WHEN MORE CAPACITY IS NOT ENOUGH: RETHINKING POTENTIAL-BASED ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Potential-based attention models organize computation through a scalar function, providing mathematical structure for analyzing model behavior. A natural goal is to use this structure to reproduce an existing attention model, preserving its outputs on the same inputs while making its behavior easier to analyze. Contrary to the intuition that scaling model capacity can overcome such limitations, we find that an unavoidable output error can persist even with arbitrarily complex potentials. Generating multiple outputs through the gradient of a single convex function means that their changes with the input cannot be chosen independently. The original attention response may not satisfy this constraint. For a class of fixed attention maps, we first prove a capacity-independent error lower bound that quantifies this structural restriction. We then give a criterion computable directly from the original model that determines the minimum number of independently corrected output directions needed to recover its outputs exactly. We also find that fewer corrections can sometimes approximate the original outputs arbitrarily well, but only as the relative scaling of output directions becomes increasingly uneven. This reveals the numerical cost of reducing the required structural change. Finally, numerical experiments with rigorous error guarantees establish that independent corrections can cross an accuracy limit that increasing potential capacity alone cannot overcome. These results provide a theoretical basis for deciding when the architecture must change instead of continuing to scale the model. We believe these insights can guide attention designs that combine mathematical structure with expressive power.

1 INTRODUCTION

Neural network design uses both scale and structure to improve learning. Larger models can fit richer computations (Kaplan et al., 2020), while architectural constraints encode prior knowledge about those computations (Bronstein et al., 2021). Energy-based attention gives this idea a concrete form: it organizes updates around a scalar objective, making their behavior amenable to mathematical analysis (Hoover et al., 2023; Hu et al., 2026; Santos et al., 2025).

Empirical results reveal both the promise and the cost of this structure. Energy Transformer reports strong results on graph anomaly detection and graph classification using attention updates designed to minimize an energy (Hoover et al., 2023). Yet in recurrent attention experiments on Sudoku, Tomihari & Karakida (2025) find that regularization toward energy-compatible weight constraints reduces accuracy relative to the unregularized model. A related tension appears in diffusion models: Cho et al. (2026) observe that stronger retrieval stability is associated with greater visual coherence but lower diversity, and use the skew-symmetric component of attention scores to control the tradeoff. These findings make the choice of structure consequential: it can support useful computations while restricting others that a task may need.

These observations raise a representation question: can we obtain the mathematical structure of a potential while retaining the behavior already learned by an attention model? Preserving that behavior requires the same inputs to produce the same outputs. Generating those outputs through the gradient of a single convex function imposes shared derivative constraints that increasing potential capacity cannot remove. For fixed attention maps with generic keys, we prove a capacity-independent error lower bound and determine the minimum number of nonlinear output directions that must be

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.

supplied independently of the potential. We also show that approximating the response can require fewer such directions than reproducing it exactly, but only through increasingly ill-conditioned output coordinates. Lower bounds and matching dimension constructions establish these results over the full readout classes, and numerical certificates demonstrate a strict separation between corrected and uncorrected readouts on a common target.

To isolate the representation constraint, we fix the keys, values, bias, and linear observation, and give the replacement the attention-weighted key mean. Under affine spanning of the keys, this mean determines the query, so it contains all the information needed to recover the response. Any obstruction therefore comes from the readout restriction. We allow a convex potential of arbitrary complexity, an affine branch, and a learned invertible change of output coordinates. An additional nonlinear branch supplies corrections in a selected output subspace. Its dimension counts the directions freed from the potential constraint.

Our answer depends on a quantity computed directly from the attention keys and values. Regress the observed values on an affine function of the keys. The remaining value components generate the part of the response that an affine readout cannot supply. For generic keys, the rank of this residual determines the minimum correction dimension. The residual therefore measures how many correction directions the target requires, independently of the potential network's width or depth.

A simple case reveals why the notion of approximation also matters. Suppose the residual has rank two and the observation has at least three independent outputs. With no correction, every admissible readout has positive error. One correction direction allows arbitrarily small error, while two allow exact representation. The one-direction saving has a cost: the learned output matrix must become increasingly ill-conditioned, stretching some directions relative to others. With any fixed bound on this imbalance, one direction again leaves positive error. Our conditioning theorem quantifies this cost and shows when approximation and exact representation require the same number of correction directions.

The experiments distinguish representation from finite implementation. The certified comparison tests whether a correction crosses the error floor of the entire uncorrected class. Geometric perturbations show that the estimated lower bound varies while the required correction dimension stays fixed. Finite polynomial and neural fits then examine attainable accuracy, and comparisons at a common error tolerance measure evaluation cost. A potential can save an output direction without reducing the number of operations. This distinction connects the representation theorems to the choices needed to fit and evaluate a readout.

2 Related Work

The closest modeling context is attention defined through potentials and energies. Hopfield–Fenchel–Young networks use a symmetric positive-definite hetero-associative readout (Santos et al., 2025). Energy Transformer and Hyper-SET develop energy-based objectives and updates (Hoover et al., 2023; Hu et al., 2026). These models motivate the study of structured attention, while our analysis concerns representation of a fixed attention response from its key mean. Input Convex Gradient Networks (Richter-Powell et al., 2021) study expressive parameterizations of convex gradients. We characterize the additional nonlinear output directions required to preserve a particular attention map, allowing the potential to have unrestricted complexity.

Jacobian analysis provides tools for studying attention dynamics (Tomihari & Karakida, 2025) and smoothness (Castin et al., 2024). Cho et al. (2026) relate symmetric and skew-symmetric score components to retrieval stability and diversity. In mean coordinates, gradient compatibility of the value response depends on the value bank and determines the representation constraint studied here. Additional degrees of freedom also play a central role in the expressive capacity of augmented Neural ODEs (Zhang et al., 2020) and low-rank model adaptation (Hu et al., 2021).

The representation bounds combine classical tools from conservative-field analysis (Im et al., 2016; Suda, 2020), volume-sampled least squares (Dereziński & Warmuth, 2017), and convex interpolation (Rockafellar, 1966; Taylor et al., 2017). These tools yield exact and approximate representation criteria for observed attention outputs, together with the minimum correction dimension and quantitative error bounds. The integral certificates make these bounds evaluable for a specified attention map and domain.

3 THEORY

3.1 ATTENTION IN MEAN COORDINATES

We fix an attention map and ask whether a structured readout can reproduce its outputs. The readout receives the attention-weighted key mean and combines a scalar potential gradient with an affine term and optional nonlinear corrections. We study how many output directions require corrections when the potential has unrestricted complexity. Let $K, V \in \mathbb{R}^{n\times d}$ have rows $k_i^\top, v_i^\top$. For inverse temperature $\beta > 0$, query x, and attention bias b_{att}, use natural coordinates $\theta = \beta x$ and $\eta = \beta b_{\mathrm{att}}$:

$$p(\theta) = \mathrm{softmax}(K\theta + \eta), \quad s(\theta) = K^\top p(\theta), \quad y(\theta) = V^\top p(\theta). \tag{1}$$

We wish to preserve the observed response $Oy(\theta)$, where $O \in \mathbb{R}^{m\times d}$ specifies which linear combinations of the value output must be reproduced; $O = I_d$ preserves the full output. A potential branch replaces it by $O\nabla\phi(s(\theta))$. We also permit an invertible output-coordinate change B, giving $OB\nabla\phi(s)$. We take ϕ to be C^1 and specify convexity for each function class. Keys and values have the same dimension because this decoder is a gradient on key-mean space.

If the keys affinely span $\mathbb{R}^d$, then

$$C(\theta) = D_\theta s = \mathrm{Cov}_p(k) \succ 0, \qquad D(\theta) = D_\theta y = \mathrm{Cov}_p(v, k). \tag{2}$$

The map s is an analytic diffeomorphism from $\mathbb{R}^d$ to $\mathrm{int\,conv}\{k_i\}$. The log-partition Hessian is positive definite, and for every interior mean the strictly convex function $\log\sum_i e^{k_i^\top\theta+\eta_i} - s^\top\theta$ is coercive and has a unique minimizer. Write $\vartheta(s)$ for the inverse query and $\widetilde{p}(s) = p(\vartheta(s))$. Thus the unrestricted decoder $y \circ \vartheta$ is always available.

The fixed-bias assumption is essential to this information-preservation property. For arbitrary probability vectors, the key mean can discard value information through $\ker[\mathbf{1}^\top; K^\top]$. At fixed bias, the attainable probabilities form an exponential family on which the key mean is injective.

Thus any representation error comes from the form of the readout, since the key mean retains the information needed to recover the output.

3.2 OBSERVATION AND VALUE RESIDUAL

We next identify which part of the target actually requires a nonlinear readout. If the observed values are affine functions of the keys, their attention-weighted average is the same affine function of the key mean. An affine branch already reproduces this part. The residual from fitting values to keys therefore isolates the part whose representation cost we need to study. We use orthonormal coordinates for the outputs retained by O, preserving Euclidean error in the original observation space. Let O have rank $2 \le r \le d$, and let C_o have orthonormal columns spanning $\mathrm{range}(O)$. Set

$$a(s) = C_o^\top Oy(\vartheta(s)), \quad W = VO^\top C_o, \quad X = [\mathbf{1}, K], \quad R = (I - P_X)W, \quad q = \mathrm{rank}\, R.$$

Here P_X is the orthogonal projector onto the column space of X. Thus R is the residual from affine least-squares regression of the observed values W on the keys, with one row per key. Writing the fit as $W = \mathbf{1}b_0^\top + KL_0^\top + R$ and using $K^\top\widetilde{p}(s) = s$ and $\mathbf{1}^\top\widetilde{p}(s) = 1$ gives

$$a(s) = L_0 s + b_0 + R^\top\widetilde{p}(s).$$

The affine branch reproduces the first two terms exactly. The remaining response is $R^\top\widetilde{p}(s)$, so R identifies the value components not supplied by the affine fit. Its rank will determine the required correction dimension. All infimal risks below use the same uniform probability measure μ on a closed mean ball $\mathcal{D} \Subset \mathrm{int\,conv}\, K$ with nonempty interior. We impose two genericity conditions on the keys: every $d + 1$ keys are affinely independent and distinct $(d + 1)$-element subsets have distinct sums. This is an open full-measure condition. Bias is finite and fixed. Theorems concern unrestricted smooth correction functions and C^1 potentials, with no fixed network-size or curvature bound. Smoothness on the closed ball means smoothness on an open neighborhood of that ball.

3.3 MINIMUM CORRECTION DIMENSION

The remaining question is how much of this nonlinear response a potential can supply. We let an unrestricted correction handle a k-dimensional output subspace and ask for the smallest sufficient k.

Larger k gives the readout more freedom outside the potential branch. The resulting model is

$$\widehat{a}(s) = Ls + b + T\nabla\phi(s) + Uz(s), \qquad U^\top U = I_k. \tag{3}$$

Here $T \in \mathbb{R}^{r\times d}$ has full row rank, ϕ is globally convex and C^1, and z is an unrestricted smooth k-output function of the same key mean. The unrestricted affine term reproduces responses affine in s. The potential branch generates multiple outputs from one scalar function, while the columns of U select k orthonormal output directions and $z(s)$ supplies corrections along them. Thus k counts correction directions, not model parameters.

