MINIMUM-KL ACTIVATION STEERING THROUGH A NORMALIZED LINEAR READOUT

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering interventions can achieve the same increase in a desired output score while changing other next-token probabilities differently. At a fixed context, we use the distribution minimizing steered-to-baseline KL under the score constraint as a method-independent benchmark. This minimum-KL target is the standard exponential tilt of the baseline distribution. We ask whether a bounded activation intervention can realize it. For interventions immediately before RMSNorm and a linear readout, we give necessary and sufficient conditions for exact attainment: the target logit direction must lie in the centered-readout range, normalization must admit a realizing state, and the minimum required intervention norm must not exceed the budget. When the score increase is feasible but the target is unattainable, we prove a positive lower bound on excess KL. For compatible directions, a nontrivial centered-readout nullspace gives a single attainment threshold, whereas an injective readout can lose and later recover attainability as the target increase grows. Experiments on three small pretrained models and four lexical event scores find large range mismatches in all twelve model–score pairs, ruling out every positive minimum-KL tilt at these final readouts. Constructed compatible targets are exactly reachable only over a narrow range of score increases at the tested budget. These results distinguish achieving a target score from achieving it with the smallest distributional change under the specified KL criterion.

1 INTRODUCTION

Activation steering changes the hidden representations of a language model during inference to favor a desired output property without changing model parameters (Turner et al., 2023; Rimsky et al., 2024). A higher score for that property, however, does not specify how the remaining next-token probabilities should change. At a fixed context, two interventions can achieve the same score increase but change other token probabilities in different ways. We study selective steering as a constrained distributional problem: achieve a prescribed score increase while minimizing departure from the baseline.

To raise the probability of digit tokens from π_0 to π, we must move probability mass from non-digit to digit tokens. The request specifies the group masses but leaves the relative probabilities within each group unconstrained. A distributional benchmark distinguishes interventions that achieve the same gain with different changes to these conditional probabilities.

We measure change from the baseline p_0 to a steered law q by steered-to-baseline KL, $D_{\mathrm{KL}}(q\|p_0) = \mathbb{E}_q \log(q_i/p_{0,i})$. This is the orientation used in KL-regularized reward optimization for language-model alignment: both reward and the log-likelihood penalty are averaged under the optimized distribution (Rafailov et al., 2023). We use the same criterion at a fixed context, with a token score f and a prescribed increase Δ. The resulting method-independent benchmark is

$$\mathcal{I}(\Delta) = \min_{q:\, \mathbb{E}_q f \geq \mathbb{E}_{p_0} f + \Delta} D_{\mathrm{KL}}(q\|p_0).$$

The inequality requires at least the requested improvement. Minimizing over all output distributions gives a lower bound on the KL cost of any steering method that meets this requirement, independently of its intervention class or budget.

For interior positive targets, the unique minimizer is the exponential tilt $q_\lambda(i) \propto p_0(i)e^{\lambda f_i}$ (Csiszár, 1975). We call this law the *minimum-KL target* under the chosen criterion. For digit tokens, it

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.

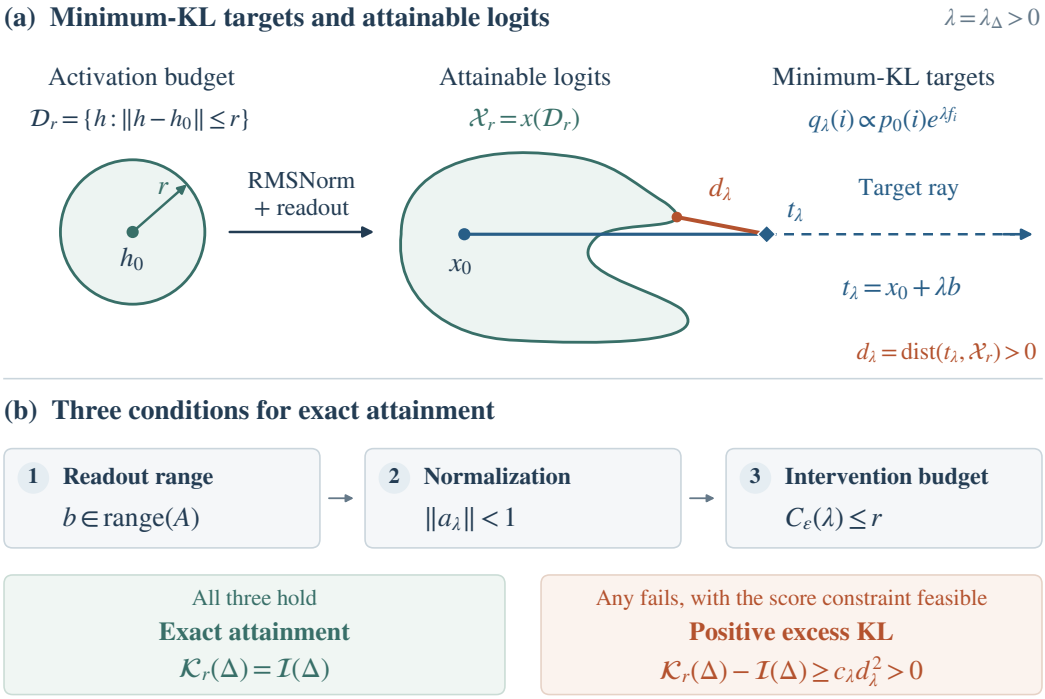

Figure 1: Minimum-KL targets and finite-budget attainability. (a) The activation ball centered at h_0 maps to $\mathcal{X}_r$, while minimum-KL targets follow a ray in centered-logit space. The orange segment joins the selected target to its nearest point in $\mathcal{X}_r$. (b) Exact attainment requires all three conditions; if the score constraint is feasible but any condition fails, excess KL is positive. Here A and b are the centered readout and score, $a_\lambda = A^\dagger t_\lambda$, $C_\varepsilon(\lambda)$ is the minimum realizing intervention norm, and $c_\lambda > 0$. The geometry is schematic; Sections 2–3 give the exact conditions.

changes only the group masses and preserves the baseline conditional laws. The KL chain rule assigns any redistribution within either group an additional cost. For general scores, an exact identity separates excess KL into deviation from the target and score overshoot.

Can bounded activation steering realize this target? We intervene on the final hidden activation immediately before RMSNorm (Zhang & Sennrich, 2019) and the subsequent linear vocabulary readout, with a Euclidean displacement budget in the model's native activation coordinates. After removing the common softmax logit offset, the minimum-KL targets form a straight ray. The normalized readout and the intervention budget determine which points on this ray can be reached (Figure 1).

Main result and consequences. We give necessary and sufficient conditions for exact attainment: the target direction must lie in the centered-readout range, normalization must admit a realizing state, and the minimum required intervention norm must not exceed the budget (Theorem 1). We compute this norm for every readout rank. When the score constraint is feasible but the target is unattainable, we prove a positive lower bound on excess KL. For compatible directions, the same minimum-norm criterion determines how attainability changes with the requested gain: a nontrivial centered-readout nullspace gives a single threshold, whereas an injective readout can lose and later recover exact attainment. We also apply the criterion to restricted interventions and earlier layers under bounded downstream errors.

Testing the obstructions. We first test the range condition on three small pretrained models and four lexical event scores. All twelve target directions have large out-of-range residuals, ruling out every positive minimum-KL tilt at these readouts (Table 1). We then construct compatible scores to isolate normalization and budget. Exact attainment holds only over a narrow range of score increases at the tested budget.

Relation to prior work. Dual steering minimizes the opposite orientation, $D_{\mathrm{KL}}(p_0\|q)$, over softmax representations on a linear-probe hyperplane (Park et al., 2026). The difference is both the KL orientation and the constraint: our benchmark fixes an output-mean requirement over all distributions, then asks whether a bounded activation intervention can reach the optimizer through the normalized readout. For event indicators, reversing KL under the same output constraint gives the same target and hence the same exact-attainment conditions. This includes all four lexical scores in our experiments. General real-valued scores give different targets (Appendix A).

Local Fisher and curvature analyses study efficient directions or specified paths (Azizi et al., 2025; Picozzi, 2026). FishBack uses a pullback Fisher metric for local intermediate-layer steering (Wang et al., 2026). Work on prompt-to-activation non-surjectivity asks whether a discrete prompt can produce a given activation (Mishra et al., 2026). Here, the relevant preimage is an activation that realizes a specified output law. Other steering and control methods use different intervention domains or objectives (You et al., 2026; Dalili et al., 2026; Nguyen et al., 2026; Skifstad et al., 2026).

2 A DISTRIBUTION-SPACE BENCHMARK FOR SELECTIVE STEERING

2.1 THE MINIMUM-KL TARGET

Fix a generation context and a strictly positive baseline next-token distribution p_0 on $m \geq 2$ tokens. Let the nonconstant vector $f \in \mathbb{R}^m$ assign score f_i to token i, and write $s_0 = \mathbb{E}_{p_0} f$. For an interior positive target increase $0 < \Delta < \max_i f_i - s_0$, define the unrestricted distribution-space benchmark by

$$\mathcal{I}(\Delta) = \min_{p:\, \mathbb{E}_p f \geq s_0 + \Delta} D_{\mathrm{KL}}(p\|p_0). \tag{1}$$

The score f specifies the token-level property to be controlled, and the inequality requires a minimum improvement in its expectation. Optimizing over all output laws makes $\mathcal{I}(\Delta)$ a lower bound for every intervention class evaluated under the same score constraint and KL orientation. Thus the benchmark depends only on p_0, f, and Δ; activation restrictions enter separately.

Strict positivity of p_0 holds for a finite-logit softmax and makes KL finite for every output law. Nonconstancy of f permits a positive improvement. The interior restriction on Δ excludes the trivial baseline target and the extreme target supported only on score-maximizing tokens; it ensures a finite positive tilt parameter. These conditions ensure feasibility in distribution space, while feasibility under an activation budget remains to be determined.

Set $\tilde{f} = f - s_0\mathbf{1}$, where $\mathbf{1}$ is the all-ones vector, and $\Lambda(\lambda) = \log \mathbb{E}_{p_0} \exp(\lambda \tilde{f})$. Information projection yields a unique $\lambda_\Delta > 0$ satisfying $\Lambda'(\lambda_\Delta) = \Delta$. The unique minimum-KL target q_{λ_Δ} then satisfies

$$q_\lambda(i) = p_0(i) e^{\lambda \tilde{f}_i - \Lambda(\lambda)}, \qquad \mathcal{I}(\Delta) = \lambda_\Delta \Delta - \Lambda(\lambda_\Delta). \tag{2}$$

The same q_λ uniquely maximizes $\lambda \mathbb{E}_p f - D_{\mathrm{KL}}(p\|p_0)$ over all output laws p, linking the benchmark to KL-regularized reward optimization. Specifying Δ selects the penalty tradeoff through λ_Δ: the constrained optimum meets the mean requirement with equality, independently of any activation direction. For every output law p, with $\lambda = \lambda_\Delta$, the exact identity

$$D_{\mathrm{KL}}(p\|p_0) - \mathcal{I}(\Delta) = D_{\mathrm{KL}}(p\|q_\lambda) + \lambda\big(\mathbb{E}_p f - s_0 - \Delta\big) \tag{3}$$

separates deviation from the target from an increase beyond the requested mean. Both terms on the right are nonnegative for score-feasible p. Hence equality with the benchmark requires $p = q_{\lambda_\Delta}$.

For an event E with $\pi_0 = p_0(E) \in (0,1)$, take $f = \mathbf{1}_E$ and $\pi = \pi_0 + \Delta < 1$. At a fixed event probability π, the target preserves both baseline conditional laws:

$$q(i) = \begin{cases} \frac{\pi}{\pi_0} p_0(i), & i \in E, \\ \frac{1-\pi}{1-\pi_0} p_0(i), & i \notin E. \end{cases}$$

Indeed, the chain rule gives

$$\begin{aligned} D_{\mathrm{KL}}(p\|p_0) = \mathrm{kl}(\pi\|\pi_0) &+ \pi D_{\mathrm{KL}}(p(\cdot \mid E)\|p_0(\cdot \mid E)) \\ &+ (1-\pi) D_{\mathrm{KL}}(p(\cdot \mid E^c)\|p_0(\cdot \mid E^c)). \end{aligned} \tag{4}$$

The first term is $\mathcal{I}(\Delta)$; the remaining terms measure redistribution within the groups. Hence conditional preservation is selected by KL minimization rather than imposed as a constraint on steering. KL supplies the distributional notion of distortion used here; it does not identify which changes in generated text are semantically undesirable. At fixed π, the chain rule for $D_{\mathrm{KL}}(p_0\|p)$ gives the same conditional-preserving optimizer. Under the inequality constraint $p(E) \geq \pi_0 + \Delta$, both orientations attain their optima at the boundary. Thus the four lexical indicator targets studied below do not depend on the KL orientation. For general real-valued scores, minimizing baseline-to-steered KL instead gives a rational tilt (Appendix A).

2.2 Activation model and reachable distributions

We now specify the intervention class and ask whether it can realize the minimum-KL target. Let $h_0 \in \mathbb{R}^d$ be the baseline activation immediately before RMSNorm and the vocabulary readout, and let $W \in \mathbb{R}^{m\times d}$ be the effective readout, with $d \geq 1$. For $\varepsilon > 0$ and intervention radius $r \geq 0$, define

$$N_\varepsilon(h) = \frac{h}{\sqrt{\|h\|^2 + \varepsilon}}, \qquad p_h = \mathrm{softmax}(WN_\varepsilon(h)), \qquad \mathcal{D}_r = \{h : \|h - h_0\| \leq r\}. \tag{5}$$

In this model, $p_0 = p_{h_0}$ is the baseline distribution defined above, and $h - h_0$ is the intervention. Fixed RMSNorm gains and implementation scales are absorbed into W and ε (Appendix H). All vector norms are Euclidean; matrix norms are the induced operator norms. The least KL cost attainable within the budget is

$$\mathcal{K}_r(\Delta) = \inf_{h\in\mathcal{D}_r:\, \mathbb{E}_{p_h} f \geq s_0 + \Delta} D_{\mathrm{KL}}(p_h\|p_0). \tag{6}$$

The difference from $\mathcal{I}(\Delta)$ is the *excess KL*; an infeasible infimum is $+\infty$. The benchmark identifies the least-distorted output law satisfying the score constraint. The activation map and budget determine whether that law is reachable. A score-feasible intervention, even one with small displacement, need not realize the unrestricted minimizer.

3 Can activation steering realize the benchmark?

The distributional benchmark fixes the target law. Exact attainment now requires an activation preimage within the budget. We resolve this inverse problem in the order imposed by the model: the readout determines compatible logit directions, normalization determines valid states, and the budget limits the displacement needed to reach them.

3.1 The target ray and the activation image

Softmax ignores common logit offsets. Let $P = I_m - \mathbf{1}\mathbf{1}^\top/m$ be the orthogonal projector onto $\mathbf{1}^\perp$, and define

$$A = PW, \quad b = Pf \neq 0, \quad x(h) = AN_\varepsilon(h), \quad x_0 = x(h_0), \quad t_\lambda = x_0 + \lambda b. \tag{7}$$

The exponential tilt in (2) satisfies $q_\lambda = \mathrm{softmax}(t_\lambda)$, so its centered logits form a straight ray. Centering is necessary because two logit vectors represent the same distribution exactly when they differ by a multiple of $\mathbf{1}$. It identifies Pf, rather than f, as the readout direction relevant to the target. By comparison, the intervention class produces

$$\mathcal{X}_r = x(\mathcal{D}_r), \qquad d_\lambda = \mathrm{dist}(t_\lambda, \mathcal{X}_r), \qquad C_\varepsilon(\lambda) = \inf_{x(h)=t_\lambda} \|h - h_0\|. \tag{8}$$

We set $C_\varepsilon = +\infty$ when the target has no activation preimage. The distance d_λ measures separation in output-logit space, whereas C_ε is the minimum activation displacement required for exact realization. A target may have a preimage in $\mathbb{R}^d$ but still have positive d_λ if every preimage lies outside the budget ball. If no preimage exists, increasing the budget cannot make the target attainable.

Let $A^\dagger$ be the Moore–Penrose pseudoinverse, $R = A^\dagger A$ the row-space projector, and $z_0 = N_\varepsilon(h_0)$. If $b \in \mathrm{range}(A)$, put

$$a_\lambda = Rz_0 + \lambda A^\dagger b, \qquad p_h = q_\lambda \iff RN_\varepsilon(h) = a_\lambda. \tag{9}$$

The first obstruction is $b \notin \mathrm{range}(A)$. For $\varepsilon > 0$, N_ε maps $\mathbb{R}^d$ bijectively onto the open unit ball. Its inverse is

$$g_\varepsilon(z) = \frac{\sqrt{\varepsilon}\, z}{\sqrt{1 - \|z\|^2}}, \qquad \|z\| < 1. \tag{10}$$

Once the range condition holds, a target preimage exists exactly when $\|a_\lambda\| < 1$. The budget determines whether a preimage is attainable by the intervention.

