

COVARIANCE LOSS, SZEMEREDI REGULARITY, AND DIFFERENTIAL PRIVACY

MARCH BOEDIHARDJO, THOMAS STROHMER, AND ROMAN VERSHYNIN

ABSTRACT. We show how randomized rounding based on Grothendieck’s identity can be used to prove a nearly tight bound on the covariance loss—the amount of covariance that is lost by taking conditional expectation. This result yields a new type of weak Szemerédi regularity lemma for positive semidefinite matrices and kernels. Moreover, it can be used to construct differentially private synthetic data.

1. INTRODUCTION

Consider a random vector X taking values in \mathbb{R}^d and a σ -algebra \mathcal{F} . Let Y denote the conditional expectation: $Y = \mathbb{E}[X|\mathcal{F}]$. In dimension $d = 1$, where X and Y are random variables, the law of total variance states that

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

Thus, taking conditional expectation underestimates the variance. This observation extends to higher dimensions. Namely, let $\Sigma_X = \mathbb{E}(X - \mathbb{E} X)(X - \mathbb{E} X)^\top$ denote the covariance matrix of X , and similarly for Σ_Y . Then

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

where \succeq denotes the Loewner order, in which $A \succeq B$ if $A - B$ is positive semidefinite. Just like in the one-dimensional case, we see that taking conditional expectation underestimates the covariance.

In [2], we asked the basic question: “How big is the *covariance loss* $\Sigma_X - \Sigma_Y$?” The answer will obviously depend on the choice of the sigma-algebra \mathcal{F} , prompting the next question: “What sigma-algebra \mathcal{F} of given complexity minimizes the covariance loss?”

It was shown in [2] that there exists a partition of the sample space into at most k sets such that for the sigma-algebra \mathcal{F} generated by this partition, the covariance loss (1.2) is upper bounded by $C\sqrt{\log \log k / \log k}$, where C is an absolute constant. In [6], Jain, Sah, and Sawhney were able to improve this bound to $C/\sqrt{\log k}$. It follows from Proposition 3.14 in [2] that this bound is optimal up to the value of the constant C . The proofs of the bounds in both [2] and [6] are somewhat technical. In this paper we give a new approach to this problem, which is conceptually related to Szemerédi regularity. This approach yields a tighter bound in terms of the constant on the covariance loss (Corollary 5.1), as well as a much shorter and more elementary proof.

The celebrated Szemerédi regularity lemma states that that for every large enough graph, the set of nodes can be divided into subsets of about the same size so that the edges between different subsets behave almost randomly [10]. Expressed in the language of linear algebra, it says that the adjacency matrix of a graph can be approximated by a sum of cut matrices¹. The Szemerédi regularity lemma, both in the combinatorial and linear-algebraic forms, has

¹A cut matrix is a rank-one matrix which is constant on a block and zero elsewhere [5].

many deep algorithmic and combinatorial applications, see e.g. [5, 8, 7, 1]. We prove a new version of (weak) Szemerédi regularity lemma, which states that any positive semidefinite kernel $K(x, y)$ whose diagonal is uniformly bounded can be decomposed into a sum of k step-functions plus an error term whose L^2 -norm is $O(1/\sqrt{\log k})$.

The outline of the paper is as follows. We will first bound the covariance loss by the “covariance increment” (Proposition 2.1), find a nice representation of the covariance increment (Proposition 2.3), and combine it with randomized rounding based on Grothendieck’s identity to bound the covariance increment (Theorem 3.1). Thus we can not only bound the covariance loss (Corollary 5.1) but also derive a new type of weak Szemerédi regularity for matrices (Theorem 5.3) and kernels (Theorem 5.6). We conclude by illustrating how our results can be used in connection with differential privacy and synthetic data to improve the accuracy bounds obtained in [2].

1.1. Notation. The subgaussian norm of a random variable X is denoted by $\|X\|_{\psi_2}$ (see e.g. [11]), $\mathbb{E}(X)$ is the expectation of X and $\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p}$. For $d \times d$ matrices A, B , define the inner product $\langle A, B \rangle = \text{Tr}(AB^T)$ and the Frobenius norm $\|A\|_F = \sqrt{\langle A, A \rangle}$.

2. THE COVARIANCE INCREMENT

We start out by bounding the covariance loss $\mathbb{E}XX^T - \mathbb{E}YY^T$ by the covariance increment.

