

Gram Moments and Pairwise Mixing of AES

Chongxu Ren  and Hongbo Yu 

Department of Computer Science and Technology,
Tsinghua University, Beijing, China
rcx23@mails.tsinghua.edu.cn, yuhongbo@tsinghua.edu.cn

Abstract. We prove that $4k+4$ rounds of AES with independent uniform round keys are $2^{-14-61k}$ -close to pairwise independent, improving on Hurst’s announced bound of $2^{-3.82-48.03k}$. Here, closeness is measured by the total variation distance between the outputs on any two distinct plaintexts and those of a uniform random permutation, with the round keys sampled once and retained across evaluations. Consequently, 12, 16, and 20 rounds suffice for statistical accuracy 2^{-128} , 2^{-192} , and 2^{-256} , respectively. Our bound also holds for the complete view of any algorithm making at most two adaptive classical forward queries.

We develop a general bound for local difference channels acting on additive MDS codes, expressed through the second moment of a Gram matrix. This bound controls the dependencies created by diffusion and refines the activity-space analysis of previous work. Combining the resulting contraction estimate with a sharper analysis of the initial difference distributions yields the stated mixing bound.

Keywords: AES · Pairwise independence · Gram matrices · MDS codes · Spectral analysis

1 Introduction

Block ciphers aim to make encryption under a secret random key resemble a uniformly chosen permutation. For the Advanced Encryption Standard (AES), proving how this behavior emerges from byte substitution, linear diffusion, and key addition remains difficult. One approach is to bound statistical properties directly from these round transformations. It underlies the Markov-cipher theory of Lai, Massey, and Murphy [15], the differential bounds of Nyberg and Knudsen [21], and Vaudenay’s decorrelation theory [23]. More recently, Liu, Tessaro, and Vaikuntanathan studied almost pairwise independence for concrete substitution-permutation networks (SPNs), including AES with independent round keys [17].

From trails to complete differentials. The wide-trail strategy of Daemen and Rijmen bounds the probability of each AES differential trail under independent round keys by ensuring sufficiently many active S-boxes [6]. A trail specifies every intermediate difference; a complete differential specifies only the endpoints. Its probability sums over all connecting trails, so a bound on individual trails need not control the sum. Nyberg and Knudsen addressed this distinction for

DES-like constructions [21]. For AES, Keliher and Sui obtained exact two-round and improved four-round differential bounds [13]; Daemen and Rijmen explained how inversion, MixColumns, and trail multiplicities determine the two-round probabilities [7].

From individual probabilities to a distributional guarantee. For a random permutation on N points, every nonzero output difference has probability $1/(N-1)$ at a fixed nonzero input difference. A pointwise upper bound substantially above this value still permits considerable unevenness. Closeness to uniformity requires controlling the accumulated deviations over all outputs. We therefore ask:

How close is the joint ciphertext distribution of two distinct plaintexts to the uniform distribution on ordered pairs of distinct blocks?

Requiring this closeness for every distinct plaintext pair gives *almost pairwise independence*. It controls every event determined by the two ciphertexts, including events collecting many output differences. This distributional viewpoint connects decorrelation theory [23] with the recent analysis of concrete SPNs [17].

The independent-round-key model. We retain the actual AES transformations and sample one tuple of independent uniform round keys for all evaluations. This isolates the mixing of the round transformations and makes the difference process exactly Markov [15]. Our bounds average over the key tuple; they do not apply to each fixed tuple or to the standard key schedule. The importance of these distinctions is illustrated by Kranz, Leander, and Wiemer’s study of key schedules and linear correlations [14] and Beyne and Rijmen’s framework for fixed-key differential probabilities [4].

Recent work reduced the rounds sufficient for pairwise accuracy 2^{-128} from roughly nine thousand in Liu, Tessaro, and Vaikuntanathan [17] to forty in Dinur [8] and twenty in Beyne, Leander, and Schütt (BLS) [3]. Two complementary estimates drive this progress: smoothing from a fixed input difference and uniform contraction thereafter. We strengthen both by controlling the dependencies created by diffusion.

1.1 Our result and approach

Let $E_{r,\mathbf{K}}$ denote AES with r substitution layers, initial whitening, and no final MixColumns. Its round keys $\mathbf{K} = (K_0, \dots, K_r)$ are independent uniform 128-bit strings, sampled once and retained across evaluations. Section 2 gives the precise round convention.

Let P be a uniformly sampled permutation of $\{0, 1\}^{128}$. We measure pairwise distance by

$$\varepsilon_r = \max_{x \neq x'} \text{TV}(\mathcal{L}(E_{r,\mathbf{K}}(x), E_{r,\mathbf{K}}(x')), \mathcal{L}(P(x), P(x'))), \quad (1)$$

where $\mathcal{L}$ denotes the distribution and total variation is half the ℓ_1 distance. Explicitly, for $N = 2^{128}$,

$$\varepsilon_r = \frac{1}{2} \max_{\substack{x \neq x' \\ y \neq y'}} \sum_{\mathbf{K}} \left| \Pr[E_{r,\mathbf{K}}(x) = y, E_{r,\mathbf{K}}(x') = y'] - \frac{1}{N(N-1)} \right|.$$

The sum runs over ordered pairs of distinct ciphertexts. The probability is over one sampled key tuple shared by both evaluations.

Theorem 1 (Pairwise mixing of AES). *For all integers $k \geq 0$ and $r \geq 4 + 4k$,*

$$\varepsilon_r < 2^{-14-61k}. \quad (2)$$

Consequently, 12, 16, and 20 rounds give distances below 2^{-136} , 2^{-197} , and 2^{-258} , respectively. They therefore suffice for statistical accuracy 2^{-128} , 2^{-192} , and 2^{-256} . These are guarantees for AES with independent round keys.

Remark 1 (Adaptive choice of plaintexts). Consider a classical algorithm that requests the encryption of at most two plaintexts. It may choose the second plaintext after seeing the first ciphertext. The bound in Theorem 1 also applies to the total variation distance between its views under $E_{r,\mathbf{K}}$ and P .

The view consists of the chosen plaintexts and received ciphertexts, together with the algorithm’s private random bits and any information it had beforehand. Those random bits and prior information must be independent of the sampled keys or permutation. Repeated plaintexts and early stopping are allowed. Section 2.2 proves this consequence.

Table 1 compares the bounds at a common round index. Hurst’s entry records an announced estimate; the underlying paper is unpublished and unavailable to us [11].

Table 1. Pairwise AES bounds after $4 + 4k$ rounds. The final three columns give sufficient round counts for the indicated accuracy. The proved bounds concern the fixed AES transformations with independent round keys. D+B combines Dinur and BLS. Our bound is Theorem 1.

Sources	Bound	2^{-128}	2^{-192}	2^{-256}
Dinur [8]	$2^{-3.82-14.65k}$	40	56	76
BLS [3]	$2^{14-39.29k}$	20	28	32
D+B [8,3]	$2^{-3.82-39.29k}$	20	24	32
Hurst [11]	$2^{-3.82-48.03k}$	16	20	28
Ours	$2^{-14-61k}$	12	16	20

Two estimates with different roles. We first smooth a fixed input difference over four rounds, exploiting structure unavailable for an arbitrary distribution. We then contract every remaining deviation uniformly over further four-round blocks.

Group the sixteen bytes into four 32-bit words, and write $q = 2^{32}$. Let $p_{a,4}$ be the difference distribution after four rounds starting from a nonzero difference a , and let u be uniform on the $q^4 - 1$ nonzero state differences. In the grouped coordinates of Section 2.3, let $\mathcal{D}_4$ be the four-round transition matrix, whose column a is $p_{a,4}$. Write $Uf = (\sum_x f(x))u$ for the projection onto the uniform distribution.

Whitening reduces pairwise distance to difference-law distance, and Cauchy-Schwarz gives $\text{TV}(p, u) \leq \frac{1}{2} \sqrt{q^4 - 1} \|p - u\|_2$. Sections 5 and 6 establish the two estimates below; the right-hand bounds use $q = 2^{32}$:

$$\begin{aligned} \frac{1}{2} \sqrt{q^4 - 1} \max_{a \neq 0} \|p_{a,4} - u\|_2 &< \frac{1}{2} \sqrt{q^4 - 1} 8q^{-5/2} < 2^{-14}, \\ \|\mathcal{D}_4 - U\| &< 8q^{-2} = 2^{-61}. \end{aligned}$$

The first line measures initial smoothing, including the conversion to TV; the second is a Euclidean operator bound.

Each further block contracts zero-sum deviations by 2^{-61} . The extra permutation of difference states at block boundaries preserves the uniform law and Euclidean norm, so the scaled error satisfies

$$\varepsilon_{4+4k} < 2^{-14} (2^{-61})^k = 2^{-14-61k}.$$

Section 7 proves the iteration and boundary identities.

The dependence created by diffusion. Substitution acts locally; diffusion combines local outputs and creates dependencies even from independent inputs. Following BLS [3], we separate the probability of each zero/nonzero word pattern (*activity*) from variation among values within that pattern (*residual variation*). A word channel preserves activity and reduces residual variation, but diffusion transfers information between the two. Bounding this interaction is the main difficulty.

Keeping the overlaps between local responses. Let T be a local channel with activity averaging removed, and t_a its column for input difference a . Its Gram matrix $G_{a,b} = \langle t_a, t_b \rangle$ records response energies on the diagonal and overlaps off the diagonal. The second Gram moment retains their total squared overlap:

$$\tau^2 = \text{Tr}(G^2) = \sum_{a,b} |\langle t_a, t_b \rangle|^2.$$

We express activity functions as combinations of coordinate-subspace indicators: some words are unrestricted and the others fixed at zero. Diffusion maps these subspaces to additive MDS codes. Their coordinates are dependent, but suitable coordinate projections are bijective. We use these projections to bound averages of products of Gram entries while retaining the second moment.

Section 3 extends this code-indicator estimate to arbitrary activity functions, giving a coupling bound for general additive MDS layers.

Using the structure of a fixed input. For initial smoothing, the word outputs before diffusion are independent. With one active input word, the output words share one source; with several, they are sums of independent contributions. We separate individual-word deviations from joint correlations and bound the latter using the Gram estimate. This use of the starting law gives the initial factor 2^{-14} .

The local AES estimates are derived in Section 4. The analysis builds on the differential properties of inversion studied by Nyberg [20] and Daemen–Rijmen [7], the differential–linear Fourier relation of Chabaud and Vaudenay [5], and the two-round probability bound of Keliher and Sui [13] used in Section 6. These provide concrete information about the AES word channel for the general diffusion argument.

1.2 Relation to previous work

Complete differential probabilities. Keliher and Sui’s exact maximum expected differential probability is $53/2^{34}$ over two AES rounds, with four-round upper bound $(53/2^{34})^4$ [13]. This supplies a local estimate for our proof, but the four-round bound is too far above $1/(2^{128} - 1)$ to imply small pairwise TV distance. Daemen and Rijmen analyze the algebra and trail multiplicities behind the two-round probabilities [7].

Dinur places every nonzero eight-round differential probability within $2^{-128}/50$ of its ideal value, and proves a four-round collision estimate used in his pairwise-independence analysis [8]. These pointwise and distributional bounds describe different aspects of convergence.

Pairwise independence for concrete AES. Liu, Tessaro, and Vaikuntanathan describe substitution output differences through structured sets and prove an extraction property by Fourier analysis [17]. This yields arbitrarily small pairwise distance for the actual AES transformations under independent keys.

BLS strengthens the quantitative analysis by expressing pairwise mixing through difference-matrix norms [3]. Their activity decomposition combines the MDS property with local singular-value bounds; the AES superbox representation then gives an iterated block contraction.

Dinur’s analysis is especially effective for distributions produced by fixed input differences, whereas BLS provides a stronger uniform contraction. Their complementarity is already identified in these works. Combining Dinur’s four-round initialization with the numerical refinement in the corrected BLS full version gives

$$\varepsilon_{4+4k} < \frac{1}{2} \sqrt{\frac{N-1}{50N}} 2^{-39.29k} < 2^{-3.82-39.29k}, \quad N = 2^{128}. \quad (3)$$

The D+B row of Table 1 includes this combination.

Our proof inherits the initialization/contraction separation, BLS’s MDS-code expansions and Frobenius estimates, and Dinur’s Gram traces and MDS marginals.

We retain the second Gram moment inside the code average and convert it to an operator bound for activity functions, strengthening both estimates.

Activity patterns and operator methods. Rønjom constructed activity-pattern transition matrices for AES diffusion, but left their exact iterated agreement with AES unresolved [22]. Here activity remains part of the full distribution: residual variation and its coupling are retained. Beyne’s geometric framework studies cryptographic properties through function spaces and operators [2]; BLS supplies the direct activity-space framework refined here.

Random substitutions and other mixing models. Baignères and Vaudenay prove convergence of expected differential and linear probabilities for AES*, whose S-boxes are independent random secret permutations [1].

Liu, Pelecanos, Tessaro, and Vaikuntanathan use layout graphs to bound higher-order independence for SPNs with random secret S-boxes, including modified AES with many diffusion layers removed [16]. Jain et al. analyze the inverse walk’s convergence to a uniform permutation of the appropriate sign, with further consequences for modified AES [12]. These different constructions and observation models are outside the fixed-AES comparison in Table 1.

Pairwise independence does not imply pseudorandomness under arbitrarily many queries. Dujmovic, Pelecanos, and Tessaro show that even constant-order independence of iterated local permutations need not suffice [9]. Our theorem concerns at most two queries to independently keyed AES.

Organization. Section 2 introduces the AES model, the difference process, and the activity decomposition. Section 3 proves the general Gram-moment and coupling bounds. Section 4 derives the local AES estimates. Sections 5 and 6 establish four-round contraction and initial smoothing, respectively, and Section 7 combines them into Theorem 1 and discusses the conclusions. For a first reading of the quantitative argument, the local estimates stated in Section 4 can be taken as given before following their applications in Sections 5–7.

Substantial AI contribution. GPT Astra 6 contributed substantially to the mathematical development and produced the initial proof of the 12-, 16-, and 20-round bounds for AES with independent round keys. The human authors subsequently worked with GPT to revise and reorganize the proof, simplify its presentation, and make the argument understandable to cryptographic readers. In accordance with the IACR Generative AI Policy, we disclose these contributions and take full responsibility for the paper’s claims, proofs, citations, and overall integrity.

2 Preliminary

We describe the AES difference process, group rounds into superboxes, and separate word activity from variation among nonzero values.

