LATE-QUERY ALGEBRAIC CHECKPOINTING AND ARBITRARY MATRIX-PRODUCT OUTPUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Exact checkpointing has two distinct costs: recomputing discarded information and combining early data with a query unavailable at commit time. We separate them using mixed derivatives of bounded-arity analytic circuits. For a bilinear target, any M-scalar checkpoint removes only an output component of dimension at most M, leaving a quotient tensor. Its ordinary rank is the exact late multiplication cost in the saturated bilinear model. For matrix multiplication, we evaluate this cost over *all* q-dimensional linear output spaces:

$$\mathsf{PMM}^{\star}_n(q) = n^{\omega(\gamma/2,1,\gamma/2)+o(1)}, \qquad q = n^{\gamma+o(1)} \geq 1, \quad 0 \leq \gamma \leq 2,$$

for ordinary, border, and asymptotic rank. A uniform version implies that a fixed saving in the late-work exponent requires storage of at least $n^2 - n^{2-\varepsilon+o(1)}$ scalars; saving any fixed fraction of full-output storage cannot improve that exponent. An explicit residual ReLU construction makes a readout gradient of unnormalized symmetric InfoNCE recover a matrix product. Beyond its fixed-key storage lower bound, constant-score key perturbations expose an independent early–late matrix product. The full late-key gradient family therefore has optimal late exponent ω, regardless of checkpoint capacity. This separates storage-limited recomputation from genuinely query-dependent arithmetic.

1 INTRODUCTION

Checkpointing trades saved intermediate values for later recomputation, but a late query creates a second obstacle: some arithmetic cannot be performed before its operands arrive. Gradient caching and rematerialization motivate the first issue (Gao et al., 2021; Chen et al., 2016; Griewank & Walther, 2000; Jain et al., 2020); query-dependent negative selection motivates the second (Xiong et al., 2021). We study an explicit information-flow cut: the early input may be reread, but only M input-dependent derived scalars survive, and the checkpoint is independent of the realized late query. This is a scalar arithmetic model, not a byte-memory or wall-clock model. In particular, ANCE motivates late key selection without being assumed to enforce our cut.

What can nonlinear checkpoints save? For a bilinear target T, our mixed-derivative normal form decomposes *any* analytic two-stage computation into $T = U + V$, where $\dim \operatorname{im} U \leq M$ and the early and late sizes bound $\mathbf{R}(U)$ and $\mathbf{R}(V)$. Thus a nonlinear checkpoint removes only an output-low-dimensional component of the bilinear part. With unrestricted preprocessing, the exact late count in the separated bilinear model is the *checkpoint quotient rank* $\mathsf{CQR}^{\mathbf{R}}_M(T) = \min_{\dim W \leq M} \mathbf{R}(\pi_W T)$. Finite preprocessing budgets admit a rank-constrained characterization as well; arbitrary analytic gates inherit constant-factor lower bounds.

An optimal frontier over arbitrary output encodings. For $\mu_n(A, B) = AB$, storage/output duality identifies a residual quotient of dimension $q = n^2 - M$ with a linear output-dual space $E \leq M_n(\mathbb{C})$. Write $T_E(A, B, C) = \operatorname{tr}(C^{\top} AB)$ for $C \in E$, and let $\mathsf{PMM}^{\star}_n(q)$ minimize its tensor complexity over all such q-planes. We prove

$$\mathsf{PMM}^{\star}_n(q) = n^{f(\gamma/2)+o(1)}, \qquad f(\alpha) = \omega(\alpha, 1, \alpha), \qquad q = n^{\gamma+o(1)},$$

for ordinary, border, and asymptotic rank. The lower bound covers the full Grassmannian; a balanced output block attains the exponent. Degeneration to Ferrers shapes, a harmonic rectangle, tensor balancing, and exactification provide the proof.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The uniform form yields an explicit storage barrier (Corollary 4.9):

$$Q \le n^{\omega-\varepsilon+o(1)} \quad \Longrightarrow \quad M \ge n^2 - n^{2-\varepsilon+o(1)}.$$

In particular, even unlimited preprocessing cannot lower the late exponent when $M \le (1-\delta)n^2$ for fixed $\delta > 0$. Constant-factor arithmetic savings and exponent improvements are different guarantees.

Separating recomputation from late arithmetic in gradients. For unnormalized symmetric InfoNCE (van den Oord et al., 2018), a residual ReLU slice hides AB while its visible query remains constant. A fixed gradient projection therefore inherits the storage frontier. We additionally exhibit key perturbations that leave *every score unchanged* but turn another gradient projection into an early–late matrix product. This gives a matching $n^{\omega+o(1)}$ late cost independently of M (Theorem 5.3): the quotient frontier describes bilinear recomputation, while the full varying-key gradient also incurs genuinely late arithmetic.

2 Uniform late-query checkpoint model

2.1 Information flow across the commit boundary

Let $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ and let $\mathcal{U}_0 \subseteq \mathbb{F}^d$, $\mathcal{K}_0 \subseteq \mathbb{F}^e$ be nonempty open neighborhood factors. For analytic $\Phi : \mathcal{U}_0 \times \mathcal{K}_0 \to \mathcal{Z}$, a uniform (P, M, Q) scheme consists of an analytic checkpoint $S : \mathcal{U}_0 \to \mathbb{F}^M$ and a decoder D, analytic near $\{(u, k, S(u))\}$, such that

$$D(u, k, S(u)) = \Phi(u, k) \qquad ((u, k) \in \mathcal{U}_0 \times \mathcal{K}_0). \tag{1}$$

Here S is independent of the realized k; M counts every input-dependent derived scalar crossing the cut. The raw u may be reread, but no uncharged buffer, second message, or query-dependent preprocessing is allowed. Fixed linear maps are free, and P, Q count nonlinear scalar nodes as in Section 3. Their coefficients cannot depend on a realized input or query: multiplying by a late key entry is charged. Uniformity means one circuit works on the product neighborhood, not algorithmic uniformity in the dimension. Discontinuous encodings and uncharged input-dependent state are excluded.

2.2 Hard-slice reduction

Lemma 2.1 (Fixed-query hard slice)**.** *Suppose $u = (x, y)$ ranges over a product neighborhood, $k_* \in \mathcal{K}_0$ is a fixed admissible query, and $L : \mathcal{Z} \to \mathcal{Z}'$ is a fixed linear or affine map. If*

$$L\Phi((x, y), k_*) = T(x, y)$$

on a nonempty product neighborhood, then every uniform scheme (1) *induces an exact two-stage scheme for T with the same checkpoint S and with late decoder*

$$D_*(x, y, s) := LD((x, y), k_*, s).$$

The late nonlinear size does not increase.

Proof. Substitute k_* into (1) and compose with L. The checkpoint is unchanged, and the fixed affine readout adds no nonlinear gate. □

Lemma 2.2 (Query-crossing mixed jet)**.** *Every scheme* (1) *satisfies, at each regular base point,*

$$\mathbf{R}\big(D^2_{uk}\Phi(u_0, k_0)\big) \le c_{\mathcal{G}} Q,$$

independently of P and M.

Proof. Replace $S(u)$ by $S(u_0) + D_u S(u_0)(u - u_0)$ inside the late circuit. This preserves the mixed u, k derivative because S has no k dependence, and adds only affine wires. Apply Lemma 3.1. □

The two lemmas isolate different costs: a fixed query exposes discarded early–early computation; a query-crossing jet exposes arithmetic that no early checkpoint can precompute.

3 MIXED JETS AND CHECKPOINT QUOTIENT RANK

3.1 TENSOR CONVENTIONS

Unless a field is displayed, algebraic tensor statements in Sections 3–4 are over $\mathbb{C}$. A bilinear map $T : X \times Y \to Z$ is identified with an element of $X^* \otimes Y^* \otimes Z$. Throughout, $\operatorname{im} T$ denotes the linear span of the values $T(x, y)$, equivalently the image of the induced linear map $X \otimes Y \to Z$. Ordinary rank $\mathbf{R}(T)$ is the smallest number of simple tensors summing to T; border rank $\underline{\mathbf{R}}(T)$ replaces such sums by their Zariski closure. Its asymptotic rank satisfies

$$\widetilde{\mathbf{R}}(T) := \inf_{m \geq 1} \underline{\mathbf{R}}(T^{\otimes m})^{1/m}, \qquad \mathbf{R}(T) \geq \underline{\mathbf{R}}(T) \geq \widetilde{\mathbf{R}}(T). \tag{2}$$

We use $\langle a, b, c \rangle$ for multiplying an $a \times b$ matrix by a $b \times c$ matrix. For $\alpha, \beta, \gamma \geq 0$, we use the standard full-sequence rectangular exponent, characterized by

$$\widetilde{\mathbf{R}}\big(\langle \lceil n^\alpha \rceil, \lceil n^\beta \rceil, \lceil n^\gamma \rceil \rangle\big) = n^{\omega(\alpha,\beta,\gamma)+o(1)}. \tag{3}$$

Blocking and padding give the full-sequence form (Bürgisser et al., 1997; Vrana, 2024). The exponent is homogeneous of degree one, $\omega(\lambda\alpha, \lambda\beta, \lambda\gamma) = \lambda\,\omega(\alpha, \beta, \gamma)$ for $\lambda \geq 0$, and invariant under the standard permutations of tensor legs. We write $\omega = \omega(1, 1, 1)$.

Coordinate monotonicity. Restricting the row, contraction, and column coordinates realizes $\langle a', b', c' \rangle$ as a restriction of $\langle a, b, c \rangle$ whenever $a' \leq a$, $b' \leq b$, and $c' \leq c$. Thus

$$\star\langle a', b', c' \rangle \leq \star\langle a, b, c \rangle, \qquad \star \in \{\mathbf{R}, \underline{\mathbf{R}}, \widetilde{\mathbf{R}}\}. \tag{4}$$

Here a, b, c are matrix side lengths, not tensor-leg dimensions; restriction commutes with closure and tensor powers.