In the sequel, X is a random vector taking values in \mathbb{R}^d , \mathcal{F} is a σ -algebra and $Y = \mathbb{E}[X|\mathcal{F}]$. Moreover, (X', \mathcal{F}', Y') is an independent copy of (X, \mathcal{F}, Y) so that the sample space is a product space for which X, \mathcal{F}, Y are based on the first component of the product space and X', \mathcal{F}', Y' are based on the second.

With a slight abuse of notation, the product σ -algebra $\mathcal{F} \times \mathcal{F}$ denotes the σ -algebra generated by \mathcal{F} and \mathcal{F}' . Moreover, the inner product $\langle Y, Y' \rangle = \mathbb{E}(\langle X, X' \rangle | \mathcal{F} \times \mathcal{F})$.

Proposition 2.1 (Covariance loss and covariance increment).

$$\left\| \mathbb{E}XX^T - \mathbb{E}YY^T \right\|_F^2 \leq \left\| \mathbb{E}XX^T \right\|_F^2 - \left\| \mathbb{E}YY^T \right\|_F^2. \quad (2.1)$$

The proof of Proposition 2.1 will be based on the following lemma:

Lemma 2.2. *Let (U, V) be a pair of random vectors taking values in \mathbb{R}^d , and let (U', V') be an independent copy. Then*

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

In particular, setting $V = U$, we have

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

Proof. The first identity readily follows if we first use the identical distribution, and then independence:

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

The lemma is proved. □

Proof of Proposition 2.1. Expanding the square of the Frobenius norm, we can express the left hand side of (2.1) as $\|\mathbb{E}XX^\top\|_F^2 - 2\langle\mathbb{E}XX^\top, \mathbb{E}YY^\top\rangle + \|\mathbb{E}YY^\top\|_F^2$. After simplification, we see that inequality (2.1) is equivalent to

$$\|\mathbb{E}YY^\top\|_F^2 \leq \langle\mathbb{E}XX^\top, \mathbb{E}YY^\top\rangle.$$

Using Lemma 2.2 we can rewrite this as

$$\mathbb{E}\langle Y, Y'\rangle^2 \leq \mathbb{E}\langle X, X'\rangle^2, \quad (2.2)$$

To check (2.2), recall that $Y = \mathbb{E}[X|\mathcal{F}]$ and apply the conditional Jensen's inequality. \square

The covariance increment has a nice representation, which will come handy in our further analysis.

Proposition 2.3.

$$\|\mathbb{E}XX^\top\|_F^2 - \|\mathbb{E}YY^\top\|_F^2 = \mathbb{E}\left(\langle X, X'\rangle - \langle Y, Y'\rangle\right)^2. \quad (2.3)$$

Proof. By Lemma 2.2, the left hand side of (2.3) equals $\mathbb{E}\langle X, X'\rangle^2 - \mathbb{E}\langle Y, Y'\rangle^2$. Note that $\langle Y, Y'\rangle = \mathbb{E}(\langle X, X'\rangle|\mathcal{F} \times \mathcal{F})$. To finish the proof, apply the law of total variance (1.1) for $\langle X, X'\rangle$ instead of X and $\langle Y, Y'\rangle$ instead of Y . \square

3. BOUNDING THE COVARIANCE INCREMENT

Theorem 3.1. *Let X be a random vector taking values \mathbb{R}^d such that $\|X\|_2 \leq 1$ a.s. Then, for any $r \in \mathbb{N}$, there exists a partition of the sample space into at most 2^r parts such that for the σ -algebra \mathcal{F} generated by this partition, the conditional expectation $Y = \mathbb{E}[X|\mathcal{F}]$ satisfies*

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

The proof of Theorem 3.1 will utilize Grothendieck's identity (see e.g. [11, Lemma 3.6.6]):

Lemma 3.2 (Grothendieck's identity). *Let x, x' be a pair of unit vectors in \mathbb{R}^d , and let $g \sim N(0, I_d)$. Then*

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

Proof of Theorem 3.1. Step 1. Let us first make a stronger assumption, namely that $\|X\|_2 = 1$ a.s., and prove a weaker conclusion, namely that there exists a random variable Z measurable with respect to the product σ -algebra $\mathcal{F} \times \mathcal{F}$ and such that

$$\|\langle X, X'\rangle - Z\|_{L^2} \leq \frac{\pi}{2\sqrt{r}}. \quad (3.1)$$

To this end, consider independent random vectors $g_1, \dots, g_k \in N(0, I_d)$, and for $x, x' \in \mathbb{R}^d$, define the random variable