Notations. We consider only discrete probability laws on finite sets and represent each law as a column vector. We use Euclidean norms with counting measure and write e_x for a coordinate unit vector. For a matrix A , A^* denotes its transpose or conjugate transpose, $\|A\|$ its Euclidean operator norm, and $\|A\|_{\text{HS}}^2 = \text{Tr}(A^*A)$ its squared Hilbert–Schmidt (Frobenius) norm, which is also the sum of the squared column norms. We write $\mathbf{1}_B$ for the indicator of a set or condition B , and $\mathbf{1}$ for the all-ones vector on the domain in question. For a probability law μ , its collision probability is $\text{Col}(\mu) = \sum_x \mu(x)^2 = \|\mu\|_2^2$. For probability laws μ and η on the same finite set Ω , their total variation distance is

$$\text{TV}(\mu, \eta) = \frac{1}{2} \|\mu - \eta\|_1 = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \eta(x)|.$$

2.1 The AES structure

AES uses a 4×4 byte state. SubBytes (SB) applies the AES S-box, ShiftRows (SR) cyclically shifts rows, and MixColumns (MC) mixes each column [19]. Addition is bitwise XOR. Figure 1 shows the states in our experiment; composition $\circ$ is read right to left:

$$\begin{aligned} z_0 &= x + K_0, \\ z_i &= (\text{MC} \circ \text{SR} \circ \text{SB})(z_{i-1}) + K_i \quad (1 \leq i < r), \\ E_r(x) &= (\text{SR} \circ \text{SB})(z_{r-1}) + K_r. \end{aligned} \tag{4}$$

The independent uniform 128-bit keys are retained across queries. Here z_0 follows input whitening, z_i follows round i , and $E_r(x)$ is the ciphertext, with final MixColumns omitted. Write $R_{K_i}(z) = (\text{MC} \circ \text{SR} \circ \text{SB})(z) + K_i$ for an ordinary round.

A word means one AES column of four bytes. We use

$$q = 2^{32}, \quad Q = q - 1, \quad N = q^4 = 2^{128}, \quad n = N - 1.$$

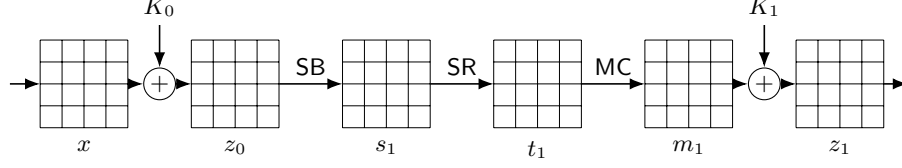
The byte alphabet is $\mathbb{F}_{256} = \mathbb{F}_2[t]/(t^8 + t^4 + t^3 + t + 1)$. MixColumns applies the matrix

$$M = \begin{pmatrix} 2 & 3 & 1 & 1 \\ 1 & 2 & 3 & 1 \\ 1 & 1 & 2 & 3 \\ 3 & 1 & 1 & 2 \end{pmatrix}, \quad \text{SR}(x)_{u,c} = x_{u,c+u},$$

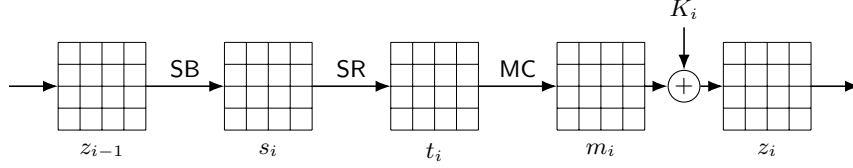
where row and column indices are modulo four and the matrix entries use the standard byte encoding. The byte S-box formula is recalled in Section 4.

The MDS property. AES diffusion follows the wide-trail strategy [6]. MixColumns is maximum distance separable (MDS): every square minor of M is invertible, so any four coordinates of (x, Mx) determine the pair. With w input bytes free and the rest zero, imposing $z < w$ output zeros leaves $w - z$ free byte values; $z \geq w$ forces the zero point. Imposing zeros is *shortening*; deleting coordinates is *puncturing* in coding theory [18]. We use these projection properties as in Dinur [8]; Section 2.3 proves their word-level counterparts.

Input whitening and round 1



Round i , $1 \leq i < r$



Round r : final round without MixColumns

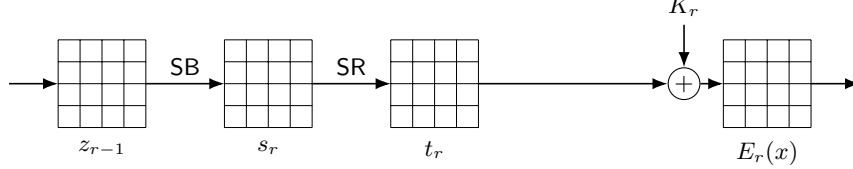


Fig. 1. AES with independent round keys. Rows show input whitening and round 1, an ordinary round i , and the final round without MixColumns. Each grid is a 4×4 byte state. Circled + denotes AddRoundKey (XOR).

2.2 Difference transitions and pairwise distance

Independent round keys make state differences a Markov chain: the next law depends only on the current difference. For a permutation s of a finite additive space V , define the normalized difference distribution table (DDT)

$$D_s(b, a) = |V|^{-1} \# \{x : s(x + a) - s(x) = b\}.$$

Its column a is the next-difference law when the first input is uniform and the difference is a . The zero difference is isolated, the nonzero block is doubly stochastic, and a change of variables gives $D_{s^{-1}} = D_s^*$.

Difference matrices need not multiply under composition without whitening. A fresh uniform key makes the first state uniform conditional on the current pair, while preserving its difference. Thus the next law depends only on that difference, and the matrices multiply [15,3]. Keys remain fixed across queries; this is an average over the retained tuple.

Distinct AES states have nonzero difference. We work on $\Omega = \mathbb{F}_{256}^{16} \setminus \{0\}$, of size $n = q^4 - 1$, with uniform law u .

Fix a round count r and write $E = E_r$. Independent input and output whitening give the exact pair identity

$$\Pr[E(x) = y, E(x + a) = y + b] = \frac{p_a(b)}{N}, \quad \varepsilon_r = \max_{a \neq 0} \frac{1}{2} \|p_a - u\|_1, \quad (5)$$

where p_a is the output-difference law for input difference $a \in \Omega$. Input whitening removes dependence on x ; output whitening makes the first ciphertext uniform independently of the difference. The ideal pair probability is $1/(N(N-1))$. Summing absolute deviations over y and $b \in \Omega$ proves the second identity [3].

This use of whitening to connect differences with two-query distinguishing also appears in Baignères and Vaudenay [1]. The same bound applies when the second query is adaptive. Condition on the adversary’s private randomness and auxiliary information, assumed independent of the oracle. Its first query is fixed, and the first answer is uniform in both experiments. Given that answer, its second query is fixed. Dividing the first identity in (5) by $1/N$ bounds the conditional second answer. A repeated query, or stopping after the first, gives identical views. Averaging yields the complete-view bound without an extra factor.

2.3 Superboxes and diffusion

We group two substitution layers using the AES *superbox* representation [7,8,3]. ShiftRows brings byte positions $(0, c), (1, c+1), (2, c+2), (3, c+3)$ into column c (indices modulo four). Figure 2 marks them for $c = 0$. Since SubBytes commutes with ShiftRows, moving ShiftRows before the first SubBytes keeps these four bytes together through SubBytes, MixColumns, key addition, and the next SubBytes. This 32-bit component is one superbox.

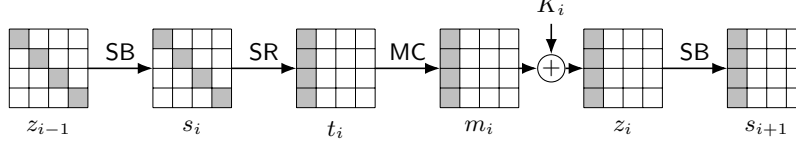
Diffusion between superboxes. Between two superbox layers, the remaining AES operations mix the four words through the linear map

$$L = \text{SR} \circ \text{MC} \circ \text{SR}, \quad (Lx)_{u,c} = \sum_v M_{u,v} x_{v,c+u+v}.$$

Columns of x are words. Every square coordinate-block minor of L is invertible: select equally many input and output words and group the equations by $c+u$, obtaining four invertible square minors of M . Thus any four word coordinates in (x, Lx) determine the pair, giving the shortening and projection properties of Section 2.1 at word level. Only additivity on $\mathbb{F}_{256}^4$ is needed, not scalar linearity over $\mathbb{F}_{2^{32}}$.

Each individual word block of L is a byte permutation followed by nonzero byte scalings, so it preserves the number of nonzero bytes. We call such a block *byte-monomial*; Section 6 uses this additional property to bound the initial difference law.

(a) AES order



(b) After commuting ShiftRows and SubBytes

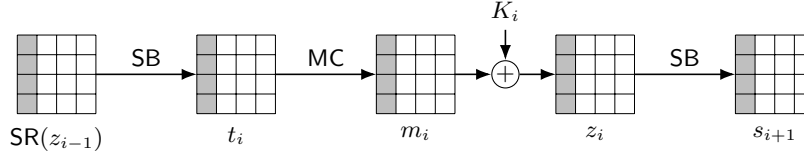


Fig. 2. One 32-bit superbox, with its four bytes shaded; unshaded bytes are arbitrary. Commuting ShiftRows and SubBytes turns (a) into the columnwise form (b). Averaging over keys gives the word difference channel K ; four parallel copies form H . Here $1 \leq i < r$.

Difference channels for grouped rounds. Let $K(b, a)$ be the probability that one superbox maps column difference a to b , averaged over its keys. Let H process four columns independently through K , and let Π apply L to differences: $\Pi e_a = e_{La}$. The word channel K includes all q differences. For state mixing, we restrict H and Π to the invariant nonzero space Ω .

Four rounds therefore consist of a superbox layer, linear mixing, and a second superbox layer, up to fixed input and output coordinate permutations. Writing $p_{a,4}$ for the resulting difference law from input difference a , we have

$$\mathcal{D}_4 = H\Pi H, \quad p_{a,4} = \mathcal{D}_4 e_a. \quad (6)$$

Sections 4–6 bound these channels; Section 7 extends the grouping to every even round count.

Compatibility with the AES round boundaries. It remains to justify the coordinate permutations used above. Put $B = \text{SB} \circ \text{MC} \circ \text{SB}$, the four parallel superboxes with key additions temporarily suppressed. Since SR commutes with parallel substitution, for every $k \geq 1$ we have

$$(\text{MC} \circ \text{SR} \circ \text{SB})^{\circ 2k} = \text{MC} \circ \text{SR} \circ B \circ (L \circ B)^{\circ (k-1)} \circ \text{SR}.$$

Here $f^{\circ n}$ denotes n -fold composition, with $f^{\circ 0}$ the identity. Moving keys through fixed invertible linear maps preserves their independent uniform law. Omitting the final MixColumns leaves only an outside ShiftRows. Thus the physical input difference a is relabeled as $\text{SR}(a)$, and a grouped output difference b corresponds to physical output difference $\text{SR}(b)$. These fixed relabelings preserve all norms and the worst input distance [7,8,3].

2.4 Activity and residual variation

An *activity pattern* specifies which words are nonzero: $(1, 0, 1, 0)$ means that only the first and third words are nonzero. A law specifies both the mass of each pattern and the values' distribution within it. We separate these as in BLS [3], using Dinur's exact zero/nonzero patterns [8].

The word channel preserves zero and the uniform law on nonzero differences. We therefore split functions of one word difference into zero, constant-nonzero, and zero-sum parts.

Regard a vector f as a function of a word difference x , including zero. The operator P_z keeps $f(0)$ and sets all other values to zero. The operator P_c sets the zero entry to zero and replaces each nonzero entry by the average $Q^{-1} \sum_{y \neq 0} f(y)$. Put

$$P = P_z + P_c, \quad P_e = I - P,$$

where I is the identity. For a probability law, P preserves its mass at zero and spreads the rest uniformly over nonzero words. The residual $P_e f = f - P f$ vanishes at zero and sums to zero. The three parts $P_z f, P_c f, P_e f$ are mutually orthogonal, and their operators are orthogonal projections. We write $\text{ran } A$ for the range of A ; thus $\text{ran } P_e$ is the *residual space*.

The word channel K fixes zero and preserves the uniform law on nonzero words, so $KP = PK = P$. Thus $R_0 = K - P = KP_e$ is its action on the residual part. A bound $\|R_0\| \leq \nu$ means that one word's residual variation satisfies

$$\|KP_e f\|_2 \leq \nu \|P_e f\|_2 \quad \text{for every } f.$$

Averaging within activity patterns. We now return to functions on Ω , the nonzero four-word differences. Let $P_0 f$ replace the values of f on each activity pattern by their average. For instance, on the pattern $(1, 0, 1, 0)$ it averages over all Q^2 choices of the two nonzero words. If f is a probability law, this preserves the probability of the pattern and spreads it uniformly over those choices. The remainder $f - P_0 f$ has sum zero within every pattern. Equivalently, P_0 applies the one-word averaging operator P separately at the four word positions.

Functions constant on each pattern form BLS's activity space [3]. Their zero/unrestricted-coordinate basis spans the same space: split each unrestricted coordinate into its zero and nonzero cases. Excluding the all-zero state leaves fifteen patterns and hence fifteen dimensions.

The uniform law u on nonzero state differences lies in the activity space. Let U keep the global uniform part: for a vector f on Ω , $Uf = (\sum_x f(x))u$. A *centered* vector has total sum zero; these vectors form the space $u^\perp$. Within this space, let $\mathcal{A}$ consist of the vectors constant on each pattern, and let $\mathcal{R}$ consist of the vectors whose sum is zero on every pattern. Their orthogonal projections are

$$u^\perp = \mathcal{A} \oplus \mathcal{R}, \quad P_{\mathcal{A}} = P_0 - U, \quad P_{\mathcal{R}} = I - P_0. \quad (7)$$

The $\mathcal{A}$ part describes excess or deficit probability in each pattern; the $\mathcal{R}$ part describes how probability varies within patterns. Centering removes one global

mean, leaving fourteen dimensions in $\mathcal{A}$. Individual pattern masses may still be nonzero, provided that their sum is zero. The parallel word channel H and its adjoint fix the activity part, since each word channel preserves zero and the uniform law on nonzero words.

Section 3.3 refines the residual part by the number of coordinates that vary. Section 6 introduces the separate uniform-reference decomposition needed for the starting law.

Remark 2 (Relation to prior-work norms). Let K_+ be the restriction of K to nonzero word differences, and list singular values σ_i and eigenvalues λ_i in decreasing order. Then

$$\|R_0\| = \sigma_3(K) = \sqrt{\lambda_2(K_+^* K_+)}, \quad \text{Tr}(R_0^* R_0) = \|K_+\|_{\text{HS}}^2 - 1.$$

Indeed, K acts as the identity on the zero and uniform-nonzero directions; the remaining action is R_0 . Restricting to K_+ removes the zero direction, and subtracting its uniform direction removes one from the squared Hilbert–Schmidt norm. Thus the residual norm used here is BLS’s third singular value, while its square is the second Gram eigenvalue used by Dinur [3,8]. Section 3.1 recalls the Gram-matrix notation.

3 Gram Moments on MDS Codes

Diffusion creates relations among word values that their activity pattern alone does not record. We retain these relations by expanding the squared output norm into overlaps between local responses, recorded in a Gram matrix. For contraction, we bound MDS-code inputs and then arbitrary activity inputs; for smoothing, we allow nonuniform common sources. Section 4 supplies the local parameters for AES.

3.1 Gram matrices and their moments

Let V be a binary vector space of size $q > 1$, with XOR as addition. For a matrix T acting on functions on V , the column Te_x is its response to input x . For AES, $T = K - P$, so this is the word-channel output minus its activity average, with entries of either sign. The general argument also allows complex entries.