The full-row-rank condition on T corresponds to observing the potential branch through O after an invertible output-coordinate change B, so that $T = C_o^\top OB$. Every full-row-rank $r \times d$ matrix has this form. It preserves the rank of the observation of that branch. This assumption matters even when the total readout has full rank: if T may be rank deficient, a rank-one nonlinear residual admits exact representation without a correction. Appendix K gives the construction.

For $0 \le k \le r$, define

$$\mathcal{E}_k^\star = \inf_{L,b,T,U,z,\phi} \int_{\mathcal{D}} \|Ls + b + T\nabla\phi(s) + Uz(s) - a(s)\|^2 \, d\mu(s). \tag{4}$$

Here $L, T \in \mathbb{R}^{r\times d}$, $b \in \mathbb{R}^r$, $U \in \mathbb{R}^{r\times k}$, $\operatorname{rank} T = r$, $U^\top U = I_k$, $z \in C^\infty(\mathbb{R}^d;\mathbb{R}^k)$, and $\phi \in C^1(\mathbb{R}^d)$ is convex. Arbitrarily small mean squared error may be achieved by a sequence of readouts, whereas exact representation requires one admissible readout to agree with the target throughout $\mathcal{D}$.

The theorem below relates the required dimension to $q = \operatorname{rank} R$, with a lower bound over the full readout class and an explicit upper construction.

Theorem 1. *For the generic keys, finite bias, rank $r \ge 2$ and mean ball defined above, for $0 \le k \le r-2$ there is $c_k = c(K,\eta,\mathcal{D},r-k) > 0$ such that*

$$c_k \sum_{j=k+2}^{r} \sigma_j(R)^2 \le \mathcal{E}_k^\star \le \sum_{j=k+2}^{r} \sigma_j(R)^2. \tag{5}$$

For $0 \le k \le r-2$, arbitrarily small mean squared error is possible if and only if $q \le k+1$, whereas exact representation is possible if and only if $q \le k$. For $k = r-1$, every smooth target on $\mathcal{D}$ admits exact representation. Consequently, the minimum dimensions for arbitrarily small error and exact representation are respectively

$$k_{\text{closure}} = \max\{q-1, 0\}, \qquad k_{\text{exact}} = \min\{q, r-1\}.$$

The affine branch leaves R unchanged, so a residual of rank at least two gives positive error when $k = 0$, regardless of potential capacity. It also allows convexification: on a compact domain, adding $M\|s\|^2/2$ makes any smooth potential ψ convex for sufficiently large M, and the affine branch cancels MTs. A global convex extension completes the construction. Thus, with unrestricted curvature, the affine branch removes the additional convexity condition that arises at residual rank one without it (Theorem 5).

The compatibility condition can be seen directly from the potential branch $g = T\nabla\phi$: its Jacobian satisfies $Dg\,T^\top = T(Dg)^\top$, since the Hessian of ϕ is symmetric. For C^1 potentials this identity holds distributionally. Integral tests with zero mean eliminate the constant Jacobian of the affine branch. For generic keys, the resulting constraints restrict the observed value residual and yield the error bound in Theorem 4.

An incompatible Jacobian alone rules out exact equality, but output error could still vanish without derivatives converging. Integration by parts transfers derivatives onto fixed smooth test functions, making each test continuous in the output L^2 norm. A nonzero test response therefore yields an error lower bound that survives limits of increasingly complex potentials and varying output matrices.

To obtain the dimension bound, let P have orthonormal columns spanning $U^\perp$ and project onto this subspace. The correction disappears, $P^\top T$ remains full row rank, and the residual becomes RP. Theorem 4 gives the lower bound uniformly over this subspace. Singular-value interlacing then leaves the tail after $k+1$ directions. For the upper bound, the correction carries k singular

components exactly, the potential branch approximates one more, and the remaining components are discarded. Appendix G proves necessity, global sufficiency, and the case $k = r - 1$. The residual spectrum therefore determines how many output directions require an unrestricted nonlinear correction.

The case $k = r - 1$ has a simple constructive interpretation. After choosing output coordinates, only one scalar response remains outside the correction subspace. Integrating that response along its matching input coordinate produces a potential with the desired partial derivative. Its other partial derivatives need not match the target, because the remaining $r-1$ output directions can be corrected independently. Convexification adds an affine response that the affine branch cancels. This construction explains why the full-residual case $q = r$ permits exact representation with one fewer correction direction, even under well-conditioned output coordinates. The exceptional case is a consequence of leaving only one derivative to prescribe.

Returning to the rank-two example from the introduction, the lower bound at $k = 0$ is positive because $\sigma_2(R) > 0$. At $k = 1$ the spectral tail vanishes, although exact representation still requires $k = 2$. The next result explains what happens to the output matrix along this one-direction approximation.

3.4 Effect of output conditioning

The limiting approximation in Theorem 1 can use output matrices with increasingly unequal singular values. We now ask how much error remains when this imbalance is bounded. Define $\mathcal{E}^\star_{k,\chi}$ by restricting the infimum in (4) to $\mathrm{cond}_2(T) = \sigma_1(T)/\sigma_r(T) \le \chi$, $\chi \ge 1$. The ratio controls the relative scaling of the largest and smallest output singular values.

Theorem 2. *For $0 \le k \le r-2$ under the assumptions of Theorem 1, a constant $\widetilde{c}_k > 0$, independent of values and χ, satisfies*

$$\mathcal{E}^\star_{k,\chi} \ge \widetilde{c}_k \left[\sum_{j=k+2}^{r} \sigma_j(R)^2 + \chi^{-2}\sigma_{k+1}(R)^2 \right]. \tag{6}$$

For every fixed finite $\chi \ge 1$, the minimum dimension for zero infimal risk is $k_{\mathrm{closure},\chi} = k_{\mathrm{exact}} = \min\{\mathrm{rank}\, R, r-1\}$. The sufficient exact constructions have $\mathrm{cond}_2(T) = 1$.

For $1 \le q < r$, unbounded conditioning reduces the dimension for arbitrarily small error from q to $q-1$, and fixed finite conditioning removes that saving. At $q = r$, exact representation with $r-1$ correction directions remains possible at condition number one. For the rank-two example with one correction direction, (6) leaves a positive term proportional to $\chi^{-2}\sigma_2(R)^2$. The construction used in Section 4 has risk C/χ^2, so the dependence on conditioning is sharp for that target: $\mathcal{E}^\star_{1,\chi} = \Theta(\chi^{-2})$.

The condition number measures relative scaling. Multiplying T by a positive scalar and dividing the potential by that scalar preserves the response and convexity, but cannot remove the imbalance between output directions. Bounding this imbalance restores the exact-representation threshold, quantifying the numerical cost of using fewer corrections.

The bound follows from retaining the smallest singular value of the output matrix in the proof of Theorem 1. After projecting away k correction directions, set $m = r - k$ and $A = P^\top T$. Then $\mathrm{cond}_2(A) \le \chi$ and $\sigma_m(A)^2/\|A\|_F^2 \ge 1/(m\chi^2)$. The coefficient estimate bounds both the residual spectral tail and its total energy weighted by this ratio. Averaging these two lower bounds, then applying singular-value interlacing to RP, gives (6) with a new constant. Full details are in Appendix H.

3.5 Ordinary nonlinear readouts

What does the potential contribute beyond an affine map and an unrestricted nonlinear correction? To answer this, remove the potential branch and consider $\widehat{a}_{\mathrm{plain}}(s) = Ls + b + Uz(s)$, $U^\top U = I_k$. Let a^μ_{aff} be the $L^2(\mu)$ affine projection of a, and set $e_\mu = a - a^\mu_{\mathrm{aff}}$ and $\Sigma_\mu = \mathbb{E}_\mu[e_\mu e_\mu^\top]$. Standard output-space projection gives the exact baseline

$$\inf_{L,b,U,z} \mathbb{E}_\mu \|Ls + b + Uz(s) - a(s)\|^2 = \sum_{j=k+1}^{r} \lambda_j(\Sigma_\mu). \tag{7}$$

This is the output projection principle underlying reduced-rank regression and its nonparametric extensions (Izenman, 1975; Foygel et al., 2013), here applied with unrestricted residual functions. It holds for square-integrable targets. In our smooth compact-domain setting the optimizing residual functions are smooth. For a full-support continuous mean law, $\operatorname{rank} \Sigma_\mu = \operatorname{rank} R = q$. The proof and the distinction from a finite empirical law are in Appendix J.

Table 1: Minimum nonlinear correction dimension k for arbitrarily small mean squared error and exact representation. Here $q = \operatorname{rank} R$ and r is the observation rank. Potential complexity and correction functions are unrestricted.

Readout class	Arbitrarily small error	Exact representation
Affine plus free nonlinear directions	q	q
Potential, affine, and correction	$\max\{q-1, 0\}$	$\min\{q, r-1\}$
Same, any fixed finite χ	$\min\{q, r-1\}$	$\min\{q, r-1\}$

Table 1 shows that adding a potential saves at most one free nonlinear output direction. The conditions for this saving depend on whether the residual occupies the entire observation space. For $1 \le q < r$, exact representation saves no correction directions. Approximation requires one fewer correction direction than exact representation, but this saving requires unbounded conditioning. For $q = r$, the case $k = r - 1$ saves one even with $\operatorname{cond}(T) = 1$. For fixed full-rank X and a generic value bank, $q = \min\{r, n-d-1\}$, so sufficiently many memories typically require $r-1$ correction directions. This genericity statement refers to value banks outside a proper algebraic subset.

The rank of R determines whether zero risk is possible, while Σ_μ measures the residual variation under the input distribution. For a nonzero tolerance δ, a certified lower bound above δ excludes a function class and a feasible upper bound below δ establishes sufficiency. The experiments examine these finite-error questions alongside the computational cost of the two readout families.

3.6 Implications for readout design

The criterion concerns the outputs that the model must preserve. Because $W = VO^\top C_o$, it incorporates the downstream linear observation before measuring the residual. A direction in the internal value response that this observation removes contributes nothing to R. Applying the criterion to the complete value bank can therefore overstate the correction dimension needed for a particular observed response. Conversely, preserving more output directions can expose additional residual components. The appropriate structural comparison starts with the required observation and then removes the affine part that a simple branch can already reproduce.

For exact representation, the remaining decision is algebraic: compute the residual rank and compare the available correction dimension with Theorem 1. For finite accuracy, the residual spectrum must be interpreted together with the domain and the condition-number budget. A small residual component can be negligible at one tolerance and decisive at another. The lower bounds identify settings in which no increase in potential capacity can meet the requested accuracy; an explicit construction or fitted readout supplies the complementary upper bound. Increasing capacity becomes relevant within the structural choices that these comparisons leave feasible.

Finally, correction dimension counts independent nonlinear output directions, not parameters or arithmetic operations. A single correction function may itself require a rich parameterization, and evaluating the gradient of a potential has a computational cost. Thus the dimension advantage in Table 1 provides a representation principle, while choosing an efficient finite model also requires a cost comparison at the desired accuracy. The experiments below examine both questions using common targets and evaluation distributions.

4 Experiments

We evaluate the representation bounds, their dependence on geometry, and their implications for finite models. The rank-two example supplies a common target for certification, geometric perturbation, and fitting. Comparisons within each experiment share the keys, values, bias, observation, and evaluation distribution.