Theorem 1 (Exact attainment of the minimum-KL target)**.** *Under* (5)*, for an interior increase Δ let $\lambda = \lambda_\Delta > 0$. Then $\mathcal{K}_r(\Delta) = \mathcal{I}(\Delta)$ if and only if all three conditions hold:*

$$b \in \mathrm{range}(A), \qquad \|a_\lambda\| < 1, \qquad C_\varepsilon(\lambda) \le r. \tag{11}$$

Here a_λ is defined when the first condition holds, and Theorem 2 computes C_ε for every readout rank. For some $c_\lambda > 0$ depending on the bounded logit region,

$$\mathcal{K}_r(\Delta) - \mathcal{I}(\Delta) \ge c_\lambda d_\lambda^2. \tag{12}$$

If the mean constraint is feasible but any of the three conditions fails, excess KL is strictly positive.

By (3), equality with the benchmark is possible only at q_λ. Because the activation ball is compact, the distance d_λ is positive whenever this law is unattainable. On the bounded logit region, softmax KL is comparable to squared centered-logit distance:

$$c_\lambda \|x - t_\lambda\|^2 \le D_{\mathrm{KL}}(\mathrm{softmax}(x) \| q_\lambda) \le \tfrac{1}{4}\|x - t_\lambda\|^2 \qquad (x \in \mathcal{X}_r). \tag{13}$$

These bounds hold over a finite logit region; they are not a Taylor approximation at the baseline. Appendix A proves the theorem and specifies c_λ. This lower constant may be small for a large vocabulary or a highly concentrated target. Thus the bound proves positive excess without implying that a large measured range residual yields a large practical KL cost. Increasing r cannot repair a range or normalization failure. Once those conditions hold, the remaining obstruction is the minimum required norm. Even when the target is unattainable, another distribution in the activation image may satisfy the mean constraint.

3.2 Computing the minimum intervention norm

Write $\mathcal{N}_A = \ker A$, $H = \|h_0\|$, $h_\parallel = Rh_0$, $h_\perp = (I_d - R)h_0$, and $\beta = \|h_\perp\|$. For a fixed normalization scale, (9) determines the row-space component of the activation. Aligning the nullspace component with $h_\perp$ minimizes displacement and reduces the calculation to one dimension.

Theorem 2 (Minimum intervention norm)**.** *Fix $\lambda > 0$. If $b \notin \mathrm{range}(A)$, or if $b \in \mathrm{range}(A)$ and $\|a_\lambda\| \ge 1$, then $C_\varepsilon(\lambda) = +\infty$. Otherwise put $\delta_\lambda = 1 - \|a_\lambda\|^2 > 0$, $\gamma_\lambda = \langle a_\lambda, h_\parallel \rangle$, and $\tau_{\min} = \sqrt{\varepsilon/\delta_\lambda}$. If $\mathcal{N}_A = \{0\}$, then*

$$C_\varepsilon(\lambda) = \|g_\varepsilon(a_\lambda) - h_0\|. \tag{14}$$

If $\dim \mathcal{N}_A \ge 1$, then

$$C_\varepsilon(\lambda)^2 = \min_{\tau \ge \tau_{\min}} \left\{ \tau^2 + H^2 - \varepsilon - 2\gamma_\lambda \tau - 2\beta\sqrt{\delta_\lambda \tau^2 - \varepsilon} \right\}. \tag{15}$$

The scalar objective is strictly convex and has a unique minimizer. For $\beta = 0$ it is $\max\{\tau_{\min}, \gamma_\lambda\}$; for $\beta > 0$ it is the unique interior solution of $\tau - \gamma_\lambda - \beta\delta_\lambda \tau / \sqrt{\delta_\lambda \tau^2 - \varepsilon} = 0$.

The scalar variable is $\tau = \sqrt{\|h\|^2 + \varepsilon}$. It gives $Rh = \tau a_\lambda$ and a nullspace-component length of $\sqrt{\delta_\lambda \tau^2 - \varepsilon}$. At fixed τ, every direction in $\ker A$ gives the same target logits. The direction closest to h_0 aligns with its baseline null component; if that component is zero, all directions of the required length tie. This choice yields (15) and recovers a realizing activation at the minimizing scale. Appendix B proves strict convexity and treats the endpoint cases. If A is injective, there is no nullspace component, and inverse normalization gives the unique realizing activation.

Let $\Pi = AA^\dagger$ project onto $\mathrm{range}(A)$ and $b_\perp = (I_m - \Pi)b$. Every attainable logit lies in $\mathrm{range}(A)$, so $d_\lambda \ge \lambda \|b_\perp\|$ and

$$\mathcal{K}_r(\Delta) - \mathcal{I}(\Delta) \ge c_\lambda \lambda^2 \|b_\perp\|^2. \tag{16}$$

The bound applies to every activation direction and budget. We therefore check the range obstruction first in pretrained readouts.

The remaining failures also yield separation. If b is in range but $\|a_\lambda\| \geq 1$, no finite activation can normalize to the required visible component. If $\|a_\lambda\| < 1$ but $C_\varepsilon(\lambda) > r$, realizing activations exist but lie outside the budget ball. In either case, t_λ lies outside the compact set $\mathcal{X}_r$, so $d_\lambda > 0$. Appendix B also bounds excess KL in terms of how far the minimum norm exceeds r. These statements concern attainment of the minimum-KL law, not score feasibility: another point in $\mathcal{X}_r$ may still give the required mean. The injective example in Section 5 makes this distinction explicit.

4 THE RANGE OBSTRUCTION IN PRETRAINED READOUTS

We first test the range condition in Theorem 1, since its failure excludes exact attainment at every intervention budget. For four lexical event indicators—digits, sentence-ending punctuation, affirmative words, and a positive-word lexicon—we project $b = Pf$ onto the centered-readout ranges of SmolLM2-135M, SmolLM2-360M (Ben Allal et al., 2025), and Qwen2.5-0.5B (Qwen Team, 2024). Table 1 reports the relative residual $\|b_\perp\|/\|b\|$. Appendix H specifies the token sets, numerical projection procedure, and rank checks.

Table 1: Relative out-of-range residuals $\|b_\perp\|/\|b\|$ for four lexical indicator scores. Rank is for the centered effective readout, at a relative singular-value cutoff of 10^{-6}; it is unchanged at 10^{-4} and 10^{-8}.

Checkpoint	m	Rank/d	Digits	Punct.	Affirm.	Positive
SmolLM2-135M	49,152	576/576	0.928	0.931	0.970	0.967
SmolLM2-360M	49,152	960/960	0.888	0.900	0.953	0.946
Qwen2.5-0.5B	151,936	896/896	0.963	0.919	0.980	0.979

All twelve residuals lie between 0.888 and 0.980, although the three readouts have numerical full column rank. Full column rank makes the map from normalized activations to centered logits injective. It does not make its d-dimensional range cover the $(m-1)$-dimensional centered vocabulary space. Each tested event-score direction has a large component outside the range. Hence $b \notin \mathrm{range}(A)$ for these stored final readouts, and Theorem 1 excludes every positive minimum-KL tilt for the lexical targets, regardless of radius or steering direction. Because both KL orientations give the same event-indicator target, this range-mismatch conclusion is independent of the orientation choice.

The test concerns exact equality with the target distribution at a fixed context. It does not rule out increasing an event probability or evaluate the performance of a particular steering vector. When the requested gain is feasible, (16) converts the residual into an excess-KL lower bound. Its constant, however, depends on the logit region and target tilt, so the residual alone does not establish a large practical KL cost. For directions that pass the range test, normalization and budget determine the remaining obstruction. We study their geometry next, then construct compatible scores to test their effect in a pretrained model.

5 GEOMETRY WHEN THE TARGET IS IN RANGE

Assume now that $b \in \mathrm{range}(A)$. The remaining question is how far the requested score can increase while its minimum-KL target stays attainable. The answer is governed by the same minimum-norm function: exact attainment holds precisely where $C_\varepsilon(\lambda) \leq r$. Normalization curves the activation preimage of the target ray, while a readout nullspace allows alternative preimages. These features determine whether the attainable gains form a single interval or can have gaps.

5.1 A NULLSPACE GIVES A SINGLE THRESHOLD

When $\ker A \neq \{0\}$, the normalized activation has a component invisible to the centered readout. This component can absorb norm slack without changing the logits, and alignment with the baseline preserves budget feasibility (Figure 3). Theorem 6 in Appendix B.3 gives an exact second-order cone representation under which $\mathcal{X}_r$ is compact and convex for $\varepsilon > 0$. Hence its intersection with the target ray is an interval.

Corollary 3 (A single threshold under a nullspace condition)**.** *Suppose* $\varepsilon > 0$ *and* $\dim \mathcal{N}_A \geq 1$. *Then* $\{\lambda \geq 0 : t_\lambda \in \mathcal{X}_r\} = [0, \lambda_c]$. *If* $r > 0$ *and* $b \in \mathrm{range}(A)$, *then* $0 < \lambda_c < \infty$. *Writing* $\Delta(\lambda) = \mathbb{E}_{q_\lambda} f - s_0$ *and* $\Delta_c = \Delta(\lambda_c)$, *zero excess holds exactly for* $0 \leq \Delta \leq \Delta_c$. *If* $b \notin \mathrm{range}(A)$, *then* $\lambda_c = 0$.

The threshold is the largest increase whose minimum-KL target can be realized. Larger increases may remain score-feasible but incur positive excess KL. The conic representation computes λ_c by maximizing λ subject to target membership (Boyd & Vandenberghe, 2004). The proof also places the baseline in the relative interior of $\mathcal{X}_r$ when $r > 0$ and b is in range. The nullspace does more than make the inverse map nonunique: it fills the slack between a relaxed norm inequality and the exact normalization equality without changing Az. Consequently, the target ray cannot leave and re-enter $\mathcal{X}_r$. This mechanism is related to hidden-convexity arguments (Huy et al., 2026).

5.2 Injective readouts can lose and recover attainability

For an injective centered readout, each normalized target has a unique activation preimage. As t_λ moves along a straight ray, its preimage under g_ε can leave and re-enter the intervention ball. Near the unit sphere, the inverse amplifies the radial component. The distance of this curved preimage path from h_0 therefore need not increase monotonically with the score gain. Figure 2 and Example 10 give a target that is unattainable at $\lambda = 1/2$ but attainable at $\lambda = 1$; another activation in the ball still achieves the intermediate score increase. Injectivity allows, but does not require, this loss and recovery.

In this construction, the chord from the baseline activation to the realizing activation at $\lambda = 1$ lies inside the ball. The mean score varies continuously along the chord, so some point attains the intermediate target mean. Its output distribution differs from the exponential tilt at $\lambda = 1/2$, whose unique activation preimage lies outside the ball. Equation (3) attributes the resulting positive excess to distributional deviation or score overshoot. The nonmonotonicity thus concerns exact optimality, even though the score can be steered.

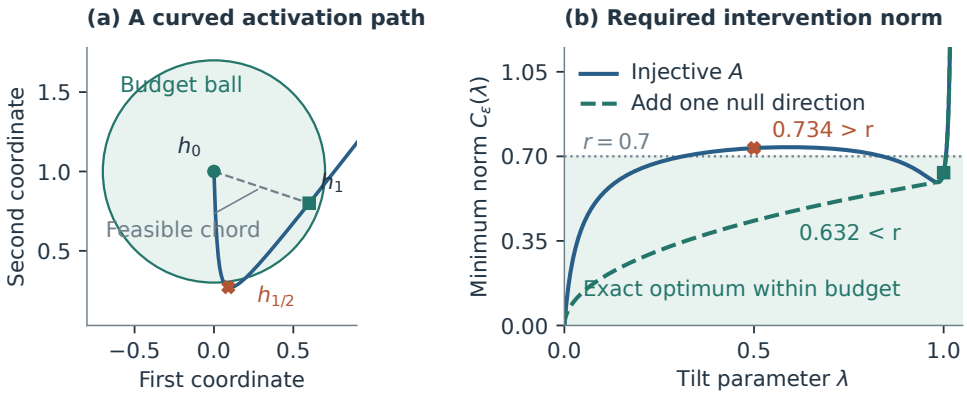

Figure 2: Loss and recovery of exact attainment (Example 10, $\varepsilon = 0.01$, $r = 0.7$). The straight logit target ray has a curved activation preimage: $h_{1/2}$ is outside the budget ball although h_1 is inside. The feasible chord reaches the intermediate mean without realizing its minimum-KL distribution. The minimum norm consequently rises above and later falls below r. The dashed nullspace comparison has one interval of attainable targets.

5.3 A controlled in-range diagnostic

The lexical scores fail the range test, so they cannot isolate the other two obstructions. For each of eight SmolLM2-135M prompts, we construct a score in $\mathrm{range}(A)$. We project a random normalized-activation direction orthogonally to the baseline normalized activation, map it through A, and rescale the resulting score to have baseline variance one. By construction, $b = Pf$ lies in the readout range.

Any subsequent failure to attain the minimum-KL target must therefore result from normalization or budget. We set $r = 0.1\|h_0\|$ and find the first positive λ satisfying $C_\varepsilon(\lambda) = r$.

The measured ratios $\varepsilon/\|h_0\|^2$ lie in $[5.05, 9.25] \times 10^{-9}$. The centered readouts are numerically injective, and (14) gives zero-excess endpoints between 5.82×10^{-5} and 1.87×10^{-4} in target-score units. Because the constructed direction is orthogonal to the baseline normalized activation, this crossing ends an interval rather than marking one point in a possible re-entry pattern. The interval is positive but narrow at the measured scale. Appendix H gives the construction; the supplementary material provides all prompt-level measurements. These scores differ from the lexical indicators in Section 4.

5.4 A SOLVABLE NORMALIZATION EXAMPLE

A four-output model isolates the normalization effect analytically. Let $(Y, Z) \in \{-1, 1\}^2$, score $f(Y, Z) = Y$, baseline $h_0 = (0, 1)$, $\alpha > 0$, and $\|u\| \le r$ with $0 < r < 1$. Define

$$p_u(Y, Z) \propto e^{a(u)Y + c(u)Z}, \qquad (a(u), c(u)) = \frac{\alpha(u_1, 1 + u_2)}{\sqrt{u_1^2 + (1 + u_2)^2 + \varepsilon}}. \tag{17}$$

The logit parameters separately control Y, the target score, and Z, whose distribution we seek to preserve. The minimum-KL tilt increases a while keeping $c = c_0$, leaving the Z marginal unchanged. For $\varepsilon = 0$, normalization constrains (a, c) to a circle, so increasing a necessarily changes c. Every positive feasible increase then incurs excess KL, although the score can rise. For $\varepsilon > 0$, the activation norm provides another degree of freedom, and exact attainment persists up to a positive threshold. Appendix E derives the cost curve and

$$\Delta_c \sim \alpha \sqrt{\varepsilon \frac{r(2 + r)}{(1 + r)^2}} \quad \text{as } \varepsilon \downarrow 0, \tag{18}$$

while the largest feasible increase tends to $\tanh(\alpha r) > 0$. Thus score feasibility persists even as the exact-attainment interval shrinks.

The example distinguishes a normalization obstruction from an inability to increase the score. At $\varepsilon = 0$, the score can increase substantially, but the normalized two-parameter image forces a change in the untargeted Z coordinate. Positive ε permits small radial adjustments and restores exact attainment near the baseline. The threshold nevertheless vanishes at the displayed rate as normalization approaches the unit-sphere limit. The constructed pretrained-score diagnostic examines the same scale dependence using measured final-state norms rather than the unit-norm coordinates of this example.

6 EXTENSIONS AND SCOPE

The exact criterion applies to the full Euclidean ball before the final normalized readout. It also bounds restricted steering methods and provides a reference geometry for interventions at earlier layers.

