# AUDIT FRONTIERS AND FINITE BUDGET RANK DECISIONS UNDER PARTIAL OBSERVATIONS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Partial observations can prove a predictive law wrong without saying what should replace it. We score updates of a finite predictive law by their worst case log score improvement over every latent world consistent with the observed law. A trust budget limits divergence from the reference; an audit budget limits divergence from the least committal law with the same observations. On a compact subanalytic set of available updates, the gain exponent, as a function of the audit precision, is eventually constant exactly when positive gain survives a zero audit. The comparison behind this criterion specializes preparation theory and bounds the gain by a finite maximum of minima of monomials in the two budgets, with constants independent of their ratio. In a conditional logit module, the vanishing orders of singular values determine the correction rank needed at each audit precision, and in a sixteen state tied network, bounds over every admissible action give exact minimum rank decisions at finite budgets. We also give certificates that hold uniformly over updates fitted to the same observations and a hardness construction whose instances share a second order jet. On a rain and sprinkler example, four of five deployed language models, told only how often the ground was wet, changed the unidentified split between rain and the sprinkler in almost every revision. In trained adapter families of two open weight models, that split moves at approximately first order over the measured edit scales, outside the module's second order assumption. This audit tests no theorem. All guarantees concern this finite interface, and no result is validated on a language model.

## 1 INTRODUCTION

A model states how likely it is that, on a summer morning, rain fell overnight and the lawn sprinkler ran, that only one of them happened, or that neither did. A record then reports only how often the ground was wet, and it shows more wet mornings than the model predicted. The record proves the prediction wrong, yet it cannot say whether the extra wet mornings came from rain, from the sprinkler or from both. What can a revision guarantee from such partial observations, and how much must it commit beyond them?

We score a revision of a finite reference law by its worst world gain, the smallest improvement in expected log score over all latent laws that agree with the observed law; in the example, an adversary chooses how the wet mornings divide among the three wet cases. A trust budget $b$ bounds the revision's KL divergence from the reference. An audit budget $\beta$ bounds its divergence from the least committal law with the same observations, the information projection of the reference (Csiszár, 1975); in the example this law is Jeffrey's update (Jeffrey, 1990), which keeps the reference division of wet mornings. The trust cost splits exactly into the least trust cost of any law with the revision's observations plus the audit (Section 2). A revision has zero audit exactly when it is the least committal law given its own observations; under a stochastic channel it can still move moments that the observations do not identify (Appendix J).

Editing or fine tuning should repair what new evidence contradicts without committing to what it cannot determine. We call the maximization of worst world gain under both budgets alignment at a finite interface; the term refers only to agreement with an observed law. Two features make this maximization hard: available updates may form a restricted, singular set that need not contain Jeffrey's update, such as the rank constrained adapter image in the designed network of Section 4, and



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.

Table 1: The central results use different information. The complete map is Table 9.

| Setting | Given information | Deliverable |
|---|---|---|
| General frontier | Channel, reference, target and tame action set | Uniform local comparison; no numerical radius |
| Tied finite model | Known architecture and target contrast; the stated prior already certifies feasible actions | Feasible factors and finite rank tests |
| Conditional logit module | Analytic displacement and corrections of bounded rank | Critical ranks, exponents and leading constants; no radius |
| Cubic score graphs | Only the reference, target, channel and second order jet | Instances with equal jets whose values differ by a factor above 200 |

the two budgets may shrink at unrelated rates. Yet the gain has a finite description in both budgets; only the finest audit precisions reveal whether gain survives a zero audit. Our contributions follow.

- **A zero audit criterion** (Section 3). On a compact subanalytic set of updates, the attainable gain is locally zero or comparable to a finite max–min of rational monomials in the two budgets, with constants that do not depend on their ratio (Theorem 3.1, a specialization of preparation theory). Along power schedules of the audit, the gain exponent has a terminal plateau exactly when positive gain survives a zero audit (Corollary 3.3), while a four phase example has an arbitrarily long earlier plateau although a zero audit permits no gain (Figure 1; Appendix E). Section 3 also expresses the exponents through the vanishing orders of analytic arcs (Appendix D).
- **Exact rank decisions in one network** (Section 4). In a sixteen state network with one shared weight and a rank constrained adapter, the singular value orders $(2, 5)$ of the displacement set the audit caps $(4\beta)^{1/4}$ and $(4\beta)^{1/10}$ of ranks zero and one (Theorem 4.1), and for $b \leq 1/512$ bounds over every admissible factor choice make ranks zero, one and two each the exact minimum at one of three audit budgets (Table 2). The orders are engineered and collapse to $(2, 2)$ under a fixed perturbation of one path (Appendix G.8). Section 5 shows that the same orders govern the generic conditional logit module: the minimum rank keeping a positive fraction of the unconstrained gain is the number of singular values vanishing at an order below the audit exponent (Theorem 5.1).
- **Second order information and certificates from data** (Sections 5 and 6). In a 3SAT construction whose instances share the reference, target, channel and second order jet, satisfiable and unsatisfiable formulas differ in value by a factor above 200 at explicit budgets, so approximating the value within a factor of two is NP hard (Theorem I.1). Without any complexity assumption, no rule seeing only the shared data approximates the value within a fixed factor at all small budgets (Corollary 5.2). Certificates from observed data answer a separate question: they hold on one event for every update and pair of budgets chosen with the same data (Theorem H.1 in Appendix H).

Section 7 computes these quantities for 1988 revisions stated by five deployed language models on the rain and sprinkler example; the audit tests no theorem. Told only the wet frequency, four of the models changed the division of wet mornings in almost every revision, even within five points of the record; 695 of the 1188 revisions under the main prompt have negative worst world gain, whereas Jeffrey's update has positive worst world gain for every record. A reasoning model mostly returned Jeffrey's update.

The ingredients have precedents (Appendix A). The worst world gain is a worst case over an identification region (Manski, 2003), a moment based ambiguity set, as in distributionally robust optimization (Delage & Ye, 2010); with unrestricted updates and a large trust budget its optimum is the robust Bayes equalizer (Grünwald & Dawid, 2004). Preparation theory already supplies comparisons uniform in the budget ratio and, over real exponent fields, feasible selectors (Buggenhout et al., 2026), and logarithmic limit sets in polynomially bounded o-minimal structures are polyhedral (Alessandrini, 2013). Theorem 3.1 specializes these results to a pure max–min of monomials, which we prove directly by resolution and linear programming; we prove no new preparation theorem. Our additions are the zero audit criterion of Corollary 3.3, exponents described by analytic arcs, and the finite rank decisions of Section 4. Singular value orders are classical (Kaveh & Makhnatch, 2022; Gantmakher, 1998), and the finite construction concerns one designed architecture, not low rank adaptation in general (Hu et al., 2022).

## 2 THE ALIGNMENT PROBLEM AT A FINITE INTERFACE

Fix finite spaces $\Omega, \mathcal{Y}$, a reference law $p_0 > 0$, a column stochastic channel $K$, and a feasible observed law $r$. Remove identically zero rows of $K$. A positive update $q$ is scored against every compatible latent world:

$$\mathcal{C}_r = \{p \in \Delta(\Omega) : Kp = r\} \neq \varnothing, \qquad \Gamma(q; r) = \min_{p \in \mathcal{C}_r} \mathbb{E}_p \log \frac{q}{p_0}. \tag{1}$$

Thus $\Gamma(q; r) > 0$ certifies $\mathrm{KL}(p_* \| p_0) - \mathrm{KL}(p_* \| q) \geq \Gamma(q; r)$ for every $p_* \in \mathcal{C}_r$. This differs from the observed gain $\Gamma_{\mathrm{obs}}(q; r) = \sum_y r_y \log[(Kq)_y / (Kp_0)_y]$: latent improvement need not reach $Kq = r$ or improve the observed risk, even at zero audit (Appendix J.3).

**Scores and costs.** Put $\mathsf{S} = \{h : \mathbb{E}_{p_0} h = 0\}$, $\langle h, g \rangle_0 = \mathbb{E}_{p_0} hg$ and $\|h\|_0^2 = \mathbb{E}_{p_0} h^2$. Every positive law has a unique score in $\mathsf{S}$, with

$$q_h(i) = p_0(i) e^{h(i) - \psi(h)}, \qquad \psi(h) = \log \mathbb{E}_{p_0} e^h, \tag{2}$$

$$\Gamma(h) := \Gamma(q_h; r) = \min_{p \in \mathcal{C}_r} \mathbb{E}_p h - \psi(h). \tag{3}$$

The fiber is a compact polytope, so $\Gamma$ is locally Lipschitz. Let $L = \mathsf{S} \cap \mathrm{range}(K^\top)$. The projection audit compares an update with the least committal law that has the same observations:

$$\pi_r = \arg\min_{v \in \mathcal{C}_r} \mathrm{KL}(v \| p_0), \qquad \mathcal{B}(q) = \mathrm{KL}(q \| \pi_{Kq}), \qquad B(h) = \mathcal{B}(q_h). \tag{4}$$

The projection is unique. Its Pythagorean identity (Csiszár, 1975) splits the trust cost as

$$D(h) := \mathrm{KL}(q_h \| p_0) = \mathrm{KL}(\pi_{Kq_h} \| p_0) + B(h), \qquad B(h) = 0 \iff h \in L. \tag{5}$$

Both costs are analytic near zero, even with redundant channel rows: $D(h) = \frac{1}{2}\|h\|_0^2 + O(\|h\|_0^3)$ and $B(h) = \frac{1}{2}\|h_\perp\|_0^2 + O(\|h\|_0 \|h_\perp\|_0^2)$, where $h_\perp$ is the orthogonal component in $L^\perp$ (Lemma B.1). The relative remainder controls the error when $\beta \ll \|h\|_0^3$. For partition observations, $\pi_r(i) = r_y p_0(i \mid y)$ is Jeffrey's update and $\Gamma \leq \Gamma_{\mathrm{obs}}$, with equality at zero audit. A general channel has no such implication; the audit preserves projection structure, not the truth of latent beliefs (Appendix J).

**Assumption 2.1** (A closed local tame set of updates)**.** The available scores $\mathcal{M} \subset \mathsf{S}$ form a compact subanalytic set containing zero, contained in a neighborhood on which $D$ and $B$ are analytic. All local values refer to this fixed $\mathcal{M}$.

Subanalytic sets are locally projections of relatively compact sets given by finitely many analytic equalities and inequalities. They allow singular rank constraints; compact local images of analytic parameterizations give examples (Appendix C). Restricting to a smaller closed score neighborhood leaves all sufficiently small trust budget values unchanged, since $D$ has its unique zero at the reference. Define

$$G(b, \beta) = \max_{\substack{h \in \mathcal{M} \\ D(h) \leq b,\ B(h) \leq \beta}} \Gamma(h), \qquad G_0(b) = G(b, 0). \tag{6}$$

The maximum is attained and nonnegative; $B \leq D$ makes the audit redundant when $\beta \geq b$. We write $G(b, \beta; r)$ if needed, and $f \asymp g$ for two sided positive constant bounds fixed by $\mathcal{M}, K, p_0, r$, never by the budgets or their ratio unless stated otherwise. Section 4 derives one such action image with explicit finite budget constants.

A declared total parser $T$ connects an output law to this interface: $p_0(i) = Q_0\{T = i\}$ and verdicts follow $K(y \mid i)$. Data processing gives $\mathrm{KL}(q_\vartheta \| p_0) \leq \mathrm{KL}(Q_\vartheta \| Q_{\vartheta_0})$, so an output level KL constraint (Yang et al., 2024) implies the finite trust constraint. Equality holds for the canonical lift in Proposition J.1, which is not assumed to lie in $\mathcal{M}$. Tables 8 and 9 in Appendix M collect notation and information requirements.

## 3 THE ATTAINABLE GAIN UNDER TWO BUDGETS

Even on a singular action set, the optimal gain has a finite asymptotic description in the two independently shrinking budgets. The comparison stays uniform when their ratio tends to zero or infinity.

**Theorem 3.1** (A uniform envelope in both budgets). *Under Assumption 2.1, either $\Gamma \le 0$ on $\mathcal{M}$ near zero and $G$ is locally zero, or there are nonempty finite sets $V_1, \dots, V_N \subset \mathbb{Q}^2_{\ge 0}$ and positive constants $c, C, b_0, \beta_0$ such that*

$$cH(b,\beta) \le G(b,\beta) \le CH(b,\beta), \qquad H(b,\beta) = \max_{1\le i\le N} \min_{(u,v)\in V_i} b^u \beta^v \tag{7}$$

*for every $0 < b < b_0$ and $0 < \beta < \beta_0$.*

Call the second alternative the *positive case*. The proof resolves the joint graph of the costs and positive gain using uniformization and normal crossings (Bierstone & Milman, 1988, Theorem 0.1 and Corollary 4.9). Full coordinate boxes map into the actual action set, including their boundary faces. On each retained face, logarithmic coordinates turn monomial comparisons into a linear program: its dual vertices give $V_i$, and a fixed contraction of a primal optimizer gives a feasible action. Zero audit faces omit the second constraint. The finite cover and contractions precede both budgets, so the constants also cover audits vanishing faster than every power of $b$ (Appendix C). The comparison is existential. Deciding its zero audit endpoint is NP hard for polynomially encoded cubic score graphs (Theorem I.1); Section 4 computes constants and feasible actions for one model.

**Corollary 3.2** (A finite spectrum of audit precisions). *In the positive case, for every real $\alpha > 0$ and $\lambda > 0$, with comparison constants allowed to depend on $\alpha, \lambda$,*

$$G(b, \lambda b^\alpha) \asymp b^{\gamma(\alpha)}, \qquad \gamma(\alpha) = \min_i \max_{(u,v)\in V_i} (u + \alpha v). \tag{8}$$

*The function $\gamma$ is continuous, nondecreasing and piecewise affine with rational coefficients and breakpoints, satisfies $\gamma(\alpha) \ge 1/2$, and is independent of $\lambda$. It need not be convex or concave.*

A larger exponent means less attainable gain. Over positive gain analytic action arcs from zero, Theorem D.1 also gives

$$\gamma(\alpha) = \inf_h \max\{d_h/a_h, \alpha d_h/c_h\},$$

where $a_h, c_h, d_h$ are the vanishing orders of trust, audit and gain, and $c_h = \infty$ for an identically zero audit. For rational $\alpha$, an arc attains the infimum and the value has a sharp leading term. The best arc can change with $\alpha$; an intermediate plateau therefore need not mean that zero audit permits improvement.

**Corollary 3.3** (The terminal plateau characterizes gain at zero audit). *In the positive case, exactly one of the following holds.*

- *(i) $G_0(b) > 0$ for all sufficiently small $b > 0$. Then $G_0(b) \sim A_0 b^{\gamma_0}$ for some $A_0 > 0$ and rational $\gamma_0 \ge 1/2$, and $\gamma(\alpha) = \gamma_0$ for all sufficiently large $\alpha$.*
- *(ii) $G_0$ is locally zero. Then, for all sufficiently large $\alpha$, $\gamma(\alpha) = u_* + v_*\alpha$ with rational $u_* \ge 0$ and $v_* > 0$.*

At fixed small $b$, compactness gives $G(b,\beta) \to G_0(b)$ as $\beta \downarrow 0$; uniform comparison identifies the envelope branches that survive. This concerns the actual set $\mathcal{M} \cap L$, not a leading Taylor kernel, and does not make every available family audit free. Appendix D proves the endpoint and recovery results. Appendix E gives a four phase example with $G_0 = 0$ but an arbitrarily long finite plateau (Figure 1; exponents verified numerically in Appendix L); no finite plateau length decides the endpoint.

## 4 One tied model and finite rank decisions

The general frontier becomes a finite decision in the following designed model. Let $\Omega = \{1,2\} \times \{-1,1\}^3$, with coordinates $(c, \ell, \xi_1, \xi_2)$ and uniform reference $p_0$. The known partition reports $(c, \ell)$; its target law is $r_\delta(c,\ell) = (1 + \delta\ell)/4$, where $1/2 \le \delta < 1$. For the one hot input $x = e_c$, a shared scalar $t \in [0, 1/8]$ defines an encoder $z = tx_1 + t(tx_2)$ and hidden head outputs $y_1 = tz$, $y_2 = t(tz) + t(t(t(t(tx_2))))$. The last summand is a separate five edge path. Every displayed edge has the same trainable weight $t$; fixed routing and exact tying are architectural assumptions. Multiplying these paths, with rows indexed by context, gives

$$C(t) = \begin{pmatrix} t^2 & t^3 \\ t^3 & t^4 + t^5 \end{pmatrix} = v(t)v(t)^\top + t^5 e_2 e_2^\top, \qquad v(t) = (t, t^2)^\top. \tag{9}$$

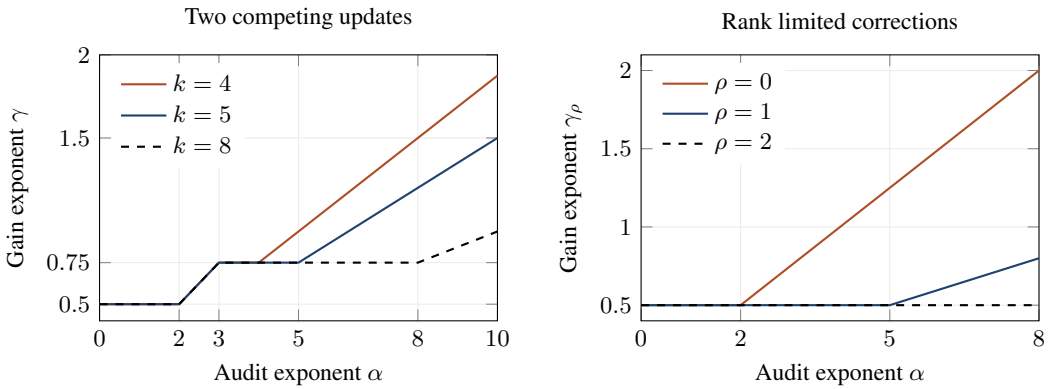


Figure 1: Exponents computed from the proved formulas with $\beta = \lambda b^\alpha$; a smaller value means larger attainable gain as $b \downarrow 0$. Left: the four phase example of Appendix E for $k = 4, 5, 8$. The axis update governs $\alpha \le 3$ and the cancellation curve governs $\alpha \ge 3$. The plateau on $[3, k]$ lengthens with $k$ but is never terminal, since only the reference has zero audit. Right: corrections of rank at most $\rho$ for the displacement (9) of Section 4, with singular orders $(2, 5)$. These curves describe that exact displacement; a fixed nonzero $t^2$ perturbation of its lower right entry changes the local orders to $(2, 2)$. The common portions of curves coincide exactly.

The observed sign has logits $\pm t$. A two layer adapter adds $\Delta W x = V U^\top x$ to the hidden heads, with $U, V \in \mathbb{R}^{2\times\rho}$, $\|U\|_F, \|V\|_F \le 1/2$, and $\rho \in \{0, 1, 2\}$. Thus its context row correction is $E = UV^\top = (\Delta W)^\top$, and $S = C(t) + E$. Rank constrains this adapter increment only: changing $t$ also changes the base paths and the observed head. The resulting law is

$$q_{t,U,V}(c, \ell, \xi) = \frac{1}{2}\frac{e^{t\ell}}{2\cosh t}\prod_{j=1}^{2}\frac{e^{S_{cj}\xi_j}}{2\cosh S_{cj}}. \tag{10}$$

**The actual action image.** Put $L_c(S) = \sum_j \log\cosh S_{cj}$ and $\bar{L} = (L_1 + L_2)/2$. The centered score and its normalizer are $h_{t,S} = t\ell + \sum_j S_{cj}\xi_j - L_c(S) + \bar{L}$ and $\psi(h_{t,S}) = \log\cosh t + \bar{L}$. Nuclear norm duality and balanced singular value factorization give the exact image

$$\mathcal{M}_\rho^{\text{tied}} = \{h_{t,C(t)+E} : 0 \le t \le 1/8,\ \operatorname{rank} E \le \rho,\ \|E\|_* \le 1/4\}. \tag{11}$$

This is not a Frobenius ball for $E$. The coefficient inverse is $t = \mathbb{E}_{p_0}[\ell h]$ and $S_{cj} = 2\mathbb{E}_{p_0}[\mathbf{1}_{\{\text{context}=c\}}\xi_j h]$. Consequently the compact parameter domain has a compact subanalytic score image, although the factors are not identifiable (Appendix G). Write $\mathsf{G}_\rho(b, \beta; \delta)$ for its optimal gain under the two budgets. With $\mathcal{J}(u) = u\tanh u - \log\cosh u$, direct calculation gives

$$B(t, E) = \tfrac{1}{2}\sum_{c,j}\mathcal{J}(S_{cj}), \qquad D(t, E) = \mathcal{J}(t) + B(t, E),$$
$$\Gamma_\delta(t, S) = \delta t - \log\cosh t - \tfrac{1}{2}\sum_{c,j}\big(|S_{cj}| + \log\cosh S_{cj}\big). \tag{12}$$

The minimum defining the gain ranges over *all* compatible worlds; their hidden signs need not be independent.

**Theorem 4.1** (Finite budget bounded factor frontier)**.** *For $0 < b \le 1/512$, $\beta \ge 0$ and $1/2 \le \delta < 1$, set*

$$T_0 = \min\{\sqrt{2b}, (4\beta)^{1/4}\}, \qquad T_1 = \min\{\sqrt{2b}, (4\beta)^{1/10}\}, \qquad T_2 = \sqrt{2b}. \tag{13}$$

*The upper bounds below hold for every admissible factor choice:*

$$\begin{aligned}
\tfrac{108}{125}\delta T_0 \le \mathsf{G}_0(b, \beta; \delta) \le \tfrac{1003}{1000}\delta T_0,\\
\tfrac{117}{125}\delta T_1 \le \mathsf{G}_1(b, \beta; \delta) \le \tfrac{1003}{1000}\delta T_1,\\
\mathsf{G}_2(b, \beta; \delta) = \delta\mathcal{J}_+^{-1}(b) - \log\cosh\big(\mathcal{J}_+^{-1}(b)\big),\\
\tfrac{15}{16}\delta T_2 \le \mathsf{G}_2(b, \beta; \delta) \le \tfrac{1003}{1000}\delta T_2.
\end{aligned} \tag{14}$$

Table 2: Exact minimum rank decisions at $b = 1/5000$, $\delta = 1/2$ and requested gain $g = 1/200$. The lower bounds use rational feasible steps, not numerical roots. The last column excludes every smaller rank.

| $\beta$ | Minimum rank | Feasible $t$ | Gain lower bound | Lower rank exclusion |
|---|---|---|---|---|
| $10^{-4}$ | 0 | $499/25000$ | $108/12500$ | none needed |
| $10^{-10}$ | 1 | $999/50000$ | $117/12500$ | $\frac{1003}{2000}(4 \cdot 10^{-10})^{1/4} < g$ |
| $10^{-24}$ | 2 | $1/50$ | $3/320$ | $\frac{1003}{2000}(4 \cdot 10^{-24})^{1/10} < g$ |

*Here $\mathcal{J}_+^{-1}$ is the inverse on nonnegative arguments. At $\beta = 0$, the values at ranks zero and one are zero, whereas rank two has the displayed positive value. At $b = 0$ every rank has value zero. All constants are exact rationals, not rounded numerical certificates.*

**Feasible parameters, not an optimization oracle.** Use $E_0 = 0$ at rank zero and the polynomial factors

$$U^{(1)} = -v, \quad V^{(1)} = v; \qquad U^{(2)} = [-v, -t^2 e_2], \quad V^{(2)} = [v, t^3 e_2]. \tag{15}$$

They obey the individual norm caps throughout $[0, 1/8]$ and leave residuals $S_0 = C(t)$, $S_1 = t^5 e_2 e_2^\top$, $S_2 = 0$. The steps $s_0 = (499/500)T_0$, $s_1 = (999/1000)T_1$, and $s_2 = T_2$ are feasible and attain the respective lower bounds in Theorem 4.1. When $T_\rho > 0$, the first two gains are at least $864/1003$ and $936/1003$ of their bounded factor optima.

For a sharper scalar certificate, let $A_\rho(t) = B(t, E_\rho(t))$ and $R_\rho(t) = \delta t - \Gamma_\delta(t, S_\rho(t))$, which is independent of $\delta$. The feasible interval ends at

$$t_\rho^* = \max\{t \in [0, 1/8] : \mathcal{J}(t) + A_\rho(t) \le b,\ A_\rho(t) \le \beta\}. \tag{16}$$

Its gain $\delta t_\rho^* - R_\rho(t_\rho^*)$ is no worse than the contracted witness, and $t_2^* = \mathcal{J}_+^{-1}(b)$ is exact. A finite search returns a validated feasible lower bracket $\widehat{t}_\rho$, checked by exact or rigorously enclosed comparisons; if $0 \le t_\rho^* - \widehat{t}_\rho \le \epsilon_t$, the gain loss is at most $\delta\epsilon_t$ (Appendix G.5).

