(\Gamma)$，其值可能有限，也可能无限，具体取决于 Γ；当 $x_c < \infty$ 时，式 (1.2) 中的函数 I 在 x_c 处是非解析的。[McK21]曾对加性变形威格纳矩阵揭示了类似现象的一部分；我们将在下文正式定义这一模型，简单来说，它既有高斯型，也有尖锐次高斯型（以及本文不讨论的其他类型）。在 [McK21], 中，本文的一位作者证明了高斯情形下对所有 x 都成立的式 (1.2) 的对应结果，但在尖锐次高斯情形下，该结果仅对 $x < x_c$ 成立。尖锐次高斯情形下的 $x \geqslant x_c$ 区间仍是一个难题。

沿着这些研究，Maillard 在高斯情形下，以物理学严谨性层面研究了本文的模型（1.1）。他还发现了一个 x_c 阈值，并提出了一种有趣的新倾斜变换，以便在1.2）所示的范围内证明$x \geqslant x_c$，这启发了我们的工作。我们验证了他的结果，并补充了我们自己的新结果，但没有采用他的技术。相反，对于满足 $x_c < +\infty$的模型，我们找到了一个近似模型序列，其中每个模型都满足 $x_c = +\infty$。在近期关于随机矩阵大偏差的研究背景下，类似的论证最早见于 [Hus22], ，但那里处理的是不同于 x_c 阈值的问题。本文表明，这种方法也适用于各种模型中的 x_c 阈值。事实上，尽管本文主体讨论样本协方差模型（1.1），附录C 则讨论变形 Wigner 情形，并完成了由 [McK21]开始的分析。

作为主定理的特例，我们重新得到齐次情形下已有的结果 $\Gamma = \mathrm{Id}$，即通常的 Wishart 矩阵情形。Majumdar 和 Vergassola 计算了高斯情形下的速率函数 [MV09]；随后，Guionnet 与第一作者近期的工作将这一结果推广到尖锐次高斯 Wishart 矩阵 [GH20]。我们的速率函数与他们的结果相符，这一计算由 Maillard 完成 [Mai21]。

我们还开发了一些技术，用以去除上述文献中存在的一项技术性假设。通常，这些文献中的结果可以非正式地表述为："假设 X 是由 μ的样本构造而成的随机矩阵，其中$\lambda_{\max}(X)$ 是尖锐次高斯测度，且该测度或者具有紧支撑，或者满足对数 Sobolev 不等式。那么 X 满足大偏差原理。"我们将说明如何去除"紧支撑或满足对数 Sobolev 不等式"这一假设，从而得到如下形式的定理："假设 μ是由 $\lambda_{\max}(X)$的样本构造而成的随机矩阵，其中 [是尖锐次高斯测度。那么20]满足大偏差原理。"例如，我们的主要结果便采用后一种表述，前面提到的变形 Wigner 结果也是如此。在附录 B中，我们还给出一个例子，对 Guionnet 和第一作者的一项结果 [GH20]作了改进。我们在备注 B.2中给出一个尖锐次高斯分布的例子，说明这一推广并非平凡：该分布既没有紧支撑，也不满足对数 Sobolev 不等式。正如我们在那里所解释的，我们的证明也可以绕过局部定律的使用。

最后，我们对尖锐次高斯条件的历史沿革作一点评述。当然，"次高斯随机变量"这一概念相当经典，但上述研究脉络中的工作都基于两个额外观点：(a) 这一类变量中的一个特定子类（下文定义 2.1 中称为"尖锐次高斯"）具有特殊性；(b) 这一子类之所以特殊，是因为由其成员构造的随机矩阵满足的界，其常数与高斯测度本身的界相同。据我们所知，这些大偏差原理相关工作旨在寻找概率论文献中其他出现过这些观点的地方，但当时并未意识到它们还曾出现于其他文献。我们最近得知了这样一处：Zhivotovskiy 最近关于独立随机矩阵之和的算子范数的工作中的一条评注 [Zhi21，评注 6]。Zhivotovskiy 指出，这些相同的观点此前已隐含地出现在 Catoni 及其合作者的工作中。

1.3 广义样本协方差矩阵的先前结果。

该模型 H_N 及其变体都相当经典。据我们所知，它们最早出现在 Marčenko 和 Pastur 于 1967 年发表的一篇论文中。该论文证明了变体 $A + \frac{1}{M} Z^T \Gamma Z$的全局定律，其中 A 是一个确定性矩阵序列，其经验测度具有某个极限，而 Γ 是随机对角矩阵，其元素独立同分布。我们这一确切模型的全局定律见 Silverstein 和 Bai 的工作（该工作更为一般） [SB95], ，其中还很好地总结了关于 H_N的各种变体的全局定律文献。

在我们的工作中，我们假设 Γ 没有离群特征值；例如，我们不允许 Γ 成为单位矩阵的有限秩扰动。人们普遍认为，并且在非常温和的限制条件下已经证明，这一限制可以防止 H_N 通常出现离群特征值。也就是说，我们不处于 Baik-Ben Arous-Péché（BBP）相变所在的情形。

因此，最大特征值趋于 H_N 的渐近支撑集的右端点。对于 H_N 的某些变体，已知其在该点附近具有 Tracy-Widom 涨落。例如，在具有行列式型结构的复高斯情形下，El Karoui [EK07] 引入了边缘正则性条件，并处理了满足该条件的所有 Γ，其中 $\frac{M}{N} \geqslant 1$；Onatski [Ona08] 将这一结果推广到了 $\frac{M}{N} < 1$。在普适性方面，Bao、Pan 和 Zhou [BPZ15] 允许 Z 具有复非高斯元素；Lee 和 Schnelli [LS16] 允许 Z 具有实非高斯元素，只要 Γ 是对角的；Knowles 和 Yin [KY17] 给出了将这一结果推广到非对角 Γ 的（非平凡）扩展。尽管我们只研究最大特征值，但考虑这样的情形也具有统计学意义： Γ 使得 H_N 具有渐近不连通支撑集，并证明支撑集的每个连通分支所对应的最右侧特征值都具有 Tracy-Widom 涨落，且彼此独立。Hachem、Hardy 和 Najim [HHN16] 在复高斯情形下获得了这一结果，随后 Fan 和 Johnstone [FJ22] 在实高斯情形下取得了相应进展。最后，Wang [Wan19] 最近首次得到了收敛到 Tracy-Widom 分布的非平凡速度，其阶为 $N^{-1/57}$ （在 $\Gamma = \mathrm{Id}$ 的情形下，Wang 发现了快得多的速度 $N^{-2/9}$，Schnelli 和 Xu [SX21] 最近又将其改进到了几乎 $N^{-1/3}$）。[1]

1.4 结构安排。

本文的结构安排如下：第 2节中，我们定义模型及其相应的 x_c 阈值，并陈述我们的

[1]许多引用的文献针对的是模型 $\frac{1}{M} \Gamma^{1/2} Z^T Z \Gamma^{1/2}$。当然，该模型与我们的模型具有相同的非零特征值，但需要 $\Gamma \geqslant 0$。

主要结果。证明分为两个阶段：首先讨论具有 $x_c = +\infty$ 的模型（见第 3节），然后通过使用 $x_c < +\infty$ 的模型进行逼近，讨论具有 $x_c = +\infty$ 的模型（见第 4 节）。第 5节中，我们概述了当 H_N 为复数域上的厄米矩阵而非实数域上的对称矩阵时所需的细微调整。

在附录 A中，我们给出了 Talagrand 关于乘积测度集中性的经典结果的一个直接推广，但据我们所知，文献中尚未出现过这一推广：他的结果针对取值于 $[-1,1]$ 的独立随机变量，而我们将其推广到取值于 d 维单位球中的独立随机变量，任意 d 均适用。估计的精度不依赖于 d （这主要是因为该球的半径不依赖于 d），这一点本身或许也值得关注。特殊情形 $d = 2$ 使我们能够考虑复随机矩阵，其上三角元素彼此独立，但实部和虚部仅仅是不相关，而非独立，这正是此前结果所要求的条件。最后，在附录 B 中，我们通过示例说明如何去除文献中相关工作的“紧支撑或对数索伯列夫”假设；在附录 C中，我们陈述并证明关于变形 Wigner 矩阵的结果。

1.5 符号约定。

我们分别将函数 $f : \mathbb{R} \to \mathbb{R}$ 的 Lipschitz 范数和有界 Lipschitz 范数记为

$$\|f\|_{\mathrm{Lip}} = \sup_{x \neq y} \left| \frac{f(x) - f(y)}{x - y} \right|,$$
$$\|f\|_{\mathcal{L}} = \|f\|_{\mathrm{Lip}} + \|f\|_{\infty}.$$

我们需要以下函数类

$$\mathcal{F}_{\mathrm{Lip}} = \{f : \mathbb{R} \to \mathbb{R} : \|f\|_{\mathcal{L}} \leqslant 1\}$$

对于给定的紧集 $\mathcal{K} \subset \mathbb{R}$,

$$\mathcal{F}_{\mathrm{Lip},\mathcal{K}} = \{f \in \mathcal{F}_{\mathrm{Lip}} : \mathrm{supp}(f) \subset \mathcal{K}\}.$$

我们还需要定义 $\mathbb{R}$ 上概率测度之间的有界 Lipschitz 距离、Wasserstein-1 距离和 Kolmogorov-Smirnov 距离，定义如下:

$$\begin{aligned} d_{\mathrm{BL}}(\mu,\nu) &= \sup_{f \in \mathcal{F}_{\mathrm{Lip}}} \left| \int_{\mathbb{R}} f(x)(\mu - \nu)(\mathrm{d}x) \right|, \\ \mathrm{W}_1(\mu,\nu) &= \sup_{f : \|f\|_{\mathrm{Lip}} \leqslant 1} \left| \int_{\mathbb{R}} f(x)(\mu - \nu)(\mathrm{d}x) \right|, \\ d_{\mathrm{KS}}(\mu,\nu) &= \sup\{|\mu((-\infty, x]) - \nu((-\infty, x])| : x \in \mathbb{R}\}, \end{aligned} \tag{1.3}$$

其中第一个距离刻画弱收敛。若概率测度 μ 具有紧支撑，我们将其斯蒂尔切斯变换记为 G_μ ，并采用如下符号约定

$$G_\mu(z) = \int \frac{\mu(\mathrm{d}\lambda)}{z - \lambda},$$

并分别将其左端点和右端点记为 $\ell(\mu)$ 和 $r(\mu)$。

我们引入 Dyson 指数 β，作为所考虑对称类的简写，即实对称（$\beta = 1$）或复 Hermitian（$\beta = 2$）。

如果 T 是一个矩阵，我们将其关于欧几里得距离的算子范数写作 $\|T\|$，将其 Frobenius 范数写作$\|T\|_F^2 = \sum_{i,j} |T_{ij}|^2$，并将其迹范数写作 $\|T\|_* = \sum_i \sigma_i(T)$。如果 T 是方阵且 $N \times N$，我们将其特征值编号为

$$\lambda_{\min}(T) = \lambda_1(T) \leqslant \lambda_2(T) \leqslant \cdots \lambda_N(T) = \lambda_{\max}(T),$$

有时在记号中省略对 T 的依赖，并将其经验谱测度写作

$$\hat{\mu}_T = \frac{1}{N} \sum_{i=1}^{N} \delta_{\lambda_i(T)}. \tag{1.4}$$

我们用 Leb(·) 表示 L 勒贝格测度，并且在附录中我们将需要归一化的半圆律 as

$$\rho_{\mathrm{sc}}(\mathrm{d}x) = \frac{\sqrt{(4 - x^2)_+}}{2\pi}\, \mathrm{d}x.$$

1.6 致谢。

我们谨向 Alice Guionnet 致谢，感谢她提出这个问题并进行了许多有益的讨论。我们也感谢 Yizhe Zhu 为我们指明了参考文献 [Zhi21]。本文所述工作得到了美国国家科学基金会 DMS-1928930 号资助；作者在 2021年秋季于加利福尼亚州伯克利的数学科学研究所驻留期间完成了这项工作。

2 结果

2.1 设定。

定理2.13是我们在实数情形下的主要结果，陈述于第2.4节；复数类比——定理2.19——见第2.5节。为恰当地陈述这些结果，我们需要在本节（第2.1节）精确定义我们的模型，以及其经验测度的极限 σ 和一个特定函数 G_σ；后者是斯蒂尔切斯变换的一种"第二分支"，在第2.2节中定义大偏差原理的速率函数时需要用到。在第2.3 节中，我们讨论若干退化情形。

我们回顾式（1.1）中的基本随机矩阵

$$H_N = \frac{1}{M} Z^T \Gamma Z = \frac{1}{M} \sum_{i=1}^{M} d_i z_i z_i^T,$$

现在我们对其进行精确定义。固定 $\alpha > 0$，并选取一个序列 $M = (M_N)_{N=1}^{\infty}$，使得

$$\lim_{N\to\infty} \frac{M_N}{N} = \alpha.$$

对于每个 M，我们考虑一个确定性的实 $M \times M$ 矩阵

$$\Gamma = \Gamma_M = \mathrm{diag}(d_1, \ldots, d_M),$$

其中 $(d_1, \ldots, d_M) = (d_1^{(M)}, \ldots, d_M^{(M)})$ 不失一般性地按序排列为 $d_1 \leqslant \cdots \leqslant d_M$。我们假设存在一个紧支撑概率测度 ρ，使得

$$\hat{\mu}_{\Gamma_M} \to \rho \quad \text{weakly.}$$

我们需要以下定义。

定义 2.1。 $\mathbb{R}$ 上的中心化概率测度 μ 称为尖锐次高斯 测度，如果其方差为单位方差且

$$\int_{\mathbb{R}} e^{tx} \mu(\mathrm{d}x) \leqslant e^{\frac{t^2}{2}} \quad \textit{for all } t \in \mathbb{R}.$$

标准高斯测度当然是尖锐次高斯测度，但拉德马赫分布$_1$ $_2(\delta_{-1} + \delta_{+1})$和$[-\sqrt{3}, \sqrt{3}]$上的均匀分布也是如此。我们固定一个尖锐次高斯测度 μ，并令 $Z \in \mathbb{R}^{M\times N}$ 为一个元素独立同分布且服从 μ的矩阵。为与文献中的记号一致，我们取 Z的行，它们是长度为 N的独立向量，并将其写成列形式：$\{z_i\}_{i=1}^{M}$。

本文全文均采用以下假设，但仅在主定理中明确陈述。

假设 1. 我们假设 Γ 渐近地不存在（外部）离群值，具体而言 at

$$\lambda_{\min}(\Gamma) \to \ell(\rho),$$
$$\lambda_{\max}(\Gamma) \to r(\rho).$$

我们允许存在内部离群值，也就是说，如果 ρ 具有不连通支撑集， Γ 可以在各个分量之间的间隙中具有特征值。

2.2 戴森方程与经验测度的极限。以下两个阈值将起到重要作用：

定义 2.2。设

$$\theta_{\max} := \begin{cases} \frac{\alpha}{r(\rho)} & \text{if } r(\rho) > 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{2.1}$$

$$x_c(\rho) := \begin{cases} r(\rho)^2 G_\rho(r(\rho)) + \left(\frac{1}{\alpha} - 1\right) r(\rho) & \text{if } r(\rho) > 0 \text{ and } G_\rho(r(\rho)) < \infty, \\ +\infty & \text{otherwise} \end{cases} \tag{2.2}$$

在这样的设定下，关于经验测度 $\hat{}\mu_{H_N}$ 的以下收敛定理基本上是经典结果：

定理 2.3。对于上述 H_N，经验测度序列 $(\hat{}\mu_{H_N})_{N=1}^{\infty}$ 依概率弱收敛到一个紧支撑测度 σ。此外，σ 在 $(r(\sigma)\ +\infty)$ 上的斯蒂尔切斯变换满足

$$H_\rho(G_\sigma(x)) = x, \tag{2.3}$$

其中 H_ρ 定义在 $(0, \theta_{\max})$ 上，如下所示

$$H_\rho(y) := \frac{1}{y} + \int_{\mathbb{R}} \frac{\alpha u}{\alpha - yu} \rho(\mathrm{d}u). \tag{2.4}$$

如果 $r(\rho) \leqslant 0$，那么 $r(\sigma) \leqslant 0$。（不过，当 $r(\sigma) < 0$ 时，可能出现 $r(\rho) > 0$；例子见备注 *2.5* 。）

证明。 [SB95]中的定理 1.1 给出了所有结论，但测度 σ 是紧支撑的，以及当 $r(\rho) \leqslant 0$时，$r(\sigma) \leqslant 0$ 这两点除外。为证明前者，注意到由于 ρ 是紧支撑的，H_ρ 在零点的某个邻域内是亚纯函数，且在零点有一个简单极点。由反函数定理，$z \mapsto {}^{1}_{H_\rho}(z)$ 因此在零点的某个邻域内全纯，所以（通过 (2.4)）G_σ 在无穷远点的某个邻域内全纯，这意味着 σ 确实是紧支撑的。为证明后者，注意到我们总能选择没有离群值的 Γ，也就是说，以这样的方式使得在有限 N处，$\ell(\hat{}\mu_\Gamma) = \ell(\rho)$ 且 $r(\hat{}\mu_\Gamma) = r(\rho)$；对于用这样的 Γ 定义的 H_N，我们有 $H_N \leqslant r(\rho) \leqslant 0$，因此 $r(\sigma) \leqslant 0$。 □

注2.4。虽然我们不会用到它，但测度 σ 可以用自由概率的语言来解释。事实上，当 ρ 支撑在 $\mathbb{R}^+$ 或 $\mathbb{R}^-$上时，σ 就是 ρ 与相应的马尔琴科-帕斯图尔定律的自由乘法卷积。自由乘法卷积通常只针对支撑在半直线上的测度定义；但当 ρ 有两个非平凡分量 $\rho^{\pm}(A) = \rho(A \cap \mathbb{R}^{\pm})$时，我们可以将 σ 写成测度 $\sigma^{\pm}$的自由加法卷积；这些测度分别是 $\rho^{\pm}$ 与马尔琴科-帕斯图尔定律的自由乘法卷积。

事实上，写成 $\rho = \rho^+ + \rho^-$ 会将 $\Gamma = \Gamma^+ + \Gamma^-$ 分解为正负 d_i，从而将 $H_N = H_N^+ + H_N^-$ 分解为 $H_N^{\pm} = \frac{1}{M} Z^T \Gamma^{\pm} Z$。矩阵 H_N^+和 H_N^- 相互独立，它们的经验测度分别收敛于 $\sigma^{\pm}$。无论 Z所服从的分布如何，这一点都成立；我们将其选为高斯分布。此时，由于高斯分布的旋转不变性，它们的联合分布与 H_N^+ 和 $OH_N^-O^T$相同，其中 O是一个与其他所有对象独立的 *Haar*正交/酉矩阵；又由于这些矩阵渐近地相互自由独立，它们之和的经验测度收敛于这些经验测度的自由加法卷积。

注2.5。注意，*w* *e* 可以有 $r(\rho) > 0$，但 $r(\sigma) < 0$。例如，我们选择 $M_N = 4N$，并取 e

$$\Gamma_N = diag(\underbrace{2, \ldots, 2}_{2N \text{ times}}, \underbrace{-2K, \ldots, -2K}_{2N \text{ times}})$$

其中 K 是稍后将确定的常数。对于该模型，我们有 $\rho = \frac{1}{2}(\delta_{-2K} + \delta_2)$。注*2.4*说明，在这种情况下，$\sigma$ 是两个测度的自由加法卷积；这两个测度分别是位于 *2* 和 $-2K$ 处的狄拉克质量与马尔琴科-帕斯图尔分布的自由乘法卷积。由于此处的自由乘法卷积只是对测度进行缩放，σ 是两个伸缩的马尔琴科-帕斯图尔分布的自由加法卷积，一个位于 *2* 附近，另一个位于 $-2K$ 附近。这里乘以 *2* 是为了使马尔琴科-帕斯图尔分布与零点保持间隔；由于一般而言 $r(\mu \boxplus \nu) \leqslant r(\mu) + r(\nu)$，因此通过取足够大的 K，我们就能使 $r(\sigma) < 0$。

方程（2.3）称为戴森方程。下面的引理定义了某个函数 G_σ，应将其视为斯蒂尔切斯变换的第二个分支。其证明将在第3.7节给出。

引理 *2.6.* 首先，若 $x_c(\rho)$ 有限，则 $x_c(\rho) = H_\rho(\theta_{\max}) := \lim_{y\uparrow\theta_{\max}} H_\rho(y)$，且 $x_c(\rho) \geqslant r(\sigma)$。其次，除非处于 $r(\rho) \leqslant 0$ 且 $r(\sigma) = 0$ 的情形（此退化情形在第 *2.3* 节中处理），存在一个实值连续函数 G_σ，其定义域为

$$D := \begin{cases} [r(\sigma), +\infty) & \text{if } r(\rho) > 0, \\ [r(\sigma), 0) & \text{if } r(\rho) \leqslant 0, \end{cases} \tag{2.5}$$

具有以下性质：

1. 若 $x \in D$ 且 $x \leqslant x_c(\rho)$，则 $H_\rho(G_\sigma(x)) = x$，并且 $\{w : H_\rho(w) = x\} = \{G_\sigma(x)\ \overline{G_\sigma}(x)\ \}$.
2. 对于 $x > \overline{r}(\sigma)$，有 $G_\sigma(x) > G_\sigma(x)$，且 $G_\sigma(r(\sigma)) = G_\sigma(r(\sigma))$。
3. G_σ 在 $D \setminus \{x_c(\rho)\}$ 上实解析。
4. G_σ 在 D上单调不减，更具体地说
   - 如果 $r(\rho) > 0$ 且 $x_c(\rho) = +\infty$，那么 G_σ 在 D上严格递增，并且 $\lim_{x\to+\infty} G_\sigma(x) = \theta_{\max}$。
   - 如果 $r(\rho) > 0$ 且 $x_c(\rho) < +\infty$，那么 G_σ 在 $(r(\sigma)\ x_c(\rho))$ 上严格递增，并且 $\lim_{x\uparrow x_c(\rho)} G_\sigma(x) = \theta max$，且对于 $x \geqslant x_c(\rho)$，$G_\sigma(x) = \theta max$。
   - 如果 $r(\rho) < 0$，那么 G_σ 在 D上严格递增，并且 $\lim_{x\uparrow 0} G_\sigma(x) = +\infty$。

2.3 退化情形。

就我们的目的而言，退化行为有两种可能。我们先对它们作非正式说明：

1. 如果 $\rho(\{0\}) > 0$，这意味着我们写作

$$H_N = \frac{1}{M}\sum_{i=1}^{N} d_i z_i z_i^T$$

其中，d_i 中有宏观比例的项（渐近地）为零。我们的结果确实适用于这种情形；不过，为求完整起见，我们将在下文检验：若改为舍弃这些零 d_i，速率函数是否仍保持不变。这相当于从 ρ 中移除零原子，将其重新归一化以使其成为概率测度，并相应地调整 α。

2. 情形 $r(\rho) \leqslant 0$ 且 $r(\sigma) = 0$ 是退化的，因为此时 $\lambda_{\max}(H_N)$ 基本上被限制在零处：一方面，$\lambda_{\max}(H_N)$ 在这个尺度下无法进入谱 bulk，因此它必须渐近地非负；另一方面，H_N（渐近地几乎）是负半定的，因此 $\lambda_{\max}(H_N)$ 必须渐近地非正。这一点可由退化大偏差原理精确表述。

We 现在将刚才阐述的结果正式表述如下。这里的所有结论均在第3.6节中证明。 .

