arXiv:2301.02705v1 [math.PR] 6 Jan 2023

协方差损失、塞梅雷迪正则性与差分隐私

MARCH BOEDIHARDJO、THOMAS STROHMER 和 ROMAN VERSHYNIN

Abstract. 我们展示了如何利用基于格罗滕迪克恒等式的随机舍入，证明协方差损失具有近乎紧的界——协方差损失是指取条件期望时损失的协方差量。这一结果为半正定矩阵和核带来了一种新型的弱塞梅雷迪正则性引理。此外，它还可用于构造差分隐私合成数据。

1. Introduction

考虑一个取值于 $\mathbb{R}^d$ 的随机向量 X，以及一个 σ-代数 $\mathcal{F}$。令 Y 表示条件期望：$Y = \mathbb{E}[X|\mathcal{F}]$。在维数为 $d = 1$ 的情形下，其中 X 和 Y 是随机变量，全方差公式指出

$$\mathrm{Var}(X) - \mathrm{Var}(Y) = \mathbb{E}\, X^2 - \mathbb{E}\, Y^2 = \mathbb{E}(X-Y)^2 \geq 0. \tag{1.1}$$

因此，取条件期望会低估方差。这一观察可推广到更高维情形。具体而言，令 $\Sigma_X = \mathbb{E}(X - \mathbb{E}\, X)(X - \mathbb{E}\, X)^{\mathsf{T}}$ 表示 X 的协方差矩阵，Σ_Y 的协方差矩阵也类似定义。那么

$$\Sigma_X - \Sigma_Y = \mathbb{E}\, XX^{\mathsf{T}} - \mathbb{E}\, YY^{\mathsf{T}} = \mathbb{E}(X-Y)(X-Y)^{\mathsf{T}} \succeq 0, \tag{1.2}$$

其中 $\succeq$ 表示洛文纳序；若 $A - B$ 为半正定，则 $A \succeq B$。与一维情形一样，我们看到取条件期望会低估协方差。

在 [2], 中，我们提出了一个基本问题：“协方差损失 $\Sigma_X - \Sigma_Y$有多大？”显然，答案取决于所选的σ代数 $\mathcal{F}$，这引出了下一个问题：“在给定复杂度下，哪个σ代数 $\mathcal{F}$ 能使协方差损失最小？”

文献 [2] 证明，存在一个将样本空间划分为至多 k 个集合的划分，使得由该划分生成的σ代数 $\mathcal{F}$ 对应的协方差损失（1.2）至多为 $C\sqrt{\log\log k/\log k}$，其中 C 是一个绝对常数。在 [6], 中，Jain、Sah 和 Sawhney 将这一上界改进为 $C/\log k$。由文献 [2] 中的命题 3.14 可知，除常数 C的取值外，这一上界已是最优的。文献 [2] 和 [6] 中的上界证明都较为繁琐。本文为这一问题提出了一种新方法，其概念上与塞梅雷迪正则性相关。这种方法给出了关于协方差损失常数的更紧上界（推论 5.1），同时也使证明大为简化，更加初等。

著名的塞梅雷迪正则性引理指出，对于任何足够大的图，其节点集合都可以划分为大小大致相同的子集，使得不同子集之间的边表现得几乎是随机的 [10]。用线性代数的语言表述，这意味着图的邻接矩阵可以近似表示为若干个割矩阵之和[1]。塞梅雷迪正则性引理无论是在组合形式还是线性代数形式下，都具有

[1]割矩阵是一种在某个块上取常数、在其他地方为零的秩一矩阵[5]。1

许多深刻的算法和组合应用，例如 [5, 8, 7, 1]。我们证明了一个新的（弱）塞梅雷迪正则性引理版本，它指出，任何对角线一致有界的半正定核 $K(x,\ y$都可以分解为若干个 k阶梯函数之和，再加上一个 L^2范数为 $O(1/\sqrt{\log k})$的误差项。

本文结构如下。首先，我们将协方差损失界定为“协方差增量”（命题 2.1），然后找出协方差增量的一种优美表示（命题 2.3），并结合基于格罗滕迪克恒等式的随机舍入来界定协方差增量（定理 3.1）。因此，我们不仅能界定协方差损失（推论 5.1），还能为矩阵（定理 5.3）和核（定理 5.6推导出一种新型的弱 Szemeredi 正则性。最后，我们将说明如何结合差分隐私与合成数据运用这些结果，以改进 [2] 中得到的精度界。