**What the rank decision certifies.** A feasible gain at least $g$ makes that rank sufficient; an all action upper bound strictly below $g$ makes it impossible. The gap between these tests remains undecided. Minimality requires excluding all smaller ranks, as in Table 2; the classes are nested by zero padding factors. Appendix G proves every bound and the table's rational comparisons; a direct search places the optimum between 0.96 and 1.00 of $\delta T_0$ and between 0.98 and 1.00 of $\delta T_1$ at $b = 1/5000$, depending on which budget binds (Appendix L). The orders $(2, 5)$ in (9) explain the two audit caps: the least entry order is two and $\det C = t^7$. These decisions require the stated ties and direct output adapter; a fixed perturbation of one path changes the orders (Section 8).

## 5 What decides the gain at fine audits

The rank decisions of Section 4 depend on the singular value orders $(2, 5)$ of $C(t)$. The order five cannot be read from the second order terms of $C(t)$. We examine the role of these orders in a conditional logit module, then the limits of second order information in the cubic score graphs of Appendix I.

The module allows $m$ contexts and $n$ hidden signs, a real analytic displacement $C(t) \in \mathbb{R}^{m \times n}$ for small edit sizes $t$, and any correction $E$ of rank at most $\rho$, with the uniform reference, the channel reporting $(c, \ell)$ and a target mean $0 < \delta < 1$ as before; its law, costs and gain are (10) and (12) with $1/m$ in place of $1/2$ and $n$ hidden signs (Lemma F.1). It assumes $C(t) = O(t^2)$, so that the worst world penalty on the hidden residual is of higher order than the reward $\delta t$; the tied network satisfies this because every path to a hidden logit passes through at least two shared edges. Write $G_\rho(b, \beta)$ for the module's value (6); the value $\mathsf{G}_\rho$ of Section 4 adds factor caps and a finite radius.

Let $\tau_\rho^2(t) = m^{-1} \sum_{j>\rho} \sigma_j(C(t))^2$ be the Frobenius tail of $C(t)$ beyond rank $\rho$. When it is not identically zero, the analytic eigenvalue branches of $C(t)^\top C(t)$ that Rellich's theorem provides yield, as $t \downarrow 0$, $\tau_\rho^2(t) = a_\rho^2 t^{2\nu_\rho}(1 + O(t))$ with $a_\rho > 0$ and an integer $\nu_\rho \ge 2$ (Proposition F.2).

**Theorem 5.1** (The frontier under a rank constraint)**.** *For a nonzero tail, put* $T_\rho(b,\beta) = \min\{\sqrt{2b}, (2\beta/a_\rho^2)^{1/(2\nu_\rho)}\}$*. For all sufficiently small* $b > 0$*, with a threshold and remainder constant uniform in* $\beta \geq 0$*,*

$$G_\rho(b,\beta) = \delta T_\rho(b,\beta) + O(T_\rho(b,\beta)^2), \tag{17}$$

*and in particular* $G_\rho(b,0) = 0$*. If instead* $\operatorname{rank} C(t) \leq \rho$ *for all small positive* $t$*, then over the same range of* $b$ *and every* $\beta \geq 0$*, including* $\beta = 0$*,*

$$G_\rho(b,\beta) = \delta \mathcal{J}_+^{-1}(b) - \log\cosh\big(\mathcal{J}_+^{-1}(b)\big) = \delta\sqrt{2b} - b + O(b^{3/2}). \tag{18}$$

Appendix F proves the theorem. Along $\beta = \lambda b^\alpha$, the exponent is $\gamma_\rho(\alpha) = \max\{1/2, \alpha/(2\nu_\rho)\}$, so rank matters for the exponent only through the order $\nu_\rho$ of the remaining tail, and only once the demanded precision exceeds it. At the critical precision $\alpha = \nu_\rho$ the leading constant matters as well, since the surviving fraction of the unconstrained gain is

$$\lim_{b\downarrow 0} \frac{G_\rho(b, \lambda b^{\nu_\rho})}{\delta\sqrt{2b}} = \min\left\{1, \left(\frac{\lambda}{2^{\nu_\rho - 1} a_\rho^2}\right)^{1/(2\nu_\rho)}\right\}. \tag{19}$$

With $\mu_1 \leq \cdots \leq \mu_{\rho_{\max}}$ the vanishing orders of the nonzero singular values and $\rho_{\max}$ the generic rank of $C$, the smallest rank that keeps a positive limiting fraction of $\sqrt{b}$ is $\rho_{\min}(\alpha) = \#\{j : \mu_j < \alpha\}$ (Corollary F.3).

These orders follow from the displacement alone: $\mu_j = \Delta_j - \Delta_{j-1}$ with $\Delta_0 = 0$ and $\Delta_j$ the least vanishing order among the nonzero minors of size $j$, the order of the $j$th determinantal divisor (Gantmakher, 1998); Proposition F.2 supplies the constants $a_\rho$. For a rational polynomial $C$, enumerating nonzero minors gives $V_\rho = \{(1/2, 0), (0, 1/(2\mu_{\rho+1}))\}$ for $\rho < \rho_{\max}$ and $V_\rho = \{(1/2, 0)\}$ for $\rho \geq \rho_{\max}$, and so decides the endpoint test of Corollary 3.3: $\gamma_\rho$ has a terminal plateau exactly when $\rho \geq \rho_{\max}$. Corollary F.4 gives the algorithm, and the same minors determine $a_\rho^2$ exactly but give no remainder bound or validity radius.

Compared with expressivity results for low rank adaptation (Zeng & Lee, 2024), which rest on Eckart–Young truncation (Eckart & Young, 1936), and with rank wise spectral tails (Conangla Planes, 2026), our increment is a threshold rank set by the vanishing orders of the tail relative to the audit exponent $\alpha$, together with an exact retention constant (Corollary F.3). In that module, when the $t^2$ coefficient of $C(t)$ has full rank, every order equals two: ranks below $\min(m, n)$ share one exponent and differ only through the constant $a_\rho^2$ in (19), which each added rank lowers. The orders $(2, 5)$ of Section 4 come from its exact ties.

The cubic construction shows what second order information cannot decide. Theorem I.1 encodes a 3SAT formula in scores whose hidden coefficients are cubic in visible parameters $(t, z)$ and all vanish at $t > 0$ exactly when $z/t$ is a satisfying assignment. Satisfiable formulas therefore keep gain of order $\sqrt{b}$ at zero audit; otherwise every action with $t > 0$ has a coefficient of size at least $3t^3/16$ and the audit caps the gain at order $\beta^{1/6}$. Deciding the zero audit endpoint is thus NP hard, as is approximating the value within a factor two at explicit rational budgets; this is not a hardness claim for the tied model's scalar witnesses. Without any complexity assumption, the same instances limit every rule that sees only a second order jet.

**Corollary 5.2** (Second order information does not fix the value)**.** *Fix numbers of variables and clauses for which satisfiable and unsatisfiable formulas both exist, for example four and eight, and let* $d$ *be one plus twice the number of variables plus the number of clauses, so that the scores live on* $2d$ *states (*$d = 17$ *in the example). The families of a satisfiable and an unsatisfiable formula share the reference, target, channel, parameter cube and second order jet at the reference, yet at* $b = 2^{-60} d^{-6}$ *and* $\beta = b^4$ *the value of the first exceeds* $200$ *times the value of the second, which is positive. Every rule whose output depends only on these shared data and the budgets therefore misses one of the two values by a factor above* $14$*, and along* $\beta = b^4$ *no fixed factor suffices at all small budgets.*

Appendix I proves the corollary. Methods that protect behavior during an update restrict it to directions orthogonal to earlier gradients (Farajtabar et al., 2020), to the exact null space of preserved keys in one layer (Fang et al., 2025), or to regions of low Gauss–Newton curvature (Ikram et al., 2026). The corollary does not evaluate those constraints; it shows that under a fine audit the attainable gain can depend on third order terms of the family of available updates, which no second order model of that family determines.

## 6 CERTIFICATES FROM OBSERVED DATA

In the tied model, observed target signs sharpen the prior certificate $\delta \geq 1/2$: with $N$ signs, one event of probability at least $1 - \eta$ gives every action, rank and pair of budgets, including those chosen with the same data, a lower certificate within $2t\sqrt{2\log(2/\eta)/N}$ of its worst world gain (Appendix G.7).

For an arbitrary finite known channel, the empirical fiber certificate instead covers every positive action by one observed law event. Theorem H.1 in Appendix H bounds the certificate loss by $C\varepsilon_n\sqrt{D(q)}$ only for $D(q) \leq p_{\min}^2/8$; the radius free version uses the oscillation of $\log(q/p_0)$. In the tied model $p_{\min} = 1/16$, so that radius is $1/2048$.

## 7 THE AUDIT APPLIED TO LANGUAGE MODEL REVISIONS

We pose the introduction's interface to language models and compute the quantities of Section 2 for their revisions. A model states its reference law $p_0$ in percent for four cases on a morning in one of twenty contexts (ten cities, two seasons): rain overnight and the sprinkler ran (A), rain only (B), the sprinkler only (C), and neither (D). The record reveals only whether the ground was wet, the partition channel with groups $\{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$ and $\{\mathrm{D}\}$. In a second turn the model is told that the ground was wet on $W\%$ of such mornings, its own stated wet percentage shifted by $-30$, $-15$, 15 or 30 points and kept within $[5, 95]$, and it states a revised law $q$. By (22), the audit $\mathcal{B}(q)$ is the revised wet frequency times the divergence of the revised split of a wet morning from the reference split; Jeffrey's update $\pi_r$ is the only revision that matches the record with zero audit.

When a reviser may state any law, a positive audit is a loss in the worst world criterion, the classical optimality of Jeffrey conditioning for partition observations (Jeffrey, 1990; Williams, 1980; Diaconis & Zabell, 1982) in its robust Bayes form under log loss (Grünwald & Dawid, 2004): by (5) and (83), Jeffrey's update at a revision's own wet frequency costs $\mathcal{B}(q)$ less trust and has worst world gain $\Gamma_{\mathrm{obs}}(q; r) \geq \Gamma(q; r)$. Sections 3 and 4 concern action sets that may exclude it, where positive gain can require a positive audit, as at ranks zero and one of Theorem 4.1.

Five OpenAI models revised their laws for twenty contexts, four shifts and three instructions: the record alone, a request to change as little else as possible, or a remark that the record does not say whether rain, the sprinkler or both wet the ground with a request to keep the proportions of A, B and C, which is a request for Jeffrey's update. Four models run at temperature zero and the reasoning model `gpt-5-mini` at its default, giving 1188 revisions under the main wording and 800 under two rewordings of the record alone (Appendix K). Because the four temperature zero models write whole percents, a revision counts as zero audit when its audit is at most that of Jeffrey's update so rounded, which is at most $0.0052$.

The four models at temperature zero change the unidentified split in almost every revision: 634 of their 640 revisions with the record alone or the request to change little exceed the rounding level, and 514 exceed $0.01$. Even among the 263 revisions within five points of the record, 260 exceed the rounding level and 216 have negative worst world gain, whereas Jeffrey's update has worst world gain $\mathrm{KL}(r\|Kp_0) > 0$, with median $0.09$ nats over all records (Figure 2). The reasoning model differs: with the record alone it keeps the reference split in 66 of 76, 79 of 80 and 58 of 80 under the three wordings (Jeffrey's update itself, within one point of the record, in 60, 76 and 49).

The record does not fix the direction of change either. When it showed more wet mornings than the model had stated, `gpt-4.1` lowered the probability that a wet morning had rain in 39 of 40 revisions, by $0.19$ on average, and raised it in 37 of 40, by $0.21$, when the four cases were listed in reverse order; across models and wordings the mean change ranges from $-0.19$ to $0.28$ (Table 6). These effects fit known incoherence of stated probabilities (Zhu & Griffiths, 2024), sensitivity to option order (Pezeshkpour & Hruschka, 2024; Zheng et al., 2024) and departures of stated belief updates from Bayes' rule (Chen et al., 2026; Imran et al., 2025) in language models; the audit locates them in the unidentified part of a revision.

Asking for Jeffrey's update does not secure it (Table 5). Under that request the five models keep the reference split in 234 of 396 revisions (`gpt-4.1` in only 18 of 80) but return Jeffrey's update itself in only 97; the request to change as little as possible keeps the split in 52 of 396.

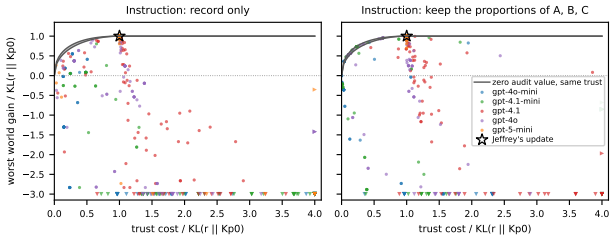


Figure 2: Revisions under the main wording against the zero audit value. Trust cost and worst world gain are divided by $\mathrm{KL}(r\|Kp_0)$, so Jeffrey's update sits at the star; the grey band is the range over records of the largest worst world gain that a zero audit revision certifies at the same trust cost. Triangles on the lower edge mark gains below $-3$, including $-\infty$; those on the right edge mark trust costs above $4$.

**Sampled laws, output laws and parameter edits.** Three further measurements each relax one feature of the main design (Appendix K). With ten sampled answers per prompt at temperature one, their mean law has audit above $0.005$ in 59 to 79 of 80 revisions per model, slightly fewer than the greedy answer (64 to 80), and the ten samples move the split the same way in only 30 to 61 of 80.

On an open weight model (Qwen2.5-1.5B), a four state law read from the next token probabilities of three yes/no questions moves a median 52% of the way toward records showing fewer wet mornings and at most 9% toward records showing more. It spends a median 6% to 8% and 57% to 67% of its trust cost on audit in the two cases and lowers $P(\text{rain} \mid \text{wet})$ in 80 of 80 revisions.

A LoRA of rank one or eight on its last MLP block, trained to match the record in all twenty contexts, has audit above $0.005$ in all 240 edited laws at shifts of $\pm 0.30$ without a KL penalty and mostly at the rounding level with KL weight $0.1$ (23 to 30 of 240 above it, all at $+0.30$); the ranks do not differ. Along the trained edits the unidentified split moves at first order in the edit size (fitted orders 0.99 to 1.01), not at second order as the module of Section 5 assumes; the drag across contexts is nearly rank one, and a shared rank one SVD correction of the output conditional log law removes 99.5% of its audit.

SmolLM2-1.7B under the same protocol has fitted drag orders 0.995 to 1.004. Without a KL penalty, all 240 output laws at the two positive shifts (two ranks, three seeds, twenty contexts) have audit above $0.005$. Across all four shifts without the penalty, the output correction lowers each rank's median context averaged audit by about 98% (Appendix K). Prompted revisions lower the wet frequency to 3% to 6% across all tested records.

These measurements test no theorem. The five OpenAI models share one provider and state laws on one four state interface. The two open weight models are small and edited only in their last block. Uncertainty statements are limited to intervals from resampling the twenty contexts (Appendix K). The counts show that the audit is computable for these models, whose revisions fill in what the record cannot identify in ways that depend on the model, instruction, wording and edit.

## 8 Scope and Outlook

The finite rank decisions concern a designed network with exact weight ties, known routing and a direct output adapter, and establish no general LoRA law. Adding a fixed nonzero $\epsilon t^2 e_2 e_2^\top$ changes the two asymptotic singular orders to $(2, 2)$; negative perturbations can also create finite rank drops (Appendix G.8).

Each result has its own range: the radius $1/512$ belongs only to Theorem 4.1; the envelope and the rank law have nonexplicit thresholds. The certificate's $\sqrt{D}$ bound needs $D \le p_{\min}^2/8$, and constants in $H_K$ and $1/p_{\min}$ can be vacuous. The target refinement of Appendix G.7 concerns one mean in a known family, not hidden factors or an unknown channel. Generated labels treated as error free can reverse a certificate (Appendix J.3). Section 7 finds approximately first order drag over the measured scales in both adapter families, outside the module of Section 5.

## AI USE STATEMENT

Large language models (LLMs) assisted with developing the theoretical framework, drafting and auditing proofs, searching the literature, language editing (e.g., grammar, spelling, and word choice), implementing standard computational methods, preparing visualizations for submission, designing and executing the experiments, and simulating reviewer feedback. LLMs also served as objects of measurement: five OpenAI model endpoints each stated a probability law over four weather cases for twenty contexts and a revised law after a record of how often the ground was wet, and two open weight models were read out through the next token probabilities of three yes/no questions and edited with LoRA adapters on their last MLP block; the stated and output laws formed the measured revisions.

LLMs did not conceive the research question. The authors made the final research decisions, interpreted the results, and take full responsibility for the manuscript.

## REFERENCES

Mikkel Abrahamsen, Linda Kleist, and Tillmann Miltzow. Training neural networks is er-complete. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 18293–18306. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/9813b270ed0288e7c0388f0fd4ec68f5-Paper.pdf.

Daniele Alessandrini. Logarithmic limit sets of real semi-algebraic sets. *Advances in Geometry*, 13 (1):155, 2013.

Miki Aoyagi and Sumio Watanabe. Stochastic complexities of reduced rank regression in bayesian estimation. *Neural Networks*, 18(7):924–933, 2005.

Giovanni Barbarino and Vanni Noferini. On the rellich eigendecomposition of para-hermitian matrices and the sign characteristics of$*$-palindromic matrix polynomials. *Linear Algebra and its Applications*, 672:1–27, 2023.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander Nicholas D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=u3U8qzFV7w.

Daniel Bertschinger, Christoph Hertrich, Paul Jungeblut, Tillmann Miltzow, and Simon Weber. Training fully connected neural networks is $\exists\mathbb{R}$-complete. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=H15KtcyHvn.

Edward Bierstone and Pierre D Milman. Semianalytic and subanalytic sets. *Publications Mathématiques de l'IHÉS*, 67:5–42, 1988.

Tijs Buggenhout, Mathias Stout, and Lisa Vandebrouck. Approximating parametric suprema for constructible and power-constructible functions. *arXiv preprint arXiv:2602.23126*, 2026.

Chacha Chen, Matthew Jörke, Adam Goliński, Masha Fedzechkina, Guillermo Sapiro, Sinead Williamson, and Nicholas Foti. Llms are not (consistently) bayesian: Quantifying internal (in) consistencies of llms' probabilistic beliefs. *arXiv preprint arXiv:2605.06915*, 2026.

Raf Cluckers and Daniel J Miller. Lebesgue classes and preparation of real constructible functions. *Journal of Functional Analysis*, 264(7):1599–1642, 2013.

Gerard Conangla Planes. How much rank does lora need? rank-error bounds for transformer attention. *arXiv e-prints*, pp. arXiv–2608, 2026.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pp. 146–158, 1975.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

Persi Diaconis and Sandy L Zabell. Updating subjective probability. *Journal of the American Statistical Association*, 77(380):822–830, 1982.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations*, volume 2025, pp. 16366–16396, 2025.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International conference on artificial intelligence and statistics*, pp. 3762–3773. PMLR, 2020.

F R (feliks Ruvimovich) Gantmakher. *The theory of matrices, volume 1*. AMS Chelsea Publishing. American Mathematical Society, Providence, RI, September 1998.

Peter D. Grunwald and Joseph Y. Halpern. Updating probabilities, 2003. URL https://arxiv.org/abs/cs/0306124.

Peter D. Grünwald and A. Philip Dawid. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *The Annals of Statistics*, 32(4), August 2004. ISSN 0090-5364. doi: 10.1214/009053604000000553. URL http://dx.doi.org/10.1214/009053604000000553.

Peter Hase, Thomas Hofweber, Xiang Zhou, Elias Stengel-Eskin, and Mohit Bansal. Fundamental problems with model editing: How should rational belief revision work in LLMs? *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=LRf19n5Ly3.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Alan J Hoffman. On approximate solutions of systems of linear inequalities. In *Selected Papers Of Alan J Hoffman: With Commentary*, pp. 174–176. World Scientific, 2003.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Zarif Ikram, Arad Firouzkouhi, Stephen Tu, Mahdi Soltanolkotabi, and Paria Rashidinejad. Crispedit: Low-curvature projections for scalable non-destructive llm editing. *arXiv preprint arXiv:2602.15823*, 2026.

Sohaib Imran, Ihor Kendiukhov, Matthew Broerman, Aditya Thomas, Riccardo Campanella, Rob Lamb, and Peter M Atkinson. Are llm belief updates consistent with bayes' theorem? *arXiv preprint arXiv:2507.17951*, 2025.

Richard C Jeffrey. *The logic of decision*. University of Chicago press, 1990.

Kiumars Kaveh and Peter Makhnatch. Invariant factors as limit of singular values of a matrix. *Arnold Mathematical Journal*, 8(3):561–571, 2022.

Tomasz Korbak, Ethan Perez, and Christopher Buckley. RL with KL penalties is better viewed as bayesian inference. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1083–1091, Stroudsburg, PA, USA, 2022. Association for Computational Linguistics.

Jean-Marie Lion and Jean-Philippe Rolin. Théoreme de préparation pour les fonctions logarithmico-exponentielles. In *Annales de l'institut Fourier*, volume 47, pp. 859–884, 1997.

S Lojasiewicz and M Coste. *Ensembles semi-analytiques*. univ-rennes.hal.science, 1965.

Charles F Manski. *Partial identification of probability distributions*. Springer, 2003.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

K G Murty and S N Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 39(2):117–129, 1987.

Lucas Monteiro Paes, Natalie Mackraz, Barry-John Theobald, and Federico Danieli. Theoretical limits of language model alignment, 2026. URL https://arxiv.org/abs/2605.07105.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024.

Franz Rellich. Störungstheorie der spektralzerlegung: I. mitteilung. analytische störung der isolierten punkteigenwerte eines beschränkten operators. *Mathematische Annalen*, 113(1):600–619, 1937.

Zhe Tao and Aditya V Thakur. Provable editing of deep neural networks using parametric linear relaxation. *Advances in Neural Information Processing Systems*, 37:113846–113883, 2024.

Thijs van Ommen, Wouter M. Koolen, Thijs E. Feenstra, and Peter D. Grünwald. Robust probability updating. *International Journal of Approximate Reasoning*, 74:30–57, 2016. ISSN 0888-613X. doi: 10.1016/j.ijar.2016.03.001. URL http://dx.doi.org/10.1016/j.ijar.2016.03.001.

Sumio Watanabe. *Algebraic geometry and statistical learning theory*, volume 25. Cambridge university press, 2009.

Peter M Williams. Bayesian conditionalisation and the principle of minimum information. *The British Journal for the Philosophy of Science*, 31(2):131–144, 1980.

Joy Qiping Yang, Salman Salamatian, Ziteng Sun, Ananda Theertha Suresh, and Ahmad Beirami. Asymptotics of language model alignment, 2024. URL https://arxiv.org/abs/2404.01730.

Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. In *International Conference on Learning Representations*, volume 2024, pp. 5078–5123, 2024.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, volume 2024, pp. 19426–19454, 2024.

Jian-Qiao Zhu and Thomas L Griffiths. Incoherent probability judgments in large language models. *arXiv preprint arXiv:2401.16646*, 2024.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

## A Related work

**Robust log score improvement and information projection.** Both starting objects are classical. Information projection minimizes divergence under linear constraints (Csiszár, 1975); under logarithmic loss, its solution is also a robust Bayes equalizer whose expected relative loss is constant across the constraint set (Grünwald & Dawid, 2004, Sections 7.3 and 10.2.2). Local geometry in these coordinates is standard in machine learning through the natural gradient (Martens, 2020). Here we maximize worst case gain over an observational fiber while restricting updates to a possibly singular action set and shrinking both an exact projection audit budget and a trust budget. In partial identification terms (Manski, 2003), the fiber is the identification region of the latent law, and the

gain is a worst case over that region. Allowing every positive prediction and a sufficiently large trust budget recovers the classical unconstrained endpoint (Appendix J). For a partition observation, the projection reduces to Jeffrey conditioning (Jeffrey, 1990; Williams, 1980; Diaconis & Zabell, 1982).

Worst case updating after a coarse observation has a different observation model. van Ommen et al. (2016) make the unknown coarsening mechanism adversarial and characterize logarithmic loss optimality by a reverse coarsening at random condition. Grunwald & Halpern (2003) characterize when Jeffrey conditioning is appropriate through a generalized coarsening at random condition, without an analogous characterization for minimum relative entropy updating. Those observations are sets containing the outcome. Ours are outputs of a known stochastic channel, and two budgets constrain the update. The earlier characterizations therefore do not transfer directly.