引理 2.7。若 ρ 是定义在 $\mathbb{R}$ 上的测度，满足 $r(\rho) \leqslant 0$ 且 $\rho(\{0\}) = 0$，那么以下各项等价 t:

- $r(\sigma) = 0$,
- $\alpha \leqslant 1$。

注 2.8。若 $\rho(\{0\}) > 0$，则我们断言，以下两种做法得到相同的结果（即相同的σ 和相同的速率函数）：（*a*）直接对 ρ 应用我们的结果；或（*b*）构造一个新的测度 τ，去除零点尖峰，将 α 改为某个 α'，并考虑相应的 H'_N。

事实上，如果 $\rho(\{0\}) = \beta > 0$，我们可以写成 $\rho = (1-\beta)\rho' + \beta\delta_0$，其中 ρ' 是一个测度，满足 $r(\rho) \leqslant 0$，且 $\rho'(\{0\}) = 0$。那么很容易看出，$H_N = H'_N + H''_N$，其中 $H'_N = Z'^T\Gamma'_N Z'$，且 $H''_N = Z''^T\Gamma''_N Z''$，其中 Γ'_N是一个 $M'_N \times M'_N$矩阵，满足 $\lim_N \frac{M'_N}{N} = \alpha(1-\beta)$，并且 Γ'_N 的经验测度收敛到 ρ'，而 $||\Gamma''_N||$ 收敛

到 0。因此，对于 $\varepsilon > 0$，$\lim \frac{1}{N} \log \mathbb{P}[||H''_N|| \geqslant \varepsilon] = -\infty$。这将问题归结为给出 H'_N的大偏差原理；此外，我们证明了 $r(\sigma) = 0$ 当且仅当 $\alpha(1-\beta) \leqslant 1$（我们将在下面的推论中陈述这一结果）。令 $\Delta_N := \frac{M'_N}{M_N}\Gamma'_N$，我们可以将 H'_N改写为 $H'_N = \frac{1}{M'_N} Z'^T \Delta_N Z'$。由于 Δ_N的经验测度收敛到 $\tau := (1-\beta) * \rho'$，我们可以将主要结果定理 2.13 应用于 H'_N，其中用 τ代替 ρ，并用 $\alpha' := \alpha(1-\beta)$ 代替 α。接下来只需证明，将 ρ 改为 ρ'、将 α 改为 α'时，定理 2.13 的陈述保持不变。为此，我们只需证明函数 H_ρ 和 H_τ 相同，因为我们将看到，大偏差原理的速率函数只通过 H_ρ依赖于 ρ。事实上，

$$\begin{aligned} H_\rho(y) &= \frac{1}{y} + \int \frac{\alpha u}{\alpha - yu}\rho(du) = \frac{1}{y} + (1-\beta)\int \frac{\alpha u}{\alpha - yu}\rho'(du) \\ &= \frac{1}{y} + (1-\beta)\int \frac{\alpha(1-\beta)^{-1}u}{\alpha - (1-\beta)^{-1}yu}\tau(du) = \frac{1}{y} + \int \frac{(1-\beta)\alpha u}{(1-\beta)\alpha - yu}\tau(du) = H_\tau(y). \end{aligned}$$

因此，通过将定理2.13 应用于 H'_N 所得到的速率函数，与将定理2.13 应用于 H_N所得到的相同。

推论 2.9。若 ρ 是 $\mathbb{R}$ 上的一个测度，且 $r(\rho) \leqslant 0$，则以下条件等价：

- $r(\sigma) = 0$,
- $\alpha(1 - \rho(\{0\})) \leqslant 1$。

定义 2.10。我们将推论2.9 中的（等价）条件概括为：称这对(ρ, α) 是退化的。

以下这一（显然的）结果为该定义提供了依据。

命题 2.11。定义 $I^{\text{degen}}: \mathbb{R} \to [0, +\infty]$ 如下

$$I^{\text{degen}}(x) = \begin{cases} 0 & \text{if } x = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

若 μ是高斯的，$\rho(\{0\}) = 0$，且二元组 (ρ, α) 是退化的，则 $\lambda_{\max}(H_N)$ 满足速度为 N、速率函数为 I^{degen}的良好大偏差原理。

在下文中，我们始终假设二元组 (ρ, α) 是非退化的。

2.4 实数情形下的主要结果。

定义 2.12。假设二元组 (ρ, α) 是非退化的。令 D 如 (2.5)所示，G_σ 为通常的斯蒂尔切斯变换，而 G_σ 为引理 2.6中的斯蒂尔切斯变换的“第二支”，定义 $I_\sigma: D \to [0, +\infty]$ 如下

$$I_\sigma(x) = \begin{cases} \frac{\beta}{2}\int_{r(\sigma)}^{x} [\overline{G}_\sigma(u) - G_\sigma(u)]\,\mathrm{d}u & \text{if } x \in D, \\ +\infty & \text{otherwise.} \end{cases}$$

我们的主定理在以下两个假设中的任一假设下成立，回顾一下，μ 是 Z中元素的共同分布。

Assu假设 A。测度 μ 是尖锐次高斯的，并且 ρ 的支撑集位于 $[0, \infty)$。

假设 B。测度 μ是高斯的。

定理 2.13。（主定理，实数版本）若假设 1 成立，则数对 (ρ, α) 是非退化的（定义 2.10中的含义），并且假设 A 或假设 B 也成立，那么 $\lambda_{\max}(H_N)$ 满足速度为 N 的大偏差原理，其良好速率函数为 I_σ。该函数是凸的，并且在 D上严格递增（特别地，它仅在 $r(\sigma)$处取零）。若另外 $r(\rho) > 0$，则

$$\lim_{x \to +\infty} \frac{I_\sigma(x)}{x} = \frac{\theta_{\max}}{2}.$$

定理2.13 将由以下两个结果推出。

命题 2.14。 若假设 *1* 成立，则数对 $(\rho,\ \alpha)$ 是非退化的，并且假设 *A* 或假设 *B* 也成立，且

$$x_c(\rho) = +\infty,$$

则 λ 最大值(H_N) 满足速度为 N 的大偏差原理，其良好速率函数为 I_σ。

命题 2.15。 命题*2.14* 蕴含定理 *2.13*。

注 2.16。 命题 *2.14* 和 *2.15* 的证明相当不同。命题 *2.14* 通过球面积分进行倾斜变换；对于每个 $x < x_c(\rho)$，我们都能找到适当的倾斜变换，使偏差$\{\lambda_{\max} \approx x\}$ 发生的可能性较大。但命题 *2.15* 既没有使用显式的倾斜变换，也几乎没有采用命题 *2.14*证明细节中的其他内容；相反，它通过利用 $x_c(\rho) < +\infty$ 模型来近似 $x_c(\rho) = +\infty$模型，并运用教科书结果：通过取一列近似大偏差原理的极限来得到大偏差原理。

注 2.17。 在非高斯情形下，要求 ρ 的支撑集包含于 $(0,\infty)$ 并非只是技术性要求。当某些 d_i 为负数时，速率函数很可能不同于高斯情形。事实上，假设所有 d_i 都等于某个 $-d$，其中 $d > 0$。那么显然 $\lambda_{\max}(-\frac{d}{M}Z^TZ) = -\lambda_{\min}(\frac{d}{M}Z^TZ)$，但 *Rademacher* 协方差矩阵最小特征值的左侧偏差与高斯情形的对应结果并不相同；*Rademacher* 速率函数在零点处必须有限，因为 $\mathbb{P}($ Z 的两列相同$) \geqslant 2^{-M}$。

2.5 复数情形下的主要结果。

我们的主要结果也适用于复数情形。不过在这种情况下，我们需要令 Z 的元素独立同分布，其分布遵循复平面上的某个中心化概率测度 μ，并满足 $\int(\Im z)^2\mu(dz) = \int(\Re z)^2\mu(dz) = 1/2$ 以及 $\int \Im z \Re z \mu(dz) = 0$。于是，我们的模型变为

$$H_N = \frac{1}{M}Z^*\Gamma Z$$

其中 Z^* 表示 Z的厄米共轭。关于 $M = M_N$ 和 Γ 的所有其他假设保持不变。我们可以将定义 2.1 扩展到复随机变量：

定义 2.18。 在 $\mathbb{C}$ 上的中心化概率测度 μ 称为关于 $\mathbb{C}$ 的尖锐次高斯测度，如果对于 X μ 分布，随机向量 $(\Re X, \Im X)$ 的协方差矩阵为 $\frac{1}{2}\begin{pmatrix}1&0\\0&1\end{pmatrix}$（实部和虚部必须不相关，但不必相互独立），并且

$$\int_{\mathbb{R}} e^{\Re(w\bar{z})}\mu(\mathrm{d}w) \leqslant e^{\frac{|z|^2}{4}} \quad \textit{for all } z \in \mathbb{C}.$$

因此，我们需要针对复数情形对假设 A 稍作更新：

假设 C。 测度 μ 关于 $\mathbb{C}$ 是尖锐次高斯的，且 ρ 的支撑集包含于 $[0,\infty)$.

于是，对于该模型，我们得到的大偏差原理与实数情形中得到的结果完全相同，只是这里的速率函数是实数情形下速率函数的两倍。

定理 2.19。 （主定理，复数版本）若假设 *1* 成立，则对 （ρ， α）这一对变量是非退化的，并且假设 *B* 或假设 *C* 成立，则 $\lambda_{\max}$（H_N）以速度 N满足大偏差原理，其良好速率函数为 $2I_\sigma$。

由于除了球面积分计算方面需要做一些调整外，证明基本保持不变，因此我们将在第5节简要说明所需的调整。

3 无限情形的证明 x_c

3.1 证明概要。本节证明命题2.14。证明首先引入速率函数的一种变分表述。虽然这种表述较为晦涩，但在技术上更为便利；随后通过球面积分进行倾斜变换，利用该变分表述证明大偏差原理。本节证明的大致思路与以往研究相似，但正如引言中所述，我们提出的新论证使我们能够绕过以往研究中的若干技术性假设。

定义 3.1。 对于 μ 上一个紧支撑概率测度，以及 $\mathbb{R}$、 $\theta \geqslant 0$ 和 $\lambda \geqslant r(\mu)$，令

$$v(\mu,\theta,\lambda) := \begin{cases} \lambda - \frac{1}{2\theta} & \text{if } G_\mu(\lambda) \leqslant 2\theta, \\ G_\mu^{-1}(2\theta) - \frac{1}{2\theta} & \text{if } G_\mu(\lambda) \geqslant 2\theta \geqslant 0, \end{cases}$$

$$J(\mu,\theta,\lambda) := \theta v(\mu,\theta,\lambda) - \frac{1}{2}\int_{\mathbb{R}} \log(1+2\theta v(\mu,\theta,\lambda) - 2\theta y)\mu(\mathrm{d}y),$$

定义 3.2。 对于 $0 \leqslant \theta < \theta_{\max}$，令

$$F(\rho,\theta) := -\frac{\alpha}{2}\int_{\mathbb{R}} \log\left(1 - \frac{\theta t}{\alpha}\right)\rho(\mathrm{d}t),$$

$$I_\sigma(x,\theta) := J\left(\sigma, \frac{\theta}{2}, x\right) - F(\rho,\theta).$$

~利用这些，定义 I_σ： $\mathbb{R} \to [0,+\infty]$ 为

$$\widetilde{I}_\sigma(x) := \begin{cases} \sup_{0\leqslant\theta<\theta_{\max}} I_\sigma(x,\theta) & \text{if } x \in D, \\ +\infty & \text{otherwise.} \end{cases}$$

引理 3.3。 （速率函数的简化）若 $x_c(\rho) = +\infty$，则有

$$I_\sigma(x) = \widetilde{I}_\sigma(x),$$

且该函数在 $[r(\sigma)\ \infty)$ 上为凹函数，并且仅在 $x = r(\sigma)$处取零。

Proof. 我们先来看第一种情况，即 $r(\rho) > 0$ 和 $G_\rho(r(\rho)) = +\infty$。我们计算 e

$$\frac{\partial}{\partial\theta}J\left(\sigma,\frac{\theta}{2},x\right) = \begin{cases} \frac{G_\sigma^{-1}(\theta)}{2} - \frac{1}{2\theta} & \text{if } \theta \leqslant G_\sigma(x) \\ \frac{x}{2} - \frac{1}{2\theta} & \text{if } \theta \geqslant G_\sigma(x) \end{cases}$$

我们还计算

$$\frac{\partial}{\partial\theta}F(\rho,\theta) = \frac{1}{2}\int_{\mathbb{R}} \frac{\alpha u}{\alpha - \theta u} d\rho(u) = \frac{H_\rho(\theta)}{2} - \frac{1}{2\theta}.$$

利用戴森方程（2.3），我们有

$$\frac{\partial}{\partial\theta}I_\sigma(x,\theta) = \begin{cases} 0 & \text{if } \theta \leqslant G_\sigma(x), \\ \frac{x}{2} - \frac{H_\rho(\theta)}{2} & \text{if } \theta \geqslant G_\sigma(x). \end{cases}$$

因此，函数 $I_\sigma(x,\cdot)$ 在 $[G_\sigma(x)G_\sigma(x)]$ 上递增，在 $[G_\sigma(x)\ +\overline{\infty})$ 上递减。因此， $\theta_x := \overline{G_\sigma}(x)$ 是使 θ 取最优值的取值，即，

$$\widetilde{I}_\sigma(x) = I_\sigma(x,\theta_x).$$

~（当 $x = r(\sigma)$ 时，我们按约定定义 $\theta_{r(\sigma)} := \theta_c$，并注意到这一论证表明 $I_\sigma(r(\sigma)) = 0$。）由此可计算导数:

$$\frac{\mathrm{d}}{\mathrm{d}x}\widetilde{I}_\sigma(x) = \frac{\partial}{\partial x}I_\sigma(x,\theta)\Big|_{\theta=\theta_x} = \frac{1}{2}(\theta_x - G_\sigma(x)) = \frac{1}{2}(\overline{G}_\sigma(x) - G_\sigma(x)).$$

~由于 I_σ 和 I_σ 的导数相同，且在 $x = r(\sigma)$ 处相等，因此结论成立。

当 $r(\sigma) < 0$ 时，对于 x 介于 $r(\sigma)$ 和 0 之间的情况，本质上类似。 □

注 3.4。这对于我们的证明并非必要，但我们注意到，当 $x_c(\rho) < +\infty$，即 $r(\rho) > 0$ 且 $G_\rho(r(\rho)) < \infty$ 时，可以验证 $\int_{\mathbb{R}} \log(r(\rho) - t)\rho(dt) > -\infty$，因此 $F(\rho, \theta_{\max})$ 和 $I_\sigma(x, \theta_{\max})$ 是有意义的。在这种情况下，可以定义 $I_\sigma^\dagger : \mathbb{R} \to [0, +\infty]$ 为

$$I_\sigma^\dagger(x) := \begin{cases} I_\sigma(x, \overline{G}_\sigma(x)) & \text{if } x \geqslant r(\sigma) \\ +\infty & \text{otherwise} \end{cases} = \begin{cases} \sup_{0 \leqslant \theta < \theta_{\max}} I_\sigma(x, \theta) & \text{if } r(\sigma) \leqslant x < x_c(\rho), \\ I_\sigma(x, \theta_{\max}) & \text{if } x \geqslant x_c(\rho), \\ +\infty & \text{otherwise.} \end{cases}$$

本注释的要点是，可以验证

$$I_\sigma(x) = I_\sigma^\dagger(x).$$

事实上，引理 3.3 的证明已经对此进行了验证：对于 $x < x_c(\rho)$，可以计算 $x \geqslant x_c(\rho)$。

$$\partial_x I_\sigma(x, \theta_{\max}) = \partial_x J(\sigma, \theta_{\max}/2, x) = \frac{1}{2}(\theta_{\max} - G_\sigma(x)) = \frac{1}{2}(\overline{G}_\sigma(x) - G_\sigma(x)),$$

这意味着 I_σ 和 $I_\sigma^\dagger$ 具有相同的导数。不过，I_σ 的定义在技术上更为方便，因此我们在证明中采用它。

假设 1 和 A 要求 $\mathrm{supp}(\rho) \subset [0, \infty)$，但允许在有限 d_i 中存在少量负的 N，这些值必须在维数趋于无穷时趋于零。然而，在技术上，处理所有d_i在有限 N时均为非负的情形更为方便。以下结果使我们可以将问题限制在这一情形；我们省略其证明，因为它与命题 2.11 的证明本质上相同。

引理 3.5。 在假设 *1*和 *A*下，定义 $\Gamma^+ = \mathrm{diag}(\max(d_1, 0) \ldots, \max(d_M, 0))$，并令 $H_N^+ = M^{-1} Z^T \Gamma^+ Z$。则 $\lambda_{\max}(H_N)$ 和 $\lambda_{\max}(H_N^+)$ 在指数意义下等价。特别地，其中一个对象的大偏差原理自动适用于另一个对象。

在下文中，必要时我们将 H_N 默认为 H_N^+。

引理 3.6。 （指数紧性） 在假设 *A* 或假设 *B*下，我们有

$$\lim_{K \to \infty} \limsup_{N \to \infty} \frac{1}{N} \log \mathbb{P}(|\lambda_{\max}(H_N)| > K) = -\infty.$$

证明。 如果 M 足够大，使得对于所有 i， $d_i \in (l(\rho) - 1, r(\rho) + 1)$，那么显然有

$$(l(\rho) - 1)\lambda_{\max}(M^{-1} Z^T Z) \leqslant \lambda_{\max}(H_N) \leqslant (r(\rho) + 1)\lambda_{\max}(M^{-1} Z^T Z)$$

从而

$$|\lambda_{\max}(H_N)| \leqslant \max(|l(\rho) - 1|, |r(\rho) + 1|)\lambda_{\max}(M^{-1} Z^T Z).$$

的指数紧性 $\lambda_{\max}$ （$M^{-1} Z^T Z$） 已在 [GH20，引理 1 中得到证明。9]。 □

定义 3.7。 （倾斜测度） 对于 $\theta \geqslant 0$，考虑 $\mathbb{P}^\theta$（Z） 在 $M \times N$ 矩阵上的倾斜测度，其密度为

$$\frac{\mathrm{d}\mathbb{P}^\theta}{\mathrm{d}\mathbb{P}}(Z) = \frac{\mathbb{E}_e[e^{N\frac{\theta}{2}\langle e, \frac{1}{M} Z^T \Gamma Z e\rangle}]}{\mathbb{E}_{e, H_N}[e^{N\frac{\theta}{2}\langle e, H_N e\rangle}]}.$$

Prop命题 3.8. (倾斜测度的弱大偏差原理上界**)** 对于 $0 \leqslant \theta < \theta_{\max}$，

$$\limsup_{\delta \downarrow 0} \limsup_{N \to \infty} \frac{1}{N} \log \mathbb{P}^\theta(|\lambda_{\max}(H_N) - x| \leqslant \delta) \begin{cases} \leqslant -(\widetilde{I}_\sigma(x) - I_\sigma(x, \theta)) & \text{if } x \in D, \\ = -\infty & \text{otherwise.} \end{cases}$$

注意到 $\mathbb{P}^0 = \mathbb{P}$，且 I_σ（x，0） $= 0$，因此特别地，我们得到了所关注测度的弱大偏差原理上界。

命题 3.9。（弱大偏差原理下界）

$$\liminf_{\delta\downarrow 0}\liminf_{N\to\infty}\frac{1}{N}\log\mathbb{P}(|\lambda_{\max}(H_N)-x|<\delta)\geqslant -\widetilde{I}_\sigma(x).$$

命题 2.14 可由此推出　按照经典方法，由引理 3.3、引理 3.6、命题 3.8和命题 3　.9.

注 3.10。实际上，本节给出的证明比刚才陈述的结果稍微更一般一些：从“在$x_c(\rho)$范围内”这个意义上说，它证明了

$$-I_\sigma(x)\leqslant\liminf_{\delta\downarrow 0}\liminf_{N\to+\infty}\frac{1}{N}\log\mathbb{P}(|\lambda_{\max}(H_N)-x|\leqslant\delta)\leqslant\limsup_{\delta\downarrow 0}\limsup_{N\to+\infty}\frac{1}{N}\log\mathbb{P}(|\lambda_{\max}(H_N)-x|\leqslant\delta)\leqslant -I_\sigma(x)$$

只要 $x<x_c(\rho)$，结论就成立，而不只是只要 $x_c(\rho)=+\infty$ 且 $x\in\mathbb{R}$ 时才成立，如前所述。不过，由于无论 $x_c(\rho)$如何，我们的最终结果都成立，因此这里不需要如此一般的结论。

3.2 退火球面积分。

本小节的目标是证明以下引理。

Lemma 3.11. 在假设 *A* 或假设 *B*下，对于每个 $0\leqslant\theta<\theta_{\max}$，我们有　　e

$$\lim_{N\to\infty}\frac{1}{N}\log\mathbb{E}_{e,H_N}[e^{N\frac{\theta}{2}\langle e,H_Ne\rangle}]=F(\rho,\theta).\tag{3.1}$$

证明。对于每个单位向量 e，我们有

$$\mathbb{E}_{H_N}[e^{N\frac{\theta}{2}\langle e,H_Ne\rangle}]=\mathbb{E}[e^{N\frac{\theta}{2}\left\langle e,\left(\frac{1}{M}\sum_{\mu=1}^M d_\mu z_\mu z_\mu^T\right)e\right\rangle}]=\mathbb{E}\left[\prod_{\mu=1}^M e^{\frac{N}{M}\frac{\theta}{2}d_\mu\langle z_\mu,e\rangle^2}\right]=\prod_{\mu=1}^M\mathbb{E}[e^{\frac{N}{M}\frac{\theta}{2}d_\mu\langle z_\mu,e\rangle^2}]$$

因为 z_μ 相互独立。应用 Hubbard-Stratonovich 变换，我们得到

$$\mathbb{E}_{H_N}[e^{N\frac{\theta}{2}\langle e,H_Ne\rangle}]=\prod_{\mu=1}^M\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}\mathbb{E}\left[e^{2x\langle z_\mu,e\rangle\sqrt{\frac{N}{M}\frac{\theta}{2}d_\mu}}\right]e^{-x^2}\,\mathrm{d}x,$$

对于那些为负的 d_μ，我们将 $\sqrt{d_\mu}=\mathrm{i}\sqrt{-d_\mu}$ 解释为

在高斯情形下（即在假设B下），证明的其余部分很简单：我们可以精确计算$\prod_{j=1}^N\mathbb{E}[e^{x\sqrt{2\frac{N}{M}\theta d_\mu}(z_\mu)_je_j}]=e^{x^2\frac{N}{M}\theta d_\mu}$，得到

$$\frac{1}{N}\log\mathbb{E}_{e,H_N}[e^{N\frac{\theta}{2}\langle e,H_Ne\rangle}]=-\frac{M}{2N}\int_{\mathbb{R}}\log\left(1-\frac{N}{M}\theta t\right)\hat{\mu}_\Gamma(\mathrm{d}t).$$

将 $\hat{\mu}_\Gamma$ 替换为 ρ，并将 $\frac{M}{N}$ 替换为 α 取极限是常规步骤，但这需要 $\theta<\theta_{\max}$。

在尖锐次高斯情形下（即在假设 A 下），上界类似：对于每个 实数 c以及每个单位 e，都有

$$\mathbb{E}[\exp(c\langle z_\mu,e\rangle)]=\prod_{j=1}^N\mathbb{E}[\exp(c(z_\mu)_je_j)]\leqslant\prod_{j=1}^N\exp\left(\frac{c^2e_j^2}{2}\right)=\exp\left(\frac{c^2}{2}\right).$$