Prescribed directions and subspaces. Suppose $h = h_0 + Uu$, where $U \in \mathbb{R}^{d \times k}$ has orthonormal columns and $\|u\| \le r$. Let $\mathcal{K}_r^U$ be the restricted minimum KL cost, and set $C_{\varepsilon,U}(\lambda) = \inf_{x(h_0 + Uu) = t_\lambda} \|u\|$. The restricted intervention class is contained in $\mathcal{D}_r$, so the projection identity gives

$$\mathcal{K}_r^U(\Delta) \ge \mathcal{K}_r(\Delta), \qquad \mathcal{K}_r^U(\Delta) = \mathcal{I}(\Delta) \iff C_{\varepsilon,U}(\lambda_\Delta) \le r. \tag{19}$$

Thus nonattainability under the full ball implies nonattainability under any prescribed direction or subspace. When the full-ball test passes, reachability within the chosen subspace still requires a separate preimage check. A target attainable somewhere in the ball may be unattainable along a fixed steering vector. The case $k = 1$ covers a prescribed direction; allowing only positive strengths imposes an additional sign restriction. General affine restrictions need not preserve the convexity of the full-ball image. Appendix G.2 gives a one-dimensional scale calculation and states when the nullspace interval result extends to a subspace.

Bounded downstream errors. For an intervention at an earlier layer, write the centered suffix-logit map on $\mathcal{D}_r$ as $G(h) = x(h) + e(h)$, where $e(h) \in \mathbf{1}^{\perp}$. The relevant quantity is the error increment relative to the baseline:

$$e_0 = e(h_0), \qquad \xi(h) = e(h) - e_0, \qquad \sup_{h \in \mathcal{D}_r} \|e(h)\| \le \eta, \qquad \zeta = \sup_{h \in \mathcal{D}_r} \|\xi(h)\| \le 2\eta. \tag{20}$$

Let $p_h^e = \mathrm{softmax}(G(h))$ and $p_0^e = \mathrm{softmax}(x_0 + e_0)$. Define $\mathcal{K}_r^e, \mathcal{I}^e$ by the analogues of (6) and (1) relative to p_0^e. For an interior perturbed increase Δ, the recalibrated tilt parameter satisfies

$$\mathbb{E}_{\mathrm{softmax}(x_0+e_0+\lambda b)} f - \mathbb{E}_{\mathrm{softmax}(x_0+e_0)} f = \Delta, \qquad T = x_0 + e_0 + \lambda b. \tag{21}$$

The resulting target is $\mathrm{softmax}(T)$, the tilt relative to the perturbed baseline.

Theorem 4 (Separation under bounded errors)**.** *No continuity of e is required. At the parameter in* (21)*, let $d_\lambda = \mathrm{dist}(t_\lambda, \mathcal{X}_r)$ be the unperturbed logit distance from* (8)*. Then*

$$\mathcal{K}_r^e(\Delta) - \mathcal{I}^e(\Delta) \ge c_e(d_\lambda - \zeta)_+^2 \ge c_e(d_\lambda - 2\eta)_+^2, \tag{22}$$

where $u_+ = \max\{u, 0\}$ and $c_e > 0$ depends on the bounded logit region. An explicit choice is given in Appendix A, and the factor 2 is sharp (Appendix D).

The identity $G(h) - T = x(h) - t_\lambda + \xi(h)$ shows why the error increment ξ controls separation: the perturbed baseline enters both the output logits and the information-projection target. If the original separation exceeds the increment, the perturbed target remains separated. Conversely, near an attainable unperturbed target with strict budget slack, every sufficiently small continuous error admits an intervention calibrated to the exact mean with excess KL at most $O(\eta^2)$. Errors in the readout range can be canceled exactly. Both continuity and slack are needed: without them, a small error can prevent mean calibration or produce linear rather than quadratic excess. Appendix C.5 and Theorem 8 give the constants and counterexamples. Appendix G.3 derives a sufficient uniform-error certificate from derivative bounds.

The fixed-target statement retains the requested increase Δ but recomputes its information projection relative to the perturbed baseline. Locally, a right inverse of the readout removes the in-range component of the logit error, and a scalar adjustment restores the exact score mean. The remaining out-of-range logit error is $O(\eta)$; the finite-region KL comparison therefore bounds excess KL by $O(\eta^2)$. The original target has unused budget, so the corrected intervention remains in the ball. This argument requires an error bound over the entire ball used for the correction. Sampled errors alone do not give such a guarantee for an arbitrary Transformer suffix.

Sampled earlier-layer approximation. Appendix H reports sampled suffix errors after blocks 15 and 29 of SmolLM2-135M at several intervention radii. The samples assess the final-readout and affine Taylor references but do not certify the uniform bounds in (20). Local Fisher information also describes only tangent behavior. Appendix F gives finite-gain cases with identical local data but different attainment.

7 Conclusion

Many output distributions can satisfy the same score constraint, so increasing a target score does not determine how the remaining next-token probabilities should change. We use steered-to-baseline KL to define a unique minimum-KL target before choosing a steering method, and characterize exactly when an intervention before an RMS-normalized linear readout can realize this target within a finite budget. Exact attainment requires the target direction to lie in the centered-readout range, a valid normalized preimage to exist, and the minimum intervention norm not to exceed the budget. We also compute this minimum norm and derive lower bounds on excess KL when the score constraint is feasible but the target is unattainable. The resulting geometry further distinguishes readouts with a nontrivial nullspace, which yield a single attainment threshold, from injective readouts, which can lose and later recover exact attainment.

For four lexical event scores on the final readouts of three small pretrained models, all twelve target directions have large out-of-range residuals, ruling out exact realization of any positive minimum-KL tilt for these scores at the tested readouts. A controlled in-range diagnostic shows that normalization and budget can remain limiting even when the range condition holds. Overall, these results separate the ability to raise a score from the ability to realize the corresponding minimum-KL output distribution.

AI USE STATEMENT

We used OpenAI GPT-6 during the research and writing of this paper. In the attainability analysis, we used it to help check the reduction of the minimum intervention norm to a one-dimensional optimization over the normalization scale. The final formula and proof are given in Theorem 2 and Appendix B. We also used GPT-6 to help check the parameter choices and activation preimages in Example 10. This example shows that exact attainment can fail at an intermediate score increase and recover at a larger one, while the intermediate score mean remains feasible.

GPT-6 was also used to draft and edit parts of the paper, including the presentation of proofs, and to assist with literature search, diagnostic code, and figure preparation. The authors take full responsibility for all mathematical claims, constructions, constants, proofs, experimental results, and the content of the paper.

REPRODUCIBILITY STATEMENT

Sections 4–6 and Appendix H specify the pretrained checkpoints, prompts, token sets, intervention radii, and evaluation procedures. The supplementary material includes code and pinned checkpoint revisions for reproducing the reported diagnostics, tables, and figures.

REFERENCES

Seyedarmin Azizi, Erfan Baghaei Potraghloo, and Massoud Pedram. Activation Steering for Chain-of-Thought Compression. *arXiv preprint arXiv:2507.04742*, 2025. URL https://arxiv.org/abs/2507.04742.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, et al. SmolLM2: When Smol Goes Big—Data-Centric Training of a Small Language Model. *arXiv preprint arXiv:2502.02737*, 2025. URL https://arxiv.org/abs/2502.02737.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. URL https://web.stanford.edu/~boyd/cvxbook/.

L. E. J. Brouwer. Über Abbildung von Mannigfaltigkeiten. *Mathematische Annalen*, 71:97–115, 1911. doi: 10.1007/BF01456931.

Imre Csiszár. I-Divergence Geometry of Probability Distributions and Minimization Problems. *The Annals of Probability*, 3(1):146–158, 1975. doi: 10.1214/aop/1176996454.

Seyed Arshan Dalili, Ajay Narayanan Sridhar, Vijaykrishnan Narayanan, and Mehrdad Mahdavi. Conditional Optimal Bridge for Riemannian Activation Steering. *arXiv preprint arXiv:2607.10517*, 2026. URL https://arxiv.org/abs/2607.10517.

Bradley Efron. Defining the Curvature of a Statistical Problem (with Applications to Second Order Efficiency). *The Annals of Statistics*, 3(6):1189–1242, 1975. doi: 10.1214/aos/1176343282. With discussion and reply.

Nguyen Quang Huy, Nguyen Huy Hung, Tran Van Nghi, Hoang Ngoc Tuan, and Nguyen Van Tuyen. Hidden convexity of quadratic systems and its application to quadratic programming. *arXiv preprint arXiv:2601.13511*, 2026. URL https://arxiv.org/abs/2601.13511.

Aayush Mishra, Daniel Khashabi, and Anqi Liu. Steered LLM Activations are Non-Surjective. *arXiv preprint arXiv:2604.09839*, 2026. URL https://arxiv.org/abs/2604.09839.

Dung Viet Nguyen, Yen Pham, Hieu Vu, Lei Zhang, and Tan Nguyen. Activation Steering with a Feedback Controller. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fa5617c176e76fee83f3f9947fdf9f3f-Abstract-Conference.html.

Kiho Park, Todd Nief, Yo Joong Choe, and Victor Veitch. The Information Geometry of Softmax: Probing and Steering. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.15293.

Dario Picozzi. The information geometry of large language models is shared, learned, and controllable. *arXiv preprint arXiv:2609.11063*, 2026. URL https://arxiv.org/abs/2609.11063.

Qwen Team. Qwen2.5: A Party of Foundation Models!, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. URL https://aclanthology.org/2024.acl-long.828/.

Julian Skifstad, Xinyue Annie Yang, and Glen Chou. Local Linearity of LLMs Enables Activation Steering via Model-Based Linear Optimal Control. *arXiv preprint arXiv:2604.19018*, 2026. URL https://arxiv.org/abs/2604.19018.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering Language Models With Activation Engineering. *arXiv preprint arXiv:2308.10248*, 2023. URL https://arxiv.org/abs/2308.10248.

Sihan Wang, Jiayi Zhao, Qingyan Cao, Hongbo Yao, and Lin Shu. FishBack: Pullback Fisher Geometry for Optimal Activation Steering in Transformers. *arXiv preprint arXiv:2605.17231*, 2026. URL https://arxiv.org/abs/2605.17231.

Zejia You, Chunyuan Deng, and Hanjie Chen. Spherical Steering: Geometry-Aware Activation Rotation for Language Models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.08169.

Biao Zhang and Rico Sennrich. Root Mean Square Layer Normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

A INFORMATION PROJECTION AND KL COMPARISONS ON BOUNDED LOGITS

A.1 THE DISTRIBUTIONAL OPTIMUM AND MEAN OVERSHOOT

Throughout this appendix, p_0 is a strictly positive probability vector on a finite alphabet, f is non-constant, $s_0 = \mathbb{E}_{p_0} f$, and $\tilde{f} = f - s_0\mathbf{1}$. The log-moment-generating function $\Lambda(\lambda) = \log \mathbb{E}_{p_0} e^{\lambda \tilde{f}}$ satisfies

$$\Lambda'(\lambda) = \mathbb{E}_{q_\lambda} \tilde{f}, \qquad \Lambda''(\lambda) = \mathrm{Var}_{q_\lambda}(f) > 0,$$

where $q_\lambda(i) = p_0(i) e^{\lambda \tilde{f}_i - \Lambda(\lambda)}$. Since $\Lambda'(0) = 0$ and $\lim_{\lambda\to\infty} \Lambda'(\lambda) = \max_i f_i - s_0$, strict monotonicity gives a unique positive tilt parameter for every interior positive gain. Taking the expectation under p of

$$\log \frac{p(i)}{p_0(i)} = \log \frac{p(i)}{q_\lambda(i)} + \lambda \tilde{f}_i - \Lambda(\lambda)$$

and using the convention $0 \log 0 = 0$, we obtain

$$D_{\mathrm{KL}}(p\|p_0) = D_{\mathrm{KL}}(p\|q_\lambda) + \lambda \mathbb{E}_p \tilde{f} - \Lambda(\lambda).$$

For a distribution p satisfying the mean constraint, set $\lambda = \lambda_\Delta$. The last two terms then equal $\mathcal{I}(\Delta)$ plus the nonnegative overshoot $\lambda(\mathbb{E}_p \tilde{f} - \Delta)$. The tilt is feasible and attains $\lambda\Delta - \Lambda(\lambda)$; equality in the constrained problem requires $p = q_\lambda$. This proves (2) and (3) as a finite-alphabet instance of information projection (Csiszár, 1975).

For an indicator score, the KL chain rule in (4) gives the conditional interpretation. A conditional term with zero weight is defined to be zero. At $\pi = \pi_0 + \Delta$, the Bernoulli term equals $\mathcal{I}(\Delta)$, and the other terms quantify changes in the two conditional distributions. For a general real-valued f, excess KL remains the difference between the two optimal costs. It need not equal the divergence of a separately specified non-target attribute.

Reversing the KL orientation. This comparison concerns the distribution-space benchmark; the rest of the paper uses the steered-to-baseline orientation. For $p_0(i) > 0$ and an interior equality constraint $\mathbb{E}_q f = c$, the unique minimizer of $D_{\mathrm{KL}}(p_0\|q)$ has full support. The stationarity conditions of the Lagrangian give

$$q_i = \frac{p_{0,i}}{\alpha + \beta f_i}, \qquad \alpha + \beta f_i > 0,$$

where the multipliers normalize q and enforce the mean constraint. This rational tilt generally differs from (2). If $f = \mathbf{1}_E$, however, the denominator is constant within each of E and E^c. Equivalently, writing $\pi = q(E)$, the chain rule for $D_{\mathrm{KL}}(p_0\|q)$ gives

$$D_{\mathrm{KL}}(p_0\|q) = \mathrm{kl}(\pi_0\|\pi) + \pi_0 D_{\mathrm{KL}}(p_0(\cdot \mid E)\|q(\cdot \mid E)) + (1-\pi_0) D_{\mathrm{KL}}(p_0(\cdot \mid E^c)\|q(\cdot \mid E^c)).$$

At fixed π, both conditional terms vanish only when the baseline conditional laws are preserved. This yields the same q as in Section 2. For $\pi \geq \pi_0 + \Delta > \pi_0$, both Bernoulli divergences increase strictly with π, so both inequality-constrained optima occur at $\pi = \pi_0 + \Delta$. Hence the event-target ray and its range compatibility are independent of the KL orientation.

A.2 SOFTMAX FISHER INFORMATION ON A BOUNDED SEGMENT

To compare logit distance with KL, define the log-partition function $L(x) = \log \sum_i e^{x_i}$ and write $p_x = \mathrm{softmax}(x)$. The Hessian of L is the Fisher information matrix in logit coordinates:

$$F(x) = \nabla^2 L(x) = \mathrm{diag}(p_x) - p_x p_x^\top. \tag{23}$$

Here $\mathrm{osc}(v) = \max_i v_i - \min_i v_i$.

Lemma 5. *Let $x, t \in \mathbf{1}^\perp$ and suppose $\max\{\mathrm{osc}(x), \mathrm{osc}(t)\} \leq M$. Then*

$$\frac{e^{-M}}{2m}\|x - t\|^2 \leq D_{\mathrm{KL}}(\mathrm{softmax}(x)\|\,\mathrm{softmax}(t)) \leq \frac{1}{4}\|x - t\|^2. \tag{24}$$

The upper inequality holds without an oscillation bound.

Proof. At a vector u with $\mathrm{osc}(u) \le M$, each softmax coordinate is at least $p_* = e^{-M}/m$. For centered v,

$$v^\top F(u)v = \min_{c\in\mathbb{R}} \sum_i p_{u,i}(v_i - c)^2 \ge p_* \min_{c\in\mathbb{R}} \sum_i (v_i - c)^2 = p_*\|v\|^2,$$
$$v^\top F(u)v = \mathrm{Var}_{p_u}(v) \le \tfrac{1}{4}\,\mathrm{osc}(v)^2 \le \tfrac{1}{2}\|v\|^2.$$

For the variance upper bound, a variable in $[a, b]$ has variance at most $(b-a)^2/4$. To see this, expand $\mathbb{E}[(X - a)(b - X)] \ge 0$ and optimize the resulting quadratic in $\mathbb{E}X$. The final inequality follows from $|v_i - v_j| \le \sqrt{2}\|v\|$.