**Editing with a preservation constraint.** Gradient subspaces, null spaces and low curvature regions are established ways to protect behavior during an edit (Farajtabar et al., 2020; Fang et al., 2025; Ikram et al., 2026). Our audit expresses the preservation concern as divergence from an information projection. Section 3 and Appendix F show that fine audit tolerances can make attainable gain depend on terms beyond a fixed local approximation. Orthogonal gradient descent uses gradients, AlphaEdit a preservation null space, and CrispEdit a second order Gauss–Newton constraint. Our examples neither equate these approximations nor show failure of those methods. Hase et al. (2024) argue that model editing is a belief revision problem whose formulation remains open. The finite interface here addresses one part of that question: how much an update commits beyond what its observations determine.

Another line quantifies what a KL budget buys. The optimal reward gain under a KL penalty equals the penalty coefficient times a Jeffreys divergence and has a base law covariance representation (Paes et al., 2026). Best-of-$n$ selection has KL and win rate bounds (Beirami et al., 2025). Korbak et al. (2022) interpret a KL penalized objective as Bayesian inference, with a tilted base law optimum. Yang et al. (2024) give the KL constrained optimum in closed form and show that methods with a comparable KL–reward tradeoff must approximate it in relative entropy. These output law KL constraints, and the KL penalty in preference fine tuning (Ziegler et al., 2020), bound our finite interface trust cost by data processing (Section 2). They score updates against a given reward. Our objective instead takes the worst world consistent with a partial observation.

**Tame geometry.** The geometric tools are inherited. Uniformization, normal crossing resolution and the subanalytic Łojasiewicz inequality originate with Hironaka and are presented in the form used here by Bierstone & Milman (1988). Preparation and finite descriptions of parametric suprema for constructible and power-constructible functions appear in Lion & Rolin (1997); Cluckers & Miller (2013); Buggenhout et al. (2026). Alessandrini (2013, Theorem 3.11) proves polyhedrality of logarithmic limit sets in regular polynomially bounded o-minimal structures, the exponent level structure behind Corollary 3.2.

For a power-constructible function on a subanalytic domain, Theorem A of Buggenhout et al. (2026) (their Theorem 6.1) compares its fiberwise bounded absolute supremum with a finite maximum of absolute values of power-constructible functions, using one constant for all parameters. For real exponent fields the comparison functions can be evaluations at feasible subanalytic selectors. Applied to $\Gamma_+$ on our compact budget feasible fibers, it already supplies budget ratio uniformity and feasible near optimal updates. As stated, however, its comparison functions need not be pure monomials, and a logarithmic limit set records exponents rather than constant level comparison.

Theorem 3.1 gives the pure max–min monomial form for these nested sublevel sets. Its witnesses come from two cost linear programs and fixed contractions in resolved charts, so the constants cover all small budget pairs, including audit tolerances below every power of the trust budget. Theorem D.1 describes the exponents by analytic arcs and gives a sharp leading term for rational schedules. Corollary 3.3 uses constant level uniformity to send $\beta$ to zero at fixed $b$. It identifies the terminal plateau with positive gain on the actual zero audit action set, which a leading Taylor kernel can misrepresent. These are specialized consequences; we prove no new preparation theorem.

**Singular learning theory.** Resolution also yields rational exponents in singular learning theory (Watanabe, 2009). There, normal crossings of the model's KL function produce a positive rational learning coefficient governing Bayes free energy and generalization error asymptotics. For reduced

rank regression, that coefficient is known exactly as a function of the true rank and dimensions (Aoyagi & Watanabe, 2005). It is an exponent of an integral over parameters; in nondegenerate cases a Newton polyhedron gives it through a linear program. Our exponents describe a maximum under two shrinking constraints. They form a spectrum indexed by the relative rate of the budgets, whose endpoint identifies positive zero audit gain. The rank in Appendix F is a correction rank acting through singular value vanishing orders, rather than the rank of a true parameter matrix.

**Rank of an adaptation.** Low rank adaptation and its expressivity motivate the correction question (Hu et al., 2022; Zeng & Lee, 2024). The 2026 preprint of Conangla Planes (2026) brackets attainable attention divergence at each rank by a task weighted spectral tail. Both analyses make a tail relevant to rank, but our constraint is an audit budget and our payoff is a worst world guarantee. The tail's vanishing order determines the resulting precision exponent. For invertible square Laurent series matrices, Kaveh & Makhnatch (2022, Theorem 1.1) identify singular value orders with invariant factors. Our rectangular analytic calculation also keeps leading constants; its determinantal divisor description is classical (Gantmakher, 1998).

Appendix F treats a prescribed conditional logit displacement with rank limited coefficient corrections; Section 4 derives such a displacement from shared path weights and a two layer output adapter $\Delta W = VU^\top$, where $\rho$ bounds the adapter increment, not the full parameter update. Its finite exclusions concern that designed architecture and establish no rank law for arbitrary LoRA insertions, trained networks or learned feature maps.

**Complexity of editing and of training.** Exact network training is complete for the existential theory of the reals (Abrahamsen et al., 2021; Bertschinger et al., 2023), and provable editing has been studied using relaxations (Tao & Thakur, 2024). Second order hardness also has classical precedents: deciding that a feasible point of an indefinite quadratic program is not a local minimum is NP complete (Murty & Kabadi, 1987). Theorem I.1 instead constructs instances sharing the reference, target, channel and second order jet. Their hardness lies in the cubic term and cannot be attributed to differences in second order local geometry. This construction is not a hardness claim for the explicit scalar witness search in Section 4.

## B Information projection and the relative audit expansion

**Lemma B.1** (Local analytic costs and relative audit error)**.** *The functions $D$ and $B$ are real analytic near zero. If $h_\perp$ is the component of $h$ in $L^\perp$, then*

$$D(h) = \tfrac{1}{2}\|h\|_0^2 + O(\|h\|_0^3), \qquad \psi(h) = \tfrac{1}{2}\|h\|_0^2 + O(\|h\|_0^3), \tag{20}$$

$$B(h) = \tfrac{1}{2}\|h_\perp\|_0^2 + O(\|h\|_0\|h_\perp\|_0^2). \tag{21}$$

*No assumption of full row rank on $K$ is required.*

### B.1 Exact identities

The set defining $\pi_{Kq}$ is compact and contains a strictly positive point $q$. Strict convexity gives uniqueness. Its minimizer is positive: if one coordinate vanished, moving a small distance toward $q$ would give a negative infinite directional derivative from that coordinate. At an interior minimizer the first order conditions put $\log(\pi_{Kq}/p_0)$ in $\mathrm{range}(K^\top)$, including the constant component. Its centered score therefore belongs to $L$.

Conversely, for $\ell \in L$ and $v$ with $Kv = Kq_\ell$,

$$\sum_i (v_i - q_\ell(i)) \log \frac{q_\ell(i)}{p_0(i)} = 0.$$

Subtracting the KL expressions gives

$$\mathrm{KL}(v\|p_0) = \mathrm{KL}(v\|q_\ell) + \mathrm{KL}(q_\ell\|p_0).$$

This proves that $q_\ell$ is the projection for its own observation, the identity (5), and the exact equivalence $B(h) = 0 \iff h \in L$. In a partition observation the projection is $\pi_{Kq}(i) = q_Y(T(i))p_0(i \mid T(i))$,

which gives

$$B(h) = \sum_y q_h(y)\,\mathrm{KL}\big(q_h(\cdot \mid y)\|p_0(\cdot \mid y)\big) \tag{22}$$

by direct substitution.

### B.2 Analyticity without a rank assumption on the channel

Work on the effective score space $L$ and define

$$A(\ell) = P_L(q_\ell/p_0 - \mathbf{1}), \qquad \ell \in L.$$

This map is analytic and $DA(0) = I_L$, because $D(q_h/p_0)|_{h=0}[u] = u$ for centered $u$. Its local analytic inverse therefore exists. Equality of the $L$ moments is equivalent to equality of observations: subtracting its reference mean from any row of $K$ gives an element of $L$. Consequently the score of the information projection is

$$a(h) = A^{-1}(P_L(q_h/p_0 - \mathbf{1}))\,, \qquad Da(0) = P_L. \tag{23}$$

No redundant observation coordinate is inverted. In particular $D$ and $B$ are analytic near zero.

### B.3 Proof of Lemma B.1

The log partition function satisfies $\psi(0) = D\psi(0) = 0$ and $D^2\psi(0)[u,v] = \langle u, v\rangle_0$. Taylor expansion gives its stated quadratic term. Since $D(h) = \mathbb{E}_{q_h} h - \psi(h)$, its quadratic term is the same. From the exact projection identity and (23),

$$B(h) = \tfrac{1}{2}\|h\|_0^2 - \tfrac{1}{2}\|P_L h\|_0^2 + O(\|h\|_0^3).$$

This identifies the Hessian at zero but is not the relative remainder used for fine audits.

Instead write $b(\ell, n) = B(\ell + n)$ for $\ell \in L$, $n \in L^\perp$. The function is analytic, is nonnegative, and satisfies $b(\ell, 0) = 0$ on the whole observable family. Thus $D_n b(\ell, 0) = 0$. The preceding Hessian calculation gives $D^2_{nn} b(0,0) = I_{L^\perp}$. Uniformly for $0 \le s \le 1$,

$$D^2_{nn} b(\ell, sn) = I_{L^\perp} + O(\|\ell\|_0 + \|n\|_0)$$

in operator norm. Taylor's integral formula in the normal variable now yields

$$b(\ell, n) = \int_0^1 (1-s) D^2_{nn} b(\ell, sn)[n,n]\, ds = \tfrac{1}{2}\|n\|_0^2 + O(\|\ell + n\|_0 \|n\|_0^2).$$

Orthogonality makes $\|\ell\|_0 + \|n\|_0$ comparable to $\|\ell + n\|_0$, and $n = h_\perp$ for $h = \ell + n$, completing the proof.

## C A finite certificate for both budgets

We give the geometric reduction and the optimization argument separately. The distinction matters for a lower bound: a useful coordinate chart must map its entire domain into the action model. A collection of candidate arcs, or a monomial formula on a domain with unresolved inequalities, would not establish the theorem.

Throughout this appendix, take the compact local representative $\mathcal{M}$ used in the main text and write

$$D(h) = \mathrm{KL}(q_h\|p_0), \qquad F(h) = \max\{\Gamma(h), 0\}.$$

The functions $D, B, F$ are continuous and subanalytic on $\mathcal{M}$; the first two are restrictions of analytic functions. They satisfy $D^{-1}(0) = \{0\}$ and $B(0) = F(0) = 0$. Since the reference is feasible, replacing $\Gamma$ by $F$ leaves every optimal value unchanged. These elementary properties, rather than a regular parameterization, are the hypotheses used below.

The compact representative introduces no extra asymptotic assumption. Intersect a closed local model with a small closed score ball inside the domain of the analytic costs; where the ball's boundary meets the model the trust cost has a positive minimum, so small budgets never reach it and a smaller representative leaves all small budget values unchanged. Compactness is a local attainment device, not a bound on a redundant factorization; we optimize action scores, the points on the joint graph.

**Parameterized models.** Let an output law $Q_\vartheta$ depend on parameters $\vartheta$ in an open subset of a Euclidean space, and suppose that its finite pushforward $q_\vartheta$ is real analytic on an open neighborhood $U$ of $\vartheta_0$, where $q_{\vartheta_0} = p_0$. This holds, for example, when $Q_\vartheta$ is a softmax over finitely many outputs whose logits are real analytic in $\vartheta$, since each $q_\vartheta(i)$ is then a finite sum of such probabilities. After shrinking $U$, the pushforward is positive on it and the centered score $h_\vartheta = \log(q_\vartheta/p_0) - \mathbb{E}_{p_0} \log(q_\vartheta/p_0)$ is real analytic there, with $h_{\vartheta_0} = 0$. A closed ball of parameters around $\vartheta_0$ inside $U$ is compact and semianalytic, and the image of a relatively compact semianalytic set under a real analytic map is subanalytic (Bierstone & Milman, 1988), so the scores of the ball form a compact subanalytic set containing zero. Once the radius is small, continuity puts this set inside the neighborhood of zero on which $D$ and $B$ are analytic, and it satisfies Assumption 2.1. Lemma F.1 applies the same image theorem to the rank constrained coefficients of Section F.1, whose parameter set is singular at the reference. The argument places such models under Theorem 3.1 and its corollaries; it neither computes their exponent sets nor supplies their constants.

*Proof of Theorem 3.1.* **A compact analytic source.** Form the joint graph

$$X = \{(h, D(h), B(h), F(h)) : h \in \mathcal{M}\}.$$

It is compact, closed, and subanalytic. The uniformization theorem (Bierstone & Milman, 1988, Theorem 0.1) supplies a real analytic manifold $N$ and a proper analytic map onto $X$. Write this map as

$$(\phi, \widetilde{D}, \widetilde{B}, \widetilde{F}) : N \longrightarrow X.$$

Its last three components are analytic and equal exactly to $D \circ \phi, B \circ \phi, F \circ \phi$. Thus the construction simultaneously makes the costs and positive gain analytic; no choice of an active worst world is needed. Properness and compactness of $X$ also imply that $N$ is compact. In particular, it has finitely many connected components.

Components on which $\widetilde{F}$ is identically zero contribute nothing to a positive optimum. On every other component, $\widetilde{D}$ is not identically zero: otherwise the action map would be identically zero and so would $\widetilde{F}$. If $\widetilde{B}$ is identically zero, apply the normal crossing theorem to $\widetilde{D}\widetilde{F}$. Otherwise apply it to $\widetilde{D}\widetilde{B}\widetilde{F}$. The relevant product is not identically zero on its connected component. Corollary 4.9 of Bierstone & Milman (1988) gives a proper surjective analytic pullback on which that product has local normal crossings. Taking the finite disjoint union of these sources gives a compact analytic manifold $P$, with an analytic action map $\phi : P \to \mathcal{M}$ whose image contains every positive gain action. Zero gain components need not be covered after they have been discarded.

A product having normal crossings makes each of its factors a coordinate monomial times a nonvanishing analytic unit. For completeness, remove from each factor its largest power of every coordinate. Coordinate orders add under multiplication, because convergent power series modulo a coordinate form an integral domain. After removing these powers the product is a unit, so each remaining factor is a unit. This proves the simultaneous assertion needed here without a further activity partition.

The normal crossing maps preserve the cover of all positive gain actions. Their sources are compact by properness; there are finitely many connected components, to which the argument applies separately.

**Full boxes and their boundary faces.** Let $Z_P = \phi^{-1}(0)$. About each point of this compact fiber choose a coordinate box $(-\varepsilon, \varepsilon)^n$, with closure inside the normal crossing neighborhood. Shrink it until the absolute values of all units have positive lower and upper bounds. These are full coordinate neighborhoods in a manifold mapping into the joint graph. Consequently the image of the entire box lies in the action model, including its coordinate faces.

Finitely many boxes cover $Z_P$. Their union contains the full preimage of every sufficiently small trust budget sublevel set. If not, compactness would give a sequence outside this open neighborhood, with $D \circ \phi \to 0$, and a subsequential limit in $Z_P$. That contradicts the definition of the neighborhood. Thus all small positive gain actions, including singular and boundary actions, are covered by finitely many boxes.

Split every box according to whether each coordinate is negative, zero, or positive. On a face substitute its zero coordinates and fix the remaining signs. Rescale the free absolute coordinates to $y \in (0, 1)^n$. Discard faces on which $F \circ \phi$ is zero. A face with $D \circ \phi = 0$ is automatically of

this kind, since its image is the reference. On each retained face there are nonzero integer vectors $a, d \in \mathbb{N}_0^n$ and fixed positive constants such that

$$c_D y^a \le D(\phi(y)) \le C_D y^a, \qquad c_F y^d \le F(\phi(y)) \le C_F y^d. \tag{24}$$

Either audit is identically zero on this face, or a nonzero $c \in \mathbb{N}_0^n$ satisfies

$$c_B y^c \le B(\phi(y)) \le C_B y^c. \tag{25}$$

The nonzero vectors follow because the corresponding nonidentical functions vanish at the original box center over the reference. Units remain bounded on the closure after restriction and rescaling. The free domain is the whole positive box. Faces may have different dimensions, and zero audit faces inside a nonzero audit component are explicitly retained.

**A uniform optimization certificate on one face.** For a positive audit face put $\Lambda_b = \log(1/b)$, $\Lambda_\beta = \log(1/\beta)$, taking both budgets below one, and consider

$$p(\Lambda_b, \Lambda_\beta) = \min_{w \ge 0} \{ d \cdot w : a \cdot w \ge \Lambda_b,\ c \cdot w \ge \Lambda_\beta \}.$$

The primal is feasible because $a, c$ are nonzero and nonnegative; it has a finite attained optimum. Its dual feasible set is

$$\mathcal{P} = \{ (u, v) \in \mathbb{R}^2_{\ge 0} : ua + vc \le d \}.$$

This is a nonempty compact rational polytope. A positive coordinate of $a$ bounds $u$, and a positive coordinate of $c$ bounds $v$. Its vertex set $V$ is finite and rational. Strong linear programming duality therefore yields

$$e^{-p(\Lambda_b, \Lambda_\beta)} = \min_{(u,v) \in V} b^u \beta^v =: H_i(b, \beta). \tag{26}$$

For an upper bound, every vertex satisfies $d - ua - vc \ge 0$ coordinatewise. Since $0 < y_j < 1$,

$$y^d \le (y^a)^u (y^c)^v.$$

Equations (24)–(25) then give, at every feasible point on this face,

$$F(\phi(y)) \le C_F c_D^{-u} c_B^{-v} b^u \beta^v.$$

Taking the minimum over the finitely many vertices and the largest of their fixed prefactors gives $F(\phi(y)) \le C_i H_i(b, \beta)$. In particular, the constant does not depend on the ratio of the budgets.

The minimum over all vertices handles the zero dual vector as well as active budget powers. The primal optimizer may be nonunique and may have zero coordinates; only one attained optimizer is needed.

For a matching lower bound choose an optimizer $w$ of the primal. Fix, once for this face, $0 < \kappa < 1$ so small that

$$C_D \kappa^{|a|} \le 1, \qquad C_B \kappa^{|c|} \le 1, \qquad |a| = \sum_j a_j.$$

Set $y_j = \kappa e^{-w_j}$. The complete box property makes this an available action. Both exact budgets hold, because

$$D(\phi(y)) \le C_D \kappa^{|a|} e^{-a \cdot w} \le b, \qquad B(\phi(y)) \le C_B \kappa^{|c|} e^{-c \cdot w} \le \beta.$$

Its gain is at least

$$c_F \kappa^{|d|} e^{-d \cdot w} = c_i H_i(b, \beta).$$

The contraction is independent of the optimizer and both budgets. Coordinates with zero weight remain at the allowed fixed value $\kappa$. Although they need not approach the chart center, the action image approaches the reference as $b \to 0$, because $D \le b$ and $D$ has a unique zero.

On a zero audit face remove the second primal constraint. Its dual is the compact interval $\{u \ge 0 : ua \le d\}$; embed its vertices as $(u, 0)$. The same estimates apply with no audit factor. Taking the maximum over all retained faces proves both uniform inequalities, since every relevant action is covered and each face provides a feasible lower bound action. Finiteness supplies global positive comparison constants.

Finally, if positive gain does not approach the reference, it vanishes on the action model in some neighborhood of zero. Small trust budgets lie there, proving the zero value alternative. Otherwise a lifted positive gain sequence has a limit in $Z_P$, so at least one retained face exists. Its recovery construction gives positive value for every sufficiently small pair of positive budgets. This also proves the claimed positivity dichotomy.

The finite atlas and fixed contraction choose all comparison constants before either budget. The same inequalities therefore cover schedules decaying faster than every power of the trust budget. Constructing the atlas efficiently is outside this existence argument. □

## D Power spectra, analytic arcs, and the value at zero audit

*Proof of Corollary 3.2.* For a fixed $\lambda > 0$, substitution of $\beta = \lambda b^{\alpha}$ in each finite envelope gives

$$H_i(b, \lambda b^{\alpha}) \asymp b^{\max_{(u,v)\in V_i}(u+\alpha v)}.$$

The factors $\lambda^v$ affect comparison constants only. The maximum over faces consequently has exponent

$$\gamma(\alpha) = \min_i \max_{(u,v)\in V_i} (u + \alpha v).$$

This formula holds for every real $\alpha > 0$, without a definability assumption on an irrational power function. Finitely many rational affine functions give a continuous, nondecreasing, piecewise affine function with rational coefficients and finite breakpoints. Finally, local Lipschitz continuity of $\Gamma$, its zero reference value, and $D(h) \asymp \|h\|_0^2$ imply $G(b, \beta) = O(\sqrt{b})$. Thus $\gamma(\alpha) \geq 1/2$. □

There is an intrinsic description that avoids choosing a particular resolution. Let $\mathcal{A}_+$ be the nonconstant real analytic score arcs extending analytically across $t = 0$, with $h(t) \in \mathcal{M}$ for $t \geq 0$ sufficiently small and $h(0) = 0$, with $\Gamma(h(t)) > 0$ for small $t > 0$. Along such an arc put

$$a_h = \operatorname{ord}_t D(h(t)), \qquad c_h = \operatorname{ord}_t B(h(t)), \qquad d_h = \operatorname{ord}_t \Gamma(h(t)). \tag{27}$$

We use $c_h = \infty$ when the audit is identically zero along the arc. The gain is eventually one analytic branch: it is a finite minimum of analytic world payoffs, whose pairwise differences have fixed signs on a sufficiently short positive arc.

Only nonnegative parameters must lie in the model. Composing with $t \mapsto t^2$ gives a two sided model arc when desired, without changing any order ratio.

**Theorem D.1** (Positive gain arcs and sharp rational schedules)**.** *In the positive value alternative,*

$$\gamma(\alpha) = \inf_{h\in\mathcal{A}_+} \max\left\{\frac{d_h}{a_h}, \frac{\alpha d_h}{c_h}\right\}, \qquad d_h/\infty = 0. \tag{28}$$

*For rational $\alpha > 0$, the infimum is attained by an analytic arc and, for each fixed $\lambda > 0$, there exist $A_{\alpha,\lambda} > 0$ and $\epsilon > 0$ such that*

$$G(b, \lambda b^{\alpha}) = A_{\alpha,\lambda} b^{\gamma(\alpha)} + O\big(b^{\gamma(\alpha)+\epsilon}\big). \tag{29}$$

*Proof of Theorem D.1.* Along a positive gain analytic action arc the robust score eventually equals one analytic world payoff branch. Indeed, each difference of two branches is analytic and is either identically zero or has a fixed sign for all sufficiently small positive parameters. The finite number of branches gives a common such interval. Hence the orders $a, c, d$ in the theorem are well defined; $c = \infty$ means audit is identically zero along the arc.

For any such arc write its nonzero leading terms as $D(h(t)) \sim At^a$, $B(h(t)) \sim B_0 t^c$, and $\Gamma(h(t)) \sim F_0 t^d$, with positive coefficients. Setting

$$t = \eta b^{\max\{1/a, \alpha/c\}}$$

and choosing a fixed sufficiently small $\eta > 0$ meets both budgets for every sufficiently small $b$. The zero audit case omits the second restriction. This yields a gain of the asserted arc order and shows that $\gamma(\alpha)$ is no larger than any arc expression. The construction covers a complete interval of budgets, not only a subsequence.

Conversely, choose a monomial face attaining the minimum spectral exponent, and an optimal normalized primal weight $w_*$ satisfying $a \cdot w_* \geq 1$, $c \cdot w_* \geq \alpha$. Add a small positive vector, then approximate by a nonnegative rational vector $p$, preserving both inequalities and changing $d \cdot w_*$ arbitrarily little. Choose an integer $m$ for which every $mp_j$ is integral and use the face coordinates $y_j(t) = \kappa t^{mp_j}$. The action map is analytic on the full original box, including the limit of this path. The image tends to zero because its trust cost does. Its three orders are $ma \cdot p, mc \cdot p, md \cdot p$. Therefore its arc expression is at most $d \cdot p$, arbitrarily close to the optimal chart value. Again a zero audit face omits $c$. This proves the infimum formula for every real precision exponent. This argument does not assert attainment for irrational schedules.

If $\alpha$ is rational, the conclusion can be strengthened. The feasible and positive optimal point graphs in variables $(b, h, v)$ are locally subanalytic: the rational power budget relation is subanalytic, and compactness allows the universal optimality condition to be expressed using bounded projections and complements. Take $b$ in a compact interval, $h, h'$ in the fixed compact representative, and $0 \leq v \leq \max F$. All these operations then take place in a bounded globally subanalytic box.