1.1. 记号。随机变量 X 的次高斯范数记作 $\|X\|_{\psi_2}$ （例如参见 [11]）， $\mathbb{E}(X)$ 表示 X 的期望，且 $\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p}$。对于 $d \times d$ 矩阵 A、B，定义内积 $\langle A、\ B\rangle = \mathrm{Tr}(AB^T)$，以及弗罗贝尼乌斯范数 $\|A\|_F = \sqrt{\langle A、\ A\rangle}$。

2. The covariance increment

我们首先着手通过协方差增量来界定协方差损失 $\mathbb{E}\,XX^{\mathsf{T}} - \mathbb{E}\,YY^{\mathsf{T}}$。

在下文中， X 是一个取值于 $\mathbb{R}^d$的随机向量， $\mathcal{F}$ 是一个 σ-代数，并且 $Y = \mathbb{E}[X|\mathcal{F}]$。此外， $(X', \mathcal{F}', Y')$是$(X, \mathcal{F}, Y)$的一个独立副本，因此样本空间是一个乘积空间，其中 $X, \mathcal{F}, Y$基于乘积空间的第一部分，而$X', \mathcal{F}', Y'$ 基于第二部分。

稍微滥用一下记号，乘积 σ-代数 $\mathcal{F} \times \mathcal{F}$ 表示由 $\mathcal{F}$和 $\mathcal{F}'$生成的 σ-代数。此外，内积 $\langle Y, Y'\rangle = \mathbb{E}(\langle X, X'\rangle | \mathcal{F} \times \mathcal{F})$。

命题 2.1（协方差损失与协方差增量）。

$$\left\|\mathbb{E}\,XX^{\mathsf{T}} - \mathbb{E}\,YY^{\mathsf{T}}\right\|_F^2 \le \left\|\mathbb{E}\,XX^{\mathsf{T}}\right\|_F^2 - \left\|\mathbb{E}\,YY^{\mathsf{T}}\right\|_F^2. \tag{2.1}$$

命题 2.1 的证明将基于以下引理：

引理 2.2. 设 (U, V) 是取值于 $\mathbb{R}^d$ 的一对随机向量，且 (U', V') 是其独立副本。则

$$\langle \mathbb{E}\,UU^{\mathsf{T}}, \mathbb{E}\,VV^{\mathsf{T}}\rangle = \mathbb{E}\langle U, V'\rangle^2.$$

特别地，令 $V = U$，我们有

$$\left\|\mathbb{E}\,UU^{\mathsf{T}}\right\|_F^2 = \mathbb{E}\langle U, U'\rangle^2.$$

证明。第一个恒等式可通过先使用同分布性，再利用独立性直接得到：

$$\langle \mathbb{E}\,UU^{\mathsf{T}}, \mathbb{E}\,VV^{\mathsf{T}}\rangle = \langle \mathbb{E}\,UU^{\mathsf{T}}, \mathbb{E}\,V'(V')^{\mathsf{T}}\rangle = \mathbb{E}\langle UU^{\mathsf{T}}, V'(V')^{\mathsf{T}}\rangle = \mathbb{E}\langle U, V'\rangle^2.$$

引理得证。 □

命题的证明 *2.1.* 展开弗罗贝尼乌斯范数的平方，可将（2.1）的左侧表示为 $\left\|\mathbb{E}\, XX^{\mathsf{T}}\right\|_F^2 - 2\langle \mathbb{E}\, XX^{\mathsf{T}},\ \mathbb{E}\, YY^{\mathsf{T}}\rangle + \left\|\mathbb{E}\, YY^{\mathsf{T}}\right\|_F^2$。经过化简，我们发现不等式（2.1）等价于

$$\left\|\mathbb{E}\, YY^{\mathsf{T}}\right\|_F^2 \le \langle \mathbb{E}\, XX^{\mathsf{T}}, \mathbb{E}\, YY^{\mathsf{T}}\rangle.$$

利用引理2.2 我们可以将其改写为

$$\mathbb{E}\langle Y, Y'\rangle^2 \le \mathbb{E}\langle X, Y'\rangle^2, \tag{2.2}$$

为验证（2.2），请回想 $Y = \mathbb{E}[X|\mathcal{F}]$ 并应用条件詹森不等式。 □

协方差增量有一个很好的表示形式，这在我们后续的分析中会派上用场。

命题 2.3。

$$\left\|\mathbb{E}\, XX^{\mathsf{T}}\right\|_F^2 - \left\|\mathbb{E}\, YY^{\mathsf{T}}\right\|_F^2 = \mathbb{E}\left(\langle X, X'\rangle - \langle Y, Y'\rangle\right)^2. \tag{2.3}$$