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

Denoting $\xi_k = \text{sign}\langle x, g_k \rangle \text{sign}\langle x', g_k \rangle$, noting that the function $\sin(\cdot)$ is 1-Lipschitz and applying Lemma 3.2, we obtain

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

By independence, this yields

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

Substitute $x = X$, $x' = X'$ and take expectation with respect to X and X' . By Fubini theorem, there exists a realization of the random vectors g_1, \dots, g_r such that

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

Fix such a realization. Let $\mathcal{F} = \sigma(V(X))$ be the σ -algebra generated by the random vector $V(X) = (\text{sign}\langle x, g_k \rangle)_{k=1}^r$. Since $V(X)$ takes at most 2^r values, \mathcal{F} satisfies the requirement of the theorem. Moreover, the random vector

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

is measurable with respect to the product sigma-algebra $\mathcal{F} \times \mathcal{F}$. Thus, we proved (3.1).

Step 2: replacing Z with $\langle Y, Y' \rangle$. The conditional expectation $\mathbb{E}[\cdot | \mathcal{F} \times \mathcal{F}]$ is an orthogonal projection in L^2 onto the subspace of random variables that are $\mathcal{F} \times \mathcal{F}$ -measurable. Since the random variable Z constructed in the previous step is $\mathcal{F} \times \mathcal{F}$ -measurable, and $\langle Y, Y' \rangle = \mathbb{E}[\langle X, X' \rangle | \mathcal{F} \times \mathcal{F}]$, it follows from (3.1) that

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

Step 3: removing the unit norm requirement. We proved the theorem under the additional assumption that X is a unit random vector. Now let X be an arbitrary random vector satisfying $\|X\|_2 \leq 1$. Let $N \in \mathbb{N}$, and consider the random vector $W = \sqrt{1 - \|X\|_2^2} \cdot \theta$, where θ is a random vector that is uniformly distributed in the unit basis $\{e_1, \dots, e_N\}$ of \mathbb{R}^N . Then the direct sum $X \oplus W = (X_1, \dots, X_d, W_1, \dots, W_N)$ is a unit random vector in \mathbb{R}^{d+N} . Applying the previous step for $X \oplus W$, we find that

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

Since $\langle \eta, \eta' \rangle = \mathbb{E}(\langle W, W' \rangle | \mathcal{F} \times \mathcal{F})$, the law of total variance (1.1) and definition of W yields

$$\| \langle W, W' \rangle - \langle \eta, \eta' \rangle \|_{L^2} \leq \| \langle W, W' \rangle \|_{L^2} \leq \| \langle \theta, \theta' \rangle \|_{L^2} = \frac{1}{\sqrt{N}}, \quad (3.4)$$

where the last identity follows since $\langle \theta, \theta' \rangle$ is Bernoulli with parameter $1/N$. Taking N to be large enough, we conclude from (3.3) that

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

Theorem 3.1 is proved. \square

Remark 3.3. In the conclusion of Theorem 3.1, we can replace the constant π^2 by $\pi^2/4 + \epsilon$ for any fixed $\epsilon > 0$.

4. SUBGAUSSIAN ERROR

We can extend the bound in Theorem 3.1 to the subgaussian norm, which is the Orlicz norm with respect to the Young function $\psi_2(x) = e^{x^2} - 1$. Thus, a random variable X is subgaussian if

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

and this quantity is called the subgaussian norm of X , see [11, Sections 2.5, 2.7.1].

Theorem 4.1. *Let X be a random vector taking values \mathbb{R}^d such that $\|X\|_2 \leq 1$ a.s. Then, for any $r \in \mathbb{N}$, there exists a partition of the sample space into at most 2^r parts such that for the σ -algebra \mathcal{F} generated by this partition, the conditional expectation $Y = \mathbb{E}[X|\mathcal{F}]$ satisfies*

$$\|\langle X, X' \rangle - \langle Y, Y' \rangle\|_{\psi_2} \leq \frac{C}{\sqrt{r}}. \quad (4.2)$$

We can prove this result by modifying the proof of Theorem 3.1. Let us explain how to do this.