More explicitly, prescribe the two closed budget inequalities and the equality $v = F(h) > 0$. Failure of optimality means that there exists another action $h'$ in the same compact representative satisfying both budgets and $F(h') > v$. This competing action set is subanalytic, as is its bounded projection; taking its complement within the candidate graph expresses optimality. The reference is in the closure of the positive optimal graph because the positivity alternative holds for every small budget and the unique zero of the trust cost forces all such optimizers to approach zero. These facts justify selecting a curve from actual optimizers rather than merely from feasible actions.

Analytic curve selection at their reference limit supplies an analytic arc of actual optimal points with

$$b(t) = A_0 t^q + O(t^{q+1}), \qquad v(t) = F_0 t^d + O(t^{d+1}), \qquad A_0, F_0 > 0.$$

The budget component is strictly increasing on a sufficiently small positive interval, and its image is an entire interval $(0, b_1)$. To justify inversion at zero, write $b(t) = t^q A(t)$ and introduce $u = tA(t)^{1/q}$. This coordinate has a nonzero derivative and an analytic inverse, while $b = u^q$. Expanding the analytic gain in $u$ gives

$$G(b, \lambda b^\alpha) = C_{\alpha,\lambda} b^{d/q} + O(b^{d/q+1/q}).$$

The exponent equals the one already determined by the envelope. Feasibility on the selected optimal action arc gives $a \geq q$ and $c \geq \alpha q$. Its arc expression is consequently at most $d/q$; the previously proved lower inequality forces equality. This establishes attainment for rational $\alpha$. The needed curve selection follows from rectilinearization (Bierstone & Milman, 1988, Theorem 0.2): lift a sequence of optimal points approaching the reference through the finite compact cover, pass to one chart and quadrant, and join the lifted limit to an interior point of that quadrant by a line segment. Its analytic image lies in the optimal point graph for all sufficiently small positive parameters. □

*Proof of Corollary 3.3.* For each fixed sufficiently small $b > 0$, compactness gives

$$\lim_{\beta \downarrow 0} G(b, \beta) = G_0(b).$$

The limit exists by monotonicity and is at least the hard value. Maximizers along any decreasing sequence of audit budgets have a convergent subsequence. Its limit is hard feasible, and continuity of the gain proves the reverse inequality.

Let $I_0$ index the faces whose dual vertices all have $v = 0$, and set $\rho_i = \max\{u : (u, 0) \in V_i\}$ for these faces. At fixed $0 < b < 1$, a face envelope tends to zero if even one of its vertices has $v > 0$; otherwise it equals $b^{\rho_i}$. Taking the limit in the uniform two budget comparison therefore gives

$$G_0(b) \asymp \max_{i \in I_0} b^{\rho_i},$$

with an empty maximum interpreted as zero. If $I_0$ is nonempty, its relevant exponents are positive, since the hard value tends to zero. The hard exponent is $\gamma_0 = \min_{i \in I_0} \rho_i$. For the positive leading coefficient, form the one budget optimal point graph on $Z = \mathcal{M} \cap L$: its triples $(b, h, v)$ satisfy $h \in Z$, $D(h) \leq b$, $v = \Gamma(h) > 0$, and there is no feasible $h' \in Z$ with $\Gamma(h') > v$. Bounded projections and complements make this graph subanalytic. Compactness gives optimizers approaching $(0, 0, 0)$. Analytic curve selection yields $b(t) = t^q A(t)$, $A(0) > 0$, and $v(t) = F_0 t^d + O(t^{d+1})$, $F_0 > 0$. The

budget coordinate is increasing near zero and covers a full interval. The analytic change $u = tA(t)^{1/q}$ gives $b = u^q$; expanding $v$ in $u$ proves $G_0(b) = A_0 b^{d/q} + O(b^{(d+1)/q})$ with $A_0 > 0$. Comparison with the envelope identifies $d/q = \gamma_0$. This argument uses neither this corollary's plateau conclusion nor Theorem D.2.

A positive audit chart can also have an audit independent optimal envelope, when a free coordinate drives its audit to zero without affecting its leading gain, so that its limiting actions approach a zero audit face. The dual certificate and the exact limit above account for this case; this is why uniform comparison constants are needed at the endpoint.

For each face let $v_i$ be its largest vertex slope and $u_i$ the largest intercept among vertices of that slope. Its spectral function is eventually $u_i + \alpha v_i$. The finite minimum over faces eventually selects the smallest slope, then the smallest intercept among the tied slopes. If $I_0 \neq \varnothing$, that slope is zero and the eventual value is exactly $\gamma_0$. Otherwise every $v_i > 0$, and the terminal spectrum is affine with strictly positive slope. This proves the equivalence between positive local zero audit gain and an eventual plateau. An intermediate plateau alone has no such implication. □

Let $Z = \mathcal{M} \cap L$ denote the actual zero audit action set.

**Theorem D.2** (Recovery of the hard value)**.** *There are $C_Z > 0$ and $\omega > 0$ such that locally on $\mathcal{M}$,*

$$\operatorname{dist}(h, Z) \leq C_Z B(h)^{\omega}. \tag{30}$$

*If $G_0(b) \sim A_0 b^{\gamma_0} > 0$ and $\alpha\omega > \gamma_0$, then, for every fixed $\lambda > 0$,*

$$\frac{G(b, \lambda b^{\alpha})}{G_0(b)} \longrightarrow 1. \tag{31}$$

*If $G_0$ is locally zero, the same error bound instead implies $G(b, \beta) \leq C\beta^{\omega}$ for small budgets.*

*Proof of Theorem D.2.* Use the actual action zero set $Z = \mathcal{M} \cap \{h : B(h) = 0\}$, not the ambient observable space. It is compact and contains zero. The subanalytic distance function and the Łojasiewicz inequality (Lojasiewicz & Coste, 1965), in the form of Bierstone & Milman (1988, Theorem 6.4 and Remark 6.5), give, locally on the action set,

$$\operatorname{dist}(h, Z) \leq CB(h)^{\omega}$$

for some $\omega > 0$. The distance is taken to the zero set inside the actual model. This distance function is continuous and subanalytic on the compact representative and vanishes at exactly the same points as $B$, which is the zero set condition for the quoted inequality. An ambient observable projection could leave a singular or curved model and would not establish this comparison. A closest action in the true zero set instead remains available, while allowing a controlled increase of the trust budget.

For a soft feasible action choose a nearest $z \in Z$. Since zero belongs to $Z$, this nearest point also remains near zero whenever $h$ does. Lipschitz continuity gives $\Gamma(h) \leq \Gamma(z) + C\beta^{\omega}$. The analytic trust cost has vanishing derivative at the reference; its locally bounded Hessian therefore gives

$$D(z) \leq b + C\sqrt{b}\,\beta^{\omega} + C\beta^{2\omega} =: b'.$$

Consequently

$$G(b, \beta) \leq G_0(b') + C\beta^{\omega}.$$

When zero audit gain is positive locally, the one budget optimal graph argument in the proof of Corollary 3.3 gives $G_0(b) \sim A_0 b^{\gamma_0}$, with $A_0 > 0$. Also $\gamma_0 \geq 1/2$. For $\beta = \lambda b^{\alpha}$ and $\alpha\omega > \gamma_0$, we have $b'/b \to 1$ and $\beta^{\omega} = o(G_0(b))$. The power asymptotic, without any derivative of $G_0$, implies $G_0(b')/G_0(b) \to 1$. Hard feasible actions supply the opposite inequality, proving the ratio limit. If the local hard value is zero instead, the same comparison only gives $0 \leq G(b, \beta) \leq C\beta^{\omega}$. It does not create a positive zero audit leading coefficient. □

## E Proof for the four phase example

**A worked example with four phases.** The following updates separate temporary and terminal plateaus. Under the uniform reference on sixteen states, the channel reveals two independent signs

$X, Y$ and hides two more, $N_1, N_2$; the product observed target has means $(\delta, 0)$ with $0 < \delta < 1$, and for an integer $k \geq 4$ the available scores on a small compact parameter neighborhood of zero are $h(s,t) = sX + tY + (s^2 - t^3)N_1 + t^k N_2$. Two updates compete: the axis $t = 0$, with trust, audit and gain orders $(2, 4, 1)$, and the cancellation curve $s = t^{3/2}$, analytic after $t = \tau^2$, with orders $(4, 4k, 3)$. Proposition E.1 shows that together they attain the optimal value over all available actions up to constant factors, so that $\gamma$ equals $1/2$ on $(0, 2]$, $\alpha/4$ on $[2, 3]$, $3/4$ on $[3, k]$ and $3\alpha/(4k)$ beyond. Zero audit forces $t^k = 0$ and $s^2 = t^3$, hence $s = t = 0$, so the plateau at $3/4$ is temporary although its length grows with $k$. No length of plateau settles the endpoint; Corollary 3.3 does (Figure 1, left).

We record the identity used in the switching path model. Under the uniform reference, let $X_1, \dots, X_d, N_1, \dots, N_e$ be independent Rademacher signs, and suppose the observation reveals all the $X_i$. Fix an observed target with means $\mathbb{E}_r X_i = \delta_i$. For the centered score $h = \sum_i a_i X_i + \sum_j c_j N_j$, independence gives

$$\begin{aligned} \psi(h) &= \sum_i \log\cosh a_i + \sum_j \log\cosh c_j, \\ D(h) &= \sum_i \mathcal{J}(a_i) + \sum_j \mathcal{J}(c_j), \qquad & B(h) &= \sum_j \mathcal{J}(c_j), \\ \Gamma(h) &= \sum_i \delta_i a_i - \sum_j |c_j| - \psi(h), \qquad & \mathcal{J}(u) &= u\tanh u - \log\cosh u. \end{aligned} \tag{32}$$

Indeed, the information projection keeps the observed marginal and restores the reference conditional distribution of the hidden signs. A compatible world can choose each hidden sign to minimize its linear contribution at every observed state, giving the absolute value penalty exactly. No approximation or restriction to product worlds is used for the guarantee. The function $\mathcal{J}$ is even, strictly increasing on the positive half line, and $\mathcal{J}(u) = u^2/2 + O(u^4)$, while $\log\cosh u = u^2/2 + O(u^4)$. Its positive inverse therefore satisfies $\mathcal{J}_+^{-1}(v) = \sqrt{2v} + O(v^{3/2})$.

### E.1 Four phases from competing profitable paths

This subsection proves the four phase claim stated above: two observed signs $X, Y$, two hidden signs $N_1, N_2$, target means $(\delta, 0)$ with $0 < \delta < 1$, an integer $k \geq 4$, and the available scores

$$h(s,t) = sX + tY + (s^2 - t^3)N_1 + t^k N_2. \tag{33}$$

Its leading hidden term $s^2 N_1$ is noncoercive, and its exact set of zero audit updates is only the reference.

**Proposition E.1** (A four phase frontier)**.** *Uniformly over sufficiently small positive $b, \beta$,*

$$G(b, \beta) \asymp \max\left\{\min(b^{1/2}, \beta^{1/4}), \min(b^{3/4}, \beta^{3/(4k)})\right\}. \tag{34}$$

*Consequently, for every fixed $\lambda > 0$,*

$$\gamma(\alpha) = \begin{cases} 1/2, & 0 < \alpha \leq 2, \\ \alpha/4, & 2 \leq \alpha \leq 3, \\ 3/4, & 3 \leq \alpha \leq k, \\ 3\alpha/(4k), & \alpha \geq k. \end{cases} \tag{35}$$

*Proof of Proposition E.1.* Write $w = s^2 - t^3$. The exact quantities are

$$\begin{aligned} D &= \mathcal{J}(s) + \mathcal{J}(t) + \mathcal{J}(w) + \mathcal{J}(t^k), \\ B &= \mathcal{J}(w) + \mathcal{J}(t^k), \\ \Gamma &= \delta s - |w| - |t|^k - \log\cosh s - \log\cosh t \\ &\quad - \log\cosh w - \log\cosh(t^k). \end{aligned} \tag{36}$$

For every feasible action, quadratic comparability gives $|s|, |t| \leq C\sqrt{b}$, $|w| \leq C\sqrt{\beta}$, and $|t| \leq C\beta^{1/(2k)}$. Put

$$T = \min\{\sqrt{b}, \beta^{1/(2k)}\}, \qquad P = \min\{\sqrt{b}, \beta^{1/4}\}, \qquad Q = T^{3/2}.$$

Since $s^2 \le |w| + |t|^3$, we have $|s| \le C(\beta^{1/4} + Q)$ as well as $|s| \le C\sqrt{b}$. The elementary inequality $\min\{M, a+q\} \le \min\{M, a\} + q$ therefore yields $\Gamma \le \delta s \le C(P+Q) \le 2C \max\{P, Q\}$. This is an upper bound over all parameter choices, including paths whose direction varies with the budget.

Two constructions match it. On the axis, take $t = 0$ and $s = cP$. Then $D = O(s^2)$, $B = O(s^4)$, and $\Gamma = \delta s + O(s^2)$. On the cancellation curve, take $t = cT > 0$ and $s = t^{3/2}$. Here $w = 0$, $D \asymp t^2$, $B \asymp t^{2k}$, and

$$\Gamma = \delta t^{3/2} - t^k - O(t^2) = \delta t^{3/2}(1 + o(1)).$$

The last comparison is essential: the auxiliary normalizer is smaller than the useful reward. One sufficiently small fixed $c$ ensures feasibility of each construction for both independent budgets. Thus $G \asymp \max\{P, Q\}$ uniformly in $b, \beta > 0$.

The contraction constant and remainder bounds are fixed on one neighborhood before either budget is chosen. Taking the better of the two feasible constructions therefore proves uniformity; they need not join into a single smooth family at their crossing.

For $\beta = \lambda b^\alpha$, its exponent is

$$\min\left\{\max\{1/2, \alpha/4\}, \max\{3/4, 3\alpha/(4k)\}\right\}.$$

For $k \ge 4$, comparing the branches gives the stated breakpoints $2, 3, k$ and the four pieces $1/2, \alpha/4, 3/4, 3\alpha/(4k)$. Their slopes both increase and decrease. In particular, this explicit polynomial image realizes the certificate

$$V_1 = \{(1/2, 0), (0, 1/4)\}, \qquad V_2 = \{(3/4, 0), (0, 3/(4k))\}.$$

These sets correspond to the axis and cancellation curve respectively. Their completeness follows from the all action upper bound above. Zero audit forces $t^k = 0$ and $s^2 - t^3 = 0$, hence $s = t = 0$. The intermediate plateau therefore coexists with $G_0 \equiv 0$ locally. □

# F Generic rank module and its proofs

## F.1 How much rank a correction needs

We compute the envelope of Theorem 3.1 in a module where every quantity is explicit: an observable edit drags hidden coordinates, and a correction of limited rank pays for the residue. Fix $m$ contexts $c$, one observed sign $\ell$, $n$ hidden signs $\xi_j$ and the uniform reference on the $m2^{n+1}$ states. The channel reports $(c, \ell)$, and the target is $r_{\mathrm{obs}}(c, \ell) = (1 + \delta\ell)/(2m)$ with $0 < \delta < 1$. An observable edit of size $t$ in a fixed small interval moves the observed sign towards the target and displaces the hidden conditional logits by a real analytic matrix $C(t) \in \mathbb{R}^{m \times n}$ with $C(t) = O(t^2)$. A correction $E$ of rank at most $\rho$ leaves the residual $S = C(t) + E$, and the module outputs

$$q_{t,E}(c, \ell, \xi) = \frac{1}{m} \frac{e^{t\ell}}{2\cosh t} \prod_{j=1}^{n} \frac{e^{S_{cj}\xi_j}}{2\cosh S_{cj}}. \tag{37}$$

In this generic module, $C$ is a chosen analytic primitive and the rank constraint acts on conditional logits. Section 4 separately derives one such displacement from tied paths and realizes its corrections by a bounded direct output adapter; that additional parameterization is not assumed by the theorem below. Write $G_\rho(b, \beta)$ for the value (6) of this module.

**Lemma F.1** (Exact costs, exact guarantee, closed image)**.** *With $\mathcal{J}(u) = u \tanh u - \log\cosh u = \frac{1}{2}u^2 - \frac{1}{4}u^4 + O(u^6)$ and positive inverse $\mathcal{J}_+^{-1}$,*

$$B(t, E) = \frac{1}{m} \sum_{c,j} \mathcal{J}(S_{cj}), \qquad D(t, E) = \mathcal{J}(t) + B(t, E), \tag{38}$$

$$\Gamma(t, E) = \delta t - \log\cosh t - \frac{1}{m} \sum_{c,j} \big(|S_{cj}| + \log\cosh S_{cj}\big). \tag{39}$$

*Every maximum at a small budget is attained, and the local image of the score map is closed and subanalytic even where the rank constraint makes it singular at the reference.*

Three features of (39) drive the rank law. The reward is linear in the edit size because the target constrains the observed mean. The residual penalty is also linear, because a compatible world places each hidden sign at its worst. The audit charges the residue quadratically, so a correction cheap for the audit can be dear for the guarantee. Only $t > 0$ earns anything, so the orders below are one sided.

Section 5 states Theorem 5.1 and the rank law that follows from it; this appendix proves them. In the theorem one budget binds, and the module converts the smaller one into an edit size at rate $\delta$. The trust budget caps the edit at $\sqrt{2b}$, while the audit budget caps it through the residue, which rank $\rho$ cannot push below the tail $\min_{\operatorname{rank} E \le \rho} \|C(t) + E\|_F^2 = \sum_{j>\rho} \sigma_j(C(t))^2$ (Eckart & Young, 1936); inverting the order of that tail gives the second entry of $T_\rho$, and the rank $\rho$ truncation of $C(t)$ at $t = T_\rho(1 - c_\rho T_\rho)$ attains the bound with one $c_\rho$ for every ratio of the budgets (Appendix F).

High orders come from exact structure rather than from genericity, and they are fragile. In the displacement (9) the least entry order is two and the determinant is $t^7$, so the orders are $(2, 5)$: rank one lifts the critical precision from two to five and rank two removes the residue (Figure 1, right). A fixed perturbation $\epsilon t^2 e_2 e_2^\top$ with $\epsilon \neq 0$ turns the determinant into $t^4(\epsilon + t^3)$ and collapses the orders to $(2, 2)$ (Appendix F); rank one then no longer improves the critical precision exponent, although it still lowers $a_\rho^2$. Exploiting high orders therefore requires proving that the structure is present and preserved. This generic local theorem alone gives no finite parameter space adapter decision; the rank decisions of Section 4 require its separately specified architecture, factor caps, radius, and all action bounds.

Let $\rho_{\max}$ be the generic local rank of $C(t)$. For $1 \le j \le \rho_{\max}$, let $\Delta_j$ be the smallest vanishing order of a nonzero $j$-by-$j$ minor, and put $\Delta_0 = 0$. The following proposition relates those orders to the determinantal divisors and records the tail constants used in the gain law. The square invertible Laurent series order identity is Kaveh & Makhnatch (2022, Theorem 1.1); the proof below covers rectangular analytic matrices and leading constants. The determinantal description is classical (Gantmakher, 1998).

**Proposition F.2** (Determinantal orders)**.** *Write the one sided nonzero singular value expansions as* $\sigma_j(C(t)) = \kappa_j t^{\mu_j}(1 + O(t))$*, where* $2 \le \mu_1 \le \cdots \le \mu_{\rho_{\max}}$ *and* $\kappa_j > 0$*. Then*

$$\mu_j = \Delta_j - \Delta_{j-1}, \qquad \nu_\rho = \mu_{\rho+1}, \qquad a_\rho^2 = \frac{1}{m} \sum_{\substack{j>\rho \\ \mu_j = \mu_{\rho+1}}} \kappa_j^2 \quad (\rho < \rho_{\max}). \tag{40}$$

**Corollary F.3** (Minimum rank at a prescribed precision)**.** *For fixed* $\alpha, \lambda > 0$*, the smallest rank satisfying*

$$\liminf_{b \downarrow 0} \frac{G_\rho(b, \lambda b^\alpha)}{\delta\sqrt{2b}} > 0$$

*is*

$$\rho_{\min}(\alpha) = \#\{j \le \rho_{\max} : \mu_j < \alpha\}. \tag{41}$$

*At a critical remaining order* $\mu_{\rho+1} = \alpha$*, retaining at least a fraction* $\zeta \in (0, 1]$ *additionally requires and is equivalent to*

$$\lambda \ge 2^{\alpha-1} a_\rho^2 \zeta^{2\alpha}. \tag{42}$$

*If* $\mu_{\rho+1} > \alpha$ *or* $\rho \ge \rho_{\max}$*, the full coefficient is retained.*

This appendix optimizes over the correction matrix itself. A factorization $E = UV^\top$ realizes every matrix of rank at most $\rho$, but the factors have a scaling symmetry and need not lie in a compact set. All asymptotic constants below may depend on the fixed channel, target, matrix family, and dimensions; the constants in the two budget theorem are independent of the two budgets.

### F.2 Exact identities and the action image

*Proof of Lemma F.1.* The state space consists of a context $c \in [m]$, an observed sign $\ell \in \{-1, 1\}$, and hidden signs $\xi \in \{-1, 1\}^n$. With $S = C(t) + E$, the model probabilities are

$$q_{t,E}(c, \ell, \xi) = \frac{1}{m} \frac{e^{t\ell}}{2\cosh t} \prod_{j=1}^{n} \frac{e^{S_{cj}\xi_j}}{2\cosh S_{cj}}. \tag{43}$$

Thus the context marginal is fixed, rather than determined by a global normalization of the raw conditional logits. To express this model in the paper's centered score coordinates, put

$$L_c(S) = \sum_j \log\cosh S_{cj}, \qquad \overline{L}(S) = \frac{1}{m}\sum_c L_c(S).$$

Its centered score and log partition function are

$$h_{t,S}(c,\ell,\xi) = t\ell + \sum_j S_{cj}\xi_j - L_c(S) + \overline{L}(S), \qquad \psi(h_{t,S}) = \log\cosh t + \overline{L}(S). \tag{44}$$

The reference expectation of $h_{t,S}$ is zero, and direct normalization verifies the displayed formula for $\psi$.

The map from coefficients to scores is analytic and injective. Indeed, conditional reference averages of $h_{t,S}\ell$ recover $t$, and those of $h_{t,S}\xi_j$ at context $c$ recover $S_{cj}$. Its inverse on the image is therefore linear. The constraint $\operatorname{rank}(S - C(t)) \le \rho$ is the common zero set of analytic minors. Restricting $t$ to a small closed interval gives a locally closed semianalytic coefficient set whose sufficiently small closed score image is subanalytic. This justifies application of the image theorem even when the rank constraint is singular at the reference. It does not require a regular factor parameterization. Restricting to a closed score ball therefore supplies the compact local representative used by the general frontier theorem. Small KL budgets keep the maximizers away from any artificial outer boundary.

Write $\mathcal{J}(u) = u\tanh u - \log\cosh u$. Conditional independence and the Bernoulli mean $\tanh u$ give

$$\mathrm{KL}(q_{t,E}\|p_0) = \mathcal{J}(t) + \frac{1}{m}\sum_{c,j}\mathcal{J}(S_{cj}).$$

The information projection with the same observed $(c,\ell)$ marginal replaces every hidden conditional by the uniform reference conditional. To verify this minimization, fix the observed marginal and apply the conditional KL chain rule. Its observed contribution is fixed; every remaining conditional KL to the uniform reference is nonnegative and vanishes exactly at the uniform conditional. The model's observed probabilities are positive, so equality determines every such conditional. Consequently the exact audit is

$$B(h_{t,S}) = \frac{1}{m}\sum_{c,j}\mathcal{J}(S_{cj}), \qquad D(h_{t,S}) = \mathcal{J}(t) + B(h_{t,S}). \tag{45}$$

For the target, the context marginal is also $1/m$ and the conditional observed mean is $\delta$. At every observed state, a compatible world can independently choose the hidden signs to minimize their linear score. Their minimum is $-\sum_j |S_{cj}|$, and these choices are simultaneously feasible. Hence

$$\Gamma(h_{t,S}) = \delta t - \log\cosh t - \frac{1}{m}\sum_{c,j}\big(|S_{cj}| + \log\cosh S_{cj}\big). \tag{46}$$

The uniform target context marginal is essential to these weights.