The Bregman identity transfers these covariance bounds to KL:

$$D_{\mathrm{KL}}(p_x\|p_t) = L(t) - L(x) - \nabla L(x)^\top (t - x) = \int_0^1 (1 - s)(t - x)^\top F(x + s(t - x))(t - x)\, ds.$$

Oscillation is convex, so every point on the segment has oscillation at most M. Integrating the covariance bounds and using $\int_0^1 (1 - s)ds = 1/2$ proves the claim. □

An admissible comparison constant is obtained as follows. With $H = \|h_0\|$ and W_i the ith row of W, set

$$\rho_r = \frac{H + r}{\sqrt{(H + r)^2 + \varepsilon}}, \quad D_W = \max_{i,j} \|W_i - W_j\|, \quad c_\lambda = \frac{1}{2m} e^{-\max\{\rho_r D_W, \mathrm{osc}(t_\lambda)\}}.$$

Every $h \in \mathcal{D}_r$ satisfies $\|h\| \le H + r$ and therefore $\|N_\varepsilon(h)\| \le \rho_r$. Centering preserves pairwise logit differences, giving

$$|x_i(h) - x_j(h)| = |(W_i - W_j)N_\varepsilon(h)| \le D_W \rho_r.$$

Applying Lemma 5 with $M = \max\{\rho_r D_W, \mathrm{osc}(t_\lambda)\}$ proves (13). Minimization over the activation ball also gives the unconstrained comparison

$$c_\lambda d_\lambda^2 \le \min_{h\in\mathcal{D}_r} D_{\mathrm{KL}}(p_h\|q_\lambda) \le d_\lambda^2/4.$$

An activation that minimizes this unconstrained KL need not satisfy the mean constraint in (6).

A.3 Proof of Theorem 1

Continuity of x and compactness of $\mathcal{D}_r$ imply that both $\mathcal{X}_r$ and the output image are compact. The mean-feasible activation set is closed in $\mathcal{D}_r$; thus $\mathcal{K}_r$ attains its minimum whenever the constraint is feasible. If this minimum equals $\mathcal{I}(\Delta)$, both nonnegative terms in (3) vanish, and the minimizing distribution is q_λ. The corresponding intervention has norm at most r. Conversely, suppose the target preimage is nonempty. It is closed, and its distance to h_0 attains a minimum: choose $\bar{h}$ in the preimage and restrict to $\|h - h_0\| \le \|\bar{h} - h_0\|$. This restricted set is nonempty and compact and has the same infimum. Therefore $C_\varepsilon(\lambda) \le r$ gives an activation in $\mathcal{D}_r$ that produces q_λ and attains $\mathcal{I}(\Delta)$.

Applying the KL lower bound for bounded logits and (3) to every feasible activation gives $\mathcal{K}_r(\Delta) - \mathcal{I}(\Delta) \ge c_\lambda d_\lambda^2$. If the attainability criterion fails, $t_\lambda \notin \mathcal{X}_r$. Compactness of $\mathcal{X}_r$ then implies $d_\lambda > 0$. If the target mean is infeasible, $\mathcal{K}_r = +\infty$ and the lower bound still holds. At $\Delta = 0$, the baseline activation h_0 has zero cost; threshold statements at zero target increase use this convention.

A.4 The comparison constant under logit perturbations

For the perturbed target T in (21), one admissible constant in Theorem 4 is

$$c_e = (2m)^{-1} \exp[-\max\{\rho_r D_W + \sqrt{2}\eta, \mathrm{osc}(T)\}].$$

Since $\mathrm{osc}(e(h)) \le \sqrt{2}\eta$, the perturbed image has logit oscillation at most $\rho_r D_W + \sqrt{2}\eta$. Lemma 5, applied to $G(h)$ and T, gives this constant. Combining the lemma with the increment identity and the perturbed information projection proves (22).

B TARGET PREIMAGES, CONIC GEOMETRY, AND THRESHOLD BEHAVIOR

B.1 REMOVING THE COMMON LOGIT OFFSET AND MINIMIZING THE INTERVENTION NORM

For logit vectors u, v, equality of their strictly positive softmax probabilities implies $u_i - v_i = L(u) - L(v)$ for every i. Thus two logit vectors give the same softmax distribution exactly when their difference is a multiple of $\mathbf{1}$. Applying this fact to $WN_\varepsilon(h)$ and $WN_\varepsilon(h_0) + \lambda f$ yields the centered-logit equation $AN_\varepsilon(h) = x_0 + \lambda b$.

Since $x_0 \in \mathrm{range}(A)$, a positive tilt requires $b \in \mathrm{range}(A)$. Under this condition, applying $A^\dagger$ gives (9); multiplication by A proves the converse. The map N_ε is a bijection onto the open unit ball. Solving $z = h/\sqrt{\|h\|^2 + \varepsilon}$ for the norm gives $\|h\|^2 = \varepsilon\|z\|^2/(1 - \|z\|^2)$ and the inverse (10). Consequently, $\|a_\lambda\| \geq 1$ rules out a preimage, whereas $\|a_\lambda\| < 1$ permits one by taking $z = a_\lambda$. If $A = 0$, the range condition fails because $b \neq 0$.

We now complete the proof of Theorem 2.

Proof. Equation (7) gives the range condition. The norm condition follows from the image of N_ε, which is the open unit ball. If $\mathcal{N}_A = \{0\}$, the preimage fixes the full normalized activation, and (10) yields (14). Otherwise, set $\tau = \sqrt{\|h\|^2 + \varepsilon}$. Equation (9) fixes $Rh = \tau a_\lambda$ and the length of the remaining component at $\sqrt{\delta_\lambda \tau^2 - \varepsilon}$. For each τ, alignment with $h_\perp$ minimizes the distance to h_0; if $\beta = 0$, any unit direction in $\mathcal{N}_A$ is equivalent. Expanding the squared distance gives (15). In the interior, its second derivative is

$$2 + \frac{2\beta\delta_\lambda\varepsilon}{(\delta_\lambda\tau^2 - \varepsilon)^{3/2}} > 0.$$

The objective diverges with τ. If $\beta > 0$, its derivative tends to $-\infty$ at the left endpoint and to $+\infty$ at infinity. If $\beta = 0$, the objective is a constrained quadratic. □

Continuity at $\tau_{\min}$ includes the endpoint in the minimization. A unit null vector exists exactly when $\dim \mathcal{N}_A \geq 1$; the injective case has no nullspace component.

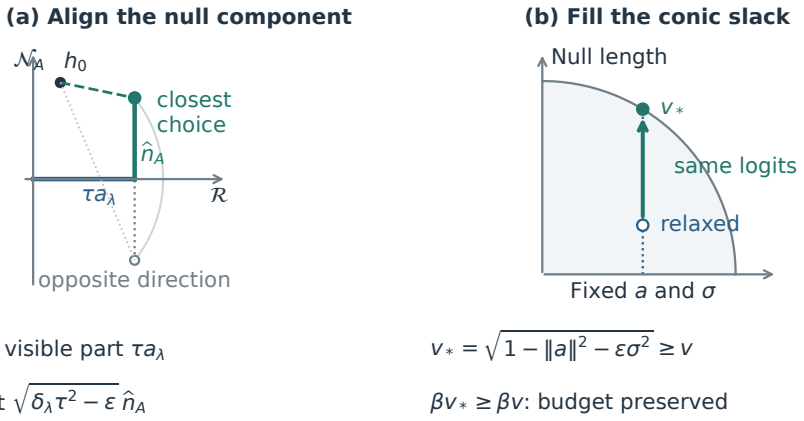

Figure 3: Two uses of a nonzero centered-readout nullspace. (a) At fixed normalization scale, the target fixes the readout-visible component; aligning the free nullspace component with the baseline null component minimizes intervention norm. (b) In the conic representation, nullspace freedom fills the remaining norm slack without changing the logits, enabling exact reconstruction. See Theorems 2 and 6 for the corresponding formulas.

B.2 EXPLICIT SEPARATION BOUNDS

Let $\Pi = AA^\dagger$ project onto $\mathcal{S} = \mathrm{range}(A)$, and set $b_\perp = (I_m - \Pi)b$. Since $\mathcal{X}_r \subseteq \mathcal{S}$ and $x_0 \in \mathcal{S}$, we have $d_\lambda \geq \lambda\|b_\perp\|$. Theorem 1 then gives the range-mismatch bound (16). When $b \in \mathcal{S}$ and

$C_\varepsilon(\lambda) > r$, excess KL has a positive lower bound controlled by $C_\varepsilon(\lambda) - r$. To obtain an explicit certificate, let $\sigma_+(A)$ be the smallest positive singular value and define

$$\alpha_r = (1-\rho_r)/2, \qquad \bar{\rho}_r = (1+\rho_r)/2, \qquad L_r = \frac{\sqrt{\varepsilon}}{(1-\bar{\rho}_r^2)^{3/2}}.$$

Then

$$\mathcal{K}_r(\Delta) - \mathcal{I}(\Delta) \geq c_\lambda \sigma_+(A)^2 \min\{\alpha_r, (C_\varepsilon(\lambda) - r)/L_r\}^2. \tag{25}$$

If $C_\varepsilon(\lambda) = +\infty$, the minimum is interpreted as α_r.

To prove (25), fix a feasible h, set $z = N_\varepsilon(h)$, and write $\delta_z = \|Rz - a_\lambda\|$. If $\delta_z < \alpha_r$, replace only the row-space component to obtain $z' = a_\lambda + (I_d - R)z$ with $\|z'\| \leq \rho_r + \delta_z < \bar{\rho}_r$. Differentiating (10) yields

$$Dg_\varepsilon(z) = \sqrt{\varepsilon}\left[(1-\|z\|^2)^{-1/2} I_d + (1-\|z\|^2)^{-3/2} z z^\top\right]. \tag{26}$$

Its operator norm on $\|z\| \leq \bar{\rho}_r$ is at most L_r. Thus $h' = g_\varepsilon(z')$ realizes the target with intervention norm at most $r + L_r\delta_z$. If $\delta_z < (C_\varepsilon(\lambda) - r)/L_r$, this contradicts the definition of the minimum intervention norm. If $C_\varepsilon(\lambda) = +\infty$, it contradicts the absence of a target preimage. Hence every feasible activation satisfies $\delta_z \geq \min\{\alpha_r, (C_\varepsilon(\lambda) - r)/L_r\}$. On the row space, $\|A(Rz - a_\lambda)\| \geq \sigma_+(A)\delta_z$. Lemma 5 and (3) then prove the bound.

B.3 Conic representation and exact-target optimality

Let $\mathcal{C}_r = \{RN_\varepsilon(h) : h \in \mathcal{D}_r\} \subseteq \mathcal{R}$. The auxiliary variables are the reciprocal normalization scale $\sigma > 0$ and the nonnegative normalized nullspace length v.

Theorem 6 (Exact conic representation)**.** *Suppose $\varepsilon > 0$ and $\dim \mathcal{N}_A \geq 1$. For $a \in \mathcal{R}$, membership $a \in \mathcal{C}_r$ holds if and only if some $\sigma > 0, v \geq 0$ satisfy*

$$\|(a, \sqrt{\varepsilon}\,\sigma, v)\| \leq 1, \qquad \langle h_\parallel, a\rangle + \beta v - \frac{H^2 - \varepsilon - r^2}{2}\sigma \geq \frac{1}{2\sigma}. \tag{27}$$

These constraints have a second-order cone representation. In particular, $\mathcal{C}_r$ and $\mathcal{X}_r = A\mathcal{C}_r$ are compact convex sets.

Proof. For a feasible h, choose $\sigma = (\|h\|^2 + \varepsilon)^{-1/2}$, $a = \sigma Rh$, and $v = \sigma\|(I_d - R)h\|$. The first constraint holds with equality. Expanding $\|h - h_0\|^2 \leq r^2$ and applying Cauchy–Schwarz to the nullspace term gives the second.

Conversely, for a feasible (a, σ, v), let $v_* = (1 - \|a\|^2 - \varepsilon\sigma^2)^{1/2} \geq v$. Choose $\widehat{n}_A = h_\perp/\beta$ when $\beta > 0$ and any unit $\widehat{n}_A \in \mathcal{N}_A$ otherwise, and set $h = (a + v_*\widehat{n}_A)/\sigma$. Then $\|h\|^2 + \varepsilon = \sigma^{-2}$ and $RN_\varepsilon(h) = a$. Because $\beta \geq 0$, replacing v by v_* does not weaken the second constraint. Expanding it gives $\|h - h_0\| \leq r$. Every point satisfying the conic constraints therefore corresponds to a feasible activation. Convexity follows from the second-order cone constraint and the epigraph of $1/(2\sigma)$, which also has a rotated second-order cone representation. Compactness follows because the set is the continuous image of $\mathcal{D}_r$. □

Under $\varepsilon > 0$ and $\dim \mathcal{N}_A \geq 1$, we can express (27) using standard cones. Let

$$u = \langle h_\parallel, a\rangle + \beta v - \tfrac{1}{2}(H^2 - \varepsilon - r^2)\sigma.$$

The second inequality becomes $2\sigma u \geq 1$ with $\sigma, u \geq 0$, a rotated second-order cone constraint on $(\sigma, u, 1)$. In particular, it enforces $\sigma > 0$. Every feasible point also satisfies

$$\frac{1}{\sqrt{(H+r)^2 + \varepsilon}} \leq \sigma \leq \frac{1}{\sqrt{(H-r)_+^2 + \varepsilon}}.$$

The reconstructed activation satisfies these bounds because $\sigma = (\|h\|^2 + \varepsilon)^{-1/2}$ and $(H - r)_+ \leq \|h\| \leq H + r$. Adding the bounds therefore leaves the feasible image unchanged. Expanding the budget constraint gives

$$\|h - h_0\|^2 = \sigma^{-2} - \varepsilon + H^2 - 2\sigma^{-1}(\langle h_\parallel, a\rangle + \beta v_*) \leq r^2.$$

Every feasible conic point therefore yields an activation within budget, proving that the representation is exact. Figure 3(b) illustrates the reconstruction.

For $r > 0$, the baseline lies in the relative interior of $\mathcal{X}_r$. Indeed, for sufficiently small $w \in \mathcal{S}$, continuity places $g_\varepsilon(z_0 + A^\dagger w)$ inside $\mathcal{D}_r$, and this activation maps to $x_0 + w$. If $b \in \mathcal{S}$, a nonzero initial segment of t_λ therefore lies in $\mathcal{X}_r$. Compactness gives a finite maximal endpoint, proving Corollary 3.

In this convex-image regime, every feasible target has a cost minimizer with exactly the requested score mean. If $x \in \mathcal{X}_r$ overshoots, the path $x_s = x_0 + s(x - x_0)$ remains feasible for $0 \le s \le 1$. Continuity of the mean gives a point $s_* < 1$ that attains the target. Along the same path,

$$\frac{d}{ds} D_{\mathrm{KL}}(p_{x_s} \| p_0) = s(x - x_0)^\top F(x_s)(x - x_0) > 0 \qquad (s > 0, x \neq x_0).$$

The point that attains the exact mean has lower cost than the overshooting endpoint. The argument uses convexity of the logit image; it does not assume monotonic behavior along arbitrary activation paths.

B.4 Explicit matching bounds past the threshold

Beyond the threshold, a supporting hyperplane gives a quadratic lower bound. A matching upper bound for the *mean-constrained* objective requires a feasible direction along which the score mean increases. Let $x_c = t_{\lambda_c}$, and suppose some $y \in \mathcal{X}_r$ satisfies $\gamma = b^\top F(x_c)(y - x_c) > 0$. Then there are explicit $\delta_* > 0$ and $c_-, c_+ > 0$ such that

$$c_- \delta^2 \le \mathcal{K}_r(\Delta(\lambda_c + \delta)) - \mathcal{I}(\Delta(\lambda_c + \delta)) \le c_+ \delta^2 \quad (0 < \delta \le \delta_*). \tag{28}$$

For the upper bound, we select a logit vector on the feasible chord from x_c to y that attains the target mean. Appendix B.3 gives the constants and proof. Without the chord condition, an upper bound based on a nearest logit point need not satisfy the mean constraint.

Proposition 7. *Assume $r > 0$, $b \in \mathcal{S}$, $\dim \mathcal{N}_A \ge 1$, and let $x_c = t_{\lambda_c}$. There is a unit supporting normal $n \in \mathcal{S}$ at x_c with $\nu = \langle n, b \rangle > 0$. For every $\delta > 0$,*

$$\mathcal{K}_r(\Delta(\lambda_c + \delta)) - \mathcal{I}(\Delta(\lambda_c + \delta)) \ge c_{\lambda_c + \delta} \nu^2 \delta^2. \tag{29}$$

Suppose additionally that $y \in \mathcal{X}_r$, $w_c = y - x_c$, and $\gamma = b^\top F(x_c) w_c > 0$. Then there are explicit $\delta_*, c_+ > 0$ such that, for $0 < \delta \le \delta_*$, the left side of (29) *is at most $c_+ \delta^2$. One admissible choice is given in the proof.*

Proof. The endpoint x_c lies on the relative boundary of the compact convex set $\mathcal{X}_r \subset \mathcal{S}$, so a supporting normal exists in $\mathcal{S}$. Because x_0 is a relative interior point, every nonzero supporting normal satisfies $\langle n, x_c - x_0 \rangle > 0$. Dividing by $\lambda_c > 0$ gives $\nu > 0$. For every $x \in \mathcal{X}_r$,

$$\| t_{\lambda_c + \delta} - x \| \ge \langle n, t_{\lambda_c + \delta} - x \rangle \ge \nu \delta.$$

Theorem 1 gives (29).