证明。 根据引理 2.2，(2.3) 的左边等于 $\mathbb{E}\langle X,\ X'\rangle^2 - \mathbb{E}\langle Y,\ Y'\rangle^2$。注意$\langle Y,\ Y'\rangle = \mathbb{E}(\langle X,$ $X'\rangle|\,\mathcal{F}\times\mathcal{F})$。为完成证明，对$\langle X,\ X'\rangle$ 应用全方差公式 (1.1)，而非对 X，并对 $\langle Y,\ Y'\rangle$ 应用全方差公式，而非对 Y。 □

3. Bounding the covariance increment

定理 3.1. 设 X 是一个取值于 $\mathbb{R}^d$ 的随机向量，且 $\|X\|_2 \le 1$ 几乎处处成立。则对于任意 $r \in \mathbb{N}$，存在一个将样本空间划分为至多 2^r 个部分的划分，使得由该划分生成的 σ-代数 $\mathcal{F}$ 的条件期望 $Y = \mathbb{E}[X|\mathcal{F}]$ 满足

$$\mathbb{E}\left(\langle X, X'\rangle - \langle Y, Y'\rangle\right)^2 \le \frac{\pi^2}{r}.$$

定理 3.1 的证明将用到格罗滕迪克恒等式（例如参见 [11, 引理 3.6。6]）：

引理 3.2（格罗滕迪克恒等式）。设 x、x' 是 $\mathbb{R}^d$ 中一对单位向量，且令 $g \sim N\in(0, I_d)$。则

$$\langle x, x'\rangle = \sin\left[\frac{\pi}{2}\,\mathbb{E}\,\mathrm{sign}\langle x, g\rangle\,\mathrm{sign}\langle x', g\rangle\right].$$

定理 *3.1* 的证明。步骤 1。 我们先作一个更强的假设，即$\|X\|_2 = 1$几乎必然成立，并证明一个较弱的结论，即存在一个随机变量 Z，它关于乘积 σ-代数 $\mathcal{F}\times\mathcal{F}$ 可测，且满足

$$\left\|\langle X, X'\rangle - Z\right\|_{L^2} \le \frac{\pi}{2\sqrt{r}}. \tag{3.1}$$

为此，考虑独立随机向量 $g_1, \dots, g_k \in N(0, I_d)$，并对 $x, x' \in \mathbb{R}^d$，定义随机变量

$$F_g(x, x') = \sin\left[\frac{\pi}{2}\cdot\frac{1}{r}\sum_{k=1}^{r}\mathrm{sign}\langle x, g_k\rangle\,\mathrm{sign}\langle x', g_k\rangle\right].$$

记作 $\xi_k = $符号$\langle x,\ g_k\rangle$符号$\langle x',\ g_k\rangle$，注意到函数 sin(·) 是 1-Lipschitz 的，并应用引理 3.2，可得

$$\left|\langle x,x'\rangle - F_g(x,x')\right| \le \frac{\pi}{2}\cdot\frac{1}{r}\left|\sum_{k=1}^{r}(\xi_k - \mathbb{E}\,\xi_k)\right|. \tag{3.2}$$

由独立性可得

$$\mathbb{E}_g\left(\langle x,x'\rangle - F_g(x,x')\right)^2 \le \left(\frac{\pi}{2r}\right)^2\cdot r\,\mathrm{Var}(\xi_1) \le \frac{\pi^2}{4r}.$$

代入 $x = X$、 $x' = X'$ ，并对 X 和 X'取期望。由富比尼定理，存在随机向量 g_1、……、 g_k 的一个实现，使得

$$\mathbb{E}\left(\langle X,X'\rangle - F_g(X,X')\right)^2 \le \frac{\pi^2}{4r},$$

固定这样一个实现。令 $\mathcal{F} = \sigma(V(X))$为由随机向量$V(X) = (\mathrm{sign}\langle x, g_k\rangle)_{k=1}^r$生成的 σ-代数。由于 $V(X)$至多取 2^r 个值， $\mathcal{F}$满足定理的要求。此外，随机向量

$$Z = F_g(X,X') = \sin\left(\frac{\pi}{2r}\langle V(X),V(X')\rangle\right)$$

关于乘积 σ-代数 $\mathcal{F}\times\mathcal{F}$可测。因此，我们证明了（3.1）。

步骤 2：将 Z替换为 $\langle Y,\ Y'\rangle$。条件期望 $\mathbb{E}\left[\cdot|\,\mathcal{F}\times\mathcal{F}\right]$ 是 L^2 中投影到由关于 $\mathcal{F}\times\mathcal{F}$可测的随机变量构成的子空间上的正交投影。由于上一步构造的随机变量 Z关于 $\mathcal{F}\times\mathcal{F}$可测，且 $\langle Y,\ Y'\rangle = \mathbb{E}\left[\langle X,X'\rangle|\,\mathcal{F}\times\mathcal{F}\right]$，由（3.1）可得

$$\left\|\langle X,X'\rangle - \langle Y,Y'\rangle\right\|_{L^2} \le \left\|\langle X,X'\rangle - Z\right\|_{L^2} \le \frac{\pi}{2\sqrt{r}}.$$