The *Gram matrix* of T is the matrix of inner products between these columns:

$$G = T^* T, \quad G_{xy} = \langle Te_x, Te_y \rangle = \sum_{z \in V} \overline{T_{zx}} T_{zy}.$$

Thus $G_{xx} = \|Te_x\|_2^2$ measures one response and G_{xy} the overlap of two, using unnormalized inner products. The matrix is *positive semidefinite*: $f^* G f = \|Tf\|_2^2 \geq 0$ for every f .

The *trace* $\text{Tr } G = \sum_x G_{xx}$ sums the squared column sizes. We also retain the sum of the squared overlaps, called the second Gram moment. Our two local parameters are

$$\tau^2 = \text{Tr}(G^2) = \sum_{x,y \in V} |G_{xy}|^2, \quad \max_x G_{xx} \leq \delta. \quad (8)$$

Here δ bounds one squared column size, whereas τ^2 retains all pairs, allowing us to average dependent inputs before taking a worst case. The equality uses $G_{yx} = \overline{G_{xy}}$.

The operator norm $\|T\|$ is the largest amplification of Euclidean length: $\|Tf\|_2 \leq \|T\| \|f\|_2$ for every f . If $\|T\| \leq \nu$ and $\text{Tr } G \leq h$, then

$$\tau^2 = \sum_x \|T^* T e_x\|_2^2 \leq \nu^2 \sum_x \|T e_x\|_2^2 = \nu^2 \text{Tr } G \leq \nu^2 h. \quad (9)$$

Thus norm and trace certify the second moment; a sharper direct estimate of τ may also be used below.

3.2 The MDS code bound

An additive $[\ell, d]$ MDS code is a subgroup $\mathcal{C} \subseteq V^\ell$ of size q^d whose projection onto any d coordinates is bijective: any d coordinate values determine exactly one codeword. Diffusing a coordinate subspace with d free words produces such a code (Section 2.3). We first bound the response to its uniform input. For $d = 1$, each codeword is $(L_1 x, \dots, L_\ell x)$ for bijections L_i : all coordinates share the same source x .

The unit indicator $c_{\mathcal{C}} = q^{-d/2} \mathbf{1}_{\mathcal{C}}$ has Euclidean norm one; a probability law would instead use q^{-d} . Write $T^{\otimes \ell}$ for the separate action on the coordinates, so $(T^{\otimes \ell})_{y,x} = \prod_i T_{y_i, x_i}$ for $x, y \in V^\ell$. We bound its squared response $\|T^{\otimes \ell} c_{\mathcal{C}}\|^2$.

For $d > \ell/2$, a Fourier change of basis exchanges the code with its dual. Fix a nondegenerate symmetric binary bilinear pairing $[\omega, x]$ on V , such as the dot product, and set $\mathcal{F}_{\omega, x} = q^{-1/2} (-1)^{[\omega, x]}$. These normalized parity functions are orthonormal, so $\mathcal{F} = \mathcal{F}^* = \mathcal{F}^{-1}$ preserves lengths. The matrix $\mathcal{F} T \mathcal{F}$ represents T in this basis; its columns need the same bound only when $d > \ell/2$.

Proposition 1 (Gram-moment bound for an MDS-code input). *Let V be a finite vector space over $\mathbb{F}_2$ with $q = |V| > 1$. Let T be a matrix indexed by V , and set $\tau = \sqrt{\text{Tr}((T^* T)^2)}$, the square root of its second Gram moment. Let $\mathcal{C} \subseteq V^\ell$ be an additive $[\ell, d]$ MDS code with $0 < d < \ell$, and put $a = \min(d, \ell - d)$. Suppose $\|T e_x\|_2^2 \leq \delta$ for every $x \in V$. If $d > \ell/2$, also suppose $\|\mathcal{F} T \mathcal{F} e_x\|_2^2 \leq \delta$ for every x . Then*

$$\|T^{\otimes \ell} c_{\mathcal{C}}\|^2 \leq \left(\frac{\tau^2}{q} \right)^a \delta^{\ell - 2a}. \quad (10)$$

Two dependent coordinates. For $\ell = 2, d = 1$, the code has the form (L_1x, L_2x) for two bijections L_1, L_2 . Its coordinates share one source. Expanding the squared response and applying Cauchy–Schwarz gives

$$\begin{aligned} \|(T \otimes T)c_{\mathcal{C}}\|_2^2 &\leq q^{-1} \sum_{x,y} |G_{L_1x, L_1y}| |G_{L_2x, L_2y}| \\ &\leq q^{-1} \sum_{u,v} |G_{uv}|^2 = \tau^2/q. \end{aligned}$$

The bijections only relabel the Gram sums. Thus the bound is $\tau^2/q \leq \nu^2 h/q$, improving on the norm-only bound ν^4 when $h < q\nu^2$. Larger codes retain such an average in each of a independent directions.

Proof of the code bound. We control the product of Gram entries by applying Hölder’s inequality to the bijective coordinate projections.

Lemma 1 (Averaging over an MDS code). *Let $\mathcal{C} \subseteq V^\ell$ project bijectively onto every set of d coordinates, where $1 \leq d \leq \ell$. For nonnegative functions $h_i : V \rightarrow \mathbb{R}$ and $p = \ell/d$,*

$$\sum_{c \in \mathcal{C}} \prod_{i=1}^{\ell} h_i(c_i) \leq \prod_{i=1}^{\ell} \left(\sum_{x \in V} h_i(x)^p \right)^{1/p}. \quad (11)$$

Proof. Let $\mathcal{B}$ contain all d -subsets of $\{1, \dots, \ell\}$. Each coordinate belongs to a fraction $d/\ell = 1/p$ of these sets, so

$$\prod_i h_i(c_i) = \prod_{B \in \mathcal{B}} \left(\prod_{i \in B} h_i(c_i)^p \right)^{1/|\mathcal{B}|}.$$

Hölder bounds the sum by the geometric mean, over B , of $\sum_{c \in \mathcal{C}} \prod_{i \in B} h_i(c_i)^p$. Bijectivity of the B -projection factors this last sum into $\prod_{i \in B} \sum_x h_i(x)^p$. Counting each coordinate’s occurrences gives (11). $\square$

Proof (of Proposition 1). The case $d \leq \ell/2$. Put $p = \ell/d \geq 2$. The definition of the Gram entries gives

$$\|T^{\otimes \ell} c_{\mathcal{C}}\|^2 = q^{-d} \sum_{x,y \in \mathcal{C}} \prod_{i=1}^{\ell} G_{x_i, y_i}.$$

Apply Lemma 1 to $\mathcal{C} \times \mathcal{C}$, viewed as a code over $V \times V$, with $h_i(u, v) = |G_{uv}|$. It gives

$$\sum_{x,y \in \mathcal{C}} \prod_i |G_{x_i, y_i}| \leq \left(\sum_{u,v \in V} |G_{uv}|^p \right)^{1/p}.$$

Cauchy–Schwarz on the two columns gives $|G_{uv}| \leq \sqrt{G_{uu}G_{vv}} \leq \delta$. Therefore

$$\|T^{\otimes \ell} c_{\mathcal{C}}\|^2 \leq q^{-d} (\delta^{p-2} \tau^2)^d,$$

which is (10).

The case $d > \ell/2$. Define the dual code

$$\mathcal{C}^\perp = \{\omega \in V^\ell : \sum_i [\omega_i, x_i] = 0 \text{ for every } x \in \mathcal{C}\}.$$

Summing a parity function over the subgroup $\mathcal{C}$ gives zero unless it is identically one there. Consequently $\mathcal{F}^{\otimes \ell} c_{\mathcal{C}} = c_{\mathcal{C}^\perp}$, where $|\mathcal{C}^\perp| = q^{\ell-d}$. The dual is an additive $[\ell, \ell-d]$ MDS code: a dual codeword vanishing on $\ell-d$ coordinates is supported on d coordinates, and the bijective projection of $\mathcal{C}$ there lets us vary those d entries independently. Nondegeneracy of the pairing then forces every entry of the dual codeword to be zero. Each projection of the dual onto $\ell-d$ coordinates is therefore injective, and its size makes it bijective. Since $\mathcal{F}$ preserves lengths,

$$\|T^{\otimes \ell} c_{\mathcal{C}}\| = \|(\mathcal{F}T\mathcal{F})^{\otimes \ell} c_{\mathcal{C}^\perp}\|.$$

Apply the first case to $\mathcal{F}T\mathcal{F}$ and this dual code. Its Gram matrix is $\mathcal{F}G\mathcal{F}$, whose second moment is $\text{Tr}((\mathcal{F}G\mathcal{F})^2) = \text{Tr}(G^2) = \tau^2$. $\square$

The form used for diffusion. In particular, when $\delta \leq q^{-1}$ and $q\tau^2 \leq 1$,

$$\|T^{\otimes \ell} c_{\mathcal{C}}\| \leq \tau q^{-(\ell-1)/2} \quad (0 < d < \ell). \quad (12)$$

Indeed, the bound in (10) equals $\tau^2 q^{-(\ell-1)} (q\tau^2)^{a-1}$ after setting $\delta = q^{-1}$. Since $a \geq 1$, the last factor is at most one. This uniform bound lets us combine activity patterns.

3.3 From code indicators to the activity space

Contraction requires arbitrary activity inputs, including signed deviations from uniformity. We express them as linear combinations of indicators that diffuse to MDS codes. We separate the response by its number of residual coordinates and account for the indicators' nonorthogonality.

The argument applies to any number $\ell \geq 2$ of parallel coordinates; AES uses four words. On V , define P as in Section 2.4: preserve $f(0)$ and replace the values $f(x)$, $x \neq 0$, by their common average. Let $P_0 = P^{\otimes \ell}$ average within each activity pattern, and let $P_{\mathcal{R}} = I - P_0$ keep the residual part. Here we include the all-zero pattern in the activity space; it contributes no residual response. Write $H = K^{\otimes \ell}$ for the parallel action of K on these coordinates.

We use the finer decomposition W_j of BLS [3] to distinguish how many coordinates contribute residual variation. Consider a product function $f(x_1, \dots, x_\ell) = f_1(x_1) \cdots f_\ell(x_\ell)$. Choose j coordinate positions where f_i lies in the one-coordinate residual space: $f_i(0) = 0$ and $\sum_{x \neq 0} f_i(x) = 0$. At every other position, let f_i lie in $\text{ran } P$, so it is constant on the nonzero values and may have a separate

value at zero. The span of these products over all choices of the j positions is W_j ; we call j the *residual degree*. For example, on the four-word pattern $(1, 0, 1, 0)$, a zero-sum fluctuation in the first word alone is in W_1 , whereas a product of zero-sum fluctuations in the first and third words is in W_2 .

The one-coordinate split into $\text{ran } P$ and its orthogonal complement gives mutually orthogonal spaces W_j , with $W_0 = \text{ran } P_0$ and $\mathcal{R} = W_1 \oplus \cdots \oplus W_\ell$. If $\|K - P\| \leq \nu$, each residual coordinate contributes a factor at most ν , while the other coordinates are fixed. Since K and K^* preserve this split, orthogonal choices of coordinate positions remain orthogonal after the parallel layer. Hence $\|Hf\|_2 \leq \nu^j \|f\|_2$ for $f \in W_j$.

The choices of residual positions and nonorthogonal spanning inputs cost the following factor:

$$C_\ell(q)^2 = \frac{2^\ell - 2}{(1 - q^{-1/2})^\ell} \sum_{j=2}^{\ell} \binom{\ell}{j} \left(\frac{2}{1 - q^{-1/2}} \right)^{\ell-j}.$$

For general ℓ , an additive ℓ -coordinate MDS map L is one for which any ℓ coordinates of an input–output pair (x, Lx) determine x .

Theorem 2 (Activity–residual coupling). *Let V be a binary vector space of size $q > 1$ and let $\ell \geq 2$. Let K isolate zero and be doubly stochastic on $V \setminus \{0\}$. Set $T = K - P$ and $\tau^2 = \text{Tr}((T^*T)^2)$, with H , P_0 , and $P_{\mathcal{R}}$ acting on V^ℓ as above. Let Π be induced by an additive ℓ -coordinate MDS map. Suppose*

$$\max_x \|Te_x\|_2^2 \leq q^{-1}, \quad \max_x \|\mathcal{F}T\mathcal{F}e_x\|_2^2 \leq q^{-1}, \quad q\tau^2 \leq 1.$$

Then

$$\|P_{\mathcal{R}}H\Pi P_0\| \leq C_\ell(q)\tau q^{-(\ell-1)/2}. \quad (13)$$

The same bound holds for $\|P_0\Pi H P_{\mathcal{R}}\|$ if the column bounds also hold for T^* and $\mathcal{F}T^*\mathcal{F}$.

These bounds control transfer from activity to residual variation and back. The word layer fixes activity vectors. The proof first treats spanning indicators, then their linear combinations.

Proof. 1. *Diffuse a spanning input.* Write $Q = q - 1$ and $\xi = q^{-1/2}$. For $B \subseteq \{1, \dots, \ell\}$, let V^B contain the tuples that are zero outside B and unrestricted inside B . Then $b_B = q^{-|B|/2} \mathbf{1}_{V^B}$ and $c_B = \Pi b_B$ are unit indicators; c_B is supported on an MDS code of dimension $|B|$. Coordinates in B may themselves be zero.

2. *Bound one residual component.* Fix an output component: apply P_e at $j \geq 1$ positions, P_z at z positions, and P_c at the remaining $u = \ell - j - z$ positions. The word layer acts as $T = K - P$ at the residual positions and fixes the zero and constant parts. Thus only the j residual positions require a code estimate. Each P_c projection takes an inner product with the normalized nonzero indicator, introducing a nonzero condition on that coordinate. Expand it as $\mathbf{1}_{x \neq 0} = 1 - \mathbf{1}_{x=0}$.

A term choosing t additional zeros leaves $d = |B| - z - t$ free symbols, unless the conditions force the zero codeword. There are three cases:

$$\begin{aligned} d \leq 0 & \quad \text{the zero point is killed by } Te_0 = 0, \\ d \geq j & \quad \text{a uniform residual marginal is killed by } T\mathbf{1} = 0, \\ 0 < d < j & \quad \text{the residual coordinates form a } [j, d] \text{ MDS code.} \end{aligned}$$

For a surviving term, projection onto the j residual positions is injective. Replacing its indicator by a unit indicator changes the original factor $q^{-|B|/2}$ by $q^{d/2}$. Each of the u normalized constant factors contributes $Q^{-1/2}$, giving

$$\underbrace{q^{-(z+t)/2}}_{\text{unit normalization}} \underbrace{Q^{-u/2}}_{\text{constant factors}}.$$

For example, take $\ell = 4$, $|B| = 2$, and $(j, z, u) = (2, 1, 1)$. The $t = 0$ term is a $[2, 1]$ code with coefficient $q^{-1/2}Q^{-1/2}$; the $t = 1$ term vanishes. Its response is at most $\tau/(q\sqrt{Q})$ by (12).