For completeness, the maximum exists at all sufficiently small budgets. The function $\mathcal{J}$ is even, strictly increasing on $[0,\infty)$, and tends to $\log 2$. If $D \le b_0 < (\log 2)/m$, each $\mathcal{J}(S_{cj}) \le mb_0$ bounds $S_{cj}$. Also $D \ge \mathcal{J}(t)$ bounds $t$ inside the chosen interval when $b_0$ is sufficiently small. Continuity of $C$ bounds $E = S - C(t)$. The feasible set in $(t,E)$ is closed and bounded, hence compact, and the continuous objective attains its maximum. The reference is feasible. Equation (46) shows that positive gain requires $t > 0$. Finally, $D = 0$ forces $t = 0, S = 0$, giving the reference distribution. □

### F.3 Uniform sharp optimization under both budgets

*Proof of Theorem 5.1.* We use the elementary expansions and inequality

$$\mathcal{J}(u) = \frac{u^2}{2} - \frac{u^4}{4} + O(u^6), \qquad 0 \le \mathcal{J}(u) \le \frac{u^2}{2}. \tag{47}$$

The inequality follows by integrating $\mathcal{J}'(u) = u\operatorname{sech}^2 u \le u$ on the positive half line. For the Frobenius tail, the classical rank approximation identity is

$$\inf_{\operatorname{rank} E \le \rho}\|C(t) + E\|_F^2 = \sum_{j>\rho}\sigma_j(C(t))^2 = m\tau_\rho^2(t). \tag{48}$$

For $\rho \geq \min(m, n)$ the identity is immediate from $E = -C(t)$, and its tail is zero. Otherwise a truncated singular value decomposition attains equality, and no matrix of rank at most $\rho$, square or rectangular, does better (Eckart & Young, 1936); this lower bound holds for every feasible correction, not only the constructed one.

Assume first that the nonzero one sided tail has the stated expansion $\tau_\rho^2(t) = a_\rho^2 t^{2\nu_\rho}(1 + O(t))$, where $\nu_\rho \geq 2$. Set

$$T_D = \sqrt{2b}, \qquad T_B = (2\beta/a_\rho^2)^{1/(2\nu_\rho)}, \qquad T = \min\{T_D, T_B\}.$$

Consider a maximizing action with positive guarantee. Its positive parameter satisfies

$$t \leq \mathcal{J}_+^{-1}(b) = \sqrt{2b} + O(b^{3/2}). \tag{49}$$

If $T_D \leq T_B$, the upper bound $\Gamma \leq \delta t$ already gives $G_\rho(b, \beta) \leq \delta T + O(T^2)$.

If $0 < T_B \leq T_D$, audit feasibility implies $\|S\|_F^2 = O(\beta)$. Applying (47) entrywise and then (48) yields, uniformly over feasible actions,

$$\begin{aligned} \beta &\geq \frac{1}{2m}\|S\|_F^2(1 - O(\beta)) \\ &\geq \frac{1}{2}\tau_\rho^2(t)(1 - O(\beta)) \\ &\geq \frac{a_\rho^2}{2}t^{2\nu_\rho}(1 - O(t + \beta)). \end{aligned} \tag{50}$$

Initially use $t = O(\sqrt{b})$ to deduce $t \leq 2T_B$ for sufficiently small $b$. Substituting this back and using $\beta = (a_\rho^2/2)T_B^{2\nu_\rho}$ sharpens the conclusion to $t \leq T_B + O(T_B^2)$. Thus the same upper bound follows. More explicitly, once $t \leq 2T_B$, the last factor in (50) differs from one by at most $K(T_B + \beta)$. Taking a fixed positive root gives $t/T_B \leq 1 + K'(T_B + \beta)$. Since $\beta$ is a constant multiple of $T_B^{2\nu_\rho}$ and $\nu_\rho \geq 2$, the additive error in $t$ is bounded by a constant times $T_B^2$. No estimate here divides by a quantity depending on $b/\beta$. This use of $\beta$, rather than only $b$, controls the error even when the audit limited step is arbitrarily smaller than the KL limited step.

For a matching action, fix a sufficiently large constant $c_\rho$, put $t = T(1 - c_\rho T)$, and take the negative rank $\rho$ truncation of $C(t)$. Its residual satisfies $\|S\|_F^2 = m\tau_\rho^2(t)$. Therefore

$$\begin{aligned} B &\leq \tfrac{1}{2}\tau_\rho^2(t) \leq \tfrac{a_\rho^2}{2}t^{2\nu_\rho}(1 + O(t)) \leq \tfrac{a_\rho^2}{2}T^{2\nu_\rho} \leq \beta, \\ D &\leq \tfrac{1}{2}t^2 + O(t^{2\nu_\rho}) \leq \tfrac{1}{2}T^2 - c_\rho T^3 + O(T^4) \leq b. \end{aligned} \tag{51}$$

One choice of $c_\rho$ and a subsequently small upper bound on $b$ works for every ratio of budgets. Since $\nu_\rho \geq 2$, $\|S\|_1 \leq \sqrt{mn}\|S\|_F = O(t^{\nu_\rho}) = O(t^2)$. The exact guarantee consequently gives $\Gamma \geq \delta t - O(t^2) = \delta T - O(T^2)$, proving the uniform formula. Truncated SVD is used only for this leading approximation; it need not exactly optimize the nonquadratic audit or the finite budget guarantee.

When $\beta = 0$, audit zero forces $S = 0$. A sufficiently small positive $t$ would then give $\operatorname{rank} C(t) \leq \rho$, contradicting the nonzero tail. Nonpositive $t$ cannot improve the guarantee, so the value is zero. If the tail instead vanishes identically, choose $E = -C(t)$. The upper bounds $D \geq \mathcal{J}(t)$ and $\Gamma \leq \delta t - \log\cosh t$ are attained together. For sufficiently small $b$, the scalar objective is increasing up to $\mathcal{J}_+^{-1}(b)$, since its derivative is $\delta - \tanh t$. Thus for every $\beta \geq 0$,

$$G_\rho(b, \beta) = \delta\mathcal{J}_+^{-1}(b) - \log\cosh(\mathcal{J}_+^{-1}(b)) = \delta\sqrt{2b} - b + O(b^{3/2}).$$

This separately verifies zero audit gain, including the zero audit boundary. The case $b = 0$ was already established by the unique zero of the total KL cost. For positive $b, \beta$ in the nonzero tail case, the constructed guarantee is positive once $b$ is sufficiently small, so the restriction to a positive maximizing parameter in the upper bound argument loses no nontrivial optimum. □

Substituting $\beta = \lambda b^\alpha$ into the uniform formula gives (19) at $\alpha = \nu_\rho$ and, for $\alpha > \nu_\rho$, the tight expansion $\delta(2\lambda/a_\rho^2)^{1/(2\nu_\rho)}b^{\alpha/(2\nu_\rho)} + O(b^{\alpha/\nu_\rho})$. The uniform statement also covers audit schedules that are not powers of the trust budget.

### F.4 Analytic orders and the minimum correction rank

*Proof of Proposition F.2.* The analytic symmetric family $C(t)^\top C(t)$ admits analytic eigenvalue branches by the classical one parameter theorem of Rellich (1937); see Barbarino & Noferini (2023, Theorem 1.1) for the modern finite dimensional formulation. Each nonzero branch is nonnegative on both sides of zero. Its leading order is therefore even, with positive coefficient. Taking its positive square root on $t > 0$ gives $\sigma_j(C(t)) = \kappa_j t^{\mu_j}(1 + O(t))$, with positive $\kappa_j$ and integer $\mu_j$. Ordering is eventually fixed because a nonzero difference of two analytic branches has an eventually fixed sign. This does not assert that sorted singular values are analytic across zero.

Let $\Delta_j$ be the least order among nonzero $j$-by-$j$ minors. The sum of their squares has order $2\Delta_j$, since their leading coefficients cannot cancel. Cauchy–Binet identifies this sum with the $j$th elementary symmetric function of the squared singular values. Ordering the singular orders increasingly therefore gives $\Delta_j = \mu_1 + \cdots + \mu_j$, and hence $\mu_j = \Delta_j - \Delta_{j-1}$. For $\rho$ below the generic rank $\rho_{\max}$,

$$\nu_\rho = \mu_{\rho+1}, \qquad a_\rho^2 = \frac{1}{m} \sum_{\substack{j>\rho \\ \mu_j = \mu_{\rho+1}}} \kappa_j^2.$$

Ranks $\rho \geq \rho_{\max}$ have identically zero tail. □

*Proof of Corollary F.3.* For $\rho < \rho_{\max}$, the theorem gives a positive limiting fraction of $\sqrt{b}$ exactly when $\alpha \leq \mu_{\rho+1}$. Therefore precisely the singular directions with $\mu_j < \alpha$ must be removed. The smallest admissible correction rank is their number. At equality, the critical constant above is positive but can be less than one. Full retention then additionally requires $\lambda \geq 2^{\alpha-1} a_\rho^2$; retention of a prescribed fraction $\zeta \in (0, 1]$ requires $\lambda \geq 2^{\alpha-1} a_\rho^2 \zeta^{2\alpha}$. If the next order is strictly larger than $\alpha$, or the tail vanishes, the limiting fraction is already one. The factors $U, V$ need that many columns to realize the corresponding matrix rank class. This count concerns representability; convergence of factor training is a separate question. □

### F.5 Verification of the changing orientation example

For the matrix in (9), the least entry order is two and the determinant is $t^7$. Its singular orders are consequently $(2, 5)$, and direct division by the leading largest singular value gives unit leading constants for both singular values. Since $m = 2$, both $a_0^2$ and $a_1^2$ equal $1/2$. The correction

$$E(t) = -\begin{pmatrix} t \\ t^2 \end{pmatrix} \begin{pmatrix} t & t^2 \end{pmatrix}$$

has rank one and leaves only the bottom right residual $t^5$. It attains the sharp leading tail constant without a numerical SVD. Rank zero has critical exponent two, rank one has exponent five, and rank two permits exact cancellation.

If both rank one factor directions are fixed, cancellation of the leading $t^2 e_1 e_1^\top$ term requires both directions to be proportional to $e_1$. The off diagonal $t^3$ entries then remain. Even with only the left direction fixed and the right factor unrestricted, the left direction must be $e_1$ to remove the leading term, leaving the bottom left $t^3$ entry; the right fixed case is analogous. Jointly changing factor directions achieves order five instead of three. This compares specified matrix rank classes. It makes no claim of equal parameter counts, identical optimization difficulty, or a universal rank requirement for arbitrary neural architectures.

**A perturbation at finite scale.** Fix $0 < \epsilon_0 < 1$, let $C_\epsilon(t) = C(t) + \epsilon t^2 e_2 e_2^\top$ with $|\epsilon| \leq \epsilon_0$, and choose $t_0 > 0$ with $t_0^2 + t_0^3 \leq (1 - \epsilon_0)/2$. Then $\det C_\epsilon(t) = t^4(\epsilon + t^3)$, and for $0 < t \leq t_0$ the larger eigenvalue $\Lambda_+$ of $C_\epsilon(t)/t^2$ satisfies $0 \leq \Lambda_+ - 1 \leq 2t^2/(1 - \epsilon_0)$ and is also the largest in absolute value. Hence, uniformly over this rectangle, $\sigma_1 = t^2 \Lambda_+ = t^2(1 + O_{\epsilon_0}(t^2))$ and $\sigma_2 = t^2|\epsilon + t^3|/\Lambda_+$, including its zeros. For every fixed $0 < |\epsilon| \leq \epsilon_0$ the orders are $(2, 2)$ and $a_1^2 = \epsilon^2/2$; at $\epsilon = 0$ they are $(2, 5)$ with $a_1^2 = 1/2$ at order five. The finite scale regimes $t^3 \gg |\epsilon|$ and $t^3 \ll |\epsilon|$ give respectively $t^5$ and $|\epsilon| t^2$, and a negative perturbation has one positive rank drop point $t = |\epsilon|^{1/3}$, inside this interval exactly when $|\epsilon| \leq t_0^3$. For an arbitrary fixed nonzero real $\epsilon$, the sorted leading coefficients at order two are $\max(1, |\epsilon|)$ and $\min(1, |\epsilon|)$, so in general $a_1^2 = \frac{1}{2} \min(1, \epsilon^2)$.

### F.6 Exact exponents and leading constants for polynomial displacements

**Corollary F.4** (A finite polynomial certificate). *In the module (37), let $C \in \mathbb{Q}[t]^{m\times n}$ have no terms of degree below two. Enumerate all nonzero minors. Let $\rho_{\max}$ be the largest size present, $\Delta_j$ the smallest order among minors of size $j$, and $\Delta_0 = 0$. If $\rho_{\max} = 0$, every rank has the trust only certificate below. Otherwise set $\mu_j = \Delta_j - \Delta_{j-1}$. For every integer $\rho \geq 0$, a valid envelope in Theorem 3.1 is a single set*

$$V_\rho = \begin{cases} \{(1/2, 0), (0, 1/(2\mu_{\rho+1}))\}, & \rho < \rho_{\max}, \\ \{(1/2, 0)\}, & \rho \geq \rho_{\max}. \end{cases}$$

*The enumeration terminates using exact rational polynomial arithmetic. It computes the exponents, not the comparison constants or a validity radius.*

*Proof.* Expand each determinant exactly, cancel equal powers, and take the smallest degree with a nonzero coefficient. Polynomial rank equals the largest size of a nonzero minor. Proposition F.2 gives the singular orders and $\nu_\rho = \mu_{\rho+1}$ for a nonzero tail. Theorem 5.1 then bounds $G_\rho$ above and below by fixed multiples of $\min\{b^{1/2}, \beta^{1/(2\mu_{\rho+1})}\}$, uniformly for small positive budgets. If $\rho \geq \rho_{\max}$, its exact zero tail formula gives $G_\rho \asymp b^{1/2}$. There are finitely many minors, so exact expansion terminates. □

The construction applies to the stated module with its polynomial displacement, not to arbitrary polynomial score graphs, and enumerating minors may be expensive. Exact coefficients matter: (9) gives $(\mu_1, \mu_2) = (2, 5)$, whereas any fixed nonzero rational perturbation $\epsilon t^2 e_2 e_2^\top$ gives $(2, 2)$. The script `checks/polynomial_orders.py` computes these and rectangular, zero and rank deficient examples in rational arithmetic, with the sums $\varpi_j$ and constants $a_\rho^2$ of the next paragraph.

**Leading constants from the same minors.** The enumeration also returns the constant $a_\rho$ of Theorem 5.1. Let $\varpi_j > 0$ be the sum of the squared leading coefficients of the $j$ by $j$ minors whose order is exactly $\Delta_j$, and put $\varpi_0 = 1$ and $\mu_0 = 0$. By the Cauchy–Binet step in the proof of Proposition F.2, $\varpi_j$ is the coefficient of $t^{2\Delta_j}$ in the $j$th elementary symmetric function of the squared singular values. A set of $j$ indices in $\{1, \dots, \rho_{\max}\}$ attains the least sum of orders $\Delta_j = \mu_1 + \cdots + \mu_j$ exactly when it contains every index $i$ with $\mu_i < \mu_j$ and its other indices have order $\mu_j$, so $\varpi_j$ is the product of the $\kappa_i^2$ with $\mu_i < \mu_j$ times an elementary symmetric function of the $\kappa_i^2$ with $\mu_i = \mu_j$. For $0 \leq \rho < \rho_{\max}$ with $\mu_\rho < \mu_{\rho+1}$ this gives $\varpi_\rho = \kappa_1^2 \cdots \kappa_\rho^2$ and $\varpi_{\rho+1} = \varpi_\rho \sum_{j:\mu_j = \mu_{\rho+1}} \kappa_j^2$, so

$$a_\rho^2 = \frac{\varpi_{\rho+1}}{m\, \varpi_\rho},$$

a rational number when $C$ is rational. If instead $\mu_\rho = \mu_{\rho+1}$, let the indices $j_0 + 1, \dots, j_1$ carry the orders equal to $\mu_{\rho+1}$, so that $j_0 < \rho < j_1$. The same count gives

$$\prod_{j=j_0+1}^{j_1} (x - \kappa_j^2) = \sum_{i=0}^{j_1 - j_0} (-1)^i \, \frac{\varpi_{j_0+i}}{\varpi_{j_0}} \, x^{j_1 - j_0 - i},$$

and $m\, a_\rho^2$ is the sum of its $j_1 - \rho$ smallest roots. For (9), $\varpi_1 = \varpi_2 = 1$ and $m = 2$ return $a_0^2 = a_1^2 = 1/2$. Under the perturbation $\epsilon t^2 e_2 e_2^\top$ with $0 < |\epsilon| < 1$ the two orders coincide, the polynomial is $x^2 - (1 + \epsilon^2)x + \epsilon^2$, and its smaller root returns $a_1^2 = \epsilon^2/2$, as found above. For a rational displacement the leading term of (17), the critical fraction (19) and the retention threshold (42) are therefore exact; the remainder constant and a validity radius are still not computed.

## G The tied model: image, finite bounds, and constructive decisions

This appendix proves Theorem 4.1 for the parameter domain in Section 4. In particular, the all action upper bounds concern the bounded factors themselves, rather than just the displayed cancellation curves. Write $e_2 = (0, 1)^\top$ and use the convention that at rank zero the factors have no columns and $E = 0$. Throughout this appendix $\mathcal{J}(u) = u \tanh u - \log\cosh u$.

### G.1 Path multiplication and the exact parameter image

The encoder produces $z = tx_1 + t^2x_2$. The two core heads are therefore $tz = t^2x_1 + t^3x_2$ and $t(tz) = t^3x_1 + t^4x_2$. The separate five edge path adds $t^5x_2$ only to the second head. Hence the matrix whose $c$th row is the hidden logit vector on input $e_c$ is precisely (9). This derives the displacement from scalar edge weights; no freely trainable matrix has been declared equal to $C(t)$. The architecture is intentionally designed, and fixed routing and exact tying are part of its specification.

For the adapter, $VU^\top e_c$ is a column of the effective weight increment $\Delta W = VU^\top$, so the matrix of context rows is $E = UV^\top$. Its rank equals $\operatorname{rank} \Delta W \le \rho$. Nuclear norm duality gives

$$\|UV^\top\|_* = \sup_{\|W\|_{\mathrm{op}} \le 1} \operatorname{tr}(W^\top UV^\top) \le \|U\|_F \|V\|_F \le \tfrac{1}{4}.$$

Conversely, if $E = P \operatorname{diag}(s_1, \dots, s_k) Q^\top$ is a thin singular value decomposition with $k \le \rho$ and $\sum_i s_i \le 1/4$, set $U = P \operatorname{diag}(\sqrt{s_i})$ and $V = Q \operatorname{diag}(\sqrt{s_i})$, adding zero columns to reach $\rho$. Then $UV^\top = E$ and both squared Frobenius norms are $\sum_i s_i \le 1/4$. This proves (11) exactly. In particular, the two individual factor caps cannot be replaced by $\|E\|_F \le 1/4$: $\operatorname{diag}(3/20, 3/20)$ satisfies that Frobenius bound but has nuclear norm $3/10 > 1/4$.

Taking the logarithm of (10) relative to $p_0 = 1/16$ gives

$$\log(q/p_0) = t\ell + \sum_j S_{cj}\xi_j - \log\cosh t - L_c(S).$$

Centering under $p_0$ gives $h_{t,S}$ and $\psi(h_{t,S})$ as in the main text. Orthogonality of the uniform reference signs yields the *context specific* inverse

$$t = \mathbb{E}_{p_0}[\ell h_{t,S}], \qquad S_{cj} = \mathbb{E}_{p_0}[\xi_j h_{t,S} \mid c] = 2\mathbb{E}_{p_0}[\mathbf{1}_{\{\mathrm{context}=c\}} \xi_j h_{t,S}]. \tag{52}$$

An unconditional average against $\xi_j$ alone would recover only the average of the two rows. Thus the coefficient to score map is injective with a linear continuous inverse; the factor representation need not be injective, even when $E = 0$.

The domain $[0, 1/8] \times \{\|U\|_F \le 1/2\} \times \{\|V\|_F \le 1/2\}$ is compact and semialgebraic. Its coefficient map is polynomial. The centered score map is real analytic on a neighborhood of this compact set, since $\log\cosh$ is real analytic at every real argument. Its graph restricted to the parameter domain is relatively compact and semianalytic, so its projection onto score coordinates is compact and subanalytic. This is a local analytic argument on a compact domain, not a claim that $\log\cosh$ is globally semialgebraic. The reference law lies in the image. Continuity of the costs and gain then gives attainment of all budget constrained maxima, including zero budgets.

### G.2 Exact information costs and the worst compatible world

The model marginal on $(c, \ell)$ is $q_{\mathrm{obs}}(c, \ell) = e^{t\ell}/(4\cosh t)$. For the partition channel, the information projection of $p_0$ with this observed marginal has the same marginal and uniform hidden signs conditionally on every $(c, \ell)$. Indeed, the KL chain rule decomposes the divergence from $p_0$ into the fixed observed divergence plus a nonnegative average of conditional divergences from the uniform hidden law, uniquely minimized by those uniform conditionals. For a tilted sign $q_u(\xi) = e^{u\xi}/(2\cosh u)$,

$$\mathrm{KL}(q_u \| \mathrm{Unif}\{-1, 1\}) = u\mathbb{E}_{q_u}\xi - \log\cosh u = \mathcal{J}(u).$$

The hidden coordinates of the model are conditionally independent, so the KL chain rule proves $B = \frac{1}{2}\sum_{c,j} \mathcal{J}(S_{cj})$ and $D = \mathcal{J}(t) + B$.

Under any world with observed law $r_\delta$, the observed linear term has expectation $\delta t$ and each context has probability $1/2$. The conditional hidden expectation of $\sum_j S_{cj}\xi_j$ is at least $-\sum_j |S_{cj}|$. Equality is attained by putting both hidden signs at their minimizing endpoints, separately at each observed state; zero coefficients can be assigned either sign. These deterministic conditionals define a compatible world, regardless of whether it is a product law in any larger sense. Averaging the normalizers proves (12) for the minimum over the complete fiber.

### G.3 Uniform localization and all action upper bounds

Evenness and $\mathcal{J}'(u) = u\,\mathrm{sech}^2\,u$ imply, for every real $u$,

$$\frac{u^2}{2} - \frac{u^4}{4} \le \mathcal{J}(u) \le \frac{u^2}{2}, \qquad 0 \le \log\cosh u \le \frac{u^2}{2}. \tag{53}$$

For $u \ge 0$, integrate $1 - u^2 \le 1 - \tanh^2 u = \mathrm{sech}^2\, u \le 1$ for the first pair, and $0 \le \tanh u \le u$ for the second. Also $\mathcal{J}$ is strictly increasing on $[0, \infty)$.

Every feasible action at $b \le 1/512$ has $\mathcal{J}(t) \le b$ and $\mathcal{J}(S_{cj}) \le 2b$. Since

$$2b \le \frac{1}{256} < \frac{199}{40000} \le \mathcal{J}(1/10),$$

it follows that $t < 1/10$ and $|S_{cj}| < 1/10$. On these arguments,

$$\mathcal{J}(u) \ge \frac{199}{400}u^2, \qquad B \ge \frac{199}{800}\|S\|_F^2. \tag{54}$$

Set $c_+ = 1003/1000$. We will use the exact rational comparisons

$$c_+^2 > \frac{200}{199}, \qquad c_+^4 > \frac{200}{199}, \qquad c_+^{10} > \frac{200}{199}\left(\frac{1011}{1000}\right)^2. \tag{55}$$

For the first, the difference is $195791/199000000 > 0$; the second follows because $c_+ > 1$. For the third, Bernoulli's inequality gives $c_+^{10} \ge 103/100$, and

$$\frac{103}{100} - \frac{200}{199}\left(\frac{1011}{1000}\right)^2 = \frac{2729}{995000} > 0.$$

All ranks obey $\Gamma_\delta(t, S) \le \delta t$ and

$$t \le \sqrt{\frac{200}{199}}\sqrt{2b} < c_+\sqrt{2b}. \tag{56}$$

At rank zero, $S = C(t)$ has entry $S_{11} = t^2$, so $B \ge (199/800)t^4$ and $t \le (200/199)^{1/4}(4\beta)^{1/4}$. Combining with (56) proves the upper bound at rank zero.

At rank one and $t > 0$, $C(t)$ is positive definite, with determinant $t^7$ and trace $t^2(1 + t^2 + t^3)$. Hence

$$\frac{t^5}{1 + t^2 + t^3} \le \sigma_2(C(t)) \le t^5. \tag{57}$$

For any possibly nonsymmetric matrix $E$ of rank at most one, choose a unit vector $w$ in its kernel. Then

$$\|C(t) + E\|_F \ge \|(C(t) + E)w\|_2 = \|C(t)w\|_2 \ge \sigma_2(C(t)).$$

Since $t < 1/10$, its denominator in (57) is at most $1011/1000$. Consequently

$$\beta \ge B \ge \frac{199}{800}\left(\frac{1000}{1011}\right)^2 t^{10}.$$

Equations (55) and (56) give $t \le c_+ T_1$ and the claimed upper gain. No optimality of a polynomial or spectral witness is used; the factor caps only restrict the matrices to which this argument applies.

For rank two, put $f_\delta(t) = \delta t - \log\cosh t$. The exact objective implies $\Gamma_\delta(t, S) \le f_\delta(t)$. On $[0, 1/10]$, $f'_\delta(t) = \delta - \tanh t > 0$. Also $\mathcal{J}_+^{-1}(b) < 1/10 < 1/8$, because $b \le 1/512 < \mathcal{J}(1/10)$. Thus every feasible action has gain at most $f_\delta(\mathcal{J}_+^{-1}(b))$. We show below that the rank two factors cancel $C$ at that step and satisfy both norm caps, proving the exact formula. Its coarser upper bound follows from (56).