For the upper bound, we calibrate a point on the feasible chord. The following constants are admissible:

$$R_f = \mathrm{osc}(f), \quad R_w = \mathrm{osc}(w_c), \quad V_f = R_f^2/4,$$
$$s_{\max} = \min\{1, \gamma/(2R_f R_w^2)\}, \quad \delta_* = \frac{\gamma s_{\max}}{2V_f},$$
$$C_u = \|b\| + \frac{2V_f \|w_c\|}{\gamma}, \quad c_+ = C_u^2/4.$$

Define $B_c(s) = \mu(x_c + s w_c)$. Along this exponential family,

$$B_c'(0) = \gamma, \qquad B_c''(s) = \mathbb{E}_{p_{x_c + s w_c}}[(f - \mathbb{E} f)(w_c - \mathbb{E} w_c)^2].$$

Thus $|B_c''(s)| \le R_f R_w^2$, and the choice of $s_{\max}$ gives $B_c'(s) \ge \gamma/2$ on $[0, s_{\max}]$. Along the target ray, the mean derivative is $\mathrm{Var}(f) \le V_f$, so the requested increment from λ_c to $\lambda_c + \delta$ is at most

$V_f\delta$. For $\delta \le \delta_*$, a point on the feasible chord attains the exact target with $0 < s \le 2V_f\delta/\gamma \le s_{\max}$. Its distance from the target point is at most

$$\|sw_c - \delta b\| \le C_u\delta.$$

At this crossing, the overshoot term in (3) is zero. Lemma 5 gives the upper bound. The positive continuous coefficient c_λ has a positive minimum on the compact interval $[\lambda_c, \lambda_c + \delta_*]$, yielding (28) with fixed constants. □

The chord condition does not require a product structure in the output distribution. For

$$A = W = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 0 \end{pmatrix}, \quad h_0 = 0, \quad \varepsilon = 1, \quad r = 1/\sqrt{3}, \quad f = (1, 0, -1)^\top,$$

the projected normalized image is the disk of radius $1/2$ in the first two coordinates, and $\lambda_c = 1/2$. Write the coordinates as (u, v), with logits $(u, v, -u - v)$. At $(u, v) = (1/2, 0)$, the derivative of the score mean in the v direction is $q_3 - (q_1 - q_3)(q_2 - q_3) > 0$, where $(q_1, q_2, q_3) = \mathrm{softmax}(1/2, 0, -1/2)$. Multiplication by $(e^{1/2} + 1 + e^{-1/2})^2$ gives $2e^{-1/2} + 2 - e^{1/2} > 0$. A sufficiently short positive tangential step, together with an inward displacement of quadratic order in the step length, stays in the disk and increases the score mean. It therefore provides a point y satisfying $\gamma > 0$.

C Perturbation bounds and exact calibration

Throughout this appendix, $\varepsilon > 0$, $b \neq 0$, and the target increase is interior. We use the perturbation quantities e_0, ξ, η, ζ from (20). Perturbations are centered, all norms are Euclidean, and $\mu(v) = \mathbb{E}_{\mathrm{softmax}(v)} f$.

C.1 Unused budget permits exact mean calibration

Suppose $C_\varepsilon(\lambda) < r$. Choose a target-realizing activation h_* with slack $\kappa = r - \|h_* - h_0\| > 0$, and set $z_* = N_\varepsilon(h_*)$, $\rho_* = \|z_*\|$, and $\bar\rho_* = (1 + \rho_*)/2$. A positive attainable tilt requires $b \in \mathrm{range}(A)$ and $A \neq 0$, so $\sigma_+(A)$ is defined. Set

$$L_* = \frac{\sqrt{\varepsilon}}{(1 - \bar\rho_*^2)^{3/2}}, \qquad \ell = \sigma_+(A) \min\left\{\frac{1 - \rho_*}{2}, \frac{\kappa}{L_*}\right\} > 0. \tag{30}$$

For $w \in \mathcal{S} := \mathrm{range}(A)$ with $\|w\| \le \ell$, the map

$$h_{\mathrm{loc}}(w) = g_\varepsilon(z_* + A^\dagger w) \quad \text{satisfies} \quad h_{\mathrm{loc}}(w) \in \mathcal{D}_r, \qquad x(h_{\mathrm{loc}}(w)) = t_\lambda + w. \tag{31}$$

The local right inverse is defined on the closed ball of radius ℓ in $\mathcal{S}$. This radius keeps the normalized vector inside the unit ball and the activation within budget. Error compensation therefore does not require global convexity of $\mathcal{X}_r$.

Let $\Pi = AA^\dagger$ project onto $\mathcal{S}$. Uniformly over the image of this local right inverse, suppose $\|\Pi\xi\| \le \zeta_\parallel$ and $\|(I_m - \Pi)\xi\| \le \zeta_\perp$. The quantity $\zeta_\perp$ bounds the error perpendicular to the readout range. We cancel the parallel component through the intervention, then adjust along b to restore the target mean. Define $L_f = \|b\|/2$, a global Lipschitz constant for μ.

Theorem 8 (Compensating errors within the readout range)**.** *Suppose $e \circ h_{\mathrm{loc}}$ is continuous on the closed ball of radius ℓ. A sufficiently small parallel error can be canceled within the remaining budget, while the uncanceled perpendicular error contributes quadratically to excess KL. Choose $a > 0$ and set*

$$v = \frac{\|b\|^2}{m} e^{-\operatorname{osc}(T) - a\operatorname{osc}(b) - \sqrt{2}\zeta_\perp}.$$

If

$$L_f\zeta_\perp \le va, \qquad \zeta_\parallel + \frac{\|b\|L_f}{v}\zeta_\perp \le \ell, \tag{32}$$

there exists $h_e \in \mathcal{D}_r$ with $\mu(G(h_e)) = \mu(T)$, and

$$0 \le \mathcal{K}_r^e(\Delta) - \mathcal{I}^e(\Delta) \le \frac{\zeta_\perp^2}{4}\left[1 + \left(\frac{\|b\|L_f}{v}\right)^2\right]. \tag{33}$$

In particular, if ξ takes values in $\mathcal{S}$ on the image of the local right inverse and $\zeta_\parallel \le \ell$, the perturbed minimum-KL target distribution is attained exactly.

C.2 Lower bound and the local right inverse

For centered logits u, $\nabla\mu(u) = F(u)b$, where F is the covariance matrix from (23). Lemma 5 gives $\|F(u)\| \le 1/2$. Hence μ is globally L_f-Lipschitz with $L_f = \|b\|/2$. Also, $\operatorname{osc}(v) \le \sqrt{2}\|v\|$ for every vector v.

For every $h \in \mathcal{D}_r$, $\operatorname{osc}(G(h)) \le \rho_r D_W + \sqrt{2}\eta$. Every point on the segment from $G(h)$ to T has oscillation at most the maximum in the definition of c_e. Lemma 5, applied on this segment, therefore gives

$$D_{\mathrm{KL}}(p_h^e \| q_\lambda^e) \ge c_e \|G(h) - T\|^2 \ge c_e (d_\lambda - \zeta)_+^2.$$

The analogue of (3) relative to p_0^e holds for every feasible distribution, even when the perturbation is discontinuous. Taking an infimum proves Theorem 4 without requiring a compact perturbed image or attainment of the perturbed optimum.

To verify the local right inverse in (31), set $u = A^\dagger w$. Since $w \in \mathcal{S}$, we have $Au = w$ and

$$\|u\| \le \frac{\|w\|}{\sigma_+(A)} \le \min\{(1-\rho_*)/2, \kappa/L_*\}.$$

The segment from z_* to $z_* + u$ lies in the closed ball of radius $\bar{\rho}_* < 1$. The Jacobian formula (26) gives $\|Dg_\varepsilon(z)\| \le L_*$ throughout this segment. Therefore $\|h_{\mathrm{loc}}(w) - h_*\| \le \kappa$, which proves budget feasibility. Inverse normalization yields $x(h_{\mathrm{loc}}(w)) = Az_* + AA^\dagger w = t_\lambda + w$. The map h_{loc} is smooth on a neighborhood of the closed ball of radius ℓ.

C.3 A scalar calibration bound

We next calibrate the score mean along the score direction and identify the role of continuity.

Proposition 9 (Calibration along the score direction)**.** *Suppose the local right inverse* (31) *has radius ℓ, and $\|\xi\| \le \zeta$ on its image. Put $L_f = \|b\|/2$. Choose $a > 0$ with $a\|b\| \le \ell$, and define*

$$v_a = \frac{\|b\|^2}{m}\exp\big[-\max\{\operatorname{osc}(T - ab), \operatorname{osc}(T + ab)\}\big].$$

If $L_f\zeta \le v_a a$ and $e(h_{\mathrm{loc}}(sb))$ is continuous for $|s| \le a$, then

$$0 \le \mathcal{K}_r^e(\Delta) - \mathcal{I}^e(\Delta) \le \frac{1}{4}\left(1 + \frac{\|b\|L_f}{v_a}\right)^2 \zeta^2.$$

Without continuity, the upper bound

$$\mathcal{K}_r^e(\Delta) - \mathcal{I}^e(\Delta) \le \frac{1}{4}\left(1 + \frac{\|b\|L_f}{v_a}\right)^2 \zeta^2 + \lambda L_f\left(1 + \frac{\|b\|^2}{2v_a}\right)\zeta$$

still holds.

Proof. Along $T + sb$, the derivative of μ is $b^\top F(T + sb)b$. The lower covariance bound makes it at least v_a on $[-a, a]$. Set $s_e = L_f\zeta/v_a$. The perturbed mean at $h_{\mathrm{loc}}(s_e b)$ is at least $\mu(T)$, while the mean at $h_{\mathrm{loc}}(-s_e b)$ is at most $\mu(T)$. By continuity, an intermediate point has the exact target mean. Its logit distance from T is at most $\|b\|s_e + \zeta$. The KL upper bound and projection identity give the first assertion because the overshoot term vanishes at this point.

Without continuity, the positive endpoint remains feasible. Its excess mean is at most $(\|b\|^2/2)s_e + L_f\zeta$, since $b^\top Fb \le \|b\|^2/2$. Combining this estimate with the logit-distance bound in the projection identity proves the second assertion. □

C.4 Proof of compensation within the readout range

For Theorem 8, attainability at $\lambda > 0$ gives $b \in \mathcal{S}$. Let

$$\mathcal{B}_\perp = \{n \in \mathbf{1}^\perp \cap \mathcal{S}^\perp : \|n\| \le \zeta_\perp\}.$$

For $n \in \mathcal{B}_\perp$ and $|s| \le a$, the oscillation of $T + sb + n$ is at most $\operatorname{osc}(T) + a \operatorname{osc}(b) + \sqrt{2}\zeta_\perp$. Hence

$$\frac{d}{ds}\mu(T + sb + n) = b^\top F(T + sb + n)b \ge v > 0.$$

At $s = 0$, the mean difference $\mu(T + n) - \mu(T)$ has magnitude at most $L_f\|n\|$. The condition $L_f\zeta_\perp \le va$ therefore brackets a unique root $s(n) \in [-a, a]$ of $\mu(T + s(n)b + n) = \mu(T)$, with $|s(n)| \le L_f\|n\|/v$. The derivative is positive for every finite s, so the root is globally unique. It also varies continuously. To see this, note that the roots lie in a compact interval. Along any convergent sequence of values of n, every subsequential limit of roots solves the limiting equation and thus equals its unique root. Define the correction map

$$w \longmapsto -\Pi\xi(h_{\mathrm{loc}}(w)) + s\big((I_m - \Pi)\xi(h_{\mathrm{loc}}(w))\big)b \tag{34}$$

The map in (34) is continuous on the closed ball of radius ℓ in the finite-dimensional space $\mathcal{S}$, and its norm is at most $\zeta_\parallel + \|b\|L_f\zeta_\perp/v \le \ell$.

The Brouwer fixed-point theorem (Brouwer, 1911) provides a fixed point w_*. Set $h_e = h_{\mathrm{loc}}(w_*)$ and $n = (I_m - \Pi)\xi(h_e)$. The fixed-point equation gives

$$G(h_e) - T = w_* + \xi(h_e) = s(n)b + n.$$

By the defining equation for the scalar correction, this distribution has the exact target mean. Since $b \perp n$, we have

$$\|G(h_e) - T\|^2 = s(n)^2\|b\|^2 + \|n\|^2 \le \zeta_\perp^2\left[1 + (\|b\|L_f/v)^2\right].$$

Lemma 5 and the projection identity give the upper bound. If $\zeta_\perp = 0$, then $s(0) = 0$, and the fixed point realizes $G(h_e) = T$ exactly. This includes the case in which the orthogonal error set contains only zero.

C.5 Transfer at a fixed gain

For the same Δ, the unperturbed and perturbed tilt parameters generally differ. Denote them by λ_0 and λ_e. If the unperturbed score variance is at least $v_J > 0$ on a finite interval containing both parameters, then

$$|\lambda_e - \lambda_0| \le \frac{\|b\|\|e_0\|}{v_J}, \qquad d_{\lambda_e} \ge d_{\lambda_0} - \|b\||\lambda_e - \lambda_0|. \tag{35}$$

Let $\Delta_0(s) = \mu(x_0 + sb) - \mu(x_0)$ and $\Delta_e(s) = \mu(x_0 + e_0 + sb) - \mu(x_0 + e_0)$. The Lipschitz estimate gives

$$|\Delta_e(s) - \Delta_0(s)| \le 2L_f\|e_0\| = \|b\|\|e_0\|.$$

If $\Delta_0(\lambda_0) = \Delta_e(\lambda_e) = \Delta$ and $\Delta_0' \ge v_J$ between these roots, the mean value theorem gives the first bound in (35). Distance to a fixed set is 1-Lipschitz, which gives the second. If $h_{\mathrm{loc},0}$ is a local right inverse at t_{λ_0}, then

$$h_{\mathrm{loc},e}(w) = h_{\mathrm{loc},0}\big((\lambda_e - \lambda_0)b + w\big)$$

is a local right inverse at t_{λ_e} whenever $\|w\| \le \ell_0 - \|b\||\lambda_e - \lambda_0|$.

We next establish a uniform bound at a fixed gain without assuming that the perturbed root is known. Fix $\lambda_0 > 0$ and an unperturbed local right inverse of radius $\ell_0 > 0$. The following construction specifies an admissible $\eta_0 > 0$ and an $O(\eta^2)$ bound. Set

$$a_{\mathrm{root}} = \min\{\lambda_0/2, \ell_0/(4\|b\|)\}, \qquad M_0 = \max_{\pm} \operatorname{osc}\big(x_0 + (\lambda_0 \pm a_{\mathrm{root}})b\big),$$
$$v_0 = \|b\|^2 e^{-M_0}/m, \qquad a_{\mathrm{corr}} = \ell_0/(8\|b\|),$$
$$v_{\mathrm{corr}} = \frac{\|b\|^2}{m}\exp[-M_0 - a_{\mathrm{corr}}\operatorname{osc}(b) - 3\sqrt{2}],$$
$$\eta_0 = \min\{1, v_0 a_{\mathrm{root}}/\|b\|, \ell_0/8, v_{\mathrm{corr}}\ell_0/(8\|b\|^2)\}.$$

For any continuous centered perturbation on $\mathcal{D}_r$ with $\sup \|e\| \le \eta \le \eta_0$, the perturbed problem at the fixed gain $\Delta = \Delta_0(\lambda_0)$ satisfies

$$0 \le \mathcal{K}_r^e(\Delta) - \mathcal{I}^e(\Delta) \le \eta^2 \left[1 + \left(\frac{\|b\|^2}{2v_{\mathrm{corr}}}\right)^2\right]. \tag{36}$$

On $[\lambda_0 - a_{\mathrm{root}}, \lambda_0 + a_{\mathrm{root}}]$, the unperturbed derivative is at least v_0. Because $\|b\|\eta \le v_0 a_{\mathrm{root}}$, the endpoint values of Δ_e bracket Δ. The strictly positive derivative of Δ_e then gives a unique root λ_e in this interval. Thus the perturbed gain is interior, and the shifted local right inverse has radius at least $3\ell_0/4$.

