MULTICLASS PROJECTED CALIBRATION: ESTIMATION LIMITS AND CROSS-ENTROPY STATIONARITY

Anonymous authors
Paper under double-blind review

ABSTRACT

When training linear classification heads on frozen feature representations, a vanishing cross-entropy gradient signals optimization progress but does not, by itself, guarantee calibrated probabilities. We identify exactly when a zero population cross-entropy gradient guarantees multiclass calibration for finite-support input distributions. We show that this guarantee holds for every conditional label distribution if and only if linear combinations of the features can isolate each group of inputs assigned the same probability vector. If they cannot, even a global population minimizer can remain miscalibrated. We then move from exact stationarity to finite training and derive a high-probability certificate for projected smooth calibration, a multiclass calibration metric that tests differences between labels and predicted probabilities using bounded Lipschitz functions of one-dimensional probability projections. The certificate separates optimization error, representation error from calibration tests that the frozen features cannot accurately represent, and finite-sample uncertainty. We also derive a computable upper bound on this representation error using only the model predictions and frozen features. Finally, we characterize the finite-sample uncertainty incurred when calibration is evaluated on held-out data and construct an exactly calibrated predictor whose empirical score matches the scaling of our upper bound up to logarithmic factors. Synthetic controls and frozen CNN features on CIFAR-10 illustrate how feature representation affects calibration, the finite-training certificate, and the statistical limits of empirical calibration auditing.

1 INTRODUCTION

A common adaptation strategy, often called linear probing, is to freeze a pretrained representation and train only a linear classification head (Radford et al., 2021). When the resulting probabilities are used as confidence estimates for thresholding, abstention, or downstream decision-making, it matters whether they accurately reflect class frequencies. A small empirical cross-entropy gradient indicates optimization progress within this fixed feature model, but it does not by itself establish such calibration. Indeed, miscalibration can persist even at a global population minimum. This raises a practical question: *when does gradient convergence provide evidence of calibration?*

By definition, a model is calibrated if its predicted probability vector matches the conditional class frequencies among inputs receiving that prediction (Vaicenavicius et al., 2019). To understand what the gradient tells us about this condition, consider the label residual: the difference between a one-hot label and the predicted probability vector. At population stationarity, meaning a zero population cross-entropy gradient, each feature-weighted residual has zero mean. Calibration instead requires the residual to have zero mean conditional on the prediction. When the input distribution has finitely many support points, this means cancellation within each group of identical predictions. If linear combinations of the features cannot isolate these groups, the gradient can miss systematic probability errors.

The representation problem thus reduces to a concrete question of expressivity: can linear combinations of the learned features capture the mean residual of each prediction group? We first provide an exact theoretical characterization of this property for finite-support distributions. We then bridge the gap between population-level stationarity and the finite-sample regime, explicitly accounting for rep-

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.

resentational constraints and finite-sample uncertainty. Specifically, our three primary contributions are as follows:

An exact geometric criterion for calibration (Theorem 2). In the finite-support setting, we prove that population stationarity guarantees calibration uniformly across all conditional label distributions if and only if the feature span contains the indicator function for every equal-prediction group, i.e., every group of inputs assigned the same probability vector. This condition ensures that a linear combination of features can isolate each group's average residual, without requiring the representation to capture variation among inputs within the group. Conversely, if this condition fails, we construct a valid conditional label distribution for which the predictor remains a global population minimizer yet is strictly miscalibrated.

A computable calibration certificate for finite training (Theorem 3). To quantify deviations from exact stationarity, we use projected smooth calibration (Gopalan et al., 2024), which measures miscalibration through witness functions—bounded Lipschitz tests of one-dimensional probability projections that reweight label residuals. For bounded features, we derive a high-probability upper bound on the population calibration metric that separates optimization error, representation error from calibration tests that the features cannot accurately represent, and finite-sample uncertainty. We also derive a computable upper bound on the representation error through the anchor operator in Proposition D.1. The resulting certificate therefore accounts for sources of calibration error that a small gradient alone does not control.

Fundamental statistical limits of empirical auditing (Theorem 1). We characterize the finite-sample uncertainty of empirical projected smooth calibration when a fixed predictor is evaluated on an independent audit dataset. We prove an upper bound on the gap between the population and empirical scores and construct an exactly calibrated predictor whose expected empirical score has the same scaling, up to logarithmic factors. Thus, even a perfectly calibrated predictor can receive a positive empirical score because the witness is selected using finite audit data. This result quantifies how much of an observed empirical calibration score can arise from finite-sample fluctuations alone.

These results characterize the feature representation required for calibration at population stationarity, quantify the additional representation and optimization errors that arise after finite training, and isolate the uncertainty caused by finite data. In Section 7, we illustrate these effects through controlled experiments and linear probing on frozen CIFAR-10 features, including a retrospective population-level analysis. Our finite-sample guarantees assume that the feature representation is bounded and fixed independently of the training data, so the theory formally applies to linear probing on frozen representations.

2 Related Work

Calibration has long been connected to proper loss minimization: unrestricted population minimization of a strictly proper loss recovers the conditional label distribution (Gneiting & Raftery, 2007), while neural networks trained on finite samples can nevertheless be miscalibrated and are often corrected post hoc, for example by temperature scaling (Guo et al., 2017). Smooth calibration provides a statistically tractable relaxation of calibration error in the binary setting (Błasiok et al., 2023a). For multiclass prediction, Gopalan et al. (2024) introduce projected smooth calibration and develop computationally efficient recalibration algorithms for this metric. We study this metric with the projection-budget restriction in (3) and address different questions: we characterize its finite-sample fluctuations when a fixed predictor is audited on held-out data, and we study when cross-entropy stationarity of a fixed multiclass feature model is sufficient to guarantee calibration.

A separate line of work asks when minimizing a proper loss over a restricted predictor family yields calibration. In the binary setting, Błasiok et al. (2023b) show that local optimality under Lipschitz post-processing of the predictions implies smooth calibration. Błasiok et al. (2024) relate population squared-loss minimization in sufficiently expressive neural-network classes to multicalibration, which requires calibrated predictions across a prescribed collection of subgroups or auditors, while Fujisawa & Futami (2025) derive finite-sample smooth-calibration guarantees for L_2-regularized empirical risk minimization, with concrete results for kernel models. Our setting and mechanism differ: for a fixed multiclass feature representation and a linear cross-entropy head, we characterize exactly when a zero gradient with respect to the head parameters implies calibration on finite support, and we show

that the determining property is whether the features can isolate groups of inputs that receive the same prediction.

Recent work also connects calibration to training dynamics and generalization. Futami & Nitanda (2026) establish uniform-convergence and functional-gradient guarantees for binary smooth calibration and analyze several learning algorithms, while Wu et al. (2025) study early stopping in well-specified high-dimensional logistic regression and show statistical advantages over asymptotic gradient descent. Our finite-training certificate instead starts from the empirical parameter gradient of a multiclass linear head and explicitly separates optimization error, calibration-test representation error, and finite-sample uncertainty. The coefficient-constrained approximation in Theorem 3 and the computable anchor operator in Proposition D.1 quantify the part of calibration that the frozen features cannot detect through the training gradient.

3 Problem Formulation

We consider a multiclass classification setting with $K \geq 2$ classes. Let $[K] = \{1, \ldots, K\}$, and assume the data $(X, Y) \sim P$ are drawn jointly from an input space $\mathcal{X}$ and a discrete label space. We represent the labels as standard one-hot basis vectors, such that $Y \in \{e_1, \ldots, e_K\} \subset \mathbb{R}^K$. We denote the marginal distribution over inputs as P_X and the true conditional class probabilities as $\eta(x) = \mathbb{E}[Y \mid X = x]$.

Given a dataset $S = \{(X_i, Y_i)\}_{i=1}^n$ containing n independent and identically distributed (i.i.d.) observations, we define the population and empirical expectations for any integrable function f as:

$$P(f) = \mathbb{E}[f(X, Y)], \qquad P_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i, Y_i). \tag{1}$$

These expectation operators apply entrywise to vectors and matrices. We assume throughout that the relevant function classes satisfy the measurability conditions needed for the stated expectations and probability bounds. A comprehensive index of our notation is provided in Appendix H.

A classification model is formalized as a measurable logit map $g : \mathcal{X} \to \mathbb{R}^K$. This map induces a predicted probability vector $p_g = \mathrm{softmax}(g)$ lying in the standard simplex $\Delta^{K-1} = \{p \in \mathbb{R}^K : p_j \geq 0,\ \sum_j p_j = 1\}$. Consequently, the predicted probability for each class is strictly positive: $p_{g,j}(x) = e^{g_j(x)} / \sum_{k=1}^K e^{g_k(x)} > 0$. We say such a model achieves *exact calibration* if its predictions match the true conditional frequencies almost surely: $\mathbb{E}[Y \mid p_g(X)] = p_g(X)$. Intuitively, among the subset of all inputs assigned a specific probability vector p, the population conditional distribution of their labels must equal p.

To train this model, we minimize the standard multiclass cross-entropy loss, defined over logits z and one-hot targets y as $\ell(z, y) = \log \sum_{j=1}^K e^{z_j} - \langle y, z \rangle$. This induces the corresponding population and empirical risks, which we assume are finite:

$$L(g) = P(\ell(g, Y)), \qquad \widehat{L}(g) = P_n(\ell(g, Y)). \tag{2}$$

Since $\nabla_z \ell(z, y) = \mathrm{softmax}(z) - y$, the logit gradient is the negative of the label residual $y - \mathrm{softmax}(z)$, directly connecting cross-entropy stationarity to residual cancellation.

Why projected residual tests? We use projected smooth calibration (Gopalan et al., 2024) because its witness-function formulation supports both finite-sample complexity bounds and approximation by the classifier features. Each scalar witness weights a label residual through a Lipschitz function of a one-dimensional projection of the predicted probability vector, allowing the metric to test structured multiclass miscalibration while retaining a form that is tractable for our analysis. For $m \geq 0$, we define

$$\mathcal{H}_m = \left\{ p \mapsto \phi(a^\top p) :\ \|a\|_\infty \leq 1,\ \|a\|_2^2 \leq m,\ \phi : [-1, 1] \to [-1, 1],\ \mathrm{Lip}(\phi) \leq 1 \right\}, \tag{3}$$

which defines a scalar calibration test by first projecting the predicted probability vector p onto the one-dimensional score $a^\top p$ and then applying a bounded 1-Lipschitz outer function ϕ to that score; the projection budget m controls the Euclidean size of the projection direction a and therefore the richness of the allowed tests. In Equation (3), the cube constraint keeps $a^\top p \in [-1, 1]$, and

1-Lipschitz means $|\phi(u) - \phi(v)| \le |u - v|$. The projection budget m bounds squared projection length, not dimension or the number of projections; for example, $m = 1$ permits $a = e_j$. Increasing m enlarges the witness class and its statistical cost, while direction selection still incurs dependence on K.

Each coordinate of $w \in \mathcal{W}_m = \mathcal{H}_m^K$ may choose its own projection direction and outer function. For $F_{g,w}(x, y) = \langle y - p_g(x), w(p_g(x))\rangle$, we define the population projected smooth calibration error, which measures the largest absolute residual–witness correlation under the data distribution, and its empirical counterpart, which measures the same quantity on the observed sample, respectively, as

$$\mathrm{psCE}_m(g) = \sup_{w \in \mathcal{W}_m} \big|P(F_{g,w})\big|, \qquad \widehat{\mathrm{psCE}}_m(g) = \sup_{w \in \mathcal{W}_m} \big|P_n(F_{g,w})\big|. \tag{4}$$

The metrics in Equation (4) lie in $[0, 2]$, and exact calibration makes every population correlation zero. The full witness class comprises all tests in this bounded geometry; it contains only constants at $m = 0$ and saturates at $m \ge K$. In Section 6, we approximate these witnesses with features to connect their residual correlations to the training gradient.

4 Estimation Error

Before analyzing the training dynamics, we first isolate the statistical limits of empirical auditing by considering a predictor fixed independently of the evaluation sample. Even in this idealized setting, computing the empirical calibration score requires taking a supremum that selects a witness from a fixed class using the data, which introduces sampling fluctuations from witness selection.

To quantify these fluctuations, we define the effective projection budget $\mu = \min\{m, K\}$ and the truncated complexity parameter $r = \min\{\mu, n\}$. We capture the statistical complexity of selecting a projected witness through the factors

$$A_\mu(n) = 1 + \sqrt{r}\log(en/r), \qquad \mathcal{C}_\mu(n) = 1 + 12(A_\mu(n) - 1),$$

where $\sqrt{r}\log(en/r) = 0$ at $r = 0$. Using these complexity measures, we bound the sampling error for a fixed predictor and construct an exactly calibrated predictor with matching expected fluctuations up to logarithmic factors in the stated regime.

Theorem 1. *Let S be an i.i.d. sample of size n, and let $g : \mathcal{X} \to \mathbb{R}^K$ be a measurable logit function fixed independently of S. For any fixed $m \ge 0$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$|\,\mathrm{psCE}_m(g) - \widehat{\mathrm{psCE}}_m(g)| \le 2\mathcal{C}_\mu(n)\sqrt{\frac{2K}{n}} + 6\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{5}$$

Conversely, for every odd $K \ge 3$, $1 \le m \le K - 1$, and $n \ge 2K^2$, there exists an input–label distribution and an exactly calibrated predictor whose predicted class probabilities are all strictly positive such that

$$\mathbb{E}\widehat{\mathrm{psCE}}_m \ge \frac{1}{12}\sqrt{\frac{mK}{n}}, \qquad \mathrm{psCE}_m = 0. \tag{6}$$

Proof. See Appendix A. □

Theorem 1 shows that selecting the projection direction for a one-dimensional test still incurs a multiclass statistical cost. In the regime where the lower bound applies, the upper bound in Equation (5) scales as $\widetilde{O}(\sqrt{mK/n})$, matching the $\Omega(\sqrt{mK/n})$ rate of the exactly calibrated construction up to logarithmic factors. In particular, this construction has a positive expected empirical score despite exact population calibration. The result characterizes the sampling behavior of this empirical supremum; it is not a minimax lower bound for all calibration estimators.

5 When Stationarity Implies Calibration

Let $\varphi : \mathcal{X} \to \mathbb{R}^d$ be a measurable feature map fixed independently of the sample S, bounded by a deterministic constant $\sup_x \|\varphi(x)\|_2^2 \le \Lambda$ with $\Lambda > 0$. Features learned from independent auxiliary

data can be treated conditionally as fixed; they must also satisfy the stated deterministic bound. For a linear head parameterized by $W \in \mathbb{R}^{K\times d}$, we define the logits $g_W = W\varphi$ and the predicted probabilities $p_W = \mathrm{softmax}(g_W)$. We also write $L(W)$ and $\widehat{L}(W)$ for the population and empirical cross-entropy risks, and similarly use W as the argument of the calibration metrics. Differentiating the cross-entropy loss with respect to the head parameters gives

$$\nabla L(W) = \mathbb{E}\big[(p_W(X) - Y)\varphi(X)^\top\big], \qquad \nabla\widehat{L}(W) = P_n\big((p_W(X) - Y)\varphi(X)^\top\big). \tag{7}$$

Setting the population gradient to zero enforces the global cancellation of feature-weighted residuals. Calibration requires the label residual to have zero mean conditional on the prediction.