由于每个 d_μ 都是正数，我们可以将此结果用于实数 $c=2\sqrt{\frac{N}{M}\frac{\theta}{2}d_\mu}$，从而得到

$$\frac{1}{N}\log\mathbb{E}_{e,H_N}[e^{N\frac{\theta}{2}\langle e,H_Ne\rangle}]\leqslant-\frac{M}{2N}\int_{\mathbb{R}}\log\left(1-\frac{N}{M}\theta t\right)\hat{\mu}_\Gamma(\mathrm{d}t),$$

which 完成了上界的证明。对于下界，固定 $0<\varepsilon<\frac{1}{4}$ 并定义　　-

$$V_N^\varepsilon=\{e:\|e\|_\infty\leqslant N^{-\frac{1}{4}-\varepsilon}\}\subset\mathbb{S}^{N-1}.$$

根据尖锐次高斯假设，我们知道，对于每个 $\delta>0$，存在 $\eta>0$，使得对于任意 $|t|\leqslant\eta$,

$$\int e^{tx}\mu(\mathrm{d}x)\geqslant e^{\frac{(1-\delta)}{2}t^2}.$$

特别地，如果我们固定 δ 足够小，使得 $f(t)=1-(1-\delta)_\alpha{}^{\theta t}$ 在 ρ 的支撑集上有下界，并记

$$\eta'=\frac{\eta}{\sqrt{2(\alpha+1)\theta(r(\rho)+1)}},$$

那么，只要 $e\in V_N^\varepsilon$ 且 $|x|\leqslant\eta' N^{\frac14+\varepsilon}$，对每个 μ，我们都有

$$\mathbb{E}\left[\exp\left(2x\langle z_\mu,e\rangle\sqrt{\frac{N}{M}\frac{\theta}{2}d_\mu}\right)\right]\geqslant\exp\left((1-\delta)x^2\frac{N}{M}\theta d_\mu\right).$$

因此，对于这样的 e，我们有

$$\mathbb{E}[e^{\frac{N}{M}\frac{\theta}{2}d_\mu\langle z_\mu,e\rangle^2}]\geqslant\frac{1}{\sqrt{\pi}}\int_{-\eta' N^{\frac14+\varepsilon}}^{\eta' N^{\frac14+\varepsilon}}\exp\left(-\left(1-(1-\delta)\frac{N}{M}\theta d_\mu\right)x^2\right)\mathrm{d}x.$$

根据标准高斯尾界，如果 c，$d>0$ 与 N 相互独立，那么

$$\frac{1}{\sqrt{\pi}}\int_{dN^{\frac14+\varepsilon}}^{\infty}\exp(-cx^2)\,\mathrm{d}x\leqslant\sqrt{c}\exp(-2cd^2N^{\frac12+2\varepsilon})$$

从而

$$\begin{aligned}\mathbb{E}[e^{\frac{N}{M}\frac{\theta}{2}d_\mu\langle z_\mu,e\rangle^2}]&\geqslant\left(1-(1-\delta)\frac{N}{M}\theta d_\mu\right)^{-\frac12}-2\sqrt{1-\frac{N}{M}\theta d_\mu}\exp\left(-2\left(1-\frac{N}{M}\theta d_\mu\right)(\eta')^2N^{\frac12+2\varepsilon}\right)\\&\geqslant\left(1-\frac{(1-\delta)\theta d_\mu}{\alpha}\right)^{-\frac12}\left(1-C\left|\frac{N}{M}-\frac{1}{\alpha}\right|\right)-C'\exp(-C''N^{\frac12+2\varepsilon})\end{aligned}$$

对于某些常数 C，C'，C''，它们通过 η 依赖于 δ，但不依赖于 μ。因此

$$\begin{aligned}\frac{1}{N}\log\mathbb{E}_{e,H_N}[e^{N\frac{\theta}{2}\langle e,H_Ne\rangle}]\geqslant\frac{M}{N}\int_{\mathbb{R}}\underbrace{\log\left[\left(1-\frac{(1-\delta)\theta t}{\alpha}\right)^{-\frac12}\left(1-C\left|\frac{N}{M}-\frac{1}{\alpha}\right|\right)-C'\exp(-C''N^{\frac12+2\varepsilon})\right]}_{=:f_N(t)}\rho(\mathrm{d}t)\\+\frac{1}{N}\log\mathbb{P}(e\in V_N^\varepsilon)\end{aligned}$$

由 [GH20 的引理 3.3],，我们有 $\mathbb{P}(e\in V_N^\varepsilon)\to1$。此外，$f_N(t)\to-\frac12\log(1-(1-\delta)_\alpha{}^{\theta t})$ 逐点收敛于 $N\to\infty$，并且根据我们对 δ 的选择，它在 ρ 的支撑集上有上界，因此由控制收敛可得

$$\liminf_{N\to\infty}\frac{1}{N}\log\mathbb{E}_{e,H_N}[e^{N\frac{\theta}{2}\langle e,H_Ne\rangle}]\geqslant-\frac{\alpha}{2}\int_{\mathbb{R}}\log\left(1-\frac{(1-\delta)\theta t}{\alpha}\right)\rho(\mathrm{d}t).$$

令 $\delta\downarrow0$，再次应用控制收敛即可完成证明。 □

3.3 集中测度。

本小节的目标是证明以下命题。

命题 3.12。 若假设 *A* 或假设 *B* 成立，则对任意 $\varepsilon>0$ 都有

$$\lim_{N\to\infty}\frac{1}{N}\log\mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N},\sigma)>\varepsilon)=-\infty.\tag{3.2}$$

注 3.13。 在我们的论文发布前不久，我们得知，本小节及附录 B 中的类似论证也将出现在 *Cook*、*Ducatez* 和 *Guionnet* 即将独立发布的工作中 *[CDG23]*。

证明。 对于假设A 和假设 B的不同情形，证明的结构是相同的。不过，有一项技术性估计在不同情形下的证明方法截然不同；我们将在下文适当位置指出这一点，除此之外，默认同时处理这两种情形。

考虑如下分解

$$Z = A + B,$$

其中

$$A_{ij} = Z_{ij}\mathbb{1}_{|Z_{ij}|\leqslant N^{\gamma}} \tag{3.3}$$

对于某个待选取的 $\gamma = \gamma(\varepsilon) > 0$。（回想一下， Z_{ij} 都是 1 阶量，因此 B 通常是稀疏的。）定义矩阵

$$H_N^A = \frac{1}{M} A^T \Gamma A$$

并且，对于较大的正数 L，定义事件

$$\mathcal{E}_L = \{\mathrm{supp}(\hat{\mu}_{H_N}) \subset (-L, L)\} \cap \{\mathrm{supp}(\hat{\mu}_{H_N^A}) \subset (-L, L)\}.$$

要证明（3.2），只需验证

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N}, \hat{\mu}_{H_N^A}) > \varepsilon, \mathcal{E}_L) = -\infty \quad \text{for every } L > 100r(\sigma), \tag{3.4}$$

$$\lim_{L\to\infty} \limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}(\mathcal{E}_L^c) = -\infty, \tag{3.5}$$

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N^A}, \mathbb{E}[\hat{\mu}_{H_N^A}]) > \varepsilon) = -\infty, \tag{3.6}$$

$$\lim_{N\to\infty} d_{\mathrm{BL}}(\mathbb{E}[\hat{\mu}_{H_N^A}], \sigma) = 0. \tag{3.7}$$

我们从（3.4）开始，证明时改编 Bordenave、Caputo 和 Chafaï 的论证（即 [BCC11，引理 C。2] 以及 [BC14，引理 2。2]）。只要 f 是一个 C^1 满足以下条件的测试函数：$\|f\|_{\mathrm{Lip}} + \|f\|_\infty \leqslant 1$，分部积分可得

$$\begin{aligned}\left|\int f(\lambda)(\hat{\mu}_{H_N} - \hat{\mu}_{H_N^A})(\mathrm{d}\lambda)\right| \mathbf{1}_{\mathcal{E}_L} &= \left|\int f'(\lambda)(F_{\hat{\mu}_{H_N}} - F_{\hat{\mu}_{H_N^A}})(\mathrm{d}\lambda)\right| \mathbf{1}_{\mathcal{E}_L} \leqslant \|f'\|_\infty \|F_{\hat{\mu}_{H_N}} - F_{\hat{\mu}_{H_N^A}}\|_1 \mathbf{1}_{\mathcal{E}_L} \\ &\leqslant 2L\|f\|_{\mathrm{Lip}} \|F_{\hat{\mu}_{H_N}} - F_{\hat{\mu}_{H_N^A}}\|_\infty \leqslant 2L d_{\mathrm{KS}}(\hat{\mu}_{H_N}, \hat{\mu}_{H_N^A}).\end{aligned}$$

若 f仅满足$\|f\|_{\mathrm{Lip}} + \|f\|_\infty \leqslant 1$，但不一定是 C^1，则存在一个 C^1函数 g，满足$\|g\|_{\mathrm{Lip}} + \|g\|_\infty \leqslant 1$且$\|f - g\|_{L^\infty([-L,L])} \leqslant 2L d_{\mathrm{KS}}(\hat{\mu}_{H_N}, \hat{\mu}_{H_N^A})$; 因此

$$\mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N}, \hat{\mu}_{H_N^A}) > \varepsilon, \mathcal{E}_L) \leqslant \mathbb{P}\Big(d_{\mathrm{KS}}(\hat{\mu}_{H_N}, \hat{\mu}_{H_N^A}) > \frac{\varepsilon}{4L}\Big).$$

它是分类 的（这是奇异值交错性的推论，参见例如 [BS10，定理 A。44]） t

$$d_{\mathrm{KS}}(\hat{\mu}_{H_N}, \hat{\mu}_{H_N^A}) \leqslant \frac{1}{N}\mathrm{rank}(Z - A) = \frac{1}{N}\mathrm{rank}(B) \leqslant \frac{1}{N}\sum_{i,j} \mathbb{1}_{|Z_{ij}|>N^{\gamma}}$$

对任意 Γ。现 在 $(\mathbb{1}_{|Z_{ij}|>N^{\gamma}})_{1\leqslant i\leqslant M,\ 1\leqslant j\leqslant N}$是一组均值为.....的 NM独立同分布的伯努利变量

$$p_N := \mathbb{P}(|Z_{ij}| > N^{\gamma}) \leqslant \exp(-cN^{2\gamma})$$

对于某个取决于 μ的次高斯范数的 c， 记作

$$\sigma^2 = NMp_N(1 - p_N) \leqslant \exp\Big(-\frac{c}{2}N^{2\gamma}\Big)$$

且 $h(x) = (x+1)\log(x+1) - x$，贝内特不等式 [Ben65] 给出

$$\mathbb{P}\left(\sum_{i\leqslant j} \mathbb{1}_{|Z_{ij}|>N^\gamma} - NMp_N \geqslant t\right) \leqslant \exp\left(-\sigma^2 h\left(\frac{t}{\sigma^2}\right)\right)$$

对任意 $t>0$。我们将选择 $t = N\frac{\varepsilon}{4L} - NMp_N \geqslant N\frac{\varepsilon}{8L}$，它具有 $\frac{t}{\sigma^2} \to +\infty$；由于 $h(x) \geqslant x\log x$ 对充分大的参数成立，因此我们有

$$\mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N}, \hat{\mu}_{H_N^A}) > \varepsilon, \mathcal{E}_L) \leqslant \exp\left(-t\log\left(\frac{t}{\sigma^2}\right)\right).$$

由于 $\gamma > 0$，这就足以证明（3.4）。

估计式（3.5）本质上是指数紧性引理 3.6 的推论，它控制 $|\lambda_{\max}(H_N)|$；同一证明也控制 $|\lambda_{\min}(H_N)|$（如果某些 d_i 为负，则该量可能为负），以及 H_N^A 的极端特征值（引理 3.6 的证明引用了 [GH20，引理 1。9]，该证明对 A 也适用）。

在尖锐次高斯情形（假设 A）和高斯情形（假设 B）下，（3.6）的验证方式差异较大。后一情形对应下文的引理 3.14，前一情形对应下文的引理 3.15（这里也会根据 γ 来选取 ε）。

对于（3.7），我们首先断言

$$\lim_{N\to\infty} d_{\mathrm{BL}}(\mathbb{E}[\hat{\mu}_{H_N^A}], \mathbb{E}[\hat{\mu}_{H_N}]) = 0. \tag{3.8}$$

只需证明 $\mathbb{E}[X_N] \to 0$，其中 $X_N = d_{\mathrm{BL}}$（$\hat{\mu}_{H_N^A}$、$\hat{\mu}_{H_N}$）。但 X_N是一个取值介于零和二之间的随机变量，而（3.4）和（3.5）表明，对于任意 $\varepsilon > 0$ 和 $N \geqslant N_0(\varepsilon)$，都有 $\mathbb{P}(X_N > \varepsilon) \leqslant \exp(-100N)$，从而得到$\mathbb{E}[X_N] \to 0$，进而得到（3.8）。剩下只需证明

$$\lim_{N\to\infty} d_{\mathrm{BL}}(\mathbb{E}[\hat{\mu}_{H_N}], \sigma) = 0.$$

这等价于断言 $\mathbb{E}[\hat{\mu}_{H_N}] \to \sigma$。这类结果最初由 Marčenko 和 Pastur 提出 [MP67]，但针对这样的模型，即其中的 d_i 是从 ρ中独立同分布抽取得到的，而不是满足性质 $\frac{1}{M}\sum\delta_{d_i} \to \rho$的确定性量；对于我们这里采用的这一模型的后一个变体，结果归功于 Silverstein 和 Bai [SB95]（实际上，他们的结果适用于更一般的情形）。 □

引理 3.14。 在假设 B下，对于每个 $\varepsilon > 0$，存在 $\gamma = \gamma(\varepsilon) > 0$ 使得，若 H_N^A 根据 γ 利用 (3.3)定义，则

$$\lim_{N\to\infty} \frac{1}{N}\log\mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N^A}, \mathbb{E}[\hat{\mu}_{H_N^A}]) > \varepsilon) = -\infty. \tag{3.9}$$

证明。 在这种情况下，A 的截断是不连续的，这在技术上处理起来不便；因此我们进一步分解

$$A = C + D,$$

其中

$$C_{ij} = \begin{cases} Z_{ij} & \text{if } |Z_{ij}| \leqslant N^\gamma, \\ N^\gamma \mathrm{sign}(Z_{ij}) & \text{if } |Z_{ij}| > N^\gamma \end{cases}, \qquad D_{ij} = \begin{cases} 0 & \text{if } |Z_{ij}| \leqslant N^\gamma, \\ -N^\gamma \mathrm{sign}(Z_{ij}) & \text{if } |Z_{ij}| > N^\gamma \end{cases},$$

并定义矩阵

$$H_N^C = \frac{1}{M} C^T \Gamma C.$$

要证明（3.9），只需验证

$$\lim_{N\to\infty} \frac{1}{N}\log\mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N^A}, \hat{\mu}_{H_N^C}) > \varepsilon, \mathcal{E}_L) = -\infty \quad \text{for every } L > 100r(\sigma), \tag{3.10}$$

$$\lim_{N\to\infty} \frac{1}{N}\log\mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N^C}, \mathbb{E}[\hat{\mu}_{H_N^C}]) > \varepsilon) = -\infty, \tag{3.11}$$

$$\lim_{N\to\infty} d_{\mathrm{BL}}(\mathbb{E}[\hat{\mu}_{H_N^C}], \mathbb{E}[\hat{\mu}_{H_N^A}]) = 0. \tag{3.12}$$

（3.10）的证明与（3.4）的证明非常相似。具体而言，与之前相同，可以证明

$$\mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N^A},\hat{\mu}_{H_N^C})>\varepsilon,\mathcal{E}_L)\leqslant\mathbb{P}\Big(d_{\mathrm{KS}}(\hat{\mu}_{H_N^A},\hat{\mu}_{H_N^C})>\frac{\varepsilon}{4L}\Big)\leqslant\mathbb{P}\Big(\frac{1}{N}\mathrm{rank}(D)>\frac{\varepsilon}{4L}\Big).$$

前面我们通过矩阵的非零元素个数来界定其秩。 B 这里我们对 D也采用同样的方法，但根据构造，B 和 D 的非零元素个数相同，因此其余部分的估计完全相同。类似地，式（3.12）的证明与式（3.8）的证明类似。

因此，只需证明式（3.11），我们先证明

$$\sup_{f\in\mathcal{F}_{\mathrm{Lip}}}\mathbb{P}\Big(\Big|\int_{\mathbb{R}}f(\lambda)(\hat{\mu}_{H_N^C}-\mathbb{E}[\hat{\mu}_{H_N^C}])(\mathrm{d}\lambda)\Big|\geqslant\delta\Big)\leqslant 2\exp\Big(-\frac{\delta^2N^{2-2\gamma}}{2}\Big).\tag{3.13}$$

事实上，我们稍微转换一下视角，定义 $C:\mathbb{R}^{M\times N}\to\mathbb{R}^{M\times N}$ 为

$$C(Z)_{ij}=\begin{cases}Z_{ij} & \text{if } |Z_{ij}|\leqslant N^{\gamma},\\ N^{\gamma}\mathrm{sign}(Z_{ij}) & \text{if } |Z_{ij}|>N^{\gamma},\end{cases}$$

并将 $H_N^{C(Z)}$ 定义为 $H_N^{C(Z)}=M^{-1}C(Z)^T\Gamma C(Z)$。取定某个 $f\in\mathcal{F}_{\mathrm{Lip}}$，并考虑映射 $h=h_f\colon\mathbb{R}^{M\times N}\to\mathbb{R}$，其定义如下

$$h(Z)=\int_{\mathbb{R}}f(\lambda)\hat{\mu}_{H_N^{C(Z)}}(\mathrm{d}\lambda).$$

我们要证明 h 是 Lipschitz 连续的。利用 L^p 形式的霍夫曼-维兰特不等式，其中 $p=1$ （例如参见[Kat87，定理 II]），并令 $\pi\in S_N$ 表示 N 个字母上的置换群中的一个置换，我们有

$$\begin{aligned}|h(Z_1)-h(Z_2)|&\leqslant\min_{\pi\in S_N}\frac{1}{N}\sum_{i=1}^N\Big|f(\lambda_i(H_N^{C(Z_1)}))-f(\lambda_{\pi(i)}(H_N^{C(Z_2)}))\Big|\leqslant\min_{\pi\in S_N}\frac{1}{N}\sum_{i=1}^N\Big|\lambda_i(H_N^{C(Z_1)})-\lambda_{\pi(i)}(H_N^{C(Z_1)})\Big|\\&\leqslant\frac{1}{N}\sum_{i=1}^N\Big|\lambda_i(H_N^{C(Z_1)}-H_N^{C(Z_2)})\Big|=\frac{1}{N}\|H_N^{C(Z_1)}-H_N^{C(Z_2)}\|_*\\&\leqslant\frac{1}{NM}(\|C(Z_1)^T\Gamma(C(Z_1)-C(Z_2))\|_*+\|(C(Z_1)-C(Z_2))^T\Gamma C(Z_2)\|_*)\end{aligned}$$

回顾一下，对于矩阵 M，$\|M\|_F=\sqrt{\sum_{i,j}|M_{ij}|^2}$ ~~表示~~其 Frobenius 范数，我们现在将使用以下结论：$\|T_1T_2\|_*\leqslant\|T_1\|_F\|T_2\|_F$；以及$\|C(Z_1)^T\Gamma\|_F\leqslant\sqrt{N}\|M\|d_{\max}N^{\gamma}$，因为其元素的模至多为 $d_{\max}N^{\gamma}$；并且

$$\|C(Z_1)-C(Z_2)\|_F\leqslant\|Z_1-Z_2\|_F$$

（这个估计正是我们需要定义 C 的原因；对于 A，类似的结论并不成立）。由此可得

$$|h(Z_1)-h(Z_2)|\leqslant 2d_{\max}\frac{N^{\gamma}}{\sqrt{NM}}\|Z_1-Z_2\|_F\leqslant\frac{4d_{\max}}{\sqrt{\alpha}}\frac{N^{\gamma}}{N}\|Z_1-Z_2\|_F.$$

由于 $\mathbb{R}^{M\times N}\cong\mathbb{R}^{MN}$ 上的高斯测度满足常数为 1 的对数 Sobolev 不等式，赫布斯特论证便给出

$$\mathbb{P}\Big(\Big|\int_{\mathbb{R}}f(\lambda)(\hat{\mu}_{H_N^C}-\mathbb{E}[\hat{\mu}_{H_N^C}])(\mathrm{d}\lambda)\Big|\geqslant\delta\Big)=\mathbb{P}(|h(Z)-\mathbb{E}[h(Z)]|\geqslant\delta)\leqslant 2\exp\Big(-\frac{\delta^2N^{2-2\gamma}}{2}\Big)$$

这就证明了（3.13）。

现在，我们希望通过在概率内部对 f 取上确界来加强结果——起初只在 $\mathcal{F}_{\mathrm{Lip},\mathcal{K}}$ 上取上确界；回顾一下，它表示支撑在某个紧集$\mathcal{K}$中的$\mathcal{F}_{\mathrm{Lip}}$ 函数集合。Guionnet 和 Zeitouni 为此给出了一种非常有用的构造（我们在下文次高斯情形的证明中也会仿照这一做法）：对于任意

$\Delta > 0$，他们构造了一个由 $2(\lceil \frac{|\mathcal{K}|}{\Delta} \rceil \pm 1)$ 个函数 $(h_k)_{2(\lceil \frac{|\mathcal{K}_\Delta|}{\Delta} \rceil + 1)k=1}$ 组成的集合，它们属于$\mathcal{F}_{\mathrm{Lip}}$，并具有如下性质：对于任意给定的$f \in \mathcal{F}_{\mathrm{Lip},\mathcal{K}}$，都可以从这些函数中选出$\lceil \frac{|\mathcal{K}|}{\Delta} \rceil + 1$个，使其和（记为 f_Δ）满足$\|f - f_\Delta\|_\infty \leqslant \Delta$。由于 f_Δ 实际上是这些有限个函数之和，而非它们的线性组合，因此我们有

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f \in \mathcal{F}_{\mathrm{Lip},\mathcal{K}}} \left| \int_{\mathbb{R}} f(\lambda)(\hat{\mu}_{H_N^C} - \mathbb{E}[\hat{\mu}_{H_N^C}])(\mathrm{d}\lambda) \right| \geqslant \delta \right) \\
&\leqslant 2\left(\left\lceil \frac{|\mathcal{K}|}{\Delta} \right\rceil + 1 \right)^{2\left(\lceil \frac{|\mathcal{K}|}{\Delta} \rceil + 1\right)} \max_{k=1} \mathbb{P}\left(\left| \int_{\mathbb{R}} h_k(\lambda)(\hat{\mu}_{H_N^C} - \mathbb{E}[\hat{\mu}_{H_N^C}])(\mathrm{d}\lambda) \right| \geqslant \frac{\delta - 2\Delta}{2(\lceil |\mathcal{K}|/\Delta \rceil + 1)} \right) \\
&\leqslant 4\left(\left\lceil \frac{|\mathcal{K}|}{\Delta} \right\rceil + 1 \right) \exp\left(-\frac{N^{2-2\gamma}}{2} \left(\frac{\delta - 2\Delta}{2(\lceil |\mathcal{K}|/\Delta \rceil + 1)} \right)^2 \right)
\end{aligned}$$

Like Guionnet 和 Zeitouni，我们选择 $\Delta = \delta/4$；那么每当 $\delta < 1$ and $|\mathcal{K}| > 1$时，我们有 e

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}_{\mathrm{Lip},\mathcal{K}}} \left| \int_{\mathbb{R}} f(\lambda)(\hat{\mu}_{H_N^C} - \mathbb{E}[\hat{\mu}_{H_N^C}])(\mathrm{d}\lambda) \right| \geqslant \delta \right) \leqslant \frac{32\delta}{|\mathcal{K}|} \exp\left(-\frac{N^{2-2\gamma}\delta^4}{128|\mathcal{K}|^2} \right). \tag{3.14}$$