There are $\binom{u}{t}$ choices of additional zeros, each contributing ξ^t to the bound. The triangle inequality sums these factors to $(1 + \xi)^u$. Together with (12), this bounds the component norm by

$$\tau q^{-(j-1)/2} q^{-z/2} Q^{-u/2} (1 + \xi)^u.$$

A term survives only if $0 < d < j$. In particular, degree one vanishes without a numerical estimate. This is the shortened-code cancellation of BLS [3]; the refinement here is the Gram-moment bound on the surviving terms.

3. *Sum the orthogonal components.* There are $\binom{\ell}{j}$ choices of residual positions. Summing the squared bounds over all zero/constant choices on the other positions gives

$$\begin{aligned} \|P_{\mathcal{R}} H c_B\|_2^2 &\leq \tau^2 q^{-(\ell-1)} \sum_{j=2}^{\ell} \binom{\ell}{j} \left[1 + \frac{q}{Q} (1 + \xi)^2 \right]^{\ell-j} \\ &= \tau^2 q^{-(\ell-1)} B_{\ell}(q), \quad B_{\ell}(q) := \sum_{j=2}^{\ell} \binom{\ell}{j} \left(\frac{2}{1 - \xi} \right)^{\ell-j}. \end{aligned}$$

The last equality uses $1 + (q/Q)(1 + \xi)^2 = 2/(1 - \xi)$.

4. *Pass from indicators to arbitrary activity inputs.* Let V_0 have the 2^ℓ vectors b_B as columns. They span activity: in each coordinate, an exact nonzero indicator is unrestricted minus zero. In one coordinate, the two unit indicators are e_0 and $q^{-1/2}\mathbf{1}$, whose inner product is ξ . Their Gram matrix and its product across coordinates give the bound used by BLS [3]:

$$V_0^* V_0 = \begin{pmatrix} 1 & \xi \\ \xi & 1 \end{pmatrix}^{\otimes \ell}, \quad \|V_0 a\|_2 \geq (1 - \xi)^{\ell/2} \|a\|_2. \quad (14)$$

Indeed, the one-coordinate eigenvalues are $1 \pm \xi$ and multiply under tensor products. The empty and full B give the all-zero and uniform unit indicators,

respectively; both have zero residual response. For $f = V_0 a$, Cauchy–Schwarz bounds the output norm by $\|a\|_2$ times the square root of the sum of the remaining $2^\ell - 2$ squared column norms. Dividing by the lower bound on $\|V_0 a\|_2$ in (14) yields

$$\|P_{\mathcal{R}} H \Pi P_0\|^2 \leq \frac{(2^\ell - 2) B_\ell(q)}{(1 - \xi)^\ell} \tau^2 q^{-(\ell-1)}.$$

This is the stated constant. For the reverse bound, apply the forward argument to K^* and the inverse MDS map, and take adjoints. $\square$

For AES, $\ell = 4$. With $\xi = q^{-1/2}$, the sum in the definition of $C_4(q)^2$ is $B_4(q) = 24/(1 - \xi)^2 + 8/(1 - \xi) + 1$. The activity basis and code expansions are inherited from BLS [3]; the Gram moment is the extra local information retained in this estimate.

3.4 Nonuniform common sources

In Section 6, a nonuniform random word drives several output coordinates. Expanding the squared norm of the weighted sum of local product responses again gives Gram products, now with source weights.

The *Schur product* is entrywise multiplication, $(A \circ B)_{xy} = A_{xy} B_{xy}$. Write $A \succeq 0$ for positive semidefiniteness and $A \preceq B$ for $B - A \succeq 0$. The Schur product theorem [10] preserves positive semidefiniteness under this product, yielding the following bound.

Lemma 2 (Weighted Schur bound). *Let $j \geq 2$. If $G_1, \dots, G_j \succeq 0$ and $G_1 \preceq \nu^2 I$, then for every vector μ indexed by their columns,*

$$\mu^*(G_1 \circ \dots \circ G_j) \mu \leq \nu^2 \sum_x |\mu(x)|^2 \prod_{i=2}^j (G_i)_{xx}. \quad (15)$$

Proof. Schur multiplication by $G_2 \circ \dots \circ G_j \succeq 0$ preserves positive semidefinite order. Apply it to $G_1 \preceq \nu^2 I$. $\square$

For a shared source, we combine one operator-norm bound, one trace bound, and the remaining diagonal bounds:

Corollary 1 (Shared source with bounded point probabilities). *Let $A_i : \mathbb{C}^V \rightarrow \mathcal{H}_i$ be matrices mapping into finite-dimensional inner-product spaces, with Gram matrices $G_i = A_i^* A_i$. Suppose $\|A_i\| \leq \nu$, $\text{Tr } G_i \leq h$, and $(G_i)_{xx} \leq \delta$ for every i, x . If $j \geq 2$ and $\|\mu\|_\infty \leq p$, then*

$$\left\| \sum_x \mu(x) \bigotimes_{i=1}^j A_i e_x \right\|^2 \leq \nu^2 p^2 h \delta^{j-2}. \quad (16)$$

Here $\|\mu\|_\infty = \max_x |\mu(x)|$; for a probability law, p bounds any one source probability. The product vector $\bigotimes_i A_i e_x$ uses the same source value x in every coordinate. Thus the displayed sum represents a common source, not independent sources at the outputs. Signed or complex weights are also allowed.

Proof. The squared norm is $\mu^*(G_1 \circ \dots \circ G_j)\mu$. Apply Lemma 2, bound $|\mu(x)|^2$ by p^2 and all diagonals except that of G_2 by δ , then sum the remaining diagonal to obtain $\text{Tr } G_2 \leq h$. $\square$

4 The Local AES Word Channel

We need three estimates for the word residual channel $R_0 = K - P$: its operator norm bounds variation from an arbitrary unit input; its trace $\text{Tr}(R_0^* R_0)$ sums residual-column energies; and an individual-column bound controls each fixed input. Norm and trace certify the second Gram moment, while the column bound supplies the pointwise hypothesis in Section 3.

We derive these estimates from the byte S-box and MixColumns, first bounding one byte, then the word norm, trace, and individual columns. A byte has 256 values and a word has $q = 2^{32}$. The results are collected below.

Lemma 3 (Local AES residual bounds). *Let $q = 2^{32}$ and $R_0 = K - P$, and let $\mathcal{F}$ be the normalized additive Fourier transform on words. For every $T \in \{R_0, R_0^*, \mathcal{F}R_0\mathcal{F}, \mathcal{F}R_0^*\mathcal{F}\}$,*

$$\begin{aligned} \|T\| &\leq \nu = 11/\sqrt{q}, & \text{Tr}(T^*T) &\leq h = 2^{-18}, \\ \|Te_x\|_2^2 &< \delta := \frac{6}{5q} - \frac{1}{q-1} < \frac{1}{5q} & \text{for every word difference } x. \end{aligned} \quad (17)$$

Norm and trace are unchanged by adjoints and Fourier conjugation; column sizes need separate arguments in each orientation. Fourier bounds support code duality in Proposition 1, and adjoints bound the reverse coupling. Sections 4.2–4.4 establish the three word estimates after the byte calculation.

4.1 The estimates for one byte

The byte estimates used throughout this section are

$$s = 9/128, \quad r_b = 65/8192, \quad p_b = 1/64. \quad (18)$$

Here s bounds residual contraction, r_b the collision of a nonzero-input law, and p_b its largest probability. The first controls arbitrary vectors; the others control fixed inputs. We derive them in both bases.

All three come from the inversion underlying the AES S-box. In the byte encoding of Section 2, AES uses

$$s_{\text{AES}}(x) = v \oplus \text{rot}_1(v) \oplus \dots \oplus \text{rot}_4(v) \oplus \mathbf{63}_{16}, \quad v = x^{-1}, \quad 0^{-1} = 0,$$

where rot_j is cyclic left rotation by j bits [19]. The lemma below first counts the possible output differences of inversion. The resulting formula depends only on the product of the input and output differences. This makes a multiplicative Fourier basis suitable for bounding the norm. The invertible affine output map then

transfers the bounds to AES. Nyberg [20] established the differential uniformity of inversion; Daemen and Rijmen [7] give its difference counts in the AES analysis. The character-sum argument follows BLS [3]; their stated lemma assumes odd n , whereas AES uses $n = 8$. We give the calculation for both parities below.

Lemma 4 (Residual contraction of inversion). *Let $m = 2^n$ with $n \geq 1$, and let D_{inv} be the difference channel of inversion on $\mathbb{F}_m$, with $0^{-1} = 0$. Let P_b project onto the zero and uniform-nonzero directions. Then*

$$\|D_{\text{inv}} - P_b\| \leq \frac{\sqrt{m} + 2}{m}.$$

For even n , each row and column of the nonzero block has one entry $4/m$, $m/2 - 2$ entries $2/m$, and $m/2$ zeros. Both conclusions persist after any invertible affine transformation of the output.

Proof. Counting differences. Let $\text{tr}(x) = x + x^2 + \dots + x^{2^{n-1}}$ be the absolute field trace from $\mathbb{F}_m$ to $\mathbb{F}_2$. This field trace is distinct from the matrix trace Tr used for Gram moments. For nonzero differences a, b , put $c = ab$ and write the first input as at . Outside $t \in \{0, 1\}$, the difference equation is $t^2 + t = c^{-1}$. The map $t \mapsto t^2 + t$ has kernel $\{0, 1\}$ and image equal to the trace-zero subspace. Indeed, trace vanishes on its image, and both spaces have size $m/2$: the trace polynomial is nonzero of degree $m/2 < m$. The two excluded inputs both give $c = 1$. Hence

$$D_{\text{inv}}(b, a) = d(ab), \quad d(c) = \frac{2}{m} (\mathbf{1}_{\text{tr}(c^{-1})=0} + \mathbf{1}_{c=1}).$$

For even n , $\text{tr}(1) = 0$; counting the nonzero trace-zero elements gives the stated histogram.

Bounding residual variation. Write $G = \mathbb{F}_m^*$ for the nonzero field elements under multiplication. A multiplicative character is a function $\chi : G \rightarrow \mathbb{C}$ with $|\chi(x)| = 1$ and $\chi(xy) = \chi(x)\chi(y)$. These are the Fourier modes of the multiplicative group; dividing their vectors by $\sqrt{m-1}$ gives an orthonormal basis. The constant character describes the uniform direction, and all other characters have sum zero. A change of variable $c = ab$ shows that D_{inv} sends the χ vector to λ_χ times the $\bar{\chi}$ vector, where $\lambda_\chi = \sum_c d(c)\chi(c)$. Thus it suffices to bound these scalar multipliers. The constant character has $\lambda_1 = 1$. For nonconstant χ , put $\psi(t) = (-1)^{\text{tr}(t)}$ and $g(\chi) = \sum_{t \in G} \chi(t)\psi(t)$. Then

$$\lambda_\chi = \frac{g(\bar{\chi}) + 2}{m}, \quad |g(\chi)|^2 = \sum_{t \in G} \chi(t) \sum_{v \in G} \psi((t-1)v) = m.$$

For the last equality, the inner sum is $m-1$ at $t=1$ and -1 otherwise, while $\sum_t \chi(t) = 0$. Thus $|\lambda_\chi| \leq (\sqrt{m} + 2)/m$. Since character conjugation permutes an orthonormal basis, this bounds the residual operator norm. At $m=2$ the residual space is zero.

Transferring the bound to an affine output map. If the output is transformed by $x \mapsto Ax + c$ with A invertible and $\mathbb{F}_2$ -linear, the difference channel becomes $\Pi_A D_{\text{inv}}$. The permutation Π_A relabels output differences, preserving lengths and the two directions selected by P_b . This proves both invariances. $\square$

The AES linear part is $I + R + R^2 + R^3 + R^4$, where R rotates a byte by one bit. It is invertible: its polynomial has value one at $X = 1$, so it is coprime to $X^8 - 1 = (X + 1)^8$, and $R^8 = I$. Lemma 4 at $m = 256$ gives $s = 9/128$. The histogram gives $r_b = (4^2 + 126 \cdot 2^2)/256^2$ for every nonzero row or column, and $p_b = 4/256$. This proves (18), the standard AES byte parameters reported by Dinur [8].

The same bounds in the Fourier basis. The Fourier orientation is needed for the dual-code argument in Proposition 1. This is the additive Fourier transform on byte differences, rather than the multiplicative basis used inside the inversion proof. Write $\mathcal{F}_b$ for its matrix, with entry $256^{-1/2}(-1)^{\text{tr}(uv)}$ at (u, v) . Define the ordinary Walsh table

$$W_{uv} = \sum_{x \in \mathbb{F}_{256}} (-1)^{\text{tr}(ux + v \text{SAES}(x))}.$$

The differential-linear Fourier relation of Chabaud and Vaudenay [5] expresses the transformed DDT through squared Walsh coefficients. Expanding the DDT over pairs $(x, x + a)$ gives, with the corresponding index convention,

$$(\mathcal{F}_b D \mathcal{F}_b)_{v,u} = W_{uv}^2 / 256^2 =: \hat{D}_{v,u}.$$

Indeed, the sum over the input difference changes variables to two independent byte inputs, and the two character sums multiply. Thus $\hat{D}$ is nonnegative and fixes the same zero and uniform-nonzero directions as D . For every row with $u \neq 0$ and every column with $v \neq 0$, direct finite summation over the other index yields

$$|W_{uv}| \leq 32, \quad \sum W_{uv}^2 = 256^2, \quad \sum W_{uv}^4 = 65 \cdot 2^{19}. \quad (19)$$

Consequently $D, D^*, \hat{D}, \hat{D}^*$ all have collision bound r_b , probability bound p_b , and residual norm s . Norms are preserved by the orthogonal Fourier change of basis; the first two properties follow from the displayed sums. Only this squared-Walsh collision is required. No Fourier coefficient bound for an individual DDT column is used in the starting-law proof. We can therefore use the same three byte parameters in all four word-channel orientations of Lemma 3.

4.2 The word norm: combining four byte channels

SubBytes contracts variation among nonzero byte values but preserves byte activity. MixColumns transfers information between these two parts. We use the residual degrees of Section 3.3: a degree- i vector varies in i byte positions and carries activity in the rest. Substitution contracts it by s^i . Bounding each transfer through MixColumns, with substitution on both sides, gives a five-by-five comparison matrix.

Let F_i project onto byte degree i . The sixteen-dimensional $\text{ran } F_0$ records all byte patterns, whereas P keeps only the zero and uniform-nonzero word directions. Thus the degree-zero block here is restricted to $\text{ran}(F_0 - P)$, of dimension fourteen. A nonzero word can have fifteen byte patterns: SubBytes preserves their masses, leaving MixColumns to mix them.

The five-by-five comparison. Rows and columns of the following matrix are indexed by the output and input residual degrees 0, 1, 2, 3, 4. Each entry of C_b bounds a transfer through MixColumns; multiplying it by s^{i+j} includes the two substitution layers:

$$C_b = \begin{pmatrix} 1/10400 & 0 & 7/500 & 9/50 & 1 \\ 0 & 27/8000 & 43/500 & 1 & 1 \\ 7/500 & 43/500 & 9/10 & 1 & 1 \\ 9/50 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}, \quad B_{ij} = s^{i+j}(C_b)_{ij}. \quad (20)$$

The activity count (23), the Gram calculation (26), and the incidence test (28) will establish these entries below. The zero entries express the absence of transfer between activity and degree one; entries equal to one use only that diffusion preserves Euclidean length.