To bridge these two conditions, we restrict our focus in Theorem 2 to marginal distributions P_X with finite support $\{x_1, \ldots, x_M\}$ and strictly positive masses $\pi_s > 0$, where M is the support size (distinct from the sample size n). For a fixed weight matrix W, let $\Phi \in \mathbb{R}^{M\times d}$ denote the feature matrix with rows $\varphi(x_s)^\top$. We partition the support into J equal-prediction groups I_j: the classifier assigns the same probability vector to every input within a given group.

Let $(v_j)_s = \mathbf{1}\{s \in I_j\}$ be the indicator vector for group I_j. The column space $\mathrm{col}(\Phi) = \{\Phi b : b \in \mathbb{R}^d\}$ represents the set of all scalar functions over the support that can be expressed via linear feature combinations. If an indicator v_j lies within this feature span—meaning there exists a vector b_j such that $v_j = \Phi b_j$—the corresponding feature combination acts as a precise switch that isolates group j. Multiplying the stationary population gradient by b_j directly yields the calibration condition for that specific group:

$$0 = \nabla L(W) b_j = \sum_{s\in I_j} \pi_s\big(p_W(x_s) - \eta(x_s)\big).$$

Thus, representing the equal-prediction group indicators lets the features test each group's mean residual. The features need not distinguish individual inputs *within* a group, because calibration permits their residuals to cancel.

Theorem 2. *For the fixed support, masses, features, and parameter matrix specified above, and for a fixed projection budget $m > 0$, the following conditions are equivalent:*

1. *For any conditional label distribution $\eta(x_s) \in \Delta^{K-1}$ $(s = 1, \ldots, M)$, stationarity of the loss, i.e., $\nabla L(W) = 0$, implies $\mathrm{psCE}_m(W) = 0$.*
2. *Each equal-prediction group indicator lies in the feature span, meaning $v_j \in \mathrm{col}(\Phi)$ for all $j \in [J]$.*

Proof. See Appendix B, where we prove the equivalence by relating witness evaluations to group indicators and constructing residuals orthogonal to the features when the criterion fails. □

In the stationarity-to-calibration implication of Theorem 2, the conditional label distribution varies, while the support masses, features, and parameter matrix W are held fixed; the loss and calibration metric are evaluated under the same conditional label distribution. Consequently, $\mathrm{rank}(\Phi) \geq J$ is necessary. When all M predictions are distinct, the criterion requires full row rank: $\mathrm{rank}(\Phi) = M$.

For any $m > 0$ on finite support, zero projected smooth calibration error is equivalent to exact calibration within each equal-prediction group. However, the rank condition alone is generally insufficient: the group indicators themselves must lie in the feature span. Failure of this criterion permits a miscalibrated stationary label distribution, but does not imply miscalibration for *every* conditional label distribution.

Stationarity without calibration. A three-class affine example exhibiting this failure is given in Appendix B.1: its two features cannot span the three distinct prediction-group indicators, and the resulting global population minimizer has zero gradient but positive projected calibration error.

6 Calibration After Finite Training

We extend Theorem 2 to finite training by accounting for nonzero gradients and imperfect witness representation. The bounds in this section do not require finite support.

Knowing that a witness can be represented is insufficient for a quantitative bound, because large feature coefficients can amplify the gradient contribution to the bound. To control this optimization-representation trade-off, we impose a coefficient budget $B \geq 0$, which bounds the Frobenius norm of the witness's feature coefficients. Furthermore, because the residual $Y - p_W$ always sums to zero, adding a common-coordinate shift $c(p)\mathbf{1}$ to a witness leaves its correlation unchanged. We therefore center the witness using the projection matrix $Q = I_K - \mathbf{1}\mathbf{1}^\top/K$. We quantify the worst-case feature approximation error over the witness class as:

$$\varepsilon_{m,S}(W; B) = \sup_{w \in \mathcal{W}_m} \inf_{\substack{A \in \mathbb{R}^{K \times d} \\ \|A\|_F \leq B}} \left(\frac{1}{n} \sum_{i=1}^{n} \|Qw(p_W(X_i)) - A\varphi(X_i)\|_2^2 \right)^{1/2}. \tag{8}$$

The inner infimum measures how well the available features can approximate a single centered witness under the coefficient budget B, while the supremum selects the hardest-to-fit witness in the class. Increasing B can reduce the approximation error but increases the multiplier of the gradient norm.

To formalize this trade-off, let $\rho_S(W) = \left(P_n \|Y - p_W\|_2^2\right)^{1/2} \leq \sqrt{2}$ denote the root-mean-square (RMS) label residual, and let $G_S(W) = \nabla \widehat{L}(W)$ denote the empirical gradient. Using the Frobenius inner product $\langle U, V \rangle_F = \mathrm{tr}(U^\top V)$, we decompose the empirical residual correlation for any witness $w \in \mathcal{W}_m$ and coefficient matrix A:

$$P_n(\langle Y - p_W, w(p_W) \rangle) = -\langle G_S(W), A \rangle_F + P_n(\langle Y - p_W, Qw(p_W) - A\varphi \rangle). \tag{9}$$

Applying the Cauchy–Schwarz inequality to Equation (9) bounds the represented component by $B\|G_S(W)\|_F$ and the unrepresented component by the RMS residual times the approximation error. Taking the infimum over coefficients A and the supremum over witnesses yields a deterministic, empirical calibration certificate:

$$\widehat{\mathrm{psCE}}_m(W) \leq B\|G_S(W)\|_F + \rho_S(W)\varepsilon_{m,S}(W; B). \tag{10}$$

We also call the empirical bound in Equation (10) a *sample certificate*. A *population certificate* bounds the population metric and must account for sampling uncertainty. For this step, we establish a uniform generalization bound over all weights within a deterministic radius $R \geq 0$. With $\delta \in (0, 1)$ and the logit Lipschitz constant $L_\mu = 1/\sqrt{2} + \min\{\sqrt{\mu}, \sqrt{2}\}$, we define the uniform statistical allowance:

$$\begin{aligned} \Delta_n(R, \delta) := 2 \min \left\{ 2, \mathcal{C}_\mu(n) \sqrt{\frac{2K}{n}} \right\} + \frac{2}{\sqrt{n}} \\ + 4\sqrt{\frac{2(K-1)d}{n} \log\left(1 + 2L_\mu R \sqrt{\Lambda n}\right)} + 2\sqrt{\frac{2\log(1/\delta)}{n}}. \end{aligned} \tag{11}$$

By Corollary A.4, with probability at least $1 - \delta$, the generalization gap $|\,\mathrm{psCE}_m(W) - \widehat{\mathrm{psCE}}_m(W)|$ is bounded by $\Delta_n(R, \delta)$ for every $\|W\|_F \leq R$.

To apply this to a trained model, consider full-batch gradient descent (GD) initialized at $W_0 = 0$, with a step size $0 < \alpha \leq 2/\Lambda$ and integer horizon $T \geq 1$ fixed independently of S:

$$W_{t+1} = W_t - \alpha G_t, \qquad G_t = \nabla \widehat{L}(W_t), \qquad t = 0, \ldots, T-1. \tag{12}$$

Standard convex optimization guarantees confine all iterates to a maximum radius $R_T = \sqrt{2\alpha T \log K}$ and ensure the final gradient $G_T = \nabla \widehat{L}(W_T)$ is bounded.

Theorem 3. *Fix the projection budget $m \geq 0$, the feature model, the step size α, the number of gradient-descent iterations T, and the confidence level δ as defined above. With probability at least $1 - \delta$, simultaneously for all $t = 0, \ldots, T$ and all budgets $B \geq 0$:*

$$\mathrm{psCE}_m(W_t) \leq B\|G_t\|_F + \rho_S(W_t)\varepsilon_{m,S}(W_t; B) + \Delta_n(R_T, \delta). \tag{13}$$

In particular, conditioning on the same high-probability event, for any $B \geq 0$:

$$\mathrm{psCE}_m(W_T) \leq B\sqrt{\frac{2\log K}{\alpha T}} + \sqrt{2}\,\varepsilon_{m,S}(W_T; B) + \Delta_n(R_T, \delta). \tag{14}$$

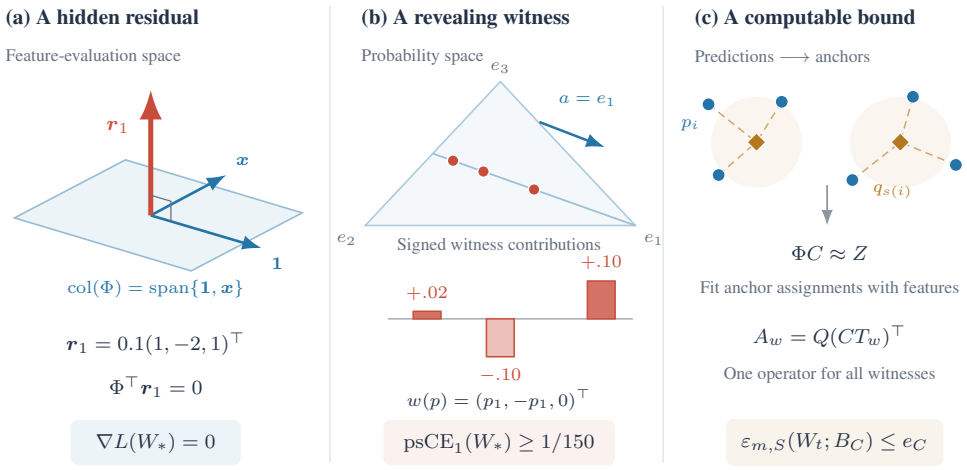

Figure 1: **From gradient cancellation to calibration certificates.** (a) A residual pattern can be orthogonal to the feature span, causing the population gradient to vanish even though groupwise calibration errors remain. (b) A feasible projected calibration witness detects this hidden error; for the three-point construction, its mean correlation is $1/150$ (Appendix B.1). (c) For finite training, the anchor operator approximates witness evaluations using the empirical features by fitting anchor assignments Z with ΦC and mapping anchor witness values T_w to coefficients A_w. Proposition D.1 provides the computable pair (B_C, e_C) used in the certificate of Theorem 3. The geometry and anchor assignments are schematic, whereas the witness contributions are exact.

Proof. See Appendix C. □

Theorem 3 separates gradient control, witness approximation, and sampling uncertainty. The explicit gradient allowance decays as $T^{-1/2}$, but approximation error may persist and $\Delta_n(R_T, \delta)$ need not decrease with T. Equation (13) uses the observed gradient and becomes numerically evaluable once we bound the approximation term. Equation (14) supplies an a priori gradient allowance; its approximation term still depends on the trained model.

From Theory to a Computable Certificate. Computing $\varepsilon_{m,S}$ directly requires optimizing over an infinite witness class. Appendix D constructs an *anchor operator* that provides a computable upper bound on this term by assigning predictions to finitely many probability anchors and approximating the resulting anchor-assignment indicators with the sample features; Proposition D.1 then produces a coefficient budget B_C and approximation bound e_C satisfying $\varepsilon_{m,S}(W; B_C) \leq e_C$. Figure 1 summarizes how this construction connects the representation criterion in Theorem 2 to the finite-training certificate in Theorem 3.

7 Empirical Evidence

We evaluate the certificates and their limitations through complementary empirical studies. First, we test the empirical certificate on frozen neural features. We then use controlled distributions to isolate the representation mechanism in Theorem 2 and examine which terms limit the finite-training certificate. We also illustrate the sampling fluctuations in Theorem 1. In Appendix G, we give the experimental protocols and numerical qualifications. The four studies reported in this section use float64 arithmetic without formal rounding-error certification.

7.1 Frozen neural features on CIFAR-10

We evaluate the anchor operator on CIFAR-10 (Krizhevsky, 2009) using five linear heads trained on a shared frozen CNN. The encoder is learned on auxiliary images disjoint from the head-training pool, and its normalized features are bounded. The heads have overlapping training subsets, so

Table 1: Frozen-feature CIFAR-10 results at $m = 1$. The two approximation columns bound $\varepsilon_{1,S}(W; B)$ at the stated budget; they are not calibration bounds themselves. The sample certificate upper-bounds the empirical metric over the complete witness class at $m = 1$. Heads share an encoder and overlapping training inputs.

Head	B	Approximation lower	Approximation upper	Sample certificate	Mean L^1 residual
0	15.988	0.2021	0.8390	0.5654	0.8272
1	16.115	0.2028	0.8584	0.5745	0.8183
2	16.168	0.2022	0.8484	0.5717	0.8283
3	16.339	0.2025	0.8516	0.5730	0.8256
4	16.148	0.2027	0.8293	0.5583	0.8241

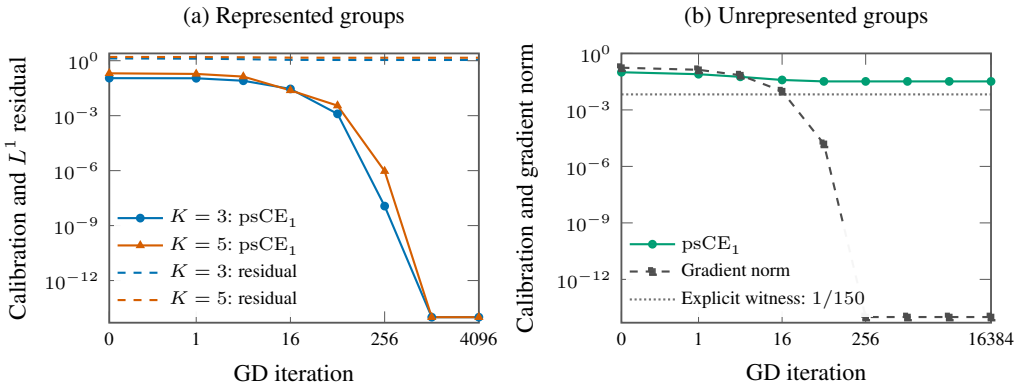

Figure 2: Represented prediction groups guarantee calibration at population stationarity; unrepresented groups can conceal miscalibration. (a) Population psCE_1 vanishes when the group indicators lie in the feature span, while mean individual L^1 label residuals persist. (b) In the affine model, the population gradient norm vanishes while psCE_1 remains positive; the dotted line denotes the witness value $1/150$ at the stationary limit. Values below 10^{-14} are thresholded.

their results are dependent. Appendix G.4 gives the architecture, data splits, training schedule, and operator construction. At $m = 1$, Table 1 reports the tightest observed-gradient empirical certificate for each head over the evaluated operator grid. Selecting this empirical bound requires no confidence correction because Equation (10) is a deterministic sample inequality. For all five heads, the empirical certificates in Table 1 are below the mean individual L^1 label-residual upper bound, while the positive approximation lower bounds show a representation cost at these coefficient budgets. Population transfer is separate: Appendix E gives the independent-audit guarantee, and Appendix G.4 reports the retrospective CIFAR-10 calculation and the conditions required for prospective coverage.