现在，对于 $L > 0$，我们定义事件

$$\mathcal{E}_{C,L} = \{\mathrm{supp}(\hat{\mu}_{H_N^C}) \subset (-L, L)\},$$

仿照上面的事件 $\mathcal{E}_L$ 。与证明$\mathcal{E}_L$ 在（3.5）中很可能发生的方法相同，可以证明

$$\lim_{L\to\infty} \limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}(\mathcal{E}_{C,L}^c) = -\infty. \tag{3.15}$$

~另一方面，固定一个较大的 L。对于任意 $f \in \mathcal{F}_{\mathrm{Lip}}$，存在 $f \in \mathcal{F}_{\mathrm{Lip}}$，它在 $(-L, L)$ 上与 f 一致，并且在 $(-2L, 2L)$ 外为零。比如说。此外，在事件 $\mathcal{E}_{C,L}$ 上，经验测度 ^$\mu_{H_N^C}$ 的支撑集位于 $(-L, L)$ 内，尽管其期望并非如此；因此令 $\mathcal{K} := (-2L, 2L)$，我们有

$$\begin{aligned}
&\mathbb{P}\left(d_{\mathrm{BL}}(\hat{\mu}_{H_N^C}, \mathbb{E}[\hat{\mu}_{H_N^C}]) \geqslant \delta, \mathcal{E}_{C,L} \right) \\
&= \mathbb{P}\left(\sup_{f \in \mathcal{F}_{\mathrm{Lip}}} \left| \int_{\mathbb{R}} f(\lambda)(\hat{\mu}_{H_N^C} - \mathbb{E}[\hat{\mu}_{H_N^C}])(\mathrm{d}\lambda) \right| \geqslant \delta, \mathcal{E}_{C,L} \right) \\
&\leqslant \mathbb{P}\left(\sup_{\widetilde{f} \in \mathcal{F}_{\mathrm{Lip},\mathcal{K}}} \left| \int_{\mathbb{R}} \widetilde{f}(\lambda)(\hat{\mu}_{H_N^C} - \mathbb{E}[\hat{\mu}_{H_N^C}])(\mathrm{d}\lambda) \right| \geqslant \frac{\delta}{2}, \mathcal{E}_{C,L} \right) + \mathbb{P}\left(\sup_{f \in \mathcal{F}_{\mathrm{Lip}}} \left| \int_{\mathbb{R}} (f(\lambda) - \widetilde{f}(\lambda)) \mathbb{E}[\hat{\mu}_{H_N^C}](\mathrm{d}\lambda) \right| \geqslant \frac{\delta}{2} \right) \\
&\leqslant \mathbb{P}\left(\sup_{\widetilde{f} \in \mathcal{F}_{\mathrm{Lip},\mathcal{K}}} \left| \int_{\mathbb{R}} \widetilde{f}(\lambda)(\hat{\mu}_{H_N^C} - \mathbb{E}[\hat{\mu}_{H_N^C}])(\mathrm{d}\lambda) \right| \geqslant \frac{\delta}{2} \right) + \mathbf{1}\left\{ \int_{\mathbb{R}} \mathbf{1}_{|\lambda| \geqslant L} \mathbb{E}[\hat{\mu}_{H_N^C}](\mathrm{d}\lambda) \geqslant \frac{\delta}{4} \right\} \\
&\leqslant \frac{8\delta}{L} \exp\left(-\frac{N^{2-2\gamma}\delta^4}{2048L^2} \right) + \mathbf{1}\left\{ \int_{\mathbb{R}} \mathbf{1}_{|\lambda| \geqslant L} \mathbb{E}[\hat{\mu}_{H_N^C}](\mathrm{d}\lambda) \geqslant \frac{\delta}{4} \right\}
\end{aligned}$$

最后一行用到了（3.14）。为了处理指示函数，我们注意到

$$\int_L^\infty \mathbb{E}[\hat{\mu}_{H_N^C}](\mathrm{d}\lambda) \leqslant \mathbb{P}(\lambda_{\max}(H_N^C) \geqslant L) \leqslant \mathbb{P}(\mathcal{E}_{C,N}^c)$$

左尾的情形类似，（3.15）表明，当 L 足够大（其大小取决于 δ）时，指示函数为零；因此

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}_{\mathrm{Lip}}} \left| \int_{\mathbb{R}} f(\lambda)(\hat{\mu}_{H_N^C} - \mathbb{E}[\hat{\mu}_{H_N^C}])(\mathrm{d}\lambda) \right| \geqslant \delta, \mathcal{E}_{C,L} \right) \leqslant \frac{8\delta}{L} \exp\left(-\frac{N^{2-2\gamma}\delta^4}{2048L^2} \right) \quad \text{for all } L > L_0(\delta).$$

合并 根据（3.15），这就完成了（3.11）的证明，从而完成了引理3.1 4. □

引理 3.15。在假设 A下，对每个 $\varepsilon>0$，存在 $\gamma=\gamma(\varepsilon)>0$ 使得，若 H_N^A 根据 γ 使用 (3.3)定义，则

$$\lim_{N\to\infty}\frac{1}{N}\log\mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N^A},\mathbb{E}[\hat{\mu}_{H_N^A}])>\varepsilon)=-\infty.$$

证明。这个证明本质上是对 Guionnet 和 Zeitouni [GZ00], 所作评论的补充练习，需要用到他们的技巧。我们需要几类测试函数 $f:\ \mathbb{R}\to\mathbb{R}$:

$$\mathcal{F}_{\mathrm{Lip}}=\left\{f:\|f\|_{\mathcal{L}}:=\|f\|_\infty+\sup_{x\neq y}\frac{|f(x)-f(y)|}{|x-y|}\leqslant 1\right\},$$

对紧集 $\mathcal{K}\subset\mathbb{R}$，$\mathcal{F}_{\mathrm{Lip},\ \mathcal{K}}=\{f\in\mathcal{F}_{\mathrm{Lip}}:\mathrm{supp}(f)\subset\mathcal{K}\}$,

$$\mathfrak{C}_1=\{f:\text{函数 }\ x\mapsto g(x):=f(x^2)\text{ 是 Lipschitz 且凸的}\}$$

以及相应的范数

$$\|f\|_{\mathfrak{C}_1}=\sup_{x\neq y}\left|\frac{f(x^2)-f(y^2)}{x-y}\right|.$$

Guionnet 和 Zeitouni [GZ00，推论 1。8] 证明了，对所有 $f\in\mathfrak{C}_1$,

$$\mathbb{P}\left(\left|\int_{\mathbb{R}}f(\lambda)(\hat{\mu}_{H_N^A}-\mathbb{E}[\hat{\mu}_{H_N^A}])(\mathrm{d}\lambda)\right|\geqslant\delta\frac{N+M}{N}\right)\leqslant 4\exp\left(-\frac{1}{16d_{\max}N^{2\gamma}\|f\|_{\mathfrak{C}_1}^2}(\delta-\delta_0(N+M))^2(N+M)^2\right)$$

其中 $\delta_0(N+M)=\frac{8\sqrt{\pi d_{\max}N^\gamma}}{\|f\|_{N+M}}\|f\|_{\mathfrak{C}_1}$。

现在固定紧集$\mathcal{K}\subset\mathbb{R}$和 $\delta>0$，并记 $k_{\min}=\min(0,\min\{x:\ x\in\mathcal{K}\})$、 $k_{\max}=\max(0,\max\{x:\ x\in\mathcal{K}\})$以及 $\Delta=\frac{\delta}{8}$ (假设 $\Delta<1$)。我们将用 $\mathfrak{C}_1$中的测试函数的有限线性组合来逼近$\mathcal{F}_{\mathrm{Lip},\mathcal{K}}$ 中的测试函数。

为此，对任意 $a\geqslant 0$，考虑函数

$$g_a(x)=\begin{cases}0 & \text{if } x\leqslant a,\\ \frac{\Delta}{\sqrt{a+\Delta}-\sqrt{a}}(\sqrt{x}-\sqrt{a}) & \text{if } a\leqslant x\leqslant a+\Delta,\\ \Delta & \text{if } x\geqslant a+\Delta.\end{cases}$$

注意， g_a 是 $\mathfrak{C}_1$中两个函数之差，即 $g_a=g_{a,1}-g_{a,2}$，其中

$$g_{a,1}(x)=\begin{cases}0 & \text{if } x\leqslant a,\\ \frac{\Delta}{\sqrt{a+\Delta}-\sqrt{a}}(\sqrt{x}-\sqrt{a}) & \text{if } x\geqslant a,\end{cases}\qquad g_{a,2}(x)=\begin{cases}0 & \text{if } x\leqslant a+\Delta,\\ \frac{\Delta}{\sqrt{a+\Delta}-\sqrt{a}}(\sqrt{x}-\sqrt{a+\Delta}) & \text{if } x\geqslant a+\Delta,\end{cases}$$

它们满足 $\|g_{a,1}\|_{\mathfrak{C}_1}=\|g_{a,2}\|_{\mathfrak{C}_1}=\frac{\Delta}{\sqrt{a+\Delta}-\sqrt{a}}\leqslant$

现在令 $f\in\mathcal{F}_{\mathrm{Lip},\mathcal{K}}$。首先注意，由于作为协方差矩阵的经验测度，$\hat{\mu}_{H_N^A}$ 和 $\mathbb{E}[\hat{\mu}_{H_N^A}]$ 的支撑集位于右半轴上，若定义

$$\widetilde{f}(x)=(f(x)-f(0))\mathbf{1}\{x\geqslant 0\},$$

then

$$\int_{\mathbb{R}}f(\lambda)(\hat{\mu}_{H_N^A}-\mathbb{E}[\hat{\mu}_{H_N^A}])(\mathrm{d}\lambda)=\int_{\mathbb{R}}\widetilde{f}(\lambda)(\hat{\mu}_{H_N^A}-\mathbb{E}[\hat{\mu}_{H_N^A}])(\mathrm{d}\lambda).$$

此外， f 的支撑集位于右半直线上，且是 1-Lipschitz 函数；虽然通常并非紧支撑，但对 $x>k_{\max}$仍为常数。实际上，我们不会直接逼近 f，而是逼近 f。我们的逼近函数将为

$$f_\Delta(x)=\sum_{i=0}^{\lceil|\mathcal{K}|/\Delta\rceil}c_i^{(\Delta)}g_{k_{\min}+i\Delta}(x),$$

其中，系数 $c_i^{(\Delta)}$ 取值于 $\{-1,\ +1\}$，并按以下方式递归定义 :

•

$$c_0^{(\Delta)} = \begin{cases} 1 & \text{if } \widetilde{f}(k_{\min}+\Delta) \geqslant 0, \\ -1 & \text{otherwise.} \end{cases}$$

• 给定 $c_{0,\,\Delta}, \ldots, c_{j,\,\Delta}$，令

$$c_{j+1}^{(\Delta)} = \begin{cases} 1 & \text{if } \widetilde{f}(k_{\min}+(j+2)\Delta) \geqslant \Delta \sum_{i=0}^{j} c_i^{(\Delta)} \\ -1 & \text{otherwise.} \end{cases}$$

注意，在每个区间上

$$I_j = [k_{\min}+j\Delta, k_{\min}+(j+1)\Delta],$$

定义 f_Δ 的和中，只有 $i=j$ 这一项非平凡；较小下标对应的项要么是 Δ，要么是 $-\Delta$，而较大下标对应的项均为零。

现在我们断言

$$\|\widetilde{f} - f_\Delta\|_\infty \leqslant 2\Delta. \tag{3.16}$$

In ~事实上，f 和 f_Δ 在 $(-\infty,\ k_{\min})$ 上均为零，我们将对 j 使用归纳法证明

$$A_j = \sup_{x\in I_j}\left|\widetilde{f}(x) - f_\Delta(x)\right| \leqslant 2\Delta \quad \text{and} \quad B_j = \left|\widetilde{f}(k_{\min}+(j+1)\Delta) - f_\Delta(k_{\min}+(j+1)\Delta)\right| \leqslant \Delta$$

对于 $j=0, \ldots, \lceil 2|\mathcal{K}|/\Delta\rceil$。在证明中，我们将经常用到以下观察

$$f_\Delta(k_{\min}+j\Delta) = \Delta\sum_{i=0}^{j-1} c_i^{(\Delta)}.$$

• 当 $j=0$ 时，情形很简单，因为对于 $x\in I_0$，我们有 $\left|\widetilde{f}(x)\right| \leqslant \Delta$（因为 $\widetilde{f}(k_{\text{最小值}})=0$ 且 $\widetilde{f}$ 是 1-Lipschitz），并且有 $|f_\Delta(x)| = \left|c_{(\Delta)\,0}\ \ g_{k_{\text{最小值}}}(x)\right| \leqslant \Delta$；由此得到 $A_0 \leqslant 2\Delta$，而 $B_0 \leqslant \Delta$ 则来自 $c_{(\Delta)\,0}$ 的定义以及 f 的 1-Lipschitz ~性质。

• 现在假设，对于 $i=0,\ldots,j-1$，有 $A_i \leqslant 2\Delta$ 和 $B_i \leqslant \Delta$。为方便起见，若 $j\geqslant 2$，令 $c_{(\Delta)} = \sum_{i=0}^{j-2} c_{(\Delta)\,i}$；否则，令 $c_{(\Delta)}=0$，其中 $j=1$。注意，无论是 $j=1$ 还是 $j\geqslant 2$，我们都有 $f_\Delta(k_{\min}+j\Delta) = (c_{(\Delta)} + c_{(\Delta)\,j-1})\Delta$；并且在 I_j 上有 $f_\Delta(x) = (c_{(\Delta)} + c_{(\Delta)\,j-1})\Delta + c_{(\Delta)\,j}\ \ g_{k_{\min}+j\Delta}(x)$。分别用 $\ell_j = k_{\min}+j\Delta$ 和 $r_j = k_{\min}+(j+1)\Delta$ 表示区间 I_j 的左端点和右端点会很有用。共有两种情况：

– **情形 1（$c_{j-1}^{(\Delta)}$ 和 $c_j^{(\Delta)}$ 相同）：** 假设 $c_{j-1}^{(\Delta)} = c_j^{(\Delta)} = 1$，另一种情形的证明类似。由此得到 $f_\Delta(\ell_j) = (c^{(\Delta)}+1)\Delta$ 以及 $f_\Delta(r_j) = (c^{(\Delta)}+2)\Delta$。

~假设我们能够证明 f 位于端点处的“正确窗口”内，也就是说，$f(\ell_j) \in [f_\Delta(\ell_j) - \Delta,\ f_\Delta(\ell_j)]$ 以及 $f(r_j) \in [f_\Delta(r_j) - \Delta,\ f_\Delta(r_j)]$。这将立即证明 $B_j \leqslant \Delta$。此外，由于 ~f 是 1-Lipschitz，这说明 f 位于某个连接这些窗口的“管状区域”内，其“高度”为 Δ；将其与 f_Δ 在 I_j 上的显式公式进行比较（它恰好位于该管状区域上方），我们还可以得到 $A_j \leqslant 2\Delta$。

因此，我们只需在端点处理解 $\widetilde{f}$。但根据 $c_{j-1}^{(\Delta)}$ 的定义，我们得到 $\widetilde{f}(\ell_j) \geqslant c^{(\Delta)}\Delta = f_\Delta(\ell_j) - \Delta$。由于 $B_j \leqslant \Delta$，这意味着 $\widetilde{f}(\ell_j) \leqslant f_\Delta(\ell_j)$。由于 $\widetilde{f}$ 是 1-Lipschitz 的，这意味着 $\widetilde{f}(r_j) \leqslant \widetilde{f}(\ell_j) + \Delta \leqslant f_\Delta(\ell_j) + \Delta = f_\Delta(r_j)$。接下来，根据 $c_j^{(\Delta)}$ 的定义，我们得到 $\widetilde{f}(r_j) \geqslant (c^{(\Delta)}+1)\Delta = f_\Delta(r_j) - \Delta$，从而完成了将 $\widetilde{f}$ 定位在 $x=\ell_j$ 和 $x=r_j$ 处正确窗口中的工作。

– **情形 2（$c_{j-1}^{(\Delta)}$ 和 $c_j^{(\Delta)}$ 不相同）：** 假设 $c_{j-1}^{(\Delta)} = 1$ 和 $c_j^{(\Delta)} = -1$，另一种情形类似。于是有 $f_\Delta(\ell_j) = (c^{(\Delta)}+1)\Delta$ 以及 $f_\Delta(r_j) = c^{(\Delta)}\Delta$。

~与第一种情形类似的论证，将 f 局限在端点处略有不同的窗口中：这里我们发现 ~$f(\ell_j) \in [f_\Delta(\ell_j) - \Delta,\ f_\Delta(\ell_j)]$ 以及 $f(r_j) \in [f_\Delta(r_j) - \Delta,\ f_\Delta(r_j) + \Delta]$。这立即表明 $B_j \leqslant \Delta$。

连接这些窗口的所有可能的 1-Lipschitz 函数路径现在呈现五边形而非“管状”的形状（如果 $f(\ell_j) = f_\Delta(\ell_j)$，那么 f 可以先在 $[\ell_j, \ell_j + \Delta/2]$ 上增大，再回落）；不过，无论如何，通过将其与 I_j 上的 f_Δ 的显式公式进行比较，仍可推出 $A_j \leqslant 2\Delta$。

这表明 $\sup_{x \leqslant k_{\min} + \Delta \lceil |\mathcal{K}|/\Delta \rceil} \left| \widetilde{f}(x) - f_\Delta(x) \right| \leqslant 2\Delta$。如果 $x \geqslant k_{\min} + \Delta \lceil |\mathcal{K}|/\Delta \rceil \geqslant k_{\min} + |\mathcal{K}|$，那么 $\widetilde{f}(x) = -f(0)$，且 $|f_\Delta(x) + f(0)| = |f_\Delta(k_{\min} + \Delta \lceil |\mathcal{K}|/\Delta \rceil) + f(0)| = \left| f_\Delta(k_{\min} + \Delta \lceil |\mathcal{K}|/\Delta \rceil) - \widetilde{f}(k_{\min} + \Delta \lceil |\mathcal{K}|/\Delta \rceil) \right| \leqslant B_{\lceil |\mathcal{K}|/\Delta \rceil} \leqslant \Delta$

根据我们上文的工作。这就完成了对（3.16）的证明。

无论 f 如何，近似项 f_Δ 都是至多由 $2(|\mathcal{K}|/\Delta + 1)$ 个 $\mathfrak{C}_1$ 中的元素 h_k 线性组合而成，且每个元素的 $\mathfrak{C}_1$-范数至多为 $\frac{\Delta}{\sqrt{k_{\max}+\Delta}-\sqrt{k_{\max}}}$。因此

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{f \in \mathcal{F}_{\mathrm{Lip},\mathcal{K}}} \left| \int f(\lambda)(\hat{\mu}_{H_N^A} - \mathbb{E}[\hat{\mu}_{H_N^A}])(\mathrm{d}\lambda) \right| \geqslant \delta \right) \\
&\leqslant 2\left(\frac{|\mathcal{K}|}{\Delta} + 1 \right) \sup_k \mathbb{P}\left(\left| \int h_k(\lambda)(\hat{\mu}_{H_N^A} - \mathbb{E}[\hat{\mu}_{H_N^A}])(\mathrm{d}\lambda) \right| \geqslant \frac{\delta - 4\Delta}{2(|\mathcal{K}|/\Delta + 1)} \right) \\
&\leqslant 8\left(\frac{|\mathcal{K}|}{\Delta} + 1 \right) \exp\left(-\frac{1}{16 d_{\max} N^{2\gamma} \frac{\Delta^2}{(\sqrt{k_{\max}+\Delta}-\sqrt{k_{\max}})^2}} \left(\frac{\delta - 4\Delta}{2(|\mathcal{K}|/\Delta + 1)} \frac{N}{N+M} - \delta_0(N+M) \right)^2 (N+M)^2 \right)
\end{aligned}
$$

其中 $\delta_0(N+M) = \frac{8\sqrt{\pi d_{\max}} N^\gamma \frac{\Delta}{\sqrt{k_{\max}+\Delta}-\sqrt{k_{\max}}}}{N+M}$。代入 $\Delta = \delta/8$ 得

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{f \in \mathcal{F}_{\mathrm{Lip},\mathcal{K}}} \left| \int f(\lambda)(\hat{\mu}_{H_N^A} - \mathbb{E}[\hat{\mu}_{H_N^A}])(\mathrm{d}\lambda) \right| \geqslant \delta \right) \\
&\leqslant 8\left(\frac{8|\mathcal{K}|}{\delta} + 1 \right) \exp\left(-\frac{4(\sqrt{k_{\max} + \delta/8} - \sqrt{k_{\max}})^2}{d_{\max} N^{2\gamma} \delta} \left(\frac{\delta}{4(8|\mathcal{K}|/\delta + 1)} \frac{N}{N+M} - \delta_0(N+M) \right)^2 (N+M)^2 \right).
\end{aligned}
$$

现在，我们用某个较大阈值 L 下的紧支撑 Lipschitz 函数 f^L 来近似一般的 $f \in \mathcal{F}_{\mathrm{Lip}}$。由于

$$
\frac{1}{N} \operatorname{tr}(H_N^A) = \frac{1}{NM} \sum_{i=1}^{N} \sum_{k=1}^{M} A_{ki}^2 d_k \leqslant d_{\max} N^{2\gamma},
$$

对于待选取的 L，我们有 $\int_{|\lambda| > L} \hat{\mu}_{H_N^A}(\mathrm{d}\lambda) \leqslant \frac{d_{\max} N^{2\gamma}}{L}$。若定义

$$
f^L(x) = \begin{cases} f(x) & \text{if } 0 \leqslant x \leqslant L, \\ f(L) - \operatorname{sign}(f(L))(x - L) & \text{if } L \leqslant x \leqslant L + |f(L)| \\ f(0) + \operatorname{sign}(f(0)) x & \text{if } -|f(0)| \leqslant x \leqslant 0 \\ 0 & \text{otherwise} \end{cases}
$$

then

$$
\left| \int f(\lambda)(\hat{\mu}_{H_N^A} - \mathbb{E}[\hat{\mu}_{H_N^A}])(\mathrm{d}\lambda) \right| \leqslant \left| \int f^L(\lambda)(\hat{\mu}_{H_N^A} - \mathbb{E}[\hat{\mu}_{H_N^A}])(\mathrm{d}\lambda) \right| + \frac{2 d_{\max} N^{2\gamma}}{L}.
$$

如果我们取引理陈述中的 $L = \frac{4 d_{\max} N^{2\gamma}}{\varepsilon}$ 和 ε，那么由于 $f^L \in \mathcal{F}_{\mathrm{Lip},\,\mathcal{K}}$，粗略地有 $\mathcal{K} = [-1, 2L]$，因此

$$
\begin{aligned}
&\mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{H_N^A}, \mathbb{E}[\hat{\mu}_{H_N^A}]) > \varepsilon) = \mathbb{P}\left(\sup_{f \in \mathcal{F}_{\mathrm{Lip}}} \left| \int f(\lambda)(\hat{\mu}_{H_N^A} - \mathbb{E}[\hat{\mu}_{H_N^A}])(\mathrm{d}\lambda) \right| \geqslant \varepsilon \right) \\
&\leqslant \mathbb{P}\left(\sup_{f \in \mathcal{F}_{\mathrm{Lip},\mathcal{K}}} \left| \int f(\lambda)(\hat{\mu}_{H_N^A} - \mathbb{E}[\hat{\mu}_{H_N^A}])(\mathrm{d}\lambda) \right| \geqslant \frac{\varepsilon}{2} \right) \\
&\leqslant 8\left(\frac{16|\mathcal{K}|}{\varepsilon} + 1 \right) \exp\left(-\frac{8(\sqrt{k_{\max} + \varepsilon/16} - \sqrt{k_{\max}})^2}{d_{\max} N^{2\gamma} \varepsilon} \left(\frac{\varepsilon}{8(8|\mathcal{K}|/\varepsilon + 1)} \frac{N}{N+M} - \delta_0(N+M) \right)^2 (N+M)^2 \right).
\end{aligned}
$$