Once these transfer bounds are established, decompose an input in the word residual space into its five orthogonal degree components and collect their lengths in a vector. The triangle inequality bounds the length in output degree i by row i of B applied to this vector. Squaring and summing over output degrees gives $\|R_0\| \leq \|B\|$. This is the residual-degree method of BLS [3].

The final matrix check is short. For the positive vector $v = (13, 3, 7, 6, 2)^\top$, rational multiplication gives $Bv < \nu v$. Scaling coordinate i by v_i makes the row sums of the resulting similar matrix equal to $(Bv)_i/v_i < \nu$. Its eigenvalues have absolute value below ν , and symmetry of B then gives $\|B\| < \nu = 11/\sqrt{q}$. We now justify the transfer bounds used in this comparison.

Separating substitution from diffusion. For this calculation, write $A = D^{\otimes 4}$ for the four independent byte channels in one substitution layer, and let $\Pi_M e_a = e_{Ma}$ describe MixColumns. The two substitution layers give $K = A\Pi_M A$ by the independent-key argument of Section 2.2. To separate substitution from diffusion, put $D_s = \sum_i s^i F_i$: this auxiliary operator multiplies every degree- i component by its known contraction bound s^i . It is not a probability transition matrix. Since A preserves degrees, $T_s = D_s^{-1}A = AD_s^{-1}$ has norm at most one and fixes the two invariant word directions. Factoring A in this way on both sides of MixColumns gives

$$R_0 = T_s W T_s, \quad W = D_s(\Pi_M - P)D_s. \quad (21)$$

The outside factors cannot increase length, so $\|R_0\| \leq \|W\|$. Each diffusion block from degree j to i is weighted by s^{i+j} . We bound the three types of block—activity to activity, activity to residual, and residual to residual—by activity counts, Gram matrices, and incidence matrices, respectively.

For these counts, let L_0 be an additive four-coordinate MDS map on $\mathcal{X}^4$, where $m = |\mathcal{X}|$. Here $\mathcal{X} = \mathbb{F}_{256}$, $m = 256$, and $L_0 = M$. Section 5 will reuse the activity and incidence formulas with $\mathcal{X} = V$, $m = q$, and $L_0 = L$. The activity and degree projections are interpreted on the chosen alphabet.

Activity to activity. First retain only the activity pattern before and after diffusion; this is what we mean by the *activity compression*. An exact support specifies which coordinates are nonzero. A graph point is a compatible input–output pair (x, L_0x) for the mixing map L_0 . For fixed input and output supports of sizes b and a , respectively, the number of graph points with exactly those supports depends only on $r = a + b$:

$$N_r(m) = \sum_{t=0}^r (-1)^{r-t} \binom{r}{t} m^{\max(t-4, 0)}.$$

To obtain this count, choose which of the r positions are allowed to vary and use the MDS property to count the remaining solutions. Inclusion–exclusion then removes solutions with unwanted zeros. These are the graph-intersection counts of BLS [3]. In particular, the normalized indicators of two input patterns of the same weight have the same averaged activity response. Their difference disappears under the compression. We need only the four nonzero weights 1, 2, 3, 4, rather than all fifteen patterns. In coordinates recording the total mass of each weight class, the matrix is

$$C_{ab} = \binom{4}{a} \frac{N_{a+b}(m)}{(m-1)^b}, \quad 1 \leq a, b \leq 4.$$

The weight classes have different sizes. Normalizing their indicators to length one changes this to the symmetric matrix

$$\hat{C}_{ab} = \sqrt{\binom{4}{a} \binom{4}{b}} \frac{N_{a+b}(m)}{(m-1)^{(a+b)/2}}.$$

It is similar to C : use the diagonal entries $\sqrt{\binom{4}{a} (m-1)^a}$. Thus the eigenvalues of C give the Euclidean norm of the symmetric matrix $\hat{C}$: this norm is its largest absolute eigenvalue. Expanding the four-by-four characteristic polynomial gives

$$(\lambda - 1) \left(\lambda - \frac{4}{(m-1)^2} \right) \left(\lambda^2 + \frac{10(m+1)}{(m-1)^3} \lambda + \frac{24}{(m-1)^4} \right). \quad (22)$$

The eigenvalue one describes the uniform nonzero law, which is removed in R_0 . The other three eigenvalues give the exact centered norm

$$\alpha(m) = \frac{5(m+1) + \sqrt{m^2 + 98m + 1}}{(m-1)^3} < \frac{6m + 54}{(m-1)^3}. \quad (23)$$

The inequality uses $\sqrt{m^2 + 98m + 1} < m + 49$. In particular, $\alpha(256) < 1/10400$ and $\alpha(q) < 7/q^2$. BLS already evaluate this centered activity compression numerically [3]; the displayed formulas evaluate the same operator symbolically and expose its $6/m^2$ scale.

Activity to residual variation. Next consider the residual variation created from an activity input. For $A \subseteq \{1, 2, 3, 4\}$, let $\mathcal{X}^A$ consist of tuples zero outside A and unrestricted inside A . The sixteen unit indicators $b_A = m^{-|A|/2} \mathbf{1}_{\mathcal{X}^A}$ span the activity space, as in Section 3.3. We use these overlapping subspaces because their constraints are linear equations. Their nonorthogonality is recorded by $G^{\text{act}} = V_0^* V_0$, where V_0 has columns b_A . Equation (14), with q replaced by m , gives $G_{AB}^{\text{act}} = m^{|A \cap B| - (|A| + |B|)/2}$.

For $j \geq 1$, both fixed word directions have zero response in degree j , so we may calculate on the full activity space. Apply L_0 to each b_A and retain degree j . The resulting vectors $F_j \Pi_{L_0} b_A$ have Gram matrix

$$\Gamma_j = V_0^* \Pi_{L_0}^* F_j \Pi_{L_0} V_0 \quad (j \geq 1).$$

For coefficients v , the quadratic form $v^* \Gamma_j v$ is the squared output norm, while $v^* G^{\text{act}} v$ is the squared input norm. Consequently, an activity-to-degree bound t_j follows if $t_j^2 G^{\text{act}} - \Gamma_j \succ 0$. The symbol $\succ 0$ means positive definite: its quadratic form is positive for every nonzero coefficient vector. This reduces the desired operator bound to a sixteen-by-sixteen matrix test.

To evaluate these entries, we count solutions to the linear conditions expressing zero or equal coordinates. For a matrix A over $\mathbb{F}_{256}$ with d columns, the basic rule is

$$\#\{x : Ax = 0\} = 256^{d - \text{rank } A}. \quad (24)$$

For example, one nonzero equation on two byte variables has 256 solutions. Products of constraint indicators are evaluated by stacking their equations. The counts are multiplied by their signed rational coefficients; no sampling is involved.

The projections can all be expressed using such conditions. On an alphabet of size m , the matrix entry, or *kernel*, of the zero/constant-nonzero projection is

$$p_m(x, y) = \frac{1 - \mathbf{1}_{x=0} - \mathbf{1}_{y=0} + m \mathbf{1}_{x=y=0}}{m - 1}. \quad (25)$$

It is one at $(0, 0)$, zero when exactly one argument is zero, and $1/(m - 1)$ when both are nonzero. Its complement is the residual projection $e_m = I - p_m$, where $I(x, y) = \mathbf{1}_{x=y}$. Use a formal variable z to mark a residual choice: multiplying four factors $p_m + ze_m$ and taking the coefficient $[z^j]$ selects exactly j residual positions. Thus

$$(\Gamma_j)_{AB} = [z^j] m^{-(|A| + |B|)/2} \sum_{x \in \mathcal{X}^A, y \in \mathcal{X}^B} \prod_{\ell=1}^4 (p_m + ze_m)((L_0 x)_\ell, (L_0 y)_\ell). \quad (26)$$

Expanding $p_m + ze_m = zI + (1 - z)p_m$ leaves only five kinds of term: equality, no constraint, left zero, right zero, and both zero. Each product of these terms is counted by (24). Byte conditions are already linear equations over $\mathbb{F}_{256}$; a word condition becomes four byte equations. This supplies every entry of Γ_j and hence the matrix test above without constructing the full diffusion matrix.

Residual to residual variation. It remains to bound transfers for which both input and output have residual variation. We count compatible values on their residual positions, keeping track of the remaining activity positions. BLS bound these transfers by shortening the MDS graph, that is, by prescribing zero coordinates [3]. We retain the exact support sizes in that argument to obtain the specific finite comparison below. The geometry is inherited; the numerical enclosures needed for our word norm are established here rather than inferred from BLS's less precise final norm. Fix an output support A of size a and choose i of its positions to carry residual variation. Similarly, fix an input support B of size b and choose j residual positions. We call these positions *marked*. Each other active position carries the constant vector of value $(m-1)^{-1/2}$ on nonzero symbols. Expand its nonzero restriction as unrestricted minus zero. Each choice in this expansion adds zero conditions to the MDS graph. The count also applies when one residual degree is zero; that case will justify the activity/degree-one zeros in (20).

A term in this expansion is a shortened MDS graph. It vanishes after residual projection if it has no free symbol, or if it has at least $i+j$ free symbols: in the latter case the marked coordinates have a uniform joint marginal. Only $1 \leq d < i+j$ free symbols can contribute. For each such graph, form a matrix indexed by assignments to the marked output and input positions. An entry is one when the two assignments extend to a graph point, and zero otherwise; this is its *incidence matrix*. There is at most one extension because the $i+j$ marked coordinates determine a graph point when $d < i+j$. Its rows and columns contain at most $m^{d-\min(i,d)}$ and $m^{d-\min(j,d)}$ ones, respectively. Its norm is at most the square root of the product of these counts, since a matrix satisfies $\|A\|^2 \leq \|A\|_1 \|A\|_\infty$, where the two norms on the right are the largest absolute column sum and row sum. Further orthogonal projections do not increase the norm.

There are $a+b-i-j$ unmarked nonzero coordinates. Selecting the unrestricted ones gives the binomial coefficient in

$$b_{ij}(a, b) = (m-1)^{-(a+b-i-j)/2} \sum_{d=1}^{i+j-1} \binom{a+b-i-j}{a+b-4-d} m^{d-(\min(i,d)+\min(j,d))/2}. \quad (27)$$

Invalid binomial coefficients are zero. The prefactor is the normalization of unmarked constants, the binomial counts the shortenings, and the final power is the incidence norm. For $i+j=1$ the sum is empty, explaining the vanishing of activity/degree-one blocks without a numerical check.

For example, take one residual coordinate on each side, so $i=j=1$. Only the term $d=1$ can survive. At the smallest possible support total $a+b=5$, the formula reduces to

$$b_{11}(a, b) = (m-1)^{-3/2}.$$

The three unmarked coordinates supply this normalization. Larger supports are accounted for by the same binomial sum. This example explains the $m^{-3/2}$ scale of the first residual interaction used below.

For a fixed pair of sizes (i, a) , there are $n_i(a) = \binom{4}{i} \binom{4-i}{a-i}$ orthogonal components: choose the i residual positions, then the $a - i$ other active positions. Combining the bounds for these components gives the at-most-four-by-four matrix

$$X_{ij}(a, b) = \sqrt{n_i(a)n_j(b)} b_{ij}(a, b), \quad i \leq a \leq 4, \quad j \leq b \leq 4.$$

Its norm bounds the entire transfer from degree j to degree i . This follows by applying the triangle inequality to the norms of the input components and then Cauchy–Schwarz over supports of each size. The individual transfers need not be equal. Every block also has the bound one, since diffusion preserves length and projections cannot increase it.

We certify a proposed smaller bound $c > 0$ using a rational matrix test. Write $X_{ij} = D_o^{1/2} R D_i^{1/2}$ with

$$D_o(a) = n_i(a)/(m-1)^{a-i}, \quad D_i(b) = n_j(b)/(m-1)^{b-j},$$

where R_{ab} is the sum in (27). Then the rational test

$$c^2 D_o^{-1} - R D_i R^* \succ 0 \tag{28}$$

proves $\|X_{ij}\| < c$: conjugating it by $D_o^{1/2}$ gives $c^2 I - X_{ij} X_{ij}^* \succ 0$. The powers of $\sqrt{m}$ are integers in both AES instances, so these tests use rational entries. Positive definiteness is checked by exact rational elimination (Section 7.2).

Verifying the entries. At $m = 256$, the Gram matrices computed by (26) satisfy $t_j^2 G^{\text{act}} - \Gamma_j \succ 0$ for both M and M^{-1} , with $t_2 = 7/500$ and $t_3 = 9/50$. The test for M bounds activity-to-residual transfer; the test for M^{-1} , by taking adjoints, bounds the reverse transfer. The incidence tests in (28) give $27/8000$, $43/500$, and $9/10$ for $(i, j) = (1, 1), (1, 2), (2, 2)$. Together with $\alpha(256) < 1/10400$, the vanishing activity/degree-one blocks, and the universal bound one, these verify every entry of (20). The comparison therefore proves $\|R_0\| < 11/\sqrt{q}$. Adjoints and Fourier conjugation preserve the norm, giving all four orientations in Lemma 3.

4.3 Total trace from the same geometry

To bound $\text{Tr}(R_0^* R_0)$, we sum squared incidence entries instead of bounding their largest amplification. Their squared Hilbert–Schmidt norms add over orthogonal blocks. Following BLS’s shortened-code argument [3], we collect the exact-support bounds into one polynomial.

In the incidence count, a surviving graph with d free symbols has m^d points. Its projection onto the marked positions is injective, so its incidence matrix has exactly m^d entries equal to one. This simple count supplies the trace estimate below. There are eight positions in all, four on each side of the diffusion map; choosing the residual positions among them explains the eighth power in the bound.

Lemma 5 (Damped MDS trace). *Let L_0 be an additive four-coordinate MDS map over an elementary abelian 2-group of size $m > 1$. For $s \geq 0$, let P_i project onto residual degree i and set $D_s = \sum_{i=0}^4 s^i P_i$, and let P_{fix} project onto the zero and uniform-nonzero directions. Put $W = D_s(\Pi_{L_0} - P_{\text{fix}})D_s$ and*

$$\eta = ms^2, \quad \beta = \frac{2}{1 - m^{-1/2}}, \quad \theta_0(m) = \frac{68m^2 + 264m + 68}{(m-1)^6}.$$

Then

$$\|W\|_{\text{HS}}^2 \leq \theta_0(m) + m^{-4} \sum_{k=2}^8 \binom{8}{k} \eta^k \beta^{8-k}. \quad (29)$$

Here η combines the number of values in a residual coordinate with its squared substitution factor, and β collects the contribution of an unmarked coordinate after summing over its activity choices. The term $\theta_0(m)$ accounts for the centered activity block. The index k counts residual positions across both input and output. The sum starts at two: degree zero is the separate activity term, and degree one vanishes. Each summand is nonnegative.