步骤 3：去除单位范数要求。我们在附加假设 X 是一个单位随机向量的条件下证明了该定理。现在令 X 为满足 $\|X\|_2 \le 1$的任意随机向量。令 $N \in \mathbb{N}$，并考虑随机向量 $W = \sqrt{1-\|X\|_2^2}\cdot\theta$，其中 θ是一个在 $\mathbb{R}^N$的单位基 $\{e_1,\dots,e_N\}$上均匀分布的随机向量。于是，直和 $X\oplus W = (X_1,\dots,X_d,W_1,\dots,W_N)$ 是 $\mathbb{R}^{d+N}$中的一个单位随机向量。对 $X\oplus W$应用上一步的结论，我们得到

$$\begin{aligned}\frac{\pi}{2\sqrt{r}} &\ge \left\|\langle X\oplus W, X'\oplus W'\rangle - \langle Y\oplus\eta, Y'\oplus\eta'\rangle\right\|_{L^2} \quad (\text{where } \eta = \mathbb{E}(W|\mathcal{F}))\\ &= \left\|\langle X,X'\rangle - \langle Y,Y'\rangle + \langle W,W'\rangle - \langle\eta,\eta'\rangle\right\|_{L^2}\\ &\ge \left\|\langle X,X'\rangle - \langle Y,Y'\rangle\right\|_{L^2} - \left\|\langle W,W'\rangle - \langle\eta,\eta'\rangle\right\|_{L^2}.\end{aligned} \tag{3.3}$$

由于 $\langle\eta,\eta'\rangle = \mathbb{E}\left(\langle W,\ W'\rangle|\,\mathcal{F}\times\mathcal{F}\right)$，全方差公式（1.1）以及 W的定义给出

$$\left\|\langle W,W'\rangle - \langle\eta,\eta'\rangle\right\|_{L^2} \le \left\|\langle W,W'\rangle\right\|_{L^2} \le \left\|\langle\theta,\theta'\rangle\right\|_{L^2} = \frac{1}{\sqrt{N}}, \tag{3.4}$$

其中，最后一个等式成立是因为 $\langle\theta,\ \theta'\rangle$ 服从参数为 $1/N$ 的伯努利分布。取 N足够大，根据（3.3）可得

$$\left\|\langle X,X'\rangle - \langle Y,Y'\rangle\right\|_{L^2} \le \frac{\pi}{\sqrt{r}}.$$

定理 3.1 得证。 □

备注 3.3. 在定理 3.1的结论中，我们可以将常数 π^2 替换为 $\pi^2/4+\epsilon$，其中任意固定的 $\epsilon>0$。

4. Subgaussian error

我们可以将定理 3.1 中的界推广到次高斯范数，即相对于杨函数 $\psi_2(x) = e^{x^2}-1$的奥利奇范数。因此，随机变量 X 是次高斯的，当且仅当

$$\|X\|_{\psi_2} = \inf\{t>0 : \mathbb{E}\,\psi_2(X/t)\le 1\} < \infty, \tag{4.1}$$

这个量称为 X 的次高斯范数，参见 [11, 第 2.5、2.7 节1]。

定理 4.1. 设 X 是一个取值于 $\mathbb{R}^d$ 的随机向量，且 $\|X\|_2\le 1$ 几乎处处成立。那么，对于任意 $r\in\mathbb{N}$，存在一个将样本空间划分为至多 2^r 个部分的划分，使得由此划分生成的 σ 代数 $\mathcal{F}$ 的条件期望 $Y=\mathbb{E}[X|\mathcal{F}]$ 满足

$$\big\|\langle X,X'\rangle - \langle Y,Y'\rangle\big\|_{\psi_2} \le \frac{C}{\sqrt{r}}. \tag{4.2}$$