4.1 CERTIFIED APPROXIMATION BOUNDS

One correction direction achieves smaller error than every uncorrected readout on the seven-key target in Appendix I. The target has three observed outputs and residual rank two. Full-domain compatibility tests give a lower bound for every uncorrected readout, while an explicit corrected readout gives an upper bound. Validated integration accounts for numerical uncertainty in both bounds:

$$E_{\mathrm{one}}(1000) \le 4.45248 \times 10^{-9} < 8.99418 \times 10^{-9} \le \mathcal{E}_0^\star. \tag{8}$$

Figure 1 shows the certified gap and its dependence on output conditioning. The lower bound allows an arbitrary affine branch, full-row-rank output matrix, and unrestricted potential. The separation therefore applies beyond any particular network architecture or fitting procedure.

The bounds in (8) use the same target and mean distribution. Their strict ordering proves that an explicit corrected readout outperforms every uncorrected readout, independently of optimization. The quadrature curve illustrates the construction's error; the certified bounds establish the separation.

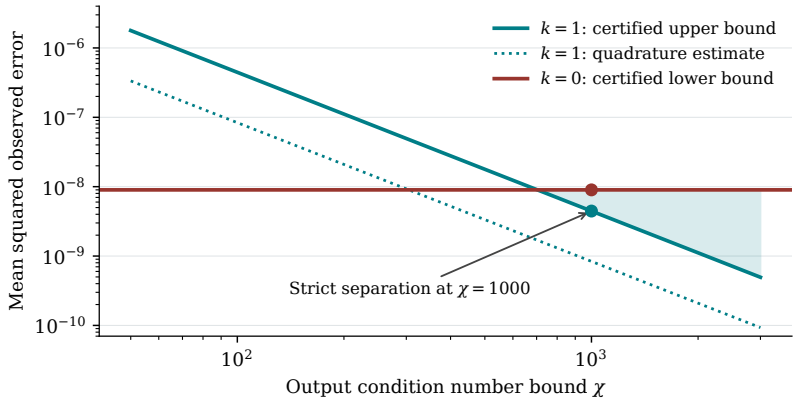


Figure 1: Certified approximation bounds for the same attention target and uniform mean distribution. The horizontal line bounds the error of every readout with $k = 0$ from below. The solid curve bounds an explicit $k = 1$ construction from above, and the dotted curve estimates its risk by numerical quadrature. In the shaded region, the corrected readout has smaller error than every uncorrected readout. Theorem 2 gives the matching order $\mathcal{E}_{1,\chi}^\star = \Theta(\chi^{-2})$ for this target.

A four-output target with residual rank two illustrates the dimension bound while retaining two outputs outside the correction subspace. Its residual is represented exactly when $k = 2$, using an output matrix with condition number one. At $k = 1$, the error is positive for every fixed condition-number bound and an explicit construction converges at the inverse-square rate. Appendix O gives the keys, value matrix, and exact construction.

4.2 DEPENDENCE ON KEY GEOMETRY

Residual rank determines whether the error can vanish. It does not specify how large a positive error must be. To isolate the role of geometry, we keep the residual spectrum fixed while deforming the keys. Starting from the three-dimensional target, we compress the third key coordinate by t and use an interior mean ball whose radius is scaled by the same factor. The value bank remains fixed. For every $t > 0$, this invertible change preserves the genericity assumptions and the residual singular values. The keys approach an affinely degenerate configuration as $t \to 0$.

The integral tests define a linear map from T to the target's violations of the gradient compatibility conditions. We represent this map by A; its smallest singular value determines the strength of the lower bound in (20). Figure 2 shows that this matrix becomes less sensitive to the nonaffine residual even though the singular values of R remain constant. Normalizing by affine approximation error preserves this effect. The experiment separates the algebraic requirement for a nonlinear correction from the magnitude of the approximation error on a specified domain.

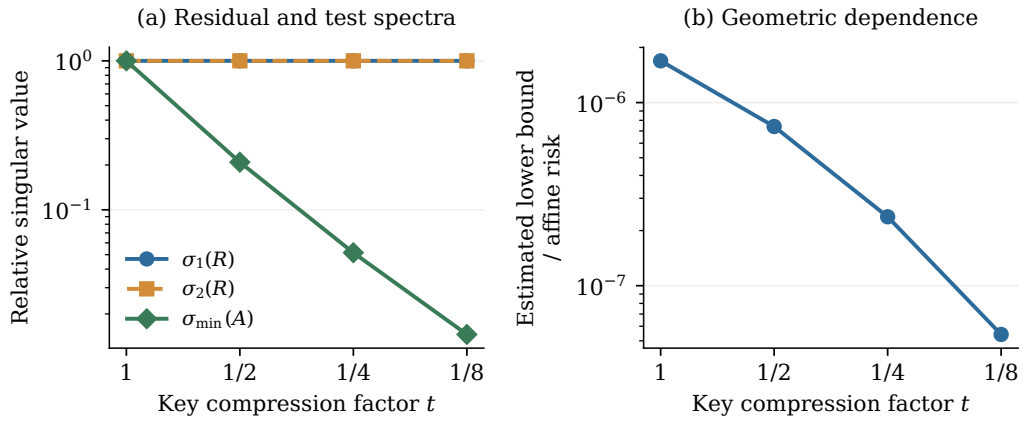


Figure 2: Sensitivity to key and domain geometry. Left: the residual singular values remain constant, while the smallest singular value of the compatibility-test matrix A decreases. All three singular values are normalized by their values at $t = 1$. This matrix maps the entries of T to the stacked antisymmetric Jacobian moments after whitening the tests (Appendix G). Right: the estimated lower bound, normalized by affine approximation risk, decreases along the same path. The mean ball contracts with the key transformation. These are order-18 quadrature estimates, with the evaluation procedure in Appendix L.

4.3 Finite-model approximation

The representation theorem allows unrestricted functions. We next examine the accuracy attainable with finite polynomial and tanh features on the same three-dimensional target. We choose correction directions from the residual SVD and fit the linear coefficients of the affine, potential, and correction branches jointly by linear least squares using a truncated SVD. The output matrix and hidden features are fixed during the solve. Both polynomial and tanh models receive only the key mean at evaluation time.

Table 2: Finite representations of the three-dimensional target with $k = 2$ and $\chi = 1000$. Tanh ranges cover all three seeds. All models use the same training means and independent order-18 quadrature. The full comparison over k and χ is in Appendix O.

Parameterization	Risk / range across seeds
Degree-seven polynomial	2.79×10^{-10}
Tanh, width 32	$(1.90\text{–}5.25) \times 10^{-5}$
Tanh, width 128	$(3.20\text{–}6.07) \times 10^{-8}$

Table 2 compares finite representations in a regime where the correction dimension permits exact representation. Increasing the tanh width reduces the error, while the polynomial representation achieves a closer approximation. This difference reflects the choice of finite features within the admissible function class. Appendix O describes the fitting and evaluation procedure.

4.4 Accuracy and computational cost

A potential can reduce the number of correction directions, but its gradient also has to be evaluated. We therefore compare the cost of ordinary and potential-based readouts that meet the same validation error tolerance under a common storage limit. The targets comprise the two synthetic examples above, three additional value banks, and a frozen Iris regressor. The synthetic targets use uniform mean distributions. The regressor is evaluated on a continuous query neighborhood. Calibration determines the directions and fitted coefficients, validation selects the least-cost feasible configuration, and independent samples measure test error. Appendix M specifies the candidate families and selection rule.

Table 3 shows that fewer correction directions do not necessarily reduce evaluation cost. Under the validation selection rule, the ordinary readout uses fewer multiplications on every target.

Table 3: For each model family, we select the readout with the fewest multiplications among those meeting the validation error tolerance and storage limit. Ordinary denotes $Ls+b+Uz$ and potential denotes $Ls+b+T\nabla\phi+Uz$. The storage cap is 1,000 scalars and δ is one percent of calibration affine error. Multiplication counts use algebraically simplified inference graphs and exclude activations and elementwise products. Values below one in the last two columns meet the independent test tolerance.

	Dimension k		Multiplications		Test risk / δ	
Target	Ordinary	Potential	Ordinary	Potential	Ordinary	Potential
Primary 3D	2	1	178	239	0.86	0.94
Four-output 4D	1	0	504	788	0.23	0.08
New values 1	2	2	178	274	0.76	0.53
New values 2	2	1	178	239	0.87	0.96
New values 3	3	2	204	274	1.01	0.95
Frozen Iris	2	1	134	168	0.10	0.07

Inference uses algebraically simplified expressions for both families. The convexifying quadratic cancels with the affine compensation, potential output coefficients are precomputed, and correction matrices are multiplied in the cheaper order. Table 3 reports the resulting multiplication counts and test errors. Appendix N describes the simplification and cost accounting.

The selected ordinary readouts meet the independent test tolerance except on New values 3, where the potential-based readout has smaller error and meets the tolerance. Table 3 reports both the cost and test error so that this accuracy difference remains visible.

The primary comparison uses calibration residual PCA for ordinary readouts and value-residual SVD for potential-based readouts. Giving both families the same PCA or SVD basis preserves the cost ordering on all targets (Table 5). The direction rule does affect accuracy, including which selected configurations meet the test tolerance. This control supports the cost comparison while identifying direction selection as a source of finite-model error.

These comparisons address finite accuracy rather than the zero-error threshold. In particular, the tolerance for the primary target exceeds its certified lower bound. Validation therefore determines which finite models are adequate here. Figure 1 answers the complementary question of whether a correction can cross a lower bound that holds for the full uncorrected function class.

5 Conclusion

Potential-based attention can face an output error that increasing potential capacity cannot remove. For the fixed attention maps studied here, we characterize the minimum number of independent correction directions needed to preserve the response using a criterion computed from the keys and observed values. Fewer directions can sometimes suffice for arbitrarily accurate approximation, at the cost of unbounded output conditioning. Certified numerical experiments verify that independent corrections can overcome the structural error bound. These results identify when preserving a learned attention response requires changing its readout structure and provide a basis for balancing mathematical structure, expressive power, and numerical conditioning.

The representation criterion can be evaluated before choosing a potential network or fitting algorithm. It separates the output freedom required by the target from the capacity used to approximate an admissible response. The finite-model comparisons further show that saving a correction direction need not reduce evaluation cost. Together, these findings distinguish structural feasibility, attainable accuracy, and computational cost when designing a readout. Extending this analysis to learned key geometry and task-dependent error criteria is a natural direction for future work.

AI USE STATEMENT

AI tools assisted literature searches, proof drafting and revision, experimental design, numerical implementation and code editing. The authors are responsible for the correctness, originality, and presentation of the submitted work.

REPRODUCIBILITY STATEMENT

The appendices provide proofs of the representation bounds, explicit synthetic targets, and the numerical integration procedures used to certify the error bounds. They also describe the data distributions, model families, fitting procedures, selection rules, and evaluation protocols for the experiments, including the treatment of numerical integration error and comparisons across random seeds.

REFERENCES

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv:2104.13478, 2021. URL https://arxiv.org/abs/2104.13478.

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5817–5840, 2024.

Hyunmin Cho, Woo Kyoung Han, and Kyong Hwan Jin. Balancing fidelity and diversity in diffusion models via symmetric attention decomposition: Hopfield perspective. In *International Conference on Machine Learning*, 2026.