7.2 Why feature representation matters

To explicitly demonstrate the geometric mechanism driving Theorem 2, we analyze two controlled scenarios using exact finite-support integration and full population gradient descent (zero initialization, step size of 1). In a two-cell model, indicator features represent every equal-prediction group but omit a latent subclass that affects the conditional label law. In Figure 2a, the projected smooth calibration error falls below 10^{-14} after 4,096 updates, while the mean individual label residual $P(\|Y - p_W\|_1)$ remains at 1.0980 and 1.4504 for $K = 3, 5$, respectively. In this model, representing the prediction groups suffices for calibration at population stationarity, despite omitted label variation and positive Bayes-excess loss.

The three-point affine model in Appendix B.1 illustrates failure of the representation criterion. After 16,384 updates, the population gradient norm is 3.46×10^{-17}, while the population projected smooth calibration error remains $2\sqrt{6}/150 \approx 0.03266$ (Figure 2b). The feature span cannot represent

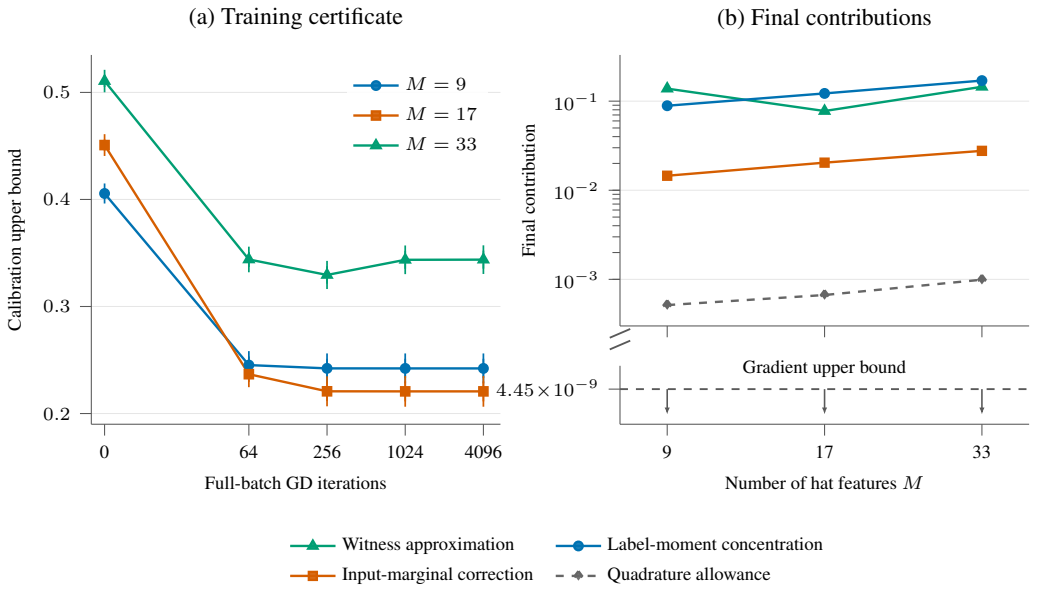

Figure 3: Known-marginal population certificates. (a) Means and sample standard deviations across three independent datasets. (b) A logarithmic view resolving the four non-gradient bound contributions. The separated lower band shows the common upper bound 4.45×10^{-9} on the final gradient contributions across all nine models, rather than per-model measurements.

the three equal-prediction group indicators, allowing feature-weighted residuals to cancel while groupwise mean residuals remain nonzero.

7.3 What limits the finite-training certificate?

We study a three-class model with a known uniform input marginal and piecewise-linear hat features at $M \in \{9, 17, 33\}$. The known-marginal specialization in Appendix F replaces the generic Δ_n term in Theorem 3 with a parameter-uniform label-moment bound. Figure 3 reports the resulting $m = 1$ certificates, and Appendix G.3 gives the training and quadrature protocols.

In Figure 3, the final certificates average 0.242, 0.221, and 0.344 for $M = 9, 17, 33$, respectively. All lie below the lower bounds on the mean individual label residual $P(\|Y - p_W\|_1)$, which exceed 1.26. Although the gradient contribution becomes small, increasing the number of features from $M = 17$ to $M = 33$ increases both the approximation contribution and the remaining terms (Table 4). Adding features alone therefore need not tighten the complete certificate.

An exactly calibrated construction also yields normalized feasible-witness scores between 0.148 and 0.190, consistent with the sampling scale in Theorem 1; Appendix G.1 gives the construction and full results.

8 Discussion and Conclusion

On finite input support, Theorem 2 characterizes exactly when cross-entropy stationarity implies calibration: the features must represent the equal-prediction group indicators, but need not recover label variation within those groups. Theorem 3 extends this picture to finite training by separating gradient control, witness-approximation error, and sampling uncertainty, while Proposition D.1 provides a computable upper bound on the approximation term. The controlled and frozen-feature studies illustrate both the usefulness and the limitations of these certificates. The exact criterion assumes finite support, the finite-training guarantees assume bounded features fixed independently of the training sample, and the statistical lower bound concerns the empirical supremum itself. Minimax estimation, continuous-support characterizations, and end-to-end training of the representation and classification head remain open directions.

AI-USE DISCLOSURE

We used OpenAI's ChatGPT for language editing, coding assistance, and identifying related works. The authors have fully reviewed and remain responsible for all content, mathematical statements, and experimental results.

ETHICS STATEMENT

We study theoretical calibration guarantees using controlled simulations and publicly available benchmark datasets. We do not collect new personal data or conduct interventions involving human participants. Our guarantees depend on the stated modeling, sampling, and numerical assumptions; they do not, by themselves, establish fairness or safety in downstream applications.

REPRODUCIBILITY STATEMENT

We provide proofs and experimental protocols in the appendices and supplementary material. The accompanying code archive includes all experimental implementations, fixed configurations, reference results, and reproduction instructions. Follow `README.md` and `REPRODUCTION_WORKFLOW.md` for setup and execution instructions. Use `export_figure_data.py` to regenerate plotting data from the supplied results and `reproduce.py` to rerun experiments in a clean working directory. To meet the submission size limit, we omit large data files and intermediate artifacts. We will make these materials publicly available through the project's GitHub repository upon acceptance.

REFERENCES

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 1727–1740. Association for Computing Machinery, 2023a. doi: 10.1145/3564246.3585182. URL https://dl.acm.org/doi/10.1145/3564246.3585182.

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. When does optimizing a proper loss yield calibration? In *Advances in Neural Information Processing Systems*, volume 36, pp. 72071–72095. Curran Associates, Inc., 2023b. doi: 10.52202/075280-3154. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e4165c96702bac5f4962b70f3cf2f136-Abstract-Conference.html.

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, Adam Tauman Kalai, and Preetum Nakkiran. Loss minimization yields multicalibration for large neural networks. In *15th Innovations in Theoretical Computer Science Conference*, volume 287 of *Leibniz International Proceedings in Informatics*, pp. 17:1–17:21. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024. doi: 10.4230/LIPIcs.ITCS.2024.17. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2024.17.

Masahiro Fujisawa and Futoshi Futami. L_2-regularized empirical risk minimization guarantees small smooth calibration error. arXiv preprint arXiv:2510.13450, 2025. URL https://arxiv.org/abs/2510.13450.

Futoshi Futami and Atsushi Nitanda. Smooth calibration error: Uniform convergence and functional gradient analysis. In *International Conference on Learning Representations*, volume 2026, pp. 117–167, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/00a5213b3479c9733eb3743459badc6b-Abstract-Conference.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Parikshit Gopalan, Lunjia Hu, and Guy N. Rothblum. On computationally efficient multi-class calibration. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1983–2026. PMLR, 2024. URL https://proceedings.mlr.press/v247/gopalan24a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/cifar.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3459–3467. PMLR, 2019. URL https://proceedings.mlr.press/v89/vaicenavicius19a.html.

Jingfeng Wu, Peter Bartlett, Matus Telgarsky, and Bin Yu. Benefits of early stopping in gradient descent for overparameterized logistic regression. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67081–67110. PMLR, 2025. URL https://proceedings.mlr.press/v267/wu25b.html.

APPENDIX GUIDE

The appendices contain the proofs and experimental details supporting the three main results. Secondary trajectory guarantees and independent full-metric audits appear in the separate supplement.

Appendix	Purpose
A	Statistical upper bounds and the calibrated lower construction.
B–C	Proofs of the representation criterion and finite-training certificate.
D	The computable anchor operator.
E	Generalization of fixed-operator certificates using an independent audit.
F	Interpolation and certificates with a known input marginal.
G	Four experimental protocols and supporting figures.
H	Notation organized by theme.

A PROOFS OF THE STATISTICAL ESTIMATION BOUNDS

We prove the upper bound in Theorem 1 by controlling scalar witnesses in Proposition A.1, combining their coordinate contributions in Lemma A.2, and applying concentration. Proposition A.3 extends this argument to a predetermined class of predictors, and Corollary A.4 specializes the result to fixed features. We then prove the lower bound with the exactly calibrated construction in Proposition A.5. This order separates the cost of selecting a witness from the additional cost of selecting a predictor.

Throughout Appendix A, $K \geq 2$ is an integer, $n \geq 1$ is the sample size, and $m \geq 0$ is the projection budget. We use the scalar witness class $\mathcal{H}_m$ from Equation (3) and its coordinatewise product $\mathcal{W}_m = \mathcal{H}_m^K$. We assume the relevant function classes are measurable and separable; otherwise, the probability statements in this appendix are interpreted in outer probability in the standard empirical-process sense.

A.1 WEIGHTED WITNESS COMPLEXITY

For a sample $S = ((x_i, y_i))_{i=1}^n$ and a class $\mathcal{F}$ of real-valued functions, we define the conditional empirical Rademacher complexity by

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) = \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i, y_i). \tag{A.1}$$

Here $\sigma_1, \ldots, \sigma_n$ are independent random signs, each uniformly distributed on $\{-1, 1\}$ and independent of S. The expectation conditions on the realized sample. All function classes are measurable and separable; otherwise, the probability statements are interpreted in the outer-probability sense. For a symmetric class containing zero, placing an absolute value inside the supremum gives the same quantity.

We use the nonnegative scalars

$$\mu = \min\{m, K\}, \qquad r = \min\{\mu, n\}, \qquad \mathcal{C}_\mu(n) = 1 + 12\sqrt{r}\log(en/r),$$

where the second term in $\mathcal{C}_\mu(n)$ is defined to be zero when $\mu = 0$. The constant part of a witness contributes the first term; the remaining term controls its variation with the prediction.

Proposition A.1 (Weighted witness complexity)**.** *For any probability vectors $p_1, \ldots, p_n \in \Delta^{K-1}$ and any signed weight vector $c = (c_1, \ldots, c_n) \in \mathbb{R}^n$,*

$$\mathbb{E}_\sigma \sup_{h \in \mathcal{H}_m} \frac{1}{n} \sum_{i=1}^n \sigma_i c_i h(p_i) \leq \min\left\{ \frac{\|c\|_1}{n}, \frac{\|c\|_2}{n} \mathcal{C}_\mu(n) \right\}. \tag{A.2}$$

When we apply Proposition A.1 with residual coordinates as the weights c_i, its bound retains their observed magnitudes. We prove it by incorporating these weights into the Euclidean distance between witness evaluations, then covering the projection values and the one-dimensional outer functions separately. This separation avoids an additional ambient-dimension factor in (A.2).

Proof. The bound $|h| \le 1$ gives the first term in (A.2); the case $c = 0$ follows immediately. Discard zero-weight observations and let $s \le n$ denote the number that remain. Define positive scalar weights

$$q_i = \frac{c_i^2}{\|c\|_2^2}, \qquad \sum_{i=1}^{s} q_i = 1.$$

Sign invariance reduces the analysis to vectors with coordinates $\sqrt{q_i}h(p_i)$. We restore the multiplier $\|c\|_2/n$ at the end.

Write a scalar witness as $h(p) = \phi(a^\top p)$. Its constant part $\phi(0)$ contributes at most

$$\mathbb{E}_\sigma \left| \sum_{i=1}^{s} \sigma_i \sqrt{q_i} \right| \le 1.$$

This proves the case $\mu = 0$. When $\mu \ge n$, the envelope and $\|c\|_1 \le \sqrt{n}\|c\|_2$ suffice. It remains to consider $0 < \mu < n$.

Define the positive scalar normalization parameters

$$b = \min\{1, \sqrt{\mu}\}, \qquad v = \mu/b^2 \ge 1.$$

Every feasible projection satisfies $|a^\top p_i| \le b$. The normalized function

$$\psi(u) = \frac{\phi(bu) - \phi(0)}{b}, \qquad u \in [-1, 1],$$

belongs to the anchored Lipschitz class

$$\Psi = \{\psi : [-1, 1] \to \mathbb{R} : \ \psi(0) = 0, \ \mathrm{Lip}(\psi) \le 1\}.$$

In particular, $|\psi(u)| \le 1$. Enlarging the choice of ψ independently of the constant part leaves the normalized evaluation class

$$\mathcal{T} = \left\{ \left(\sqrt{q_i}\psi(a^\top p_i/b)\right)_{i=1}^{s} : a \in \mathcal{A}_m, \ \psi \in \Psi \right\} \subseteq B_2^s,$$

$$\mathcal{A}_m = \{a \in \mathbb{R}^K : \ \|a\|_\infty \le 1, \ \|a\|_2^2 \le m\}.$$

Here B_2^d denotes the Euclidean unit ball in $\mathbb{R}^d$. The nonconstant part of the original witness contributes an additional factor b.

We first cover the projection evaluations. Let $M \in \mathbb{R}^{s\times K}$ have row $M_i = \sqrt{q_i}p_i^\top/b$. The projection vectors lie in the ellipsoid $E = \sqrt{\mu}MB_2^K$. If β_j are its semiaxis lengths, then

$$\sum_j \beta_j^2 = \mu\|M\|_F^2 = \frac{\mu}{b^2}\sum_i q_i\|p_i\|_2^2 \le v.$$

For a radius $\epsilon > 0$, retain the semiaxes larger than $2\epsilon/5$. The projection of each ellipsoid point onto the discarded subspace has norm at most $2\epsilon/5$, so the discarded part of a difference has norm at most $4\epsilon/5$. The Pythagorean identity therefore shows that an ϵ-separated set remains $3\epsilon/5$-separated after projection onto the retained subspace. Balls of radius $r_0 = 3\epsilon/10$ around these projected points are disjoint.

Let d be the number of retained axes and E_d the corresponding ellipsoid. The support-function form of Young's inequality, with coefficient $3/4$, gives

$$E_d + r_0 B_2^d \subseteq \mathrm{diag}\left(\sqrt{\frac{7}{4}\beta_j^2 + \frac{7}{3}r_0^2}\right)_{j=1}^{d} B_2^d.$$

Volume comparison therefore bounds the logarithm of the packing size by

$$\frac{1}{2}\sum_{j:\beta_j > 2\epsilon/5} \log\left(\frac{175\beta_j^2}{9\epsilon^2} + \frac{7}{3}\right) \le P\frac{v}{\epsilon^2}, \qquad P = \frac{25}{4}\log\frac{7}{3} < 5.296.$$

To verify the last inequality, the function $t^{-2}\log(175t^2/9 + 7/3)/2$ decreases for $t > 0$ and equals P at $t = 2/5$. Its derivative has the sign of

$$z^{-1}(z - 7/3) - \log z < 0, \qquad z = 175t^2/9 + 7/3 > 1.$$

If there are no retained axes, the packing is a singleton. A maximal packing of the feasible projection evaluations yields an internal cover, so its centers correspond to feasible projection vectors with unweighted coordinates in $[-1, 1]$.