In Step 1, instead of using additivity of variance, we can use Hoeffding's inequality to control the sum of independent Bernoulli random variables ξ_k . Denoting $\zeta_k = \xi_k - \mathbb{E} \xi_k$, we have $\|\sum_{k=1}^r \zeta_k\|_{\psi_2} \lesssim \sqrt{r}$, see [11, Proposition 2.6.1]. Using this in (3.2), we get

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

By definition of the subgaussian norm, this means that

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

where $c > 0$ is some absolute constant. Substituting here $x = X$ and $x' = X'$ and applying the Fubini inequality as in Step 1, we obtain the following version of (3.1):

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

In Step 2, although the conditional expectation is not a metric projection in the subgaussian norm, it is an approximate metric projection:

Lemma 4.2. *Let X be a random variable and \mathcal{F} be a σ -algebra. Then, for any random variable Z that is \mathcal{F} -measurable, and any $p \geq 1$, we have*

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

Proof. Subtracting and adding Z and using triangle inequality, we get

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

Since Z is \mathcal{F} -measurable, we have

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

where the last step follows from definition of subgaussian norm (4.1) and conditional Jensen's inequality. Combine the two bounds to complete the proof. \square

Using this lemma for $\langle X, X' \rangle$ instead of X , we obtain in Step 2 that

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

In Step 3, we argue in a similar way about the subgaussian norm. The bound (3.4) becomes

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

Here in the first step we use Lemma 4.2 for $X = \langle W, W' \rangle$ and $Z = 0$, and the last step is a straightforward bound on the subgaussian norm of a Bernoulli random variable with parameter $1/N$. This completes the proof.

5. IMPLICATIONS

5.1. Covariance loss. Combining Proposition 2.1, Proposition 2.3 and Theorem 3.1, we obtain:

Corollary 5.1 (Covariance loss). *Let X be a random vector taking values \mathbb{R}^d such that $\|X\|_2 \leq 1$ a.s. Then, for any $r \in \mathbb{N}$, there exists a partition of the sample space into at most 2^r parts such that for the σ -algebra \mathcal{F} generated by this partition, the conditional expectation $Y = \mathbb{E}[X|\mathcal{F}]$ satisfies*

$$\left\| \mathbb{E} X X^\top - \mathbb{E} Y Y^\top \right\|_F \leq \frac{\pi}{\sqrt{r}}.$$

In particular, for any k with $\log_2 k \in \mathbb{N}$, there exists a partition of the sample space into at most k parts such that for the σ -algebra \mathcal{F} generated by this partition, we have

$$\left\| \mathbb{E} X X^\top - \mathbb{E} Y Y^\top \right\|_F \leq \frac{\pi}{\sqrt{\log_2 k}}. \quad (5.1)$$

Remark 5.2 (Optimality). Proposition 3.14 in [2] implies that there exists a random vector X taking values in the unit ball of \mathbb{R}^d such $\left\| \mathbb{E} X X^\top - \mathbb{E} Y Y^\top \right\|_F \geq \frac{\sqrt{\log_2(e)}}{80\sqrt{\log_2 k}}$ for any σ -algebra \mathcal{F} generated by a partition into at most k parts. Thus the bound in (5.1) is sharp up to an absolute constant.

5.2. Weak Szemerédi regularity.

Theorem 5.3 (Weak Szemerédi regularity). *Let A be an $n \times n$ positive semidefinite matrix such that $A_{ii} \leq 1$ for all i . Then, for any $r \in \mathbb{N}$, there exists a partition $[n] = I_1 \cup \dots \cup I_k$ with $k \leq 2^r$, and a matrix B that is constant on each block $I_i \times I_j$ and such that*

$$\frac{1}{n} \|A - B\|_F \leq \frac{\pi}{\sqrt{r}}. \quad (5.2)$$

Moreover, B can be computed by averaging the entries of A in each block.

Proof. By assumption, A can be represented as the Gram matrix of some vectors x_1, \dots, x_n satisfying $\|x_i\|_2 \leq 1$ for all i . Thus $A = [\langle x_i, x_j \rangle]_{i,j=1}^n$. Apply Theorem 3.1 for the random vector X that is uniformly distributed on $\{x_1, \dots, x_n\}$ to complete the proof. \square

Remark 5.4. Rewriting the approximation error in Theorem 5.3 as

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

we can interpret it as a bound on the *mean squared error* of the entries.

Remark 5.5. Applying Theorem 4.1 instead of Theorem 3.1, we can replace the Frobenius norm in (5.2) by stronger matrix norms, such as the ℓ_p norm of the entries.

Theorem 5.6 (Weak Szemerédi regularity, analytic form). *Let (Ω, μ) be a probability measure space. Let $K : \Omega \times \Omega \rightarrow \mathbb{R}$ be a measurable positive definite kernel such that $K(t, t) \leq 1$ for all $t \in \Omega$. Then for any $r \in \mathbb{N}$, there exists a partition $\Omega = I_1 \cup \dots \cup I_k$ with $k \leq 2^r$, and a function $L : \Omega \times \Omega \rightarrow \mathbb{R}$ that is constant on each block $I_i \times I_j$ such that*

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

Moreover, L can be computed by averaging K on each block.