我们可以通过修改定理 3.1 的证明来证明这个结果。下面说明具体做法。

在步骤 1 中，我们可以使用霍夫丁不等式来控制独立伯努利随机变量之和 ξ_k，而不是利用方差的可加性。记 $\zeta_k=\xi_k-\mathbb{E}\,\xi_k$，则有 $\big\|\sum_{k=1}^r\zeta_k\big\|_{\psi_2}\lesssim\sqrt{r}$，参见 [11, 命题 2.6。1]。将其代入式（3.2），得到

$$\big\|\langle x,x'\rangle - F_g(x,x')\big\|_{\psi_2}\lesssim 1/\sqrt{r}.$$

根据次高斯范数的定义，这意味着

$$\mathbb{E}_g\,\psi_2\left((\langle x,x'\rangle - F_g(x,x'))c\sqrt{r}\right)\le 1,$$

其中 $c>0$ 是某个绝对常数。在此代入 $x=X$ 和 $x'=X'$，并如步骤 1 中那样应用富比尼不等式，我们得到式（3.1）的以下形式：

$$\big\|\langle X,X'\rangle - Z\big\|_{\psi_2}\lesssim\frac{1}{\sqrt{r}}.$$

在步骤 2 中，尽管条件期望在次高斯范数下不是度量投影，但它是一个近似度量投影：

引理 4.2。设 X 是一个随机变量，$\mathcal{F}$ 是一个 σ-代数。那么，对于任何 $\mathcal{F}$-可测的随机变量 Z，以及任意 $p\ge 1$，我们有

$$\big\|X-\mathbb{E}(X|\mathcal{F})\big\|_{\psi_2}\le 2\|X-Z\|_{\psi_2}.$$

证明。减去并加上 Z，再利用三角不等式，我们得到

$$\big\|X-\mathbb{E}(X|\mathcal{F})\big\|_{\psi_2}\le\|X-Z\|_{\psi_2}+\big\|\mathbb{E}(X|\mathcal{F})-Z\big\|_{\psi_2}.$$

由于 Z 是 $\mathcal{F}$-可测的，我们有

$$\big\|\mathbb{E}(X|\mathcal{F})-Z\big\|_{\psi_2}=\big\|\mathbb{E}(X-Z|\mathcal{F})\big\|_{\psi_2}\le\|X-Z\|_{\psi_2},$$

其中最后一步由次高斯范数的定义（4.1）和条件詹森不等式得出。结合这两个界即可完成证明。
□

在步骤 2 中，使用该引理处理 $\langle X、\ X'\rangle$ 而不是 X，我们得到

$$\big\|\langle X,X'\rangle-\langle Y,Y'\rangle\big\|_{\psi_2}\le 2\big\|\langle X,X'\rangle-Z\big\|_{\psi_2}\lesssim\frac{1}{\sqrt{r}}.$$

在步骤 3 中，我们用类似的方法处理次高斯范数。界 (3.4) 变为

$$\big\|\langle W,W'\rangle-\langle \eta,\eta'\rangle\big\|_{\psi_2}\le 2\big\|\langle W,W'\rangle\big\|_{\psi_2}\le\big\|\langle \theta,\theta'\rangle\big\|_{\psi_2}\lesssim\frac{1}{\sqrt{\log N}},$$

这里，第一步中我们对 4.2 应用引理 $X=\langle W、\ W'\rangle$ 和 $Z=0$；最后一步则是对参数为 $1/N$的伯努利随机变量的次高斯范数作直接估计。证毕。

5. Implications

5.1. 协方差损失。 结合命题 2.1、命题 2.3 和定理 3.1，可得：

推论 5.1（协方差损失）。设 X 是一个取值于 $\mathbb{R}^d$ 的随机向量，且$\|X\|_2\le 1$ 几乎处处成立。那么，对任意 $r\in\mathbb{N}$，存在一个将样本空间划分为至多2^r 个部分的划分，使得由该划分生成的 σ-代数 $\mathcal{F}$ 的条件期望$Y=\mathbb{E}[X|\mathcal{F}]$ 满足

$$\big\|\mathbb{E}\,XX^{\mathsf{T}}-\mathbb{E}\,YY^{\mathsf{T}}\big\|_F\le\frac{\pi}{\sqrt{r}}.$$

特别地，对于任意满足 $\log_2 k\in\mathbb{N}$ 的 k，都存在一个将样本空间划分为至多 k 个部分的划分，使得对于由该划分生成的 σ代数 $\mathcal{F}$，我们有

$$\big\|\mathbb{E}\,XX^{\mathsf{T}}-\mathbb{E}\,YY^{\mathsf{T}}\big\|_F\le\frac{\pi}{\sqrt{\log_2 k}}.\tag{5.1}$$