We next cover the anchored outer functions. For $0 < \epsilon \le 1$, define the integer $M_0 = \lceil 1/\epsilon \rceil$ and grid spacing $h_0 = 1/M_0$. On the grid $\{jh_0 : -M_0 \le j \le M_0\}$, round each value to

$$h_0 \left\lfloor \frac{\psi(jh_0)}{h_0} + \frac{1}{2} \right\rfloor$$

using a common tie rule. The code at zero is fixed, and adjacent codes differ by at most one. There are thus at most 3^{2M_0} codes. Their piecewise linear interpolants belong to Ψ. Endpoint interpolation contributes at most $h_0/2$ error, and rounding contributes at most another $h_0/2$. These interpolants give an internal ϵ-cover and hence

$$\log N(\Psi, \|\cdot\|_\infty, \epsilon) \le 2\lceil 1/\epsilon \rceil \log 3 \le \frac{O}{\epsilon}, \qquad O = 4 \log 3.$$

Here $N(\mathcal{C}, d, \epsilon)$ denotes the minimum number of radius-ϵ balls, with centers in $\mathcal{C}$, required to cover the class $\mathcal{C}$ under the metric or pseudometric d.

Combining a projection cover of radius $4\epsilon/5$ with an outer-function cover of radius $\epsilon/5$ yields

$$\log N(\mathcal{T}, \|\cdot\|_2, \epsilon) \le H(\epsilon) = \frac{Av}{\epsilon^2} + \frac{D}{\epsilon}, \qquad A = \frac{25P}{16}, \qquad D = 5O.$$

The two errors add because each interpolant is 1-Lipschitz.

We finish by chaining these covers. For a finite set of vectors V, the Rademacher moment-generating-function bound gives

$$\mathbb{E}_\sigma \max_{z \in V} \langle \sigma, z \rangle \le \max_{z \in V} \|z\|_2 \sqrt{2 \log |V|}.$$

When $v < n$, let

$$x = \log(n/v) > 0, \qquad \rho = 1/4, \qquad J = \left\lceil \frac{x}{2 \log 4} \right\rceil.$$

Choose internal nets at radii $\epsilon_j = \rho^j$, starting the chain at zero. The first increment has norm at most one and log-cardinality at most $H(\rho)$. Subsequent increments have norm at most $(1 + 1/\rho)\epsilon_j$ and log-cardinality at most $H(\epsilon_j) + H(\epsilon_{j-1})$. The terminal contribution is at most $\epsilon_J \sqrt{s} \le \sqrt{v}$.

Separating the square roots of the projection and outer-function entropy terms gives

$$\mathbb{E}_\sigma \sup_{t \in \mathcal{T}} \langle \sigma, t \rangle \le C_0 + \sqrt{v}(C_1 + C_2 x),$$

where the scalar coefficients are

$$c_P = (1 + 1/\rho)\sqrt{2A(1 + \rho^2)}, \qquad c_O = (1 + 1/\rho)\sqrt{2D(1 + \rho)},$$

$$C_0 = \sqrt{2D/\rho} + \frac{c_O \rho}{1 - \sqrt{\rho}} < 31.787, \qquad C_1 = 1 + \frac{\sqrt{2A}}{\rho} < 17.273, \qquad C_2 = \frac{c_P}{2 \log 4} < 7.562.$$

Indeed, $J - 1 \le x/(2 \log 4)$, and the outer-function terms from levels $j \ge 2$ sum to at most $\rho/(1 - \sqrt{\rho})$. Thus $C_0 + C_1 < 50$ and $C_2 < 7.6$. Since $v \ge 1$, we can replace C_0 by $C_0\sqrt{v}$. The unit-ball envelope also gives

$$\mathbb{E}_\sigma \sup_{t \in \mathcal{T}} \langle \sigma, t \rangle \le \sqrt{n} = \sqrt{v} e^{x/2}.$$

For $0 \le x \le 9$,

$$\frac{e^{x/2}}{1 + x} \le \max\{1, e^{9/2}/10\} < 12,$$

whereas for $x \ge 9$, $50 + 7.6x \le 12(1 + x)$. Combining the two bounds gives

$$\mathbb{E}_\sigma \sup_{t \in \mathcal{T}} \langle \sigma, t \rangle \le 12\sqrt{v}(1 + x).$$

The envelope handles $v = n$. This covers every case with $0 < \mu < n$: if $\mu \le 1$, then $v = 1 \le n$; otherwise, $v = \mu < n$. Finally,

$$b\sqrt{v} = \sqrt{\mu}, \qquad x = \log(nb^2/\mu) \le \log(n/\mu).$$

Restoring the constant part and the multiplier $\|c\|_2/n$ proves (A.2). □

A.2 From witness complexity to estimation error

For a vector-valued logit function $g : \mathcal{X} \to \mathbb{R}^K$, write $p_g = \mathrm{softmax}(g)$. A vector witness $w \in \mathcal{W}_m$ defines the real-valued residual–witness function

$$F_{g,w}(x, y) = \langle y - p_g(x), w(p_g(x)) \rangle.$$

For one predictor g and for a nonempty logit class $\mathcal{G}$, respectively, define

$$\mathcal{F}_{m,g} = \{F_{g,w} : w \in \mathcal{W}_m\}, \qquad \mathcal{F}_{m,\mathcal{G}} = \{F_{g,w} : g \in \mathcal{G},\ w \in \mathcal{W}_m\}. \tag{A.3}$$

These classes are symmetric and contain zero. Inserting an absolute value inside the supremum in (A.1) therefore leaves their conditional complexities unchanged.

Lemma A.2 (Conditional complexity for fixed predictions)**.** *For every realized sample and every fixed set of logit predictions on that sample,*

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_{m,g}) \le B_{m,K,n} := \min\left\{2, \mathcal{C}_\mu(n)\sqrt{\frac{2K}{n}}\right\}.$$

In Lemma A.2, "fixed" means that the predictions remain fixed during the conditional sign expectation. Population validity for a predictor selected using the sample additionally requires accounting for that selection, as in Proposition A.3.

Proof. Write $p_i = p_g(x_i)$, and let p_{ij} and y_{ij} denote the jth coordinates of p_i and y_i. Let $r_{ij} = y_{ij} - p_{ij}$ be the scalar residual for observation i and class j, and let $r_{\cdot j} \in \mathbb{R}^n$ collect the residuals for class j. The Cartesian product structure of $\mathcal{W}_m$ and Proposition A.1 imply

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{F}_{m,g}) &= \sum_{j=1}^{K} \mathbb{E}_\sigma \sup_{h \in \mathcal{H}_m} \frac{1}{n} \sum_i \sigma_i r_{ij} h(p_i) \\
&\le \frac{\mathcal{C}_\mu(n)}{n} \sum_{j=1}^{K} \|r_{\cdot j}\|_2 \\
&\le \frac{\mathcal{C}_\mu(n)}{n} \sqrt{K \sum_{i,j} r_{ij}^2} \le \mathcal{C}_\mu(n)\sqrt{\frac{2K}{n}}.
\end{aligned}$$

The same signs may appear in every coordinate: linearity of expectation justifies the sum. Since $\|y_i - p_i\|_2^2 \le 2$, the final inequality holds without independence across coordinate residuals. The envelope $|F_{g,w}| \le 2$ gives the other bound. □

To extend this control to a learned predictor, we need continuity of the witness correlations as the logits change. The softmax Jacobian at a probability vector p is $J(p) = \mathrm{diag}(p) - pp^\top$. For $u \in \mathbb{R}^K$,

$$\|J(p)u\|_1 = \mathbb{E}_p|u_j - \mathbb{E}_p u_j| \le \sqrt{\mathrm{Var}_p(u_j)} \le \frac{\max_j u_j - \min_j u_j}{2} \le \frac{\|u\|_2}{\sqrt{2}}.$$

The subscript p means that the class index j is sampled with probabilities p_j. Comparing the mean with the midpoint of the range proves the variance bound. The Jacobian's absolute row sums are $2p_j(1 - p_j) \le 1/2$, so symmetry also gives the spectral bound $1/2$. Integrating these Jacobian bounds along a line segment yields, for any logits $z, z' \in \mathbb{R}^K$,

$$\begin{aligned}
\|\,\mathrm{softmax}(z) - \mathrm{softmax}(z')\|_1 &\le \frac{\|z - z'\|_2}{\sqrt{2}}, \\
\|\,\mathrm{softmax}(z) - \mathrm{softmax}(z')\|_2 &\le \frac{\|z - z'\|_2}{2}.
\end{aligned} \tag{A.4}$$

Proposition A.3 (Uniform estimation for a deterministic logit class)**.** *Let $\mathcal{G}$ be a nonempty deterministic measurable logit class. For a realized sample S, define its empirical logit pseudometric and the residual-test Lipschitz constant by*

$$d_{\mathcal{G},S}(g, g')^2 = \frac{1}{n}\sum_{i=1}^{n} \|g(x_i) - g'(x_i)\|_2^2, \qquad L_\mu = \frac{1}{\sqrt{2}} + \min\{\sqrt{\mu}, \sqrt{2}\}.$$

Then

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_{m,\mathcal{G}}) \le B_{m,K,n} + \inf_{\epsilon>0}\left\{ L_\mu \epsilon + 2\sqrt{\frac{2\log N(\mathcal{G}, d_{\mathcal{G},S}, \epsilon)}{n}} \right\}. \tag{A.5}$$

For an i.i.d. sample and $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for all $g \in \mathcal{G}$,

$$\left|\mathrm{psCE}_m(g) - \widehat{\mathrm{psCE}}_m(g)\right| \le 2B_{m,K,n} + 2\inf_{\epsilon>0}\left\{ L_\mu \epsilon + 2\sqrt{\frac{2\log N(\mathcal{G}, d_{\mathcal{G},S}, \epsilon)}{n}} \right\} + 6\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{A.6}$$

The right-hand side can be capped at 2.

The covering term in Proposition A.3 accounts for selecting a predictor from the same observations used to evaluate calibration. Equation (A.5) bounds conditional complexity; Equation (A.6) converts that bound into a population statement. All predictors are controlled on one event, so the population bound remains valid after fitting the model or selecting an iterate within the predetermined class.

Proof. Every scalar witness satisfies

$$|h(p) - h(p')| \le \min\{\sqrt{\mu}\|p-p'\|_2, \|p-p'\|_1\}.$$

At an input x, write $p = p_g(x)$ and $p' = p_{g'}(x)$. Adding and subtracting $\langle y - p', w(p)\rangle$, and applying (A.4), gives

$$\begin{aligned}|F_{g,w}(x,y) - F_{g',w}(x,y)| &\le \|p-p'\|_1 + 2\min\{\sqrt{\mu}\|p-p'\|_2, \|p-p'\|_1\} \\ &\le L_\mu \|g(x) - g'(x)\|_2.\end{aligned}$$

Fix a radius $\epsilon > 0$ with finite covering number; radii with infinite covering number give a vacuous bound. Before drawing the signs, choose an internal empirical cover $g_1, \dots, g_N$, conditional on the sample, with $N = N(\mathcal{G}, d_{\mathcal{G},S}, \epsilon)$. For each g, a covering center g_j satisfies

$$\frac{1}{n}\sum_i |F_{g,w}(x_i,y_i) - F_{g_j,w}(x_i,y_i)| \le L_\mu d_{\mathcal{G},S}(g,g_j) \le L_\mu \epsilon.$$

For each center, define the scalar sign-dependent random variable

$$Z_j(\sigma) = \sup_w \frac{1}{n}\sum_i \sigma_i F_{g_j,w}(x_i,y_i).$$

Changing one sign changes Z_j by at most $4/n$. The bounded-difference moment-generating-function inequality yields

$$\mathbb{E}_\sigma \exp\{\lambda(Z_j - \mathbb{E}_\sigma Z_j)\} \le \exp(2\lambda^2/n).$$

The exponential maximum inequality, followed by optimization over λ, gives

$$\mathbb{E}_\sigma \max_{j\le N} Z_j \le \max_{j\le N} \mathbb{E}_\sigma Z_j + 2\sqrt{2\log N/n}.$$

Lemma A.2 bounds every center by $B_{m,K,n}$. This proves (A.5), including for sample-dependent covers and dependent coordinate predictions.

For the population step, let P denote population expectation, let P_n denote empirical averaging, and define the scalar deviation

$$D(S) = \sup_{F\in\mathcal{F}_{m,\mathcal{G}}} |P(F) - P_n(F)|. \tag{A.7}$$

The symmetry of the classes in (A.3) and symmetrization with an independent copy of the sample give

$$\mathbb{E}D(S) \le 2\mathbb{E}_S \widehat{\mathfrak{R}}_S(\mathcal{F}_{m,\mathcal{G}}). \tag{A.8}$$

Both $D(S)$ and $\widehat{\mathfrak{R}}_S(\mathcal{F}_{m,\mathcal{G}})$ change by at most $4/n$ when one observation changes, since F takes values in $[-2,2]$. Applying bounded differences to $D(S)$ and to the lower tail of the empirical complexity, each with failure probability $\delta/2$, and using (A.8) yields

$$D(S) \le 2\widehat{\mathfrak{R}}_S(\mathcal{F}_{m,\mathcal{G}}) + 6\sqrt{2\log(2/\delta)/n}.$$

Finally,

$$\sup_{g\in\mathcal{G}}\left|\mathrm{psCE}_m(g)-\widehat{\mathrm{psCE}}_m(g)\right|\le D(S).$$

Combining this inequality with (A.5) proves (A.6). □

The upper bound in Theorem 1 follows by taking $\mathcal{G}$ to be a singleton. Its covering number is one at every radius, so the infimum in (A.6) is zero. The same conclusion applies conditionally to a predictor fitted on an independent training sample.

A.3 The explicit deviation bound for finite features

Corollary A.4 makes the covering contribution in Proposition A.3 explicit for a fixed feature map. The centering matrix $Q = I_K - \mathbf{1}\mathbf{1}^\top/K$ removes a common shift of all logits, which leaves softmax probabilities unchanged. We can therefore cover centered logit representatives, using only $(K-1)d$ parameter directions.