Michał Dereziński and Manfred K. Warmuth. Unbiased estimates for linear regression via volume sampling. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Rina Foygel, Michael Horrell, Mathias Drton, and John Lafferty. Nonparametric reduced rank regression. arXiv:1301.1919, 2013. URL https://arxiv.org/abs/1301.1919.

Benjamin Hoover, Yuchen Liang, Bao Pham, Rameswar Panda, Hendrik Strobelt, Duen Horng Chau, Mohammed J. Zaki, and Dmitry Krotov. Energy transformer. In *Advances in Neural Information Processing Systems*, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. arXiv:2106.09685, 2021. URL https://arxiv.org/abs/2106.09685.

Yunzhe Hu, Difan Zou, and Dong Xu. Hyper-SET: Designing transformers via hyperspherical energy minimization. In *International Conference on Learning Representations*, 2026.

Daniel Jiwoong Im, Mohamed Ishmael Belghazi, and Roland Memisevic. Conservativeness of untied auto-encoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, pp. 1694–1700, 2016. doi: 10.1609/aaai.v30i1.10268.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975. doi: 10.1016/0047-259X(75)90042-1.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv:2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

Jack Richter-Powell, Jonathan Lorraine, and Brandon Amos. Input convex gradient networks. arXiv:2111.12187, 2021. URL https://arxiv.org/abs/2111.12187.

R. Tyrrell Rockafellar. Characterization of the subdifferentials of convex functions. *Pacific Journal of Mathematics*, 17(3), 1966.

Saul Santos, Vlad Niculae, Daniel McNamee, and André F. T. Martins. Hopfield-Fenchel-Young networks: A unified framework for associative memory retrieval. *Journal of Machine Learning Research*, 26(265):1–51, 2025. URL https://jmlr.org/papers/v26/24-1961.html.

Tomoharu Suda. Application of Helmholtz–Hodge decomposition to the study of certain vector fields. *Journal of Physics A: Mathematical and Theoretical*, 2020. doi: 10.1088/1751-8121/aba657. URL https://arxiv.org/abs/1911.10382.

Adrien B. Taylor, Julien M. Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161: 307–345, 2017. doi: 10.1007/s10107-016-1009-3. URL https://arxiv.org/abs/1502.05666.

Akiyoshi Tomihari and Ryo Karakida. Recurrent self-attention dynamics: An energy-agnostic perspective from jacobians. In *Advances in Neural Information Processing Systems*, 2025.

Han Zhang, Xi Gao, Jacob Unterman, and Tom Arodz. Approximation capabilities of neural odes and invertible residual networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 11086–11095, 2020. URL https://proceedings.mlr.press/v119/zhang20h.html.

A Exact representation under linear observation

A.1 Exact representation of observed outputs

Let $O \in \mathbb{R}^{m\times d}$ have rank $r \geq 2$. Assume that every $d+1$ keys are affinely independent and that distinct $(d+1)$-element subsets have distinct sums. These *generic* conditions hold on an open full-measure set.

Theorem 3. *For generic keys and any finite bias, there exists a global convex C^1 potential satisfying $O\nabla\phi(s(\theta)) = Oy(\theta)$ on a nonempty open query set if and only if*

$$Ov_i = O(Ak_i + c) \quad \textit{for all } i, \qquad A = A^\top \succeq 0. \tag{9}$$

Without convexity, drop $A \succeq 0$. If an invertible output matrix B may also be learned, exact representation by $OB\nabla\phi$ is equivalent to

$$R_{\mathrm{obs}} := (I - P_{[\mathbf{1},K]})VO^\top = 0. \tag{10}$$

Values may differ arbitrarily in $\ker O$. The restriction acts on the components used by the head. For observation rank at least two, symmetry of the partial Jacobian imposes the affine condition. A scalar observation has no such symmetry constraint and can preserve nonlinear responses. This case includes binary centered logits.

For an orthonormal frame U spanning $\mathrm{range}(O^\top)$, write $a(s) = U^\top y(\theta(s))$. Partial gradients require $D_s a\, U$ to be symmetric. Weighted least squares decomposes this Jacobian into affine interpolants on the key simplices. Distinct subset sums force each interpolant to satisfy the same partial symmetry. Replacing one vertex of an affine basis then forces every observable residual to vanish when $r \geq 2$. Convexity is exactly a positive-semidefinite completion condition. Appendix B proves these steps, including the C^1 case, and attributes the least-squares identity.

B Proof of the exact representation criterion

The key step is to turn a condition on derivatives into a condition on the finite value bank. A weighted least-squares identity from volume sampling (Dereziński & Warmuth, 2017) writes the attention Jacobian as a sum of affine interpolants. Symmetry then constrains each interpolant, which yields the classification below.

B.1 A WEIGHTED SIMPLEX IDENTITY

Let $X = [\mathbf{1}, K]$, $P = \mathrm{diag}(p)$, and let $W \in \mathbb{R}^{n\times r}$ contain arbitrary observable values. Write $a(s) = W^\top p(\theta(s))$. For I of size $d+1$, define c_I, A_I by $w_i = A_I k_i + c_I$ for $i \in I$. Then

$$\det C\, D_s a = \sum_{|I|=d+1} p_I \det(X_I)^2 A_I, \qquad p_I = \prod_{i\in I} p_i. \tag{11}$$

Indeed, the weighted regression coefficients are $T = (X^\top PX)^{-1}X^\top PW$. Its slope block transposed is $\mathrm{Cov}_p(w,k)C^{-1} = D_s a$, by eliminating the intercept in the normal equations. Also $\det(X^\top PX) = \det C$. To check the numerator identity entrywise, use Cramer's rule for one column of W and one coefficient: replace the corresponding column of $X^\top PX$ by $X^\top Pw$. Cauchy–Binet gives a sum of $p_I \det(X_I)\det(X_I^{(j\leftarrow w)})$. Each determinant ratio is the jth coefficient of $X_I^{-1}w_I$. This proves the identity, without invoking a probabilistic assumption on the keys.

B.2 CHARACTERIZATION OF COMPATIBLE PARTIAL GRADIENTS

Fix $U \in \mathbb{R}^{d\times r}$ with $U^\top U = I_r$ and suppose $D_s a\, U$ is symmetric on a nonempty open mean set. Multiplying (11) by the softmax normalizer to power $d+1$ and using independence of distinct exponential frequencies yields $A_I U = (A_I U)^\top$ for every I. Analyticity transfers the open-set identity to all queries. Finite biases only multiply coefficients by strictly positive scalars.

Fix an affine basis with barycentric coordinates $\lambda_j(s)$ and its interpolant $A_0 s + c_0$. For any extra key k_* let $e = w_* - A_0 k_* - c_0$. The genericity assumptions implies $\lambda_j(k_*) \neq 0$ for every basis vertex. Replacing vertex j gives slope $A_0 + e\nabla\lambda_j^\top/\lambda_j(k_*)$. Thus $e(U^\top\nabla\lambda_j)^\top$ is symmetric for every j. If $e \neq 0$, symmetry of this rank-one matrix forces $U^\top\nabla\lambda_j \in \mathrm{span}(e)$. But the barycentric gradients span $\mathbb{R}^d$, and their projections span $\mathbb{R}^r$, contradicting $r \geq 2$. Hence $e = 0$ for each extra key. The exact kernel is therefore

$$\mathcal{L}_U = \{W : w_i = Ak_i + c,\ AU = (AU)^\top\}. \tag{12}$$

Conversely this condition gives $a(s) = As + c$ with symmetric partial Jacobian. Its dimension is $r(d+1) - r(r-1)/2$: evaluation on spanning keys is injective and $A \mapsto \mathrm{skew}(AU)$ is onto.

B.3 CONVEX COMPLETION FOR C^1 POTENTIALS

For the original observation matrix take U spanning its row space and complete it to an orthogonal matrix $[U, Z]$. Preservation is equivalent to prescribing $U^\top\nabla\phi = Hu + Tz + c$, where $s = Uu + Zz$ and $H = H^\top$. If a convex C^1 potential exists on the open set, mollify it inside a ball compactly contained in that set using a symmetric smooth kernel. Its prescribed first partial derivatives remain the same affine functions. The smooth convex Hessian consequently has block form $\left[\begin{smallmatrix} H & T \\ T^\top & D\end{smallmatrix}\right] \succeq 0$ at an interior point. Thus $H \succeq 0$ and $\mathrm{range}(T) \subseteq \mathrm{range}(H)$: test the quadratic form on (v, tz) for $v \in \ker H$ and both signs of t. Conversely choose $D = T^\top H^\dagger T$. This block matrix is PSD and defines a global quadratic potential with the prescribed partial gradient. Without convexity, any symmetric completion works. This proves (9), including necessity for local agreement and sufficiency with a global potential.

For a free B, apply this result to observation OB and values $VB^{-\top}$. It implies that $VO^\top$ is affine in the keys. Conversely let $Oy(s) = Ts + c_o$. Both T and c_o take values in $\mathrm{range}(O)$. Set $A = O^\dagger T$ and $c = O^\dagger c_o$. then $O(As + c) = Oy$. A square polar factorization $A = BP$ exists with B orthogonal and $P \succeq 0$, including singular A. The global convex quadratic with gradient $Ps + B^\top c$ gives the desired equality. This proves (10).

Fix any spanning generic bank with $n > d+1$ and a nonzero $w \perp \mathrm{col}\, X$. Observe only $e_1^\top$, and set $V = K + \delta w e_1^\top$. On a compact interior mean ball let $f(s) = w^\top p(\theta(s))$. Extend it smoothly with compact support and form $\Psi(s) = \int_0^{s_1} \widetilde{f}(t, s_2, \ldots, s_d)dt$. Its Hessian is globally bounded: outside the support interval in s_1 it is constant in s_1, with bounded derivatives in the other coordinates. For sufficiently small nonzero δ, $\phi(s) = \|s\|^2/2 + \delta\Psi(s)$ is globally strongly convex and $\partial_1\phi = s_1 + \delta f(s)$ on the ball. The observed value residual is $\delta w \neq 0$, giving a rank-one exception to the affine classification with a global strongly convex potential.

Table 4: All entries concern the same continuous mean law and unrestricted invertible output layers under the genericity assumptions. Residual rank is $\operatorname{rank} R$. No affine or nonlinear correction branch is included.

Residual rank	General gradient + output layer	Convex gradient + output layer
0	Exact representation	Exact representation
1	Zero infimum, unattained	Zero iff (14); otherwise positive
≥ 2	Positive floor	Positive floor

C Exact and approximate representation

We first study readouts $g(s) = T\nabla\phi(s)$, without affine or nonlinear correction branches. These results are used below to analyze the larger class in (3). Use a closed mean ball $\mathcal{D} \Subset \operatorname{int}\operatorname{conv} K$ with nonempty interior and its uniform law. Queries are obtained by the inverse mean map. Let C_o have orthonormal columns spanning $\operatorname{range}(O)$, $a(s) = C_o^\top O y(\vartheta(s))$, $W = VO^\top C_o$, $X = [\mathbf{1}, K]$, and $R = (I - P_X)W$. Its nonzero singular values equal those of R_{obs}. Write $\mathcal{E}^\star_{\mathrm{grad}}$ and $\mathcal{E}^\star_{\mathrm{cvx}}$ for the infimal observed mean-square risks on this same law, over all invertible output layers and global C^1 potentials, respectively without and with convexity. The observation rank r stays fixed.