Proof. When Ω is finite, using Theorem 3.1 (see also Remark 3.3), one can easily establish $\|K - L\|_{L^2(\mu \times \mu)} \leq \frac{\pi}{2\sqrt{r}} + \epsilon$ for any fixed $\epsilon > 0$. The general case follows by a standard approximation argument, which we leave to the reader. \square

5.3. Comparison to the existing work on weak Szemerédi regularity. Our result is more restrictive but stronger than the classical weak regularity lemma by Frieze and Kannan [5]. There the approximation error $R = K - L$ is measured in the *cut norm*

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

The cut norm is equivalent to the operator norm $L^1 \rightarrow L^\infty$ (see e.g. [8]) and clearly satisfies $\|R\|_{\square} \leq \|R\|_{L^2(\mu \times \mu)}$. So the Hilbert-Schmidt norm, which is the focus of the current paper, is *stronger* than the cut norm. As a result, our bound (5.3) automatically extends to the cut norm. A similar bound for the cut norm was established in the original work of Frieze and Kannan [5] in wider generality: it holds only for kernels but for an arbitrary measurable function $K(x, y)$ that is pointwise bounded by 1, see [8, Lemma 3.1]. In contrast to this, positive semidefiniteness is required for any nontrivial bound on the error in the stronger Hilbert-Schmidt norm, such as the one in Theorem 5.6. We will see this in the next section.

A matrix decomposition in the spirit of Theorem 5.3, i.e. with error bounded in the Frobenius norm, appears in [3, Theorem 7] by Deshpande, Kannan, and Srivastava. Their theorem, which does not require the matrix to be positive semidefinite, is only nontrivial for low-rank matrices A , namely for matrices whose rank is at most logarithmic² in the dimension n . Our results do not have a rank restriction.

6. OPTIMALITY

The decay rate $1/r$ in the conclusion of Theorem 3.1 is optimal, because the decay rate $1/\sqrt{r}$ in the conclusion of Corollary 5.1 is optimal.

All previous results in the literature on weak Szemerédi regularity do not need the assumption that the matrix A is positive semidefinite, see e.g. [5, 8, 1]. Can it be removed from our Theorem 5.3? The following result says that it cannot be removed.

²Theorem 7 in [3] approximates A by a sum of roughly $t = \text{rank}(A) \log^4 n$ cut-matrices; they can be further broken down into 2^t smaller cut-matrices with disjoint support. This ultimate decomposition is nontrivial only if $2^t < n^2$.

Proposition 6.1 (Positive semidefiniteness is essential). *Let A be an $n \times n$ Hadamard matrix, and let $[n] = I_1 \cup \dots \cup I_k$ be any partition. Then, for any matrix B that is constant on each block $I_i \times I_j$, we have*

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

Thus, the error does not vanish unless the number of parts k is extremely large, namely $k = n - o(n)$.

The proof will use the following elementary bound:

Lemma 6.2. *Let x_1, \dots, x_m are orthonormal vectors in \mathbb{R}^n . Then for each $k = 1, \dots, m$ we have*

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

Proof. The norm of the vector $x_k - \frac{1}{m} \sum_{t=1}^m x_t$ is bounded below by the inner product of that vector and x_k , which equals $1 - 1/m$. \square

Proof of Proposition 6.1. The minimum of $\|A - B\|_F$ over all matrices B in the proposition can only get smaller if we minimize over all matrices that are constant on smaller blocks, namely on the sets $\{\ell\} \times I_j$, where $\ell = 1, \dots, n$ and $j = 1, \dots, k$. The latter minimum is attained³ for the matrix B that is obtained by averaging the entries of A in each block $\{\ell\} \times I_j$. Equivalently, B is obtained by averaging the columns of A in each set I_j . Formally, the columns of B are

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

where B_ℓ is the ℓ column of B and A_t is the t th column of A . Note that there at least $n - k$ indices belong to the non-singleton blocks I_j , i.e.

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

(Indeed, since T^c is the union of the singleton blocks, its cardinality equals the number of such blocks I_j , which is bounded by the total number of blocks k .)