Corollary A.4 (Deterministic finite-feature deviation)**.** *Let $d \ge 1$ be an integer. Fix, independently of the sample, a feature map $\varphi : \mathcal{X} \to \mathbb{R}^d$ satisfying $\sup_x \|\varphi(x)\|_2 \le \sqrt{\Lambda}$, where $\Lambda \ge 0$ is deterministic. For a deterministic radius $R \ge 0$, define*

$$\mathcal{G}_R = \{x \mapsto W\varphi(x) : W \in \mathbb{R}^{K\times d},\ \|W\|_F \le R\}.$$

For an i.i.d. sample of size n and $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for every $g \in \mathcal{G}_R$,

$$\left|\mathrm{psCE}_m(g)-\widehat{\mathrm{psCE}}_m(g)\right| \le \Delta_n(R,\delta),$$

where the deviation bound defined in (11) is

$$\Delta_n(R,\delta) = 2\min\left\{2, \mathcal{C}_\mu(n)\sqrt{\frac{2K}{n}}\right\} + \frac{2}{\sqrt{n}} + 4\sqrt{\frac{2(K-1)d}{n}\log\bigl(1+2L_\mu R\sqrt{\Lambda n}\bigr)} + 2\sqrt{\frac{2\log(1/\delta)}{n}}. \tag{A.9}$$

The right-hand side can be capped at 2.

In order, the four terms in (A.9) account for selecting a witness at fixed predictions, approximating logits by a finite cover, selecting a predictor within that cover, and obtaining confidence $1-\delta$ by concentration. The cover has dimension $(K-1)d$ because common-logit shifts leave the predictions unchanged.

Proof. The centered representative QW satisfies $\|QW\|_F \le \|W\|_F$ and has the same softmax probabilities as W. Define the centered logit class

$$\mathcal{G}_R^0 = \{x \mapsto W\varphi(x) : QW = W,\ \|W\|_F \le R\}.$$

Although $\mathcal{G}_R^0$ need not contain every function in $\mathcal{G}_R$, their residual-test classes coincide: $\mathcal{F}_{m,\mathcal{G}_R^0} = \mathcal{F}_{m,\mathcal{G}_R}$. The centered parameter ball has dimension $(K-1)d$, and its empirical logit distance is at most $\sqrt{\Lambda}\|W-W'\|_F$. A Euclidean ball cover therefore gives

$$\log N(\mathcal{G}_R^0, d_{\mathcal{G}_R^0,S}, \epsilon) \le (K-1)d\log(1+2R\sqrt{\Lambda}/\epsilon). \tag{A.10}$$

Apply (A.5) to $\mathcal{G}_R^0$ and take $\epsilon = (L_\mu\sqrt{n})^{-1}$ in (A.10). This gives the same deterministic bound on the empirical complexity of $\mathcal{F}_{m,\mathcal{G}_R}$ and its expectation. Equation (A.8) and bounded differences applied to $D(S)$ from (A.7), now with $\mathcal{G} = \mathcal{G}_R$, yield

$$D(S) \le 2\mathbb{E}_S\widehat{\mathfrak{R}}_S(\mathcal{F}_{m,\mathcal{G}_R}) + 2\sqrt{2\log(1/\delta)/n}$$

with probability at least $1-\delta$. Substitution proves (A.9). □

The deterministic cover radius can also be optimized. Specifically, the sum of the two terms $2/\sqrt{n}$ and the covering term in (A.9) can be replaced by

$$\inf_{\eta>0}\left\{2L_\mu\eta+4\sqrt{\frac{2(K-1)d}{n}\log\left(1+\frac{2R\sqrt{\Lambda}}{\eta}\right)}\right\}.$$

Every radius bounds the same expected complexity before concentration, so neither this infimum nor minimization over a finite predetermined grid requires a union bound. Selection of R or of the features using these observations requires a predetermined larger class, a selection-complexity bound, or an independent sample. Within a fixed class, the uniform event permits data-dependent predictors and stopping iterates.

A.4 An exactly calibrated lower-bound construction

Proposition A.5 supplies the lower bound in Theorem 1. Its construction also retains the constant-witness contribution when the projection budget is zero. We write Π for the random prediction vector, keeping it distinct from the population averaging operator P.

Proposition A.5 (Calibrated empirical fluctuations)**.** *Let $K \geq 3$ be odd, $m \geq 0$, and $n \geq 2K^2$. There is a distribution of (Π, Y) with Π in the interior of Δ^{K-1} and $\mathbb{E}[Y \mid \Pi] = \Pi$ such that*

$$\mathrm{psCE}_m(\Pi) = 0, \qquad \widehat{\mathbb{E}\mathrm{psCE}}_m(\Pi) \geq \frac{1}{12}\sqrt{\frac{K}{n}}\max\left\{1, \sqrt{\min\{m, K-1\}}\right\}.$$

Here the metrics in (4) are evaluated directly from the joint law of the random prediction Π and the label Y. The expectation is over an i.i.d. sample $((\Pi_i, Y_i))_{i=1}^n$ of size n.

To prove Proposition A.5, we pair prediction groups and assign opposite projection coefficients within each pair. Each output coordinate can then select a feasible witness that follows its own residual fluctuations. These separately selected coordinate witnesses accumulate in the empirical supremum even though every population witness correlation is zero.

Proof. Let $u = (1/K, \ldots, 1/K) \in \mathbb{R}^K$ be the uniform probability vector. Draw an index L uniformly from $[K]$ and define

$$p_\ell = (e_\ell + u)/2, \qquad \Pi = p_L.$$

Conditional on $L = \ell$, draw the one-hot label vector Y with class probabilities p_ℓ. The vectors p_ℓ are distinct and strictly positive. Hence $\mathbb{E}[Y \mid \Pi] = \Pi$, and every population witness correlation vanishes. Taking $X = L$ and defining $g(\ell)$ by the coordinatewise logarithms of p_ℓ realizes this construction as the softmax predictor required by Theorem 1.

First suppose $0 < m \leq K - 1$. Define the integer number of pairs and the scalar direction amplitude by

$$q = (K-1)/2, \qquad b = \sqrt{m/(K-1)} \leq 1.$$

For a realized sample, define the scalar residual sums

$$S_{j\ell} = \sum_{i:L_i=\ell}(Y_{ij} - p_{\ell,j}).$$

For each output coordinate j, partition $[K] \setminus \{j\}$ into q pairs (ℓ_r, ℓ'_r). Choose a vector $a^{(j)} \in \mathbb{R}^K$ with $a^{(j)}_j = 0$ and

$$a^{(j)}_{\ell_r} = b\,\mathrm{sgn}(S_{j\ell_r} - S_{j\ell'_r}), \qquad a^{(j)}_{\ell'_r} = -b\,\mathrm{sgn}(S_{j\ell_r} - S_{j\ell'_r}),$$

using either sign at zero. This vector has coordinate sum zero, infinity norm at most one, and squared Euclidean norm m. Thus

$$h_j(p) = \langle a^{(j)}, p\rangle$$

is in $\mathcal{H}_m$, using the identity outer function on $[-1,1]$. Moreover, $h_j(p_\ell) = a_\ell^{(j)}/2$. Selecting a separate witness for each coordinate gives

$$\widehat{\mathrm{psCE}}_m(\Pi) \geq \frac{b}{2n} \sum_{j=1}^{K} \sum_{r=1}^{q} |S_{j\ell_r} - S_{j\ell'_r}|. \tag{A.11}$$

We next bound each paired fluctuation by its second and fourth moments. Fix j and distinct $\ell, \ell' \neq j$, and write

$$S_{j\ell} - S_{j\ell'} = \sum_{i=1}^{n} Z_i, \qquad Z_i = (\mathbf{1}\{L_i = \ell\} - \mathbf{1}\{L_i = \ell'\}) \left(Y_{ij} - \frac{1}{2K}\right).$$

The independent scalar variables Z_i have mean zero, satisfy $|Z_i| \leq 1$, and have variance

$$v = \mathbb{E}Z_i^2 = \frac{2}{K}\frac{1}{2K}\left(1 - \frac{1}{2K}\right) = \frac{1 - 1/(2K)}{K^2}.$$

In particular, $nv \geq 1$. For the scalar sum $U = \sum_i Z_i$, independence and $\mathbb{E}Z_i^4 \leq v$ imply

$$\mathbb{E}U^4 = n\mathbb{E}Z_1^4 + 3n(n-1)v^2 \leq nv + 3(nv)^2 \leq 4(nv)^2.$$

Hölder's inequality, written as $\mathbb{E}U^2 \leq (\mathbb{E}|U|)^{2/3}(\mathbb{E}U^4)^{1/3}$, then gives

$$\mathbb{E}|U| \geq \frac{(\mathbb{E}U^2)^{3/2}}{(\mathbb{E}U^4)^{1/2}} \geq \frac{1}{2}\sqrt{nv}.$$

Taking expectations in (A.11), we obtain

$$\begin{aligned}
\mathbb{E}\widehat{\mathrm{psCE}}_m(\Pi) &\geq \frac{1}{2n}\sqrt{\frac{m}{K-1}}\, K\, \frac{K-1}{2}\, \frac{\sqrt{nv}}{2} \\
&= \frac{1}{8}\sqrt{\frac{m(K-1)(1-1/(2K))}{n}} \\
&\geq \frac{1}{12}\sqrt{\frac{mK}{n}}.
\end{aligned}$$

The last inequality uses $((K-1)/K)(1-1/(2K)) \geq 5/9$ for $K \geq 3$ and $\sqrt{5}/24 > 1/12$.

It remains to retain the fluctuation already available to constant witnesses. Set $R_{ij} = Y_{ij} - \Pi_{ij}$, where Π_{ij} is coordinate j of observation i's prediction. Since every $\mathcal{H}_m$ contains all constant functions with values in $[-1,1]$,

$$\widehat{\mathrm{psCE}}_m(\Pi) \geq \frac{1}{n}\sum_j \left|\sum_i R_{ij}\right|.$$

On the same distribution, $\mathbb{E}R_{ij} = 0$, $|R_{ij}| \leq 1$, and

$$v_0 = \mathbb{E}R_{ij}^2 = \mathbb{E}[\Pi_{ij}(1-\Pi_{ij})] = \frac{3(K-1)}{4K^2}.$$

Because $nv_0 \geq 1$, the same fourth-moment argument yields

$$\mathbb{E}\widehat{\mathrm{psCE}}_m(\Pi) \geq \frac{K}{2n}\sqrt{nv_0} = \frac{\sqrt{3(K-1)}}{4\sqrt{n}} \geq \sqrt{\frac{K}{8n}}.$$

For $m > 0$, monotonicity of $\mathcal{H}_m$ in m combines the projection bound at $\min\{m, K-1\}$ with this constant-witness bound. At $m = 0$, the constant-witness bound alone proves the claim. □

B Proof of the representation criterion

We prove both the stationarity criterion and its equal-prediction group characterization in Theorem 2. The input support consists of distinct points $x_1, \ldots, x_M$. Their masses are the positive scalars

$\pi_s = P(X = x_s)$, with $\sum_{s=1}^{M} \pi_s = 1$. Throughout this proof, the feature map $\varphi : \mathcal{X} \to \mathbb{R}^d$ and parameter matrix $W \in \mathbb{R}^{K \times d}$ remain fixed. We write $p_s = p_W(x_s) \in \Delta^{K-1}$ for the probability vector at support point x_s, with p_{sj} denoting its jth coordinate.

All conditional label laws considered below share these input masses; only the conditional distribution of Y given X varies. The relevant function space consists of maps from this finite support into $\mathbb{R}^K$. We equip it with the probability-weighted inner product and its associated norm,

$$\langle u, v \rangle_\pi = \sum_{s=1}^{M} \pi_s u(x_s)^\top v(x_s), \qquad \|u\|_\pi^2 = \langle u, u \rangle_\pi.$$

The matrix $Q = I_K - \mathbf{1}\mathbf{1}^\top / K$ projects onto the class-coordinate subspace orthogonal to the all-ones vector $\mathbf{1} \in \mathbb{R}^K$. We define the centered feature span as

$$\mathcal{T} = \{x \mapsto QA\varphi(x) : A \in \mathbb{R}^{K \times d}\}.$$

This subspace contains the centered logit changes available through the fixed features. It is closed because it is finite dimensional, and every element is pointwise orthogonal to $\mathbf{1}$. We denote the orthogonal projection onto this subspace, with respect to $\langle \cdot, \cdot \rangle_\pi$, by $\Pi_{\mathcal{T}}$.

Proof of Theorem 2. Suppose first that every centered witness belongs to $\mathcal{T}$. Fix a conditional label law η and a vector witness $w \in \mathcal{W}_m$. By assumption, there exists a coefficient matrix $A \in \mathbb{R}^{K \times d}$ such that $Qw(p_W) = QA\varphi$ on the support. Both $\eta(x)$ and $p_W(x)$ sum to one, so their difference is orthogonal to $\mathbf{1}$. Consequently, the gradient identity in Equation (7) gives

$$P(\langle \eta - p_W, w(p_W) \rangle) = P(\langle \eta - p_W, QA\varphi \rangle) = -\langle \nabla L(W), A \rangle_F.$$

Here $\langle U, V \rangle_F = \operatorname{tr}(U^\top V)$ is the Frobenius inner product on class-by-feature matrices. A zero population gradient therefore makes every witness correlation vanish.

For the converse, suppose that $f = Qw(p_W)$ lies outside $\mathcal{T}$ for some $w \in \mathcal{W}_m$. Its orthogonal remainder

$$u = f - \Pi_{\mathcal{T}} f$$

is nonzero, is orthogonal to $\mathcal{T}$, and satisfies $\mathbf{1}^\top u(x_s) = 0$ at every support point. We construct a conditional label law by perturbing the fixed predictions in this undetected direction. Since every softmax coordinate is strictly positive and the support is finite, the following scalars are well defined:

$$p_{\min} = \min_{s,j} p_{sj} > 0, \qquad U = \max_{s,j} |u_j(x_s)| > 0, \qquad \tau = \frac{p_{\min}}{2U}.$$

Define $\eta(x_s) = p_s + \tau u(x_s)$. Its coordinates sum to one and are at least $p_{\min}/2$, so it is an interior conditional probability vector. Orthogonality gives, for every $A \in \mathbb{R}^{K \times d}$,

$$0 = P(\langle u, QA\varphi \rangle) = \langle P(u\varphi^\top), A \rangle_F.$$

Thus $P(u\varphi^\top) = 0$ and $\nabla L(W) = -\tau P(u\varphi^\top) = 0$. The same law has a strictly positive witness correlation:

$$P(\langle \eta - p_W, w(p_W) \rangle) = \tau \langle u, f \rangle_\pi = \tau \|u\|_\pi^2 > 0.$$

Thus the centered-witness span condition is necessary for the implication in Theorem 2 to hold uniformly over conditional label laws. Convexity of the population cross-entropy risk in W also makes this stationary point a global minimizer in the fixed feature model. The construction establishes the existence of a miscalibrated label law when the span condition fails; it does not assert failure for every label law.

We next identify this span condition when $m > 0$. Let $p^{(1)}, \ldots, p^{(J)}$ be the distinct prediction vectors on the support. For each $j \in \{1, \ldots, J\}$, the corresponding equal-prediction group is the set of support points with prediction $p^{(j)}$. Every scalar witness is constant on each such set. We show that scalar witness evaluations span all functions with this property.