3.2 SCALAR ANALYTIC DAGS AND THE MIXED-JET LEMMA

Fix a finite analytic gate library $\mathcal{G}$ of arity at most a. A circuit is a scalar directed acyclic graph followed by a fixed linear readout of scalar node coordinates. Linear and affine nodes are free. The nonlinear scalar size $\operatorname{nl}(F)$ counts the non-affine scalar gates. A gate whose Hessian vanishes at the chosen point contributes zero to the bound, but is still counted. The library and its arity bound are fixed independently of the input dimension. Gates are evaluated at a regular point where their analytic branches are defined.

Lemma 3.1 (Bounded-arity mixed-jet rank)**.** *There is a constant $c_{\mathcal{G}}$, depending only on $\mathcal{G}$, such that for every vector-valued circuit $F : X \times Y \to Z$ and every regular base point (x_0, y_0),*

$$\mathbf{R}\big(D^2_{xy} F(x_0, y_0)\big) \leq c_{\mathcal{G}}\ \operatorname{nl}(F).$$

The rank on the left is tensor rank in $X^* \otimes Y^* \otimes Z$.

Proof. Second-order reverse accumulation sums node contributions $\sigma_g \otimes ((J^x_g)^\top H_g J^y_g)$, where σ_g is the output sensitivity and H_g the local scalar Hessian. Each Hessian has rank at most a, so each node contributes at most a simple tensors. Appendix A gives the full argument. □

An *exact two-stage scheme for T* satisfies $D(x, y, S(x, y)) = T(x, y)$ on a product neighborhood, with all gates on their analytic branches. It is the fixed-query specialization of Section 2.

Theorem 3.2 (Finite analytic checkpoint normal form)**.** *Let $T : X \times Y \to Z$ be bilinear and suppose T admits an exact two-stage scheme with preprocessing size at most P, checkpoint dimension M, and late size at most Q. Then there are bilinear maps $U, V : X \times Y \to Z$ such that*

$$T = U + V, \qquad \dim \operatorname{im} U \leq M, \qquad \mathbf{R}(U) \leq c_{\mathcal{G}} P, \qquad \mathbf{R}(V) \leq c_{\mathcal{G}} Q. \tag{5}$$

Proof. Translate a regular base point to $(0, 0)$ and write $s_0 = S(0, 0)$, $A = D_x S$, $B = D_y S$, and $C = D^2_{xy} S$ there. The mixed-jet lemma gives $\mathbf{R}(C) \leq c_{\mathcal{G}} P$. Define $\overline{D}(x, y) := D(x, y, s_0 + Ax + By)$. The chain rule gives

$$D^2_{xy}[D(x, y, S(x, y))] = D_s D(0, 0, s_0) \circ C + D^2_{xy} \overline{D}(0, 0). \tag{6}$$

The first summand is U, which factors through $\mathbb{F}^M$ and has rank at most $c_{\mathcal{G}}P$; the second is V and has rank at most $c_{\mathcal{G}}Q$ because affine wires add no nonlinear nodes. The left side is T after taking the mixed jet, proving the stated normal form. □

Proposition 3.3 (Finite preprocessing–storage–late tradeoff)**.** *Every exact scheme in Theorem 3.2 satisfies*

$$Q \geq c_{\mathcal{G}}^{-1}\mathbf{R}(T) - P \qquad \text{and} \qquad Q \geq c_{\mathcal{G}}^{-1}\mathsf{CQR}_M^{\mathbf{R}}(T). \tag{7}$$

The first bound records the finite preprocessing budget; the second is independent of how much preprocessing work is spent on a component whose output span is at most M.

Proof. Rank subadditivity gives $\mathbf{R}(T) \leq c_{\mathcal{G}}(P+Q)$. Quotienting (5) by $W = \operatorname{im} U$ gives $\mathbf{R}(\pi_W T) \leq c_{\mathcal{G}}Q$. □

3.3 Finite exact checkpoint quotient rank

For a tensor $T \in X^* \otimes Y^* \otimes Z$ and an output subspace $W \leq Z$, let $\pi_W : Z \to Z/W$ be the quotient map. Define the finite exact quantity

$$\mathsf{CQR}_M^{\mathbf{R}}(T) := \min_{\substack{W \leq Z \\ \dim W \leq M}} \mathbf{R}\big((\mathrm{Id} \otimes \mathrm{Id} \otimes \pi_W)T\big). \tag{8}$$

We also use the relaxations

$$\mathsf{CQR}_M^{\underline{\mathbf{R}}}(T) := \min_{\dim W \leq M} \underline{\mathbf{R}}(\pi_W T), \qquad \mathsf{CQR}_M^{\widetilde{\mathbf{R}}}(T) := \inf_{\dim W \leq M} \widetilde{\mathbf{R}}(\pi_W T). \tag{9}$$

Proposition 3.4 (Finite-budget bilinear characterization)**.** *In the separated bilinear model, charged products have the form $a(x)b(y)$, checkpoint coordinates are bilinear, and the late readout is linear in stored coordinates and new products. Its optimal late count with budgets (P, M) is*

$$\mathsf{F}_{P,M}(T) := \min_{\substack{\dim \operatorname{im} U \leq M \\ \mathbf{R}(U) \leq P}} \mathbf{R}(T - U). \tag{10}$$

With unrestricted preprocessing this equals $\mathsf{CQR}_M^{\mathbf{R}}(T)$. Every analytic scheme also satisfies $\mathsf{F}_{\lfloor c_{\mathcal{G}}P \rfloor, M}(T) \leq c_{\mathcal{G}}Q$.

Proof. A scheme gives $U = LS$ with $\mathbf{R}(U) \leq P$ and $\mathbf{R}(T-U) \leq Q$. Conversely, store coordinates of a minimizing U and compute $T - U$ late. With unrestricted P, quotient by $\operatorname{im} U$; a linear section of a minimizing quotient gives the reverse construction. The analytic assertion follows from (5). □

Border and asymptotic rank supply degeneration bounds; exactification below recovers the ordinary-rank exponent.

4 Arbitrary linear outputs of one matrix product

4.1 The dual bridge and the three finite quantities

Let $\mu_n(A, B) = AB$ on $M_n(\mathbb{C})$. Under the trace pairing $\langle C, Z \rangle = \operatorname{tr}(C^\top Z)$, a subspace $E \leq M_n(\mathbb{C})$ defines

$$T_E(A, B, C) := \operatorname{tr}(C^\top AB), \qquad C \in E. \tag{11}$$

For $1 \leq q \leq n^2$, set

$$\mathsf{PMM}_n^\star(q) := \inf_{\substack{E \leq M_n(\mathbb{C}) \\ \dim E = q}} \star(T_E), \qquad \star \in \{\mathbf{R}, \underline{\mathbf{R}}, \widetilde{\mathbf{R}}\}, \qquad \mathsf{PMM}_n^\star(0) := 0. \tag{12}$$

For ordinary and border rank the infimum is a minimum, since the attainable ranks are nonnegative integers; no attainment is needed for asymptotic rank.

Proposition 4.1 (Storage/output duality). *For $0 \le M \le n^2$ and $q = n^2 - M$,*

$$\mathsf{CQR}^{\star}_{M}(\mu_n) = \mathsf{PMM}^{\star}_{n}(q), \qquad \star \in \{\mathbf{R}, \underline{\mathbf{R}}, \widetilde{\mathbf{R}}\}. \tag{13}$$

For $M \ge n^2$, both sides are interpreted as zero.

Proof. A quotient by W has output-dual space $W^{\perp}$. Enlarging W to dimension M only restricts this dual space and cannot increase any rank. The correspondence $W \leftrightarrow W^{\perp}$ then gives the equality. □

4.2 Borel–Ferrers degeneration

Let B_+ be the upper-triangular Borel subgroup of $GL_n(\mathbb{C})$ and B_- the lower-triangular Borel subgroup. The action on the output-dual matrix space is fixed throughout this section as

$$(g, h) \cdot C = gCh^{-1}, \qquad (g, h) \in GL_n \times GL_n. \tag{14}$$

Cyclicity of trace gives $\mathrm{tr}((gCh^{-1})^{\top}AB) = \mathrm{tr}(C^{\top}(g^{\top}A)(Bh^{-\top}))$. Thus moving E gives an isomorphic tensor through invertible input changes.

Lemma 4.2 (Borel–Ferrers degeneration). *For every q-plane $E \le M_n(\mathbb{C})$, the tensor T_E degenerates to a coordinate-output tensor T_S for a northwest-closed Ferrers diagram $S \subseteq [n] \times [n]$ with $|S| = q$. In particular,*

$$\widetilde{\mathbf{R}}(T_E) \ge \widetilde{\mathbf{R}}(T_S), \qquad \underline{\mathbf{R}}(T_E) \ge \underline{\mathbf{R}}(T_S). \tag{15}$$

Proof. A Borel fixed point in the projective orbit closure is a matrix-unit span, by distinct torus weights; upper/lower unipotents force northwest closure (Borel, 1991; Eisenbud, 1995). A curve to this fixed point lifts to a tensor degeneration, proving both inequalities. Appendix A.2 details the lift. □

4.3 A harmonic rectangle and tensor balancing

Lemma 4.3 (Harmonic rectangle). *Let S be northwest Ferrers with row lengths $\mu_1 \ge \cdots \ge \mu_n \ge 0$ and $|S| = q$. There are integers $p, r \le n$ such that S contains a $p \times r$ rectangle and*

$$pr \ge \frac{q}{H_n}, \qquad H_n := \sum_{i=1}^{n} \frac{1}{i}. \tag{16}$$

Proof. Put $A = \max_i i\mu_i$. Then $q \le AH_n$; at a maximizing p, $r = \mu_p$ gives the contained rectangle with $pr = A$. □

Lemma 4.4 (Balancing by tensor products). *For all p, r, n,*

$$\widetilde{\mathbf{R}}(\langle p, n, r\rangle) \ge \widetilde{\mathbf{R}}(\langle pr, n^2, pr\rangle)^{1/2}. \tag{17}$$

Ordinary and border rank are also bounded below by the displayed asymptotic quantity through (2).