备注 5.2（最优性）。 命题 3.14 [2] 表明，存在一个随机向量 X取值于 $\mathbb{R}^d$的单位球内，使得 $\big\|\mathbb{E}\,XX^{\mathsf{T}}-\mathbb{E}\,YY^{\mathsf{T}}\big\|_F\ge\sqrt{\log_2(e)}\,\overline{80\sqrt{\log_2 k}}$对于任意由至多 σ 个部分的划分生成的代数$\mathcal{F}$都成立。因此，式（5.1）中的界在绝对常数因子范围内是紧的。

5.2. 弱 Szemerédi 正则性。

定理 5.3（弱 Szemerédi 正则性）。设 A 是一个 $n\times n$ 半正定矩阵，满足 $A_{ii}\le 1$ 对所有 i 均成立。那么，对于任意 $r\in\mathbb{N}$，存在一个划分 $[n]=I_1\cup\ldots\cup I_k$，满足 $k\le 2^r$，以及一个矩阵 B，它在每个块 $I_i\times I_j$ 上均为常数，并且满足

$$\frac{1}{n}\|A-B\|_F\le\frac{\pi}{\sqrt{r}}.\tag{5.2}$$

此外， B 可以通过对每个块中 A的各项取平均来计算。

证明。 根据假设， A 可以表示为某些向量的格拉姆矩阵 x_1，...， x_n满足 $\|x_i\|_2\le 1$ 对所有 i均成立。因此 $A=[\langle x_i,x_j\rangle]_{i,\ j=1}^n$。对在 3.1 中所述的随机向量 X 应用定理，该向量在 $\{x_1,\ldots,x_n\}$ 上服从均匀分布，从而完成证明。□

注 5.4。 将定理5.3 中的近似误差改写为

$$\frac{1}{n^2}\sum_{i,j=1}^n(A_{ij}-B_{ij})^2\le\frac{\pi^2}{r},$$

我们可以将其解释为对各个元素均方误差的界。

注 5.5。 应用定理 4.1 而非定理 3.1，我们可以将（5.2）中的弗罗贝尼乌斯范数替换为更强的矩阵范数，例如各元素的 ℓ_p 范数。

定理 5.6（弱 Szemerédi 正则性，解析形式）。设 (Ω, μ) 是一个概率测度空间。设 $K : \Omega \times \Omega \to \mathbb{R}$ 是一个可测正定核，满足 $K(t,t) \le 1$ 对所有 $t \in \Omega$ 都成立。那么对于任意 $r \in \mathbb{N}$，存在一个划分 $\Omega = I_1 \cup \ldots \cup I_k$，使得 $k \le 2^r$，以及一个函数 $L : \Omega \times \Omega \to \mathbb{R}$，它在每个块 $I_i \times I_j$ 上为常数，使得

$$\|K - L\|_{L^2(\mu\times\mu)} \le \frac{\pi}{\sqrt{r}}. \tag{5.3}$$

此外， L可以通过在每个块上对 K 取平均来计算。

证明。当 Ω 有限时，利用定理 3.1 （另见注记 3.3），很容易证明$\|K - L\|_{L^2(\mu\times\mu)} \le \frac{\pi}{2\sqrt{r}} + \epsilon$ 对任意固定的 $\epsilon > 0$ 都成立。一般情形可由标准的逼近论证得到，此处留给读者。□

5.3. 与弱 Szemerédi 正则性现有研究的比较。我们的结果比 Frieze 和 Kannan [5] 提出的经典弱正则性引理限制更多，但也更强。该文中的近似误差 $R = K - L$ 是用割范数度量的

$$\|R\|_\square = \sup_{S,T\subset\Omega} \left| \int_{S\times T} R(x,y)\, d\mu(x) d\mu(y) \right|.$$

割范数与算子范数等价 $L^1 \to L^\infty$ (例如参见 [8])，并且显然满足$\|R\|_\square \le \|R\|_{L^2(\mu\times\mu)}$。因此，本文关注的 Hilbert-Schmidt 范数比割范数更强。因此，我们的界（5.3）自动也适用于割范数。Frieze 和 Kannan 的原始工作中给出了一个适用范围更广的割范数界 [5]： 它仅适用于核，但函数 $K(x,y)$ 可以是任意逐点有界于 1 的可测函数，参见 [8, 引理 3。1]与此不同，若要在更强的 Hilbert-Schmidt 范数下对误差给出任何非平凡界，例如定理 5.6中的界，则必须要求半正定性。我们将在下一节看到这一点。