If $J = 1$, constant witnesses already span the required space. Otherwise, choose a direction $a \in \mathbb{R}^K$ outside the finite union of hyperplanes

$$\{a : a^\top (p^{(s)} - p^{(t)}) = 0\}, \qquad s \neq t.$$

A positive rescaling makes $\|a\|_\infty \le 1$ and $\|a\|_2^2 \le m$ while preserving the separation of the projected values $t_s = a^\top p^{(s)}$. These scalars are distinct and belong to $[-1, 1]$. Given arbitrary target values $b_1, \dots, b_J \in \mathbb{R}$, choose a scalar $\rho > 0$ small enough that

$$|\rho b_s| \le 1, \qquad \rho|b_s - b_t| \le |t_s - t_t| \quad \text{for every } s, t.$$

Such a choice exists because there are finitely many targets and every separation between distinct projected values is positive. Linear interpolation of the sorted pairs $(t_s, \rho b_s)$, continued constantly to the endpoints -1 and 1, defines a bounded 1-Lipschitz outer function. The resulting scalar witness has values ρb_s. Rescaling in the linear span recovers any prescribed function that is constant within each equal-prediction group. In particular, if every witness correlation vanishes, these indicator tests force each class's conditional residual mean to vanish within every equal-prediction group. Thus zero projected smooth calibration error is equivalent to exact calibration on this finite support when $m > 0$.

Let $\Phi \in \mathbb{R}^{M \times d}$ be the feature evaluation matrix, with row $\Phi_{s\cdot} = \varphi(x_s)^\top$. Its column space is the set of scalar feature evaluations on the support. If this space contains all equal-prediction group indicators, it contains every coordinate of every centered witness, giving the required inclusion in $\mathcal{T}$.

Conversely, assume the required inclusion in $\mathcal{T}$ and fix a nonzero vector $b \in \mathbb{R}^K$ with $\mathbf{1}^\top b = 0$. The interpolation argument and linearity of $\mathcal{T}$ imply that $bf \in \mathcal{T}$ for every scalar function f that is constant within each equal-prediction group: scalar witnesses can be placed in individual vector coordinates and their centered evaluations can then be combined linearly. Hence $bf = QA\varphi$ for some A. Multiplication by $b^\top / \|b\|_2^2$ represents f in the scalar feature span. This applies to each group indicator.

The J equal-prediction group indicators are linearly independent, so their inclusion in $\mathrm{col}(\Phi)$ requires $\mathrm{rank}(\Phi) \ge J$. If all M predictions are distinct, the indicators form the standard basis of $\mathbb{R}^M$; their inclusion is equivalent to $\mathrm{rank}(\Phi) = M$. At $m = 0$, all scalar witnesses are constant, and representing the constant function suffices. □

Theorem 2 therefore gives an exact representation criterion. However, the criterion does not control the required coefficient norms uniformly over feature models. The approximation profile in Equation (8) measures this quantitative cost under a specified coefficient budget.

B.1 Stationarity without calibration

To illustrate how stationarity fails to ensure calibration when the representation is insufficiently expressive, consider a classification task with $K = 3$ classes. Let X be drawn uniformly from $\{-1, 0, 1\}$, and define the feature map $\varphi(x) = (1, x)^\top$. We construct the target model and true conditional probabilities as follows:

$$g_*(x) = (\log(2/3) + x \log 3, 0, 0)^\top, \qquad p_* = \mathrm{softmax}(g_*),$$
$$\eta(x) = p_*(x) + 0.1 b(x)(1, -1, 0)^\top, \qquad (b(-1), b(0), b(1)) = (1, -2, 1).$$

At the ordered support points, the first prediction coordinates are $(0.1, 0.25, 0.5)$. Let W_* be the weight matrix that produces g_*. The label probabilities are strictly interior. Because $\mathbb{E}[b(X)] = \mathbb{E}[Xb(X)] = 0$, the conditional-mean residuals are orthogonal to the features, yielding $\nabla L(W_*) = 0$. By convexity, W_* is a global population minimizer for this linear feature model. Nevertheless, selecting the feasible $m = 1$ witness $w(p) = (p_1, -p_1, 0)^\top$ exposes the miscalibration:

$$\mathrm{psCE}_1(W_*) \ge \mathbb{E}\langle \eta - p_*, w(p_*) \rangle = 1/150 > 0.$$

Figure 1 shows that feature-weighted residuals cancel while the witness correlation remains positive. The two-dimensional feature span cannot contain all three equal-prediction group indicators.

C Proof of the training certificate

To prove Theorem 3, we first establish the deterministic sample bound stated in Equation (10). This separates the representation argument from the two ingredients needed for the population training guarantee: a bound on the final gradient and a uniform sampling event.

For a parameter matrix W, let $r = Y - p_W(X) \in \mathbb{R}^K$ denote the label-minus-prediction residual. Define its sample root-mean-square norm and the empirical gradient matrix by

$$\rho_S(W) = \big(P_n(\|Y - p_W\|_2^2)\big)^{1/2}, \qquad G_S(W) = P_n((p_W - Y)\varphi^\top) = \nabla \widehat{L}(W) \in \mathbb{R}^{K\times d}.$$

The coefficient matrix in Equation (8) may be restricted to $QA = A$. Indeed, replacing A by QA increases neither its Frobenius norm nor its approximation error, since, at every input x,

$$\|Qw(p_W(x)) - A\varphi(x)\|_2^2 = \|Qw(p_W(x)) - QA\varphi(x)\|_2^2 + \|(I_K - Q)A\varphi(x)\|_2^2.$$

Lemma C.1 (Deterministic signed-residual certificate)**.** *For every sample, parameter matrix W, and scalar budget $B \geq 0$,*

$$\widehat{\mathrm{psCE}}_m(W) \leq B\|G_S(W)\|_F + \big(P_n(\|Y - p_W\|_2^2)\big)^{1/2}\varepsilon_{m,S}(W; B).$$

The residual factor is at most $\sqrt{2}$.

Proof. Fix a witness $w \in \mathcal{W}_m$ and a matrix $A \in \mathbb{R}^{K\times d}$ with $\|A\|_F \leq B$. Since $\mathbf{1}^\top r = 0$, Equation (9) decomposes the witness correlation as

$$P_n(\langle r, w(p_W)\rangle) = -\langle G_S(W), A\rangle_F + P_n(\langle r, Qw(p_W) - A\varphi\rangle).$$

The first term pairs the represented component with the parameter gradient. The second contains the part left outside that representation. Frobenius duality and Cauchy–Schwarz bound the absolute correlation by

$$B\|G_S(W)\|_F + \big(P_n(\|r\|_2^2)\big)^{1/2}\big(P_n(\|Qw(p_W) - A\varphi\|_2^2)\big)^{1/2}.$$

Taking the infimum over A, using approximate minimizers if necessary, and then the supremum over w proves the inequality in Lemma C.1. Finally, for a one-hot label $Y = e_j$ and a probability vector p,

$$\|Y - p\|_2^2 = 1 - 2p_j + \|p\|_2^2 \leq 2.$$

□

To pass to the population, fix a scalar radius $R \geq 0$ independently of the sample. Corollary A.4 provides an event $\mathcal{E}_R$ of probability at least $1 - \delta$ on which the population and empirical calibration metrics differ by at most $\Delta_n(R, \delta)$ from Equation (11), uniformly over $\|W\|_F \leq R$. Lemma C.1 therefore gives, simultaneously for all such W and all $B \geq 0$,

$$\mathrm{psCE}_m(W) \leq B\|G_S(W)\|_F + \sqrt{2}\,\varepsilon_{m,S}(W; B) + \Delta_n(R, \delta).$$

The same inequality holds with $\rho_S(W)$ in place of $\sqrt{2}$. No union bound over B is needed because the sample inequality is deterministic for every budget.

Proof of Theorem 3. The softmax Jacobian $\mathrm{diag}(p_W) - p_W p_W^\top$ is positive semidefinite and has operator norm at most $1/2$. Consequently, the empirical cross-entropy Hessian has Frobenius operator norm at most $\Lambda/2$. In a matrix direction $A \in \mathbb{R}^{K\times d}$, its quadratic form is bounded by

$$\frac{1}{2}P_n(\|A\varphi\|_2^2) \leq \frac{\Lambda}{2}\|A\|_F^2.$$

For the update in Equation (12) with $0 < \alpha \leq 2/\Lambda$, smoothness gives

$$\widehat{L}(W_{t+1}) \leq \widehat{L}(W_t) - \alpha\left(1 - \frac{\alpha\Lambda}{4}\right)\|G_t\|_F^2 \leq \widehat{L}(W_t) - \frac{\alpha}{2}\|G_t\|_F^2.$$

The zero initialization has loss $\widehat{L}(W_0) = \log K$. Summing the decrease and using nonnegativity of the loss yields

$$\sum_{t=0}^{T-1}\|G_t\|_F^2 \leq \frac{2\log K}{\alpha}, \qquad \min_{t<T}\|G_t\|_F \leq \sqrt{\frac{2\log K}{\alpha T}}.$$

Convexity upgrades this best-iterate estimate to a final-iterate estimate. Let vec stack a matrix into a vector in $\mathbb{R}^{Kd}$, and let $H_t \in \mathbb{R}^{Kd\times Kd}$ be the empirical Hessian averaged along the segment from W_t to W_{t+1}. The gradient difference identity gives

$$\operatorname{vec} G_{t+1} = (I_{Kd} - \alpha H_t)\operatorname{vec} G_t, \qquad 0 \preceq H_t \preceq (\Lambda/2) I_{Kd}.$$

Because $\alpha\Lambda/2 \le 1$, the matrix $I_{Kd} - \alpha H_t$ has operator norm at most one. Hence the gradient norms are nonincreasing and

$$T\|G_T\|_F^2 \le \sum_{t<T} \|G_t\|_F^2 \le 2\log K/\alpha.$$

The same descent estimate places the whole trajectory in a deterministic ball. For every $t \le T$, Cauchy–Schwarz gives

$$\|W_t\|_F \le \alpha \sum_{s<t} \|G_s\|_F \le \alpha\sqrt{t\sum_{s<t}\|G_s\|_F^2} \le \sqrt{2\alpha T\log K} = R_T.$$

We may therefore apply the single uniform event $\mathcal{E}_{R_T}$ to every iterate. Combining this event with Lemma C.1 proves Equation (13) for every $0 \le t \le T$. Substituting the final-gradient estimate and $\rho_S(W_T) \le \sqrt{2}$ then proves Equation (14). The argument covers stochastic labels and misspecified conditional label laws. □

Because the event in Theorem 3 is simultaneous over the fixed training horizon, the observed-gradient bound in Equation (13) may also guide checkpoint selection within that horizon.

D A COMPUTABLE ANCHOR OPERATOR

The approximation profile in Equation (8) asks for a coefficient matrix for every witness. We construct these matrices through a common linear map, the anchor operator. For each scalar witness, the operator maps its values at finitely many anchors (reference points in the simplex) to a feature coefficient vector. This construction keeps the approximation uniform over the full witness class while reducing its evaluation to finite matrices.

Fix a sample of size n, a parameter matrix W, and an anchor count $M \ge 1$. Let $\Phi \in \mathbb{R}^{n\times d}$ be the sample feature matrix, with row $\Phi_{i\cdot} = \varphi(X_i)^\top$, and let $p_i = p_W(X_i) \in \Delta^{K-1}$ be its prediction vectors. The construction uses probability anchors $q_1, \dots, q_M \in \Delta^{K-1}$, an assignment map $s : \{1, \dots, n\} \to \{1, \dots, M\}$, and a coefficient matrix $C \in \mathbb{R}^{d\times M}$. We form the assignment matrix $Z \in \mathbb{R}^{n\times M}$ with entries $Z_{ij} = \mathbf{1}\{s(i) = j\}$ and define

$$E = Z - \Phi C, \qquad D = E^\top E/n, \qquad R = C^\top C, \qquad V = \frac{1}{n}\sum_i (p_i - q_{s(i)})(p_i - q_{s(i)})^\top.$$

The matrix $E \in \mathbb{R}^{n\times M}$ records the errors when the original features approximate anchor assignments. The positive semidefinite matrices $D, R \in \mathbb{R}^{M\times M}$ measure this error and the coefficient size, respectively. The positive semidefinite matrix $V \in \mathbb{R}^{K\times K}$ records the displacement between predictions and their assigned anchors.

For a real symmetric matrix $H \in \mathbb{R}^{M\times M}$, we use the scalar $\sigma(H)$ and symmetric matrix $\mathcal{L}(H) \in \mathbb{R}^{K\times K}$,

$$\sigma(H) = \sum_j [(H\mathbf{1})_j]_+, \qquad \mathcal{L}(H) = \sum_{j<\ell} [-H_{j\ell}]_+ (q_j - q_\ell)(q_j - q_\ell)^\top.$$

Here $\mathbf{1} \in \mathbb{R}^M$, and $[u]_+ = \max(u, 0)$ is the positive part of a real scalar. Every summand defining $\mathcal{L}(H)$ is positive semidefinite. Thus its largest eigenvalue, denoted by $\lambda_{\max}$, can control quadratic forms in the projection direction.

Proposition D.1 (Uniform approximation from an anchor operator)**.** *Let $\mu = \min\{m, K\}$ and let $\Gamma \subset (0, \infty)$ be any nonempty finite set. Define the nonnegative scalars B_C and e_C by*

$$B_C^2 = K \min\left\{\sum_{j,\ell} |R_{j\ell}|, \sigma(R) + \mu\lambda_{\max}(\mathcal{L}(R))\right\}$$

and

$$e_C^2 = K \min_{\gamma \in \Gamma} \left\{ (1+\gamma)\sigma(D) + \mu\lambda_{\max}\left((1+\gamma)\mathcal{L}(D) + (1+\gamma^{-1})V\right)\right\}.$$

Then $\varepsilon_{m,S}(W; B_C) \le \min\{\sqrt{K}, e_C\}$. *The bound holds for every sample, including a sample used to choose* W, *the anchors* q_j, *the assignments* s, *and the operator* C.

The two scalars in Proposition D.1 control different costs of the same construction. The coefficient budget B_C limits the feature coefficients and therefore their multiplier of the empirical gradient. The approximation upper bound e_C combines anchor-assignment approximation and prediction-to-anchor displacement. Substituting $B = B_C$ and $\varepsilon_{m,S}(W; B_C) \le \min\{\sqrt{K}, e_C\}$ into Equation (13) makes these costs explicit. Thus a smaller approximation error alone need not give a smaller calibration bound.

Proof. Fix a scalar witness $h(p) = \phi(a^\top p) \in \mathcal{H}_m$. Its direction $a \in \mathbb{R}^K$ satisfies $\|a\|_\infty \le 1$ and $\|a\|_2^2 \le m$, and its outer function $\phi : [-1,1] \to [-1,1]$ is 1-Lipschitz. Define the anchor-value vector $z \in \mathbb{R}^M$ by $z_j = h(q_j)$. For a symmetric matrix H, the exact identity

$$z^\top H z = \sum_j (H\mathbf{1})_j z_j^2 - \sum_{j<\ell} H_{j\ell}(z_j - z_\ell)^2$$

and the bounds $|z_j| \le 1$ and $|z_j - z_\ell| \le |a^\top(q_j - q_\ell)|$ give

$$z^\top H z \le \sigma(H) + a^\top \mathcal{L}(H) a.$$

All pairwise differences use the same projection vector a. Keeping this common direction is what allows one spectral bound to control their combined contribution.