Proof. Tensor $\langle p, n, r\rangle$ with its leg permutation $\langle r, n, p\rangle$ to obtain $\langle pr, n^2, pr\rangle$. Submultiplicativity and the rank hierarchy give (17). □

4.4 Finite sandwich

Put

$$t_0 := \max\left\{1, \left\lceil \frac{q}{H_n} \right\rceil\right\}, \qquad s := \lceil \sqrt{q} \rceil.$$

Proposition 4.5 (Finite arbitrary-output sandwich). *For $1 \le q \le n^2$ and each $\star \in \{\mathbf{R}, \underline{\mathbf{R}}, \widetilde{\mathbf{R}}\}$,*

$$\max\left\{n\lceil \sqrt{q} \rceil,\ \widetilde{\mathbf{R}}\big(\langle t_0, n^2, t_0\rangle\big)^{1/2}\right\} \le \mathsf{PMM}^{\star}_{n}(q) \le \star(\langle s, n, s\rangle)\,. \tag{18}$$

Proof. Degenerate to a Ferrers shape, extract the harmonic rectangle, and balance its tensor. Independently, joint column/row spans of dimensions a, b satisfy $q \le ab$ and give input flattenings of ranks na, nb, proving $n\lceil \sqrt{q} \rceil$. A q-plane inside an $s \times s$ output block gives the upper bound; see Appendix A.2. □

4.5 Rectangular exponents and Bini exactification

Define

$$f(\alpha) := \omega(\alpha, 1, \alpha), \qquad 0 \le \alpha \le 1.$$

Lemma 4.6 (2-Lipschitz continuity)**.** *For* $0 \le \alpha \le \beta \le 1$*,*

$$0 \le f(\beta) - f(\alpha) \le 2(\beta - \alpha). \tag{19}$$

Proof. Restriction gives monotonicity; partitioning the larger output into row and column blocks gives $f(\beta) \le f(\alpha) + 2(\beta - \alpha)$. □

Lemma 4.7 (Rectangular Bini regularization and exactification)**.** *For every fixed* $\alpha \in [0, 1]$*,*

$$\mathbf{R}_{\mathbb{C}}(\langle \lceil n^\alpha \rceil, n, \lceil n^\alpha \rceil \rangle) = n^{f(\alpha)+o(1)}. \tag{20}$$

Consequently ordinary, border, and asymptotic rank have the same exponent for these rectangular families.

Proof. The rank hierarchy gives the lower bound. For the upper bound, fix a base size and a tensor power with border rank close to its asymptotic rank. Interpolation of a fixed degeneration gives only an $O(k)$ overhead on its kth power (Bini et al., 1980; Bürgisser et al., 1997); padding covers every size. Appendix A.2 specifies the order of choices and the exact interpolation. □

Theorem 4.8 (Arbitrary linear-output matrix-product frontier)**.** *Let* $q = q(n)$ *be integer-valued, with* $1 \le q(n) \le n^2$*, and satisfy* $q = n^{\gamma+o(1)}$ *for some* $0 \le \gamma \le 2$*. Then, for each* $\star \in \{\mathbf{R}, \underline{\mathbf{R}}, \widetilde{\mathbf{R}}\}$*,*

$$\mathsf{PMM}^\star_n(q) = n^{f(\gamma/2)+o(1)} = n^{\omega(\gamma/2,1,\gamma/2)+o(1)}. \tag{21}$$

For $q = 0$*,* $\mathsf{PMM}^\star_n(0) = 0$ *exactly.*

Proof. For $\gamma > 0$, $t_0 = n^{\gamma+o(1)}$ and $s = n^{\gamma/2+o(1)}$. Homogeneity turns the lower exponent into $\frac{1}{2}\omega(\gamma, 2, \gamma) = f(\gamma/2)$; Lemmas 4.6–4.7 control rounding and give the upper exponent. At $\gamma = 0$ the lower dot-product restriction costs n, matching $f(0) = 1$. The complete endpoint squeeze is in Appendix A.2. □

Corollary 4.9 (Uniform frontier and near-full storage barrier)**.** *Uniformly over integers* $1 \le q \le n^2$ *and each of the three rank notions,*

$$\log_n \mathsf{PMM}^\star_n(q) = f\left(\frac{\log_n q}{2}\right) + o(1). \tag{22}$$

Consequently, any exact analytic scheme for AB *with* $q = (n^2 - M)_+ \ge 1$ *obeys*

$$Q \ge c_{\mathcal{G}}^{-1} n \lceil \sqrt{q} \rceil, \qquad Q \ge q\, n^{\omega-2-o(1)}. \tag{23}$$

For fixed $\delta > 0$*,* $M \le (1 - \delta)n^2$ *forces* $Q \ge n^{\omega-o(1)}$*. For fixed* $\varepsilon > 0$*, achieving* $Q \le n^{\omega-\varepsilon+o(1)}$ *requires* $M \ge n^2 - n^{2-\varepsilon+o(1)}$*, without any power-law assumption on* M*.*

Proof. If uniformity failed, a subsequence of $\log_n q \in [0, 2]$ would converge and contradict Theorem 4.8 and continuity of f. Since $f(\alpha) \ge \omega - 2 + 2\alpha$ by Lemma 4.6, (22) and the finite flattening bound give (23). Rearrangement gives the storage statements; $q = 0$ satisfies the necessary storage condition trivially. □

5 Two costs in a commit-before-key gradient

This section maps one concrete commit-before-key protocol to the algebraic model. All loss and gradient calculations are over $\mathbb{R}$ with the Frobenius pairing. The Borel and tensor lower bounds are over $\mathbb{C}$; Appendix A.4 transfers them to the real protocol.

5.1 AN INVERTIBLE COTANGENT FOR SYMMETRIC INFONCE

For $N \geq 2$, let $Q, K \in \mathbb{R}^{N\times N}$ have rows q_i, k_i, and define the unnormalized symmetric loss

$$L_{\mathrm{sym}}(Q,K) := \sum_{i=1}^{N} \log \sum_{j=1}^{N} e^{\beta q_i^\top k_j} - \beta \sum_{i=1}^{N} q_i^\top k_i + \sum_{j=1}^{N} \log \sum_{i=1}^{N} e^{\beta q_i^\top k_j} - \beta \sum_{j=1}^{N} q_j^\top k_j, \qquad \beta > 0. \tag{24}$$

Let P_{r} be the row-softmax of $\beta QK^\top$ and P_{c} the column-softmax, so the latter is normalized down each column. Direct differentiation gives

$$G(Q,K) := \nabla_Q L_{\mathrm{sym}}(Q,K) = \beta(P_{\mathrm{r}} + P_{\mathrm{c}} - 2I_N)K. \tag{25}$$

The row-only coefficient has a forced right null direction, $(P_r - I)\mathbf{1} = 0$, and the column-only coefficient a forced left one. Each yields a rank-deficient cotangent; symmetrization permits full row rank.

Lemma 5.1 (An explicit invertible cotangent)**.** *The finite witness $(Q(t), K(t))$ in Appendix A.3 has invertible G for every $t > 0$ and $\beta > 0$. Full row rank persists on a key neighborhood.*

Proof. For this witness, $P_r = \mathbf{1}p^\top$ and $P_c = p\mathbf{1}^\top$, where $p = (e^{2\beta t}, 1, \ldots, 1)^\top/(e^{2\beta t} + N - 1)$. The determinant lemma gives $\det(P_r + P_c - 2I) = (-2)^{N-2}(1 - N\|p\|_2^2) \neq 0$, and $\det K(t) = \pm t$. Rank is open. □

5.2 AN EXPLICIT RESIDUAL RELU MATRIX-PRODUCT SLICE

Take the explicit witness of Lemma 5.1 at $N = n$ and $t \geq 1$, padded to $d = 2n$:

$$\overline{Q} = [Q_*, 0], \qquad \overline{K} = [K_*, 0] \in \mathbb{R}^{n\times 2n}.$$

The slice is a two-layer residual network on an input $X \in \mathbb{R}^{n\times d}$,

$$H = \sigma\big(XW_1 + \mathbf{1}_n b_1^\top\big) \in \mathbb{R}^{n\times m}, \qquad Y = X + HW_{2,0} \in \mathbb{R}^{n\times d}, \tag{26}$$

with $W_1 \in \mathbb{R}^{d\times m}$, $W_{2,0} \in \mathbb{R}^{m\times d}$, $b_1 \in \mathbb{R}^m$, and $m = 4n$. Set

$$W_{2,0} = \begin{bmatrix} I_d \\ I_d \end{bmatrix}, \qquad W_{2,0}^+ = \tfrac{1}{2}[I_d, I_d]. \tag{27}$$

Thus $W_{2,0}^+ W_{2,0} = I_d$. For isometric $E \in \mathbb{R}^{d\times n}$, set

$$N_0 = 2^{-1/2} \begin{bmatrix} E \\ -E \end{bmatrix} \in \mathbb{R}^{m\times n}. \tag{28}$$

Then $N_0^\top N_0 = I_n$, $W_{2,0}^\top N_0 = W_{2,0}^+ N_0 = 0$, and $\mathbf{1}_m^\top N_0 = 0$.

For hard variables $A, B \in \mathbb{R}^{n\times n}$, use

$$X(A) = [Q_*, A], \qquad L_0 = \mathrm{diag}(0_n, -I_n),$$

and the independent slice

$$W_1(B) = L_0 W_{2,0}^+ + \begin{bmatrix} 0_{n\times m} \\ BN_0^\top \end{bmatrix}. \tag{29}$$

Choose any $\rho > 0$ and take $b_1 = \rho\mathbf{1}_m$. At

$$Q_\rho = [Q_*, 0] + 2\rho\mathbf{1}_{n\times d}, \qquad K = \overline{K} \tag{30}$$

the determinant calculation in Appendix A.3 shows that $G_\rho := G(Q_\rho, \overline{K})$ has row rank n for every $\rho > 0$. Shrink a product neighborhood of $(A, B) = (0, 0)$ so that

$$\big\|[0, -A]W_{2,0}^+ + ABN_0^\top\big\|_\infty < \rho \tag{31}$$

holds, where $\|\cdot\|_\infty$ is the maximum absolute entry. This sufficient condition makes every pre-activation strictly positive. Since $\rho\mathbf{1}_{n\times m}W_{2,0} = 2\rho\mathbf{1}_{n\times d}$, ReLU is the identity and

$$H(A,B) = \rho\mathbf{1}_{n\times m} + [0,-A]W_{2,0}^{+} + ABN_0^\top,$$
$$X(A) + H(A,B)W_{2,0} = Q_\rho, \tag{32}$$

using the displayed nullspace identities. At $W_2 = W_{2,0}$,

$$\nabla_{W_2} L_{\mathrm{sym}} = H(A,B)^\top G_\rho.$$

Since $H(A,B)N_0 = AB$, a fixed right inverse $G_\rho J_\rho = I_n$ gives

$$\boxed{N_0^\top(\nabla_{W_2} L_{\mathrm{sym}})J_\rho = (AB)^\top.} \tag{33}$$

The coordinates W_1, W_2 are independent, and the complexified mixed jet is taken on this strict real chamber.