With $a = a_{\mathrm{corr}}$ in Theorem 8, we have $\mathrm{osc}(T) \le M_0 + \sqrt{2}\eta$ and $\zeta_\parallel, \zeta_\perp \le 2\eta$. The lower variance constant in that theorem is therefore at least v_{corr}. Moreover,

$$L_f\zeta_\perp \le \|b\|\eta \le v_{\mathrm{corr}} a_{\mathrm{corr}}, \qquad \zeta_\parallel + \frac{\|b\|L_f}{v_{\mathrm{corr}}}\zeta_\perp \le 2\eta + \frac{\|b\|^2\eta}{v_{\mathrm{corr}}} \le \frac{\ell_0}{4} + \frac{\ell_0}{8} < \frac{3\ell_0}{4}.$$

These inequalities verify the hypotheses of the upper bound, so (33) yields (36). The constants depend on the unperturbed target and its budget slack. No derivative bound or modulus of continuity for e is required.

D Counterexamples and boundary cases

The following examples distinguish the effects of geometry, continuity, and budget slack. Every output set is finite, every readout matrix is centered, and each positive normalization constant is fixed unless stated otherwise.

D.1 Loss and recovery of exact attainability

Example 10 (Loss and recovery of attainability)**.** Let $W = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{smallmatrix}\right)$, $h_0 = (0,1)$, $\varepsilon = 1/100$, $r = 7/10$, and $f = (6,-2,-4)^\top/\sqrt{101}$. This centered readout is injective. The minimum-KL target distribution at $\lambda = 1$ is realized by $h_1 = (3/5, 4/5)$, with intervention norm $\sqrt{2/5} < r$. At $\lambda = 1/2$, its unique preimage is $h_{1/2} = (3,9)/(10\sqrt{11})$, and

$$\|h_{1/2} - h_0\|^2 = \frac{119}{110} - \frac{9}{5\sqrt{11}} > \frac{49}{100}.$$

The intermediate target mean is feasible by continuity along the segment $[h_0, h_1] \subset \mathcal{D}_r$. Hence its excess is strictly positive, although the larger minimum-KL target distribution has zero excess.

To verify Example 10, let $z_i = N_\varepsilon(h_i)$, with $h_0 = (0,1)$, $h_1 = (3/5, 4/5)$, and $\varepsilon = 1/100$. Both activations have norm one. Then

$$z_0 = (0, 10/\sqrt{101}), \qquad z_1 = (6/\sqrt{101}, 8/\sqrt{101}), \qquad b = W(z_1 - z_0) = (6,-2,-4)/\sqrt{101}.$$

The centered matrix W has rows $(1,0), (0,1), (-1,-1)$ and is injective. Thus the unique activation realizing the minimum-KL target distribution at parameter λ is $g_\varepsilon((1-\lambda)z_0 + \lambda z_1)$ whenever this normalized vector lies in the open unit ball. At $\lambda = 1$, its distance from h_0 is $\sqrt{2/5} < 7/10$. At $\lambda = 1/2$, it equals

$$h_{1/2} = \frac{(3,9)}{10\sqrt{11}}, \qquad \|h_{1/2} - h_0\|^2 = \frac{119}{110} - \frac{9}{5\sqrt{11}} > \frac{49}{100}.$$

The minimum-KL target is therefore unattainable at $1/2$ but attainable at 1. The intermediate mean remains feasible: it lies strictly between the endpoint means of the ideal family, and continuity gives an activation with that mean on $[h_0, h_1]$. Compactness then implies strictly positive excess at $1/2$. Appending a zero column to W and a zero coordinate to h_0 produces the nonzero-nullspace comparison in Figure 2 without changing the target ray.

D.2 THE FACTOR TWO IN THE PERTURBATION MARGIN

Let $v = (1, -1)/\sqrt{2}$, $W = f = v$, $h_0 = 0$, $\varepsilon = 1$, and $r = 1$. The activation is scalar; write $a(h) = h/\sqrt{1+h^2}$ and $a_r = 1/\sqrt{2}$. Choose the smooth error

$$e_\eta(h) = \eta(2h^2 - 1)v, \qquad |h| \le 1.$$

The error norm is at most η, while $e_0 = -\eta v$ and $e_\eta(1) = \eta v$. At $\lambda = a_r + 2\eta$, the unperturbed target is at distance $d_\lambda = 2\eta$ from the unperturbed interval $[-a_r, a_r]v$. The perturbed target is $T = e_0 + \lambda v = (a_r + \eta)v = G(1)$. Thus the perturbed target is attained exactly, with zero excess at the corresponding positive interior gain. Replacing 2η by $c\eta$ for any universal $c < 2$ would incorrectly give a positive lower bound in this example.

D.3 A SMOOTH ORTHOGONAL ERROR WITH QUADRATIC EXCESS

Use three outputs and orthonormal centered vectors

$$v = (1, -1, 0)/\sqrt{2}, \qquad n = (1, 1, -2)/\sqrt{6}.$$

Again set $W = f = v$, $h_0 = 0$, $\varepsilon = r = 1$. Fix $0 < \lambda < a_r$ and let $e_\eta(h) = \eta(a(h)/a_r)^2 n$. Its baseline error is zero. For every $a \in [-a_r, a_r]$,

$$\|av + \eta(a/a_r)^2 n - \lambda v\|^2 = (a - \lambda)^2 + \eta^2 a^4/a_r^4 \ge \min\{\lambda/2, \eta\lambda^2/(4a_r^2)\}^2.$$

Indeed, either $|a - \lambda| \ge \lambda/2$ or $a > \lambda/2$. For sufficiently small positive η, the displayed lower bound is a positive constant times η^2. The logits remain uniformly bounded, so Lemma 5 and the projection identity give an $\Omega(\eta^2)$ lower bound from the perturbed image. Because the unperturbed target has positive slack, Theorem 8 or Proposition 9 gives an $O(\eta^2)$ upper bound at the same gain. Hence excess KL is $\Theta(\eta^2)$. The bound in (22) vanishes because $d_\lambda = 0$; separation must instead be measured from the perturbed image.

D.4 A DISCONTINUOUS ERROR WITH LINEAR EXCESS

Return to the binary model above and fix $0 < \theta < a_r$. For $0 < \eta < \min\{\theta, a_r - \theta\}$, define

$$e_\eta(h) = \begin{cases} 0, & a(h) \le \theta - \eta, \\ -\eta v, & \theta - \eta < a(h) < \theta, \\ +\eta v, & a(h) \ge \theta. \end{cases}$$

The baseline error is zero. The attainable scalar logit coefficients omit $(\theta - \eta, \theta + \eta)$, so the smallest coefficient whose mean is at least that of $p_\theta = \mathrm{softmax}(\theta v)$ is $\theta + \eta$. For the binary family $p_s = \mathrm{softmax}(sv)$, the divergence $D_{\mathrm{KL}}(p_s \| p_0)$ increases strictly for $s > 0$ because its derivative is $s \operatorname{Var}_{p_s}(v) > 0$. Thus the perturbed optimum is $p_{\theta+\eta}$, with excess

$$D_{\mathrm{KL}}(p_{\theta+\eta} \| p_\theta) + \theta\big(\mathbb{E}_{p_{\theta+\eta}} v - \mathbb{E}_{p_\theta} v\big) = \theta \operatorname{Var}_{p_\theta}(v)\eta + O(\eta^2).$$

This excess is $\Theta(\eta)$ despite positive slack at the unperturbed target. A uniform error bound alone therefore cannot replace continuity in the upper-bound theorem.

D.5 LOSS OF FEASIBILITY WITHOUT BUDGET SLACK

In the same binary model, take the boundary target $\lambda = a_r$ and the smooth perturbation $e_\eta(h) = -\eta h v$. For $0 < \eta < 1/(2\sqrt{2})$,

$$\frac{d}{dh}(a(h) - \eta h) = (1 + h^2)^{-3/2} - \eta > 0 \quad \text{on } [-1, 1].$$

The largest attainable logit coefficient is $a_r - \eta$, below the target a_r. The baseline error is zero. Since the binary mean increases strictly with its logit coefficient, the requested gain is infeasible. Thus an arbitrarily small smooth error can destroy feasibility when the minimum intervention norm equals the budget.

D.6 Zero normalization constant and a nonclosed attainable image

When $\varepsilon = 0$, the activation $h = 0$ is excluded. Assume $H = \|h_0\| > 0$. A unit direction z is reachable if some $s > 0$ satisfies $\|sz - h_0\| \le r$. Minimizing squared distance along the ray gives the exact directional domains

$$\begin{cases} \|z\| = 1, & \langle h_0, z\rangle \ge \sqrt{H^2 - r^2}, & r < H, \\ \|z\| = 1, & \langle h_0, z\rangle > 0, & r = H, \\ \|z\| = 1, & & r > H. \end{cases} \tag{37}$$

For $r < H$, every reachable direction z has a positive projection of the baseline onto its ray. Minimizing at that projection proves necessity and sufficiency. For $r = H$, the inequality becomes $s^2 \le 2s\langle h_0, z\rangle$, which has a positive solution exactly when the inner product is positive. For $r > H$, every direction is feasible for sufficiently small $s > 0$.

If $\ker A \neq \{0\}$ and $r < H$, the projected directional domain in the row space is

$$\|a\| \le 1, \qquad \langle h_\parallel, a\rangle + \beta\sqrt{1 - \|a\|^2} \ge \sqrt{H^2 - r^2}.$$

For $r = H$, the last inequality is strict and has right side zero; for $r > H$, the domain is the row-space unit ball. These descriptions follow by maximizing the inner product over nullspace components of length $\sqrt{1 - \|a\|^2}$. The projected domain can be nonclosed when $r = H$.

To show how strict separation can fail when $r = H$, take $h_0 = (0, 1)$, $r = 1$, and a centered injective matrix with rows $(1, 1/2), (0, 1/4), (-1, -3/4)$. Let $f = W((1, 0) - h_0) = (1/2, -1/4, -1/4)$. The ideal parameter $\lambda = 1$ requires direction $(1, 0)$, which violates the strict inequality in (37). It is unattainable. Nevertheless, for $0 < \phi < \pi/2$, set

$$z_\phi = (\cos\phi, \sin\phi), \qquad h_\phi = \sin\phi\, z_\phi.$$

Then $\|h_\phi - h_0\| = \cos\phi < 1$, and the resulting distributions converge to the minimum-KL target as $\phi \downarrow 0$. They satisfy the mean constraint for all sufficiently small $\phi > 0$. Indeed, f is an affine function of the indicator of output 1, so it suffices to show that the probability of output 1 increases at $\phi = 0$. With $p = \mathrm{softmax}(W(1, 0)^\top)$ denoting the limiting distribution, the derivative is $p_1(\frac{1}{2} - \mathbb{E}_p W_{\cdot,2}) > 0$, because every other entry in the second column is below $1/2$. Thus the infimum of excess KL is zero despite nonattainment. The example shows why the compactness supplied by a positive normalization constant is essential to Theorem 1.

E An exactly solvable four-output model

We solve the four-output model in (17). Recall that $f(Y, Z) = Y$, $h_0 = (0, 1)$, $0 < r < 1$, and $\|u\| \le r$. Denote the two logit parameters by $a(u)$ and $c(u)$. The parameter c controls the non-target mean $\mathbb{E}Z = \tanh c$. At baseline, $a_0 = 0$ and $c_0 = \alpha/\sqrt{1 + \varepsilon}$. Independence of Y and Z gives

$$\begin{aligned} \mathbb{E}_{p_u} Y &= \tanh a(u), \\ D_{\mathrm{KL}}(p_u \| p_0) &= k(\tanh a(u)) + \psi(c(u) \| c_0), \\ k(\Delta) &= \Delta \operatorname{arctanh}\Delta + \tfrac{1}{2}\log(1 - \Delta^2), \\ \psi(c\|c_0) &= (c - c_0)\tanh c - \log\cosh c + \log\cosh c_0. \end{aligned} \tag{38}$$

The unrestricted cost is $\mathcal{I}(\Delta) = k(\Delta)$. Throughout this appendix, $0 \le \Delta < 1$ and $\theta_\Delta = \operatorname{arctanh}\Delta$, where $\operatorname{arctanh}$ is the inverse hyperbolic tangent. We also use $\operatorname{sech} c = 1/\cosh c$.

E.1 Zero normalization constant

If $\varepsilon = 0$, the restriction $r < 1$ keeps h away from zero and makes its second coordinate positive. Hence $a^2 + c^2 = \alpha^2$ and $c > 0$. Attaining an exact target requires $a = \theta_\Delta$, which fixes $c = \sqrt{\alpha^2 - \theta_\Delta^2}$. With $x = \theta_\Delta/\alpha$, the nearest activation on the required directional ray is

$$h = (x\sqrt{1 - x^2}, 1 - x^2), \qquad u = (x\sqrt{1 - x^2}, -x^2), \qquad \|u\| = x.$$

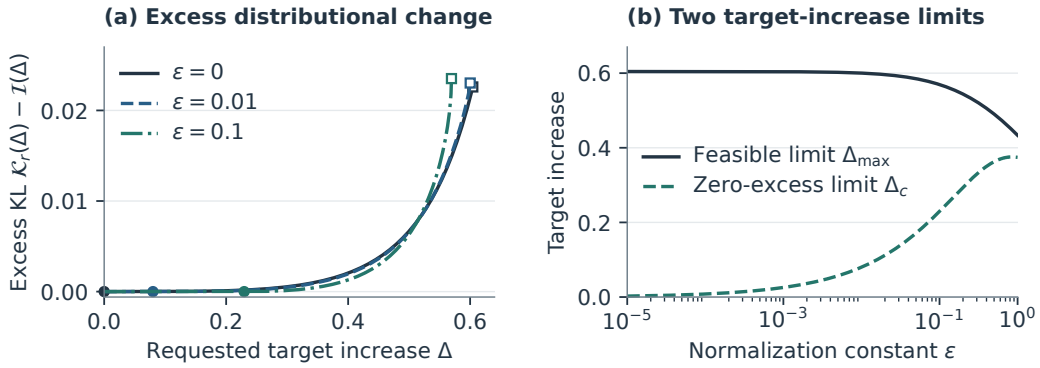

Figure 4: Exact attainment and target-increase feasibility in the four-output model of Section 5.4, with $\alpha = 1$ and $r = 0.7$. (a) Excess KL is measured relative to the baseline of each model. Filled circles mark the last zero-excess target increase; open squares mark the largest feasible increase. (b) As $\varepsilon \downarrow 0$, the zero-excess threshold Δ_c vanishes while the largest feasible increase approaches $\tanh(0.7) > 0$. The curves use the exact formulas below.

Thus the target is feasible exactly when $\Delta \le \tanh(\alpha r)$, and

$$\mathcal{K}_r(\Delta) = k(\Delta) + \psi\big(\sqrt{\alpha^2 - \theta_\Delta^2}\,\|\alpha\big). \tag{39}$$

Overshooting cannot lower the cost. The term $k(\tanh a)$ increases for $a \ge 0$, while increasing a decreases $c \le \alpha$, and $\partial_c \psi(c\|\alpha) = (c - \alpha)\operatorname{sech}^2 c \le 0$.

For fixed α, expansion at $\Delta = 0$ gives

$$\mathcal{K}_r(\Delta) - \mathcal{I}(\Delta) = \frac{\operatorname{sech}^2 \alpha}{8\alpha^2}\Delta^4 + O(\Delta^6).$$

More explicitly, with $c = \sqrt{\alpha^2 - \theta_\Delta^2}$,

$$\tfrac{1}{2}\operatorname{sech}^2 \alpha\, \frac{\theta_\Delta^4}{(\alpha + c)^2} \le \mathcal{K}_r(\Delta) - \mathcal{I}(\Delta) \le \tfrac{1}{2}\operatorname{sech}^2 c\, \frac{\theta_\Delta^4}{(\alpha + c)^2}.$$

These bounds follow from $\psi(c\|\alpha) = \int_c^\alpha (\alpha - s)\operatorname{sech}^2 s\, ds$ and $\alpha - c = \theta_\Delta^2/(\alpha + c)$. If interventions are restricted to the horizontal line $u = (v, 0)$, $|v| \le r$, the maximum score mean is $\tanh(\alpha r/\sqrt{1 + r^2})$, strictly below $\tanh(\alpha r)$. For an exact target, the restriction does not change the required output distribution but changes the intervention norm.