Apply this inequality with $H = R$. Since $z^\top R z = \|Cz\|_2^2$, it gives the second expression in B_C^2/K after using $\|a\|_2^2 \le \mu$. The first expression follows directly from $|z_j| \le 1$.

We now control the approximation error. Define $u \in \mathbb{R}^n$ by $u_i = h(p_i) - h(q_{s(i)})$. The Lipschitz property and the quadratic-form bound with $H = D$ imply

$$\frac{1}{n}\|u\|_2^2 \le a^\top V a, \qquad \frac{1}{n}\|Ez\|_2^2 \le \sigma(D) + a^\top \mathcal{L}(D) a.$$

The sample vector of scalar witness values is $u + Zz$, so its difference from the feature approximation $\Phi C z$ is $u + Ez$. For every scalar $\gamma > 0$, Young's inequality therefore gives

$$\frac{1}{n}\|u + Ez\|_2^2 \le (1+\gamma)\sigma(D) + a^\top \left[(1+\gamma)\mathcal{L}(D) + (1+\gamma^{-1})V\right] a.$$

The constraint $\|a\|_2^2 \le \mu$ bounds the last quadratic form by μ times its largest eigenvalue. Minimizing over Γ yields the asserted scalar error bound.

For a vector witness $w \in \mathcal{W}_m$, collect its anchor values in $T_w \in \mathbb{R}^{M \times K}$, whose kth column contains the values of the kth scalar witness. As in Section 6, define

$$A_w = Q(CT_w)^\top \in \mathbb{R}^{K \times d}.$$

The scalar coefficient bounds, summed across the K columns, and contraction by the orthogonal projection Q give $\|A_w\|_F \le B_C$. To see how this also bounds the centered approximation error, let $H_w \in \mathbb{R}^{n \times K}$ have row i equal to $w(p_i)^\top$. Then

$$P_n(\|Qw(p_W) - A_w\varphi\|_2^2) = \frac{1}{n}\|(H_w - \Phi C T_w)Q\|_F^2 \le \frac{1}{n}\|H_w - \Phi C T_w\|_F^2 \le e_C^2.$$

This construction applies to every w, so Equation (8) gives $\varepsilon_{m,S}(W; B_C) \le e_C$. Finally, the zero matrix $A = 0$ always has error at most $\sqrt{K}$ and is feasible within the same coefficient budget. □

Alternatively, applying the triangle inequality before bounding the two error sources gives the valid replacement for e_C,

$$\sqrt{K}\left[\sqrt{\sigma(D) + \mu\lambda_{\max}(\mathcal{L}(D))} + \sqrt{\mu\lambda_{\max}(V)}\right].$$

The shared-direction bound in Proposition D.1 can improve on this separate bound. Both cover the complete witness class. They are deterministic sample approximation bounds; the population training statement additionally uses the uniform deviation term in Theorem 3.

E GENERALIZATION OF A FIXED APPROXIMATION OPERATOR

Proposition D.1 bounds the complete witness approximation error on the observed inputs. A population guarantee requires a further statistical step. An independent audit sample permits this step after the predictor and its approximation operator have been selected. We first transfer the error of that fixed operator in Proposition E.2. We then combine it with an audit estimate of the parameter gradient in Corollary E.3.

Throughout this section, $N \geq 1$ is the audit sample size. The audit observations are independent; they may have different laws, in which case P denotes their average population law. We write P_N for their empirical average and P_X for the corresponding average input marginal. All predictors, feature maps, and witness-to-coefficient maps under consideration are fixed independently of these audit observations. The results may equivalently be applied conditionally on an independent construction sample.

For a feature map $\varphi : \mathcal{X} \to \mathbb{R}^d$, a parameter matrix $W \in \mathbb{R}^{K\times d}$, and a scalar coefficient budget $B \geq 0$, the population approximation profile is

$$\varepsilon_{m,P}(W; B) = \sup_{w\in\mathcal{W}_m} \inf_{\substack{A\in\mathbb{R}^{K\times d}\\ \|A\|_F\leq B}} \left(P_X\left(\|Qw(p_W) - A\varphi\|_2^2\right)\right)^{1/2}. \tag{E.1}$$

The population profile in Equation (E.1) has the same order of optimization as the sample profile in Equation (8): each witness receives its own coefficient matrix within a common budget, and the outer supremum ranges over the full class. A prescribed map $w \mapsto A_w$ therefore provides a simultaneous candidate for every inner optimization.

E.1 ONE-SIDED CONCENTRATION OF A COMPLETE APPROXIMATION ERROR

The transfer argument exploits nonnegativity of the squared approximation errors. For nonnegative scalars u, R and a positive scalar s, define

$$\mathcal{U}(u, R, s) = \min\left\{R,\ u + Rs + \sqrt{(Rs)^2 + 2Rsu}\right\}. \tag{E.2}$$

In Equation (E.2), u is an empirical upper bound, R is a uniform range bound, and s is the logarithm of the inverse failure probability divided by the audit size. The correction grows with the error range as well as with the empirical error. Lemma E.1 explains why this correction suffices for a supremum of nonnegative errors.

Lemma E.1 (Transfer of a nonnegative supremum)**.** *Let $\{f_\theta : \theta \in \Theta\}$ be a fixed measurable class of functions satisfying $0 \leq f_\theta \leq R$, and let u be any upper bound on $\sup_\theta P_N(f_\theta)$. For every scalar $t > 0$, with probability at least $1 - e^{-t}$,*

$$\sup_{\theta\in\Theta} P(f_\theta) \leq \mathcal{U}(u, R, t/N).$$

In Lemma E.1, the empirical bound must cover the complete function class, but the concentration proof needs only a function that maximizes, or nearly maximizes, the population expectation. That function is fixed before the audit sample is drawn. Consequently, the lemma transfers a full-class upper bound without a separate complexity term for selecting a witness on the audit sample. The empirical bound u itself may depend on the audit observations.

Proof. Fix a population maximizer, initially assuming one exists, and denote its population expectation by the scalar $\mu = P(f_\theta)$. The lower-tail Chernoff bound for a nonnegative random variable bounded by R gives

$$\mu \leq P_N(f_\theta) + \sqrt{2R\mu t/N}$$

with probability at least $1 - e^{-t}$. The empirical supremum, and therefore u, bounds $P_N(f_\theta)$. Solving the resulting quadratic inequality in $\sqrt{\mu}$ gives

$$\mu \leq u + Rt/N + \sqrt{(Rt/N)^2 + 2Rtu/N}.$$

The deterministic bound $\mu \leq R$ supplies the minimum in (E.2). The cases $R = 0$ or $\mu = 0$ are immediate.

For completeness, the same lower-tail argument accommodates independent observations with different laws. For a random variable $Z \in [0, R]$ and a scalar $\lambda \geq 0$, convexity and $e^{-x} \leq 1-x+x^2/2$ imply

$$\mathbb{E}e^{-\lambda Z} \leq 1 + \frac{\mathbb{E}Z}{R}(e^{-\lambda R} - 1) \leq \exp\left\{-\lambda \mathbb{E}Z + \lambda^2 R \mathbb{E}Z/2\right\}.$$

Multiplying these bounds across the independent observations and optimizing λ gives

$$\mathbb{P}\{P_N(Z) \leq \mu - a\} \leq \exp\{-Na^2/(2R\mu)\},$$

where μ is the average expectation and $a \geq 0$. This proves the stated lower-tail inequality.

If the population supremum is unattained, choose a deterministic sequence of functions whose population expectations increase to it. Each member's population expectation exceeds $\mathcal{U}(u, R, t/N)$ with probability at most e^{-t}. These exceedance events increase with the sequence of population expectations, and their union is precisely the event that the supremum exceeds $\mathcal{U}(u, R, t/N)$. Continuity of probability for increasing events therefore gives the same failure bound for the supremum. □

Proposition E.2 (Transfer of a fixed approximation operator)**.** *Fix $\delta \in (0, 1)$. Condition on a predictor W, a feature map with $\|\varphi(x)\|_2^2 \leq \Lambda$, and a map $w \mapsto A_w$ satisfying $\|A_w\|_F \leq B$ for every $w \in \mathcal{W}_m$. These objects must be fixed independently of the audit observations. Suppose the audit computation certifies*

$$\sup_{w \in \mathcal{W}_m} P_N\big(\|Qw(p_W) - A_w\varphi\|_2^2\big) \leq u.$$

For any uniform range bound R for these squared errors, fixed independently of the audit observations, the population approximation certificate holds with probability at least $1 - \delta$:

$$\varepsilon_{m,P}(W; B) \leq \min\left\{\sqrt{K},\ \sqrt{\mathcal{U}\big(u, R, \log(1/\delta)/N\big)}\right\}. \tag{E.3}$$

One always may take $R = (\sqrt{K} + B\sqrt{\Lambda})^2$.

Proposition E.2 distinguishes approximation quality from its transfer cost. A smaller empirical error improves the population bound in Equation (E.3), but a larger coefficient budget may enlarge the uniform range and its confidence correction. The zero coefficient matrix supplies the separate cap $\sqrt{K}$ after the fixed operator has been transferred.

Proof. Apply Lemma E.1 to

$$f_w(x) = \|Qw(p_W(x)) - A_w\varphi(x)\|_2^2.$$

The triangle inequality gives the stated range because $\|Qw\|_2 \leq \sqrt{K}$ and $\|A_w\varphi\|_2 \leq B\sqrt{\Lambda}$. For each witness, the population RMS approximation error obtained with its prescribed matrix A_w upper-bounds the infimum in Equation (E.1). The separate choice $A = 0$ gives the additional upper bound $\sqrt{K}$. □

The scalar u must bound the error of the same fixed operator A_w used in the range calculation. Capping its empirical squared error at K by choosing $A = 0$ would change that operator. Accordingly, transfer precedes comparison with the zero-operator alternative. Optimizing the spectral and Young bounds from Proposition D.1 on the audit sample is compatible with this requirement when it only sharpens the error bound for an unchanged map $w \mapsto A_w$.

E.2 Combining approximation transfer with the audit gradient

To obtain a calibration certificate, define the residual vector $r = Y - p_W \in \mathbb{R}^K$, consistently with Equation (4). The population and audit gradient matrices are $G = -P(r\varphi^\top)$ and $\widehat{G} = -P_N(r\varphi^\top)$, respectively, both in $\mathbb{R}^{K \times d}$. The minus sign follows from the cross-entropy gradient in Equation (7). In addition, the audit supplies a residual-weighted feature second-moment matrix and two nonnegative scalar summaries:

$$\widehat{V} = P_N\big(\|r\|_2^2 \varphi\varphi^\top\big) \in \mathbb{R}^{d \times d}, \qquad \widehat{v} = \lambda_{\max}(\widehat{V}), \qquad \widehat{q} = P_N\big(\|r\|_2^2\big).$$

The largest eigenvalue $\widehat{v}$ measures the directional variation relevant to gradient concentration. The scalar $\widehat{q}$ records the empirical residual second moment used to control the approximation contribution.

Consider a positive integer J of predictor–operator pairs, all fixed independently of the audit sample, and a total failure probability $\delta \in (0,1)$. For each pair, use its certified u, R, B and define

$$s = \frac{\log(4J/\delta)}{N}, \qquad e_+ = \min\{\sqrt{K}, \sqrt{\mathcal{U}(u,R,s)}\},$$

$$v_+ = \mathcal{U}(\widehat{v}, 2\Lambda, s), \qquad q_+ = \mathcal{U}(\widehat{q}, 2, s), \qquad \gamma_+ = \sqrt{2v_+ s} + \frac{4\sqrt{2\Lambda}}{3}s.$$

On the concentration events used in Corollary E.3, the scalars e_+, v_+, and q_+ upper-bound population approximation error, the largest eigenvalue of the population residual-weighted feature second moment, and the residual second moment, respectively. The scalar γ_+ is the resulting gradient sampling allowance.

Corollary E.3 (Independent-audit calibration certificate)**.** *With probability at least $1-\delta$, all J pairs simultaneously satisfy*

$$\mathrm{psCE}_m(W) \le \min\left\{2,\ B(\|\widehat{G}\|_F + \gamma_+) + \sqrt{q_+}\, e_+\right\}. \tag{E.4}$$

Corollary E.3 evaluates the same optimization–representation decomposition as Theorem 3, using an independent sample to control its population quantities. In particular, the approximation contribution in Equation (E.4) remains visible even when the training gradient is small. Simultaneous coverage permits comparison among the prespecified predictor–operator pairs on the common audit event.

Proof. Proposition E.2 gives the approximation event. To bound

$$v = \lambda_{\max}\left(P\big(\|r\|_2^2 \varphi\varphi^\top\big)\right),$$

fix a unit eigenvector maximizing the population quadratic form. Its scalar integrand lies in $[0, 2\Lambda]$, and its empirical mean is at most $\widehat{v}$. Lemma E.1 therefore gives $v \le v_+$. The same lemma gives $P(\|r\|_2^2) \le q_+$.

For the gradient, suppose $G \neq 0$ and fix the population direction matrix $H = G/\|G\|_F \in \mathbb{R}^{K\times d}$. Cauchy–Schwarz gives

$$P\big(\langle H, r\varphi^\top\rangle_F^2\big) \le P\big(\|r\|_2^2 \|H\varphi\|_2^2\big) \le v.$$

Apply scalar Bernstein concentration to the variables $-\langle H, r\varphi^\top\rangle_F$, whose average expectation is $\|G\|_F$. Each centered scalar summand has absolute value at most $2\sqrt{2\Lambda}$, so

$$\|G\|_F \le \|\widehat{G}\|_F + \sqrt{2vs} + \frac{4\sqrt{2\Lambda}}{3}s$$

outside an event of probability at most e^{-Ns}. On the second-moment event, replace v by v_+. When $G = 0$, the gradient bound is immediate. Averaging the second moments gives the same argument for independent observations with different laws.

The population residual pairing and Cauchy–Schwarz imply

$$\mathrm{psCE}_m(W) \le B\|G\|_F + \big(P(\|r\|_2^2)\big)^{1/2} \varepsilon_{m,P}(W; B).$$

Combining the four events proves (E.4) for a given pair. A union bound over the four events and J pairs gives total failure probability at most δ; independence between these events or between the candidate pairs is unnecessary. Finally, bounded witnesses and $\|Y - p_W\|_1 \le 2$ give the deterministic cap two. □

Writing G_S for the training gradient at the same fixed predictor W, the triangle inequality gives

$$\|\widehat{G}\|_F \le \|G_S\|_F + \|\widehat{G} - G_S\|_F.$$

This separates the observed training gradient, its change on the audit sample, the gradient sampling allowance, and the population approximation error. We evaluate this certificate for frozen CIFAR-10 features in Appendix G.4, where we also specify the retrospective scope of the calculation.

F KNOWN-MARGINAL CERTIFICATES FROM INTERPOLATION FEATURES

The independent-audit approach in Corollary E.3 applies when population expectations must be estimated from fresh observations. When the input marginal is known, we can instead evaluate the population input expectation directly, up to controlled numerical integration error. The remaining sampling uncertainty is a finite-dimensional label moment that does not depend on the parameter matrix. This yields the specialization in Proposition F.3: one concentration event covers every parameter and every training iterate, while explicit interpolation features provide the witness approximation.