5.3 The uniform commit-before-key protocol

In the commit-before-key protocol, (A,B) precede the key and only $S(A,B) \in \mathbb{R}^M$ crosses the cut; after a key near $\overline{K}$, one decoder must output $\nabla_{W_2} L_{\mathrm{sym}}$ for every key (raw A, B may be reread).

At the fixed key $\overline{K}$, the projection $Z \mapsto (N_0^\top Z J_\rho)^\top$ recovers AB by (33). The hard-slice lemma and storage/output duality therefore give:

Theorem 5.2 (Fixed-key recomputation lower bound)**.** *Any uniform exact commit-before-key protocol containing this slice satisfies, over the complexified mixed jet,*

$$Q \geq c_{\mathcal{G}}^{-1}\mathsf{CQR}_M^{\mathbf{R}}(\mu_n) = c_{\mathcal{G}}^{-1}\mathsf{PMM}_n^{\mathbf{R}}((n^2-M)_+), \qquad x_+ := \max(x,0). \tag{34}$$

For $q = (n^2-M)_+ = n^{\gamma+o(1)} \geq 1$,

$$Q \geq n^{\omega(\gamma/2,1,\gamma/2)-o(1)}$$

and Corollary 4.9 applies. At $M \geq n^2$ this quotient is zero, not necessarily the cost of the full varying-key gradient.

Proof. Apply the hard-slice lemma with (33); the finite tradeoff gives the quotient inequality, and the dual bridge plus Theorem 4.8 gives the exponent. □

Complexifying a real jet cannot increase rank, so the complex tensor lower bounds apply to real protocols (Appendix A.4). Saved activations, products, and partial gradients all count toward M; raw A, B and fixed network constants may be reread.

5.4 A genuine late-key separation

Theorem 5.3 (Unavoidable late-key arithmetic)**.** *For the same hard family with keys ranging over a neighborhood of $\overline{K}$, every uniform exact analytic protocol satisfies, for every P, M,*

$$Q \geq c_{\mathcal{G}}^{-1}\mathbf{R}_{\mathbb{C}}\langle n,n,n-1\rangle. \tag{35}$$

In particular, $Q \geq n^{\omega-o(1)}$. With scalar multiplication, exponentiation, and reciprocal gates, the optimal late exponent is ω for every checkpoint budget, attained even with no preprocessing. If the key is instead available before commit, storing the $8n^2$ gradient entries permits zero late gates.

Proof sketch. Put $C_\rho = \beta(P_r + P_c - 2I)$ at $(Q_\rho, \overline{K})$; it is invertible because $G_\rho = C_\rho[K_*, 0]$ has row rank n. Let $F = [0, I_n]^\top$ and choose $R \in \mathbb{R}^{n\times(n-1)}$ with $R^\top R = I$ and $\mathbf{1}^\top R = 0$. For late $V \in \mathbb{R}^{n\times(n-1)}$, set

$$K(V) = \overline{K} + [0, C_\rho^{-1}VR^\top].$$

All scores remain constant, while $G(Q_\rho, K(V))FR = V$. For $Z = \nabla_{W_2} L_{\mathrm{sym}}$ a fixed projection is

$$F^\top W_{2,0}^\top Z F R = (2\rho\mathbf{1}_{n\times n} - A)^\top V. \tag{36}$$

Fix $B = 0$. Its mixed A, V derivative is $-A^\top V$, and Lemma 2.2 gives the lower bound independently of stored state. Computing AB, the softmax cotangent, and $H^\top G$ directly takes $n^{\omega+o(1)}$ late gates. Appendix A.5 gives the complete proof and the real-field upper bound. □

This is an exact separation for the specified protocol family: storage can eliminate the fixed-key recomputation, but not the arithmetic involving a genuinely late key. It does not assert a lower bound for unrestricted training schedules or normalized InfoNCE.

6 EXPERIMENTS

Construction and conditioning. We reran all 16,000 CPU PyTorch trials over five sizes $n \in \{2, 4, 8, 16, 32\}$, ten seeds, two precisions, five cotangent scales, four bias ratios, and eight chamber radii. Each baseline uses the same $(Q_\rho, \overline{K}, G_\rho, J_\rho)$ for differentiation and recovery. All 12,000 rows before the first chamber boundary pass the structural checks. On radii $\lambda \in [0.25, 0.9]$, float64 passes $4{,}000/4{,}000$ at 10^{-9} (worst error 2.55×10^{-11}); float32 passes $3{,}786/4{,}000$ at 10^{-3}. All 214 float32 misses have target norm at most 1.29×10^{-5}. We retain these failures rather than reinterpret a conditioning diagnostic as a certified tolerance. Figure 1 shows the canonical setting.

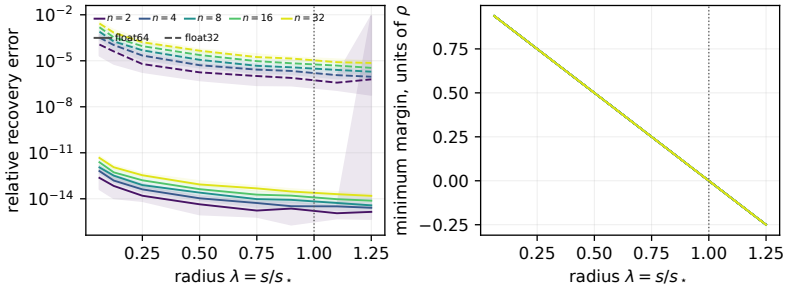


Figure 1: Recovery error and actual ReLU margin at $t = 16$, $\rho/\sqrt{t} = 2^{-8}$. Curves are seed medians with 10–90% bands; color denotes n and line style denotes precision. The dotted line is the first chamber boundary, not a guarantee about every later radius.

The 700 ablations test the fixed decoder under a column-space projection, embedding normalization, and perturbed keys (Table 1). Pointwise refits diagnose decoder mismatch, not a uniform extension. Chamber exit creates large errors in some, not all, instances. All 50 new constant-score key tests pass, with maximum mixed-product error 8.19×10^{-14}; ten exact symbolic identities at $n = 2, 3$ also pass.

Arithmetic and causal accounting. All 140 admissible routes in the 165-case int64 sweep have zero residual and costs (M, nM, nq); the 25 key-aware routes fail the commit-independence audit. At $q = n^2/4$, late classical work is $n^3/4$: a constant-factor saving, not an exponent improvement. Equal-area shapes have equal classical counts. These checks audit identities and information flow, not optimal tensor ranks. Appendix B specifies the full reproducible workflow.

7 RELATED WORK

Subspace minimization already appears in border substitution. For a concise tensor, Landsberg & Michałek (2025, Eq. (2.1)) bound border rank using $\min_{\dim E=q}(a - q + \underline{\mathbf{R}}(T|_E))$. At fixed q, the penalty is constant: for the output factor of matrix multiplication this contains precisely $n^2 - q + \mathsf{PMM}_n^{\mathbf{R}}(q)$. Our contribution is the checkpoint interpretation and the matching exponent evaluation of this existing type of minimization, not the removal of its penalty.

Restriction, degeneration, partial products, and rectangular exponents are classical (Strassen, 1987; Bürgisser et al., 1997; Schönhage, 1981). We use Borel fixed points, tensor balancing, and Bini exactification as tools, with proofs specialized to the full-Grassmannian output problem. Support functionals and support-rank/weighted-product results optimize different objects (Strassen, 1991; Bläser et al., 2018; Vrana, 2024); prescribed support alone does not determine ordinary tensor rank. Matrix-space rank bounds are complementary (Atkinson & Stephens, 1979; Bini, 1986; Bshouty, 1990). The two mixed-jet arguments identify exactly which early–early and early–late obstructions these algebraic tools control.

AI USE STATEMENT

Generative AI tools assisted with research-direction exploration, mathematical definitions and candidate results, proof development and checking, literature searches, experimental design and implementation, interpretation of results, and manuscript and artifact preparation. The supplementary material records executable finite checks and the full experimental rerun underlying the displayed results. These checks do not replace mathematical proofs or certify novelty. The authors are responsible for the correctness, originality, citations, code, and presentation of the submission, including all AI-assisted content.

REPRODUCIBILITY STATEMENT

The main text and Appendix A give complete mathematical arguments. The anonymous supplementary package includes all experiment scripts, raw and aggregate CSV files for the $16{,}000$ baseline, 700 ablation, 165 accounting, and 50 late-key trials, exact symbolic checks, pinned dependencies and a data-to-manuscript generation script. Appendix B specifies every grid, seed rule, ablation, metric, and threshold. Figure 1, both tables, and the numerical statements are generated or audited against this bundled rerun. Numerical checks test finite identities and conditioning; they do not establish tensor-rank or exponent lower bounds.

REFERENCES

M. D. Atkinson and N. M. Stephens. On the maximal multiplicative complexity of a family of bilinear forms. *Linear Algebra and its Applications*, 27:1–8, 1979. doi: 10.1016/0024-3795(79)90026-0.

Dario Bini. Border rank of $m \times n \times (mn - q)$ tensors. *Linear Algebra and its Applications*, 79: 45–51, 1986. doi: 10.1016/0024-3795(86)90291-0.

