Superlogarithmic Real Blocky Rank

Abstract

Boolean matrices can require more than logarithmically many blocky summands even when arbitrary real coefficients are allowed. We give a self-contained proof for a family obtained by polynomial evaluation over a binary finite field followed by binary Hadamard encoding. A distance argument, a count across disjoint degree intervals, and an integer-minor argument give real blocky rank of order at least $\log N \log\log N$ along the constructed sequence of square dimensions N. The Lean formalization proves the finite bound and its superlogarithmic consequence without an external graph theorem. The field basis is chosen noncomputably, so the result makes no claim about efficient entry computation.

1 Blocky decompositions and the result

A blocky summand combines disjoint all-one rectangles. The number of summands measures how many such patterns are needed to express a matrix by a real linear combination. The definition follows Avraham and Yehudayoff's On Blocky Ranks of Matrices. The coefficients may have either sign.

A zero-one matrix B is **blocky** if its ones are a union of rectangles $R_i \times C_i$ whose row sets are pairwise disjoint and whose column sets are pairwise disjoint. Zero rows and columns are allowed. For example, the matrix

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

is blocky. Its action sends x to $(x_1 + x_2, x_1 + x_2, x_4, 0)$. The **real blocky rank** of a real matrix A is the least $r \in \mathbb{N}$ for which real coefficients c_s and blocky matrices B_s satisfy

$$A = \sum_{s=1}^{r} c_s B_s.$$

A decomposition exists because each individual entry is a scalar multiple of a singleton rectangle.

Fix an integer $v \geq 20$ and define

$$t = 2^{2v}, \quad q = 2^t, \quad N = q^2, \quad m = t\frac{q}{2}.$$

All logarithms below have base two unless written ln.

Theorem 1.1 (Finite lower bound). There is a zero-one matrix $M_v \in \mathbb{R}^{N\times N}$ such that

$$\mathrm{br}_{\mathbb{R}}(M_v) > vt = \frac{\log_2 N(\log_2 \log_2 N - 1)}{4}.$$

The matrices M_v are the zero-column paddings of the generator defined in Section 3.

The proof uses a rectangular generator whose nonzero real inputs produce many nonzero outputs. Write $\mathrm{wt}(y)$ for the number of nonzero coordinates of y. The two ingredients are the following statements.

Lemma 1.2 (Distance forces blocky rank). Every real $N \times m$ matrix G satisfying

$$x \neq 0 \Rightarrow \mathrm{wt}(Gx) \geq \frac{N}{4}$$

has $\mathrm{br}_{\mathbb{R}}(G) > vt$.

Lemma 1.3 (Binary generator). A zero-one $N \times m$ matrix G satisfies the distance hypothesis of Lemma 1.2.

2 Counting across degree intervals

Proof. Suppose G has $r \leq R = vt$ blocky summands. Each rectangle first forms the sum of its input coordinates, called its middle value. Each output in that rectangle then uses that sum multiplied by the coefficient of its summand. Thus $G = PQ$ for suitable real matrices P, Q. Each input belongs to at most one rectangle in each summand, and each output belongs to at most one rectangle in each summand. There are at most mr connections from inputs to middle values and at most Nr connections from middle values to outputs. We count the latter connections even when a summand has coefficient zero.

The degree of a middle value is the number of output coordinates in its rectangle. Put

$$a = v2^v, \quad B = t^2, \quad C = B^2 = t^4.$$

For each $j \in \{0, ..., a-1\}$ use the scale $k = C^{j+1}$ and the interval

$$\left[\frac{N}{Bk}, N\frac{B}{k}\right).$$

A middle value is medium at this scale when its degree lies in this interval. It is high when its degree is at least $N\frac{B}{k}$, and low when its degree is below $\frac{N}{Bk}$. These classes depend on k. Consecutive intervals meet at an excluded upper endpoint because $N\frac{B}{Ck} = \frac{N}{Bk}$. Hence the medium classes at distinct scales are disjoint.

The parameter estimates proved in Section 4 give

$$B \geq 8R, \quad k \geq 16(a+1), \quad 2kC^a \leq m.$$

They also ensure that the interval endpoints and $\frac{k}{4}$ are positive integers.

Fix a set X of k inputs. There are at most $r\frac{k}{B} \leq \frac{k}{8}$ high middle values because each uses at least $N\frac{B}{k}$ output connections. Suppose X reaches at most $\frac{k}{4}$ medium values. Set all high middle

values and all medium values reached by X to zero. These are fewer than k homogeneous linear equations on the k inputs in X. Their common kernel contains a nonzero vector. Extend this vector by zero outside X.