E.2 A POSITIVE NORMALIZATION CONSTANT

Now suppose $\varepsilon > 0$. Put

$$\begin{aligned} \chi_{\max} &= \frac{1 + \varepsilon + r^2 - \sqrt{(1 + \varepsilon + r^2)^2 - 4r^2}}{2}, & \chi &= (\theta_\Delta/\alpha)^2, \\ y_+(\chi) &= 1 - \chi + \sqrt{(1 - \chi)(r^2 - \chi) - \varepsilon\chi}, & & \\ c_+(\chi) &= \alpha\sqrt{1 - \chi}\,\frac{y_+(\chi)}{\sqrt{y_+(\chi)^2 + \varepsilon}}, & c^*(\chi) &= \min\{c_0, c_+(\chi)\}. \end{aligned} \tag{40}$$

The target is feasible if and only if $\chi \le \chi_{\max}$. In this regime,

$$\mathcal{K}_r(\Delta) = k(\Delta) + \psi(c^*(\chi)\|c_0). \tag{41}$$

To prove the formula, fix $a = \alpha\sqrt{\chi} \ge 0$ and set $y = 1 + u_2 > 0$. Solving $a(u) = \alpha\sqrt{\chi}$ for u_1 gives

$$u_1 = \sqrt{\chi}\sqrt{\frac{y^2 + \varepsilon}{1 - \chi}}, \qquad \|u\|^2 = \frac{(y - (1 - \chi))^2}{1 - \chi} + \chi + \frac{\varepsilon\chi}{1 - \chi}. \tag{42}$$

The squared intervention norm is minimized at $y = 1-\chi$, which is positive for $\chi < 1$. Its minimum is $\chi + \varepsilon\chi/(1-\chi)$, and the budget inequality gives $\chi \le \chi_{\max} < r^2 < 1$. The feasible interval for y contains $1-\chi$, has upper endpoint $y_+(\chi)$, and lies in $(0,\infty)$ because $r < 1$. The non-target logit parameter $c(y) = \alpha\sqrt{1-\chi}\, y/\sqrt{y^2+\varepsilon}$ increases strictly with y, and its value at $y = 1-\chi$ is at most c_0. Because $\psi(c\|c_0)$ is minimized at $c = c_0$, the optimal reachable value is $c^*(\chi)$.

The radicand in (40) is $r^2 - (1+\varepsilon+r^2)\chi + \chi^2$. Its derivative is $2\chi - (1+\varepsilon+r^2) < 0$ on $[0, \chi_{\max}]$. Hence y_+ and c_+ decrease strictly, and the optimized non-target distortion is nondecreasing in χ. The target-score cost increases strictly, so overshooting cannot improve (41).

An optimal activation can be constructed explicitly. If $c_+(\chi) \ge c_0$, choose

$$y = \sqrt{\frac{\varepsilon}{\varepsilon - (1+\varepsilon)\chi}};$$

Otherwise, choose $y = y_+(\chi)$. Equation (42) then determines u_1, with $u_2 = y - 1$. In the first case, the denominator is positive, and the chosen y is feasible because the increasing function $c(y)$ crosses c_0 between $1-\chi$ and $y_+(\chi)$.

The threshold for preserving $c = c_0$ and attaining the ideal cost is

$$\begin{aligned} y_r &= \frac{1+\sqrt{\varepsilon^2 + (1+\varepsilon)r^2}}{1+\varepsilon}, \\ \chi_c &= \frac{\varepsilon}{1+\varepsilon}(1 - y_r^{-2}), \qquad\qquad \Delta_c = \tanh(\alpha\sqrt{\chi_c}), \qquad (43) \\ \mathcal{K}_r(\Delta) = \mathcal{I}(\Delta) \quad &\iff \quad 0 \le \Delta \le \Delta_c. \end{aligned}$$

Preserving $c = c_0$ requires $u_1^2 = \varepsilon(y^2-1)$ and $y \ge 1$. The squared intervention norm is $(1+\varepsilon)y^2 - 2y + 1 - \varepsilon$, which increases strictly for $y \ge 1$. Equating it to r^2 gives y_r; solving for χ yields (43). The construction also proves $0 < \chi_c < \chi_{\max}$: a point with $c = c_0$ has $y \ge 1 > 1 - \chi_c$, whereas the minimum-norm point at $\chi_{\max}$ has $y = 1 - \chi_{\max}$. The function c_+ is smooth and has a strictly negative derivative at χ_c. Taylor expansion gives

$$\mathcal{K}_r(\Delta) - \mathcal{I}(\Delta) = \frac{\operatorname{sech}^2 c_0}{2}\left[\frac{d}{d\Delta} c_+(\chi(\Delta))\Big|_{\Delta_c}\right]^2 (\Delta - \Delta_c)^2 + O((\Delta - \Delta_c)^3)$$

as $\Delta \downarrow \Delta_c$ from above, with strictly positive leading coefficient.

For fixed α, r and $\varepsilon \downarrow 0$,

$$\Delta_c \sim \alpha\sqrt{\varepsilon\frac{r(2+r)}{(1+r)^2}}, \qquad \Delta_{\max} = \tanh(\alpha\sqrt{\chi_{\max}}) \longrightarrow \tanh(\alpha r) > 0.$$

The first asymptotic proves (18). At fixed positive ε, the radial correction required to keep $c = c_0$ is $u_2 = (1+\varepsilon)\theta_\Delta^2/(2\alpha^2\varepsilon) + O(\theta_\Delta^4)$ as $\theta_\Delta \to 0$, with a remainder constant that depends on ε. The coefficient grows as the normalization constant decreases, explaining the shrinking zero-excess interval. At baseline, the parameter derivatives are $\partial_{u_1} a = \alpha/\sqrt{1+\varepsilon}$ and $\partial_{u_2} c = \alpha\varepsilon/(1+\varepsilon)^{3/2}$. For every $\varepsilon > 0$, the Fisher information matrix in activation coordinates is positive definite, and the behavior score belongs to the tangent space. Local Fisher efficiency is therefore one despite the finite threshold.

F Local Fisher efficiency and higher-order excess

Let $u \in \mathbb{R}^k$ parametrize a strictly positive family p_u on a finite output set, with baseline $u = 0$. This section does not require a normalized linear readout. We compare the finite attainability results with local information geometry on a sufficiently small fixed closed activation ball.

F.1 The regular local optimum

Work in the Hilbert space of baseline-centered functions, with inner product $\langle a, c\rangle_0 = \mathbb{E}_{p_0}[ac]$ and norm $\|a\|_0$. Define

$$\tilde{f} = f - \mathbb{E}_{p_0} f, \qquad V = \|\tilde{f}\|_0^2, \qquad Jv = \frac{d}{dt}\log p_{tv}\Big|_{t=0}, \qquad F_0 = J^*J, \qquad g = J^*\tilde{f}.$$

Here J^* is the adjoint with respect to the Euclidean inner product on parameter space and $\langle\cdot,\cdot\rangle_0$ on output functions. Suppose p_u is C^3, $F_0 \succ 0$, and $g \neq 0$. Set

$$\kappa_F = g^\top F_0^{-1} g, \qquad \mathcal{T} = \operatorname{range} J, \qquad \eta_F = \kappa_F / V.$$

The subspace $\mathcal{T}$ is the tangent space of the statistical model, spanned by its likelihood score functions. We have $\kappa_F = \|P_{\mathcal{T}} \tilde{f}\|_0^2$, where $P_{\mathcal{T}} = J F_0^{-1} J^*$ is the orthogonal projection, and $0 < \eta_F \leq 1$. For the cost $\mathcal{K}_r$ under the mean inequality constraint,

$$\begin{aligned} u_*(\Delta) &= \frac{\Delta}{\kappa_F} F_0^{-1} g + O(\Delta^2), \\ \mathcal{K}_r(\Delta) &= \frac{\Delta^2}{2\kappa_F} + O(\Delta^3), \qquad\qquad \mathcal{I}(\Delta) = \frac{\Delta^2}{2V} + O(\Delta^3), \\ \mathcal{K}_r(\Delta) - \mathcal{I}(\Delta) &= \frac{\Delta^2}{2V}(\eta_F^{-1} - 1) + O(\Delta^3). \end{aligned} \tag{44}$$

Let $K(u) = D_{\mathrm{KL}}(p_u \| p_0)$ and $s(u) = \mathbb{E}_{p_u} f$. Their expansions are $K(u) = \frac{1}{2} u^\top F_0 u + O(\|u\|^3)$ and $s(u) - s(0) = g^\top u + O(\|u\|^2)$. Fix a ball on which $\nabla^2 K \succ 0$ and $\nabla s \neq 0$. The scalar inverse function theorem along $F_0^{-1} g$ gives a feasible point with norm $O(\Delta)$ and cost $O(\Delta^2)$. Compactness gives a minimizer, and strong convexity bounds its norm by $O(\Delta)$ uniformly. Thus it lies in the interior for sufficiently small Δ. The mean constraint must be active: if it were slack, $\nabla K = 0$ would force the minimizer to be the baseline, which does not attain the requested increase.

The multiplier ν satisfies

$$\nabla K(u) = \nu \nabla s(u), \qquad s(u) - s(0) = \Delta.$$

At $(u, \nu, \Delta) = (0, 0, 0)$ the Jacobian in (u, ν) is $\left(\begin{smallmatrix} F_0 & -g \\ g^\top & 0 \end{smallmatrix}\right)$. This matrix is invertible because its Schur complement is $\kappa_F > 0$. Every minimizing multiplier is $O(\Delta)$ because $\nabla K = O(\Delta)$ and $\|\nabla s\|$ is bounded below. Local uniqueness from the implicit function theorem therefore places all minimizers on the same branch, with $u_*'(0) = F_0^{-1} g / \kappa_F$. Expanding the log-partition function at baseline gives the formula for $\mathcal{I}$ and completes (44).

F.2 A finite comparison in log-density coordinates

For a baseline-centered function φ, write $p_\varphi(y) = p_0(y) e^{\varphi(y)} / \mathbb{E}_{p_0} e^{\varphi}$. If a, c are centered and $\|a\|_\infty, \|c\|_\infty \leq L_\infty$, then

$$\frac{e^{-2L_\infty}}{2} \|a - c\|_0^2 \leq D_{\mathrm{KL}}(p_a \| p_c) \leq \frac{e^{2L_\infty}}{2} \|a - c\|_0^2. \tag{45}$$

At every point φ on the segment between them, the density ratio p_φ / p_0 lies in $[e^{-2L_\infty}, e^{2L_\infty}]$. For a centered direction v,

$$\operatorname{Var}_{p_\varphi}(v) = \min_{d \in \mathbb{R}} \mathbb{E}_{p_\varphi}(v - d)^2 \geq e^{-2L_\infty} \min_d \mathbb{E}_{p_0}(v - d)^2 = e^{-2L_\infty} \|v\|_0^2,$$

The upper bound follows by taking $d = 0$. The Hessian of $\log \mathbb{E}_{p_0} e^{\varphi}$ is the same covariance. Integrating the Hessian bounds against the Taylor remainder weight, whose integral is $1/2$, proves (45) in the stated KL orientation.

F.3 The optimized fourth-order coefficient

Assume now that the family is C^5, $F_0 \succ 0$, $V > 0$, and $\tilde{f} \in \mathcal{T}$. Define its centered log-density chart and second derivative by

$$\varphi(u) = \log(p_u / p_0) - \mathbb{E}_{p_0} \log(p_u / p_0), \qquad \mathcal{B}[v, w] = D^2 \varphi(0)[v, w].$$

Let $w = F_0^{-1} g / V$, so $Jw = \tilde{f}/V$ and $g^\top w = 1$, and let $B_\perp = (I - P_{\mathcal{T}}) \mathcal{B}[w, w]$. Then there exist $r, \delta, C > 0$, depending on the family, such that the optimum over $\|u\| \leq r$ satisfies, for $0 < \Delta \leq \delta$,

$$\left| \mathcal{K}_r(\Delta) - \mathcal{I}(\Delta) - \frac{\Delta^4}{8} \|B_\perp\|_0^2 \right| \leq C \Delta^5. \tag{46}$$

If $B_\perp = 0$, a feasible activation calibrated to the target mean has excess $O(\Delta^6)$. The fourth-order coefficient is the squared norm of the second-derivative component orthogonal to the tangent space. This role of normal components is related to statistical curvature (Efron, 1975); here the mean constraint fixes the direction and normalization.

On a small compact neighborhood, probabilities are bounded below and the required derivatives of φ, K, and s are bounded. The implicit-function branch is C^4 and gives uniform Taylor remainders:

$$u_*(\Delta) = \Delta w + \Delta^2 a + O(\Delta^3), \qquad \lambda_\Delta = \Delta/V + \lambda_2 \Delta^2 + O(\Delta^3), \qquad \lambda_2 = -\frac{\mathbb{E}_{p_0} \tilde{f}^3}{2V^3}.$$

Write $q_\Delta = q_{\lambda_\Delta}$ for the minimum-KL target distribution. With $\delta_\Delta = \varphi(u_*) - \lambda_\Delta \tilde{f}$,

$$\delta_\Delta = \Delta^2 \left(Ja + \tfrac{1}{2}\mathcal{B}[w,w] - \lambda_2 \tilde{f} \right) + O(\Delta^3), \qquad (I - P_{\mathcal{T}})\delta_\Delta = \tfrac{1}{2}\Delta^2 B_\perp + O(\Delta^3).$$

Both chart points have sup norm $O(\Delta)$, and their difference is $O(\Delta^2)$. Along the segment between them, the covariance Hessian differs from its baseline value by $O(\Delta)$ uniformly in operator norm. The integral Taylor remainder gives

$$D_{\mathrm{KL}}(p_{u_*} \| q_\Delta) = \tfrac{1}{2}\|\delta_\Delta\|_0^2 + O(\Delta^5) \geq \tfrac{1}{8}\Delta^4 \|B_\perp\|_0^2 - O(\Delta^5).$$

The optimizer attains the target mean exactly, so the projection identity equates this divergence with excess cost.

For the matching upper bound, we cancel the tangent component of the second-order discrepancy. Set

$$a_2 = F_0^{-1} J^* \big(\lambda_2 \tilde{f} - \tfrac{1}{2}\mathcal{B}[w,w] \big), \qquad \widetilde{u} = \Delta w + \Delta^2 a_2.$$

The chart discrepancy is $\Delta^2 B_\perp/2 + O(\Delta^3)$. Its mean discrepancy is $O(\Delta^3)$: the baseline derivative $\langle \tilde{f}, \cdot \rangle_0$ annihilates $B_\perp$ and varies by only $O(\Delta)$ along the chart segment. The derivative of the score mean along w is $1 + O(\Delta)$, so a scalar adjustment $\widetilde{u} + s_\Delta w$ with $s_\Delta = O(\Delta^3)$ attains the exact mean. For small Δ, this derivative is at least $1/2$. The correction therefore exists uniformly, remains within the fixed radius, and preserves the second-order chart discrepancy. Applying the covariance expansion at the calibrated point gives the matching upper coefficient in (46). Derivative bounds on the compact neighborhood make the remainders uniform. If $B_\perp = 0$, the corrected point has chart error $O(\Delta^3)$, and (45) gives excess $O(\Delta^6)$.

If interventions are restricted to the fixed ray $u = sw$, scalar mean calibration can cancel only the component parallel to $\tilde{f}$. The same calculation gives the leading excess

$$\frac{\Delta^4}{8} \|P_{\tilde{f}^\perp} \mathcal{B}[w,w]\|_0^2,$$

with error $O(\Delta^5)$. Here $P_{\tilde{f}^\perp}$ denotes orthogonal projection in the baseline-centered Hilbert space. Because $\tilde{f} \in \mathcal{T}$, the difference from the optimized coefficient is $\|P_{\mathcal{T} \cap \tilde{f}^\perp} \mathcal{B}[w,w]\|_0^2/8$. Second-order activation adjustments remove this tangent component.

F.4 What local data cannot determine

Let $Y, Z_1, \ldots, Z_M$ be independent uniform signs at the baseline, and take

$$p_u \propto \exp\left\{ uY + \gamma u^2 \sum_{j=1}^{M} Z_j \right\}, \qquad f = Y, \quad \gamma > 0.$$

For every M, γ, the scalar local data are $F_0 = g = V = 1$. If $\theta_\Delta = \operatorname{arctanh} \Delta \leq r$, independence and monotonicity for $u \geq \theta_\Delta$ give

$$\mathcal{K}_r(\Delta) = k(\Delta) + M\psi(\gamma\theta_\Delta^2 \| 0),$$

where k and ψ are defined in (38). At any fixed positive feasible Δ, increasing M makes excess KL unbounded without changing the local data or r. Choosing γ proportional to Δ^{-2} also keeps the excess nonzero at arbitrarily small gains. Thus these local quantities alone cannot yield a uniform finite-gain bound.