Dario Bini, Grazia Lotti, and Francesco Romani. Approximate solutions for the bilinear form computational problem. *SIAM Journal on Computing*, 9(4):692–697, 1980. doi: 10.1137/0209053.

Markus Bläser, Matthias Christandl, and Jeroen Zuiddam. The border support rank of two-by-two matrix multiplication is seven. *Chicago Journal of Theoretical Computer Science*, 2018(5), 2018. doi: 10.4086/cjtcs.2018.005.

Armand Borel. *Linear Algebraic Groups*. Springer, second edition, 1991. doi: 10.1007/978-1-4612-0941-6.

Nader H. Bshouty. Maximal rank of $m \times n \times (mn - k)$ tensors. *SIAM Journal on Computing*, 19 (3):467–471, 1990. doi: 10.1137/0219031.

Peter Bürgisser, Michael Clausen, and M. Amin Shokrollahi. *Algebraic Complexity Theory*. Springer, 1997. doi: 10.1007/978-3-662-03338-8.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

David Eisenbud. *Commutative Algebra with a View Toward Algebraic Geometry*. Springer, 1995. doi: 10.1007/978-1-4612-5350-1.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. Scaling deep contrastive learning batch size under memory limited setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP*, pp. 316–321, 2021. doi: 10.18653/v1/2021.repl4nlp-1.31.

Andreas Griewank and Andrea Walther. Algorithm 799: Revolve: An implementation of checkpointing for the reverse or adjoint mode of computational differentiation. *ACM Transactions on Mathematical Software*, 26(1):19–45, 2000. doi: 10.1145/347837.347846.

Paras Jain, Ajay Jain, Aniruddha Nrusimha, Amir Gholami, Pieter Abbeel, Kurt Keutzer, Ion Stoica, and Joseph E. Gonzalez. Checkmate: Breaking the memory wall with optimal tensor rematerialization. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 497–511, 2020.

Joseph M. Landsberg and Mateusz Michałek. Towards finding hay in a haystack: Explicit tensors of border rank greater than $2.02m$ in $\mathbb{C}^m \otimes \mathbb{C}^m \otimes \mathbb{C}^m$. *Theory of Computing*, 21(13):1–17, 2025. doi: 10.4086/toc.2025.v021a013.

Arnold Schönhage. Partial and total matrix multiplication. *SIAM Journal on Computing*, 10(3): 434–455, 1981. doi: 10.1137/0210032.

Volker Strassen. Relative bilinear complexity and matrix multiplication. *Journal für die reine und angewandte Mathematik*, 375/376:406–443, 1987. doi: 10.1515/crll.1987.375-376.406.

Volker Strassen. Degeneration and complexity of bilinear maps: Some asymptotic spectra. *Journal für die reine und angewandte Mathematik*, 413:127–180, 1991.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. doi: 10.48550/arXiv.1807.03748.

Péter Vrana. Partial and weighted matrix multiplication. *arXiv preprint arXiv:2408.15728*, 2024. doi: 10.48550/arXiv.2408.15728.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=zeFrfgyZln.

A COMPLETE PROOF DETAILS

This appendix gives complete proofs and the field comparisons used in the main text. All computations use the information-flow model of Section 2.

A.1 MIXED JETS AND THE FINITE NORMAL FORM

Proof of Lemma 3.1. Order the scalar nodes topologically. For a nonlinear node g with parent tuple z_g, let J_g^x, J_g^y be the Jacobians of that tuple with respect to x, y at the base point, let H_g be the Hessian of its scalar gate function, and let $\sigma_g \in Z$ be the reverse output sensitivity of that node. Second-order reverse accumulation gives the node contribution

$$\sigma_g \otimes \big((J_g^x)^\top H_g J_g^y\big).$$

The parent tuple has dimension at most a, so H_g has rank at most a. Factoring it into at most a rank-one terms writes this contribution as at most a simple tensors in $X^* \otimes Y^* \otimes Z$. Linear nodes have zero local Hessian. Summing the contributions proves the claim with $c_{\mathcal{G}} = a$ (or a larger library-dependent constant when several scalar primitives are grouped into one gate). □

Proof of Lemma 2.2. Translate a regular (u_0, k_0) to $(0, 0)$ and let $s_0 = S(0)$ and $A = D_u S(0)$. Define

$$\overline{D}(u, k) = D(u, k, s_0 + Au).$$

In the second-order chain rule for $D(u, k, S(u))$, the only possible term involving second derivatives of S is the output sensitivity $D_s D$ applied to that second derivative. Its mixed u, k component vanishes because S is independent of k. The first derivatives of S agree with those of $s_0 + Au$. Therefore

$$D^2_{uk}\Phi(0, 0) = D^2_{uk}\overline{D}(0, 0).$$

The new circuit has at most Q non-affine gates, since all added coefficients are fixed derivatives at the base point. Lemma 3.1 proves the rank bound. The argument allows any number of stored coordinates and does not charge their preprocessing. □

Proof of Theorem 3.2. Choose a regular base point and translate it to $(0, 0)$. Put

$$s_0 = S(0, 0), \qquad A = D_x S(0, 0), \qquad B = D_y S(0, 0), \qquad C = D^2_{xy} S(0, 0).$$

The mixed-jet lemma applied to the vector output S gives $\mathbf{R}(C) \le c_{\mathcal{G}} P$. Define the late circuit with an affine summary input

$$\overline{D}(x, y) := D(x, y, s_0 + Ax + By).$$

The second-order chain rule at the base point separates the only term involving the mixed derivative of S:

$$D^2_{xy}\big[D(x, y, S(x, y))\big] = D_s D(0, 0, s_0) \circ C + D^2_{xy}\overline{D}(0, 0).$$

All terms involving $D_x S$ or $D_y S$ are already present in the affine-summary circuit. Translating the base point adds only affine and constant terms, so the mixed derivative on the left is the original bilinear tensor T. Set

$$U := D_s D(0, 0, s_0) \circ C, \qquad V := D^2_{xy}\overline{D}(0, 0).$$

The map U factors through $\mathbb{F}^M$, hence $\dim \operatorname{im} U \le M$ and $\mathbf{R}(U) \le c_{\mathcal{G}} P$. Adding affine wires does not add nonlinear nodes to the late circuit, so the mixed-jet lemma gives $\mathbf{R}(V) \le c_{\mathcal{G}} Q$. □

Proof of Proposition 3.3. Tensor rank is subadditive, so the normal form gives $\mathbf{R}(T) \le c_{\mathcal{G}}(P + Q)$. If $W = \operatorname{im} U$, then $\pi_W T = \pi_W V$ and output projection cannot increase ordinary rank. Taking the minimum over all W of dimension at most M gives the second inequality. □

Proof of Proposition 3.4. First keep P finite. Every separated bilinear preprocessing circuit with at most P multiplications yields a checkpoint tensor

$$S = \sum_{i=1}^{P} a_i \otimes b_i \otimes s_i.$$

A late linear readout of the checkpoint and at most Q new products writes $T = LS + V$, with $\mathbf{R}(V) \le Q$. Put $U = LS$. Then $\dim \operatorname{im} U \le M$ and $\mathbf{R}(U) \le P$, so $\mathsf{F}_{P,M}(T) \le Q$.

Conversely, take a minimizing tensor U in (10). The set is nonempty because it contains $U = 0$, and its attainable objective values are nonnegative integers. Let $W = \operatorname{im} U$, choose a basis for W, and compute the $\mathbf{R}(U) \le P$ separated products in a rank decomposition of U. Free linear combinations yield its $\dim W \le M$ coordinates, which are stored. Late computation of a rank decomposition of $T - U$, followed by free addition of the stored component, uses exactly $\mathbf{R}(T - U)$ multiplications.

With unrestricted P, any decomposition $T = U + V$ of this kind satisfies $\mathbf{R}(\pi_W T) \le \mathbf{R}(V)$ for $W = \operatorname{im} U$. Thus the late cost is at least $\mathsf{CQR}^{\mathbf{R}}_M(T)$. Conversely, choose a minimizing quotient subspace W, a linear section $\sigma : Z/W \to Z$, and

$$V = \sigma \pi_W T, \qquad U = T - V.$$

Then $\operatorname{im} U \subseteq W$. Unrestricted preprocessing stores coordinates of U, while $\mathbf{R}(V) = \mathbf{R}(\pi_W T)$: one inequality follows from σ, and the other from $\pi_W \sigma = \mathrm{Id}$. This proves the saturated equivalence. Theorem 3.2 supplies an admissible U with $\mathbf{R}(U) \le \lfloor c_{\mathcal{G}} P \rfloor$ and $\mathbf{R}(T - U) \le c_{\mathcal{G}} Q$ for every analytic scheme, proving its stated lower bound. □

A.2 Arbitrary-output matrix multiplication

Proof of Proposition 4.1. If $W \le M_n(\mathbb{C})$ has dimension $m \le M$, its annihilator under the trace pairing is $E = W^{\perp}$ of dimension $n^2 - m \ge q$. Restricting the output dual from E to any q-dimensional subspace cannot increase ordinary, border, or asymptotic rank. Thus the infimum is unchanged when W is enlarged to dimension M, equivalently when E is restricted to dimension q. Conversely, every q-plane is $W^{\perp}$ for the M-plane $W = E^{\perp}$; $q = 0$ is the zero quotient. □

Proof of Lemma 4.2. Consider the projective orbit closure

$$\mathcal{Y} = \overline{(GL_n \times GL_n) \cdot E} \subseteq \mathrm{Gr}(q, M_n).$$

It is nonempty and projective and is stable under $B_+ \times B_-$. The Borel fixed-point theorem supplies a fixed subspace $E_0 \in \mathcal{Y}$. The diagonal torus has distinct characters $p_i q_j^{-1}$ on the matrix units, so E_0 is spanned by a set S of matrix units. For $a < b$, invariance under the elementary upper and lower unipotents gives

$$(I + tE_{ab})E_{bj} = E_{bj} + tE_{aj}, \qquad E_{ib}(I + tE_{ba})^{-1} = E_{ib} - tE_{ia}.$$

Consequently S is closed under moving north and west, hence is a northwest Ferrers diagram.