Theorem 4. *For generic keys, finite bias and $r \geq 2$, set $\tau^2 = \sum_{j=2}^{r} \sigma_j(R)^2$. There is a constant $c = c(K, \eta, \mathcal{D}, r) > 0$, independent of V and O apart from r, such that*

$$c\tau^2 \leq \mathcal{E}^\star_{\mathrm{grad}} \leq \tau^2, \qquad \mathcal{E}^\star_{\mathrm{cvx}} \geq c\tau^2. \tag{13}$$

In particular, $\mathcal{E}^\star_{\mathrm{grad}} = 0$ if and only if $\operatorname{rank} R \leq 1$. When $\operatorname{rank} R = 1$, this infimum is not attained.

Thus two independent nonaffine residual directions suffice for a positive approximation error, even if the observation rank is larger. For generic values the integrability obstruction can already occur at $n \geq d + 3$, independently of $r \geq 2$.

Every observed readout can be written $g = T\nabla\phi$ for an arbitrary full-row-rank $T \in \mathbb{R}^{r\times d}$. Positive scaling can be absorbed into the potential, so normalize $\|T\|_F = 1$. Its compatibility is $DgT^\top = T(Dg)^\top$, also distributionally for C^1 potentials. The simplex identity then imposes $A_I T^\top = TA_I^\top$ on every target affine interpolant. Vertex replacement gives $e_i(Tb_j)^\top - (Tb_j)e_i^\top = 0$ for each extra-memory residual e_i and each barycentric gradient b_j. One nonzero residual confines $\operatorname{range} T$ to its span. Two independent residuals force $T = 0$. A quantitative wedge-product inequality and fixed integral tests yield the spectral tail in (13), without losing compatibility by first inverting T.

Theorem 5. *Under the same assumptions, suppose $\operatorname{rank} R = 1$ and write $a(s) = A_0 s + c + qf(s)$ with $\|q\| = 1$. Then*

$$\mathcal{E}^\star_{\mathrm{cvx}} = 0 \quad \Longleftrightarrow \quad \exists v \neq 0 :\ (I - qq^\top)A_0 v = 0, \quad \partial_v(q^\top a(s)) \geq 0 \quad \text{on } \mathcal{D}. \tag{14}$$

The criterion is unchanged by moving an affine scalar part between f and A_0. It requires a direction in which the other outputs are constant and the remaining scalar output is monotone. If it fails, convexity leaves positive risk even though general gradients have zero infimal risk. The proof constructs a global convex extension.

At zero bias let $K = ((0,0),(1,0),(0,1),(2,2))^\top$, $w = (3,-2,-2,1)^\top$, and $v_i = (w_i, k_{i2})$ with $O = I_2$. The residual has rank one and $a(s) = (f(s), s_2)$. At $s_0 = (3/4, 3/4)$, exact rational calculation gives $\nabla f(s_0) = 0$ and $\partial_{11} f(s_0) = -92/27$. Compatible directions must be parallel to e_1, but $\partial_1 f$ changes sign across s_0. Every sufficiently small centered ball thus has $\mathcal{E}^\star_{\mathrm{grad}} = 0 < \mathcal{E}^\star_{\mathrm{cvx}}$. Appendix E proves the closure results.

D Integral lower bounds

No genericity assumption is needed for the following test of a particular smooth response. Choose nonnegative compactly supported tests ψ_ℓ inside the mean domain and define

$$F_\ell = \mathbb{E}[\psi_\ell Da] = -\mathbb{E}[a(\nabla\psi_\ell)^\top], \qquad C_\psi = \sum_\ell \mathbb{E}\|\nabla\psi_\ell\|^2.$$

For normalized T and any convex potential, integration by parts gives

$$\mathbb{E}\|T\nabla\phi - a\|^2 \geq C_\psi^{-1} \sum_\ell \operatorname{dist}_F(F_\ell T^\top, \mathbb{S}_+^r)^2. \tag{15}$$

For each entry k and sign ε, minimize the numerator over $|T_{ij}| \leq 1$, $T_k = \varepsilon$, using PSD auxiliaries. Let its optimum be $\gamma_{k,\varepsilon}$. Homogeneity gives the bound uniform over output matrices

$$\mathcal{E}^\star_{\mathrm{cvx}} \geq \frac{\min_{k,\varepsilon} \gamma_{k,\varepsilon}}{rd\, C_\psi}. \tag{16}$$

Dual feasible matrices give lower bounds on these convex programs. We include moment uncertainty when evaluating the resulting bounds. Replacing the PSD cone by symmetric matrices gives a linear singular-value certificate for the general-gradient class. Zero certificates are inconclusive. Appendix F gives dual formulas and analytic integration-error control.

If the same law has a valid convex-class floor $c_0 > 0$, a globally Λ-Lipschitz convex gradient with maximum calibration error ϵ_N and fill distance h_N satisfies

$$\sqrt{c_0} \leq \epsilon_N + (\|O\|_{\mathrm{op}}\Lambda + L_h)h_N. \tag{17}$$

Here $h(s) = Oy(s)$ is L_h-Lipschitz. For exact calibration this forces $\Lambda \geq [\sqrt{c_0}/h_N - L_h]_+/\|O\|_{\mathrm{op}}$. Under the positive error lower bound in (17), any sequence of exact interpolants on increasingly dense calibration sets must have diverging gradient Lipschitz constants. This follows by combining the continuous error bound with the classical coverage estimate for a gradient in fixed coordinates.

E Spectral-tail bounds and approximation with rank-one residuals

The exact classification identifies when compatibility can hold. To obtain an approximation lower bound, we also need to control how far a target is from satisfying it. We first extract quantitative bounds on the interpolation coefficients. A rank-one construction and a global convex extension then establish the corresponding upper bounds.

E.1 Quantitative coefficient extraction

Compress observations using C_o as in Section C. Every full-row-rank T equals $C_o^\top OB$ for some $B \in GL(d)$: two full-row-rank $r \times d$ matrices lie in the same right $GL(d)$ orbit, as seen by extending their row-space bases. Positive scaling of T is absorbed by inverse scaling of the potential and preserves convexity.

For each key simplex I write the target affine interpolant as $A_I s + c_I$. The simplex identity gives

$$Da(s)T^\top - TDa(s)^\top = \sum_I h_I(s)(A_I T^\top - TA_I^\top), \qquad h_I(s) = p_I \det(X_I)^2/\det C.$$

The functions h_I are linearly independent on every open mean region: multiply a putative relation by $\det C$ and the softmax normalizer to power $d+1$, then use the distinct subset frequencies. Choose a smooth bump χ positive on an interior ball and tests $\psi_I = \chi h_I$. Their Gram matrix $G_{IJ} = \mathbb{E}\chi h_I h_J$ is positive definite. Integration by parts for $e = T\nabla\phi - a$ bounds the sum of squared Frobenius norms of the compatibility moments $\mathbb{E}[\psi_I(DaT^\top - TDa^\top)]$ by $4C_\psi\|T\|_F^2\mathbb{E}\|e\|^2$. Inverting G yields

$$\mathbb{E}\|e\|^2 \geq c_1\|T\|_F^{-2} \sum_I \|A_I T^\top - TA_I^\top\|_F^2, \tag{18}$$

One may take $C_\psi = \sum_I \mathbb{E}\|\nabla\psi_I\|^2$ and $c_1 = \lambda_{\min}(G)^2/(4C_\psi) > 0$, depending only on the fixed keys, bias and domain. Subsequent constants absorb the fixed barycentric-coordinate factors. The tests are fixed independently of the values and T, which ensures that the constant is uniform in both.

Fix an affine basis and its interpolant $A_0 s + c_0$. Stack extra-point residuals e_i as rows of E. Replacing basis vertex j changes the slope by $e_i b_j^\top/\lambda_j(k_i)$, with $b_j = \nabla\lambda_j$. All denominators are nonzero. Boundedness of this finite coefficient extraction gives a key-dependent $c_2 > 0$ such that

the coefficient sum in (18) dominates $c_2 \sum_{i,j} \|e_i(Tb_j)^\top - (Tb_j)e_i^\top\|_F^2$. Let $S = \sum_j b_j b_j^\top \succ 0$. The wedge sum is

$$2\sum_i \{\|e_i\|^2 \operatorname{tr}(TST^\top) - e_i^\top TST^\top e_i\} \geq 2\lambda_{\min}(S)(\|E\|_F^2 - \|E\|_{\mathrm{op}}^2)\|T\|_F^2.$$

For justification, the relevant quadratic form in each column of T is $\|E\|_F^2 I - E^\top E \succeq 0$, whose smallest eigenvalue is the displayed spectral tail. Finally $R = (I - P_X)JE$, where J embeds the extra rows and is an isometry. Applying this contraction to a best rank-one approximation of E shows $\tau(R) \leq \tau(E)$. This proves the lower bound in Theorem 4.

E.2 Approximation of targets with rank-one residuals

Write a target with rank-one residual as $a(s) = A_0 s + c + qf(s)$ and rotate the first output coordinate to q. The remaining affine slope $A_\perp$ has at most $r - 1 < d$ rows, so choose a unit $v \in \ker A_\perp$. Translate the mean ball to the origin and rotate input coordinates to $s = (x, z)$. After subtracting its constant output, the target is $(h(x, z), Pz)$. Perturb P arbitrarily slightly to a full-row-rank P_ϵ if necessary. This is possible since $r - 1 \leq d - 1$. Define locally

$$\Phi_M(x, z) = \int_0^x h(t, z)dt + \tfrac{M}{2}\|z\|^2, \qquad T_M = \begin{pmatrix} 1 & 0 \\ 0 & P_\epsilon/M \end{pmatrix}.$$

The resulting readout is $(h, P_\epsilon z + P_\epsilon \partial_z \int_0^x h(t, z)dt/M)$ and converges uniformly to the target on the compact ball as $M \to \infty$, then $\epsilon \to 0$. Constants are restored by adding a linear potential, since T_M has full row rank. A smooth cutoff extends the local potential to a global smooth potential without changing its gradient on the ball. The potentials in this construction may be nonconvex.

For a general target discard all but the largest singular component of R. The retained affine-plus-rank-one target lies in this closure. The removed response has norm at most the removed bank's Frobenius norm because $\|p\|_2 \leq 1$. Hence the infimal squared risk is at most τ^2. Exact attainment at nonzero residual rank would contradict Theorem 3. Equality almost everywhere extends to the open domain by continuity. Rank-zero targets admit exact convex realization by polar decomposition.

E.3 Necessity of directional monotonicity

Normalize an approximating convex sequence as $g_j = T_j \nabla \phi_j$, $\|T_j\|_F = 1$. A subsequence has limit $T_* \neq 0$. Distributionally, $Dg_j T_j^\top$ is a positive-semidefinite symmetric matrix-valued measure. Testing and passing to the L^2 limit shows $DaT_*^\top$ is symmetric PSD. Since a is smooth this holds pointwise in the interior, and by continuity on the closed ball. The simplex and vertex argument for a rank-one residual confines $T_* = qv^\top$ with $v \neq 0$. Symmetry requires $A_\perp v = 0$, and the remaining PSD scalar is $\partial_v(q^\top a) \geq 0$. This proves necessity without derivative convergence of the approximants.