只需令 γ 足够小即可。 □

3.4 弱大偏差原理上界的证明

定义 3.16. 考虑以下确定性的 $M \times N$ 实矩阵集合:

$$\mathcal{A}^L_{x,\delta,\varepsilon} = \left\{ Z : \ with\ H_N = H_N(Z) = \frac{1}{M} Z^T \Gamma Z, \quad |\lambda_{\max}(H_N) - x| < \delta, \quad d_{\mathrm{BL}}(\hat{\mu}_{H_N}, \sigma) < \varepsilon, \ and\ \|H_N\| \leqslant L \right\}.$$

引理 3.17。 对于每个 $x \geqslant r(\sigma)$ 和 $0 \leqslant \theta < \theta_{\max}$，以及所有充分大的 L，我们有

$$\begin{aligned} &\lim_{\varepsilon \downarrow 0} \limsup_{\delta \downarrow 0} \limsup_{N \to \infty} \sup_{Z \in \mathcal{A}^L_{x,\delta,\varepsilon}} \left| \frac{1}{N} \log \mathbb{E}_e[e^{N \frac{\theta}{2} \langle e, H_N e \rangle}] - J\left(\sigma, \frac{\theta}{2}, x\right) \right| \\ &= \lim_{\delta \downarrow 0} \limsup_{\varepsilon \downarrow 0} \limsup_{N \to \infty} \sup_{Z \in \mathcal{A}^L_{x,\delta,\varepsilon}} \left| \frac{1}{N} \log \mathbb{E}_e[e^{N \frac{\theta}{2} \langle e, H_N e \rangle}] - J\left(\sigma, \frac{\theta}{2}, x\right) \right| = 0. \end{aligned}$$

证明。 这是 [GH22]（以及更强结果）的一个直接推论。固定（较小的） $t > 0$ 和（较大的） $L > 0$；其定理 6.2 表明存在 $N_0(t, L)$ 使得对于 $N \geqslant N_0(t, L)$ 我们有

$$\sup_{Z \in \mathcal{A}^L_{x,\delta,\varepsilon}} \left| \frac{1}{N} \log \mathbb{E}[e^{N \frac{\theta}{2} \langle e, H_N e \rangle}] - J\left(\hat{\mu}_{H_N}, \frac{\theta}{2}, x\right) \right| \leqslant t.$$

另一方面，其定理 21 表明， $J(\mu, \frac{\theta}{2}, x)$ 是定义在 $[-L, L]$ 上的紧支撑概率测度集合中关于 μ 的连续函数；因此，对于 $\varepsilon \leqslant \varepsilon_0(t)$，我们有

$$\sup_{Z \in \mathcal{A}^L_{x,\delta,\varepsilon}} \left| J\left(\hat{\mu}_{H_N}, \frac{\theta}{2}, x\right) - J\left(\sigma, \frac{\theta}{2}, x\right) \right| \leqslant t.$$

我们按正确的顺序结合这两个估计及其量词，然后取 $t \downarrow 0$。 □

引理 3.18. 对于每个 $x \geqslant r(\sigma)$、 $0 \leqslant \theta < \theta_{\max}$，以及充分大的 L，有

$$\limsup_{\delta \downarrow 0} \limsup_{\varepsilon \downarrow 0} \limsup_{N \to \infty} \frac{1}{N} \log \mathbb{P}^\theta(\mathcal{A}^L_{x,\delta,\varepsilon}) \leqslant -(\widetilde{I}_\sigma(x) - I_\sigma(x, \theta)).$$

证明。 对于任意 $0 \leqslant \theta' < \theta_{\max}$,

$$\begin{aligned} \mathbb{P}^\theta(\mathcal{A}^L_{x,\delta,\varepsilon}) &= \frac{1}{\mathbb{E}_{e,H_N}[e^{N \frac{\theta}{2} \langle e, H_N e \rangle}]} \mathbb{E}_{H_N} \left[\mathbb{1}_{H_N \in \mathcal{A}^L_{x,\delta,\varepsilon}} \mathbb{E}_e[e^{N \frac{\theta}{2} \langle e, H_N e \rangle}] \frac{\mathbb{E}_e[e^{N \frac{\theta'}{2} \langle e, H_N e \rangle}]}{\mathbb{E}_e[e^{N \frac{\theta'}{2} \langle e, H_N e \rangle}]} \right] \\ &\leqslant \frac{\mathbb{E}_{e,H_N}[e^{N \frac{\theta'}{2} \langle e, H_N e \rangle}]}{\mathbb{E}_{e,H_N}[e^{N \frac{\theta}{2} \langle e, H_N e \rangle}]} \left(\sup_{Z \in \mathcal{A}^L_{x,\delta,\varepsilon}} \mathbb{E}_e[e^{N \frac{\theta}{2} \langle e, H_N e \rangle}] \right) \left(\sup_{Z \in \mathcal{A}^L_{x,\delta,\varepsilon}} \frac{1}{\mathbb{E}_e[e^{N \frac{\theta'}{2} \langle e, H_N e \rangle}]} \right) \end{aligned}$$

由引理 3.11 和 3.17，可得

$$\limsup_{\delta \downarrow 0} \limsup_{\varepsilon \downarrow 0} \limsup_{N \to \infty} \frac{1}{N} \log \mathbb{P}^\theta(\mathcal{A}^L_{x,\delta,\varepsilon}) \leqslant I_\sigma(x, \theta) - I_\sigma(x, \theta').$$

对右端关于 $0 \leqslant \theta' < \theta_{\max}$ 取下确界，即完成证明。 □

引理 3.19. 设 $A_{N,M}$ 是一个双指标事件序列，且 $\lim_{M \to \infty} \lim_{N \to \infty} \frac{1}{N} \log \mathbb{P}(A_{N,M}) = -\infty$。那么对于每个 $\theta < \theta_{\max}$，有

$$\lim_{M \to \infty} \lim_{N \to \infty} \frac{1}{N} \log \mathbb{P}^\theta(A_{N,M}) = -\infty.$$

特别地（取$A_{N,M}$与 M无关），若$\lim_{N \to \infty} \frac{1}{N} \log \mathbb{P}(A_N) = -\infty$,则$\lim_{N \to \infty} \frac{1}{N} \log \mathbb{P}^\theta(A_N) = -\infty$。

命题 3.8 的证明。若 $x < r(\sigma)$，则对足够小的 δ，我们有$\{|\lambda_{\max}(H_N) - x| \leqslant \delta\} \subset \{d_{\mathrm{BL}}(\hat{\mu}_{H_N}, \sigma) > \varepsilon\}$，其中某个 $\varepsilon = \varepsilon(\delta)$；由引理 3.19 和命题 3.12 可知这已足够。下文中我们假设 $x \geqslant r(\sigma)$。当 L足够大时，我们有

$$\mathbb{P}^\theta(|\lambda_{\max}(H_N) - x| \leqslant \delta) \leqslant \mathbb{P}^\theta(\mathcal{A}^L_{x,\delta,\varepsilon}) + \mathbb{P}^\theta(d_{\mathrm{BL}}(\hat{\mu}_{H_N}, \sigma) > \varepsilon) + \mathbb{P}^\theta(\|H_N\| > L).$$

然后，我们对两边取归一化对数；取 $N \to +\infty$，并再次应用引理3.19 和命题3.12，得到

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}^\theta(|\lambda_{\max}(H_N) - x| \leqslant \delta) \leqslant \max\left\{ \limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}^\theta(\mathcal{A}^L_{x,\delta,\varepsilon}), \limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}^\theta(\|H_N\| > L) \right\}.$$

然后取 $\varepsilon \downarrow 0$，再取 $\delta \downarrow 0$，并应用引理3.18 ，得到

$$\limsup_{\delta\downarrow 0} \limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}^\theta(|\lambda_{\max}(H_N) - x| \leqslant \delta) \leqslant \max\left\{ -(\widetilde{I}_\sigma(x) - I_\sigma(x,\theta)), \limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}^\theta(\|H_N\| > L) \right\}.$$

取 $L \to \infty$ 并应用引理3.19 和3.6，证明即告完成。 □

引理*3.19*的证明。 对于每个单位向量 e，引理3.11 的证明表明

$$\mathbb{E}_{H_N}[e^{N\frac{\theta}{2}\langle e, H_N e\rangle}] = \prod_{\mu=1}^{M} \mathbb{E}[e^{\frac{N}{M}\frac{\theta}{2} d_\mu \langle z_\mu, e\rangle^2}] \geqslant 1,$$

因此，若 A 是一个事件且 $\varepsilon > 0$，则有

$$\mathbb{P}^\theta(A) \leqslant \mathbb{E}_{H_N}[\mathbf{1}_A \mathbb{E}_e[e^{N\frac{\theta}{2}\langle e, H_N e\rangle}]] \leqslant \mathbb{P}(A)^{\frac{\varepsilon}{1+\varepsilon}} \mathbb{E}_{e,H_N}[e^{N\frac{(1+\varepsilon)\theta}{2}\langle e, H_N e\rangle}]^{\frac{1}{1+\varepsilon}}$$

若 ε 足够小，使得$(1+\varepsilon)\theta < \theta_{\max}$，则应用引理3.11 即可完成证明。 □

3.5 弱大偏差原理下界的证明

命题 3.20。 假设 (ρ, α) 是非退化的，并设 D° 为集合 D 的内部，该集合定义于 (2.5)中。若$r(\rho) > 0$，则还需假设 $G_\rho(r(\rho)) = +\infty$。那么，对每个 $x \in D^\circ$，都存在唯一的 θ_x 满足 $0 \leqslant \theta_x < \theta_{\max}$，使得

$$\widetilde{I}_\sigma(x) = \sup_{0\leqslant\theta<\theta_{\max}} I_\sigma(x,\theta) = I_\sigma(x,\theta_x).$$

映射 $x \mapsto \theta_x$ 是单射。

证明。 这是引理 3.3 证明的推论。特别地，$\theta_x = G_\sigma(x)$。 □

引理 3.21。 设 $x \geqslant r(\sigma)$，并设 θ_x 如命题 *3.20*中所定义。那么，对所有充分大的 L 以及 δ，ε只要足够小（取决于 x），就有

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}^{\theta_x}(\mathcal{A}^L_{x,\delta,\varepsilon}) = 0.$$

证明。 我们断言，实际上 $\mathbb{P}^{\theta_x}(\mathcal{A}^L_{x,\delta,\varepsilon})$ 趋于 1。正如命题 3.8 的证明所示，对所有 $0 \leqslant \theta < \theta_{\max}$，量 $\mathbb{P}^\theta(d_{\mathrm{BL}}(\hat{\mu}_{H_N}, \sigma) \geqslant \varepsilon)$ 和 $\mathbb{P}^\theta(\|H_N\| \geqslant L)$ 都趋于零（实际上以指数速度趋于零），当 L 足够大且所有 $\varepsilon > 0$ 均满足条件时。因此，只需证明

$$\mathbb{P}^{\theta_x}(|\lambda_{\max}(H_N) - x| \geqslant \delta) = \mathrm{o}(1). \tag{3.17}$$

但是，命题 3.8 给出了变量 $\lambda_{\max}(H_N)$ 在测度 $\mathbb{P}^{\theta_x}$ 下的弱大偏差原理上界，该变量具有~指数紧性，其速率函数 $J_x(y)$ 在 $y < r(\sigma)$ 处为无穷大，否则等于 $J_x(y) = I_\sigma(y) - I_\sigma(y,\theta_x)$。命题 3.20 表明，$J_x$ 非负，且仅在 x 处取零；事实上，若 $r(\sigma) \leqslant y \neq x$，则$J_x(y) = \sup_{0\leqslant\theta<\theta_{\max}} I(y,\theta) - I(x,\theta_x)$，且上确界仅在 θ_y 处取得，而该点不同于 θ_x，因为$y \neq x$。结合 $\mathbb{P}^{\theta_x}$ 的指数紧性（见引理 3.6 和引理 3.19），即可得到 (3.17)。 □

命题3.9的证明。 我们有

$$\mathbb{P}(|\lambda_{\max}(H_N)-x|\leqslant\delta)\geqslant\mathbb{P}(\mathcal{A}^L_{x,\delta,\varepsilon})\geqslant\frac{\mathbb{E}[\mathbf{1}_{\mathcal{A}^L_{x,\delta,\varepsilon}}\mathbb{E}_e[e^{N\frac{\theta_x}{2}\langle e,H_Ne\rangle}]}{\mathbb{E}[e^{N\frac{\theta_x}{2}\langle e,H_Ne\rangle}]}\mathbb{E}[e^{N\frac{\theta_x}{2}\langle e,H_Ne\rangle}]\left(\inf_{Z\in\mathcal{A}^L_{x,\delta,\varepsilon}}\frac{1}{\mathbb{E}_e[e^{N\frac{\theta_x}{2}\langle e,H_Ne\rangle}]}\right).$$
$$=\mathbb{P}^{\theta_x}(\mathcal{A}^L_{x,\delta,\varepsilon})\mathbb{E}[e^{N\frac{\theta_x}{2}\langle e,H_Ne\rangle}]\left(\inf_{Z\in\mathcal{A}^L_{x,\delta,\varepsilon}}\frac{1}{\mathbb{E}_e[e^{N\frac{\theta_x}{2}\langle e,H_Ne\rangle}]}\right).$$

根据引理3.11、3.17和3.21，我们有

$$\liminf_{\delta\downarrow0}\liminf_{N\to\infty}\frac{1}{N}\log\mathbb{P}(|\lambda_{\max}(H_N)-x|\leqslant\delta)\geqslant F(\rho,\theta_x)-J\left(\sigma,\frac{\theta_x}{2},x\right)=-I_\sigma(x,\theta_x)\geqslant-\widetilde{I}_\sigma(x).$$

□

3.6 退化情形

引理2.7的证明。 设 ρ 是 $\mathbb{R}$ 上的紧支撑测度，满足 $r(\rho)\leqslant0$。定理 2.3 已经表明 $r(\sigma)\leqslant0$；与该定理的证明一样，我们考虑一个无离群值选取的序列 $(\Gamma_M)_{M=1}^\infty$，以及利用这些 构造的矩阵 Γ_M。有两种情形：

1. 若 $\alpha\leqslant1$，由于 $H_N\geqslant-KZ^TZ$其中 $K=\max_M(-\lambda_{\min}(\Gamma_M))$. Z^TZ的经验测度收敛于马尔琴科-帕斯图尔分布 MP_α，因此 σ 的累积分布函数处处小于分布函数 $(-K)*\mathrm{MP}_\alpha$，其中 $(-K)*\mathrm{MP}_\alpha$ 是 MP_α 经乘以 $(-K)$所得的推前测度。由于 $r(-K*\mathrm{MP}_\alpha)=-K\ell(\mathrm{MP}_\alpha)=0$，我们有 $r(\sigma)\geqslant0$。

2. 如果 $\alpha>1$，我们可以找到一个 $\delta>1$ 使得 $\alpha-\delta>1$。若令 $\delta_M=\lfloor\delta M\rfloor$，则令 Γ'_M 为 Γ_M中由前 $M-\delta_M$ 行和列组成的子矩阵，令 Γ''_M 为 Γ_M 中由最后 δ_M 行和列组成的子矩阵。我们还令 Z' 为 Z 中由前 $M-\delta_N$ 行组成的子矩阵，令 Z'' 为 Z 中由最后δ_N行组成的子矩阵。我们有 $H_N=H'_N+H''_N$，其中 $H'_N=Z'^T\Gamma'_MZ'$且 $H''_N=Z''^T\Gamma''_MZ''$。由于 $\rho(\{0\})=0$，因此 $\limsup_N\lambda_{\max}(\Gamma'_M)=c<0$；又由于 $H''_N\leqslant0$，对于足够大的 N，有 $H_N\leqslant H'_N\leqslant cZ'^TZ'$。所以$r(\sigma)\leqslant c\ell(MP_{\alpha-\delta})<0$。

□

命题 2.11 的证明。 与非退化情形相同的证明表明， $\lambda_{\max}$ 无法以此速度推入谱 bulk，从而处理了小球概率 $\mathbb{P}(\lambda_{\max}\approx x)$ 对于 $x<0$。

如果 Γ 是负半定的，那么其余论证便显而易见，因为 $H_N\leqslant0$，因此 $\lambda_{\max}(H_N)\leqslant0$ 确定成立。

因此，假设 Γ 在有限的 N 处有少数正特征值，也就是说， H_N 不一定是负半定的。定义 $\Gamma^{\mathrm{nsd}}:=\mathrm{diag}(d_1^{(\mathrm{nsd})},\dots,d_M^{(\mathrm{nsd})})$，其中 $d_i^{(\mathrm{nsd})}=\min(d_i,0)$，令 $\varepsilon_N=\|\Gamma-\Gamma^{(\mathrm{nsd})}\|$；根据假设 1，该值趋于零。再令 $H_N^{(\mathrm{nsd})}:=\frac{1}{M}Z^T\Gamma^{(\mathrm{nsd})}Z$，并使用相同的噪声 Z 将其与 H_N 耦合，以此定义二者。注意到 $\|H_N-H_N^{(\mathrm{nsd})}\|\leqslant\frac{\varepsilon_N}{M}\|Z\|^2$，因此，对于每个 $\delta>0$，

$$\lim_{N\to\infty}\mathbb{P}(|\lambda_{\max}(H_N^{(\mathrm{nsd})})-\lambda_{\max}(H_N)|>\delta)\leqslant\lim_{N\to\infty}\mathbb{P}(\sqrt{M}^{-1}\|Z\|>\delta/\varepsilon_N)=-\infty$$

（其详细内容见下文引理 4.3 的证明）。这意味着序列 $(\lambda_{\max}(H_N))_{N=1}^\infty$和 $(\lambda_{\max}(H_N^{(\mathrm{nsd})}))_{N=1}^\infty$ 是指数等价的；由于根据上述论证，后一个序列满足所需的大偏差原理，根据经典结论（例如参见 [DZ10，定理 4.2。13]），前一个序列也满足该原理。 □

3.7 斯蒂尔切斯变换的第二分支

引理的证明 2.6。若 $x_c(\rho)$ 是有限的，则

$$x_c(\rho) = r(\rho)^2 G_\rho(r(\rho)) + \left(\frac{1}{\alpha} - 1\right) r(\rho) = \frac{1}{\theta_{\max}} + r(\rho)\left(\int_{\mathbb{R}} \frac{r(\rho) - (r(\rho) - u)}{r(\rho) - u}\rho(\mathrm{d}u)\right)$$
$$= \frac{1}{\theta_{\max}} + \int_{\mathbb{R}} \frac{\alpha u}{\alpha - \frac{\alpha}{r(\rho)}u}\rho(\mathrm{d}u) = H_\rho(\theta_{\max}).$$

我们将在证明过程中逐步证明断言 $x_c(\rho) \geqslant r(\sigma)$。最终需要分三种情况讨论。它们都需要计算

$$f_\rho(\theta) := \theta^2 H_\rho'(\theta) = -1 + \alpha \int_{\mathbb{R}} \frac{u^2\theta^2}{(\alpha - u\theta)^2}\rho(\mathrm{d}u).$$

注意，$\lim_{\theta\downarrow 0} f_\rho(\theta) = -1$。我们断言，对于 $\theta \in (0, \theta_{\max})$，$f_\rho$ 严格递增。事实上，它等于函数 $f_{u,\rho}(\theta) := \frac{u^2\theta^2}{(\alpha - u\theta)^2}$ 在 u 上的平均值（再加上一个常数）；函数 $f_{0,\rho}$ 是常数，而对于每个 $u \neq 0$，函数 $f_{u,\rho}$ 都严格递增，因为它们的导数 $2(\frac{\alpha}{u})\theta(\frac{\alpha}{u} - \theta)(\frac{\alpha}{u} - \theta)^4$ 与 $\frac{\alpha}{u}(\frac{\alpha}{u} - \theta)$ 同号，而后者的符号可根据 u 的符号手动判定（在 $u > 0$ 的情形下，这依赖于 $\theta < \theta_{\max} = \frac{\alpha}{r(\rho)}$）。

这三种情形分别是：

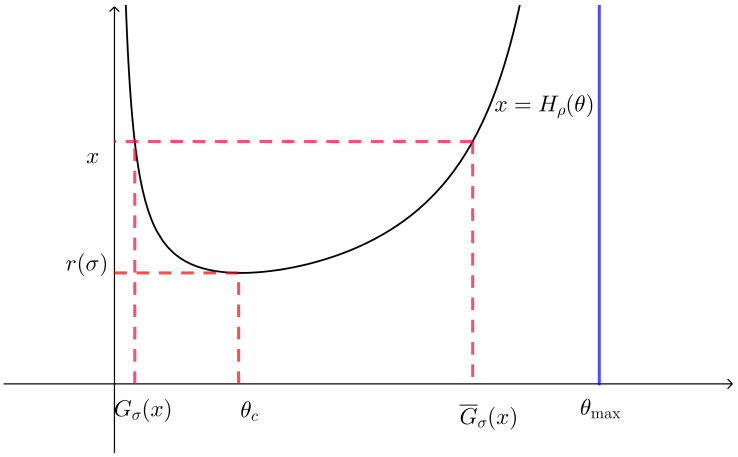

图 1：函数 H_ρ 的图像，其中 $r(\rho) > 0$ 且 $G_\rho(r(\rho)) = +\infty$，并标出了 $r(\sigma)$ θ_c，$G_\sigma(x)$ $G_\sigma(x)$ 的表示。

1. 情形 1（$r(\rho) > 0$ 且 $G_\rho(r(\rho)) = +\infty$），如图 1 所示：容易看出，$H_\rho(\theta_{\max}) = +\infty$。因此，对于某个 θ，$f_\rho(\theta)$ 为正；由于它还严格递增，且在零处趋于 -1，所以存在唯一的 $\theta_c \in (0, \theta_{\max})$ 使其为零，即存在唯一的 $\theta_c \in (0, \theta_{\max})$，使得 H_ρ 在 $(0, \theta_c)$ 上递减，并在 $(\theta_c, \theta_{\max})$ 上递增。利用解析延拓的唯一性，可以推知$H_\rho(\theta_c) = r(\sigma)$ 且 $G_\sigma(r(\sigma)) = \theta_c$。此外，可以看出方程 $H_\rho(y) = x$，视为由 $x \in \mathbb{R}$ 参数化的 $y \in (0, \theta_{\max})$ 的函数，

   (a) 当 $x < r(\sigma)$ 时无解。(b) 当 $x = r(\sigma)$ 时有一个解。该解为 θ_c，并令 $G_\sigma(r(\sigma)) := \theta_c$。(c) 当 $x > r(\sigma)$ 时有两个解 y_1 和 y_2，满足 $0 < y_1 < \theta_c < y_2 < \theta_{\max}$。此外，根据戴森方程 (2.3)，显然有 $y_1 = G_\sigma(x)$。我们用 $G_\sigma(x)$ 表示第二个解 y_2。特别地，以此方式定义在 $[r(\sigma), +\infty)$ 上的 G_σ 是解析递增的，且 $\lim_{x\to\infty} G_\sigma(x) = \theta_{\max}$。

2. 情形 2（$r(\rho) > 0$ 和 $G_\rho(r(\rho)) < +\infty$），见图 2：再一次，$H_\rho$ 在 $(0, \theta_c)$ 上递减，在 $(\theta_c, \theta_{\max})$ 上递增，但当且仅当 $x_c = r(\sigma)$ 时，$\theta_c = \theta_{\max}$。与之前一样，方程 $H_\rho(y) = x$，视为关于 $y \in (0, \theta_{\max})$ 且以 $x \in \mathbb{R}$ 为参数的函数，