Proof. Consider a support block with $k = i + j > 0$ marked residual positions among the eight input/output positions and u unmarked active positions. As in (27), only shortened dimensions $1 \leq d < k$ survive. The projection onto the marked positions is injective, so the incidence matrix has m^d unit entries and Hilbert–Schmidt norm $m^{d/2}$. Residual projections cannot increase this norm. Inclusion–exclusion gives for the block the bound

$$(m-1)^{-u/2} \sum_{d=1}^{k-1} \binom{u}{u+k-4-d} m^{d/2}.$$

There are $\binom{8}{k} \binom{8-k}{u}$ such blocks. Their squared norms add, and damping multiplies them by s^{2k} . For $k = 1$ the sum is empty. For $k \geq 2$, divide the inner sum by $m^{(u+k-4)/2}$. It is a subsum of $\sum_{v=0}^u \binom{u}{v} m^{-v/2} = (1 + m^{-1/2})^u$. Thus the undamped total at degree k is at most

$$\binom{8}{k} m^{k-4} \left[1 + \frac{m}{m-1} (1 + m^{-1/2})^2 \right]^{8-k} = \binom{8}{k} m^{k-4} \beta^{8-k}.$$

Summing over $k \geq 2$ gives the binomial expression in (29). At degree zero, the symmetric activity compression has the three centered eigenvalues in (22). Their squared sum is $\theta_0(m)$; the zero and uniform-nonzero directions have both been removed. $\square$

The factorization (21) gives $\text{Tr}(R_0^* R_0) \leq \|W\|_{\text{HS}}^2$. Apply Lemma 5 with $m = 256$ and $s = 9/128$, so $\eta = 81/64$ and $\beta = 32/15$. Evaluating the sum as $(\eta + \beta)^8 - \beta^8 - 8\eta\beta^7$ gives

$$\text{Tr}(R_0^* R_0) \leq \theta_0(256) + 256^{-4} [(\eta + \beta)^8 - \beta^8 - 8\eta\beta^7] < 2^{-18}. \quad (30)$$

For an elementary rational check, $256^4\theta_0(256) < 128$, $\eta + \beta < 17/5$, $\beta > 17/8$, and $\eta > 5/4$. Hence the expression in (30), multiplied by 256^4 , is less than

$$128 + (17/5)^8 - (17/8)^8 - 10(17/8)^7 < 2^{14}.$$

Adjoint and Fourier conjugation preserve trace, so this proves the second estimate in Lemma 3 in all four orientations.

4.4 Individual columns: following a fixed input

A column of K starts from a fixed word difference. After the first SubBytes layer, its four byte differences are independent: a zero input byte stays zero, and a nonzero input byte produces a law with collision at most r_b and largest probability at most p_b . MixColumns then makes the output bytes dependent. To bound the final column, we must follow this dependence through the second SubBytes layer.

We separate the output according to its zero positions and residual positions. For example, suppose two input bytes are active and one output byte is prescribed to be zero. The MDS property leaves one free input symbol; the other input is determined by the zero equation. We keep the probability of this event in the restricted law, so its mass may be below one. The remaining source can then be bounded using the byte collision and norm estimates. This follows the MDS marginal method of Dinur [8]. We state the version for restricted, unnormalized marginals because both the column estimate here and the smoothing argument require that form.

A *marginal* keeps selected coordinates and sums over the others. The following lemma retains the zero-event probability without renormalizing, so contributions from different supports can be added. Its next-coordinate channel A will be D or one of its three other orientations.

Lemma 6 (MDS zero events and restricted responses). *Let V be an elementary abelian 2-group of size m . Feed $w \in \{1, 2, 3, 4\}$ independent laws supported on $V \setminus \{0\}$ into w positions of an additive four-coordinate MDS map, with the other inputs zero. Each input law has collision at most ρ and largest point probability at most p . Write Y for the output tuple.*

(a) Specified zero events. *Any specified set of $z \geq w$ output zeros is impossible. For a specified set D of $z = 1, 2, 3$ positions, $\Pr[Y_D = 0] \leq U_z$, where*

w	U_1	U_2	U_3	
1	0	0	0	
2	ρ	0	0	
3	ρ	$p\rho$	0	
4	ρ	ρ^2	$p^2\rho$	(31)

The other output coordinates are unrestricted.

(b) Residual response. *Let A isolate zero and be doubly stochastic on nonzero symbols. Let P fix the mass at zero and average the other entries over $V \setminus \{0\}$,*

as in Section 2.4. Suppose $R = A - P$ has norm at most s_0 and each nonzero column of A has collision at most ρ . Put $\Delta = \rho - 1/(m - 1)$. For disjoint output sets D, S with $|D| = z < w$ and $|S| = j \geq 1$, define

$$h_D(y) = \Pr[Y_S = y, Y_D = 0].$$

With $d = w - z$ and $a = \min(d, j)$,

$$\|R^{\otimes j} h_D\|_2^2 \leq p^{2z} (\rho s_0^2)^a \Delta^{j-a}. \quad (32)$$

The same bound holds with $h_D(y) = \Pr[Y_S = y, Y_D = 0, \mathcal{E}]$ for any further indicator event $\mathcal{E}$ on the input tuple. These marginals are not renormalized following restriction.

The parameters have a direct interpretation. There are w active inputs; z prescribed output zeros leave $d = w - z$ free symbols. We observe j residual outputs. Up to $\min(d, j)$ of them are bounded using the operator norm s_0 and the input collision ρ ; any remaining outputs use Δ , the squared-column bound for $A - P$. The factor p^{2z} pays for the probability factors lost when z input variables cease to be free.

Proof. Restricting the source. After z outputs are fixed, any $d = w - z$ original input coordinates parameterize the shortened MDS space. The other z probability factors are at most p . Thus the restricted density is pointwise dominated by p^z times the product of the d free input laws, and its squared norm is at most $p^{2z} \rho^d$. A further indicator only decreases this density.

Probabilities of zero events. An output coordinate is a sum of independently transformed input symbols. Its law is therefore a convolution, the law of a sum of independent variables. For one zero, a convolution of two input laws has largest probability at most ρ by Cauchy–Schwarz; convolving further laws cannot increase it. For $w = 3$ and two zeros, one free symbol remains: bound one factor by p and sum the other two by Cauchy–Schwarz. For $w = 4$ and two zeros, two disjoint pairs both parameterize the shortened space, giving ρ^2 . For three zeros use two point bounds and a collision bound. MDS gives the impossible cases.

Response of the next substitution layer. Suppose first that $j \geq d$. Any d marked outputs parameterize the shortened space bijectively. On these coordinates the parallel residual channel has norm at most s_0^d , so its Gram matrix is bounded by $s_0^{2d} I$. Each further output contributes a Schur factor with diagonal at most Δ : a nonzero residual column has squared norm at most $\rho - 1/(m - 1)$, and the zero column vanishes. By the Schur argument of Section 3.4, each such factor costs at most Δ . Multiplying this bound by the source squared norm gives (32).

If $j < d$, fix any $d - j$ free inputs. The map from the remaining j inputs to the j marked outputs is bijective: together these prescriptions specify d coordinates of the shortened MDS code. Under the unrestricted free product law, the marked marginal is therefore the convolution of a bijective image of j independent inputs with the contribution of the remaining inputs. Its collision is at most ρ^j , since averaging translates cannot increase the Euclidean norm. Pointwise domination gives $\|h_D\|_2^2 \leq p^{2z} \rho^j$; applying $R^{\otimes j}$ proves the same formula. $\square$

The lemma bounds the residual coordinates. To recover an entire activity component, we must also account for its constant-nonzero coordinates. On an exact support of size $b = 4 - z$, each such output factor contributes $1/(m - 1)$ in squared norm. Thus the response in (32) is multiplied by $(m - 1)^{j-b}$. Orthogonality permits summing over the $\binom{4}{b}$ supports and $\binom{b}{j}$ marked sets. We use this form in the byte calculation below.

Applying the restriction bound to AES. We can now finish the individual-column bound. The same argument applies to all four orientations because $D, D^*, \widehat{D}, \widehat{D}^*$ have the common parameters in (18); each uses one of the mixing maps $M, M^{-1}, M^\top, M^{-\top}$, all MDS.

For an input of byte weight t and $z < t$ zero output coordinates, put $\Delta_b = r_b - 1/255$. The specified-zero bounds are $U_{t0} = 1$, $U_{t1} = r_b$ when $t \geq 2$, $U_{3,2} = p_b r_b$, $U_{4,2} = r_b^2$, and $U_{4,3} = p_b^2 r_b$. They follow by parameterizing the shortened MDS space and using Cauchy–Schwarz, as in Lemma 6. At $t = 1$ there are no zero outputs, so only the $z = 0$ term is present.

Sum the degree contributions. We add the bounds over exact output supports and over choices of residual positions. Let μ be the word-output law for an input of byte weight t . Here F_0 includes the uniform word component: we bound the whole law first and subtract that component at the end. The quantity $\|F_j \mu\|^2$ is its squared contribution in byte degree j . For degree zero, only the probabilities of the zero conditions are needed. For positive degrees, use (32). The resulting upper bounds are

$$\begin{aligned} c_{t,0} &= \sum_{z=0}^{t-1} \binom{4}{z} \frac{U_{tz}^2}{255^{4-z}}, \\ c_{t,j} &= \sum_{z=0}^{\min(t-1, 4-j)} \binom{4}{z} \binom{4-z}{j} \frac{p_b^{2z} (r_b s^2)^\ell \Delta_b^{j-\ell}}{255^{4-z-j}} \quad (1 \leq j \leq 4), \end{aligned} \tag{33}$$

Here $\ell = \min(t - z, j)$ in each summand. In the second line, the numerator is the restricted-response bound with $w = t$, $s_0 = s$, $\rho = r_b$, and $p = p_b$. Requiring all remaining coordinates to be nonzero selects an exact support. This is the further indicator restriction allowed by Lemma 6. There are $\binom{4}{z}$ exact supports and $\binom{4-z}{j}$ choices of residual coordinates; their images are orthogonal. The unmarked active coordinates supply the displayed powers of 255.

Bound collision. Orthogonality now gives

$$\|F_j \mu\|^2 \leq c_{t,j}, \quad \text{Col}(\mu) \leq \rho_t^{\text{all}} := \sum_{j=0}^4 c_{t,j} < \frac{6}{5q}. \tag{34}$$

The last inequality is a rational check for $t = 1, 2, 3, 4$, covering every nonzero input in every orientation.

Remove the uniform word component. For a nonzero input, μ is supported on the $Q = q - 1$ nonzero words and has mass one. Its squared distance from their uniform law is therefore $\text{Col}(\mu) - 1/Q$. This is the residual column energy,

giving $(T^*T)_{xx} < 6/(5q) - 1/Q < 1/(5q)$; the zero column vanishes. This proves the third estimate and completes Lemma 3. $\square$

The local information required by Section 3 is now available. In particular, the norm and trace bounds give $\text{Tr}((T^*T)^2) \leq 121 \cdot 2^{-50} < 2^{-43}$ by (9), and the individual-column estimate is stronger than the required $(T^*T)_{xx} \leq 1/q$. Section 6 reuses the separate degree energies in (33) to bound the second response of an actual word-output law.

5 Four-Round Contraction

We bound $\mathcal{D}_4 = HHH$ on every centered input, using the orthogonal decomposition $u^\perp = \mathcal{A} \oplus \mathcal{R}$ from Section 2.4. Activity records excess or deficit mass in each zero/nonzero word pattern; the residual component records variation within patterns. Each *block* restricts the input to one space and projects the output onto another. The two within-space transfers use MDS counts (Section 4.2); the two cross-space, or coupling, transfers use the Gram bound (Section 3). Both use the local AES estimates

$$\|R_0\| \leq \nu = \frac{11}{\sqrt{q}}, \quad \text{Tr}(R_0^*R_0) \leq 2^{-18}, \quad \max_x (T^*T)_{xx} < \frac{1}{5q}, \quad (35)$$

where the diagonal bound holds for $T \in \{R_0, R_0^*, \mathcal{F}R_0\mathcal{F}, \mathcal{F}R_0^*\mathcal{F}\}$. Lemma 3 gives these estimates for $q = 2^{32}$: the norm controls residual words, while the trace and column bounds supply the coupling hypotheses in both directions.

Proposition 2 (Uniform four-round contraction). *On vectors indexed by the nonzero state differences Ω , the grouped AES operator $\mathcal{D}_4 = HHH$ satisfies*

$$\|\mathcal{D}_4 - U\| < \frac{8}{q^2} = 2^{-61}. \quad (36)$$

Since $Uf = 0$ for centered f , this reduces its Euclidean length by a factor below 2^{-61} . We assemble the four transfer bounds in a two-by-two matrix acting on the lengths of the activity and residual components.

5.1 The transfers within each space

The word layers preserve activity and contract residual variation. We reuse the geometry of Section 4.2, now with words as coordinates and alphabet size q instead of 256.

Lemma 7 (Four-round diagonal-block bounds). *Suppose $\|K - P\| \leq \nu = 11/2^{16}$, and let $q = 2^{32}$. Then*

$$\|P_{\mathcal{A}}(\mathcal{D}_4 - U)P_{\mathcal{A}}\| < 7/q^2, \quad (37)$$

$$\|P_{\mathcal{R}}\mathcal{D}_4P_{\mathcal{R}}\| < 15/(2q^2). \quad (38)$$

Proof. Activity to activity. Since H and H^* fix activity, its diagonal block equals the activity compression of $\Pi - U$. Equation (23) gives $\alpha(q) < 7/q^2$. This conclusion uses only the MDS graph counts and therefore holds for both orientations of the additive word map.

Residual variation to residual variation. Refine $\mathcal{R}$ according to the number of residual word coordinates, using $W_1, \dots, W_4$ from Section 3.3. An input in W_j is contracted by ν^j before diffusion; an output in W_i is contracted by ν^i afterwards. We therefore need the undamped diffusion bound for each pair (i, j) .

Use the support comparison X_{ij} from Section 4.2. Its squared Hilbert–Schmidt norm, the sum of its squared entries, is

$$\sum_{a,b} n_i(a) n_j(b) b_{ij}(a,b)^2$$

where $i \leq a \leq 4$ and $j \leq b \leq 4$ are the output and input activity sizes, $b_{ij}(a,b)$ is given by (27), and $n_i(a) = \binom{4}{i} \binom{4-i}{a-i}$ counts the choices of residual and other active positions. The operator norm is at most this Hilbert–Schmidt norm, so it suffices to evaluate the displayed sum at $m = q$. Its bounds are $20^2/q^3$ at $(1, 1)$, $32^2/q^2$ at $(1, 2)$, $16^2/q$ at $(1, 3)$, and $18^2/q$ at $(2, 2)$. The transposed pairs have the same bounds. For every remaining pair, we bound the diffusion block’s operator norm by one: diffusion preserves length and orthogonal projections cannot increase it. Taking square roots of the four squared-sum bounds gives the undamped comparison

$$C_+ = \begin{pmatrix} 20q^{-3/2} & 32q^{-1} & 16q^{-1/2} & 1 \\ 32q^{-1} & 18q^{-1/2} & 1 & 1 \\ 16q^{-1/2} & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}. \quad (39)$$

Rows and columns are indexed by the output and input residual degrees 1, 2, 3, 4. Reinstating the two word layers multiplies entry (i, j) by ν^{i+j} . The resulting symmetric, nonnegative matrix has all four row sums below $15/(2q^2)$, as follows by rational substitution with $q\nu^2 = 121$. Its norm is therefore below the same bound. This proves (38). $\square$

5.2 Coupling and the final contraction bound

For the off-diagonal entries, Theorem 2 uses the local norm and trace to control the Gram moment, and hence the variation created by diffusion between the two spaces.