### G.4 Polynomial feasible witnesses and exact finite constants

Multiplying (15) gives $U^{(1)}V^{(1)\top} = -vv^\top$ and $U^{(2)}V^{(2)\top} = -vv^\top - t^5 e_2 e_2^\top$, proving the residual formulas. For $0 \le t \le 1/8$ their factor norms satisfy

$$\begin{aligned}\|U^{(1)}\|_F^2 = \|V^{(1)}\|_F^2 &= t^2 + t^4 \le 65/4096 < 1/4,\\ \|U^{(2)}\|_F^2 &= t^2 + 2t^4 \le 33/2048 < 1/4,\\ \|V^{(2)}\|_F^2 &= t^2 + t^4 + t^6 \le 4161/262144 < 1/4.\end{aligned}$$

Thus these are admissible parameter curves on the whole declared interval.

For rank zero set $c_0 = 499/500$, $x = 1/16$, $a(s) = 1 + 2s^2 + (s^2 + s^3)^2$, and $a_* = a(x) = 16908577/16777216$. At $t = c_0 T_0$ we have $t \le T_0 \le x$, and $\|C(t)\|_F^2 = t^4 a(t) \le t^4 a_*$. The elementary rational inequalities

$$c_0^4 a_* < 1, \qquad c_0^2 + \frac{a_*}{512} < 1 \tag{58}$$

imply, by (53),

$$B \le \frac{t^4 a_*}{4} \le \frac{T_0^4}{4} \le \beta, \qquad D \le \frac{t^2}{2} + \frac{t^4 a_*}{4} \le \frac{T_0^2}{2}\left(c_0^2 + \frac{a_*}{512}\right) \le b.$$

For completeness, the exact positive differences in (58) are

$$1 - c_0^4 a_* = \frac{218896934745423}{1048576000000000000}, \qquad 1 - c_0^2 - \frac{a_*}{512} = \frac{272137525463}{134217728000000}.$$

All entries of $C(t)$ are nonnegative. Therefore

$$\Gamma_\delta(t, C(t)) \ge \delta t - \frac{t^2}{2} - \frac{t^2 + 2t^3 + t^4 + t^5}{2} - \frac{t^4 a_*}{4}.$$

For $T_0 > 0$, divide by $\delta T_0$ and use $\delta \ge 1/2$, $c_0 \le 1$, and $T_0 \le 1/16$ to obtain

$$\frac{\Gamma_\delta(c_0 T_0, C(c_0 T_0))}{\delta T_0} \ge c_0 - \frac{2}{16} - \frac{2}{16^2} - \frac{1}{16^3} - \frac{1}{16^4} - \frac{a_*}{2 \cdot 16^3} > \frac{108}{125}. \tag{59}$$

The exact difference in the final comparison is $13831074531/17179869184000 > 0$.

For rank one put $c_1 = 999/1000$ and $t = c_1 T_1$. Since the only residual is $S_{22} = t^5$,

$$B = \tfrac{1}{2}\mathcal{J}(t^5) \le t^{10}/4 \le \beta, \qquad D \le t^2/2 + t^{10}/4 \le \frac{T_1^2}{2}\left(c_1^2 + \frac{1}{2 \cdot 16^8}\right) \le b.$$

Here $1 - c_1^2 = 1999/10^6 > 1/(2 \cdot 16^8)$. Also

$$\Gamma_\delta(t, t^5 e_2 e_2^\top) \ge \delta t - t^2/2 - t^5/2 - t^{10}/4.$$

For $T_1 > 0$ this gives

$$\frac{\Gamma_\delta(c_1 T_1, (c_1 T_1)^5 e_2 e_2^\top)}{\delta T_1} \ge c_1 - \frac{1}{16} - \frac{1}{16^4} - \frac{1}{2 \cdot 16^9} > \frac{117}{125}. \tag{60}$$

Indeed the last difference is $1/2000 - 1/65536 - 1/137438953472 > 0$.

At rank two, $S = 0$ and $t = T_2 \le 1/16$ give $D = \mathcal{J}(T_2) \le T_2^2/2 = b$ and $B = 0$. Its gain obeys

$$\Gamma_\delta(T_2, 0) \ge \delta T_2 - T_2^2/2 \ge \frac{15}{16}\delta T_2,$$

because $T_2/(2\delta) \le 1/16$. The same admissible cancellation curve at $t = \mathcal{J}_+^{-1}(b)$ attains the exact rank two upper bound.

If $\beta = 0$, the nonnegativity and strict zero set of $\mathcal{J}$ imply $S = 0$. At any positive $t$, $\det C(t) = t^7 > 0$, so no rank zero or rank one $E$ can cancel it. Only $t = 0$, $E = 0$ remains at those ranks, with gain zero. Rank two retains its positive value for $b > 0$. If $b = 0$, $D = 0$ implies the reference law and gain zero at every rank; redundant factors need not themselves vanish. These cases avoid division by $T_\rho = 0$ and complete the proof of Theorem 4.1.

### G.5 Scalar endpoints, finite brackets, and rank tests

The exact witness costs and penalties appearing in (16) are

$$\begin{aligned} A_0(t) &= \tfrac{1}{2}\{\mathcal{J}(t^2) + 2\mathcal{J}(t^3) + \mathcal{J}(t^4 + t^5)\}, \\ A_1(t) &= \tfrac{1}{2}\mathcal{J}(t^5), \qquad A_2(t) = 0, \end{aligned} \tag{61}$$

$$\begin{aligned} R_0(t) &= \log\cosh t + \tfrac{1}{2}\{t^2 + 2t^3 + t^4 + t^5 \\ &\qquad + \log\cosh(t^2) + 2\log\cosh(t^3) + \log\cosh(t^4 + t^5)\}, \\ R_1(t) &= \log\cosh t + \tfrac{1}{2}\{t^5 + \log\cosh(t^5)\}, \\ R_2(t) &= \log\cosh t. \end{aligned} \tag{62}$$

The trust costs are $F_\rho(t) = \mathcal{J}(t) + A_\rho(t)$. Every $F_\rho$ is continuous and strictly increasing for $t \geq 0$; $A_0, A_1$ are also strictly increasing, while $A_2$ is zero. The feasible set in (16) is consequently a closed interval containing zero. Its endpoint is at least $c_\rho T_\rho$ for $\rho = 0, 1$, and $t_2^* = \mathcal{J}_+^{-1}(b)$ in the theorem's range.

The curve gain $g_\rho(t) = \delta t - R_\rho(t)$ is strictly increasing on $[0, 1/8]$. For rank zero, if $u$ is any of its nonnegative residual entries, differentiation gives $\frac{d}{dt}(u + \log\cosh u)/2 = u'(1 + \tanh u)/2 \leq u'$. Thus

$$\begin{aligned} g_0'(t) &\geq \delta - 3t - 6t^2 - 4t^3 - 5t^4 \\ &\geq \frac{1}{2} - \frac{3}{8} - \frac{6}{64} - \frac{4}{512} - \frac{5}{4096} = \frac{91}{4096} > 0. \end{aligned}$$

Similarly $g_1'(t) \geq \delta - t - 5t^4 > 0$ and $g_2'(t) = \delta - \tanh t > 0$ on this interval. Each derivative is at most $\delta$, because every penalty is nondecreasing. Hence the endpoint improves the contracted witness and, for any feasible lower bracket satisfying $0 \leq t_\rho^* - \widehat{t}_\rho \leq \epsilon_t$,

$$0 \leq g_\rho(t_\rho^*) - g_\rho(\widehat{t}_\rho) \leq \delta\epsilon_t. \tag{63}$$

Such a bracket can be obtained by scalar monotone search with exact real comparisons or rigorously enclosed function evaluations. At a comparison whose sign remains unresolved, one must refine the enclosure or retain the existing certified bracket; an ordinary floating point sign test is not a proof of feasibility. Zero budget cases have the exact endpoints already described. A coarse finite bracket does not automatically retain the endpoint's multiplicative guarantee: its actual gain or the error in (63) must be used.

For $T_0, T_1 > 0$, dividing the constructive lower bound by the universal upper bound gives respectively $(108/125)/(1003/1000) = 864/1003$ and $(117/125)/(1003/1000) = 936/1003$. Rank two has the exact endpoint optimizer. No assertion that the rank one curve optimizes the full factor class is needed. In fact, minimizing audit at a fixed $t$ and maximizing gain under both budgets are different objectives.

For any requested gain $g$, let $L_\rho$ be the gain of a certified feasible point, including any bracket allowance, and take $U_\rho = (1003/1000)\delta T_\rho$, or the exact rank two value. Then $L_\rho \geq g$ certifies sufficiency and $U_\rho < g$ certifies impossibility. Neither test decides the case $L_\rho < g \leq U_\rho$. The factor classes are nested by zero padding columns. Therefore a certified rank is minimal only when every lower rank has also been excluded; a constructive upper rank witness alone does not establish it.

### G.6 The three rational minimum rank decisions

Fix $b = 1/5000$, $\delta = 1/2$, and $g = 1/200$. For $\beta = 10^{-4}$, the rank zero trust cap gives $T_0 = 1/50$, so $t = 499/25000$ is feasible and its gain is at least $108/12500 > g$. Rank zero is already the least available rank. For $\beta = 10^{-10}$, $T_1 = 1/50$, so $t = 999/50000$ is a feasible rank one step with gain at least $117/12500 > g$. Rank zero is impossible because

$$\frac{1003}{2000}(4 \cdot 10^{-10})^{1/4} < \frac{1}{200}, \qquad \left(\frac{10}{1003}\right)^4 - \frac{4}{10^{10}} = \frac{23987945891919}{2530135270202500000000} > 0.$$

For $\beta = 10^{-24}$, rank two at $t = 1/50$ is feasible with gain at least $3/320 > g$. Rank one is impossible because

$$\frac{1003}{2000}(4 \cdot 10^{-24})^{1/10} < \frac{1}{200},$$

as verified by the exact positive difference

$$\left(\frac{10}{1003}\right)^{10} - \frac{4}{10^{24}} = \frac{249896959174292861064726455811095 1}{257602064267847338183860472262250000000000000000000000} > 0.$$

Nesting then excludes rank zero as well. These comparisons and the rational contracted steps prove Table 2; no rounded inverse-$\mathcal{J}$ evaluation is involved. As an alternative direct check at these rational points, (53) gives the rational certificates

$$D \le \frac{t^2}{2} + \frac{1}{4}\sum_{c,j} S_{cj}^2, \quad B \le \frac{1}{4}\sum_{c,j} S_{cj}^2, \quad \Gamma_{1/2}(t,S) \ge \frac{t}{2} - \frac{t^2}{2} - \frac{1}{2}\sum_{c,j}|S_{cj}| - \frac{1}{4}\sum_{c,j} S_{cj}^2.$$

### G.7 One target event

The statistical experiment here is restricted to the known family $r_\delta(c,\ell) = (1+\delta\ell)/4$, $1/2 \le \delta < 1$. In particular the context is uniform and the observed mean is common to both contexts. Neither condition is estimated from the data. Fix an integer $N \ge 1$ and $\eta \in (0,1)$ in advance, and draw $N$ independent observed pairs. Set

$$\widehat{\delta} = N^{-1}\sum_{i=1}^{N} \ell_i, \qquad e_N = \sqrt{2\log(2/\eta)/N},$$

$\underline{\delta} = \max\{0, \widehat{\delta} - e_N\}$ and $\overline{\delta} = \min\{1, \widehat{\delta} + e_N\}$. On one event, with $L_N$ as in the proposition below, simultaneously for every action, rank and pair of budgets,

$$\Gamma_\delta(t,S) \;\ge\; L_N(t,S) \;\ge\; \Gamma_\delta(t,S) - 2e_N t. \tag{64}$$

**Proposition G.1** (One simultaneous target event)**.** *With probability at least $1-\eta$, the event*

$$\mathcal{E}_N = \{|\widehat{\delta} - \delta| \le e_N\} \tag{65}$$

*holds. On this event $\underline{\delta} \le \delta \le \overline{\delta}$ and $0 \le \delta - \underline{\delta} \le 2e_N$. Consequently, for all $t \in [0, 1/8]$ and every feasible factor pair, simultaneously,*

$$L_N(t,S) := \underline{\delta}\, t - \log\cosh t - \frac{1}{2}\sum_{c,j}\big(|S_{cj}| + \log\cosh S_{cj}\big)$$

*satisfies (64). The assertion holds jointly over all three ranks and all budgets, including data selected choices.*

*On the same event, when $0 < b \le 1/512$ and $\beta \ge 0$, write $c_+ = 1003/1000$ and $\tau = \mathcal{J}_+^{-1}(b)$. Then*

$$\begin{aligned} &\mathsf{G}_\rho(b,\beta;\delta) \le c_+\overline{\delta}\, T_\rho, && \rho = 0,1,\\ &\mathsf{G}_2(b,\beta;\delta) \le \overline{\delta}\,\tau - \log\cosh\tau. \end{aligned} \tag{66}$$

*Proof.* The signs have mean $\delta$ and range of length two, so Hoeffding's inequality gives $\Pr\{|\widehat{\delta} - \delta| > e_N\} \le 2\exp(-Ne_N^2/2) = \eta$. On $\mathcal{E}_N$, $\widehat{\delta} - e_N \le \delta$ and $\delta - (\widehat{\delta} - e_N) \le 2e_N$. Clipping the lower endpoint at zero preserves both inequalities because $\delta \ge 0$. Clipping the upper endpoint at one preserves $\delta \le \overline{\delta}$ because $\delta < 1$. The exact gain identity gives

$$\Gamma_\delta(t,S) - L_N(t,S) = (\delta - \underline{\delta})t \in [0, 2e_N t].$$

This deterministic identity holds for every $(t,S)$ on a single event about $\widehat{\delta}$. No union bound over actions, ranks or budgets is required, and no independence between a selected action and these observations is required. Their trust and audit costs are known functions of $(t,S)$, not estimates of $\delta$.

For $\rho = 0, 1$, first apply Theorem 4.1 at the true $\delta < 1$, and then use $c_+\delta T_\rho \le c_+\overline{\delta} T_\rho$. For rank two use its exact formula at the true $\delta$ and $\tau \ge 0$. Thus (66) remains a valid comparison if the clipped upper endpoint equals one; it never invokes a theorem whose target assumption requires $\delta < 1$ at that endpoint. □

A feasible action with $L_N \geq g$ is sufficient for target gain $g$ on $\mathcal{E}_N$. An upper bound in (66) strictly below $g$ excludes that rank on the same event. Minimality requires excluding every smaller rank; a gap between the certificates is unresolved. The bounds also permit reporting $\max\{L_N(t,S), \Gamma_{1/2}(t,S)\}$, since $\Gamma_\delta - \Gamma_{1/2} = (\delta - 1/2)t \geq 0$ deterministically. The data dependent term improves this baseline only if $\underline{\delta} > 1/2$ (and $t > 0$). Positivity of the population witness does not require observing any target samples.

The argument uses the model's exact identity, which holds throughout $0 \leq t \leq 1/8$, rather than the generic empirical theorem, whose radius $p_{\min}^2/8$ is $1/2048$ here; the upper bounds in (66) keep their stated radius $1/512$. The event is for fixed $N$, not an arbitrary stopping time.

The probability guarantee in Proposition G.1 is over target observations; it is not a posterior guarantee conditional on a favorable certificate.

### G.8 Which structure the finite decision requires

The finite theorem concerns the exact tied architecture, the specified factor caps and the direct output adapter, where $\operatorname{rank} \Delta W = \operatorname{rank} E$ because the adapter acts directly on the context indicator and outputs hidden logits; a nonlinear or context dependent layer after the adapter can change that image. Since $t$ also changes the base network and the observed head, $\rho$ is not the rank of all weight changes.

**A two edge perturbation changes the orders.** Add a path whose effective contribution is $\epsilon t^2 x_2$ to the second hidden head, with fixed coefficient $\epsilon$. Then

$$C_\epsilon(t) = C(t) + \epsilon t^2 e_2 e_2^\top, \qquad \det C_\epsilon(t) = t^4(\epsilon + t^3), \qquad t^{-2} C_\epsilon(t) \longrightarrow \operatorname{diag}(1, \epsilon). \tag{67}$$

For every fixed $\epsilon \neq 0$, the limiting matrix is nonsingular. Continuity of singular values gives both orders equal to two, with sorted leading coefficients $\max(1, |\epsilon|)$ and $\min(1, |\epsilon|)$. At $\epsilon = 0$ the orders are instead $(2, 5)$. Consequently an arbitrarily small fixed perturbation of this type does not preserve the fifth order tail as $t \downarrow 0$. The unperturbed finite constants are not uniform robustness guarantees over such paths.

If $\epsilon < 0$, the positive rank drop root $t = (-\epsilon)^{1/3}$ lies in the declared interval $[0, 1/8]$ exactly when $-1/512 \leq \epsilon < 0$, and lies in its interior when the first inequality is strict. At that point $C_\epsilon(t) = v(t)v(t)^\top$, so the same rank one factors $U = -v(t), V = v(t)$ cancel it. For an explicit finite budget counterexample to exporting the original zero audit separation, take $\epsilon = -1/8000$, $t = 1/20$, and $\delta = 1/2$. Then $S = 0$, $B = 0$, and (53) gives

$$D = \mathcal{J}(1/20) \leq \frac{1}{800} < \frac{1}{512}, \qquad \Gamma_{1/2}(1/20, 0) \geq \frac{1}{40} - \frac{1}{800} = \frac{19}{800} > 0.$$

The factors remain within their caps. Rank one can therefore have positive zero audit gain in this perturbed architecture, although it cannot in the unperturbed theorem. Roots outside $[0, 1/8]$ make no statement about admissible finite actions.

**A narrower, exactly calculable skip variation.** If only the five edge skip coefficient changes to a fixed real $a$, so that $C_a(t) = v(t)v(t)^\top + at^5 e_2 e_2^\top$, the rank one factors $U = -v, V = v$ remain admissible and leave $S = at^5 e_2 e_2^\top$. Substitution in (12) gives the exact audit $\frac{1}{2}\mathcal{J}(at^5)$, trust cost $\mathcal{J}(t) + \frac{1}{2}\mathcal{J}(at^5)$ and gain $\delta t - \log\cosh t - \frac{1}{2}\{|a|t^5 + \log\cosh(at^5)\}$. Both costs are nondecreasing in $t \geq 0$, so any step verified against these exact functions is a feasible certificate. We assert neither the original constants nor the original gain monotonicity bound for arbitrary $a$; at $a = 0$ rank one cancels everything, which is why the original rank one zero audit bound cannot be retained. This is not robustness to arbitrary untied weights.

**Local rank theory versus finite parameter decisions.** The generic conditional logit class of Appendix F allows a rank constrained matrix $E$ rather than bounded factors. On the nonnegative-$t$ side near $(t, S) = (0, 0)$, $E = S - C(t)$ tends to zero and hence has $\|E\|_* < 1/4$; balanced factorization then realizes it in the tied class. Thus the classes agree on this local germ, which is enough to compare singular orders. It does not identify their complete finite feasible sets, establish the finite radius, or replace the explicit estimates in Appendices G.3–G.4. The derivative of $UV^\top$ at $U = V = 0$ is zero; the nonlinear image, not just a tangent space rank calculation, is needed for the witness.

**What partial observations do not learn.** The observed marginal of (10) depends only on $t$. Observed sign likelihood therefore cannot identify useful hidden factors $U, V$, let alone discover the tied architecture. The construction chooses factors from the known paths and certifies their consequences. Uniform contexts and a common target mean are also substantive assumptions: different target context weights would change the worst world hidden penalty while the model's audit weights remain uniform. The target data refinement in Appendix G.7 does not cover that different family or an unknown channel.

# H Proof of the certificate from observations

## H.1 From observations to attainable gain

Section 3 treats the observed law $r$ as known; here we estimate it from the same sample used to choose the update. One confidence event must therefore cover every candidate at once. Fix $K$, $p_0$, $n$ and $\eta \in (0,1)$ in advance, write $m_Y = |\mathcal{Y}|$ and $p_{\min} = \min_i p_0(i) > 0$, and let $Y_1, \dots, Y_n$ be independent draws from $r_* = Kp_*$ with empirical law $\widehat{r}_n$. Set

$$\varepsilon_n = \min\left\{2, \sqrt{\frac{2(m_Y \log 2 + \log(1/\eta))}{n}}\right\}, \qquad \mathcal{C}_n = \{p \in \Delta(\Omega) : \|Kp - \widehat{r}_n\|_1 \le \varepsilon_n\}, \quad (68)$$

and, when $\mathcal{C}_n$ is nonempty, let $\Gamma_n(q) = \min_{p \in \mathcal{C}_n} \mathbb{E}_p \log(q/p_0)$ with $G_n(b,\beta)$ its maximum over the same updates and budgets as in (6); if $\mathcal{C}_n$ is empty the rule returns $p_0$.

**Theorem H.1** (A uniform certificate from observations)**.** *For fixed $K, p_0$ and finite spaces put $\overline{b} = p_{\min}^2/8$ and $C = 4\sqrt{2}H_K/p_{\min}$, with $H_K$ the Hoffman constant of the fiber constraints of $K$ (Appendix H). On one event of probability at least $1-\eta$, the population fiber satisfies $\mathcal{C}_{r_*} \subseteq \mathcal{C}_n$ and, simultaneously for every positive $q$ with $D(q) = \mathrm{KL}(q\|p_0) \le \overline{b}$,*

$$0 \le \Gamma(q; r_*) - \Gamma_n(q) \le C\varepsilon_n \sqrt{D(q)}. \quad (69)$$

*Consequently, on that event, simultaneously for all $0 < b \le \overline{b}$ and all $\beta \ge 0$,*

$$[G(b,\beta; r_*) - C\varepsilon_n\sqrt{b}]_+ \le G_n(b,\beta) \le G(b,\beta; r_*). \quad (70)$$

*Neither constant depends on $n$, $\eta$, the budgets or the feasible target $r_*$. On that same event any measurably selected empirical maximizer has true improvement in KL risk at least $G_n(b,\beta)$.*

The confidence event concerns the observed law, not a selected update, so the same sample may choose the update and both budgets. Hoeffding's inequality (Hoeffding, 1963) covers the population fiber and Hoffman's error bound (Hoffman, 2003) compares it with the enlarged fiber. Their oscillation bound (73) holds for every positive $q$; only its passage to $\sqrt{D(q)}$ uses the radius $\overline{b}$. With $\eta$ replaced by $6\eta/(\pi^2 n^2)$ the event covers every $n$ of one stream at once, hence any almost surely finite stopping time (Corollary H.2).

With a population schedule of value $G(b, \lambda b^\alpha; r_*) \ge c_g b^{\gamma(\alpha)}$, the theorem shows that

$$n \ge \frac{8C^2}{c_g^2}\left[m_Y \log 2 + \log(1/\eta)\right] b^{1-2\gamma(\alpha)} \quad (71)$$

observations retain at least half of the population value. At the floor $\gamma = 1/2$ the sample size does not grow as $b$ shrinks and at larger exponents it does; the constant is loose, and the empirical certificate needs neither $H_K$ nor $c_g$. The scale is sufficient and not necessary: in the rank module the feasible witness has trust cost $O(T_\rho^2)$ rather than $b$, so a fixed sample size retains a fixed fraction of $G_\rho$ along every shrinking pair of budgets.

The sample size $n$ is fixed in this result. The simultaneous statements concern actions and budgets for this sample, not an arbitrary stopping time. All logarithms are natural.

**Coverage of the population fiber.** For each $v \in \{-1,1\}^{m_Y}$, $v_{Y_i} \in [-1,1]$. Hoeffding's inequality gives

$$\Pr\{v^\top(\widehat{r}_n - r_*) > e\} \le e^{-ne^2/2}.$$

Since $\|x\|_1 = \max_v v^\top x$, a union bound over $2^{m_Y}$ signs gives $\Pr\{\|\widehat{r}_n - r_*\|_1 > e\} \le 2^{m_Y} e^{-ne^2/2}$. If the untruncated radius is at least two, use the simplex diameter instead. Thus the event $\mathcal{E} = \{\|\widehat{r}_n - r_*\|_1 \le \varepsilon_n\}$ has probability at least $1-\eta$. Every $p \in \mathcal{C}_{r_*}$ belongs to $\mathcal{C}_n$ on this event, so $\mathcal{C}_n$ is nonempty there.