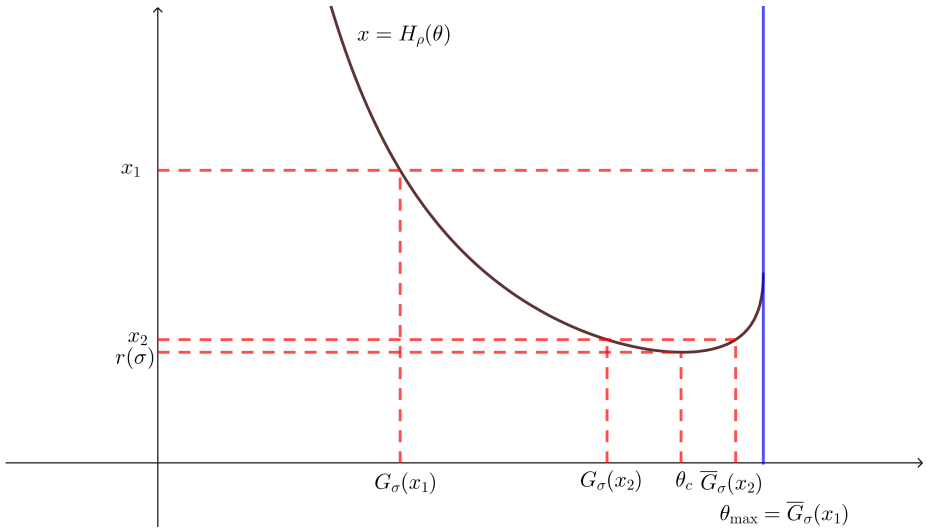

图 2：函数 H_ρ 在 $r(\rho) > 0$ 和 $G_\rho(r(\rho)) < +\infty$ 时的图像，并表示出 $r(\sigma)$ θ_c，$G_\sigma(x)$ $G_\sigma(x)$，其中 $x_1 > x_c$ 和 $x_2 < \overline{x}_c$。

(a) 当 $x < r(\sigma)$ 时无解。(b) 当 $x = r(\sigma)$ 时有一个解。该解为 θ_c，并令 $G_\sigma(r(\sigma)) := \theta_c$。(c) 有两个解 y_1 和 y_2，当 $r(\sigma) < x \leqslant x_c$ 时满足 $0 < y_1 < \theta_c < y_2 < \theta_{\max}$。再一次，我们有$y_1 = G_\sigma(x)$，并令 $G_\sigma := y_2$。(d) 有一个解 y，当 $x > x_c(\rho)$ 时满足 $0 < y < \theta_c$。然而，在这种情况下，我们将定义 $G_\sigma(x) := \theta_{\max}$。

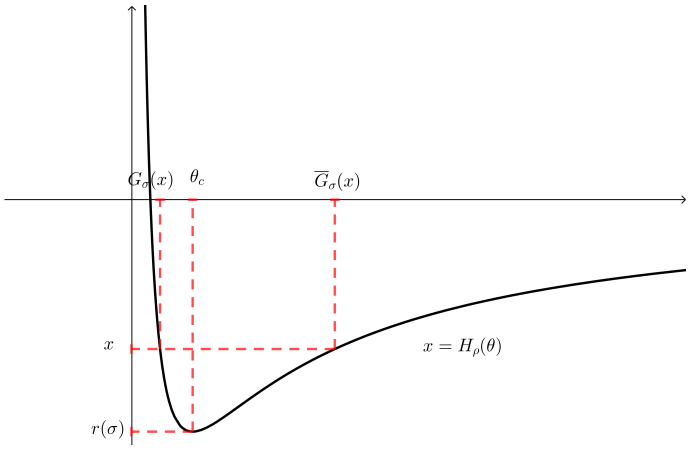

图 3：函数 H_ρ 在 $r(\rho) < 0$ 和 $\alpha > 1$ 时的图像，并表示出 $r(\sigma)$ θ_c，$G_\sigma(x)$ $G_\sigma(x)$。

3. 情形 3（$r(\rho) \leqslant 0$）：利用注 2.8 和引理 2.7，我们只需考虑 $\rho(\{0\}) = 0$且 $\alpha > 1$ 的情形（见图 3）。由于 f_ρ 严格递增，$H'_\rho(\theta)$ 要么对所有 θ 都为负，要么存在 $\theta_c \in (0, \theta_{\max})$，使得 $H'_\rho(\theta)$ 在 $\theta \in (0, \theta_c)$ 上为负，而在 $\theta \in (\theta_c, \theta_{\max})$ 上为正。但由于

$$H_\rho(\theta) = (1-\alpha)\theta^{-1} + o_{\theta \to +\infty}(\theta^{-1}) \tag{3.18}$$

且 $\alpha > 1$，我们必然处于后一种情形，而且确实必有 $H_\rho(\theta_c) < 0$，且 $H_\rho < 0$ 在 $(\theta_c, +\infty)$ 上。于是，方程 $H_\rho(y) = x$，视为由 $x \in \mathbb{R}$ 参数化的 $y \in (0, +\infty)$ 的函数，

(a) 关于 $x \geqslant 0$ 有一个解。该解等于 $G_\sigma(x)$。(b) 关于 $0 > x > r(\sigma)$ 有两个解 y_1， y_2 ，其中 $0 < y_1 < \theta_c < y_2$。我们有 $y_1 = G_\sigma(x)$，并将第二个解 y_2 记为 $G_\sigma(x)$。再次地， G_σ 在 $r(\sigma)$ 与 0 之间解析递增，且 $\lim_{x\uparrow 0} G_\sigma = +\infty$。(c) 关于 $x = r(\sigma)$ 有一个解，即 $y = \theta_c$。(d) 关于 $x < r(\sigma)$ 无解。

□

4 有限情形的证明 x_c

在本节中，我们证明命题 2.15。我们指出，根据定义（2.2），只有在 $x_c(\rho) < +\infty$成立时，才能有 $r(\rho) > 0$。

首先，由其定义和引理 2.6，可知速率函数具有如下形式： $I_\sigma(x) = \frac{1}{2}\int_{r(\sigma)}^{x} g(y)\mathrm{d}y$，其中 g 严格递增，对任意小的自变量都为正，且 $\lim_{x\to+\infty} \frac{g(x)}{\theta_{\max}} = 1$；这证明了 I_σ所声称的性质。

现在我们固定一个 ρ，使得 $x_c(\rho) < +\infty$，并尝试证明相应的大偏差原理，假设我们已知每个满足 $x_c = +\infty$的模型的大偏差原理。证明采用逼近法。具体而言，我们将对 ρ 的右边缘进行离散化，把大于 $r(\rho) - \varepsilon$ 的 d_i替换为 $r(\rho)$。

定义 4.1。 对于 $\varepsilon > 0$，我们定义 $\Gamma^{(\varepsilon)} := \mathrm{diag}(d_1^{(\varepsilon)}, ..., d_M^{(\varepsilon)})$ 其中若 $d_i \leqslant r(\rho) - \varepsilon$，则 $d_i^{(\varepsilon)} = d_i$；并且对于$d_i > r(\rho) - \varepsilon$，有 $d_i^{(\varepsilon)} = r(\rho)$。同理，我们将 $H_N^{(\varepsilon)}$ 定义为：

$$H_N^{(\varepsilon)} := \frac{1}{M} Z^T \Gamma^{(\varepsilon)} Z.$$

之后我们需要将 $H_N^{(\varepsilon)}$ 与 H_N 耦合，即使用相同的噪声 Z 来定义二者。我们还将 $\rho^{(\varepsilon)}$ 定义为 $\mathbb{R}$ 上由下式给出的概率测度：

$$\rho^{(\varepsilon)}(A) := \rho(A \cap]-\infty, r(\rho) - \varepsilon]) + \rho(]r(\rho) - \varepsilon, r(\rho)])\delta_{r(\rho)} \tag{4.1}$$

对于 *Borel* A。

我们注意到

$$\lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} \delta_{d_i^{(\varepsilon)}} = \rho^{(\varepsilon)}$$

只要 ρ 在 $r(\rho) - \varepsilon$ 处没有原子。由于一个概率测度至多有可数个原子，我们可以沿着某个依赖于 ρ 的序列选取 $\varepsilon \to 0$，使其避开这些原子；在后续证明中我们将默认这样做。

那么，若假设我们避开了这类原子， $H_N^{(\varepsilon)}$ 的经验测度将收敛到测度 $\sigma^{(\varepsilon)}$，其特征是它的斯蒂尔切斯变换为 $H_{\rho^{(\varepsilon)}}$的反函数。

由于 $\rho^{(\varepsilon)}$ 在其右端点处有一个原子，我们有 $G_{\rho^{(\varepsilon)}}(r(\rho^{(\varepsilon)})) = +\infty$，因此 $\lambda_{\max}(H_N^{(\varepsilon)})$满足大偏差原理，其速率函数 $I^{(\varepsilon)}$ 定义如下

$$I^{(\varepsilon)}(x) = \begin{cases} \frac{1}{2}\int_{r(\sigma^{(\varepsilon)})}^{x} \left(\overline{G_{\sigma^{(\varepsilon)}}}(t) - G_{\sigma^{(\varepsilon)}}(t)\right) \mathrm{d}t & \text{if } x \geqslant r(\sigma^{(\varepsilon)}), \\ +\infty & \text{otherwise.} \end{cases}$$

为证明我们的结果，我们需要以下三个引理。

引理 4.2. 函数 $\varepsilon \mapsto r(\sigma^{(\varepsilon)})$ 是单调不减的，并且

$$\lim_{\varepsilon\to 0} r(\sigma^{(\varepsilon)}) = r(\sigma). \tag{4.2}$$

此外，该函数 函数 $I^{(\varepsilon)}$ 在 $(r(\sigma), +\infty)$ 的所有紧子集上一致收敛于 I，当 $\varepsilon \to 0$.

引理 4.3。 对于每个 $K > 0$，如果比值 $\frac{\eta}{\varepsilon}$ 足够大（其大小取决于 K），则

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}[\|H_N - H_N^{(\varepsilon)}\| \geqslant \eta] \leqslant -K$$

这里回顾一下，$k \cdot k$ 表示算子范数（在此情形下也就是谱半径）。

引理 4.4. 定义 $J\colon \mathbb{R} \to \mathbb{R}$ 如下

$$J(x) = \sup_{\delta>0} \liminf_{\varepsilon\downarrow 0} \inf_{y\in(x-\delta,x+\delta)} I^{(\varepsilon)}(y). \tag{4.3}$$

则 $I = J$。此外，I 是一个良好速率函数，并且对于任意闭集 $F \subset \mathbb{R}$，我们有

$$\inf_{y\in F} I(y) \leqslant \limsup_{\varepsilon\downarrow 0} \inf_{y\in F} I^{(\varepsilon)}(y). \tag{4.4}$$

暂且假设这 三个引理成立，并证明它们蕴含大偏差原理。

命题 *2.15*的证明。 这将直接由 [DZ10], 中的定理 4.2.16 推出，该定理说明了如何从 $\lambda_{\max}(H_N^{(\varepsilon)})$ 的 $\varepsilon \downarrow 0$ 大偏差原理推出 $\lambda_{\max}(H_N)$ 的大偏差原理。[2] 他们所定义的“指数良好逼近”条件，用我们的记号表述为

$$\lim_{\varepsilon\downarrow 0} \limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\Big(\Big|\lambda_{\max}(H_N^{(\varepsilon)}) - \lambda_{\max}(H_N)\Big| > \delta\Big) = -\infty,$$

这由 引理4.3。我们在上文的引理 4.4 中检查了其结果的其余条件 . □

引理 *4.2*的证明。根据构造，$r(\rho^{(\varepsilon)}) = r(\rho)$，因此 $\theta_{\max}$ 与 ε相互独立。因此

$$r(\sigma^{(\varepsilon)}) = \min_{0<\theta<\theta_{\max}} H_{\rho^{(\varepsilon)}}(\theta).$$

由于 $\rho^{(\varepsilon)}$ 收敛到 ρ，因此 $H_{\rho^{(\varepsilon)}}$ 在 $(0,\ \theta_{\max})$ 的所有紧子集上一致收敛到 H_ρ，这意味着

$$\limsup_{\varepsilon\to 0} r(\sigma^{(\varepsilon)}) \leqslant r(\sigma).$$

另一方面，我们对这一收敛还有更多了解：我们声称，对于每个 $\theta \in (0,\ \theta_{\max})$ ，

$$H_\rho(\theta) \leqslant H_{\rho^{(\varepsilon)}}(\theta) \tag{4.5}$$

并且，函数 $\varepsilon \mapsto H_{\rho^{(\varepsilon)}}(\theta)$ 实际上在 $\varepsilon \in (0, r(\rho))$ 上是非递减的。（注意，这意味着 $\varepsilon \mapsto r(\sigma^{(\varepsilon)})$ 是非递减的。）确实，我们可以写成

$$H_{\rho^{(\varepsilon)}}(\theta) = \frac{1}{\theta} + \int_{-\infty}^{r(\rho)-\varepsilon} \frac{\alpha u}{\alpha-\theta u}\rho(\mathrm{d}u) + \frac{\alpha r(\rho)}{\alpha-\theta r(\rho)}\rho((r(\rho)-\varepsilon, r(\rho)]) = \frac{1}{\theta} + \int_{\mathbb{R}} f_{\alpha,\theta}^{(\varepsilon)}(u)\rho(\mathrm{d}u)$$

其中 $f_{\alpha,\theta}^{(\varepsilon)}\colon \mathrm{supp}(\rho) \to \mathbb{R}$ 定义为

$$f_{\alpha,x}^{(\varepsilon)}(u) = \begin{cases} \frac{\alpha u}{\alpha-\theta u} & \text{if } u < r(\rho)-\varepsilon, \\ \frac{\alpha r(\rho)}{\alpha-\theta r(\rho)} & \text{if } u \geqslant r(\rho)-\varepsilon. \end{cases}$$

由于 $u \mapsto \frac{\alpha u}{\alpha-\theta u}$ 在 ρ 的支撑集上严格递增，且对 $u > 0$为正，因此对于每个 $u \in \mathrm{supp}(\rho)$，映射 $\varepsilon \mapsto f_{\alpha,\theta}^{(\varepsilon)}(u)$ 在集合 $\varepsilon \in (0,\ r(\rho))$ 上非递减。这表明，对于足够小的ε （且关于 θ一致）， $\varepsilon \mapsto H_{\rho^{(\varepsilon)}}(\theta)$ 非递减，从而

$$\liminf_{\varepsilon\downarrow 0} r(\sigma^{(\varepsilon)}) \geqslant r(\sigma),$$

[2]转换一下记号：他们的 m 是我们的 ε^{-1}，他们的 ε 是我们的 N^{-1}。因此，他们的 $\widetilde{\mu_\varepsilon}$ 是 $\lambda_{\max}(H_{\varepsilon^{-1}})$ 的分布，而他们的 $\mu_{\varepsilon,m}$ 是 $\lambda_{\max}(H^{(m^{-1})}{}_{\varepsilon^{-1}})$ 的分布。

完成 (4.2) 的证明。

现在我们证明 $I^{(\varepsilon)}$ 到 I 在 $(r(\sigma)\ +\infty)$ 的紧集上一致收敛。回顾一下， $\theta_{\max}$ 不依赖于 ε。如果 $x > r(\sigma)$，那么当 ε 充分小时，我们有 $x > r(\sigma^{(\varepsilon)})$，从而

$$\begin{aligned} I^{(\varepsilon)}(x) &= \frac{1}{2}\int_{r(\sigma^{(\varepsilon)})}^{x}\left(\overline{G}_{\sigma^{(\varepsilon)}}(t) - G_{\sigma^{(\varepsilon)}}(t)\right)\mathrm{d}t = \frac{1}{2}\int_{r(\sigma^{(\varepsilon)})}^{x}\int_{0}^{\theta_{\max}} \mathbb{1}_{G_{\sigma^{(\varepsilon)}}(t)\leqslant u\leqslant \overline{G}_{\sigma^{(\varepsilon)}}(t)}\,\mathrm{d}u\,\mathrm{d}t \\ &= \frac{1}{2}\int_{0}^{x}\int_{0}^{\theta_{\max}} \mathbb{1}_{H_{\rho^{(\varepsilon)}}(u)\leqslant t}\,\mathrm{d}u\,\mathrm{d}t. \end{aligned}$$

类似地，

$$I(x) = \frac{1}{2}\int_{0}^{x}\int_{0}^{\theta_{\max}} \mathbb{1}_{H_{\rho}(u)\leqslant t}\,\mathrm{d}u\,\mathrm{d}t.$$

定义

$$\begin{aligned} D^{(\varepsilon)} &:= \{(t,u)\in(0,+\infty)\times(0,\theta_{\max}) : H_{\rho^{(\varepsilon)}}(u)\geqslant t > H_{\rho}(u)\}, \\ R_x &:= (0,x)\times(0,\theta_{\max}), \\ D_x^{(\varepsilon)} &:= D^{(\varepsilon)}\cap R_x. \end{aligned}$$

由（4.5），我们实际上有 $I^{(\varepsilon)}(x) \leqslant I(x)$ 以及

$$I(x) - I^{(\varepsilon)}(x) = \frac{1}{2}\int_{0}^{x}\int_{0}^{\theta_{\max}} \mathbb{1}_{H_{\rho}(u)<t\leqslant H_{\rho^{(\varepsilon)}}(u)}\,\mathrm{d}u\,\mathrm{d}t = \frac{1}{2}\operatorname{Leb}(D_x^{(\varepsilon)}).$$

因此，如果 $[a,b]\subset(r(\sigma)\ +\infty)$，那么对于所有 $x\in[a,b]$ 以及所有 $\varepsilon<\varepsilon_0(a)$，我们都有

$$|I^{(\varepsilon)}(x) - I(x)| \leqslant \operatorname{Leb}(D_b^{(\varepsilon)}).$$

由于 $H_{\rho^{(\varepsilon)}}$ 递减至 H_{ρ}，集合 $D_b^{(\varepsilon)}$ 构成嵌套集，并且它们对所有 $\varepsilon>0$ 的交集为空。由于它们的勒贝格测度被 $\operatorname{Leb}(R_b) < \infty$从上方界定，我们有 $\lim_{\varepsilon\downarrow 0}\operatorname{Leb}(D_b^{(\varepsilon)}) = 0$，这就证明了 $I^{(\varepsilon)}$ 在 $(r(\sigma)\ +\infty)$ 的紧集上一致收敛于 I。

□

引理 *4.3* 的证明。确定性地，我们有

$$\|H_N - H_N^{(\varepsilon)}\| = \frac{1}{M}\|Z^T(\Gamma-\Gamma^{(\varepsilon)})Z\| \leqslant \frac{1}{M}\|Z\|^2\|\Gamma-\Gamma^{(\varepsilon)}\| \leqslant \frac{\varepsilon\|Z\|^2}{M}.$$

因此，只需证明：对于每个 $K>0$，都存在 $t_K>0$，使得对所有 $t>t_K$，

$$\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}[\sqrt{M}^{-1}\|Z\|\geqslant t] \leqslant -K.$$

利用 Z 的元素的次高斯性质以及 极限$_N\ \frac{M}{N} = \alpha$，例如采用 [GH20 第 2] 节中的论证，即可推出这一点。

□

引理 *4.4* 的证明。定义

$$J_\delta(x) = \liminf_{\varepsilon\downarrow 0}\ \inf_{y\in(x-\delta,x+\delta)} I^{(\varepsilon)}(y),$$

它关于 δ 单调不增。若 $x < r(\sigma)$，则由引理 4.2 可知，存在 $\delta>0$，使得对所有充分小的 ε， $x+\delta < r(\sigma^{(\varepsilon)})$，因此 $J_\delta(x) = +\infty$，从而 $J(x) = +\infty$。若 $x > r(\sigma)$，则存在 $\delta>0$，使得对所有充分小的 ε， $x-\delta > r(\sigma^{(\varepsilon)})$。由于每个 $I^{(\varepsilon)}$ 都是单调不减的，因此有 $J_\delta(x) = \liminf_{\varepsilon\downarrow 0} I^{(\varepsilon)}(x-\delta) = I(x-\delta)$；再次由引理 4.2 可得， $J(x) = I(x)$。最后，令 $x = r(\sigma)$。于是对每个 $\delta>0$ 以及所有 $\varepsilon<\varepsilon_0(\delta)$，都有$r(\sigma^{(\varepsilon)}) \in (x-\delta, x+\delta)$，因此 $J_\delta(x) = 0 = J(x)$。这就完成了 $I = J$ 的证明；显然， I 是良好速率函数，因为它在 $(-\infty, r(\sigma))$ 上取无穷大，在 $r(\sigma)$处唯一取零，并且严格递增。 α =

现在我们验证 (4.4)，并根据 $+\infty$ $F\subset\ -\infty\ r\ \sigma$ $y\ \ y\in F\ \ <r\ \sigma$ 4 2 $_{y\in F}\, I^{(\varepsilon)}\ y\ = +\infty$ (F)

$\inf_{y\in F} I(\,)$ 是无穷还是有限分情况讨论。若 (F)，则必有 $(\,,(\,))$，且 $\sup\{:\}(\,)$。于是由引理 . 可知，对所有充分小的 ε，$\inf(\,)$。若 $\alpha(F) = 0$，则无需证明。若 $0<\alpha(F)<\infty$，则对某个 $\delta>0$，有 $F\subset(r(\sigma)+\delta,+\infty)$，并令 $y_F = \min\{y:\ y\in F\}$，便有 $\alpha(F) = I(y_F)$。只要 ε 足够小，使得 $y_F > r(\sigma^{(\varepsilon)})$，就有 $\inf_{y\in F} I^{(\varepsilon)}(y) = I^{(\varepsilon)}(y_F)$；当 $\varepsilon\downarrow 0$ 趋于 $I(y_F)$ 时，由引理 4.2 可得此结论。

□

5 复数情形

本节将回顾为将定理 2.13 的证明调整为适用于定理 2.19.