E.4 Sufficiency and a global extension lemma

In the preceding rotated coordinates the criterion is $h_x \geq 0$. For each $\epsilon > 0$, use

$$\Phi_{\epsilon,M}(x, z) = \int_0^x h(t, z)dt + \tfrac{\epsilon}{2}x^2 + \tfrac{M}{2}\|z\|^2.$$

On a slightly larger ball, $h_x + \epsilon \geq \epsilon/2$ by continuity. All other derivatives are bounded there. A sufficiently large M makes the Hessian strictly positive by the Schur complement. The same full-row-rank T_M gives errors tending to zero as $\epsilon \to 0$, $P_\epsilon \to P$, and $M \to \infty$.

Here is the needed extension argument. Suppose a smooth potential u on a neighborhood of a convex compact set D obeys $mI \preceq D^2 u \preceq LI$ there, with $m > 0$. For all $s, t \in D$, strong convexity and Lipschitz continuity of the gradient imply

$$u(t) - u(s) - \nabla u(s)^\top (t - s) \geq \tfrac{m}{2}\|t - s\|^2 \geq \frac{\|\nabla u(t) - \nabla u(s)\|^2}{2\Lambda}, \quad \Lambda = L^2/m.$$

Define the following supremum of quadratic functions

$$H(q) = \sup_{s \in D} \left\{ s^\top q - u(s) + \frac{\|q - \nabla u(s)\|^2}{2\Lambda} \right\}. \tag{19}$$

Compactness makes this supremum finite everywhere and $1/\Lambda$-strongly convex. The inequalities ensure that at $q = \nabla u(s)$ the corresponding quadratic is active and $s \in \partial H(q)$. Hence H^* is global convex C^1 and has gradient exactly ∇u on D. This is the same classical conjugate interpolation construction (Taylor et al., 2017), here justified for a compact continuum of interpolation data. Applying it to each $\Phi_{\epsilon,M}$ completes the global sufficiency proof.

E.5 A COUNTEREXAMPLE WITH RANK-ONE RESIDUAL AND RISK SCALING WITH THE SECOND SINGULAR VALUE

For the stated four-key example, $w^\top X = 0$ and at uniform probabilities $\nabla_\theta f = 0$. Thus

$$D_s^2 f = C_0^{-1} \left(\tfrac{1}{4} K^\top \operatorname{diag}(w) K \right) C_0^{-1}, \quad C_0 = \begin{pmatrix} 11/16 & 7/16 \\ 7/16 & 11/16 \end{pmatrix}, \quad (D_s^2 f)_{11} = -92/27.$$

No orientation of e_1 has a nonnegative derivative on a centered ball. The compactness necessity argument therefore proves a strictly positive convex infimum.

For orthonormal residual directions w_1, w_2, q_1, q_2 about a full-row-rank affine target, sufficiently small fixed δ makes the target with residual $\delta w_1 q_1^\top$ satisfy strict directional monotonicity: choose a right-inverse direction with affine derivative q_1, then the scalar perturbation preserves its positive sign on the compact domain. Adding $\epsilon w_2 q_2^\top$, $0 < |\epsilon| \le \delta$, gives $\tau^2 = \epsilon^2$. Dropping this second component and using the convex closure of the first gives a convex risk upper bound ϵ^2. Theorem 4 supplies the matching order. The bounds concern the same fixed mean law and the readout class with unrestricted full-row-rank output matrices.

F FINITE MOMENT CERTIFICATES AND THEIR NUMERICAL ENCLOSURES

For a convex C^1 potential and nonnegative test ψ_ℓ, $\mathbb{E}\psi_\ell D(T\nabla\phi)T^\top \succeq 0$ in the distributional sense. Integration by parts and Cauchy–Schwarz bound its distance from $F_\ell T^\top$ by $\|T\|_{\mathrm{op}} \sqrt{\mathbb{E}\|a - g\|^2 \mathbb{E}\|\nabla\psi_\ell\|^2}$. Summing gives (15). For unit-Frobenius T, its largest absolute entry β is at least $1/\sqrt{rd}$, so T/β lies on one box face, and the objective is positively homogeneous of degree two. This proves (16), without requiring generic keys.

For any matrices Y_ℓ with $\operatorname{sym} Y_\ell \preceq 0$,

$$\operatorname{dist}_F(X, \mathbb{S}_+^r)^2 \ge 2\langle Y_\ell, X \rangle - \|Y_\ell\|_F^2.$$

Set $H = 2\sum_\ell Y_\ell^\top F_\ell$. On face (k, ε) a valid dual lower bound is $\varepsilon H_k - \sum_{j \ne k} |H_j| - \sum_\ell \|Y_\ell\|_F^2$. We project numerical residuals into the polar cone, then, for rational center matrices in the controlled experiments, verify polar feasibility and this dual expression with exact rational arithmetic.

If $\|F_\ell - \widehat{F}_\ell\|_F \le \delta_\ell$ and the computed box-face minimum has valid lower bound $\widehat{\gamma}$, the true minimum is at least

$$[\sqrt{\widehat{\gamma}} - \sqrt{rd}\,(\sum_\ell \delta_\ell^2)^{1/2}]_+^2.$$

This follows from the 1-Lipschitz distance to a closed convex cone and $\|T\|_F \le \sqrt{rd}$ on the box. The uncertainty term includes integration error.

Our test is the compact C^1 bump $\psi = c(1 - \|s - s_\ell\|^2/h^2)_+^2$, normalized by $\mathbb{E}_{\mathcal{D}}\psi = 1$. It can equivalently be approximated in the required Sobolev norm by smooth compactly supported tests. For a d-ball of radius R, $c = (R/h)^d (d+2)(d+4)/8$ and its gradient energy is

$$\mathbb{E}\|\nabla\psi\|^2 = \frac{8dc^2 h^{d-2}}{R^d} \mathrm{B}(d/2 + 1, 3).$$

If $\|D^2 a\| \le H_a$ on the test support, then $\|F_\ell - Da(s_\ell)\|_F \le \sqrt{d}\, H_a h$ when H_a bounds the bilinear Hessian norm. We include this dimension factor in calculations. Rational keys and rational e^{θ_j} give exact rational s_ℓ and $Da(s_\ell)$, avoiding inverse-map and derivative differencing error.

For completeness, let D_k bound key diameter, D_v value diameter, and $C \succeq \mu I$ on the neighborhood. Central-moment differentiation gives the conservative bound

$$H_a \le D_v(D_k^2/\mu^2 + D_k^4/\mu^3).$$

Subtracting an affine response before this bound is allowed since its Hessian vanishes. At zero bias, on a query ball of radius t, $C(\theta) \succeq e^{-D_k t}C(0)$. Its image contains a mean ball of radius $t\lambda_{\min}(C(0))e^{-D_k t}$. We use the rational lower bound $\lambda_{\min}C(0) \ge 1/\sum_{ij}|(C(0)^{-1})_{ij}|$ and $e^{-x} \ge 2^{-\lceil 2x \rceil}$ to certify the mean domains. The resulting error bound depends on these local geometric constants.

For any point select a nearest calibration mean. The triangle inequality gives pointwise observed error at most $\epsilon_N + (\|O\|_{\mathrm{op}}\Lambda + L_h)h_N$. Integrating and comparing with c_0 proves (17). If $h_N \to 0$ and calibration is exact, every feasible sequence has $\Lambda_N \to \infty$. The inequality gives a necessary growth rate for the smoothness bound.

G Proof of Theorem 1

We first establish the convexification step and then prove the lower bound, upper bound, and exact representation criterion.

Convexification. Let ψ be smooth near the closed ball $\mathcal{D}$. On a slightly larger closed ball, choose M so that $\nabla^2\psi + MI \succ 0$. The global extension lemma in Appendix E gives a global convex C^1 potential whose gradient equals $\nabla\psi + Ms$ on $\mathcal{D}$. Replacing L by $L - MT$ therefore preserves the response exactly. A global C^1 nonconvex potential can first be mollified: its gradients converge uniformly on a neighborhood of $\mathcal{D}$. For each approximant choose its own M. Thus, without a curvature bound, the convex and general-gradient versions of (3) have the same infimal risk. The convexifying constants may vary along the approximating sequence.

Lower bound. Fix U, and let P have orthonormal columns spanning $U^\perp$. Since T is surjective, $P^\top T$ has full row rank $r - k$. The projected target after subtracting the affine branch is a frozen attention response with value bank $WP - KL^\top P - \mathbf{1}b^\top P$. Its residual is exactly RP. Theorem 4 applies with observation rank $r - k \ge 2$. Its constant depends on keys, bias, domain, and this rank, and is uniform in values, P, L, and b. Orthogonal projection contracts the error, so every competitor has risk at least $c_k \sum_{j=2}^{r-k}\sigma_j(RP)^2$. By the min–max characterization, $\sigma_j(RP) \ge \sigma_{j+k}(R)$. Equality in the sum is attained by discarding the first k right singular directions. This proves the lower bound after optimizing U.

Upper bound. Decompose $W = XA + R$ and use the affine branch for $p^\top XA$. Choose U as the first k right singular vectors of R and set $z(s) = U^\top R^\top \widetilde{p}(s)$ on a neighborhood of $\mathcal{D}$. These functions have global smooth extensions by a cutoff inside the mean hull. The next singular component gives a target with rank-one residual in $\mathbb{R}^r$. The general-gradient closure construction of Theorem 4 supplies full-row-rank T and smooth potentials approximating it uniformly. Convexify each potential with affine compensation as above. The discarded response is bounded in squared norm by $\sum_{j=k+2}^{r}\sigma_j(R)^2$, since $\|p\|_2 \le 1$. The approximation error tends uniformly to zero, proving the upper bound.

Exact representation for $k \le r-2$. If an exact representation exists, project onto $U^\perp$ and subtract the affine branch. The exact classification, Theorem 3, forces $RP = 0$. Hence $\operatorname{rank} R \le k$. Conversely, if $\operatorname{rank} R \le k$, choose U containing the entire row space of R and implement its response in z. Choose any full-row-rank T and $\phi(s) = \|s\|^2/2$. The affine branch realizes the affine target minus Ts. This gives a global convex potential and exact agreement on the mean domain. Equality almost everywhere is enough for necessity, by continuity and the domain's nonempty interior.

The case $k = r - 1$. Rotate output coordinates so U spans the last $r - 1$ coordinates. Translate the ball center to zero, extend the target smoothly to a neighborhood, and put $\psi(s) = \int_0^{s_1} a_1(t, s_2, \ldots, s_d)dt$. The integration segment stays in the centered ball. Take T to select the first r input coordinates. Convexify ψ and cancel MTs using the affine branch. The first output

now equals a_1 exactly. The remaining $r-1$ discrepancies are arbitrary smooth functions carried by z. Smooth cutoff and the global convex extension provide global admissibility without altering $\mathcal{D}$. Finally, the two minimum-dimension formulas follow by checking $q=0$, $1 \le q < r$ and $q = r$.