For the tensor statement, choose an algebraic curve through E_0 with generic point in the dense orbit, normalize it, and take a local discrete valuation ring (DVR) at a point above E_0. After a finite extension if necessary, the generic orbit point lifts to a pair (g, h) in $GL_n \times GL_n$. Pulling back the tautological bundle supplies a free rank-q module, whose fibers define the restricted tensors. The generic fiber is isomorphic to T_E: the changes $A \mapsto g^{\top} A$ and $B \mapsto B h^{-\top}$ realize the action, and a basis of the pulled-back bundle fixes the third leg. This is a tensor degeneration. It implies $\underline{\mathbf{R}}(T_S^{\otimes m}) \le \underline{\mathbf{R}}(T_E^{\otimes m})$ for each $m \ge 1$, so taking infima of the mth roots gives the asserted asymptotic inequality. □

Proof of Lemma 4.3. Let $A = \max_i i\mu_i$. Since $\mu_i \le A/i$,

$$q = \sum_{i=1}^{n} \mu_i \le A H_n.$$

Choose p attaining the maximum and put $r = \mu_p$. The first p rows all have length at least r, so they contain a $p \times r$ rectangle with $pr = A \ge q/H_n$. □

Proof of Lemma 4.4. Transposition and permutation of tensor legs identify $\langle p, n, r \rangle$ with $\langle r, n, p \rangle$. Their tensor product is

$$\langle p, n, r \rangle \otimes \langle r, n, p \rangle \cong \langle pr, n^2, pr \rangle.$$

Submultiplicativity of border rank and the definition of asymptotic rank give the displayed inequality. Ordinary and border rank dominate asymptotic rank, so the same lower bound applies to both finite levels. □

Proof of Proposition 4.5. For any q-plane E, Borel–Ferrers degeneration and the harmonic rectangle yield a contained $\langle p, n, r\rangle$ with $pr \geq q/H_n$. Tensor balancing and coordinate restriction give the asymptotic-rank lower bound in (18), hence the same bound for ordinary and border rank.

For the independent flattening bound, let

$$L = \operatorname{span}\{Cz : C \in E,\ z \in \mathbb{C}^n\}, \qquad R = \operatorname{span}\{C^\top z : C \in E,\ z \in \mathbb{C}^n\},$$

with $a = \dim L$ and $b = \dim R$. Since $E \subseteq L \otimes R$ under the matrix-space identification, $q \leq ab$. Write $A_{:k}$ for the kth column of A. The A-flattening sends A to the bilinear form

$$(B, C) \longmapsto \sum_{k=1}^{n} \sum_{i,j=1}^{n} C_{ij} A_{ik} B_{kj}.$$

It vanishes exactly when $C^\top A_{:k} = 0$ for every $C \in E$ and every k. This kernel has dimension $n(n-a)$, so the flattening rank is na. The analogous B-flattening has rank nb. Flattening ranks bound ordinary and border rank and multiply under tensor powers, so they also bound asymptotic rank. Since $\max(a, b) \geq \lceil\sqrt{q}\rceil$, every T_E has complexity at least $n\lceil\sqrt{q}\rceil$ for all three notions.

For the upper bound choose q coordinate outputs inside an $s \times s$ block, $s = \lceil\sqrt{q}\rceil$. Compute the block $\langle s, n, s\rangle$ and project to the selected coordinates. This also proves the exact endpoint $\mathsf{PMM}^{\star}_n(1) = n$. □

Proof of Lemma 4.6. Restriction gives $f(\alpha) \leq f(\beta)$. Partition the $\lceil n^\beta\rceil \times \lceil n^\beta\rceil$ output into at most $L = n^{2(\beta-\alpha)+o(1)}$ blocks of side lengths at most $\lceil n^\alpha\rceil$. The larger tensor is a restriction of a direct sum of L copies of the α-rectangle tensor T. For fixed n, taking tensor powers of this direct sum gives

$$\underline{\mathbf{R}}\big((T^{\oplus L})^{\otimes m}\big) \leq L^m \underline{\mathbf{R}}(T^{\otimes m}),$$

so $\widetilde{\mathbf{R}}(T^{\oplus L}) \leq L\widetilde{\mathbf{R}}(T)$. Applying (3) proves $f(\beta) \leq f(\alpha) + 2(\beta - \alpha)$. This argument uses only tensor powers and does not presuppose ordinary-rank exactification. □

Proof of Lemma 4.7. The ordinary-rank lower bound follows from the rank hierarchy and the full-sequence asymptotic-rank convention. For the upper bound fix $\eta > 0$. Choose an integer $n_0 \geq 2$ so large that

$$T_{n_0} := \langle\lceil n_0^\alpha\rceil, n_0, \lceil n_0^\alpha\rceil\rangle, \qquad \widetilde{\mathbf{R}}(T_{n_0}) \leq n_0^{f(\alpha)+\eta}.$$

By the definition $\widetilde{\mathbf{R}}(T) = \inf_{m\geq 1} \underline{\mathbf{R}}(T^{\otimes m})^{1/m}$, choose a fixed positive integer m with

$$S = T_{n_0}^{\otimes m} = \langle A_0, B_0, A_0\rangle, \quad B_0 = n_0^m, \quad A_0 = \lceil n_0^\alpha\rceil^m, \quad r := \underline{\mathbf{R}}(S) \leq B_0^{f(\alpha)+2\eta}.$$

This step uses a tensor power; it does not assume a one-copy border-rank bound from the definition of f. A finite-order border degeneration, after finite reparametrization, clearing poles in its rank-one factors, and truncating them, gives a fixed polynomial family

$$F(\varepsilon) = \varepsilon^d S + O(\varepsilon^{d+1}), \qquad \deg F \leq D,$$

whose ordinary rank is at most r at every nonzero parameter (Bini et al., 1980; Bürgisser et al., 1997). The coefficient of ε^{dk} in $F(\varepsilon)^{\otimes k}$ is exactly $S^{\otimes k}$. Evaluating at $kD + 1$ distinct nonzero complex parameters and applying Vandermonde interpolation yields

$$\mathbf{R}(S^{\otimes k}) \leq (kD + 1)r^k. \tag{37}$$

Now let k be the least integer with $B_0^k \geq N$. Then $B_0^k < B_0 N$ and $A_0^k \geq (B_0^k)^\alpha \geq N^\alpha$. The matrix side lengths of $S^{\otimes k}$ therefore dominate $(\lceil N^\alpha\rceil, N, \lceil N^\alpha\rceil)$, so coordinate restriction gives

$$\mathbf{R}(\langle\lceil N^\alpha\rceil, N, \lceil N^\alpha\rceil\rangle) \leq (kD + 1)B_0^{k(f(\alpha)+2\eta)} \leq N^{f(\alpha)+2\eta+o(1)}.$$

Here n_0, m, B_0, D are fixed before $N \to \infty$, and $k = O(\log N)$. Letting $\eta \downarrow 0$ proves the upper exponent. The rank hierarchy gives the common exponent for all three rank notions. The interpolation is exact, not a finite-precision numerical approximation. □

Proof of Theorem 4.8. Set $\alpha = \gamma/2$. For $\alpha < 1$ and every fixed $0 < \varepsilon < 1 - \alpha$, the upper sandwich and Lemma 4.7 give

$$\limsup_{n\to\infty} \log_n \mathsf{PMM}^{\star}_n(q) \le f(\alpha + \varepsilon).$$

Let $\varepsilon \downarrow 0$ and apply Lemma 4.6. When $\alpha = 1$, use $s = \lceil \sqrt{q} \rceil \le n$ and Lemma 4.7 at $\alpha = 1$; no value of f outside $[0, 1]$ is needed. For $\gamma > 0$, $t_0 = n^{\gamma + o(1)}$. For fixed $0 < \delta < \gamma$, eventually $t_0 \ge \lceil n^{\gamma-\delta} \rceil$. The finite lower sandwich and coordinate restriction imply

$$\liminf_{n\to\infty} \log_n \mathsf{PMM}^{\star}_n(q) \ge \tfrac{1}{2}\omega(\gamma - \delta, 2, \gamma - \delta) = f((\gamma - \delta)/2).$$

Let $\delta \downarrow 0$. At $\gamma = 0$, $t_0 \ge 1$ gives the dot-product restriction $\langle 1, n^2, 1\rangle$, of asymptotic rank n^2. Thus the lower exponent is at least one, matching $f(0) = 1$ from the upper squeeze. The zero-output case is exactly the zero tensor. □

Proof of Corollary 4.9. Fix one of the three rank notions. Suppose the error in (22) did not tend to zero uniformly in q. There would be $\eta > 0$, a sequence $n_j \to \infty$, and $1 \le q_j \le n_j^2$ along which its absolute value is at least η. Pass to a subsequence with $\gamma_j = \log_{n_j} q_j \to \gamma \in [0, 2]$. The proof of Theorem 4.8 applies along this subsequence: the finite sandwich, harmonic loss, and rectangular exactification imply

$$\log_{n_j} \mathsf{PMM}^{\star}_{n_j}(q_j) \longrightarrow f(\gamma/2).$$

Continuity gives $f(\gamma_j/2) \to f(\gamma/2)$, a contradiction. There are only three rank notions, so a single error sequence $\eta_n \downarrow 0$ can bound them all.

The normal form, storage/output duality, and uniform frontier now give, for $q \ge 1$,

$$Q \ge c_{\mathcal{G}}^{-1} n^{f((\log_n q)/2) - \eta_n} \ge c_{\mathcal{G}}^{-1} q\, n^{\omega - 2 - \eta_n},$$

because $f(1) - f(\alpha) \le 2(1 - \alpha)$. Absorbing the fixed constant into $n^{-o(1)}$ proves the second bound in (23); the first follows directly from the finite flattening bound. If $q \ge \delta n^2$, the late exponent is at least ω. If $Q \le n^{\omega - \varepsilon + \zeta_n}$ with $\zeta_n \to 0$, rearrangement yields

$$q \le c_{\mathcal{G}} n^{2 - \varepsilon + \eta_n + \zeta_n} = n^{2 - \varepsilon + o(1)},$$

proving the necessary storage condition. When $q = 0$, $M \ge n^2$ and that condition is immediate. The implication is necessary, not sufficient, and applies without assuming any limit for $\log_n q$. □

Attainment at the exponent level. For the bilinear recomputation task, store the $n^2 - q$ coordinates outside a balanced block containing q selected outputs; recompute those outputs using the rectangular algorithm. Preprocessing can compute the entire product and retain only the desired checkpoint, so an early budget $P = n^{\omega + o(1)}$ already suffices for the saturated exponent upper bound. There is no assertion of a matching finite rank formula or of a matching exponent for every individual output subspace.