按照定理 5.3的思路进行的矩阵分解，即误差以弗罗贝尼乌斯范数为界的分解，见 [3, 定理 7] Deshpande、Kannan 和 Srivastava 提出的定理。该定理不要求矩阵为半正定，但只有对低秩矩阵 A才有实质意义，也就是说，矩阵的秩至多为对数级[2] 于维数n。我们的结果没有秩的限制。

6. Optimality

定理 $1/r$ 3.1 结论中的衰减率是最优的，因为推论 5.1 结论中的衰减率$1/\sqrt{r}$ 也是最优的。

文献中此前所有关于弱 Szemerédi 正则性的结果都不需要假设矩阵 A 是半正定的，参见例如 [5, 8, 1]。这一假设能否从我们的定理 5.3中去掉？以下结果表明，这一假设不能去掉。

[2]定理 7 在 [3] 中将 A 近似为约 $t = \mathrm{rank}(A)\log^4 n$ 个割矩阵之和；这些割矩阵还可以进一步拆分为 2^t 个支撑集互不相交的更小割矩阵。只有当 $2^t < n^2$时，这种最终分解才有实质意义。

命题 6.1（半正定性至关重要）。设 A 是一个 $n\times n$ 哈达玛矩阵， $[n]=I_1\cup\ldots\cup I_k$ 是任意一个划分。那么，对于任意在每个块 $I_i\times I_j$上取常数值的矩阵 B，我们有

$$\frac{1}{n}\|A-B\|_F\geq\frac{1}{2}\sqrt{1-\frac{k}{n}}.$$

因此，除非部分数 k 极大，即 $k=n-o(n)$，否则误差不会消失。

证明将用到以下初等界：

引理 6.2。设 $x_1,\ldots,x_m$ 是 $\mathbb{R}^n$ 中的正交归一向量。那么对于每个 $k=1,\ldots,m$，有

$$\Big\|x_k-\frac{1}{m}\sum_{t=1}^m x_t\Big\|_2\geq 1-\frac{1}{m}.$$

证明。向量 $x_k-\frac{1}{m}\sum_{t=1}^m x_t$ 的范数不小于该向量与 x_k 的内积，而该内积等于 $1-1/m$。□

命题 *6.1* 的证明。 对 $\|A-B\|_F$在命题中所有矩阵 B 上的最小值，如果改为在所有在更小的块上为常数的矩阵上取最小值，只会变小；具体而言，这些块是集合 $\{\ell\}\times I_j$，其中 $\ell=1,\ldots,n$ 且 $j=1,\ldots,k$。后一个最小值由矩阵 [3] 取得，该矩阵 B 是通过对 A 每个块中的元素取平均$\{\ell\}\times I_j$得到的。等价地， B是通过对 A每个集合 I_j中的列取平均得到的。具体来说， B 的列为

$$B_\ell=\frac{1}{|I_j|}\sum_{t\in I_j}A_t\quad\text{for }\ell\in I_j,$$

其中 B_ℓ是 B的 ℓ列， A_t是 A的第 t列。注意，至少有 $n-k$个指标属于非单元素块 I_j，即

$$T:=\bigcup_{j:|I_j|>1}I_j\quad\text{satisfies}\quad|T|\geq n-k.$$

（事实上，由于 T^c 是所有单元素块的并集，其基数等于这类块的数量 I_j，而该数量不超过块的总数 k。）

任取一个指标 $\ell\in T$，于是 $\ell\in I_j$ 其中$|I_j|\geq 2$。由于 A 是一个哈达玛矩阵，其各列 A_t 相互正交，且欧几里得范数等于 $\sqrt{n}$。因此，应用引理6.2并重新缩放，可得

$$\|A_\ell-B_\ell\|_2\geq\Big(1-\frac{1}{|I_j|}\Big)\sqrt{n}\geq\frac{\sqrt{n}}{2}.$$

由此可得

$$\|A-B\|_F^2=\sum_{\ell=1}^n\|A_\ell-B_\ell\|_2^2\geq\sum_{\ell\in T}\|A_\ell-B_\ell\|_2^2\geq|T|\left(\frac{\sqrt{n}}{2}\right)^2\geq\frac{n(n-k)}{4}.$$

□

[3]要说明这一点，可仿照定理3.1证明中的步骤 2 进行论证。

7. Application to differential privacy and synthetic data

本节简要介绍如何将本文的发现应用于改进 [2] 关于生成隐私合成数据的结果。推论 5.1 中协方差损失界背后的划分结果，以及定理5.3中的塞梅雷迪正则性界，为实现与 k-匿名性和差分隐私相关的数据隐私提供了一种自然的方法。虽然[4, 9] 中提出将塞梅雷迪正则性引理用作数据匿名化机制，但这些论文中的结果仅为经验性结果，并未提供任何理论保证。