Likewise, $p_u(Y,Z) \propto \exp\{uY + \gamma u^3 Z\}$ has $F_0 = g = V = 1$ and $B_\perp = 0$, but its optimized excess is $\psi(\gamma(\operatorname{arctanh}\Delta)^3\|0) = \gamma^2\Delta^6/2 + O(\Delta^8) > 0$. Thus a zero fourth-order coefficient does not imply zero excess. A sufficient condition is a feasible path satisfying $\varphi(u(\lambda)) = \lambda\tilde{f}$. For a differentiable path starting at zero, this condition follows by integrating $D\varphi(u(\lambda))u'(\lambda) = \tilde{f}$ over the full interval. Accessibility of the target score at baseline guarantees neither such a path nor its feasibility within budget. The local theorem also excludes singular Fisher matrices: for $p_u \propto p_0 e^{u^2\tilde{f}}$, both F_0 and g vanish at zero, although positive mean gains are attainable with u of order $\sqrt{\Delta}$.

G Restricted interventions and the scope of the readout model

G.1 Range mismatch and the nullspace condition

If $\operatorname{rank}(A) < m-1$, then $\operatorname{range}(A)$ is a proper linear subspace of $\mathbf{1}^\perp$. A projected score direction drawn from an absolutely continuous distribution on $\mathbf{1}^\perp$ therefore lies outside $\operatorname{range}(A)$ almost surely. This dimension argument establishes nonmembership but does not bound the residual. Event indicators form a discrete collection, so their distances to the readout range require a separate calculation. Appendix H reports those distances.

For full-column-rank W, the centered version has a nontrivial nullspace exactly when $\mathbf{1} \in \operatorname{range}(W)$. Indeed, $PWv = 0$ implies $Wv = c\mathbf{1}$. For nonzero v, injectivity requires $c \neq 0$; conversely, a preimage of $\mathbf{1}$ gives a centered null vector. Centering can therefore reduce the rank of a full-column-rank readout by at most one. Vocabulary size alone does not determine this condition. Theorem 6 also applies to rank-deficient effective readouts with a nontrivial nullspace.

G.2 An exact criterion for affine intervention subspaces

Let $U \in \mathbb{R}^{d\times k}$ have orthonormal columns, with $1 \le k \le d$, and restrict the activation to $h = h_0 + Uu$, $\|u\| \le r$. Write $\mathcal{K}_r^U$ for the resulting objective and define

$$C_{\varepsilon,U}(\lambda) = \inf_{x(h_0+Uu)=t_\lambda} \|u\|.$$

The compactness and projection arguments in Appendix A prove (19). The lower KL bound also applies with distance measured to the restricted logit image. Thus nonattainability under the full ball extends to every such method. Conversely, a target attainable in the full ball need not be attainable in the prescribed subspace.

The restricted minimum norm can be computed without optimizing over all activation coordinates. Set

$$A_U = AU, \quad R_U = A_U^\dagger A_U, \quad u_0 = U^\top h_0, \quad h_{\perp U} = h_0 - Uu_0, \quad \beta_U = \|(I_k - R_U)u_0\|.$$

For the target point t_λ and a scale $s > 0$, define

$$v_s = A_U^\dagger(st_\lambda - Ah_0),$$
$$\rho_s^2 = s^2 - \varepsilon - \|h_{\perp U}\|^2 - \|R_U u_0 + v_s\|^2.$$

In each case below, the infimum is over scales $s > 0$ satisfying $st_\lambda - Ah_0 \in \operatorname{range}(A_U)$ and the stated additional condition. In the second case, $\rho_s = \sqrt{\rho_s^2}$.

$$C_{\varepsilon,U}(\lambda)^2 = \begin{cases} \inf_s \|v_s\|^2, & \ker A_U = \{0\}, \quad \rho_s^2 = 0, \\ \inf_s\{\|v_s\|^2 + (\rho_s - \beta_U)^2\}, & \dim\ker A_U \ge 1, \quad \rho_s^2 \ge 0. \end{cases} \tag{47}$$

An empty scale set gives $+\infty$. These are one-dimensional characterizations; they do not require a general convexity claim, unlike (15).

Proof. For a realizing activation, set $s = \sqrt{\|h_0 + Uu\|^2 + \varepsilon}$. The logit equation is $A_U u = st_\lambda - Ah_0$. The range condition is necessary and sufficient for its solvability, and every solution has the form $u = v_s + n$ with $n \in \ker A_U$. Since $h_{\perp U} \perp \operatorname{range}(U)$, the normalization equation becomes

$$\|(I_k - R_U)u_0 + n\|^2 = \rho_s^2.$$

For a trivial kernel, the equation requires $\rho_s^2 = 0$ and fixes $u = v_s$. Otherwise, the vector on the left can have any nonnegative length. At a fixed scale, alignment with $(I_k - R_U)u_0$ gives the minimum squared norm of n, namely $(\rho_s - \beta_U)^2$. Choose a unit vector $\widehat{n}_U$ along that vector when $\beta_U > 0$, and any unit vector in $\ker A_U$ otherwise. Then

$$u = v_s + \rho_s \widehat{n}_U - (I_k - R_U)u_0$$

realizes the target at the cost in (47). Orthogonality of the row-space and nullspace components proves optimality at each scale; taking the infimum proves the formulas. A finite minimum intervention norm is attained. Indeed, the target preimage in parameter space is closed, and restricting it to a ball through any preimage point preserves the infimum on a compact set. □

For a fixed unit steering direction, take $k = 1$. If strengths must be nonnegative, retain only reconstructed scalars $u \geq 0$ and minimize u^2 over them. When $A_U = 0$, both nullspace orientations must be checked before imposing this restriction. A two-sided strength interval and a one-sided ray define different intervention domains.

There is also a direct reduction to the original theorems. If $Ah_{\perp U} = 0$, the fixed perpendicular component is annihilated by the centered readout, and

$$x(h_0 + Uu) = \frac{A_U(u_0 + u)}{\sqrt{\|u_0 + u\|^2 + \varepsilon + \|h_{\perp U}\|^2}}.$$

The original model therefore applies in dimension k, with baseline u_0, centered readout A_U, and normalization constant $\varepsilon + \|h_{\perp U}\|^2$. This case includes $h_0 \in \mathrm{range}(U)$. The conic theorem applies to the reduced model when $\ker A_U \neq \{0\}$. For a general affine subspace, the scalar certificate remains valid, but intersection with the activation ball alone does not imply a convex logit image.

G.3 From derivative bounds to a uniform logit-error bound

Fix a context and an intervention layer, and hold all other activation coordinates at baseline. Let G be the centered downstream map and x a specified normalized-readout reference. Set $e = G - x$ and suppose e is C^2 on a neighborhood of $\mathcal{D}_r$. If

$$\|De(h_0)\| \leq a_e, \qquad \sup_{h \in \mathcal{D}_r} \|D^2 e(h)\| \leq M_e,$$

where the second norm is the induced bilinear operator norm, the integral form of the Taylor remainder gives

$$\sup_{h \in \mathcal{D}_r} \|e(h) - e(h_0)\| \leq a_e r + \tfrac{1}{2} M_e r^2, \qquad \sup_{h \in \mathcal{D}_r} \|e(h)\| \leq \|e(h_0)\| + a_e r + \tfrac{1}{2} M_e r^2.$$

To obtain these bounds, expand along $h_0 + tu$ and integrate the Hessian term with weight $1 - t$. They can be used as ζ and η in Theorem 4. If the local right inverse and smallness hypotheses also hold, they apply in Theorem 8. The uniform Hessian assumption provides control over the entire ball.

For comparison, the affine Taylor reference $G(h_0) + DG(h_0)(h - h_0)$ has a quadratic remainder when the downstream Hessian is uniformly bounded. Measured accuracy of this reference does not establish the geometry of a normalized linear readout. Appendix H reports the two reference maps separately.

G.4 Average objectives across contexts

Consider finitely many contexts j that share the normalized readout map x. Let $\omega_j > 0$ sum to one, and let $p_{0,j}$ and f_j be their baseline distributions and output scores. The objective is the weighted sum of KL divergences, subject to a weighted score increase of at least Δ. The unrestricted minimizers are contextwise exponential tilts $q_{j,\lambda}$ with a common parameter. Define

$$\tilde{f}_j = f_j - \mathbb{E}_{p_{0,j}} f_j, \qquad \overline{\Lambda}(\lambda) = \sum_j \omega_j \log \mathbb{E}_{p_{0,j}} e^{\lambda \tilde{f}_j}.$$

For an interior positive target increase with at least one nonconstant score, the unique parameter satisfies $\overline{\Lambda}'(\lambda) = \Delta$, and $\overline{\mathcal{I}}(\Delta) = \lambda\Delta - \overline{\Lambda}(\lambda)$. Summing (3) over contexts gives

$$\sum_j \omega_j D_{\mathrm{KL}}(p_j \| p_{0,j}) - \overline{\mathcal{I}}(\Delta) = \sum_j \omega_j D_{\mathrm{KL}}(p_j \| q_{j,\lambda}) + \lambda \left(\sum_j \omega_j \mathbb{E}_{p_j} \tilde{f}_j - \Delta \right). \tag{48}$$

Zero excess therefore requires simultaneous attainment of the minimum-KL target in every context. Let $h_{0,j}$ be the baseline activation in context j and set $x_{0,j} = x(h_{0,j})$. Under a shared displacement u, the joint target preimage consists of displacements satisfying $x(h_{0,j} + u) = x_{0,j} + \lambda P f_j$ for every j. If the shared intervention domain is compact, zero excess is equivalent to nonemptiness of this set.

With independent interventions and separate budgets, the product of convex attainable logit sets is convex, so attainable parameters along the common target ray form an interval. Under a shared displacement, the joint attainable set need not equal that product; convexity of each marginal image alone does not imply a threshold. A multi-token property similarly requires a sequence distribution and its intervention map. The fixed-context readout theorem does not replace either additional model.

H Pretrained-model diagnostics

H.1 Checkpoints, token sets, and projection calculation

We use the base checkpoints SmolLM2-135M, SmolLM2-360M, and Qwen2.5-0.5B. Exact checkpoint revisions and complete reproduction records are provided in the supplementary material. The checkpoints have vocabulary sizes $49{,}152$, $49{,}152$, and $151{,}936$, with hidden dimensions 576, 960, and 896, respectively. All three use tied embeddings. Let V_{out} be the output projection (unembedding) matrix and g_{RMS} the learned final normalization gain. We form

$$W = \sqrt{d}\, V_{\mathrm{out}} \operatorname{diag}(g_{\mathrm{RMS}}), \qquad \varepsilon = d\varepsilon_{\mathrm{RMS}}, \qquad A = PW.$$

The implementation constants are 10^{-5} for both SmolLM2 checkpoints and 10^{-6} for Qwen2.5-0.5B. The conversion uses the exact mean-square normalization formula; the implementation constant is not the parameter of the unit-norm example in Section 5.4.

For each token ID, we decode the token in isolation, strip surrounding whitespace, and apply case folding. We exclude special tokens and IDs without a tokenizer entry while retaining every model output row in A. The four indicator sets contain: nonempty strings of ASCII decimal digits; nonempty strings using only `.`, `!`, and `?`; the words `yes`, `true`, `correct`; and the words `good`, `great`, `excellent`, `wonderful`, `happy`, `love`, `amazing`, `beautiful`, `best`, `pleasant`. Both SmolLM2 vocabularies give set sizes $10, 22, 12, 30$; Qwen gives $10, 111, 17, 42$. These sets define token-level lexical indicators, not sequence-level semantic attributes.

We convert the stored bfloat16 entries exactly to float64, form the Gram matrix in vocabulary chunks, and solve the resulting positive-definite linear system for each score projection. The smallest-to-largest singular-value ratios are 0.018059, 0.012741, and 0.00039323. At the 10^{-6} relative cutoff used in Table 1, all three stored centered effective readouts have full column rank. The numerical ranks are unchanged at 10^{-4} and 10^{-8}. Table 1 reports the projection residuals.

H.2 Minimum intervention norms at the measured normalization scale

We evaluate SmolLM2-135M on eight short English completion prompts without a chat template, covering factual, arithmetic, descriptive, and programming contexts. For each prompt, h_0 is the last-position residual immediately before the final RMSNorm. Set $z_0 = N_\varepsilon(h_0)$. We draw a Gaussian vector, project it onto $z_0^\perp$, and rescale the resulting vector v so that $f = Av$ has baseline variance one. Hence $Pf = f \in \operatorname{range}(A)$, and the target normalized activation is $z_0 + \lambda v$. These constructed scores isolate the budget obstruction; the lexical scores remain outside the readout range.

Because the centered readout has full column rank, Theorem 2 reduces C_ε to the injective formula (14). We solve $C_\varepsilon(\lambda) = 0.1\|h_0\|$ by bracketing before $z_0 + \lambda v$ reaches the unit-sphere boundary, then compute the score increase from the exact softmax tilt. In exact arithmetic, $v \perp z_0$ makes both radial and orthogonal displacements increase with positive λ. The crossing therefore ends the zero-excess interval. Increases are measured in units of the baseline standard deviation of the constructed score, not changes in lexical event probability.

H.3 Sampled intermediate-layer errors

All downstream calculations use float64. We intervene at the final position immediately after block 15 or 29, holding all other positions at their baseline values at that layer. Let $h_{0,\ell}$ be the baseline activation after block ℓ. The suffix includes all remaining blocks, the final RMSNorm, and the unembedding. For each of the same eight prompts and each layer, we sample eight Gaussian unit directions and add a radial direction and the normalized gradient of the next-token digit probability. We evaluate both signs at relative radii $0.001, 0.01, 0.05$, and 0.1, for 1,280 perturbed evaluations.

Let G_ℓ denote the centered suffix logits. Using the final effective matrix A, define the direct normalized-readout reference $x(h) = AN_\varepsilon(h)$. For each displacement u, we record

$$e_\ell(h) = G_\ell(h) - x(h), \qquad \xi_\ell(u) = e_\ell(h_{0,\ell} + u) - e_\ell(h_{0,\ell}),$$
$$S_\ell(u) = \|G_\ell(h_{0,\ell} + u) - G_\ell(h_{0,\ell})\|, \quad R_\ell(u) = \|G_\ell(h_{0,\ell} + u) - G_\ell(h_{0,\ell}) - DG_\ell(h_{0,\ell})u\|.$$

Here R_ℓ measures affine Taylor error, not the error of x. Figure 5 reports the largest error increments and Taylor remainders over 160 prompt/direction pairs for each layer and radius. At relative radius 0.1, median $\|\xi_\ell(u)\|/S_\ell(u)$ ratios are 2.69 (block 15) and 0.832 (block 29); median $R_\ell(u)/S_\ell(u)$ ratios are 0.183 and 0.0718. Each median is over per-pair ratios, not a ratio of maxima.

The direct reference has an error increment comparable to the output change even near the final block. The affine Taylor reference is more accurate at small radii, but that local accuracy does not imply accuracy of the normalized-readout reference. The samples do not certify the uniform bounds in Section 6; sampled maxima are lower bounds on the corresponding suprema.

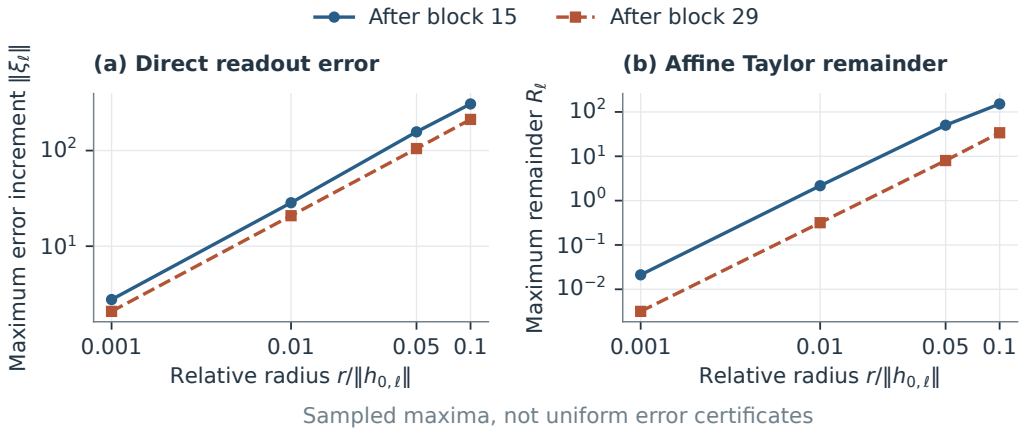

Figure 5: Sampled downstream errors for eight prompts and 20 signed directions per prompt in SmolLM2-135M. (a) Error increments for the direct normalized-readout reference. (b) Remainders for the affine Taylor reference. Each point is the maximum over the same 160 prompt/direction pairs at that layer and radius; connecting lines guide the eye. Both panels use logarithmic axes. These measurements do not certify the supremum over the activation ball.