G.1 Full-domain integral tests that eliminate affine branches

To eliminate the affine branch, choose signed C^1 tests vanishing with their first derivatives on $\partial\mathcal{D}$ and satisfying $\mathbb{E}\psi_\ell = 0$. Put $F_\ell = -\mathbb{E}[a(\nabla\psi_\ell)^\top]$ and let $\mathcal{C}(T)$ stack the entries above the diagonal of $F_\ell T^\top - T F_\ell^\top$. For $g = Ls + b + T\nabla\phi$, integration by parts eliminates L because $\mathbb{E}\psi_\ell = 0$. For $\|T\|_F = 1$, Cauchy–Schwarz yields the conservative bound

$$\mathbb{E}\|g-a\|^2 \ge \frac{\sigma_{\min}(\mathcal{C})^2}{2\sum_\ell \mathbb{E}\|\nabla\psi_\ell\|^2}. \tag{20}$$

Signed tests enforce symmetry, whereas the positive-semidefinite cone test requires nonnegative tests. If their gradient Gram matrix is H, an invertible change of test basis is allowed. We whiten by $H^{-1/2}$ before forming $\mathcal{C}$. We integrate these tests over the full mean ball. Appendix I bounds the integration error and propagates it through the whitening and singular-value calculations.

H Proof of Theorem 2

We retain the smallest singular value in the coefficient bound used for Theorem 4.

Fix $P \in \mathbb{R}^{r\times m}$ with orthonormal columns, $m = r-k \ge 2$, and write $A = P^\top T$. The Rayleigh bounds on $TT^\top$ imply $\sigma_m(A) \ge \sigma_r(T)$ and $\sigma_1(A) \le \sigma_1(T)$. Therefore $\mathrm{cond}_2(A) \le \chi$ and

$$\frac{\sigma_m(A)^2}{\|A\|_F^2} \ge \frac{1}{m\chi^2}.$$

Subtract the projected affine branch. As in the proof of Theorem 1, the residual is $R_P = RP$ and the quantitative extraction constants depend only on keys, bias, mean domain and m. Let E collect the residuals relative to one key simplex and let $S = \sum_j b_j b_j^\top \succ 0$. With $Q = \|E\|_F^2 I_m - E^\top E \succeq 0$, the squared wedge sum is

$$2\,\mathrm{tr}(QASA^\top) \ge 2\lambda_{\min}(S)\,\mathrm{tr}(QAA^\top).$$

It simultaneously dominates

$$2\lambda_{\min}(S)\left(\sum_{j=2}^m \sigma_j(E)^2\right)\|A\|_F^2 \quad \text{and} \quad 2\lambda_{\min}(S)(m-1)\|E\|_F^2\sigma_m(A)^2,$$

because $\lambda_{\min}(Q) = \sum_{j=2}^m \sigma_j(E)^2$ and $\mathrm{tr}\,Q = (m-1)\|E\|_F^2$. The fixed extraction and division by $\|A\|_F^2$ therefore give, for a common positive key-dependent constant c_m,

$$\text{risk} \ge c_m \sum_{j=2}^m \sigma_j(E)^2, \qquad \text{risk} \ge c_m \frac{m-1}{m\chi^2}\|E\|_F^2.$$

The contraction $R_P = (I - P_X)JE$ implies the same inequalities with E replaced by R_P. The maximum of two bounds is at least half their sum. Finally $\sigma_j(RP) \ge \sigma_{j+k}(R)$ gives (6), for example with $\widetilde{c}_k = c_m(m-1)/(2m)$ after retaining only the leading term from $\|RP\|_F^2$. All constants are uniform in P, the affine branch, values and χ.

If $q = \mathrm{rank}\,R > k$ and $k \le r-2$, the second term is strictly positive for any fixed finite χ. If $q \le k$, use the exact construction with T selecting r orthonormal input coordinates and compensate its quadratic potential by the affine branch. At $k = r-1$, the scalar integration construction also uses such a T. Both have condition number one, proving the minimum-dimension statement including $q = r$. For the stated rank-two one-dimensional correction construction the existing upper bound is C/χ^2, while the second term above is positive and of the same order. This proves the matching rate for that target.

I VALIDATED NUMERICAL INTEGRATION

We derive the interval bounds in (8) for the rank-two target and uniform mean distribution specified below.

The primary key bank and rational value construction are

$$K = \begin{pmatrix} 3 & 2 & -3 \\ 2 & 0 & -3 \\ -3 & 2 & -2 \\ -2 & -1 & 3 \\ 0 & -3 & -3 \\ 3 & -2 & 3 \\ 1 & -1 & -1 \end{pmatrix}, \qquad V_{\mathrm{rat}} = \begin{pmatrix} 52/17 & -195/131 & 0 \\ -82/17 & -80/131 & 0 \\ -3/17 & 330/131 & 0 \\ 1/34 & -380/131 & 0 \\ 65/34 & 0 & 0 \\ 0 & 325/131 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Set $\eta = 0$, $O = I_3$, $c = (4, -3, -6)^\top/7$, and let R be 0.65 times the smallest distance from c to a facet of $\operatorname{conv} K$, numerically 0.792623020481267. The law is uniform on the closed ball $\|s - c\| \le R$. The rational construction has $[\mathbf{1}, K]^\top V_{\mathrm{rat}} = 0$ and rank two. For the numerical certificate, each entry of V_{rat} and c, and the decimal radius 0.792623020481267, is rounded to nearest IEEE binary64 (ties to even). These rounded values are then treated as exact inputs.

In normalized coordinates $x = (s - c)/R$, use all 19 monomials of degrees one through three and $\psi_\alpha = (1 - \|x\|^2)^2(x^\alpha - m_\alpha)$, where m_α makes the uniform-ball expectation zero. These tests and their first derivatives vanish on the boundary. Uniform-ball polynomial moments are exact rationals: odd moments vanish and, for even multi-indices,

$$\mathbb{E}x^{2\alpha} = \frac{\prod_j (2\alpha_j - 1)!!}{\prod_{i=1}^{|\alpha|}(3 + 2i)}.$$

Thus the raw gradient Gram matrix H is evaluated exactly, treating the binary64 radius as an exact binary rational. We choose a numerical whitening matrix B, then regard its binary64 entries as exact binary coefficients. Directed ball arithmetic accounts for the whitening error and evaluates $C_\psi = \operatorname{tr}(BHB^\top)$, with upper bound 19.0000000001. The Frobenius upper bound $\|B\|_F < 43.599$ propagates raw moment uncertainty conservatively.

At a numerical inverse candidate θ_0, directed arithmetic encloses $s(\theta_0)$ and C_0. A lower bound is $\mu_0 = 1/\|C_0^{-1}\|_F$. If D is the key diameter, exponential tilting and the variational characterization of covariance imply

$$e^{-D\|h\|} C_0 \preceq C(\theta_0 + h) \preceq e^{D\|h\|} C_0.$$

Consequently the mean image of the query ball of radius a contains the mean ball of radius $\mu_0(1 - e^{-Da})/D$. To see this, integrate the outward directional derivative along every radius. The strictly convex log-partition objective for a mean in that ball has its minimizer inside the query ball. This validates an inverse enclosure for an entire mean box. We enlarge the solved radius and subdivide boxes that do not satisfy the test. The inverse covariance is enclosed using $e^{-Da}C_0^{-1} \preceq C^{-1} \preceq e^{Da}C_0^{-1}$ and the entrywise bound $|(C^{-1} - C_0^{-1})_{ij}| \le (e^{Da} - 1)\sqrt{(C_0^{-1})_{ii}(C_0^{-1})_{jj}}$.

Parameterize the uniform ball by $(u, v, w) \in [0, 1]^3$:

$$x = u(\sin(\pi v)\cos(2\pi w), \sin(\pi v)\sin(2\pi w), \cos(\pi v)), \quad d\mu = \tfrac{3\pi}{2} u^2 \sin(\pi v)\, du\, dv\, dw.$$

On each parameter box we enclose the eighth directional derivatives of every raw moment integrand using interval Taylor series. The implicit derivatives are generated from $C\theta' = s'$ and $p_i' = p_i(k_i - s)^\top \theta'$, with interval inverse covariance at order zero. Recursion encloses all higher coefficients over the entire box. This parameterization is smooth on each integration box.

Four-point Gauss integration on a one-dimensional interval of length h, applied to a function whose eighth Taylor coefficient has bound M, has mean-error bound $h^8 M(4!)^4/[9(8!)^2]$. Tensorizing positive normalized rules gives the sum of the three directional bounds. Adaptive dyadic subdivision therefore bounds the vector moment error by summing box volumes times these remainders. The run uses 137,438 boxes. Gauss nodes and weights are evaluated as exact radical expressions in Arb ball arithmetic. At each node, softmax and its inverse-mean residual are enclosed. The inverse-radius

estimate bounds the remaining response uncertainty. Floating-point accumulation is bounded using $\gamma_n = nu/(1-nu)$ and an analytic bound on raw moment magnitudes. The final raw moment error, including accumulation and matrix-product allowance, is at most 1.045×10^{-6} in the Frobenius norm of the complete $19 \times 2 \times 3$ raw-moment tensor (114 entries). Each directional interval bound is first aggregated in the Euclidean norm of those 114 entries. Box contributions are then summed by the triangle inequality. This tensor norm is the error propagated through whitening and the Jacobian compatibility operator.

Let A represent the linear map from T to the whitened moments of $Da\,T^\top - TDa^\top$, and let Y be a numerical right preconditioner. Directed arithmetic bounds $\delta = \|(AY)^\top AY - I\|_F$ and $\nu = \|Y^\top Y\|_\infty$. Thus $\sigma_{\min}(A) \geq \sqrt{1-\delta}/\sqrt{\nu}$ for the numerical matrix. Subtracting $\sqrt{2}\|B\|_F$ times the raw moment error gives a true operator singular-value lower bound exceeding 0.0005846188. Equation (20) then gives

$$\mathcal{E}_0^\star \geq 8.99418 \times 10^{-9}.$$

This includes arbitrary affine branches, full-row-rank output matrices, and general C^1 potentials, including nonconvex ones.

For the explicit potential $\psi = \int_0^{s_2-c_2} f_2(s_1, c_2+t, s_3)dt$, the only uncanceled error is $\chi^{-1}\partial_3\psi$. Convexification and global extension prove admissibility. To bound its risk, partition the mean cube into 128^3 boxes and discard only boxes wholly outside the ball. At fixed first and third coordinates, accumulate interval enclosures of $\partial_3 f_2$ along the second coordinate, including partial terminal cells, then square the enclosed integral. Boundary cells use their full volume, yielding an upper bound under the same ball law. The Jacobian enclosures are sharpened by an interval mean-value bound using

$$\nabla_s^2 f_2 = C^{-1}\mathbb{E}_p[(v_2 - f_2 - D_s f_2(k-s))(k-s)(k-s)^\top]C^{-1}.$$

Both direct covariance and mean-value enclosures are intersected. The sum over 1,136,840 intersecting boxes gives the squared-error coefficient upper bound 0.004452477. Hence at $\chi = 1000$,

$$E_{\text{one}} \leq 4.45248 \times 10^{-9} < 8.99418 \times 10^{-9} \leq \mathcal{E}_0^\star.$$

The ratio of the certified lower and upper bounds exceeds two.