Proof (of Proposition 2). The coupling estimates. The Gram moment of R_0 satisfies

$$\tau^2 = \text{Tr}((R_0^* R_0)^2) \leq \nu^2 \text{Tr}(R_0^* R_0) \leq 121 \cdot 2^{-50} < 2^{-43}.$$

In particular $q\tau^2 < 1$; the local diagonal bounds are stronger than the required $1/q$. Thus all hypotheses of Theorem 2 hold. Apply it with four parallel coordinates

($\ell = 4$) and $\xi = 2^{-16}$. Its constants obey

$$B_4(q) < 34, \quad \frac{14}{(1-\xi)^4} < 15, \quad C_4(q)^2 < 512.$$

Consequently

$$\|P_{\mathcal{R}} H H P_0\|^2 < 512 \nu^2 2^{-18} q^{-3} = \frac{121}{512 q^4} < \frac{1}{4 q^4}. \quad (40)$$

Thus the coupling norm is below $1/(2q^2)$. Here P_0 retains all activity patterns, whereas $\mathcal{A}$ is their centered subspace on nonzero differences. Restricting to $\mathcal{A}$ cannot increase the norm, and the first word layer fixes this input. Hence (40) bounds the activity-to-residual block of $\mathcal{D}_4$. The adjoint channel has the same Gram moment, and its column bounds are included in (35). Applying the theorem to that channel and inverse diffusion bounds the reverse block by the same constant.

Combining all four transfers. Together with Lemma 7, the coupling estimates give the following comparison, with output activity and residual lengths as rows and input lengths as columns. The block-norm comparison of BLS [3] yields

$$\|\mathcal{D}_4 - U\| < \frac{1}{q^2} \left\| \begin{pmatrix} 7 & 1/2 \\ 1/2 & 15/2 \end{pmatrix} \right\| \leq \frac{8}{q^2}.$$

The symmetric nonnegative comparison matrix has maximum row sum eight. $\square$

The q^{-2} scale. Let v be the unit-length indicator of differences with exactly two active words. There are $6Q^2$ such differences, where $Q = q - 1$, among $n = q^4 - 1$ nonzero states. As in the truncated-differential viewpoint of BLS [3], both H and H^* fix v . The MDS property requires at least five active input and output words in total, ruling out a two-active-word output from such an input. Thus $\mathcal{D}_4$ contributes zero and U contributes $6Q^2/n$ to

$$\langle v, (\mathcal{D}_4 - U)v \rangle = -\frac{6Q^2}{n}.$$

Cauchy-Schwarz gives $\|\mathcal{D}_4 - U\| \geq 6Q^2/n$, matching the q^{-2} scale of the upper bound. This concerns the operator norm, not the minimum round count for a prescribed TV distance.

6 Four-Round Smoothing

For a fixed input difference, the first word layer produces independent word laws. Following them through diffusion and the next word layer gives a stronger starting estimate than the uniform operator bound alone.

Proposition 3 (Four-round smoothing). *For every input difference $a \in \Omega$, with $q = 2^{32}$,*

$$\|p_{a,4} - u\|_2^2 < 64q^{-5} = 2^{-154}. \quad (41)$$

Equivalently, $\|p_{a,4} - u\|_2 < 8q^{-5/2} = 2^{-77}$. Squared norms let us add the orthogonal error contributions below.

The decomposition guiding the proof. We separate the final error by the output words that deviate from uniformity. One-word terms depend on marginal laws; joint terms also reflect the dependence created by diffusion. Write $a = (a_1, a_2, a_3, a_4)$, with $1 \leq w \leq 4$ nonzero words. Each active word produces a law Ke_{a_i} supported on nonzero words; inactive words stay zero. With $g = \Pi He_a$ and $p = Hg = p_{a,4}$, the stages are

$$e_a \xrightarrow{H} \bigotimes_{\ell=1}^4 Ke_{a_\ell} \xrightarrow{\Pi} g \xrightarrow{H} p_{a,4}.$$

Thus the w sources are independent before diffusion; the four word coordinates of g generally are not.

A product reference law. We follow the coordinate-subset decomposition of Dinur [8], keeping the zero-versus-nonzero bias separate from residual variation. Unlike u , the uniform law on all q^4 states is a product, so we use it first and correct the reference afterwards. For one word, J averages over all q values including zero, while $Z = P - J$ retains the bias between zero and nonzero values. With $Q = q - 1$, write

$$\begin{aligned} J &= q^{-1} \mathbf{1} \mathbf{1}^*, & Z &= P - J = \varphi \varphi^*, \\ \varphi(x) &= \frac{q \mathbf{1}_{x=0} - 1}{\sqrt{qQ}}, & K &= J + Z + R_0. \end{aligned} \tag{42}$$

The unit vector φ describes zero-versus-nonzero bias. The orthogonal images of J , Z , and R_0 record the uniform component, this bias, and variation among nonzero values. Unlike U , J averages one word including zero. Put $\tilde{K} = K - J = Z + R_0$. We must retain Z because a word can be zero in a nonzero state. Indeed, $(K - J)\varphi = \varphi$ and $\|K - J\| = 1$: only R_0 contracts by ν .

Separate one-word and joint deviations. Expand the last stage as $p = (J + \tilde{K})^{\otimes 4} g$, with one term for each set of coordinates using $\tilde{K}$. A term using it only at position i depends solely on that marginal; J averages the other three words. The all- J term is $u_* = q^{-4} \mathbf{1}$. Let E_1 sum the squared norms of terms with one $\tilde{K}$ factor, and $E_{\geq 2}$ those with at least two. These latter terms include products of marginal biases as well as dependence between words. The terms have orthogonal images, so $E_1 + E_{\geq 2} = \|p - u_*\|^2$. Since $p(0) = 0$, its squared distance to either reference equals its collision minus that reference's squared norm. Thus

$$\|p - u\|^2 = E_1 + E_{\geq 2} - \left(\frac{1}{q^4 - 1} - \frac{1}{q^4} \right) < E_1 + E_{\geq 2}. \tag{43}$$

After collecting the source-law estimates, Section 6.2 proves $E_1 < 60q^{-5}$. For $E_{\geq 2}$, one active input word gives a common source and a bound of $4q^{-5}$ (Section 6.3); several active words give independent sources and a bound of q^{-5} (Section 6.4). These are alternative cases, so $E_1 + E_{\geq 2} < 64q^{-5}$.

6.1 Estimates for the first word layer

For $\mu = Ke_a$, we need its collision, largest point probability, and response to a second word channel after a byte-monomial relabeling. Such a relabeling permutes byte positions and multiplies bytes by nonzero field elements; each one-word block of L has this form (Section 2.3). We write T_i for such a relabeling and also for its action on laws.

Lemma 8 (Forward word-output estimates). *For a nonzero input word a , let $\mu = Ke_a$. The collision and point-probability bounds are*

$$\text{Col}(\mu) < \rho := \frac{6}{5q}, \quad \|\mu\|_\infty \leq \frac{20}{q}. \quad (44)$$

For every byte-monomial bijection T_i ,

$$\|R_0 T_i \mu\|^2 < \frac{13}{q^2}. \quad (45)$$

For one active source, $T_i \mu$ is the marginal at output word i , and the last bound controls its next response. Collision controls sums of independent sources; the point-probability bound controls joint responses to one source.

Proof. The collision bound is exactly the estimate in (34). We first prove the point-probability bound, then the second-response bound.

Maximum point probability. Keliher–Sui [13] prove that the largest expected differential probability of the two-round AES superbox is $53/2^{34}$. Their reduced network consists of four S-boxes, MixColumns, and four S-boxes, with independent uniform subkeys averaged. It therefore has difference channel K , up to the fixed outer relabelings described in Section 2.3. These relabelings preserve its largest entry, so

$$\|Ke_a\|_\infty \leq \frac{53}{2^{34}} = \frac{53}{4q} < \frac{20}{q} \quad (a \neq 0).$$

We use only the coarser $20/q$ below. The collision and second-response bounds require separate estimates.

A second word response. Keeping the five byte degrees separate retains stronger damping on higher degrees. Let t be the byte weight of a and let u_Q be uniform on nonzero words. Since $F_0 u_Q = u_Q$ and $\langle \mu, u_Q \rangle = 1/Q$,

$$\|F_0(\mu - u_Q)\|^2 = \|F_0 \mu\|^2 - 1/Q, \quad F_j(\mu - u_Q) = F_j \mu \quad (j \geq 1).$$

Centering removes only the uniform part of degree zero. Thus the constants from (33) bound the lengths of the five degree components of $\mu - u_Q$ by the entries of

$$v_t = (\sqrt{c_{t,0} - 1/Q}, \sqrt{c_{t,1}}, \dots, \sqrt{c_{t,4}})^\top.$$

The first radicand is positive because $c_{t,0} \geq 255^{-4} > 1/Q$. A byte-monomial map preserves each total byte degree and fixes u_Q . Consequently $T_i(\mu - u_Q)$ obeys the same five norm bounds. Applying the nonnegative block comparison B from (20) bounds the output component lengths by Bv_t . Also $R_0T_iu_Q = 0$, so centering does not change the response being estimated. Hence

$$\|R_0T_i\mu\|^2 \leq \|Bv_t\|^2 < \frac{13}{q^2} \quad (t = 1, 2, 3, 4). \quad (46)$$

Substitution with positive square roots enclosed upward, as specified in Section 7.2, proves the four scalar inequalities and hence (45). $\square$

6.2 One-coordinate deviations

For E_1 , only one word's marginal matters. With one active source it is a relabeling of μ ; with several sources it is their sum, whose law gains additional smoothing through convolution.

Several sources: a convolution bound. For $w \geq 2$ independent sources with collision at most ρ , let g_i be an output marginal, with no restrictions on the other output coordinates. For an MDS map on an alphabet of size m , it satisfies

$$\|g_i - m^{-1}\mathbf{1}\|^2 \leq \frac{(m\rho - 1)^w}{m}. \quad (47)$$

This is the one-coordinate case of Dinur's Fourier marginal bound [8]. We retain its short proof to make the use of additive word maps explicit. Each nonzero block of the MDS map is an additive bijection. The marginal is therefore a convolution of relabeled independent source laws, say $\mu_1, \dots, \mu_w$. For an additive character χ , define the unnormalized Fourier coefficient $\hat{\mu}_\ell(\chi) = \sum_x \mu_\ell(x)\chi(x)$. These characters are the parity functions of Section 3.2; the constant one has coefficient one for a probability law. Parseval's identity therefore gives $\sum_{\chi \neq 1} |\hat{\mu}_\ell(\chi)|^2 = m \text{Col}(\mu_\ell) - 1 \leq m\rho - 1$. Fourier transformation turns convolution into a product, so

$$m \|g_i - m^{-1}\mathbf{1}\|^2 = \sum_{\chi \neq 1} \prod_{\ell=1}^w |\hat{\mu}_\ell(\chi)|^2 \leq \prod_{\ell=1}^w \sum_{\chi \neq 1} |\hat{\mu}_\ell(\chi)|^2 \leq (m\rho - 1)^w.$$

The middle inequality retains only some of the nonnegative terms in the expanded product. This proves (47) using independence of the sources; the output coordinates may remain dependent.

The zero bias and residual variation. Let g_i be the marginal of g at coordinate i , and put $t = q/Q < 2$. We must bound the two orthogonal parts $Kg_i = Zg_i + R_0g_i$. The first depends only on the probability of zero. With one active source, $g_i = T_i\mu$ and $g_i(0) = 0$. With at least two, Lemma 6 gives $g_i(0) \leq 6/(5q) < 2/q$. Thus in both cases

$$\|Zg_i\|^2 = \frac{q}{Q} (g_i(0) - 1/q)^2 \leq \frac{t}{q^2}.$$

For one source, (45) gives $\|R_0 g_i\|^2 < 13/q^2$. For $w \geq 2$ sources, $q\rho - 1 = 1/5$, and R_0 kills the uniform vector. The convolution bound (47) gives

$$\|R_0 g_i\|^2 \leq \frac{\nu^2}{q} (1/5)^w < \frac{128}{25q^2} < \frac{13}{q^2}.$$

Since Z and R_0 have orthogonal images, their squared norms add. The three coordinates averaged by J each contribute $1/q$ in squared norm. Summing over the four choices of the remaining coordinate therefore gives

$$E_1 = q^{-3} \sum_{i=1}^4 \left\| \tilde{K} g_i \right\|^2 < 4(13+t)q^{-5} < 60q^{-5}. \quad (48)$$

6.3 Joint deviations from one common source

With one active source $X \sim \mu = K e_x$, the last layer receives $T_1 X, \dots, T_4 X$, all nonzero. Corollary 1 controls their joint responses using the point-probability bound and local trace. Define

$$A_i = \tilde{K} T_i (I - P_z), \quad G_i = A_i^* A_i.$$

Since $\mu(0) = 0$, restricting to nonzero input columns leaves the response unchanged. This restriction removes the large zero-input response of $\tilde{K}$. The resulting Gram matrix is

$$G_i = T_i^* R_0^* R_0 T_i + \frac{1}{qQ} \mathbf{1}_{\neq 0} \mathbf{1}_{\neq 0}^*. \quad (49)$$

The two summands act on orthogonal spaces: the residual space and the constant-nonzero direction. The latter has eigenvalue $1/q \leq \nu^2$. With $h = 2^{-18}$, Lemma 3 therefore gives

$$\|A_i\| \leq \nu, \quad \text{Tr } G_i \leq h + 1/q < 3 \cdot 2^{-19}, \quad (G_i)_{yy} < 1/q.$$

The trace adds the eigenvalue $1/q$; the diagonal adds $1/(qQ)$ to each nonzero residual column bound and is zero at zero. Thus the shared-source bound applies with trace $h + 1/q$ and diagonal parameter $1/q$.

For a term with $j \geq 2$ coordinates using $\tilde{K}$, apply Corollary 1 with $\|\mu\|_\infty \leq 20/q$. Its unmarked uniform factors contribute q^{j-4} in squared norm, so

$$E_{\geq 2} \leq \nu^2 (20/q)^2 (h + 1/q) \sum_{j=2}^4 \binom{4}{j} q^{j-4} q^{-(j-2)} = 11\nu^2 (20/q)^2 (h + 1/q) q^{-2}.$$

The power of q in the summand is q^{-2} for every j . The factor $11 = \binom{4}{2} + \binom{4}{3} + \binom{4}{4}$ counts the possible coordinate sets. Using $q\nu^2 < 128$ yields

$$q^5 E_{\geq 2} < 11 \cdot 128 \cdot 400 \cdot 3 \cdot 2^{-19} < 4.$$

6.4 Joint deviations from independent sources

With several active words, the sources are independent, though the outputs of diffusion need not be. Expand each $\tilde{K} = Z + R_0$. The orthogonal factors J , Z , and R_0 allow squared norms to add; the following lemma bounds them using the MDS restrictions of Section 4.4.