The output of this vector can use only low middle values. The inputs in X reach at most kr middle values, and each low middle value reaches fewer than $\frac{N}{Bk}$ outputs. Thus the output has at most

$$kr\frac{N}{Bk} = N\frac{r}{B} \le \frac{N}{8} < \frac{N}{4}$$

nonzero coordinates. Overlap and cancellation can only reduce this support. The distance hypothesis gives a contradiction. Every k inputs therefore reach more than $\frac{k}{4}$ medium values.

Put $s = \frac{k}{4}$, and let U be the set of medium values. Counting output connections gives

$$|U| \le rBk \le C\frac{k}{8} \le C(s-a).$$

The last inequality follows from $k \ge 16(a+1)$, which implies $\frac{k}{8} \ge a$. A set of k inputs exists because the parameter inequalities give $k \le m$, so the preceding paragraph gives $|U| > s$.

Choose uniformly an s-element subset T of U. An input having $e \le a$ medium neighbors has all these neighbors in T with probability

$$\prod_{i=0}^{e-1} \frac{s-i}{|U|-i} \ge \left(\frac{s-a}{|U|}\right)^{e} \ge C^{-e} \ge C^{-a}.$$

Here $s > a$, and the product is one for $e = 0$. The product formula counts the subsets containing the input's neighborhood.

A fixed T contains the entire medium neighborhood of at most $k-1$ inputs. A set of k such inputs would reach at most s medium values, contrary to the expansion just proved. If L inputs have at most a medium neighbors, averaging the number captured by T gives

$$LC^{-a} \le k-1, \quad L < kC^{a} \le \frac{m}{2}.$$

This is a finite double count of pairs formed by an input and a subset containing its neighborhood. At least half the inputs have at least $a+1$ medium neighbors. Each scale therefore uses at least $\frac{m(a+1)}{2}$ input connections.

The a medium classes are disjoint, so the total number of input connections is at least $m\frac{a(a+1)}{2}$. Comparing this count to mr and cancelling $m > 0$ yields

$$r \ge \frac{a(a+1)}{2} > \frac{a^2}{2} = v^2\frac{t}{2} > vt = R.$$

This contradicts $r \le R$ and proves Lemma 1.2. □

3 A polynomial and Hadamard generator

Proof. Choose a field F of cardinality $q = 2^t$ and a basis $b_0, \dots, b_{t-1}$ over $\mathbb{F}_2$. Let $\mathrm{bit}_i(u)$ be the ith binary coordinate of $u \in F$. Index the rows by $(\alpha, \beta) \in F \times F$ and the columns by (j, s) for $0 \le j < \frac{q}{2}$ and $0 \le s < t$.

Define the zero-one entries by

$$G_{(\alpha,\beta),(j,s)} = \mathrm{bit}_0(\beta b_s \alpha^j).$$

This gives $q^2 = N$ rows and $(\frac{q}{2})t = m$ columns.

A binary input x determines coefficients $a_j = \sum_{s<t} x_{(j,s)} b_s$ and a polynomial

$$f(X) = \sum_{j<\frac{q}{2}} a_j X^j.$$

A nonzero x gives a nonzero f because the b_s form a basis. Linearity of the coordinate map shows that the output over $\mathbb{F}_2$ is

$$(\mathrm{bit}_0(\beta f(\alpha)))_{(\alpha,\beta)\in F\times F}.$$

A nonzero polynomial of degree below $\frac{q}{2}$ has fewer than $\frac{q}{2}$ roots, so $f(\alpha) \neq 0$ at more than $\frac{q}{2}$ field elements.

For each such α, multiplication by $f(\alpha)$ permutes F. Addition of b_0 pairs the field elements whose first bit is zero with those whose first bit is one. Thus exactly $\frac{q}{2}$ values of β give first bit one. Summing these disjoint row blocks gives

$$x \neq 0 \Rightarrow \mathrm{wt}(Gx) \geq \left(\frac{q}{2}\right)\left(\frac{q}{2}\right) = \frac{N}{4} \quad \text{over} \quad \mathbb{F}_2.$$

The map $u \mapsto (\mathrm{bit}_0(\beta u))_{\beta\in F}$ is the binary Hadamard encoding used here.