**A fiber error bound uniform in the target.** Let $d = |\Omega|$ and encode the compatible world constraints as

$$A = \begin{bmatrix} K \\ -K \\ \mathbf{1}^\top \\ -\mathbf{1}^\top \\ -I_d \end{bmatrix}, \qquad c(r) = \begin{bmatrix} r \\ -r \\ 1 \\ -1 \\ 0 \end{bmatrix}, \qquad \mathcal{C}_r = \{p : Ap \le c(r)\}.$$

Hoffman's inequality supplies a constant $H_K$ depending only on this fixed matrix and the norms, not on a consistent right hand side. For $p \in \Delta(\Omega)$ the normalization and nonnegativity residuals vanish, while the two channel residuals have total positive part $\|Kp - r\|_1$. Consequently

$$\operatorname{dist}_1(p, \mathcal{C}_r) \le H_K \|Kp - r\|_1. \tag{72}$$

This includes redundant channel rows and targets on the boundary of $K\Delta(\Omega)$.

For completeness, the usual projection proof makes the constant explicit. The normal from $p$ to its Euclidean projection $y$ onto $\{z : Az \le c\}$ is a nonnegative combination $A_I^\top \mu$ of linearly independent active rows, so $\|p-y\|_2^2 = \mu^\top(A_I p - c_I) \le \sigma_I^{-1}\|p-y\|_2 \|(Ap-c)_+\|_2$, with $\sigma_I$ the smallest singular value of $A_I^\top$. Norm conversion gives $H_K = \sqrt{d}\max_I \sigma_I^{-1}$ over nonempty independent row subsets of $A$, independent of $c$. For rational $K$ this is a finite computation, with no polynomial running time asserted, and it can be large for an ill conditioned $K$; computing $\Gamma_n(q)$ is a separate linear program that needs neither $H_K$ nor $c_g$.

**Uniform comparison of worst world gains.** On $\mathcal{E}$, each $p \in \mathcal{C}_n$ has $\|Kp - r_*\|_1 \le 2\varepsilon_n$. By (72) there is $p' \in \mathcal{C}_{r_*}$ with $\|p - p'\|_1 \le 2H_K\varepsilon_n$. For $f = \log(q/p_0)$ and any positive $q$, subtracting the midpoint of the largest and smallest coordinates of $f$ gives

$$|\mathbb{E}_p f - \mathbb{E}_{p'} f| \le \tfrac{1}{2}\operatorname{osc}(f)\|p - p'\|_1 \le H_K\varepsilon_n \operatorname{osc}(f).$$

Taking minima and using $\mathcal{C}_{r_*} \subseteq \mathcal{C}_n$ yields, on this same event for every positive $q$,

$$0 \le \Gamma(q; r_*) - \Gamma_n(q) \le H_K\varepsilon_n \operatorname{osc}\log(q/p_0). \tag{73}$$

No union bound over the action class is used.

On the same event the value obeys a radius free bound as well. For $b > 0$ and $\beta \ge 0$ let $h$ attain $G(b, \beta; r_*)$; since $\log(q_h/p_0) = h - \psi(h)$ has the oscillation of $h$, (73) gives

$$G(b,\beta;r_*) \ge G_n(b,\beta) \ge \Gamma_n(q_h) \ge G(b,\beta;r_*) - H_K\varepsilon_n \operatorname{osc}(h),$$

so the empirical value falls short of the population value by at most $H_K\varepsilon_n$ times the largest oscillation of an available score with trust cost at most $b$.

Put $p_{\min} = \min_i p_0(i) > 0$ and choose $\overline{b} = p_{\min}^2/8$. Pinsker's inequality gives $\|q - p_0\|_1 \le \sqrt{2D(q)} \le p_{\min}/2$, so $q_i \ge p_0(i)/2$. The mean value theorem for the logarithm gives

$$|\log(q_i/p_0(i))| \le \frac{2|q_i - p_0(i)|}{p_{\min}} \le \frac{2\sqrt{2D(q)}}{p_{\min}}, \qquad \operatorname{osc}\log(q/p_0) \le \frac{4\sqrt{2D(q)}}{p_{\min}}.$$

Equation (69) follows with $C = 4\sqrt{2}H_K/p_{\min}$.

**Optimization and risk.** For nonempty $\mathcal{C}_n$, its compactness makes $\Gamma_n(q_h)$ continuous on the fixed compact score image. Both budget problems therefore attain their maxima. The feasible actions are identical in the empirical and population problems: the model, $K, p_0, D$ and $B$ are fixed. Taking maxima in (69), and using $\Gamma_n(p_0) = 0$, proves (70) simultaneously for every stated budget pair, including $\beta = 0$. For an empirical optimizer $q_n$, coverage of $p_*$ and the risk identity give

$$\mathrm{KL}(p_*\|p_0) - \mathrm{KL}(p_*\|q_n) \ge \Gamma_n(q_n) = G_n(b,\beta).$$

The empirical histogram takes finitely many values at fixed $n$; choosing one optimizer for each histogram gives a measurable rule. If $\mathcal{C}_n$ is empty the rule returns $p_0$ without evaluating $\Gamma_n$; this branch is outside $\mathcal{E}$. Neither existence nor measurability implies efficient optimization.

**A sufficient sample size.** For a fixed positive population schedule, choose $c_g > 0$ and $b_0 > 0$ such that $G(b, \lambda b^\alpha; r_*) \geq c_g b^\gamma$ for $0 < b \leq b_0$. The constants here may depend on the fixed model, target, $\alpha$ and $\lambda$. Since $\varepsilon_n \leq \sqrt{2(m_Y \log 2 + \log(1/\eta))/n}$, it suffices that

$$n \geq \frac{8C^2}{c_g^2}(m_Y \log 2 + \log(1/\eta))b^{1-2\gamma}$$

to make $C\varepsilon_n\sqrt{b} \leq G/2$ and hence $G_n \geq G/2$. This proves (71) as a sufficient scale. There is no matching lower bound here, and $c_g$ is not asserted uniform over targets. Since $c_g$ may be unknown, this expression is not a data computable stopping rule. The reported value $\Gamma_n(q)$ is the certificate for an actual returned update.

**Why the generic sample scale can be conservative.** For the nonzero tail rank construction in Appendix F, write $T = T_\rho(b, \beta) > 0$. With fixed constants, its feasible SVD witness satisfies, uniformly for small $b$ and every $\beta > 0$, $D(q^{\mathrm{svd}}) \leq d_0 T^2$, $\Gamma(q^{\mathrm{svd}}; r) \geq \delta T - a_0 T^2$ and $G_\rho(b, \beta) \leq \delta T + a_1 T^2$, so (69) gives $G_{n,\rho}(b, \beta) \geq \delta T - a_0 T^2 - C\sqrt{d_0}\varepsilon_n T$ for the empirical value restricted to the same rank model. A fixed $\varepsilon_n \leq \delta/(4C\sqrt{d_0})$ and a small budget neighborhood with $(a_0 + a_1/2)T \leq \delta/4$ then give $G_{n,\rho} \geq G_\rho/2$, so a fixed $n$ suffices across shrinking budgets, whereas (71) uses $D \leq b$ and loses this information. No necessity conclusion follows.

### H.2 Validity at a finite random sample size

**Corollary H.2** (A fixed stream stopping certificate)**.** *Let $(Y_i)_{i\geq 1}$ be one iid stream from the fixed law $r_* = Kp_*$, with known fixed $K, p_0$. Fix $0 < \eta < 1$, put $\eta_n = 6\eta/(\pi^2 n^2)$, and use*

$$\varepsilon_n^{\mathrm{all}} = \min\left\{2, \sqrt{\frac{2[m_Y \log 2 + \log(\pi^2 n^2/(6\eta))]}{n}}\right\}$$

*in the empirical set at time $n \geq 1$. With probability at least $1-\eta$, the containment and both gain/value inequalities of Theorem H.1 hold simultaneously for every $n$, every small positive action and every stated budget pair, with $\varepsilon_n^{\mathrm{all}}$ in place of $\varepsilon_n$. In particular they hold at any almost surely finite random index $\tau \geq 1$, including a stopping time.*

*Proof.* Apply Theorem H.1 at each $n$ with failure probability $\eta_n$. The intersection of its events has probability at least $1 - \sum_{n\geq 1}\eta_n = 1 - \eta$. On that event the inequalities hold for every index, so substituting a finite random index is valid pathwise. The radius now contains $2\log n$; this corollary does not retain the fixed-$n$ sample size expression without that change. □

Adaptively choosing which population to observe can change the conditional law of $Y_i$ and invalidate the iid premise; that problem requires a different confidence construction.

## I The computational construction

### I.1 Computing the value is hard at a fixed jet

A *regular cubic score graph* is an injective score map $h(u) = \sum_i u_i L_i + \sum_j F_j(u)N_j$ in which the $L_i \in L$ are linearly independent, the $N_j$ lie in $L^\perp$ and the $F_j$ are homogeneous cubics. Its second order jet is its Taylor polynomial through degree two, which here is just the visible linear term.

For a 3SAT formula $\Phi$ with $n$ variables and $m$ clauses put $d = 1 + 2n + m$, give each of $X, Y_1, \ldots, Y_n, N_1, \ldots, N_{n+m}$ a pair of states on which that feature is $\pm 1$ and the others vanish, and let $p_0$ be uniform on the $2d$ states. The channel reveals the individual states of the $X$ and $Y_i$ pairs and only the labels of the $N_j$ pairs; with $\delta = 1/2$ the target gives the two $X$ states masses $(1 \pm \delta)/(2d)$ and leaves every other output at its reference mass. The formula enters only through $n + m$ cubic residuals on $(t, z) \in [-1/4, 1/4]^{n+1}$, a Boolean residual $F_i = tz_i(z_i - t)$ per variable and, per clause, the product of its three unsatisfied literal factors, $t - z_i$ for a positive and $z_i$ for a negative literal. The available scores are

$$h(t, z) = tX + \sum_i z_i Y_i + \sum_j F_j(t, z)N_j, \tag{74}$$

a regular cubic score graph with value $G_\Phi$.

**Theorem I.1** (A gap at fixed budgets)**.** *The construction is computable in polynomial time and all its data have polynomial bit length. For fixed $n$ and $m$, every instance has the same reference, target, channel, tangent space and second order jet. For every $0 < b \le (256d^2)^{-1}$ and every $\beta \ge 0$,*

$$G_\Phi(b,\beta) \asymp \begin{cases} \sqrt{b}, & \Phi \textit{ satisfiable}, \\ \min\{\sqrt{b}, \beta^{1/6}\}, & \Phi \textit{ unsatisfiable}, \end{cases} \tag{75}$$

*with constants depending on $d$ but not on either budget. Hence it is NP hard to decide whether positive gain survives a zero audit, or whether the exponent $\gamma_\Phi(4)$ on the schedule $\beta = \lambda b^4$, for any fixed $\lambda > 0$, equals $1/2$ or $2/3$; at $b = 2^{-60}d^{-6}$ and $\beta = b^4$ the two cases differ by a factor above $200$, so approximating the value within a factor of two is NP hard as well.*

This appendix proves Theorem I.1 and, in its last paragraph, Corollary 5.2. Membership in NP is not asserted. The endpoint and exponent claims assert hardness of exact decisions; the factor two claim asserts hardness of approximating the value.

**The model and its encoding.** The visible coefficients recover $(t, z)$, so this is a regular graph over $H = \mathrm{span}\{X, Y_i\} \subseteq L$ with cubic displacement in $L^\perp$. For fixed $n, m$, its two jet is independent of the formula. There are $2d$ explicit states, a polynomial size zero one channel, and at most eight expanded monomials per clause; all rational coefficients and probabilities have polynomial bit length. The compact parameter cube ensures that the value is attained.

**Exact probability identities.** Write $a = (t, z_1, \dots, z_n, F_1, \dots, F_{n+m})$ and $Z = \sum_{k=1}^{d} \cosh(a_k)$. Then

$$\begin{aligned} \psi &= \log(Z/d), & D(h) &= \frac{\sum_k a_k \sinh(a_k)}{Z} - \psi, \\ \Gamma(h) &= \frac{\delta t - \sum_j |F_j|}{d} - \psi, & B(h) &= \frac{\sum_j \cosh(F_j)\mathcal{J}(F_j)}{Z}. \end{aligned} \tag{76}$$

Indeed, pair $k$ has model mass $\cosh(a_k)/Z$ and conditional probabilities $e^{\pm a_k}/(2\cosh(a_k))$. The information projection resets each hidden conditional to uniform and preserves all channel outputs, giving the audit formula. The target fixes the visible contribution to $\delta t/d$, while each hidden pair has world mass $1/d$ that can be placed on its lower score. This gives the exact robust gain. Since $Z \ge d$, positive gain requires $t > 0$, and $\Gamma \le \delta t/d$.

**Uniform local bounds.** On the whole cube, $|F_i| \le 1/32$ for Boolean residuals and $|F_j| \le 1/8$ for clauses. Disjoint supports give $\|h\|_\infty \le 1/4$. We claim

$$\frac{1}{4d}\sum_k a_k^2 \le D(h) \le \frac{1}{d}\sum_k a_k^2, \qquad B(h) \ge \frac{1}{8d}\sum_j F_j^2. \tag{77}$$

For $\varphi(u) = \log \mathbb{E}_0 e^{uh}$, integration by parts yields $D(h) = \int_0^1 u \,\mathrm{Var}_{q_{uh}}(h)\, du$. The tilted density relative to $p_0$ is between $e^{-1/2}$ and $e^{1/2}$. Using $\mathrm{Var}_q(h) = \inf_c \mathbb{E}_q(h-c)^2$ bounds the tilted variance by these same multiples of $\mathrm{Var}_{p_0}(h) = d^{-1}\sum_k a_k^2$. Integration gives the KL inequalities because $e^{-1/2}/2 > 1/4$ and $e^{1/2}/2 < 1$. For the audit, $Z < 2d$ and $\cosh(F_j) \ge 1$. Also $\mathcal{J}(v) \ge v^2/4$ for $|v| \le 1/8$, by integrating $\mathcal{J}'(v) = v\,\mathrm{sech}^2\, v$ and using evenness. This proves (77). We shall also use $\mathcal{J}(v) \le v^2/2$ and $\cosh v \le 1 + v^2$ on $|v| \le 1/4$.

**An upper bound over every unsatisfiable action.** If $\Phi$ is unsatisfiable, then for every $t > 0$ and every $z \in \mathbb{R}^n$,

$$\max_j |F_j(t,z)| \ge 3t^3/16. \tag{78}$$

To see this, put $w_i = z_i/t$. If some $w_i$ is at distance at least $1/4$ from $\{0,1\}$, then $|w_i(w_i - 1)| \ge 3/16$, proving the claim via its Boolean residual. Otherwise round each $w_i$ to its unique nearest Boolean value. Some clause is false at this assignment. All three of its unsatisfied literal factors, divided by $t$,

have magnitude at least $3/4$, so its residual is at least $27t^3/64$. Equality in the distance test belongs to the first case. In particular, no boundedness assumption on $z/t$ was needed. Equations (76)–(78) give

$$G_\Phi(b,\beta) \le \min\left\{\sqrt{b/d},\ \frac{1}{2d}\left(\frac{2048d\beta}{9}\right)^{1/6}\right\} \quad (\Phi \text{ unsatisfiable}). \tag{79}$$

Indeed, the two constraints imply $t \le 2\sqrt{db}$ and $t^6 \le 2048d\beta/9$; nonpositive gain actions cannot improve the baseline. At $\beta = 0$, the optimal gain is zero. There may still be safe points with $t = 0$; their gain is nonpositive.

**Matching feasible updates.** Let $b_0 = (256d^2)^{-1}$. If $u \in \{0,1\}^n$ satisfies $\Phi$, set $t = \sqrt{b}$ and $z_i = tu_i$. All residuals vanish, $B = 0$, and $D \le (n+1)t^2/d \le b$. Likewise, $\psi \le (n+1)t^2/d \le b$, so for $b \le b_0$,

$$\frac{\sqrt{b}}{4d} \le G_\Phi(b,\beta) \le \sqrt{b/d} \quad (\Phi \text{ satisfiable},\ \beta \ge 0). \tag{80}$$

Here the lower bound uses $\delta t/d - b \ge \delta t/(2d)$, and the upper bound uses total KL as before. Conversely, positive gain with $B = 0$ requires $t > 0$ and all $F_j = 0$. The Boolean equations force $z_i/t \in \{0,1\}$ and the clause equations make this assignment satisfying. This proves the hard endpoint claim.

For a matching lower bound in the unsatisfiable case, set $z = 0$. Each clause residual is zero or $t^3$; write $q \le m \le d$ for the number of nonzero ones. For $0 \le t \le 1/(16d)$,

$$D \le (t^2 + qt^6)/d \le 2t^2, \qquad B \le t^6, \qquad \Gamma \ge \frac{t}{d}\{\delta - qt^2 - t - qt^5\} \ge \frac{t}{4d}.$$

The audit estimate follows from $Z \ge d$, $\cosh(t^3) < 2$, and $\mathcal{J}(t^3) \le t^6/2$; the gain uses $\psi \le (t^2 + qt^6)/d$ and $qt^2 + t + qt^5 \le 1/256 + 1/16 + 1/16^5 < 1/4$. Thus $t = \frac{1}{2}\min\{\sqrt{b}, \beta^{1/6}\}$ is feasible when $b \le b_0$ and gives $G_\Phi \ge (8d)^{-1}\min\{\sqrt{b}, \beta^{1/6}\}$. Together with (79) and (80), this proves the full two budget profile, including $\beta = 0$. For $\beta = \lambda b^\alpha$ with fixed $\lambda > 0$, the exponents are respectively $1/2$ and $\max\{1/2, \alpha/6\}$.

**A rational finite budget gap.** We use standard binary Turing encoding for the rational data, budgets and score polynomials. The approximation task returns a rational $\widehat{G}$ satisfying $G_\Phi/2 \le \widehat{G} \le 2G_\Phi$; the exponentiated model probabilities need not be rational. Choose the explicit positive rational budgets and threshold

$$b = 2^{-60}d^{-6}, \qquad \beta = b^4, \qquad L_d = 2^{-32}d^{-4}. \tag{81}$$

For (81), $b < b_0$ and the satisfiable lower bound is $L_d$. In the unsatisfiable case let $U = (2d)^{-1}(2048d\beta/9)^{1/6}$. Direct substitution gives

$$\frac{U}{L_d} = 2\left(\frac{2048}{9}\right)^{1/6} 2^{-10}d^{-5/6} < \frac{5}{1024} < \frac{1}{200}. \tag{82}$$

The strict bound uses $(2048/9)^{1/6} < 5/2$ and requires no floating point comparison. A factor two estimate is at least $L_d/2$ on satisfiable inputs and below $L_d/100$ on unsatisfiable inputs; the rational threshold $L_d/4$ separates them. An estimate with absolute error $L_d/8$ is separated instead by $L_d/2$. These comparisons decide 3SAT in polynomial time if either promised estimator runs in polynomial time. The budgets and thresholds have $O(\log d)$ bits, completing the approximation hardness claim.

**Proof of Corollary 5.2.** Fix $n$ and $m$, a satisfiable $\Phi$ and an unsatisfiable $\Phi'$. Their reference, target, channel, parameter cube and second order jet coincide, so a rule whose output depends only on these data and the budgets returns one value $\widehat{G}$ for both. At the budgets (81), the bounds (80), (79) and (82) give $G_\Phi \ge L_d > 200\,U \ge 200\,G_{\Phi'}$, and $G_{\Phi'} > 0$ by the lower bound of the unsatisfiable case in Theorem I.1. If $\widehat{G}$ were within a factor 14 of both values, then $G_\Phi \le 14\,\widehat{G} \le 196\,G_{\Phi'} < G_\Phi$; hence $\widehat{G}$ misses one of them by a factor above 14. Along $\beta = b^4$, (75) gives $G_\Phi/G_{\Phi'} \asymp b^{-1/6}$ with constants depending on $d$, so no fixed factor suffices at all small budgets. Neither conclusion uses a complexity assumption: both concern what a rule sees, whereas Theorem I.1 concerns polynomial time rules that see the whole instance. Such pairs exist, for instance with $n = 4$ and $m = 8$: the eight clauses on $x_1, x_2, x_3$ with all sign patterns are unsatisfiable, and eight distinct clauses that contain the literal $x_4$ are satisfied by setting $x_4$ true.

# J WHAT THE OBSERVATIONAL GUARANTEE PRESERVES

Alignment at a finite interface separates observed data improvement from preservation of reference prediction structure. The calculations below specify their relation.

**The classical unconstrained endpoint.** Suppose $\mathcal{C}_r$ contains a strictly positive distribution, so $\pi_r > 0$. For every positive $q$,

$$\Gamma(q; r) \le \mathbb{E}_{\pi_r} \log(q/p_0) = \mathrm{KL}(\pi_r \| p_0) - \mathrm{KL}(\pi_r \| q).$$

The logarithm $\log(\pi_r/p_0)$ lies in $\mathrm{range}(K^\top)$, so its expectation is constant on $\mathcal{C}_r$. Hence $q = \pi_r$ attains the bound. If all positive distributions are allowed and $b \ge \mathrm{KL}(\pi_r \| p_0)$, the optimum with zero audit is $\mathrm{KL}(\pi_r \| p_0)$: the classical robust Bayes equalizer (Grünwald & Dawid, 2004, Sections 7.3 and 10.2.2). The problem in Section 3 concerns a restricted local image with both budgets shrinking, where that endpoint need not be available.

## J.1 PARTITION OBSERVATIONS

Let $K$ reveal the group $y = T_K(i)$. Empty observation groups may be removed. Every compatible world has group law $r$ and arbitrary conditional laws within groups. Therefore

$$\begin{aligned} \Gamma(q; r) &= \sum_y r_y \log \frac{(Kq)_y}{(Kp_0)_y} + \sum_y r_y \min_{i: T_K(i) = y} \log \frac{q(i \mid y)}{p_0(i \mid y)} \\ &\le \mathrm{KL}(r \| Kp_0) - \mathrm{KL}(r \| Kq). \end{aligned} \tag{83}$$

The minimum conditional ratio is at most one, with equality only when the two conditionals coincide. Thus the second sum is nonpositive. If $B(q) = 0$, all conditionals coincide by (22), and equality holds. Conversely, equality forces this agreement only for groups with $r_y > 0$. The audit weights all groups by $(Kq)_y > 0$, including groups to which the target assigns zero mass. Thus the converse requires positive target mass on every nonempty group.

## J.2 A ZERO AUDIT DOES NOT FIX EVERY UNIDENTIFIED MOMENT

Let $\ell, n_1, n_2$ be independent uniform signs on eight states under the reference, observed through the stochastic channel

$$K(y \mid \ell, n_1, n_2) = \frac{1 + y\chi(\ell + \theta n_1)}{2}, \qquad r(y) = \frac{1 + y\chi\delta}{2}, \qquad y \in \{-1, 1\}, \tag{84}$$

with $0 < \theta < 1$, $0 < \chi < 1/(1+\theta)$ and a target mean $0 < \delta < 1$. Put

$$g = \ell + \theta n_1, \qquad f = n_1 - \theta \ell, \qquad h = tg.$$

Here $L = \mathrm{span}(g)$, $f \in L^\perp$ and $h \in L$, so $B(q_h) = 0$ exactly. Nevertheless independence under the reference gives

$$\mathbb{E}_{q_h} f = \tanh(\theta t) - \theta \tanh t = \frac{\theta(1 - \theta^2)}{3} t^3 + O(t^5) \ne 0 \tag{85}$$

for sufficiently small positive $t$. This moment is not identified by the observation. Indeed, $p_* = p_0(1 + \delta\ell)$ is positive, and $p_* \pm a p_0 f$ remain probability distributions for sufficiently small $a > 0$. Their observations agree because $\mathbb{E}_{p_0} f = 0$ and $\mathbb{E}_{p_0} fg = 0$, but their $f$ moments differ by $2a\mathbb{E}_{p_0} f^2 > 0$. At the same time,

$$\Gamma(q_h; r) = \delta t - \log\cosh t - \log\cosh(\theta t) > 0$$

for small positive $t$. Positive worst world gain and zero projection audit are consequently compatible with changing an unidentified moment. The audit controls reference projection structure; it does not make the reference's latent beliefs true.

## J.3 THE OUTPUT LEVEL, A STOCHASTIC CHANNEL AND MEASUREMENT ERROR

For positive semantic $q$, define its canonical lift

$$\mathcal{L}(q)(z) = Q_0(z) \frac{q(T(z))}{p_0(T(z))}. \tag{86}$$

It changes the frequencies of semantic states and preserves the reference conditional output distribution within each state.