Pick any index $\ell \in T$, so $\ell \in I_j$ where $|I_j| \geq 2$. Since A is an Hadamard matrix, its columns A_t are orthogonal, and their Euclidean norms equal \sqrt{n} . Thus, applying Lemma 6.2 and rescaling, we have

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

It follows that

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

\square

³To see this, argue like in Step 2 of the proof of Theorem 3.1.

7. APPLICATION TO DIFFERENTIAL PRIVACY AND SYNTHETIC DATA

In this section we briefly describe how the findings in this paper can be applied to improve the results in [2] on generating private synthetic data. The partitioning result behind the covariance loss bound in Corollary 5.1 and the Szemerédi regularity bound in Theorem 5.3 provide a natural technique towards data privacy related to k -anonymity and differential privacy. While the Szemerédi regularity lemma has been proposed in [4, 9] as a mechanism for data anonymization, the results in those papers are merely empirical and come without any theoretical guarantees.

The findings in this paper can be readily used to improve upon the data privacy and utility guarantees for k -anonymity and differential privacy in [2]. For instance, using Corollary 5.1 in lieu of Theorem 1.2 in [2], but otherwise following the same procedure, we arrive at the following theorem which gives an improved accuracy guarantee compared to Theorem 5.14 in [2]. We leave the details to the reader and refer to [2] for the precise problem statement and a detailed description of the underlying concepts.

Theorem 7.1. *Let K be a convex set in \mathbb{R}^p that lies in the unit Euclidean ball B_2^p . Let $\alpha > 0$. If $n, p \in \mathbb{N}$, $\varepsilon > 0$ satisfy $n > C(\alpha)\frac{p}{\varepsilon}$, then there is an ε -differentially private algorithm with input $x_1, \dots, x_n \in K$ and output $u_1, \dots, u_{m(\alpha)} \in K$ such that*

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

for all $d \in \mathbb{N}$. The run time of the algorithm is $C(\alpha)(np + q)$, where q is the complexity to find a best approximation element of K to a given vector in \mathbb{R}^p .

ACKNOWLEDGEMENT

M.B. acknowledges support from NSF DMS-2140592. T.S. acknowledges support from NIH R01HL16351, NSF DMS-2027248, and NSF DMS-2208356. R.V. acknowledges support from NSF DMS-1954233, NSF DMS-2027299, U.S. Army 76649-CS, and NSF+Simons Research Collaborations on the Mathematical and Scientific Foundations of Deep Learning.

REFERENCES

- [1] Greg Bodwin and Santosh Vempala. A unified view of graph regularity via matrix decompositions. *Random Structures & Algorithms*, 61(1):62–83, 2022.
- [2] March Boedihardjo, Thomas Strohmer, and Roman Vershyn. Covariance’s Loss is Privacy’s Gain: Computationally Efficient, Private and Accurate Synthetic Data. *Foundations of Computational Mathematics*, to appear.
- [3] Amit Deshpande, Ravindran Kannan, and Nikhil Srivastava. Zero-one rounding of singular vectors. In *International Colloquium on Automata, Languages, and Programming*, pages 278–289. Springer, 2012.
- [4] Daniele Foffano, Luca Rossi, and Andrea Torsello. You can’t see me: Anonymizing graphs using the Szemerédi regularity lemma. *Frontiers in big Data*, 2:7, 2019.
- [5] Alan Frieze and Ravi Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.
- [6] Vishesh Jain, Ashwin Sah, and Mehtaab Sawhney. Optimal minimization of the covariance loss. *arXiv preprint arXiv:2205.01773*, 2022.
- [7] János Komlós, Ali Shokoufandeh, Miklós Simonovits, and Endre Szemerédi. The regularity lemma and its applications in graph theory. *Summer school on theoretical aspects of computer science*, pages 84–112, 2000.
- [8] László Lovász and Balázs Szegedy. Szemerédi’s lemma for the analyst. *GFA Geometric And Functional Analysis*, 17(1):252–270, 2007.

- [9] Giorgia Minello, Luca Rossi, and Andrea Torsello. k-anonymity on graphs using the Szemerédi regularity lemma. *IEEE Transactions on Network Science and Engineering*, 8(2):1283–1292, 2020.
- [10] Endre Szemerédi. Regular partitions of graphs. In J.-C. Bermond, J.-C. Fournier, M. Las Vergnas, and D. Sotteau, editors, *Proc. Colloque Inter. CNRS*, page 399–401, 1978.
- [11] Roman Vershynin. *High-dimensional probability. An introduction with applications in data science*. Cambridge University Press, 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`