Regard G as an integer matrix and delete any set D of fewer than $\frac{N}{4}$ rows. The remaining matrix is injective over $\mathbb{F}_2$. Otherwise a nonzero vector in its kernel would have its original output supported on D, contradicting the binary distance. Full column rank gives an $m \times m$ minor whose determinant is nonzero modulo two. The determinant of the corresponding integer minor is odd. Its real determinant is therefore nonzero, so the remaining matrix is injective over $\mathbb{R}$.

If a nonzero real input had fewer than $\frac{N}{4}$ nonzero outputs, delete the support of that output. The remaining matrix would kill the input, contrary to its real injectivity. This proves the real distance required by Lemma 1.3. □

4 Parameter estimates and square padding

The parameter inequalities follow from two inequalities,

$$B \geq 8R \quad \text{and} \quad C^{2a} \leq q.$$

Indeed, $v \leq 2^v$ gives $R = v2^{2v} \leq 2^{3v}$, so

$$8R \leq 2^{3v+3} \leq 2^{4v} = B.$$

The inequality $16v^2 \leq 2^v$ holds at $v = 20$. It propagates by induction because $(v+1)^2 \leq 2v^2$ for $v \geq 3$. Consequently,

$$C^{2a} = 2^{16v^2 2^v} \leq 2^{2^{2v}} = q.$$

For every scale, $k \geq C$ and $k \leq C^a$. Since $a \leq R$ and $a \geq 1$,

$$k \geq B^2 \geq 64a^2 \geq 16(a+1).$$

The inequality $t \geq 4$ gives $2kC^a \leq 2C^{2a} \leq 2q \leq t\frac{q}{2} = m$. Also $B \leq C^a$, so $Bk \leq C^{2a} \leq q \leq N$. Both Bk and N are powers of two. Thus $\frac{N}{Bk}$ is a positive integer, and $N\frac{B}{k} = B^2\frac{N}{Bk}$ is an integer. Finally k is a power of two at least 16, so $\frac{k}{4}$ is an integer.

Proof. Take G from Lemma 1.3. The inequality $t \leq q$ gives $m = t\frac{q}{2} \leq q^2 = N$. Append $N - m$ zero columns to obtain M_v. Each blocky summand extends by zero columns. Restricting a blocky summand to the original columns also leaves a blocky matrix. Extension and restriction show $\mathrm{br}_{\mathbb{R}}(M_v) = \mathrm{br}_{\mathbb{R}}(G) > vt$ by Lemma 1.2.

The parameter identities give

$$\log_2 N = 2t, \quad \log_2 \log_2 N = 2v + 1.$$

Substitution yields the stated expression for vt. □

Corollary 4.1 (Superlogarithmic rank). The dimensions $N_v = 2^{2 \cdot 2^{2v}}$ tend to infinity and

$$\frac{\log_2 N_v}{\mathrm{br}_{\mathbb{R}}(M_v)} \to 0 \quad (v \to \infty).$$

Equivalently, $\log_2 N_v = o(\mathrm{br}_{\mathbb{R}}(M_v))$. The same statement holds for $\ln N_v$.

Proof. The powers defining N_v increase without bound. For $v \geq 20$, Theorem 1.1 gives

$$0 < \frac{\log_2 N_v}{\mathrm{br}_{\mathbb{R}}(M_v)} < \frac{2t}{vt} = \frac{2}{v}.$$

The upper bound tends to zero. Multiplication by the constant $\ln 2$ proves the assertion for natural logarithms. □

5 Formal scope

The bound also gives $\log_2 N_v < \mathrm{br}_{\mathbb{R}}(M_v)$ for every $v \geq 20$, and hence eventually as v tends to infinity. Indeed, $\log_2 N_v = 2t < vt < \mathrm{br}_{\mathbb{R}}(M_v)$ because $v \geq 20$. The threshold 20 proves the eventual assertion.

The statements concern the sequence of dimensions N_v. They do not assert a lower bound at every integer dimension. The construction uses a chosen basis of a finite field. The mathematical argument works for every binary basis. The formal construction chooses its basis noncomputably and does not specify an efficient executable representation or prove an entry-computation running time.

The Lean development formalizes the generator, its binary and real distance, the degree-interval count, padding, and the quantitative and asymptotic bounds. The argument assumes no graph expansion theorem or complexity conjecture. Lean and its mathematical library are described in the Lean project and the mathlib documentation.