A.3 The InfoNCE cotangent and the ReLU slice

Proof of Lemma 5.1. For $t > 0$, let $r = \sqrt{t}$. Set

$$Q(t) = \begin{bmatrix} 2r & 2r & 0 & \cdots & 0 \\ 0 & 2r & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 2r & 0 & \cdots & 0 \end{bmatrix},$$

and define $K(t)$ by listing its columns as

$$\big[r\mathbf{1},\ re_1,\ e_2, \ldots, e_{N-1}\big].$$

Thus the key vectors are the rows of this matrix. Direct multiplication gives

$$Q(t)K(t)^{\top} = \begin{bmatrix} 4t & 2t\mathbf{1}^{\top} \\ 2t\mathbf{1} & 0 \end{bmatrix}.$$

Every row-softmax has the same row $p^\top$, and every column-softmax has the same column p, where

$$z = e^{2\beta t} > 1, \qquad p = \frac{(z, 1, \dots, 1)^\top}{z + N - 1}.$$

The matrix determinant lemma, applied to $-2I + \mathbf{1}p^\top + p\mathbf{1}^\top$, yields

$$\begin{aligned}\det(P_r + P_c - 2I) &= (-2)^N \det\begin{bmatrix} 1/2 & -\|p\|_2^2/2 \\ -N/2 & 1/2 \end{bmatrix} \\ &= (-2)^{N-2}(1 - N\|p\|_2^2).\end{aligned}$$

The last factor is strictly negative because p is a nonuniform probability vector. Also $\det K(t) = \pm t \neq 0$. Thus the cotangent is invertible at every finite $t > 0$, rather than only for sufficiently large t. Full row rank persists on a key neighborhood by continuity. □

Positive bias preserves rank for the chosen witness. For the padded witness at $t \geq 1$, write $r = \sqrt{t} \geq 1$. Adding 2ρ to every visible query entry adds the same vector to each score row. Hence $P_r = \mathbf{1}p_\rho^\top$ for a probability vector p_ρ, whereas $P_c = p\mathbf{1}^\top$ keeps the vector p above. Before the common factor β, the column logits defining p_ρ, up to a common shift, are

$$b_1 = 2t + 4\rho r, \qquad b_j = 2\rho(r+1)\ (2 \leq j \leq n-1), \qquad b_n = 2\rho r.$$

The middle range is empty for $n = 2$. Since $r \geq 1$ and $\rho \geq 0$, b_1 exceeds every other b_j, so $(p_\rho)_1 > 1/n$. Put $p = (a, b, \dots, b)^\top$ with $a > b$ and $a + (n-1)b = 1$. Then

$$p_\rho^\top p = b + (a - b)(p_\rho)_1 > 1/n.$$

Another application of the determinant lemma gives

$$\det(\mathbf{1}p_\rho^\top + p\mathbf{1}^\top - 2I) = (-2)^{n-2}(1 - np_\rho^\top p) \neq 0.$$

Therefore G_ρ has row rank n for *every* $\rho > 0$ at $t \geq 1$. The strict ReLU chamber is obtained independently by shrinking the neighborhood of $A = B = 0$; no infinitesimal-bias assumption is needed.

A.4 Real and complex tensor ranks

Proposition A.1 (Field bridge)**.** *Let T be a real tensor and $T_\mathbb{C}$ its complexification. For real border rank use Euclidean limits of real rank-bounded tensors, and define real asymptotic rank by their tensor powers. Then*

$$\star(T_\mathbb{C}) \leq \star(T), \qquad \star \in \{\mathbf{R}, \underline{\mathbf{R}}, \widetilde{\mathbf{R}}\}. \tag{38}$$

Hence a lower bound established over $\mathbb{C}$ for $T_\mathbb{C}$ bounds the real jet T that the protocol actually produces, so Theorem 5.2 applies to the real protocol and not only to its complexification.

Proof. Every real rank decomposition is also a complex one. Every Euclidean limit of real rank-at-most-r tensors lies in the complex Zariski closure of the rank-at-most-r locus, giving the border-rank inequality. Apply it to tensor powers and take infima for asymptotic rank. In particular, complexifying the real checkpoint subspace and its quotient cannot invalidate a complex lower bound. □

Real upper bounds from complex algorithms. For every real tensor T, ordinary ranks also satisfy

$$\mathbf{R}_\mathbb{R}(T) \leq 4\mathbf{R}_\mathbb{C}(T_\mathbb{C}).$$

Indeed, take a complex rank decomposition of the real tensor and take real parts term by term. For $a = a_0 + ia_1$, $b = b_0 + ib_1$, $c = c_0 + ic_1$, the real part of $a \otimes b \otimes c$ is

$$a_0 \otimes b_0 \otimes c_0 - a_1 \otimes b_1 \otimes c_0 - a_0 \otimes b_1 \otimes c_1 - a_1 \otimes b_0 \otimes c_1.$$

Thus exact real and complex matrix-multiplication algorithms have the same exponent, including fixed-aspect-ratio rectangles. This comparison justifies the real upper bound in Theorem 5.3.

A.5 CONSTANT-SCORE KEYS AND THE LATE-QUERY SEPARATION

Proof of Theorem 5.3. All matrices below are at the fixed parameters of Section 5. Put

$$C_\rho = \beta\big(P_r(Q_\rho, \overline{K}) + P_c(Q_\rho, \overline{K}) - 2I_n\big).$$

Because $G_\rho = C_\rho[K_*, 0]$ has row rank n and K_* is invertible, C_ρ is invertible. Let $F = [0, I_n]^\top \in \mathbb{R}^{2n\times n}$ select the last n coordinates. Choose $R \in \mathbb{R}^{n\times(n-1)}$ with $R^\top R = I_{n-1}$ and $\mathbf{1}_n^\top R = 0$. For an arbitrary sufficiently small late matrix $V \in \mathbb{R}^{n\times(n-1)}$, define

$$\Delta(V) = C_\rho^{-1} V R^\top, \qquad K(V) = \overline{K} + [0, \Delta(V)].$$

The key map is affine and sends a neighborhood of $V = 0$ into the admissible key neighborhood. It is independent of A, B. Since the last n coordinates of Q_ρ are all 2ρ,

$$Q_\rho(K(V) - \overline{K})^\top = 2\rho \mathbf{1}_n \mathbf{1}_n^\top R V^\top C_\rho^{-\top} = 0.$$

Hence the logits and both softmax matrices are exactly constant on this key plane. Formula (25) consequently gives

$$G(Q_\rho, K(V)) = C_\rho K(V) = [C_\rho K_*, V R^\top], \qquad G(Q_\rho, K(V))FR = V.$$

On the strict chamber, $Z = H^\top G$ and $HW_{2,0} = Q_\rho - X(A)$, so

$$\begin{aligned} F^\top W_{2,0}^\top Z F R &= (HW_{2,0}F)^\top G F R \\ &= (2\rho \mathbf{1}_{n\times n} - A)^\top V. \end{aligned}$$

This proves (36). In particular, at $B = 0$ its mixed A, V derivative is exactly $(A, V) \mapsto -A^\top V$, which has complex tensor rank $\mathbf{R}_{\mathbb{C}}\langle n, n, n-1\rangle$. Specializing $B = 0$, composing the late key input with the displayed affine map, and applying the fixed output projection add no nonlinear gates. The checkpoint remains independent of V. Lemma 2.2, followed by complexification, proves the finite lower bound for every P, M. The tensor contains $\langle n-1, n-1, n-1\rangle$ and is a restriction of $\langle n, n, n\rangle$, so its rank is $n^{\omega+o(1)}$.

For the upper bound, no stored state is needed. Recompute AB and form H by (32). The visible query is the fixed Q_ρ, so $Q_\rho K^\top$ is a free linear map of the late key. Form the row and column softmaxes using $O(n^2)$ exponentials, reciprocal gates, and scalar multiplications. Compute $G = \beta(P_r + P_c - 2I)K$ and $H^\top G$ using a fixed number of square matrix multiplications of side length n. The real/complex comparison above bounds the total by $n^{\omega+o(1)} + O(n^2) = n^{\omega+o(1)}$. This circuit works for all (A, B) in the chosen chamber and all keys in the full neighborhood, not just the special key plane.

Finally, if the same variable key is provided before commit, compute $Z \in \mathbb{R}^{4n\times 2n}$ early and store its $8n^2$ entries. The late decoder is the identity and has zero nonlinear gates. Thus at the same storage capacity the late-key and early-key protocols genuinely differ. Separately, at a *single fixed key* known in advance, storing just AB (n^2 scalars) suffices: the remaining terms of H are affine in reread A, and multiplication by fixed G_ρ is free. This explains why zero quotient rank for the fixed-key task does not contradict the full-family late lower bound. □

A.6 GENERAL NULLSPACE RECOVERY

Let $W_0 \in \mathbb{R}^{m\times d}$ have full column rank and choose

$$s \le m - d, \qquad W_0^\top N_0 = 0, \qquad N_0^\top N_0 = I_s. \tag{39}$$

With $W_0^+ = (W_0^\top W_0)^{-1} W_0^\top$, set

$$H = Y_{\text{vis}} W_0^+ + C N_0^\top. \tag{40}$$

At $W = W_0$,

$$Y = X + HW_0 = X + Y_{\text{vis}}.$$

If $G = \nabla_Y L$ has fixed right inverse J_G, then

$$N_0^\top (\nabla_W L) J_G = C^\top, \tag{41}$$

since $W_0^+ N_0 = 0$. Thus recovery is fixed when the visible output and late query are fixed on the hard slice.