- 我们保留定义3.1 和定义3.2。不过，我们需要修改定义3.7 ，将 $\theta/2$ 替换为 θ:

$$\frac{\mathrm{d}\mathbb{P}^{\theta}}{\mathrm{d}\mathbb{P}}(Z)=\frac{\mathbb{E}_{e}[e^{N\theta\left\langle e,\frac{1}{M}Z^{*}\Gamma Ze\right\rangle}]}{\mathbb{E}_{e,H_N}[e^{N\theta\langle e,H_Ne\rangle}]}.$$

- 在命题3.8 和3.9 中，将2 乘到两个等式的右边:

$$\limsup_{\delta\downarrow 0}\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}_N^{\theta}(|\lambda_{\max}(H_N)-x|\leqslant\delta)\begin{cases}\leqslant -2(\widetilde{I}_{\sigma}(x)-I_{\sigma}(x,\theta)) & \text{if } x\in D,\\ =-\infty & \text{otherwise,}\end{cases}$$

and

$$\liminf_{\delta\downarrow 0}\liminf_{N\to\infty}\frac{1}{N}\log\mathbb{P}_N(|\lambda_{\max}(H_N)-x|<\delta)\geqslant -2\widetilde{I}_{\sigma}(x).$$

- 在引理3.11 中，方程（3.1）替换为

$$\lim_{N\to\infty}\frac{1}{N}\log\mathbb{E}_{e,H_N}[e^{N\theta\langle e,H_Ne\rangle}]=2F(\rho,\theta).$$

在该引理的证明中，Hubbard-Stratonovich 变换变为

$$\mathbb{E}_{H_N}[e^{N\theta\langle e,H_Ne\rangle}]=\prod_{\mu=1}^{M}\frac{1}{\pi}\int_{w\in\mathbb{C}}\mathbb{E}\Big[e^{2\Re(\overline{w}\langle z_\mu,e\rangle)\sqrt{\frac{N}{M}\theta d_\mu}}\Big]e^{-|w|^2}\,\mathrm{d}w.$$

在高斯情形下，我们有:

$$\prod_{j=1}^{N}\mathbb{E}[e^{2\Re(\overline{w}\overline{(z_\mu)}_je_j)\sqrt{\frac{N}{M}\theta d_\mu)}}]=e^{|w|^2\frac{N}{M}\theta d_\mu}$$

然后:

$$\frac{1}{N}\log\mathbb{E}_{e,H_N}[e^{N\theta\langle e,H_Ne\rangle}]=-\frac{M}{N}\int_{\mathbb{R}}\log\Big(1-\frac{N}{M}\theta t\Big)\hat{\mu}_{\Gamma}(\mathrm{d}t).$$

在尖锐次高斯情形下，对任意 $w\in\mathbb{C}$ 和 $c\in\mathbb{R}$，我们得到:

$$\mathbb{E}[\exp(c\Re(\overline{w}\langle z_\mu,e\rangle))]=\prod_{j=1}^{N}\mathbb{E}[\exp(c\Re(\overline{w(z_\mu)_j}e_j))]\leqslant\prod_{j=1}^{N}\exp\Big(\frac{c^2|w|^2|e_j|^2}{4}\Big)=\exp\Big(\frac{c^2|w|^2}{4}\Big),$$

从而得到:

$$\frac{1}{N}\log\mathbb{E}_{e,H_N}[e^{N\theta\langle e,H_Ne\rangle}]\leqslant-\frac{M}{N}\int_{\mathbb{R}}\log\Big(1-\frac{N}{M}\theta t\Big)\hat{\mu}_{\Gamma}(\mathrm{d}t).$$

下界也有类似的修改。

- 在引理 3.17中，我们将方程修改为

$$\begin{aligned}&\lim_{\varepsilon\downarrow 0}\limsup_{\delta\downarrow 0}\limsup_{N\to\infty}\sup_{Z\in\mathcal{A}^L_{x,\delta,\varepsilon}}\Big|\frac{1}{N}\log\mathbb{E}_e[e^{N\theta\langle e,H_Ne\rangle}]-2J\Big(\sigma,\frac{\theta}{2},x\Big)\Big|\\&=\limsup_{\delta\downarrow 0}\limsup_{\varepsilon\downarrow 0}\limsup_{N\to\infty}\sup_{Z\in\mathcal{A}^L_{x,\delta,\varepsilon}}\Big|\frac{1}{N}\log\mathbb{E}_e[e^{N\theta\langle e,H_Ne\rangle}]-2J\Big(\sigma,\frac{\theta}{2},x\Big)\Big|=0.\end{aligned}$$

证明实际上完全相同，我们只需使用复数版本（即 $\beta=2$ 定理 6.2 中的 [GH22]）。需要注意的是，函数 J 在本文与 [GH22]中的约定有所不同。如果我们将J^{GH} 函数 J 在 [GH22],

$$J^{GH}(\mu,\theta,x)=2J\Big(\mu,\frac{\theta}{2},x\Big).$$

- 在引理 3.18 中，我们再次需要将右侧乘以 2：

$$\limsup_{\delta\downarrow 0}\limsup_{\varepsilon\downarrow 0}\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}^{\theta}(\mathcal{A}^{L}_{x,\delta,\varepsilon})\leqslant -2(\widetilde{I}_{\sigma}(x)-I_{\sigma}(x,\theta)).$$

  对引理 3.11 和 3.17 所做的修改同样适用于证明，其余部分保持不变。

- 在命题 3.20 中， $\theta_{\max}$ 的表达式保持不变。

- 在命题 3.9的证明中，同样地，对引理 3.11、3.17所做的修改仍然适用，因此我们得到

$$\liminf_{\delta\downarrow 0}\liminf_{N\to\infty}\frac{1}{N}\log\mathbb{P}(|\lambda_{\max}(H_N)-x|\leqslant\delta)\geqslant 2F(\rho,\theta_x)-2J\left(\sigma,\frac{\theta_x}{2},x\right)=-2I_{\sigma}(x,\theta_x)\geqslant -2\widetilde{I}_{\sigma}(x).$$

  .

附录 A 多维乘积测度的集中

本节介绍 Talagrand [Tal96] 和 Guionnet-Zeitouni[GZ00] 关于乘积测度集中性的一系列经典结果的直接推广，以便研究实部和虚部不一定相互独立的复厄米随机矩阵。

在 20 世纪 90 年代，Talagrand 建立了关于紧支撑测度乘积的集中理论，得到了如下形式的结果："若 $f:[-1,1]^N\to\mathbb{R}$ 是 Lipschitz 函数且具有凸次水平集，并且 $(X_i)_{i=1}^N$ 是取值于 $[-1,1]$ 的独立随机变量，则随机变量 $f(X_1,\dots,X_N)$ 集中于其中位数附近"[Tal96，定理 6.6]。Guionnet 和 Zeitouni 将他的结果推广到随机矩阵，并据此证明了如下形式的结果："若实对称或复厄米 Wigner 矩阵 W_N 的元素具有紧支撑，则 $\hat{}\mu_{W_N}$ 在 Wasserstein-1 距离下集中于其均值 $\mathbb{E}[\hat{}\mu_{W_N}]$ 附近"[GZ00，推论 1.4(b)]。然而，由于 Talagrand 的结果是针对最容易处理的情形 $f:[-1,1]^N\to\mathbb{R}$ 得出的，Guionnet 和 Zeitouni 的结果在复厄米情形下要求 W_N 的元素具有相互独立的实部和虚部；这样一来， W_N 的线性统计量确实可以是 N^2 独立随机变量 $(\mathrm{Re}\ W_{ij},\mathrm{Im}\ W_{ij})_{1\leqslant i<j\leqslant N}\cup(W_{ii})_{i=1}^N$ 的良好函数。

我们希望证明关于稍微更一般的 Wigner矩阵的结果 W_N，其中每个 W_{ij} 的实部和虚部可以彼此相关，只要不同 $W_{ij}\in\mathbb{C}$ 的上三角位置取值之间仍相互独立 i 且 j。为此，我们需要推广 [GZ00], 中的推论 1.4(b)，而这又需要推广 [Tal96]中的定理 6.6。我们沿用 Talagrand 的措辞和大部分符号，以便读者更容易对照，但我们引入了 d维欧几里得单位球

$$B_d=\{x\in\mathbb{R}^d:\|x\|_2^2\leqslant 1\}.$$

命题 A.1。设 f 一个定义在 f 上的实值函数$(B_d)^N$。我们假设，对于每个实数 a ，

$$\textit{the set } \{f\leqslant a\} \textit{ is convex.}$$

设凸集 $C\subset(B_d)^N$，取 $\sigma>0$，并假设 f 在 C 上的限制的利普希茨常数至多为 σ；也就是说，

$$\forall\, x,y\in C,\qquad |f(x)-f(y)|\leqslant\sigma\|x-y\|,$$

其中，kxk 表示范数 $k(x_1,\dots,x_N)k^2=\sum_{i=1}^N kx_ik_2^2$。

考虑取值于 B_d 的独立随机变量$(X_i)^{i\leqslant N}$，以及随机变量

$$h=f(X_1,\dots,X_N).$$

那么，如果 M 是 h 的中位数，则对于所有 $t>0$，有

$$\mathbb{P}(|h-M|\geqslant t)\leqslant 4c+\frac{4}{1-2c}\exp\left(-\frac{t^2}{16\sigma^2}\right)\tag{A.1}$$

其中我们假设

$$c=\mathbb{P}((X_1,\dots,X_N)\notin C)<\frac{1}{2}.$$

我们省略证明，因为这只是对 Talagrand 原始结果的直接改写。值得注意的是，(A.1) 与 d 无关，这一点乍看可能令人意外，因为我们没有对每个 X_i 的 d 元素之间的相关性作出任何假设。然而，缺少对 d 的依赖，本质上是因为我们选择将 Talagrand 的紧集 $[-1,1]$ 扩展到 B_d，对于每个 d，其欧几里得直径均为 2；而不是例如用 $[-1,1]^d$ 替换$[-1,1]$，其欧几里得直径为 $2\sqrt{d}$。（如果我们改为考虑 $f\colon([-1,1]^d)^N \to \mathbb{R}$ 以及变量$X_i \in [-1,1]^d$，则会得到 (A.1) 的一个变体，其右端为 $4c + \frac{4}{1-2c}\exp\left(-\frac{t^2}{16d\sigma^2}\right)$。）

我们猜想，这种形式的扩展已经在文献中出现过，或许不止一次，但我们未能找到它。

通过思考 $\mathbb{C} \cong \mathbb{R}^2$ 并利用 $d=2$ 情形下的命题 A.1，我们得到 [GZ00] 的推论 1.4(b) 的如下扩展。为便于比较，我们再次沿用他们的措辞和记号。我们考虑由下式给出的非齐次复厄米随机矩阵 X_A：

$$X_A = ((X_A)_{ij})_{1\leqslant i,j\leqslant N}, \quad X_A = X_A^*, \quad (X_A)_{ij} = \frac{1}{\sqrt{N}} A_{ij}\omega_{ij}$$

with

$$\omega = (\omega^R + \mathrm{i}\omega^I) = (\omega_{ij})_{1\leqslant i,j\leqslant N} = (\omega_{ij}^R + \sqrt{-1}\omega_{ij}^I)_{1\leqslant i,j\leqslant N}, \quad \omega_{ij} = \overline{\omega_{ji}},$$
$$A = (A_{ij})_{1\leqslant i,j\leqslant N}, \quad A_{ij} = \overline{A_{ji}}.$$

这里 $\{\omega_{ij}, 1 \leqslant i \leqslant j \leqslant N\}$ 是相互独立、分布分别为$\{P_{ij}, 1 \leqslant i \leqslant j \leqslant N\}$的复随机变量，而 P_{ij}' 是 $\mathbb{C}$ 上的概率测度；但现在我们不对其实部和虚部边缘分布之间的关系作任何假设，唯一的条件是 P_{ii} 支撑在 $\mathbb{R}$ 上，以确保 X_A 为厄米矩阵。这里 A 是一个元素$\{A_{ij}, 1 \leqslant i \leqslant j \leqslant N\}$一致有界（例如以 a 为界）的非随机复矩阵。（当然，若令所有 P_{ij}' 都支撑在 $\mathbb{R}$ 上，并令 A 的所有元素均为实数，我们便可得到关于实对称随机矩阵的结果。）

命题 A.2。假设$(P_{ij}, i \leqslant j)$ 一致紧支撑，也就是说，存在一个紧集 $K \subset \mathbb{C}$，使得对于任意 $1 \leqslant i \leqslant j \leqslant N$, $P_{ij}(K^c) = 0$。记

$$\|K\| = \max\{\|x\| : x \in K\}.$$

固定 $\delta_1(N) = 8\|K\|\sqrt{\pi a/N}$以及 $M = a\|K\|\sqrt{8}$。对于任意 $\delta > (128(M+\sqrt{\delta})\delta_1(N))^{2/5}$,

$$\mathbb{P}(\mathrm{W}_1(\hat{\mu}_{X_A}, \mathbb{E}[\hat{\mu}_{X_A}]) > \delta) \leqslant \frac{128(M+\sqrt{\delta})}{\delta^{3/2}} \exp\left(-N^2 \frac{1}{16\|K\|^2 a^2}\left[\frac{\delta^{5/2}}{128(M+\sqrt{\delta})} - \delta_1(N)\right]^2\right).$$

我们再次省略证明，它只需仿照 Guionnet 和 Zeitouni 的证明；唯一需要指出的是，此处 $\|K\|$的定义方式使得 $\frac{K}{\|K\|} \subset B_2$，因此在 Guionnet 和 Zeitouni 使用 [Tal96，定理 6.6]。

附录 B “紧致对数索伯列夫”假设

本节将说明，如何从次高斯情形下关于最大特征值大偏差原理的既有倾斜结果中去除一个特定的技术性假设，我们将其称为“紧支撑或对数 Sobolev”假设。该假设在文献中以多种形式出现：较早的论文往往明确要求某个底层测度具有紧支撑，或满足对数 Sobolev 不等式；而较晚的论文则往往要求经验测度集中方面的某种表述，这在紧支撑或对数 Sobolev 情形下很容易验证。

我们用于去除这一假设的技术，利用了 Guionnet 与第一作者 [GH22]近期证明的球面积分连续性性质的强化版本。具体如下：令 $\hat{\mu}_N$ 表示所研究矩阵的经验谱测度（定义见 (1.4)），μ_∞ 表示其确定性极限，d_{BL} 表示 (1.3) 中的有界-利普希茨距离。以往的结果要求满足如下形式的估计

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_N, \mu_\infty) > N^{-\delta}) = -\infty \tag{B.1}$$

其中 $\delta > 0$ 足够小。在更好的连续性性质下，只需证明

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_N, \mu_\infty) > \varepsilon) = -\infty \tag{B.2}$$

对所有 $\varepsilon > 0$. 我们看到，相较于（B.2），（B.1）有两项主要优势：第一，我们即将证明，（B.2）通常无需紧支撑或对数 Sobolev 假设即可得到证明，其基本思路是仔细截断矩阵元素，并按照 Talagrand 的方法，利用 Guionnet 和 Zeitouni 针对紧支撑乘积测度的集中结果。第二，通常可以在不依赖局部定律的情况下证明（B.2）；此前的结果曾借助局部定律来验证（B.1）（例如参见 [McK21、Hus22]）。由于局部定律仅在某些情况下适用，我们认为，对于未来的大偏差原理结果而言，能够绕过对局部定律的使用或许会有所帮助。

我们在尖锐次高斯 Wigner 矩阵这一最简单的情形下展示这些思路，并证明以下结果；该结果从 [GH20] 的主要结果中去除了“紧支撑或对数 Sobolev”假设。

定理 B.1. 设 μ 是定义在 $\mathbb{R}$ 上的中心化概率测度，且方差为单位方差；设 X_N 是相应的 *Wigner* 矩阵，即 X_N 是一个实对称随机矩阵，其元素除对称性外独立同分布，且服从 $\frac{\mu}{\sqrt{N}}$。若 μ 是尖锐次高斯，则 $\lambda_{\max}(X_N)$ 以 N 为速度满足大偏差原理，其良好速率函数为

$$I(x) = \begin{cases} +\infty & \text{if } x < 2, \\ \frac{1}{2}\int_2^x \sqrt{y^2-4}\,\mathrm{d}y & \text{if } x \geqslant 2. \end{cases}$$

证明. 我们声称，对于每个 $\varepsilon > 0$，都有

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{X_N}, \rho_{\mathrm{sc}}) > \varepsilon) = -\infty. \tag{B.3}$$

引理 [GH20] 证明了一个稍强的结论：在紧支撑或对数 Sobolev 假设下，存在某个足够小的 $\kappa > 0$ 使得 $\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{X_N}, \rho_{\mathrm{sc}}) > N^{-\kappa}) = -\infty$。不过，仿照第 3.4 节和 3.5 节中的论证可知，(B.3) 已经足够。

为证明 (B.3)，我们仿照命题 3.12 的证明，将 $X = A+B$ 分解为 $A_{ij} = X_{ij}\mathbb{1}\{|X_{ij}| \leqslant N^{\gamma-1/2}\}$，其中 $\gamma = \gamma(\varepsilon) > 0$ 待定，然后证明 (3.4)、(3.5)、(3.6) 和 (3.7) 的类似结论。(3.4) 的证明与样本协方差情形相同，不过我们使用经典结果 $d_{\mathrm{KS}}(\hat{\mu}_{X_N}, \hat{\mu}_{A_N}) \leqslant \frac{1}{N}\mathrm{rank}(X_N - A_N)$[BS10, Theorem A.43], ，该结果由特征值交错性得出，而非 [BS10, Theorem A.44], 后者如正文中所述基于奇异值交错性。(3.5) 和 (3.7) 的估计与正文中的相同，而在 Wigner 情形下，(3.6) 的估计实际上更容易：定义$P = \sqrt{8N^\gamma}$和 $\varepsilon_1(N) = 16\sqrt{\pi N^{\gamma-1}}$。由于 $\sqrt{N}A_N$的元素在 $[-N^\gamma, N^\gamma]$中紧支撑， [GZ00, Theorem 1.4(a)] 给出

$$\mathbb{P}(\mathrm{W}_1(\hat{\mu}_{A_N}, \mathbb{E}[\hat{\mu}_{A_N}]) > \varepsilon) \leqslant \frac{128(P+\sqrt{\varepsilon})}{\varepsilon^{3/2}} \exp\left(-N^2 \frac{1}{64N^{2\gamma}}\left[\frac{\varepsilon^{5/2}}{128(P+\sqrt{\varepsilon})} - \varepsilon_1(N)\right]^2\right)$$

只要 ε 满足隐式方程 $\varepsilon > (128(P+\sqrt{\varepsilon})\varepsilon_1(N))^{2/5}$，其右端的阶为$N^{(2\gamma-1)(2/5)}$，因此当 N 足够大时，隐式方程便成立。于是，指数函数中的项的阶为 $N^{2-2\gamma+2(-\gamma)}$，且当 γ 足够小时，指数中的幂至少为 1，这就足够了。 □

注 B.2。 这个证明允许 X_N 的元素服从尖锐次高斯分布，而无需具有紧支撑或满足对数 *Sobolev* 不等式。下面给出一个这样的分布的例子。为此，考虑 Z 一个标准高斯变量，$\mathcal{F} = \sigma(\{Z \in [n, n+1[\} : n \in \mathbb{Z})$ 并定义

$$\tilde{Z} = sign(Z)\sqrt{\mathbb{E}[Z^2|\mathcal{F}]}.$$

˜变量 Z 是中心化的，方差为 1。显然，它不是紧支撑的；又由于它的支撑集是一个˜离散点集，因此不可能满足任何对数 *Sobolev* 不等式。接下来只需证明 Z 是尖锐次高斯的。˜为此，我们来考察它的矩。由于 Z 的分布关于 $k \in \mathbb{N}$对称，因此有：

$$\mathbb{E}[\tilde{Z}^{2k+1}] = 0$$

对于偶数阶矩，应用詹森不等式可得

$$\mathbb{E}[\tilde{Z}^{2k}] = \mathbb{E}[\mathbb{E}[Z^2|\mathcal{F}]^k] \leqslant \mathbb{E}[Z^{2k}]$$

以及 $\mathrm{Var}(\tilde{Z}) = \mathbb{E}[\tilde{Z}^2] = \mathbb{E}[\mathbb{E}[Z^2|F]] = \mathbb{E}[Z^2] = 1$。因此，由于 Z 是高斯的，对于 $t \in \mathbb{R}$

$$\mathbb{E}[\exp(t\tilde{Z})] \leqslant \mathbb{E}[\exp(tZ)] \leqslant \exp\left(\frac{t^2}{2}\right),$$

˜这意味着 Z 确实是尖锐次高斯的。

附录 C 变形 Wigner矩阵

在本附录中，我们运用自身的技术改进第二作者先前关于所谓“变形 Wigner 矩阵”的结果 [McK21]。在这一模型中，考虑一个 $N \times N$ 随机矩阵

$$X_N = \frac{W_N}{\sqrt{N}} + D_N,$$

它可以是实对称矩阵，也可以是复 Hermitian 矩阵，其中 W_N 除对称性外，其元素独立同分布，且均服从某个中心化概率测度 μ （定义在 $\mathbb{R}$ 上，若 $\beta = 1$ ；或定义在 $\mathbb{C}$ 上，若 $\beta = 2$），并且确定性矩阵 D_N满足以下假设，该假设弱于 [McK21] 中相应的假设。

假设 2. 矩阵 D_N 是实对角确定性矩阵，其经验测度 $\hat{}\mu_{D_N}$ 随着 $N \to \infty$ 趋于无穷而弱收敛到某个紧支撑概率测度 μ_D，且渐近地不存在外部离群值，具体而言：

$$\lambda_{\max}(D_N) \to r(\mu_D),$$
$$\lambda_{\min}(D_N) \to \ell(\mu_D).$$

注 C.1. 关于符号的一点说明：我们在研究广义样本协方差情形时使用的许多对象，在这里也有对应物。我们选择复用符号，而不是让记号变得冗杂。例如，广义样本协方差情形有一个阈值 x_c，定义见(2.2)。这里也需要一个类似的阈值，尽管定义(C.1) 有所不同，我们仍将这个新版本称为 x_c ，而不另用例如 $x_c^{(\mathrm{dw})}$这样的记号。我们也同样复用了 H、速率函数 I等记号，只有一个显著例外：定义 $J(\mu, \theta, \lambda)$ 中的特殊函数*3.1*在两个模型中完全相同。

经验测度的典型极限 该模型下经验测度的典型收敛行为是经典结果 [Pas72，Voi91]：我们有 e

$$\hat{\mu}_{X_N} \to \rho_{\mathrm{sc}} \boxplus \mu_D =: \mu_D^{\mathrm{sc}},$$

其中 ρ_{sc} 是半圆律 $\rho_{\mathrm{sc}}(\mathrm{d}x) = \sqrt{(4-x^2)_{2\pi}}$ $^+$ $\overline{\mathrm{d}x,\ 记}$号 $\boxplus$ 表示两个紧支撑概率测度的自由加法卷积，而 μ_D^{sc} 是简写。回顾一下， $\boxplus$ 是用 Voiculescu R-变换定义的，这对我们很重要：对于紧支撑概率测度 μ，回顾斯蒂尔切斯变换 G_μ，它是从 $(r(\mu) + \infty)$ 到 $(0, G_\mu(r(\mu)))$ 的递减双射。我们将其反函数记为$K_\mu : (0, G_\mu(r(\mu))) \to (r(\mu) + \infty)$，并将其 R-变换记为 $R_\mu(y) = K_\mu(y) - \frac{1}{y}$；该变换使自由卷积线性化，即 $R_{\mu \boxplus \nu} = R_\mu + R_\nu$。

下面的引理定义了我们编写速率函数所需的函数。

引理 C.2。 函数 H： $(0, \infty) \to \mathbb{R}$ 定义为

$$H(y) := \begin{cases} y + K_{\mu_D}(y) & \text{if } 0 \leqslant y \leqslant G_{\mu_D}(r(\mu_D)), \\ y + r(\mu_D) & \text{if } y \geqslant G_{\mu_D}(r(\mu_D)), \end{cases}$$

具有以下性质：

- H 是连续且凸的。
- H 在 $y_c = G_{\rho_{\mu^{\mathrm{sc}}D}}(r(\mu_{\mathrm{sc}D}))$ 处取得唯一最小值，该点位于 $(0, G_{\mu_D}(r(\mu_D))]$,，且 $H(y_c) = r(\mu_{\mathrm{sc}D})$。
- 存在一个连续、严格递增的函数 $\mathfrak{G}$： $[r(\mu_{\mathrm{sc}D}) \ +\infty) \ \overline{\to} \mathbb{R}$，具有以下性质：
  - *If* $0 \leqslant y \leqslant G_{\mu_D}(r(\mu_D))$, *then* $H(\overline{\mathfrak{G}}(y)) = y$, *and* $\{w : H(w) = y\} = \{G_{\mu_D^{\mathrm{sc}}}(y), \overline{\mathfrak{G}}(y)\}$.
  - *We have* $\overline{\mathfrak{G}}(y) > G_{\mu_D^{\mathrm{sc}}}(y)$ *for* $y > r(\mu_D^{\mathrm{sc}})$, *and* $\overline{\mathfrak{G}}(r(\mu_D^{\mathrm{sc}})) = G_{\mu_D^{\mathrm{sc}}}(r(\mu_D^{\mathrm{sc}}))$.