**Proposition J.1** (Canonical lift and projection audit)**.** *The law $\mathcal{L}(q)$ is the unique minimum KL output law with semantic pushforward $q$. For positive $q, w$,*

$$\mathrm{KL}(\mathcal{L}(q)\|Q_0) = \mathrm{KL}(q\|p_0), \qquad \mathrm{KL}(\mathcal{L}(q)\|\mathcal{L}(w)) = \mathrm{KL}(q\|w). \tag{87}$$

*The minimum KL output law with observation $Kq$ is $\Pi_{Kq} = \mathcal{L}(\pi_{Kq})$, so its full output projection audit is*

$$\mathrm{KL}(\mathcal{L}(q)\|\Pi_{Kq}) = B(q). \tag{88}$$

*If an output target $P_*$ has pushforward $p_*$ and $\mathrm{KL}(P_*\|Q_0) < \infty$, then*

$$\mathrm{KL}(P_*\|Q_0) - \mathrm{KL}(P_*\|\mathcal{L}(q)) = \mathbb{E}_{p_*} \log(q/p_0). \tag{89}$$

*Proof.* For any output law $P$ with pushforward $q$, the conditional KL chain rule is

$$\mathrm{KL}(P\|Q_0) = \mathrm{KL}(q\|p_0) + \sum_i q_i \,\mathrm{KL}(P(\cdot \mid T = i)\|Q_0(\cdot \mid T = i)).$$

This identity holds with extended nonnegative values, including when $P$ charges baseline null outputs. Each conditional term is minimized uniquely at the reference conditional law. This proves the minimum lift assertion and the first identity; applying the same calculation between two lifts gives the second. Under the observation constraint, first minimize the conditional terms, then minimize semantic KL over $Kw = Kq$. The resulting law is $\mathcal{L}(\pi_{Kq})$, and the second identity gives the full output audit. Finally, $\log[\mathcal{L}(q)/Q_0]$ is the bounded function $\log(q/p_0) \circ T$ on the baseline support. Finite target KL makes both target divergences finite, so subtracting them gives (89). Without that condition the log ratio expectation remains defined for $P_* \ll Q_0$, but a difference of two infinite KL values is not a risk improvement. □

**Zero audit does not transfer latent gain through a stochastic channel.** Take $p_0 = (1,1,1)/3$, $p_* = (1/2, 1/3, 1/6)$, and

$$K = \begin{pmatrix} 3/5 & 1/10 & 3/10 \\ 1/10 & 3/5 & 3/10 \\ 3/10 & 3/10 & 2/5 \end{pmatrix}, \qquad q = (110, 121, 100)/331.$$

The positive channel has eigenvalues $1, 1/2, 1/10$, hence each fiber is a singleton and $B(q) = 0$; the centered score $h = (0, 1, -1)\log(11/10)$ lies in $L$ because $K$ is invertible. Direct calculation gives

$$\Gamma(q; Kp_*) = \tfrac{1}{6}\log(11/10) - \log(331/330) > 2/165 > 0.$$

But $Kp_0 = (1,1,1)/3$, $Kp_* = (23, 18, 19)/60$, and $Kq = (1081, 1136, 1093)/3310$. Using $\log x \le x - 1$ yields

$$\Gamma_{\mathrm{obs}}(q; Kp_*) \le \sum_y (Kp_*)_y \left[\frac{(Kq)_y}{(Kp_0)_y} - 1\right] = -61/33100 < 0.$$

The strict latent bound follows from $x/(1+x) < \log(1+x) < x$ for $x > 0$. Thus positive worst world gain at zero audit need not improve the observed risk, even through a positive invertible channel.

**Ignoring error can reverse the certificate.** On $\Omega = \{-1, 1\}$ let $p_0$ be uniform and the true world have $\mathbb{E}X = -1/2$. A proxy reports $Z = -X$, so $\mathbb{E}Z = 1/2$. If this label is incorrectly treated as an error free observation of $X$, the update $q_t(x) = e^{tx}/(2\cosh t)$ receives the purported gain $t/2 - \log\cosh t > 0$ for small $t > 0$. Its actual gain is $-t/2 - \log\cosh t < 0$, despite arbitrarily small update KL. More labels only estimate the wrong measurement law more precisely. A fitted or context dependent labeling mechanism requires identification and uncertainty control beyond the known channel observation theorem; sample splitting does not supply them.

## K AUDITING REVISIONS STATED BY LANGUAGE MODELS

**Protocol.** The interface, record, shifts and instructions are those of Section 7, with $r = (W/100, 1 - W/100)$; 299 of the 300 stated priors are positive and the other is excluded. The models are `gpt-4o-mini-2024-07-18`, `gpt-4.1-mini-2025-04-14`, `gpt-4.1-2025-04-14`, `gpt-4o-2024-08-06` at temperature zero and `gpt-5-mini-2025-08-07` at its default with low reasoning effort; the two rewordings list the cases in reverse order and describe the record as a soil moisture sensor.

**Sampled laws.** Every prior and revision prompt of the four temperature controlled models (main wording, record alone, the $W$ of Section 7) was sampled ten times at temperature one in one request. Table 3 compares the mean of the ten sampled laws, which approximates the law the model outputs, with the single samples and the greedy answer. The sampled priors have a standard deviation of the wet frequency between 0.059 and 0.082 across the ten draws.

Table 3: Sampled laws, ten samples per prompt at temperature one, 80 revisions per model. Mean law: the average of the ten sampled laws. Same direction: all ten samples move $P(\text{rain} \mid \text{wet})$ the same way. $\ell_1$: distance between the greedy answer and the mean law.

| Model | mean law $\mathcal{B} > 0.005$ | mean law $\mathcal{B}$ | sample $\mathcal{B}$ | mean law $\Gamma < 0$ | sd wet | sd $P(\text{rain} \mid \text{wet})$ | same dir. | $\ell_1$ |
|---|---|---|---|---|---|---|---|---|
| `gpt-4o-mini` | 70/80 | 0.058 | 0.094 | 61/80 | 0.104 | 0.091 | 38/80 | 0.24 |
| `gpt-4.1-mini` | 79/80 | 0.122 | 0.139 | 69/80 | 0.040 | 0.050 | 61/80 | 0.13 |
| `gpt-4.1` | 59/80 | 0.047 | 0.051 | 54/80 | 0.024 | 0.052 | 48/80 | 0.09 |
| `gpt-4o` | 64/80 | 0.028 | 0.040 | 47/80 | 0.037 | 0.064 | 30/80 | 0.10 |

**Qwen2.5-1.5B-Instruct.** For `Qwen2.5-1.5B-Instruct` the four state law is read from token probabilities rather than from generated text: three yes/no questions (did it rain; given rain, did the sprinkler run; given no rain, did the sprinkler run) give $P(\text{rain})$, $P(\text{sprinkler} \mid \text{rain})$ and $P(\text{sprinkler} \mid \text{no rain})$ as the renormalized probabilities of the yes and no tokens (median coverage 0.985 of the next token law), and the chain rule gives the joint law; the record enters all three questions. A single letter answer is degenerate at this size (the model puts 97% on D whatever the context) and is not used. Prompted revisions use the twenty contexts and four shifts of Section 7.

Parameter edits train a LoRA $\Delta W = BA$ of rank one or eight on the down projection of the last MLP block, shared by the three questions and the twenty contexts, by Adam on the mean over contexts of $-[W_c \log q_c(\text{wet}) + (1 - W_c) \log q_c(\text{dry})] + \lambda \,\mathrm{KL}(q_c \| p_{0,c})$ for 300 steps, three seeds, $\lambda \in \{0, 0.1, 1\}$, learning rate $2 \cdot 10^{-3}$, $A$ Gaussian with variance $1/d_{\text{ff}}$ and $B = 0$ at initialization, scaling $1/\rho$, gradient norm clipped at one, trained and audited on the same twenty contexts; everything after the last attention block acts position by position, so the activations leaving that block are cached and the edited laws are exact. Table 4 reports both.

For the drag, each trained edit is scaled by $t \in [2^{-8}, 1]$, the changes of the centered within wet log conditionals across the twenty contexts form a $20 \times 3$ matrix $C(t)$, and the orders $\mu_j$ of its two singular values are fitted on $t \leq 1/16$: every trained edit gives $\mu_1 = \mu_2 = 1.00$ (first order; the fitted values over all 72 edits lie between 0.99 and 1.01), the second singular value is 7% to 12% of the first in the medians of the six rank and $\lambda$ groups (6% to 44% over single edits), and the best rank one correction of $C(1)$ shared by all contexts lowers the audit, averaged over contexts and taken at its median over the twelve rank one $\lambda = 0$ edits, from 0.0136 to 0.00007. A random direction of the same Frobenius size moves the wet log odds 66 times less than the trained edit and has audit below $10^{-4}$; per unit of observed movement its drag is comparable (3.8 against 3.2), so the Frobenius size of an edit does not set its drag. The SVD correction subtracts the rank one component from the output conditional log law while keeping the wet mass fixed; it does not establish that a neural parameter edit can realize the corrected law.

**SmolLM2-1.7B-Instruct.** We use `HuggingFaceTB/SmolLM2-1.7B-Instruct` with the same yes/no interface and training protocol. For the prior queries, the yes/no tokens cover a median 0.904 of the next token law (minimum 0.866). Its prior wet frequencies range from 7.7% to 13.3%, so both negative shifts are clipped to a target wet frequency of 5%. The 80 prompted records therefore contain 60 distinct prompts, and the two negative training conditions have the same target. Prompted wet frequencies range from 3.132% to 5.968% across all tested records, below their corresponding priors.

The 72 stored runs yield 144 fitted singular value orders, from 0.995 to 1.004 after rounding to three decimal places. Without a KL penalty, all 240 output laws at the two positive shifts have audit above 0.005 (two ranks, three seeds, twenty contexts). For each rank at $\lambda = 0$, we average audits over contexts, then take separate before and after medians across twelve runs (four nominal shifts, three seeds). The rank one SVD correction of $C(1)$ lowers these medians by 98.1% for rank one and 98.0% for rank eight; it acts on output conditional log laws, with no additional adapter trained.

Table 4: Qwen2.5-1.5B-Instruct: prompted revisions of the exact output law (80 in total, 20 per shift) and LoRA edits of the last MLP block (240 per row: 20 contexts, 4 shifts, 3 seeds). Fit: mean absolute difference between the wet frequency and the record. $\Delta$ rain: mean change of $P(\text{rain} \mid \text{wet})$.

| Revision | $\mathcal{B} > 0.005$ | median $\mathcal{B}$ | $\Gamma < 0$ | fit | mean $D$ | $\Delta$ rain |
|---|---|---|---|---|---|---|
| prompt, shift $-0.30$ | 1/20 | 0.0017 | 11/20 | 0.079 | 0.030 | -0.046 |
| prompt, shift $-0.15$ | 1/20 | 0.0017 | 20/20 | 0.065 | 0.020 | -0.044 |
| prompt, shift $+0.15$ | 3/20 | 0.0023 | 20/20 | 0.158 | 0.006 | -0.046 |
| prompt, shift $+0.30$ | 5/20 | 0.0028 | 20/20 | 0.280 | 0.008 | -0.049 |
| LoRA rank 1, $\lambda = 0$ | 186/240 | 0.0128 | 57/240 | 0.003 | 0.133 | +0.098 |
| LoRA rank 8, $\lambda = 0$ | 186/240 | 0.0136 | 58/240 | 0.003 | 0.134 | +0.099 |
| LoRA rank 1, $\lambda = 0.1$ | 23/240 | 0.0007 | 47/240 | 0.015 | 0.101 | -0.005 |
| LoRA rank 8, $\lambda = 0.1$ | 30/240 | 0.0010 | 48/240 | 0.016 | 0.099 | -0.006 |
| LoRA rank 1, $\lambda = 1$ | 0/240 | 0.0001 | 0/240 | 0.096 | 0.027 | -0.001 |
| LoRA rank 8, $\lambda = 1$ | 0/240 | 0.0001 | 9/240 | 0.095 | 0.029 | -0.001 |

**Further counts.** With the record alone, at least 77 of 80 revisions in each of the twelve cells of the four models at temperature zero and the three wordings exceed the rounding level. The request to change as little as possible accounts for 56 of the 74 revisions under the main wording that give zero probability to a case the record does not exclude, whose worst world gain is $-\infty$. Reading a stated $0\%$ as $0.5\%$ instead changes the number of negative worst world gains under the main wording from 695 to 691. Resampling the twenty contexts $10{,}000$ times gives 95% percentile intervals of $0.98$ to $1.0$ for 634 of 640 and 260 of 263, $0.77$ to $0.87$ for 216 of 263, $0.54$ to $0.64$ for 234 of 396, $0.11$ to $0.16$ for 52 of 396, and $0.93$ to $1.0$ and $0.85$ to $1.0$ for the direction counts of `gpt-4.1`, 39 and 37 of 40. At audit thresholds of $0.01$ and $0.02$, 514 and 385 of the 640 revisions and 211 and 166 of the 263 within five points of the record remain above.

Table 5: Revisions under the main wording, per model and instruction. Matched: recorded wet frequency reproduced within one point. Zero audit: audit at most that of Jeffrey's update written in whole percents.

| Model | Instruction | Matched | Zero audit | Median audit | $\Gamma < 0$ | $\Gamma = -\infty$ |
|---|---|---|---|---|---|---|
| `gpt-4o-mini` | record | 0/80 | 0 | 0.050 | 66 | 0 |
| | change little | 0/80 | 1 | 0.029 | 60 | 20 |
| | keep proportions | 18/80 | 42 | 0.000 | 60 | 0 |
| `gpt-4.1-mini` | record | 5/80 | 0 | 0.087 | 67 | 6 |
| | change little | 14/80 | 1 | 0.034 | 64 | 9 |
| | keep proportions | 14/80 | 55 | 0.000 | 46 | 1 |
| `gpt-4.1` | record | 58/80 | 0 | 0.018 | 53 | 1 |
| | change little | 63/80 | 3 | 0.025 | 60 | 9 |
| | keep proportions | 60/80 | 18 | 0.011 | 44 | 6 |
| `gpt-4o` | record | 36/80 | 1 | 0.017 | 50 | 0 |
| | change little | 39/80 | 0 | 0.033 | 59 | 3 |
| | keep proportions | 40/80 | 44 | 0.000 | 30 | 0 |
| `gpt-5-mini` | record | 66/76 | 66 | 0.000 | 12 | 4 |
| | change little | 73/76 | 47 | 0.000 | 24 | 15 |
| | keep proportions | 76/76 | 75 | 0.000 | 0 | 0 |

Table 6: Record alone, per model and wording (main, reversed order, reworded with a sensor), 80 revisions per cell (76 for `gpt-5-mini` under the main wording). Above rounding: audit larger than that of Jeffrey's update written in whole percents. $\Delta$ rain: mean change in the probability that a wet morning had rain when the record showed more wet mornings than the model had stated; Jeffrey's update leaves it unchanged.

| | Above rounding | | | $\Delta$ rain | | |
|---|---|---|---|---|---|---|
| Model | main | reversed | sensor | main | reversed | sensor |
| `gpt-4o-mini` | 80 | 80 | 77 | 0.13 | 0.28 | $-0.04$ |
| `gpt-4.1-mini` | 80 | 80 | 80 | 0.27 | 0.14 | 0.11 |
| `gpt-4.1` | 80 | 80 | 78 | $-0.19$ | 0.21 | $-0.01$ |
| `gpt-4o` | 79 | 80 | 80 | 0.02 | 0.11 | 0.07 |
| `gpt-5-mini` | 10 | 1 | 22 | 0.01 | 0.00 | $-0.06$ |

## L NUMERICAL CHECKS OF THE FINITE BOUNDS AND OF THE FOUR PHASE EXPONENTS

**The tied model.** For the image (11) and the exact costs and gain (12), the value $\mathsf{G}_\rho(b, \beta; 1/2)$ was maximized directly over $t$ and the factors: a scan over $t$ (at rank zero, where $E = 0$, the feasible endpoint in $t$ is located by bisection), for each $t$ a grid over the singular directions and values of $E$ followed by a derivative free local search started from the grid optimum and from the witnesses (15), and a reevaluation of the best point in fifty digit arithmetic. Table 7 places the values inside the intervals of Theorem 4.1; the position is $(\mathsf{G}_\rho - \text{lower})/(\text{upper} - \text{lower})$. At $b = 1/5000$ the optimum of ranks zero and one lies between 0.66 and 0.98 of the way from the lower to the upper bound over the nine audit budgets computed; at $b = 1/512$, the edge of the radius, it lies between 0.02 and 0.98, close to the witness lower bound when the audit constraint is slack and within $10^{-5}$ of the upper bound for rank zero when the audit binds. At rank two the search comes within a relative $3 \cdot 10^{-5}$ of the exact value. At $b = 1/5000$ the numerical values reproduce the three decisions of Table 2: the minimum rank whose value reaches $g = 1/200$ is zero, one and two at $\beta = 10^{-4}$, $10^{-10}$ and $10^{-24}$. These are floating point searches, not certificates; the certificates are the rational inequalities of Appendix G.

Table 7: Numerical optimum of the tied model against the bounds of Theorem 4.1, $\delta = 1/2$, at the three budget pairs of Table 2; eleven further pairs at $b = 1/5000$ and $b = 1/512$ are in the supplement.

| $b$ | $\beta$ | rank | $\mathsf{G}_\rho$ | lower | upper | $\mathsf{G}_\rho/(\delta T_\rho)$ | position |
|---|---|---|---|---|---|---|---|
| $2.0 \cdot 10^{-4}$ | $1 \cdot 10^{-4}$ | 0 | $9.59160 \cdot 10^{-3}$ | $8.64000 \cdot 10^{-3}$ | $1.00300 \cdot 10^{-2}$ | 0.9592 | 0.685 |
| $2.0 \cdot 10^{-4}$ | $1 \cdot 10^{-4}$ | 1 | $9.80072 \cdot 10^{-3}$ | $9.36000 \cdot 10^{-3}$ | $1.00300 \cdot 10^{-2}$ | 0.9801 | 0.658 |
| $2.0 \cdot 10^{-4}$ | $1 \cdot 10^{-4}$ | 2 | $9.80072 \cdot 10^{-3}$ | $9.37500 \cdot 10^{-3}$ | $1.00300 \cdot 10^{-2}$ | 0.9801 | 0.650 |
| $2.0 \cdot 10^{-4}$ | $1 \cdot 10^{-10}$ | 0 | $2.21585 \cdot 10^{-3}$ | $1.93196 \cdot 10^{-3}$ | $2.24278 \cdot 10^{-3}$ | 0.9910 | 0.913 |
| $2.0 \cdot 10^{-4}$ | $1 \cdot 10^{-10}$ | 1 | $9.80072 \cdot 10^{-3}$ | $9.36000 \cdot 10^{-3}$ | $1.00300 \cdot 10^{-2}$ | 0.9801 | 0.658 |
| $2.0 \cdot 10^{-4}$ | $1 \cdot 10^{-10}$ | 2 | $9.80072 \cdot 10^{-3}$ | $9.37500 \cdot 10^{-3}$ | $1.00300 \cdot 10^{-2}$ | 0.9801 | 0.650 |
| $2.0 \cdot 10^{-4}$ | $1 \cdot 10^{-24}$ | 0 | $7.07105 \cdot 10^{-7}$ | $6.10940 \cdot 10^{-7}$ | $7.09228 \cdot 10^{-7}$ | 1.0000 | 0.978 |
| $2.0 \cdot 10^{-4}$ | $1 \cdot 10^{-24}$ | 1 | $2.27576 \cdot 10^{-3}$ | $2.14019 \cdot 10^{-3}$ | $2.29338 \cdot 10^{-3}$ | 0.9953 | 0.885 |
| $2.0 \cdot 10^{-4}$ | $1 \cdot 10^{-24}$ | 2 | $9.80072 \cdot 10^{-3}$ | $9.37500 \cdot 10^{-3}$ | $1.00300 \cdot 10^{-2}$ | 0.9801 | 0.650 |

**The four phase example.** Along $\beta = b^\alpha$ the value of the example of Appendix E was computed in eighty digit arithmetic on the two arcs that Proposition E.1 proves optimal up to constant factors, the axis $t = 0$ and the cancellation curve $s = \tau^3$, $t = \tau^2$, for $b = 10^{-6}$ to $10^{-14}$. The local exponent $\Delta \log G / \Delta \log b$ between consecutive values of $b$ was compared with the proved formula for eight values of $\alpha$ per $k$ (the table is in the supplement); the last two slopes differ from it by at most 0.0006, the remaining difference being the approach to the asymptotic regime.

## M STANDING ASSUMPTIONS AND NOTATION

The general frontier concerns a fixed compact subanalytic score image containing the reference, inside a neighborhood where both costs are analytic (Assumption 2.1). The bounded factor construction, generic rank module and computational construction have their own action sets. A result's constants and sampling assumptions do not transfer between these sets without a separate argument.

**Bounded factor model and scalar searches.** The font distinguishes the finite value $\mathsf{G}_\rho$ from the general value $G$ and its hard audit restriction $G_0$. The notation $G_\rho$ without the sans serif font is reserved for the generic module in Appendix F. Their common coefficient germ on the nonnegative-$t$ side does not identify their finite feasible sets. The symbols of the tied model are defined in Section 4 and Appendix G where they are introduced.

**Other local notation.** Symbols that belong to one appendix (the module orders $\mu_j$, $\nu_\rho$, $a_\rho$ and $\rho_{\max}$ of Appendix F; the observation count $n$, radius $\varepsilon_n$, set $\mathcal{C}_n$ and constants $H_K$, $p_{\min}$, $c_g$ of Appendix H; $\operatorname{osc}(f) = \max f - \min f$) are defined where they are used and keep their meaning within that appendix; $p$ denotes a compatible latent world, not the model update.

Table 8: Core notation. Further model specific symbols are described in this section.

| Symbol | Meaning |
|---|---|
| $\Omega, \mathcal{Y}$ | Finite latent prediction and observation spaces |
| $p_0, p_*, q$ | Reference law, true world and updated prediction |
| $K, r, \mathcal{C}_r$ | Channel, observed target law and compatible-world fiber |
| $h, \psi, \mathsf{S}$ | Centered score, log partition and score space |
| $\langle h, g\rangle_0, \|h\|_0$ | Reference-weighted inner product and norm, $\|h\|_0^2 = \mathbb{E}_{p_0} h^2$ |
| $L$ | Centered observable scores, $\mathsf{S} \cap \mathrm{range}(K^\top)$ |
| $\pi_r, \mathcal{B}$ | Information projection and projection audit of a law |
| $D, B$ | Trust cost and projection audit as functions of a score |
| $\Gamma, \Gamma_{\mathrm{obs}}$ | Worst-world gain and observed-data gain |
| $\mathcal{M}, b, \beta$ | Available score image, trust budget and audit budget |
| $G(b,\beta), G_0(b)$ | General optimal gain and its zero-audit value |
| $H, V_i, \gamma$ | Max–min monomial envelope, rational exponent sets and gain exponent |
| $\alpha, \lambda$ | Power-schedule parameters, $\beta = \lambda b^\alpha$ |

Table 9: What each result assumes, uses and delivers. Distinct rows need not describe a single realizable model or a shared data source.

| Result | Actions, known quantities and data access | Output and guarantee |
|---|---|---|
| Envelope (Thm. 3.1) | Compact subanalytic $\mathcal{M}$; known $K, p_0, r$; no data required. | Uniform local comparison of $G$ with $H$ and the zero-audit endpoint. No numerical constants, radius or efficient optimizer are supplied. |
| Tied finite model (Thm. 4.1) | Specified sixteen-state network; uniform reference, known partition and target family $r_\delta$, $1/2 \le \delta < 1$; individually bounded factors. No data required for the prior certificate. | Bounds for every factor choice at $0 < b \le 1/512$, explicit feasible factors and steps, and sufficient/impossible rank tests. Exact scalar endpoints are distinct from validated finite returned points. |
| Optional target refinement (Sec. G.7) | The tied model and fixed-$N$ iid observations from its known target family; the common mean $\delta$ is refined, not the hidden factors. | One event for all actions, ranks and budgets; optional gain and rank exclusion certificates. The known $\delta \ge 1/2$ already certifies a baseline. |
| Generic rank module (Thm. 5.1) | Specified analytic conditional-logit displacement, uniform context and coefficient corrections of rank $\le \rho$; no data required. | Critical ranks, exponents and leading constants. The validity threshold and remainder constant are nonexplicit; the finite feasible set is not the bounded-factor image. |
| Observation certificate (Thm. H.1) | The envelope's action set; known $K, p_0$; iid observation counts from $r_*$. Corollary H.2 handles stopping times. | Empirical lower bounds simultaneous over actions and budgets. The $\sqrt{D}$ form requires $D \le p_{\min}^2/8$; the oscillation form holds for every positive update. |
| Computational construction (Thm. I.1) | Fixed cubic score graphs with explicitly encoded known instances; no data required. | Value-approximation hardness at a common second-order jet. Jet-only rules cannot give a fixed-factor approximation across these instances. |