本文的研究结果可直接用于改进 [2] 中 k-匿名性和差分隐私的数据隐私与效用保证。例如，使用推论 5.1 代替 [2]，中的定理 1.2，但其余步骤保持不变，我们便可得到以下定理。与 [2] 中的定理 5.14 相比，该定理给出了更优的准确性保证。具体细节留给读者，并请参阅 [2] ，以了解精确的问题陈述和相关概念的详细说明。

定理 7.1。设 K 是 $\mathbb{R}^p$ 中的一个凸集，且位于单位欧几里得球 B_2^p 内。设 $\alpha > 0$。若 n、$p \in \mathbb{N}$、$\varepsilon > 0$ 满足 $n > C(\alpha)\frac{p}{\varepsilon}$，则存在一个 ε-差分隐私算法，其输入为 x_1、……、$x_n \in K$，输出为 u_1、……、$u_{m(\alpha)} \in K$，使得

$$\mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^{n} x_i^{\otimes d} - \frac{1}{m(\alpha)}\sum_{i=1}^{m} u_i^{\otimes d} \right\|_2^2 \lesssim 16^d \alpha,$$

对所有 $d \in \mathbb{N}$ 均成立。该算法的运行时间为 $C(\alpha)(np+q)$，其中 q 是求出 K 中与给定向量 $\mathbb{R}^p$ 对应的最佳近似元素所需的复杂度。

Acknowledgement

M.B. 感谢美国国家科学基金会（NSF）DMS-2140592 项目的支持。T.S. 感谢美国国立卫生研究院（NIH）R01HL16351、美国国家科学基金会（NSF）DMS-2027248 和 NSF DMS-2208356 项目的支持。R.V. 感谢 NSF DMS-1954233、NSF DMS-2027299、美国陆军 76649-CS，以及 NSF+西蒙斯数学与科学深度学习基础研究合作项目的支持。

References

[1] Greg Bodwin 和 Santosh Vempala。通过矩阵分解统一理解图的正则性。 *Random Structures & Algorithms* 61 1 62 83, 2022 , ( ): – 。[2]

March Boedihardjo、Thomas Strohmer 和 Roman Vershyin。协方差损失即隐私收益：计算高效、保护隐私且准确的合成数据。 *Foundations of Computational Mathematics*，即将出版。[3] Amit Deshpande、Ravindran Kannan 和 Nikhil Srivastava。奇异向量的零一舍入。载于 *International Colloquium on Automata Languages and Programming*，第 278–289 页。施普林格，2012。[4] Daniele Foffano、Luca Rossi 和 Andrea Torsello。你看不见我：使用塞梅雷迪正则性引理对图进行匿名化。 *Frontiers in big Data*，2:7，2019。[5] Alan Frieze 和 Ravi Kannan。矩阵的快速近似及其应用。 *Combinatorica*，19 2 175 220，1999 ( ): – 。[6] Vishesh Jain、Ashwin Sah 和 Mehtaab Sawhney。协方差损失的最优最小化。*preprint arXiv:2205.01773* 2022 arXiv , 。

[7] J´anos Koml´os、Ali Shokoufandeh、Mikl´os Simonovits 和 Endre Szemer´edi。正则性引理及其在图论中的应用。 *Summer school on theoretical aspects of computer science* 84 112, , 第 – 页。[8] 2000 L´ ´ ´ ´ ´ *GAFA Geometric And Functional* aszlo *Analysis* 17 1 252 270, 2007 Lovasz 和 Balazs Szegedy。分析学家的塞梅雷迪引理。, ( ): – 。

[9] Giorgia Minello、Luca Rossi 和 Andrea Torsello。使用塞梅雷迪正则性引理实现图上的 k-匿名性。
IEEE Transactions on Network Science and Engineering，8(2):1283–1292，2020。[10] Endre Szemer´edi。
图的正则划分。载于 J.-C. Bermond、J.-C. Fournier、M. Las Vergnas 和 D. Sotteau 编，
Proc. Colloque Inter. CNRS，第 399–401 页，1978。[11] Roman Vershynin。
High-dimensional probability. An introduction with applications in data science。剑桥大学出版社，2018。

Department of Mathematics, ETH Zürich
Email address: march.boedihardjo@ifor.math.ethz.ch

Department of Mathematics, University of California, Davis and Center of Data Science and Artificial Intelligence Research, UC Davis*Email address*: strohmer@math.ucdavis.edu

Department of Mathematics, University of California, Irvine
Email address: rvershyn@uci.edu