Lemma 9 (Higher-order error from independent sources). *Let $q = 2^{32}$, $c = 6/5$, and $\rho = c/q$. Feed $2 \leq w \leq 4$ independent laws supported on nonzero words through an additive four-coordinate MDS map. Suppose their collisions are at most ρ and their point probabilities at most $20/q$. Let the final channel K isolate zero, be doubly stochastic on nonzero symbols, and satisfy*

$$\|R_0\| \leq \nu, \quad q\nu^2 \leq 128, \quad \max_{x \neq 0} \|Ke_x\|^2 \leq \rho, \quad R_0 = K - P.$$

Let g be the law after diffusion. In the tensor expansion $(J + Z + R_0)^{\otimes 4}g$, let $E_{\geq 2}$ be the sum of squared norms of terms with at least two Z or R_0 factors. Then $E_{\geq 2} < q^{-5}$.

Proof. Write $Y = (Y_1, \dots, Y_4) \sim g$. We bound a residual response with prescribed zeros, then apply it to the three possible kinds of term. Put $Q = q - 1$ and $b = 21^2/Q$ for the cost of a Z factor; $21 = 1 + 20$ arises from expanding its zero-versus-nonzero condition.

A residual response with prescribed zeros. For $r \geq 1$ marked residual coordinates and $k < w$ prescribed zeros on other coordinates, put $d = w - k$. Let h_D be the unnormalized marginal on the r marked coordinates with zeros prescribed on the set D , where $|D| = k$. Here d counts the free source symbols, and the mass of h_D is the probability of the prescribed-zero event. In (32), substitute $\rho = c/q$, $p = 20/q$, and $\rho - 1/Q < 1/q$. Taking square roots and using $c^{\min(d,r)/2} \leq c^2$ gives

$$\|R_0^{\otimes r} h_D\| \leq c^2 20^k q^{-k-r/2} \nu^{\min(d,r)}. \quad (50)$$

If $k \geq w$, the prescribed-zero event is impossible.

The three kinds of term. Consider one tensor term with r residual factors R_0 , z bias factors Z , and $4 - r - z$ uniform factors, where $r + z \geq 2$. Its squared norm, including the uniform factors, is bounded by

$$q^{-4} \begin{cases} c^4 \nu^4, & r \geq 2, z = 0, \\ c^4 \nu^2 b^z, & r \geq 1, z \geq 1, \\ c^2 b^z, & r = 0, z \geq 2. \end{cases} \quad (51)$$

For the first case, use (50) with $k = 0$; $\min(w, r) \geq 2$ supplies two contractions.

For the mixed case, each Z factor projects onto the unit vector φ . Its coefficient is found by weighting the corresponding coordinate by $\varphi(Y_i)$. Let S contain these z coordinates and expand

$$\prod_{i \in S} \varphi(Y_i) = (qQ)^{-z/2} \sum_{D \subseteq S} (-1)^{z-|D|} q^{|D|} \mathbf{1}_{Y_D=0}.$$

Apply (50) termwise. Every surviving term supplies at least one contraction, because at least one free source symbol and one residual output remain. The factor $q^{|D|}$ in the expansion cancels the q^{-k} in (50). Hence the weighted residual norm is at most

$$c^2 \nu q^{-r/2} (qQ)^{-z/2} \sum_{k=0}^z \binom{z}{k} 20^k = c^2 \nu q^{-r/2} (qQ)^{-z/2} 21^z.$$

The unmarked uniform factors contribute q^{r+z-4} in squared norm, which gives the second case of (51).

For Z factors only, let S be their positions. Before including the uniform factors, their coefficient is $\mathbb{E}[\prod_{i \in S} \varphi(Y_i)]$. Bound this expectation using the zero-event probabilities in Lemma 6. They give $\Pr[Y_D = 0] \leq c 20^{k-1} q^{-k}$ for $k = |D| \geq 1$. The expansion bounds the absolute value of this coefficient by

$$(qQ)^{-z/2} \left[1 + \frac{c}{20} (21^z - 1) \right] \leq c 21^z (qQ)^{-z/2}.$$

This gives the third case, again including the uniform factors.

Summing the orthogonal terms. Now $c < 2$, $\nu^2 \leq 128/q$, and $b < 512/q < 1$. Including the common factor q^{-4} , each bound in (51) is therefore strictly below $2^{20} q^{-6}$. There are at most $3^4 = 81$ choices of J , Z , or R_0 across the four words, and only a subset contributes to $E_{\geq 2}$. Orthogonality yields

$$E_{\geq 2} < 81 \cdot 2^{20} q^{-6} = \frac{81}{4096} q^{-5} < q^{-5}. \quad \square$$

For $\mu = Ke_a$, Lemmas 8 and 3 supply all hypotheses. Thus $E_{\geq 2} < q^{-5}$ with several sources, and $E_{\geq 2} < 4q^{-5}$ in either source case. Adding $E_1 < 60q^{-5}$ from (48) in (43) proves Proposition 3. $\square$

7 The Mixing Bound and Conclusions

7.1 Combining the four-round estimates

We combine smoothing and contraction as in BLS and Dinur [3,8]. The extra diffusion at group boundaries preserves Euclidean length and uniformity; we make it explicit below before converting to total variation.

Proof (of Theorem 1). For even $r \geq 2$, write the grouped difference law and its deviation from uniformity as

$$p_{a,r} = H(\Pi H)^{r/2-1} e_a, \quad f_{a,r} = p_{a,r} - u.$$

Appending two rounds adds a diffusion and superbox layer, both preserving u , so

$$f_{a,r+2} = H \Pi f_{a,r} \quad (r \geq 4 \text{ even}).$$

By (6), four more rounds apply $(H\Pi)^2 = \mathcal{D}_4\Pi$. Since Π preserves the centered subspace and its norm, Propositions 2 and 3 give

$$\|f_{a,4+4k}\|_2 < \frac{8}{q^{5/2}} \left(\frac{8}{q^2}\right)^k.$$

There are $n = q^4 - 1$ nonzero differences. Cauchy–Schwarz gives $\|f\|_1 \leq \sqrt{n} \|f\|_2$ on this space, and whitening identifies the pairwise total variation with half the difference-law ℓ_1 distance. Consequently, using $n < q^4$,

$$\varepsilon_{4+4k} \leq \frac{1}{2} \sqrt{n} \max_{a \neq 0} \|f_{a,4+4k}\|_2 < \frac{4}{\sqrt{q}} \left(\frac{8}{q^2}\right)^k = 2^{-14-61k}.$$

The strict inequality also holds for $k = 0$. Section 2 proves the complete adaptive two-query statement. Taking $k = 2, 3, 4$ gives the advertised 12, 16, and 20 sufficient rounds.

Additional rounds. To append a round while still omitting final MixColumns, apply MC to both current outputs, then append $z \mapsto (\text{SR} \circ \text{SB})(z) + K_{r+1}$ with a fresh uniform key. The old last key becomes $\text{MC}(K_r)$, which is still independent uniform. This common Markov kernel constructs the next cipher family and preserves the uniform distinct-pair law: for every fixed new key, the appended map is a permutation. Applying a common Markov kernel cannot increase total variation, so $\varepsilon_{r+1} \leq \varepsilon_r$. Hence the stated bound holds for every $r \geq 4 + 4k$, completing the theorem. $\square$

7.2 Finite arithmetic and analytic inputs

The analytic inputs are character orthogonality for the byte norm (Lemma 4), MDS restrictions for byte-degree energies, and BLS activity geometry with the centered spectrum evaluated in (22). The point-probability estimate uses the computer-assisted Keliher–Sui bound [13]; our checks do not rerun their exhaustive calculation.

For the local norm, trace, collision bounds in all four orientations, and starting energies in Sections 4–6, finite sums use exact integers and rationals. Positive square roots in upper bounds use

$$\text{rt}(x) = 2^{-256} \lceil 2^{256} \sqrt{x} \rceil.$$

Squaring verifies $\text{rt}(x)^2 \geq x$ and, when $x > 0$, $(\text{rt}(x) - 2^{-256})^2 < x$. The matrix inequalities are checked by exact rational elimination or positive-vector bounds, as in Section 4.2. These finite evaluations of the displayed formulas instantiate the preceding analytic estimates.

7.3 Conclusions

The rate $2^{-14-61k}$ combines the product structure of the first word-output laws with a uniform contraction for every subsequent four rounds. Thus 12, 16, and

20 rounds suffice for distances below 2^{-128} , 2^{-192} , and 2^{-256} in the model of Theorem 1.

Both estimates retain aggregate information about local columns: the second Gram moment controls overlaps along an MDS code, while the trace controls a common source with bounded point probabilities. Averaging before taking a worst case serves both contraction and smoothing. The general code and coupling results also apply to other additive MDS layers once the required local and Fourier-basis bounds are established.

The bounds concern actual AES with independent uniform round keys sampled once and retained, and at most two adaptive classical forward queries. No such bounds are established here for the standard key schedule or more queries. The matching q^{-2} operator scale (Section 5) does not determine minimum round counts for total-variation targets.

References

1. Baignères, T., Vaudenay, S.: Proving the security of AES substitution-permutation network. In: Selected Areas in Cryptography—SAC 2005. LNCS, vol. 3897, pp. 65–81. Springer (2006). https://doi.org/10.1007/11693383_5
2. Beyne, T.: A geometric approach to linear cryptanalysis. In: Advances in Cryptology—ASIACRYPT 2021. LNCS, vol. 13090, pp. 36–66. Springer (2021). https://doi.org/10.1007/978-3-030-92062-3_2
3. Beyne, T., Leander, G., Schütt, I.: Pairwise independence of AES-like block ciphers. In: Daemen, J., Thomé, E. (eds.) Advances in Cryptology—EUROCRYPT 2026. LNCS, vol. 16546, pp. 181–210. Springer (2026). https://doi.org/10.1007/978-3-032-25333-0_7, corrected full version: ePrint 2025/1495, 27 September 2026
4. Beyne, T., Rijmen, V.: Differential cryptanalysis in the fixed-key model. In: Advances in Cryptology—CRYPTO 2022. LNCS, vol. 13509, pp. 687–716. Springer (2022). https://doi.org/10.1007/978-3-031-15982-4_23
5. Chabaud, F., Vaudenay, S.: Links between differential and linear cryptanalysis. In: Advances in Cryptology—EUROCRYPT ’94. LNCS, vol. 950, pp. 356–365. Springer (1995). <https://doi.org/10.1007/BFb0053450>
6. Daemen, J., Rijmen, V.: The wide trail design strategy. In: Cryptography and Coding. LNCS, vol. 2260, pp. 222–238. Springer (2001). https://doi.org/10.1007/3-540-45325-3_20
7. Daemen, J., Rijmen, V.: Understanding two-round differentials in AES. In: Security and Cryptography for Networks—SCN 2006. LNCS, vol. 4116, pp. 78–94. Springer (2006). https://doi.org/10.1007/11832072_6, full report: ePrint 2006/039, titled *Two-Round AES Differentials?*
8. Dinur, I.: New techniques for analyzing differentials with application to AES. In: Daemen, J., Thomé, E. (eds.) Advances in Cryptology—EUROCRYPT 2026. LNCS, vol. 16546, pp. 211–242. Springer (2026). https://doi.org/10.1007/978-3-032-25333-0_8, full version: ePrint 2025/1326, 5 March 2026
9. Dujmovic, J., Pelecanos, A., Tessaro, S.: When simple permutations mix poorly: Limited independence does not imply pseudorandomness. In: Advances in Cryptology—EUROCRYPT 2026. LNCS, vol. 16546, pp. 424–450. Springer (2026). https://doi.org/10.1007/978-3-032-25333-0_15

10. Horn, R.A., Johnson, C.R.: Topics in Matrix Analysis. Cambridge University Press (1991). <https://doi.org/10.1017/CBO9780511840371>, chapter 5: The Hadamard product
11. Hurst, N.: Improving pairwise independence bounds for AES via numerical spectral computation. COSIC seminar, KU Leuven (August 2026), 25 August 2026. Public seminar abstract
12. Jain, V., Liu, T., Mizgerd, C., Pelecanos, A., Tessaro, S., Vaikuntanathan, V.: How fast does the inverse walk approximate a random permutation? In: Advances in Cryptology—CRYPTO 2026. LNCS, vol. 16805, pp. 257–285. Springer (2026). https://doi.org/10.1007/978-3-032-35412-9_9
13. Keliher, L., Sui, J.: Exact maximum expected differential and linear probability for two-round Advanced Encryption Standard. IET Information Security **1**(2), 53–57 (2007). <https://doi.org/10.1049/iet-ifs:20060161>, full report: ePrint 2005/321
14. Kranz, T., Leander, G., Wiemer, F.: Linear cryptanalysis: Key schedules and tweakable block ciphers. IACR Transactions on Symmetric Cryptology **2017**(1), 474–505 (2017). <https://doi.org/10.46586/tosc.v2017.i1.474-505>, full version: ePrint 2017/154
15. Lai, X., Massey, J.L., Murphy, S.: Markov ciphers and differential cryptanalysis. In: Advances in Cryptology—EUROCRYPT ’91. LNCS, vol. 547, pp. 17–38. Springer (1991). https://doi.org/10.1007/3-540-46416-6_2
16. Liu, T., Pelecanos, A., Tessaro, S., Vaikuntanathan, V.: Layout graphs, random walks and the t -wise independence of SPN block ciphers. In: Advances in Cryptology—CRYPTO 2023. LNCS, vol. 14083, pp. 694–726. Springer (2023). https://doi.org/10.1007/978-3-031-38548-3_23
17. Liu, T., Tessaro, S., Vaikuntanathan, V.: The t -wise independence of substitution-permutation networks. In: Advances in Cryptology—CRYPTO 2021. LNCS, vol. 12828, pp. 454–483. Springer (2021). https://doi.org/10.1007/978-3-030-84259-8_16
18. MacWilliams, F.J., Sloane, N.J.A.: The Theory of Error-Correcting Codes, North-Holland Mathematical Library, vol. 16. North-Holland (1977)
19. National Institute of Standards and Technology: Advanced encryption standard (AES). Federal Information Processing Standards Publication 197-upd1, U.S. Department of Commerce (May 2023). <https://doi.org/10.6028/NIST.FIPS.197-upd1>, updated 9 May 2023
20. Nyberg, K.: Differentially uniform mappings for cryptography. In: Advances in Cryptology—EUROCRYPT ’93. LNCS, vol. 765, pp. 55–64. Springer (1994). https://doi.org/10.1007/3-540-48285-7_6
21. Nyberg, K., Knudsen, L.R.: Provable security against a differential attack. Journal of Cryptology **8**(1), 27–37 (1995). <https://doi.org/10.1007/BF00204800>
22. Rønjom, S.: A short note on a weight probability distribution related to SPNs. Cryptology ePrint Archive, Paper 2019/750 (2019), <https://eprint.iacr.org/2019/750>
23. Vaudenay, S.: Decorrelation: A theory for block cipher security. Journal of Cryptology **16**(4), 249–286 (2003). <https://doi.org/10.1007/s00145-003-0220-6>