B EXPERIMENTAL DETAILS

B.1 ARTIFACT AND EXECUTION ENVIRONMENT

The anonymous supplementary package includes the full source, executable experiments, current raw CSV files, generated figures and tables, exact checks, and dependency pins. The reported rerun used Python 3.13.5, PyTorch 2.10.0+cpu, NumPy 2.3.5, Matplotlib 3.10.8, and SymPy 1.14.0 on CPU; PyTorch experiments set one CPU thread. The package-version pins are in `requirements.txt`. From the package root:

```
python -m pip install -r requirements.txt
python scripts/run_hard_slice_experiment.py
python scripts/run_checkpoint_tradeoff.py
python scripts/check_late_key_separation.py
python scripts/proof_sanity_checks.py
python scripts/build_experiment_assets.py
sh paper/build.sh
```

The first two experiments also support `-quick`, which is a smoke test rather than the reported full sweep. Figures, tables, and the numerical audit are regenerated from the CSVs by `build_experiment_assets.py`. Finite-precision results can differ slightly across library versions and CPU backends; the paper reports the bundled rerun, not an assertion of bitwise agreement with a historical run. The exact int64 accounting contains no floating-point residual.

B.2 HARD-SLICE SWEEP AND NUMERICAL DEFINITIONS

The full Cartesian grid is

$$n \in \{2,4,8,16,32\}, \quad \text{seed} \in \{0,\dots,9\}, \quad \text{precision} \in \{\text{float64}, \text{float32}\},$$
$$t \in \{1,2,4,8,16\}, \qquad \eta = \rho/\sqrt{t} \in \{2^{-12}, 2^{-8}, 2^{-4}, 2^{-2}\},$$
$$\lambda = \varsigma/\varsigma_\star \in \{1/16, 1/8, 1/4, 1/2, 3/4, 0.9, 1.1, 1.25\}.$$

This gives 16,000 rows. All experiments use $\beta = 1$ and the concrete isometry $E = [I_n, 0]^\top$ in (28). The cotangent point is the witness in Appendix A.3. NumPy's default generator, seeded by $104729n + \text{seed}$, generates two Gaussian direction matrices in float64; each is Frobenius-normalized and then cast to the working precision. Thus both precisions share the same underlying directions. Set $A = \varsigma U$, $B = \varsigma V$.

The first positive root over all scalar entries of

$$\rho\mathbf{1} + \varsigma[0, -U]W_{2,0}^{+} + \varsigma^2(UV)N_0^\top$$

defines $\varsigma_\star$. The script computes it in float64 using the stable quadratic formula and treats coefficients below $64\epsilon_{64}\max(1, |c_2|, |c_1|, |c_0|)$ as numerically zero. For $0 \le \lambda < 1$, the exact pre-activations stay in the initial strict chamber. At or beyond the boundary, we inspect their actual minimum rather than infer all later chamber membership from λ alone.

The right inverse J is formed by an SVD, with numerical-rank threshold $\max(n, d)\epsilon\sigma_{\max}$. The principal error is

$$e_{\text{rec}} = \frac{\|N_0^\top(\nabla_{W_2} L)J_\rho - (AB)^\top\|_F}{\max(\|AB\|_F, \text{tiny}_{\text{dtype}})}.$$

The comparison product AB is computed in the same working precision; this is a numerical consistency metric, not comparison with an exact-real oracle. We also record the minimum pre-activation, its ratio to ρ, numerical rank and condition number of G, $\|GJ - I\|_F/\|I\|_F$, fixed-query residual $\|Q_{\text{out}} - Q_\rho\|_F/\|Q_\rho\|_F$, target and hidden norms, loss, radius, and finite-value flags.

Finite values are checked separately. Structural checks require full row rank, positive margin, and inverse/query residuals below 10^{-10} (float64) or 10^{-5} (float32) for $\lambda < 1$. The main recovery thresholds are separately fixed at 10^{-9} and 10^{-3} on $\lambda \in [0.25, 0.9]$. All 8,000 evaluated rows have $\kappa_2(G) < 352$, so the script's conditioning filter $\kappa_2(G) \le 10^4$ excludes none. The optional diagnostic

$$50n\epsilon\kappa_2(G)\max\{1, \|H\|_F/\|AB\|_F\}$$

is not a proved error bound and does not reclassify failures. At the canonical point $t = 16$, $\eta = 2^{-8}$, $\lambda = 0.25$, the maximum relative errors are 4.85×10^{-13} (float64) and 2.53×10^{-4} (float32).

B.3 Ablations

All seven variants use $t = 16$, $\eta = 2^{-8}$, and $\lambda = 0.5$, except the outside-chamber variant at $\lambda = 1.25$. Each crosses five sizes, ten seeds, and two precisions, giving 700 rows in total. The baseline keeps the original fixed (N_0, J_ρ) decoder. The column-space variant replaces only its left projection by $P_+ = 2^{-1/2}[E^\top, E^\top]^\top$; the network itself is unchanged.

The key perturbation is $K' = \overline{K} + 10^{-2}\Delta$, with a Gaussian Δ generated using seed $999983n + \text{seed}$, then scaled in float64 so that $\|\Delta\|_F = \|\overline{K}\|_F$ and cast to the working precision. The normalized variant divides every query and key row by its Euclidean norm before computing logits; its denominator is clamped at the smallest positive normal value of the dtype. This is embedding normalization, not averaging the loss. Both use the same unnormalized baseline decoder in their fixed rows. Their refit rows instead recompute the right inverse of the actual cotangent at that finite point. The full-row-rank observations in Table 1 concern these sampled points. The table pools both precisions, so its medians should not be read as precision-specific error estimates.

Variant	Decoder	Median error	Maximum error	Min. margin/ρ
Baseline	fixed	1.28×10^{-7}	6.95×10^{-5}	0.50
Column-space readout	fixed	1.55×10^{2}	4.40×10^{2}	0.50
Perturbed late key	fixed	4.05×10^{-2}	7.45×10^{-2}	0.50
Perturbed late key	refit	2.51×10^{-7}	5.81×10^{-5}	0.50
Normalized embeddings	fixed	9.76×10^{-1}	1.01	0.50
Normalized embeddings	refit	1.73×10^{-7}	3.26×10^{-5}	0.50
Outside chamber	fixed	5.69×10^{-8}	8.42×10^{-2}	-0.25

Table 1: Hard-slice ablations at $t = 16$, $\eta = 2^{-8}$, and $\lambda = 0.5$ ($\lambda = 1.25$ outside the chamber). Each row pools five sizes, ten seeds, and both precisions. All sampled cotangents have full row rank. Fixed rows test the baseline decoder; refits are pointwise diagnostics.

B.4 Exact checkpoint accounting

The accounting sweep crosses $n \in \{4, 8, 16, 32, 64\}$ and seeds $0, \ldots, 4$, giving 165 rows: 140 admissible routes and 25 deliberately key-aware audit routes. Inputs have small int64 entries, and all arithmetic stays within int64 range. An admissible route fixes a set Ω of q output coordinates before the key, computes and stores the complement, and recomputes Ω late. Every row records the scalar counts $M = n^2 - q$, nM, and nq and compares the reconstructed product with direct integer multiplication. The invalid audit routes compare two keys selecting different pre-cut masks, so matching their arithmetic counts does not make them admissible. The auxiliary CPU timing in the CSV is not used in a performance claim.

n	q	M	nM	nq	admissible route	inadmissible route
4	0	16	64	0	full storage	—
4	4	12	48	16	balanced, unbalanced	keys $0, 1$; 2×2
4	16	0	0	64	no storage	—
8	0	64	512	0	full storage	—
8	8	56	448	64	balanced, unbalanced	—
8	16	48	384	128	balanced, unbalanced	keys $0, 1$; 4×4
8	64	0	0	512	no storage	—
16	0	256	4096	0	full storage	—
16	16	240	3840	256	balanced, unbalanced	—
16	64	192	3072	1024	balanced, unbalanced	keys $0, 1$; 8×8
16	256	0	0	4096	no storage	—
32	0	1024	32768	0	full storage	—
32	32	992	31744	1024	balanced, unbalanced	—
32	256	768	24576	8192	balanced, unbalanced	keys $0, 1$; 16×16
32	1024	0	0	32768	no storage	—
64	0	4096	262144	0	full storage	—
64	64	4032	258048	4096	balanced, unbalanced	—
64	1024	3072	196608	65536	balanced, unbalanced	keys $0, 1$; 32×32
64	4096	0	0	262144	no storage	—

Table 2: Exact checkpoint and late-work accounting, one row per distinct (n, q). Admissible routes fix the q coordinate outputs Ω independently of the late key, and nM and nq are counts of scalar multiplications. Admissible routes store the $M = n^2 - q$ outputs outside Ω before the key and compute the q selected outputs after it: the stored work is nM, the late work is nq, and the two sum to n^3 in every row, with all integer residuals zero. Inadmissible routes are not performance baselines: keys 0 and 1 select different pre-cut checkpoints, violating commit-before-key. Balanced and unbalanced splits of one q share a single accounting, and an inadmissible route can reach the same accounting as an admissible one.

B.5 Constant-score late-key checks

The additional script checks (36) at the canonical t, η for five sizes and ten seeds in float64, with $B = 0$. It uses an explicit orthonormal contrast basis R of $\mathbf{1}^{\perp}$. Gaussian A and late V are generated using seed $1299709n +$ seed and rescaled to $\|A\|_F = \rho/4$ and $\|V\|_F = 10^{-2}$. It constructs $K(V)$, differentiates the actual InfoNCE loss with PyTorch, and checks constant scores, $GFR = V$, the projected gradient, and the isolated mixed product. The maximum score change is 4.26×10^{-14}, projected-gradient relative error 4.90×10^{-14}, and mixed-product relative error 8.19×10^{-14}. All 50 margins are positive.

SymPy additionally verifies five identities at each of $n = 2, 3$: zero-sum contrasts, the dual-basis identity, zero score change, the projected-gradient formula, and its bilinear component. These exact checks use integer contrast columns and their rational dual rather than an orthonormal basis. The full mathematical proof above covers arbitrary n and arbitrary invertible C_ρ; finite symbolic or numerical checks do not establish the rank lower bound.