定义 C.3。定义 I： $\mathbb{R} \to [0, +\infty]$ 为

$$I(x) = \begin{cases} +\infty & \text{if } x < r(\rho_{\mathrm{sc}} \boxplus \mu_D), \\ \frac{1}{2}\int_{r(\rho_{\mathrm{sc}} \boxplus \mu_D)}^{x} (\overline{\mathfrak{G}}(y) - G_{\rho_{\mathrm{sc}} \boxplus \mu_D}(y))\,\mathrm{d}y & \text{if } x \geqslant r(\rho_{\mathrm{sc}} \boxplus \mu_D). \end{cases}$$

定理 C.4。 假设 *2* 成立，且 μ 是尖锐次高斯的（若 *2.1*，则按定义 $\beta = 1$；若 *2.18*，则按定义 $\beta = 2$）。则 $\lambda_{\max}(X_N)$ 满足速度为 N、良好速率函数为 $I^{(\beta)} = \beta I$ 的大偏差原理。该函数是凸的，并且在 $[r(\rho_{\mathrm{sc}} \boxplus \mu_D) \ +\infty)$ 上严格递增（特别地，仅在 $r(\rho_{\mathrm{sc}} \boxplus \mu_D)$ 处取零），并且

$$\lim_{x \to +\infty} \frac{I^{(\beta)}(x)}{\frac{\beta x^2}{4}} = 1.$$

如果 μ实际上是高斯分布，那么根据旋转不变性，我们无需假设 D_N 是对角的；只需假设它是对称的（若 $\beta = 1$）或厄米的（若 $\beta = 2$），并满足假设 *2*中的其余条件。

注 C.5。 这改进了 *[McK21]*, 的结果；该结果给出了定理 *C.4*的一个较弱版本，并要求额外满足三个假设：首先，μ 实际上要么是高斯测度（即 $\frac{W_N}{\sqrt{N}}$ 是高斯正交系综/高斯酉系综），要么重要的阈值 $\overline{}$

$$x_c = x_c(\mu_D) := \begin{cases} r(\mu_D) + G_{\mu_D}(r(\mu_D)) & \text{if } G_{\mu_D}(r(\mu_D)) < +\infty, \\ +\infty & \text{otherwise} \end{cases} \tag{C.1}$$

是无穷大；其次， μ 要么是紧支撑的，要么满足对数 *Sobolev* 不等式；第三，变形以某种温和的多项式速度趋于其极限，$d_{\mathrm{BL}}(\hat{}\mu_{D_N}, \mu_D) \lesssim N^{-\varepsilon}$ 其中某个 $\varepsilon > 0$。然而，该处给出的速率函数形式不同；下面我们将证明这两种形式等价。

我们利用附录*B*中的方法去掉第二和第三个假设。与正文中一样，我们通过用一列 $x_c < +\infty$ 模型逼近 $x_c = +\infty$ 模型，然后利用关于逼近大偏差原理的教科书结果，去掉第一个假设。

定义 C.6。 对于任意 $x \geqslant r(\rho_{\mathrm{sc}} \boxplus \mu_D)$ 以及任意 $\theta \geqslant 0$，定义

$$I(x, \theta) := J(\rho_{\mathrm{sc}} \boxplus \mu_D, \theta, x) - \theta^2 - J(\mu_D, \theta, r(\mu_D)).$$

~利用这一点，定义速率函数 I： $\mathbb{R} \to [0, +\infty]$ 为

$$\widetilde{I}(x) := \begin{cases} +\infty & \text{if } x < r(\rho_{\mathrm{sc}} \boxplus \mu_D), \\ \sup_{\theta \geqslant 0} I(x, \theta) & \text{if } x \geqslant r(\rho_{\mathrm{sc}} \boxplus \mu_D). \end{cases}$$

~**引理 C.7**。若 $x_c(\mu_D) = +\infty$，则 $I = I$。

引理 C.8。 若假设*2* 成立，且 μ 是尖锐次高斯的，并且

$$x_c(\mu_D) = +\infty,$$

then $\lambda_{\max}(X_N)$ 满足速度为 N、良好速率函数为.....的大偏差原理 I.

引理 C.9。 引理 *C.8*推出定理 *C.4*。

引理 C.2 的证明。连续性很容易验证。由于 G_{μ_D} 在 $(r(\mu_D)\ +\infty)$ 上严格凸，函数 K_{μ_D} 在 $(0, G_{\mu_D}(r(\mu_D)))$ 上严格凸，因此 H 是凸的，事实上在 $(0, G_{\mu_D}(r(\mu_D)))$ 上严格凸。由于边界行为为 $\lim_{y\to 0} H(y) = \lim_{y\to+\infty} H(y) = +\infty$，因此 H 在某个 y_c 处取得唯一最小值，而该点必在 $(0, G_{\mu_D}(r(\mu_D))]$ 内。另一方面，令 $y_s := G_{\mu_D^{\mathrm{sc}}}(r(\mu_D^{\mathrm{sc}}))$。HAN 的引理 6.1 给出 $y_s \leqslant \min(G_{\rho_{\mathrm{sc}}}(r(\rho_{\mathrm{sc}}))\ G_{\mu_D}(r(\mu_D))$，而 McK 的命题 6.1 证明中的一个计算表明 $G'_{\mu_D}(y_s) = -1$，因此，通过对 $y = K_{\mu_D}(G_{\mu_D}(y))$ 关于 y 求导并在 y_s 处求值，可得 $H'(y_s) = 0$。因此 $y_s = y_c$。McK 的命题 6.1 证明中的另一个计算表明 $H(y_s) = r(\mu_D^{\mathrm{sc}})$。

现在我们研究 $\overline{\mathfrak{G}}$。由于在我们的归一化中，$R_{\rho_{\mathrm{sc}}}(y) = y$，我们有 $H(y) = R_{\rho_{\mathrm{sc}}}(y) + R_{\mu_D}(y) + \frac{1}{y} = K_{\rho_{\mathrm{sc}}\boxplus\mu_D}(y)$，其中 $y < y_c$。由于 H 在 $(0, b)$ 上连续且严格凸（其中某个 b），在 $(b, +\infty)$ 上仿射递增，并在 $y_c \in (0, b]$ 处取得最小值，因此不难看出，对于 $y < H(y_c)$，$\#\{w : H(w) = y\}$ 为零；对于 $y = H(y_c)$，其为一；对于 $y > H(y_c)$，其为二；这些元素中较小的一个是 $G_{\rho_{\mathrm{sc}}\boxplus\mu_D}(y)$；并且，如果 $\mathfrak{G}$ 表示 H 在 $(y_c, +\infty)$ 上的反函数，那么 $\mathfrak{G}$ 具有所声称的性质。 □

~引理 C.7 的证明。证明与广义样本协方差情形中的证明相同，即证明 I 和 I 在 $(r(\rho_{\mathrm{sc}} \boxplus \mu_D)\ +\infty)$ 上具有相同的导数，并且二者都在 $r(\rho_{\mathrm{sc}} \boxplus \mu_D)$ 处为零。我们依赖于此前在 [McK21] 的命题 6.1 中完成的两项计算。第一项计算表明，I 在 $r(\rho_{\mathrm{sc}} \boxplus \mu_D)$ 处为零。第二项表明，对于每个 $x > r(\rho_{\mathrm{sc}} \boxplus \mu_D)$，都有 $I(x) = I(x, \theta_x)$，其中 θ_x 是约束优化问题的唯一解

$$H(2\theta_x) = 2\theta_x + K_{\mu_D}(2\theta_x) = x \quad \text{subject to} \quad 2\theta_x \in (G_{\rho_{\mathrm{sc}}\boxplus\mu_D}(r(\rho_{\mathrm{sc}} \boxplus \mu_D)), G_{\mu_D}(r(\mu_D))),$$

我们用斯蒂尔切斯变换分支的新语言将其识别为 $\theta_x = \frac{1}{2}\mathfrak{G}(x)$；并且该极大值点~是唯一的，即 $I(x,\ \theta) < I(x)$，如果 $\theta \neq \theta_x$。因此

$$\frac{\mathrm{d}}{\mathrm{d}x}\widetilde{I}(x) = \frac{\partial}{\partial x} I(x,\theta)\Big|_{\theta=\theta_x} = \frac{\partial}{\partial x} J(\rho_{\mathrm{sc}} \boxplus \mu_D, \theta, x)\Big|_{\theta=\theta_x} = \theta_x - \frac{1}{2} G_{\rho_{\mathrm{sc}}\boxplus\mu_D}(x) = \frac{1}{2}(\overline{\mathfrak{G}}(x) - G_{\rho_{\mathrm{sc}}\boxplus\mu_D}),$$

由此完成证明。 □

引理 *C.8* 的证明。如上所述，这是 [McK21]，的主要结果，除了 (a) 速率函数在该文中~写作 βI（引理 C.7表明这无关紧要），以及 (b) 该文要求 μ 要么具有紧支撑，要么满足对数索伯列夫条件，并且要求

$$d_{\mathrm{BL}}(\hat{\mu}_{D_N}, \mu_D) \leqslant CN^{-\varepsilon} \tag{C.2}$$

对于某个 C 和 ε。在附加假设 (b) 下，[McK21] 的引理 5.3 表明

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{X_N}, \rho_{\mathrm{sc}} \boxplus \mu_D) > N^{-\kappa}) = -\infty$$

对于 $\kappa > 0$ 足够小的情形。要获得这一较小的多项式速度，(a) 需要局部定律 [EKS19]，的一个（非常弱的推论）；而 (b) 是要求满足 (C.2) 的根本原因。但附录 B 说明，实际上只需证明

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{X_N}, \rho_{\mathrm{sc}} \boxplus \mu_D) > \varepsilon) = -\infty \tag{C.3}$$

对每个 $\varepsilon > 0$ 都成立；这使我们可以舍去 (C.2) 的要求，并给出一种绕过局部定律的证明。为此，按照附录 B 中的做法，我们将 $W_N N^{-1/2} = A + B = A_N + B_N$with $A_{ij} = W_{ij} N^{-1/2} \mathbb{1}\{|W_{ij} N^{-1/2}| \leqslant N^{\gamma-1/2}\}$ 拆分，其中 $\gamma = \gamma(\varepsilon) > 0$ 取为待定值，然后分解 $X_N = (A + D_N)\ + (B + D_N)$，并证明 (3.4)、(3.5)、(3.6) 和 (3.7) 的类似结果。(3.4)、(3.5) 和 (3.7) 的类似结果与之前完全相同。(3.6) 的类似结果是如下估计

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(d_{\mathrm{BL}}(\hat{\mu}_{A+D_N}, \mathbb{E}[\hat{\mu}_{A+D_N}]) > \varepsilon) = -\infty.$$

在附录B中，当 $D_N = 0$时，我们指出 $\sqrt{N}A$的元素在 $[-N^\gamma, N^\gamma]$中紧支撑，并应用了 [GZ00]的结果。当 $D_N = 0$ 时，矩阵 $\sqrt{N}(A + D_N)$的元素紧支撑于边长为N^γ量级的盒子中，但这些盒子的中心已偏离零点。这需要对[GZ00]，的结果作一个直接的修改；这一修改已作为引理 5.9 在 [McK21]中陈述。至此，(C.3) 的证明完成。 □

引理C.9的证明。这个证明也与广义样本协方差情形中的证明相同。我们全程使用引理C.2的结果，为简便起见，只写 $\beta = 1$ 的情形。

根据其定义和引理 C.2，可知速率函数具有形式 $I(x) = \frac{1}{2}\int_{r(\rho_{\mathrm{sc}}\boxplus\mu_D)}^{x} g(y)\,\mathrm{d}y$，其中$g$ 严格递增，对任意小的自变量均为正，且 $\lim_{x\to+\infty}\frac{g(x)}{x} = 1$；这证明了关于 I 的所述性质。

固定一个 μ_D，使得 $x_c(\mu_D) < +\infty$，并将 $d_i = d_i^{(N)}$ 记为 D_N 的对角元素，即$D_N = \mathrm{diag}(d_1,\dots,d_N)$。对于 $\varepsilon > 0$，定义 $D_N^{(\varepsilon)} = \mathrm{diag}(d_1^{(\varepsilon)},\dots,d_N^{(\varepsilon)})$，其中若 $d_i \leqslant r(\mu_D) - \varepsilon$，则 $d_i^{(\varepsilon)} = d_i$；否则$d_i^{(\varepsilon)} = r(\mu_D)$。然后令

$$X_N^{(\varepsilon)} = N^{-1/2}W_N + D_N^{(\varepsilon)},$$

通过使用相同的随机性 W_N，将其与 X_N 耦合。令 $\mu_D^{(\varepsilon)}$ 如 (4.1) 中所定义，并注意，若 $\varepsilon \to 0$ 避开了 $r(\mu_D)$ 附近存在的任何原子，则 $X_N^{(\varepsilon)}$ 的经验测度收敛至 $\rho_{\mathrm{sc}}\boxplus\mu_D^{(\varepsilon)}$，相应的函数为$H^{(\varepsilon)}$。根据构造，$r(\mu_D^{(\varepsilon)}) = r(\mu_D)$，但对每个 ε，$G_{\mu_D^{(\varepsilon)}}(r(\mu_D^{(\varepsilon)})) = +\infty$，因此 $\lambda_{\max}(X_N^{(\varepsilon)})$ 满足速度为 N、良好速率函数为 $I^{(\varepsilon)}$ 的大偏差原理，其定义为

$$I^{(\varepsilon)}(x) = \begin{cases} +\infty & \text{if } x < r(\rho_{\mathrm{sc}}\boxplus\mu_D^{(\varepsilon)}), \\ \frac{1}{2}\int_{r(\rho_{\mathrm{sc}}\boxplus\mu_D)}^{x}(\overline{\mathfrak{G}}^{(\varepsilon)}(y) - G_{\rho_{\mathrm{sc}}\boxplus\mu_D^{(\varepsilon)}}(y))\,\mathrm{d}y & \text{if } x \geqslant r(\rho_{\mathrm{sc}}\boxplus\mu_D^{(\varepsilon)}), \end{cases}$$

其中 $\mathfrak{G}^{(\varepsilon)}$ 如引理C.2 中对测度 $\mu_D^{(\varepsilon)}$ 的定义。

为了完成证明，我们只需要引理4.2、4.3和4.4的类似结论。引理4.3 的类似结论在这里甚至更容易，因为$\|X_N - X_N^{(\varepsilon)}\| \leqslant \varepsilon$ 是确定性的。引理4.4 本质上是引理4.2的推论，这里仍然如此，因此我们只需要引理4.2的类似结论。

关于这一结果：与正文中类似的论证表明，对每个 $x > r(\mu_D) = r(\mu_D^{(\varepsilon)})$，函数 $\varepsilon \mapsto G_{\mu_D^{(\varepsilon)}}(x)$ 在 $\varepsilon > 0$ 较小时单调不减，因此 $\varepsilon \mapsto K_{\mu_D^{(\varepsilon)}}(y)$ 单调不减，从而 $\varepsilon \mapsto H^{(\varepsilon)}(y)$ 单调不减。因此 $\varepsilon \mapsto r(\rho_{\mathrm{sc}}\boxplus\mu_D^{(\varepsilon)})$ 单调不减，且 $\liminf_{\varepsilon\downarrow 0} r(\rho_{\mathrm{sc}}\boxplus\mu_D^{(\varepsilon)}) \geqslant r(\rho_{\mathrm{sc}}\boxplus\mu_D)$。对于另一个不等式，显然 $H^{(\varepsilon)}$ 在 $(0, G_{\mu_D}(r(\mu_D))]$ 的所有紧子集上一致收敛到 H。但是 $H^{(\varepsilon)}(y) - y$关于 y 严格递减，并且当 $y \to +\infty$ 时趋于 $r(\mu_D)$；因此 $H^{(\varepsilon)}$ 在 $(0, +\infty)$ 的所有紧子集上一致趋于 H，从而 $\lim_{\varepsilon\to 0} r(\rho_{\mathrm{sc}}\boxplus\mu_D^{(\varepsilon)}) = r(\rho_{\mathrm{sc}}\boxplus\mu_D)$。最后，对于 $r > r(\rho_{\mathrm{sc}}\boxplus\mu_D)$，且当 ε 足够小时，我们有如下表示式

$$I^{(\varepsilon)}(x) = \frac{1}{2}\int_0^x\int_0^\infty \mathbb{1}_{H^{(\varepsilon)}(u)\leqslant t}\,\mathrm{d}u\,\mathrm{d}t \quad \text{and} \quad I(x) = \frac{1}{2}\int_0^x\int_0^\infty \mathbb{1}_{H(u)\leqslant t}\,\mathrm{d}u\,\mathrm{d}t,$$

因此再次有 $I^{(\varepsilon)}(x) \leqslant I(x)$，且 $\sup_{x\in[a,b]}|I^{(\varepsilon)}(x) - I(x)| \leqslant \mathrm{Leb}(D_b^{(\varepsilon)})$，其中 $D_b^{(\varepsilon)}$ 已作适当的重新定义。集合 $D_b^{(\varepsilon)}$ 再次构成递减序列，且交集为空，因此若其勒贝格测度有限，则该测度趋于零。此前这一有限性显而易见，但这里需要稍加思考：由于 $\lim_{y\to+\infty} H(y) = +\infty$，存在 c 使得$H(y) \geqslant b$对所有 $y \geqslant c$都成立，于是 $D_b^{(\varepsilon)} \subset [0,b]\times[0,c]$，因此 $\mathrm{Leb}(D_b^{(\varepsilon)}) < +\infty$。证明的其余部分与正文中的证明相同。

□

参考文献

[AGH21] Fanny Augeri、Alice Guionnet 和 Jonathan Husson。次高斯矩阵最大特征值的大偏差。《数学物理通讯》，383(2):997–1050，2021 年。[BS10] Zhidong Bai 和 Jack W. Silverstein。大维随机矩阵的谱分析。Springer统计学系列。Springer，纽约，第二版，2010 年。[BPZ15] Zhigang Bao、Guangming Pan 和 Wang Zhou。一般总体样本协方差矩阵最大特征值的普适性。《统计年刊》，43(1):382–421，2015 年。

[BGH20] Serban Belinschi、Alice Guionnet 和 Jiaoyang Huang。通过球面积分得到的大偏差原理，2020 年。arXiv:2004.07117v2。[Ben65] George Bennett。独立有界随机变量之和的矩上界与概率不等式。《Biometrika》，52:559–569，1965 年。[BG20] Giulio Biroli 和 Alice Guionnet。尖峰高斯随机矩阵的最大特征值与特征向量的大偏差。《电子概率通讯》，25：第 70 号论文，13 页，2020 年。[BC14] Charles Bordenave 和 Pietro Caputo。无高斯尾部的 Wigner矩阵的大偏差原理。《概率年刊》，42(6):2454–2496，2014 年。[BCC11] Charles Bordenave、Pietro Caputo 和 Djalil Chafaï。非厄米重尾随机矩阵的谱。《数学物理通讯》，307(2):513–560，2011 年。[CDG23] Nicholas Cook、Raphaël Ducatez 和 Alice Guionnet。次高斯 Wigner矩阵最大特征值的大偏差，2023 年。即将在 arXiv 发布。[DZ10] Amir Dembo 和 Ofer Zeitouni。大偏差技术及其应用，随机建模与应用概率第 38 卷。Springer-Verlag，柏林，2010 年。1998 年第二版的修订重印本。[EK07] Noureddine El Karoui。一大类复样本协方差矩阵最大特征值的 Tracy-Widom 极限。《概率年刊》，35(2):663–714，2007 年。[EKS19] László Erdős、Torben Krüger 和 Dominik Schröder。慢相关衰减的随机矩阵。《Forum Math. Sigma》，7:e8，89 页，2019 年。[FJ22] Zhou Fan 和 Iain M. Johnstone。实协方差与MANOVA估计量各谱边缘处的 Tracy-Widom 分布。《应用概率年刊》，32(4):2967–3003，2022 年。[GH20] Alice Guionnet 和 Jonathan Husson。Rademacher矩阵最大特征值的大偏差。《概率年刊》，48(3):1436–1465，2020 年。[GH22] Alice Guionnet 和 Jonathan Husson。k 维球面积分的渐近性质及其应用。《ALEA 拉丁美洲概率与数理统计杂志》，19(1):769–797，2022 年。[GM20] Alice Guionnet 和 Mylène Maïda。两个随机矩阵之和的最大特征值的大偏差。《电子概率杂志》，25：第 14 号论文，24 页，2020 年。[GZ00] Alice Guionnet 和 Ofer Zeitouni。大型矩阵谱测度的集中。《电子概率通讯》，5:119–136，2000 年。[HHN16] Walid Hachem、Adrien Hardy 和 Jamal Najim。大型复相关Wishart矩阵：谱边缘处的涨落与渐近独立性。《概率年刊》，44(3):2264–2348，2016 年。[Hus22] Jonathan Husson。方差剖面矩阵最大特征值的大偏差。《电子概率杂志》，27：第 74 号论文，44 页，2022 年。[Kat87] Tosio Kato。离散谱的变化。《数学物理通讯》，111(3):501–504，1987 年。[KY17] Antti Knowles 和 Jun Yin。随机矩阵的各向异性局部定律。《概率论及相关领域》，169(1-2):257–352，2017 年。[LS16] Ji Oon Lee 和 Kevin Schnelli。一般总体实样本协方差矩阵最大特征值的 Tracy-Widom 分布。《应用概率年刊》，26(6):3786–3839，2016 年。[Mai21] Antoine Maillard。广义样本协方差矩阵极端特征值的大偏差。《欧洲物理快报》，133(2):20005，2021 年 1 月。[MV09] Satya N. Majumdar 和 Massimo Vergassola。Wishart矩阵与高斯随机矩阵最大特征值的大偏差。《物理评论快报》，102:060601，2009 年 2 月。[MP67] V. A. Marčenko 和 L. A. Pastur。若干类随机矩阵的特征值分布。《苏联数学文集》，1(4):457–483，1967 年。[McK21] Benjamin McKenna。变形Wigner随机矩阵极端特征值的大偏差。《电子概率杂志》，26（论文 34）:1–37，2021 年。[MP22] Pierre Mergny 和 Marc Potters。不变随机矩阵之和或乘积的顶端特征值的右大偏差原理。《统计力学：理论与实验》，2022(6)：第 063301 号论文，65 页，2022 年。

[Ona08] Alexei Onatski。《奇异复 Wishart 矩阵最大特征值的 Tracy-Widom 极限》。《应用概率年刊》，18(2):470–490，2008 年。[Pas72] L. A. Pastur。《随机矩阵的谱》。《理论与数学物理》，10(1):102–112，1972 年。[SX21] Kevin Schnelli 和 Yuanyuan Xu。《样本协方差矩阵最大特征值收敛至 Tracy–Widom 定律的收敛速率》，2021 年。arXiv:2108.02728v1。[SB95] Jack W. Silverstein 和 Z. D. Bai。《一类大维随机矩阵特征值的经验分布》。《多元分析杂志》，54:175–192，1995 年。[Tal96] Michel Talagrand。《重新审视独立性》。《概率年刊》，24(1):1–34，1996 年。[Voi91] Dan Voiculescu。《随机矩阵与自由积的极限定律》。《数学发明》，104(1):201–220，1991 年。[Wan19] Haoyu Wang。《样本协方差矩阵最大特征值的定量普适性》，2019 年。arXiv:1912.05473v1。[Zhi21] Nikita Zhivotovskiy。《通过变分原理得到的独立矩阵之和与简单张量的无维数依赖界》，2021 年。arXiv:2108.08198v3。