The endpoint verification evaluates the polynomial Gram matrix and the final inequalities in directed arithmetic. It combines the moment and integration enclosures described above, with floating-point inputs interpreted as exact binary64 values. The full-domain integration validates the derivative enclosures and remainder bounds on every box.

J PROJECTION FORMULA FOR ORDINARY NONLINEAR READOUTS

We include the elementary projection argument to fix the comparison class. Let $\mathcal{A}$ be the finite-dimensional closed subspace of affine scalar functions in $L^2(\mu)$, and apply its orthogonal projector componentwise. For fixed U, projection onto $U^\perp$ eliminates z. The optimal remaining affine error is $\mathbb{E}\|(I-UU^\top)e_\mu\|^2 = \operatorname{tr}\Sigma_\mu - \operatorname{tr}(U^\top\Sigma_\mu U)$. It is attained by $Ls + b = a^\mu_{\text{aff}}$ and $z = U^\top e_\mu$. Maximizing the captured trace chooses the top k eigenvectors and gives (7). The formula also holds for an empirical measure, with its corresponding covariance spectrum.

For the full-support law on the mean ball, $v \in \ker\Sigma_\mu$ means that $v^\top a(s) = h_0 + h^\top s$ almost everywhere. Continuity makes the identity hold throughout the ball. Multiplication by the partition function and analytic continuation in natural coordinates give

$$\sum_i (w_i^\top v - h_0 - h^\top k_i)e^{k_i^\top\theta+\eta_i} = 0.$$

The keys are distinct. Choose a line direction with all $k_i^\top u$ distinct. Differentiating along that line gives a nonsingular Vandermonde system. Thus every coefficient vanishes, $Wv \in \operatorname{col} X$, equivalently $Rv = 0$. The converse is immediate. Hence $\ker\Sigma_\mu = \ker R$, and the ordinary exact and approximation dimensions both equal q. This argument requires only affine spanning and distinct keys, weaker than the generic hypotheses needed for our mixed class boundary.

Finally, $(I - P_X)W$ is a linear map into an $(n-d-1)$-dimensional space. As W varies freely its maximum rank is $\min\{r, n-d-1\}$. Lower-rank matrices are the common zero set of maximal minors. Since fixed rank-r observation permits arbitrary W, this proves the stated generic rank formula.

K FULL-RANK AND RANK-DEFICIENT OUTPUT MATRICES

Write $A = C_o^\top O$, which has full row rank. Invertible right actions are transitive on full-row-rank $r \times d$ matrices: extend the rows of A and T to invertible square matrices $\bar{A}$ and $\bar{T}$, and take $B = \bar{A}^{-1}\bar{T}$. Then $AB = T$. Hence learning any full-row-rank T is precisely learning an invertible output-coordinate change before the fixed observation. Affine and correction branches can restore rank to the total response, but do not enforce this property on its potential branch.

Consider a smooth target on $\mathcal{D}$ of the form $a(s) = A_0 s + c + uf(s)$, where $u \neq 0$. Extend f smoothly outside a neighborhood of $\mathcal{D}$ and define

$$\psi(s) = \int_0^{s_1} f(t, s_2, \ldots, s_d)\, dt, \qquad T = ue_1^\top.$$

Then $T\nabla\psi = uf$ on $\mathcal{D}$. Choose M so that $\psi + M\|s\|^2/2$ is strongly convex on a slightly larger ball. The extension construction in Appendix E gives a global convex C^1 potential ϕ with that gradient on $\mathcal{D}$. Taking $L = A_0 - MT$ yields $Ls + c + T\nabla\phi = a$ with $k = 0$ and $\operatorname{rank} T = 1$. In particular, the generic $q = 1 < r$ exact correction dimension changes from one to zero if this rank loss is allowed. The full-row-rank assumption therefore changes the exact representation criterion.

L DIRECTION CONTROLS AND GEOMETRIC SENSITIVITY

We repeat both readout families with each of two common output bases: calibration residual PCA and value-residual SVD. For the potential-based readout the same basis rotates both T and the correction. Splits, feature widths, seeds, least-squares cutoff, validation tolerance, storage cap and cost accounting after algebraic simplification are unchanged. This control refits all candidates with the same validation-only selection rule. Each entry in Table 5 is the least-cost validation-feasible

Table 5: Shared direction selection. Each entry is the multiplication count, followed by independent test risk divided by tolerance in parentheses. Both readout families use the same output basis. Values below one meet the test tolerance.

	Common PCA		Common SVD	
Target	Ordinary	Potential	Ordinary	Potential
Primary 3D	178 (0.864)	239 (0.957)	178 (0.864)	239 (0.942)
Four-output 4D	504 (0.227)	788 (0.083)	504 (0.227)	788 (0.083)
New values 1	178 (0.760)	239 (0.906)	204 (0.693)	274 (0.526)
New values 2	178 (0.867)	239 (0.586)	204 (1.030)	239 (0.964)
New values 3	204 (1.013)	274 (0.858)	204 (1.013)	274 (0.949)
Frozen Iris	134 (0.095)	168 (0.089)	134 (0.095)	168 (0.067)

member of its family under the storage cap.

For the geometry experiment, set $D_t = \operatorname{diag}(1, 1, t)$ and $K_t = KD_t$, retain V, and use the uniform ball centered at $D_t c$ with radius tR. This ball lies inside the transformed original ball and hence inside the new key hull. Every $t > 0$ preserves the genericity conditions and $\operatorname{col}[\mathbf{1}, K_t]$, so the value-residual singular values stay fixed. The limiting keys at $t = 0$ fail to affinely span. The experiment therefore varies key and domain geometry together.

Figure 2 uses order-18 quadrature, with order 12 as a refinement check. The matrix A uses the same 19 whitened polynomial tests as the primary certificate. These estimates describe the geometric dependence of the bound.

M FINITE-MODEL COMPARISON PROTOCOL

The comparison uses five synthetic targets and one frozen regressor. The synthetic targets are the certified three-dimensional example, the four-output example, and three value banks on the same three-dimensional keys, generated with seeds 92201–92203. The sixth target is a trained six-key

Iris regressor held fixed throughout. The candidate families and selection rule were specified before these baseline runs.

Each continuous distribution supplies 1,024 calibration, 512 validation, and 4,096 independent test inputs. Synthetic inputs are uniform on the specified mean balls. Iris inputs are obtained from uniform queries on $[-2, 2]^2$ through the mean map.

For the ordinary class, calibration residual PCA supplies U. For the potential-based class, value-residual SVD supplies U. Mean coordinates are centered and normalized by a scalar radius. The two families share targets, samples, feature seeds, and fitting tolerances. Tanh widths are 32 and 96, with seeds 0, 1, and 2, Gaussian weight scale 0.7, and bias scale 0.2. Output coefficients are fitted by least squares with relative SVD cutoff 10^{-12}. The ordinary family uses $k = 0, \ldots, r$ and the potential-based family uses $k = 0, \ldots, r - 1$. The latter uses output condition numbers 10 and 1,000, with the largest singular direction chosen as the first direction outside U.

The error tolerance is one percent of the calibration affine risk, and the storage limit is 1,000 scalars. Among validation-feasible configurations, selection minimizes dense multiplications, then stored scalars, then enumeration index. Independent test inputs are used only for evaluation. The cost comparison uses the simplified expressions in Appendix N.

The regressor predicts petal length and width from sepal length and width using six keys. A fixed random permutation (seed 9182027) gives 100 training and 50 test examples. Inputs and outputs are standardized using training statistics. Keys and values are initialized independently from $N(0, 0.5^2)$ and biases at zero. Nonlinear least squares minimizes mean squared response error plus a 10^{-3} squared parameter penalty, using an analytic Jacobian, at most 300 function evaluations, and function, step, and gradient tolerances of 10^{-10}. The fitted regressor is held fixed in all readout comparisons.

N Inference and numerical least squares

The representation used during fitting can contain canceling terms that need not be evaluated at inference. We simplify both readout families before comparing their computational costs. We evaluate algebraically equivalent inference expressions for all 330 fitted models, using the same calibration, validation, and test inputs. Table 3 reports the resulting multiplication counts.

Write $v_j(x) = \tanh(w_j^\top x + b_j)$ and $\psi = \sum_j c_j \tanh(w_j^\top x + b_j)$. Quadratic convexification and affine compensation cancel identically. Precomputing $G_{j,:} = c_j (T w_j)^\top$ gives $T \nabla \psi = \sum_j G_{j,:} - (v \odot v) G$, with the constant absorbed into the affine intercept. The nonlinear correction is evaluated as either $(vH)U^\top$ or $v(HU^\top)$, choosing the smaller multiplication count and then storage count. Both model families use this rule.

Runtime storage counts the contracted matrices and can increase when arithmetic decreases. Common input normalization is excluded from both storage counts. For width h, both families evaluate h tanh activations, and potential-based models additionally evaluate h squarings. These operations are excluded from dense multiplication counts. Validation and test errors use the simplified expressions.

Every fit uses a relative singular-value cutoff of 10^{-12}. The fitted models provide feasible approximation errors. Truncation can remove small nonzero singular directions, so these errors need not equal the full least-squares optimum.

O Synthetic targets and finite-model experiments

The three-dimensional target is specified in Appendix I. It is shared by the certified comparison, polynomial references, and fixed-feature models. Directions are chosen from the value bank before evaluation.

For the four-output example, the keys and first two value columns are

$$K = \begin{pmatrix} -1 & 3 & -1 & 3 \\ -3 & -1 & -1 & -2 \\ 2 & 3 & 0 & -1 \\ 3 & 0 & -1 & -2 \\ 2 & 2 & -1 & 1 \\ -2 & -2 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 2 & -1 & 3 & -3 \end{pmatrix}, \qquad V_{:,1:2} = \begin{pmatrix} 740/307 & 695/288 \\ -290/307 & -15/16 \\ -10/307 & -5/144 \\ 790/307 & 185/72 \\ -1235/307 & -145/36 \\ 5/307 & 0 \\ 0 & 5/288 \\ 0 & 0 \end{pmatrix},$$

and the other two value columns are zero. All 56 simplex and subset-sum tests hold. The mean-ball center is $(1/2, 3/8, 1/8, -1/4)^\top$. The mean-ball radius is 0.370894219812. The observation is I_4 and the bias is zero. Sphere quadrature uses Gauss–Chebyshev weights for the first spherical coordinate, Gauss–Legendre for the second and radius, and a uniform azimuth. We compare quadrature orders 6 and 9.

Exact representation assigns the two residual scalar functions to the correction branch and cancels a quadratic potential affinely. With one correction direction, the construction in the proof of Theorem 1 instead leaves an error proportional to χ^{-2}. The two-correction identity is checked by evaluating the original softmax output and the replacement from the key mean.

For the finite-model comparison in Table 2, all models use the same 2,048 training means and an independent order-18 evaluation rule. The residual SVD fixes an output rotation, with deterministic signs. In that frame the first k directions are correction directions and $T = \mathrm{diag}(1/\chi, 1, 1/\chi)$. Polynomial degree is seven. Frozen tanh feature widths are 32 and 128, with seeds 0, 1, 2 and nonzero hidden biases. The affine terms and the potential and correction linear coefficients are solved jointly by SVD least squares at relative cutoff 10^{-12}. The output matrix and hidden features remain fixed during fitting. Convexification and affine compensation preserve every finite fit on the ball.