F.1 CONTINUITY AND THE INTERPOLATION CONSTRUCTION

The interpolation argument begins in logit space. Set $\mu = \min\{m, K\}$ and define the nonnegative scalar

$$c_\mu = \min\{\sqrt{\mu}/2, 1/\sqrt{2}\}.$$

Lemma F.1 uses this constant to control the change of any scalar projected witness under a change of logits.

Lemma F.1 (Scalar witness continuity in logits)**.** *Every $h \in \mathcal{H}_m$ satisfies*

$$|h(\mathrm{softmax}(z)) - h(\mathrm{softmax}(z'))| \le c_\mu \|z - z'\|_2 \qquad \textit{for all } z, z' \in \mathbb{R}^K.$$

The bound in Lemma F.1 has two complementary sources. The Euclidean size of the projection direction gives the factor $\sqrt{\mu}/2$. The bounded coordinates of that direction give a second, dimension-independent factor $1/\sqrt{2}$. Taking the smaller factor preserves the benefit of a small projection budget without allowing the continuity constant to grow indefinitely with m.

Proof. Write $h(p) = \phi(a^\top p)$, where $a \in \mathbb{R}^K$ is an admissible direction and the scalar outer function ϕ is 1-Lipschitz. The Euclidean softmax Jacobian bound gives the constant $\|a\|_2/2 \le \sqrt{\mu}/2$.

For the second bound, let $J_p = \mathrm{diag}(p) - pp^\top$ be the softmax Jacobian at a probability vector p. Its $2 \to 1$ norm equals $\sup_{a \in [-1,1]^K} \|J_p a\|_2$. Convexity permits taking this maximum at a cube vertex $a \in \{-1, 1\}^K$. Let the scalar s be the total probability on the coordinates for which $a_j = 1$. The entries of $J_p a$ are $2(1-s)p_j$ on this set and $-2sp_j$ on its complement. Consequently,

$$\|J_p a\|_2^2 \le 4(1-s)^2 s^2 + 4s^2(1-s)^2 = 8s^2(1-s)^2 \le \frac{1}{2}.$$

Integrating the Jacobian along the segment from z to z' and using outer-function Lipschitzness gives the second bound. Taking the smaller constant proves Lemma F.1, including at $m = 0$. □

Lemma F.1 first gives an interpolation certificate on a general fixed finite triangulation. Here the input law is supported on the triangulated subset of a finite-dimensional Euclidean space. Let its vertices be $v_1, \ldots, v_M$ in the input domain, and write the hat features as $\lambda_j(x) = \varphi_j(x)$ for $j = 1, \ldots, M$. These piecewise-linear coordinates satisfy $\lambda_j(x) \ge 0$ and $\sum_j \lambda_j(x) = 1$, with nonzero weights only at vertices of a simplex containing x. Thus $\|\varphi(x)\|_2^2 \le 1$. For $W \in \mathbb{R}^{K \times M}$, denote the vertex logit vectors by $W_j = W_{\cdot j}$ and define

$$\bar{W}(x) = \sum_{j=1}^{M} \lambda_j(x) W_j = g_W(x), \qquad \xi_W(x) = \sum_{j=1}^{M} \lambda_j(x) \|W_j - \bar{W}(x)\|_2.$$

The vector $\bar{W}(x) \in \mathbb{R}^K$ is the interpolated logit, and the nonnegative scalar $\xi_W(x)$ measures its average distance from the active vertex logits. Let $\mathrm{osc}(W)$ be the largest Euclidean difference between logits at vertices of a common simplex.

Proposition F.2 (Calibration from vertex interpolation)**.** *For every sample and parameter matrix W,*

$$\widehat{\mathrm{psCE}}_m(W) \le \sqrt{KM}\|G_S(W)\|_F + c_\mu P_n\big(\|Y - p_W\|_1 \xi_W\big) \le \sqrt{KM}\|G_S(W)\|_F + 2c_\mu \,\mathrm{osc}(W).$$

On the uniform event in Theorem 3, with $d = M$ and $\Lambda = 1$, every iterate $t \le T$ therefore satisfies

$$\mathrm{psCE}_m(W_t) \le \sqrt{KM}\|G_t\|_F + 2c_\mu \,\mathrm{osc}(W_t) + \Delta_n(R_T, \delta). \tag{F.1}$$

Proposition F.2 provides a coefficient budget directly: every witness is bounded at every vertex, so its vertex-value matrix has Frobenius norm at most $\sqrt{KM}$. Approximation error depends instead on the variation between neighboring vertex logits. These are different costs, and both remain in Equation (F.1).

Proof. For a vector witness w, form the uncentered matrix $C \in \mathbb{R}^{K\times M}$ with columns $C_j = w(p_W(v_j))$. Its entries have absolute value at most one, giving $\|C\|_F \le \sqrt{KM}$. Lemma F.1 gives, coordinatewise in $k \in [K]$,

$$\begin{aligned}|w_k(p_W(x)) - (C\varphi(x))_k| &\le \sum_{j=1}^{M} \lambda_j(x)|w_k(\mathrm{softmax}(\bar{W}(x))) - w_k(\mathrm{softmax}(W_j))| \\ &\le c_\mu \xi_W(x).\end{aligned}$$

Pairing this estimate with the residual vector $r = Y - p_W$ yields

$$\begin{aligned}|P_n(\langle r, w(p_W)\rangle)| &\le |\langle G_S(W), C\rangle_F| + c_\mu P_n\big(\|r\|_1 \xi_W\big) \\ &\le \sqrt{KM}\|G_S(W)\|_F + c_\mu P_n\big(\|r\|_1 \xi_W\big).\end{aligned}$$

Using the uncentered interpolation error and the ℓ^1 residual avoids a $\sqrt{K}$ factor in this error term while respecting the centered approximation: QC has no larger norm and $\langle r, QC\varphi\rangle = \langle r, C\varphi\rangle$ because $\mathbf{1}^\top r = 0$.

For each active vertex index j,

$$\|W_j - \bar{W}(x)\|_2 \le \sum_{\ell=1}^{M} \lambda_\ell(x)\|W_j - W_\ell\|_2 \le \mathrm{osc}(W).$$

Hence $\xi_W \le \mathrm{osc}(W)$. Take the supremum over witnesses and use $\|Y - p_W\|_1 \le 2$ to obtain the empirical bounds. Adding the radius-R_T uniform deviation proves (F.1). The sharper observable term $c_\mu P_n(\|Y - p_W\|_1 \xi_W)$ may also be retained. □

If the realized logit interpolant is L-Lipschitz and every simplex has diameter at most h, then adjacent vertex logits differ by at most Lh, so $\mathrm{osc}(W) \le Lh$. A bound on L that remains uniform under mesh refinement requires additional control beyond GD and $\|W_t\|_F \le R_T$; these alone permit positive representation error.

F.2 A PARAMETER-UNIFORM CERTIFICATE FOR A KNOWN INPUT MARGINAL

Now let $X \sim \mathrm{Unif}[0,1]$. Fix an integer $M \ge 2$ and vertices $v_j = (j-1)/(M-1)$ for $j = 1, \dots, M$. The feature vector $\varphi(x) \in \mathbb{R}^M$ contains the continuous piecewise-linear hat functions: its jth coordinate is one at v_j, zero at the other vertices, and linear between adjacent vertices. The coordinates are nonnegative and sum to one, so $\|\varphi(x)\|_2 \le 1$. Fix this feature map before drawing n independent observations under an arbitrary conditional categorical label law.

For $W \in \mathbb{R}^{K\times M}$, the logits $W\varphi(x)$ interpolate adjacent columns of W. Define the scalar coefficient budget, adjacent-column differences, and their average by

$$B = \sqrt{KM}, \qquad d_j(W) = \|W_{\cdot,j+1} - W_{\cdot,j}\|_2, \qquad \bar{d}(W) = \frac{1}{M-1}\sum_{j=1}^{M-1} d_j(W).$$

Each scalar $d_j(W)$ measures the logit change across one interpolation interval. The average $\bar{d}(W)$ controls the integrated witness error.

The empirical gradient and input-marginal correction are the matrices

$$G_S(W) = P_n\big((p_W - Y)\varphi^\top\big) \in \mathbb{R}^{K\times M},$$

$$D_S(W) = (P_X - P_{n,X})\big(p_W\varphi^\top\big) \in \mathbb{R}^{K\times M},$$

where $P_{n,X}$ averages over the training inputs. The correction requires the known input marginal and the training sample, but not the conditional label law. Suppose numerical integration produces a matrix $\widetilde{D}_S(W)$ and nonnegative scalar $e_q(W)$ satisfying, for every W, the deterministic real-arithmetic enclosure

$$\|\widetilde{D}_S(W) - D_S(W)\|_F \le e_q(W).$$

For $\delta \in (0,1)$, define the scalar label-moment allowance

$$\rho_n(\delta) = \sqrt{\frac{2}{3n}} + \sqrt{\frac{\log(1/\delta)}{n}}.$$

This allowance is distinct from the observed residual RMS $\rho_S(W)$ in Theorem 3.

Proposition F.3 (Known-marginal training certificate)**.** *With probability at least $1-\delta$, simultaneously for every $W \in \mathbb{R}^{K\times M}$ and every $m \ge 0$,*

$$\begin{aligned}\mathrm{psCE}_m(W) \le \min\Bigg\{2,\ B\left[\|G_S(W) + \widetilde{D}_S(W)\|_F + e_q(W) + \rho_n(\delta)\right] \\ + \frac{2c_\mu}{3}\bar{d}(W)\Bigg\}.\end{aligned} \tag{F.2}$$

Replacing the norm of the sum by $\|G_S(W)\|_F + \|\widetilde{D}_S(W)\|_F$ gives a decomposition into optimization, observable input-marginal correction, statistical, numerical-integration, and witness-approximation terms. Moreover, the population approximation profile satisfies

$$\varepsilon_{m,P}(W;B) \le \sqrt{K}\, c_\mu \left(\frac{2}{15(M-1)} \sum_{j=1}^{M-1} d_j(W)^2\right)^{1/2}. \tag{F.3}$$

For J prespecified feature-map/sample pairs, replacing δ by δ/J gives simultaneous coverage of all pairs and all their training iterates; independence between the J pairs is unnecessary.

Two different simplifications make Proposition F.3 possible. Knowledge of the input marginal isolates all label sampling error in one parameter-independent matrix. The hat features then approximate each witness using its values at the vertices, so the representation term depends on adjacent logit differences. Consequently, a finer feature grid changes both the coefficient budget and these differences; it need not improve the complete bound in Equation (F.2) automatically.

Proof. Consider the random matrix $U = Y\varphi(X)^\top \in \mathbb{R}^{K\times M}$. Its entries are nonnegative and sum to one, so the set of its possible values has Frobenius diameter at most $\sqrt{2}$. Equal spacing and the uniform input marginal give

$$\mathbb{E}\|U\|_F^2 = \mathbb{E}\|\varphi(X)\|_2^2 = \frac{2}{3}.$$

Jensen's inequality and independence imply

$$\mathbb{E}\|P_n(U) - P(U)\|_F \le \left(\mathbb{E}\|P_n(U) - P(U)\|_F^2\right)^{1/2} \le \sqrt{\frac{2}{3n}}.$$

Replacing one observation changes $\|P_n(U) - P(U)\|_F$ by at most $\sqrt{2}/n$. McDiarmid's inequality therefore bounds this norm by $\rho_n(\delta)$ with probability at least $1-\delta$.

This event does not depend on the parameter matrix. Writing $G_P(W) = P((p_W - Y)\varphi^\top)$ for the population gradient, the identity

$$G_P(W) = G_S(W) + D_S(W) + (P_n(U) - P(U))$$

holds for every W. The same event consequently controls all iterates and parameter selections from the sample, without a parameter-radius cover.

It remains to bound witness representation. For a vector witness $w \in \mathcal{W}_m$, form a matrix $C \in \mathbb{R}^{K\times M}$ with columns $C_j = w(p_W(v_j))$. Every entry has absolute value at most one, giving $\|C\|_F \le B$. On interval j, use the local coordinate $t \in [0,1]$ and define

$$\xi_W(x) = 2t(1-t)d_j(W).$$

Lemma F.1, applied to the two vertex logits and their convex interpolation, bounds each coordinate of $w(p_W(x)) - C\varphi(x)$ by $c_\mu \xi_W(x)$. Direct integration gives

$$P_X(\xi_W) = \frac{\bar{d}(W)}{3}, \qquad P_X(\xi_W^2) = \frac{2}{15(M-1)} \sum_{j=1}^{M-1} d_j(W)^2.$$

Pair the uncentered interpolation error with $Y - p_W$ and use $\|Y - p_W\|_1 \le 2$. The represented part pairs with $G_P(W)$, so taking the supremum over witnesses yields

$$\mathrm{psCE}_m(W) \le B\|G_P(W)\|_F + \frac{2c_\mu}{3}\bar{d}(W).$$

Centering C by Q contracts its Frobenius norm. The Euclidean interpolation error is at most $\sqrt{K}c_\mu\xi_W(x)$, which proves (F.3). Substituting the uniform label-moment bound and the integration enclosure proves (F.2). A union bound gives the assertion for J prespecified pairs. □

F.3 Numerical integration and the final GD iterate

Proposition F.3 permits any justified enclosure for the input expectation. In the experiment in Appendix G.3, each hat interval uses an integer $q \ge 1$ of midpoint nodes to integrate $p_W\varphi^\top$; its empirical counterpart is evaluated on the training inputs. The resulting real-arithmetic allowance is

$$e_q(W) = \frac{\bar{d}(W)/2 + \sqrt{2}}{4q}. \tag{F.4}$$

To verify it, let $h = 1/(M-1)$ denote the width of an interval. On interval j, the feature derivative has norm $\|\varphi'\|_2 = \sqrt{2}/h$, while the softmax Jacobian bound gives $\|p'_W\|_2 \le d_j(W)/(2h)$. The Frobenius Lipschitz constant of $p_W\varphi^\top$ is therefore at most $(d_j(W)/2 + \sqrt{2})/h$. Its midpoint integration error is at most this constant times $h^2/(4q)$. Summing over intervals proves (F.4). This bounds discretization error in real arithmetic; floating-point evaluation has the separate numerical scope stated with the experiment.

For unregularized full-batch cross-entropy GD from $W_0 = 0$, fix $0 < \alpha \le 2$. The feature norm bound makes the empirical objective convex and $1/2$-smooth. The proof of Theorem 3 gives

$$\sum_{t=0}^{T-1} \|G_S(W_t)\|_F^2 \le \frac{2\log K}{\alpha}.$$

To recall why this controls the final iterate, let H_t be the empirical Hessian averaged along the segment from W_t to W_{t+1} after vectorization. Then

$$\operatorname{vec} G_S(W_{t+1}) = (I - \alpha H_t)\operatorname{vec} G_S(W_t), \qquad 0 \preceq H_t \preceq \tfrac{1}{2}I.$$

Since $\alpha/2 \le 1$, the gradient norms are nonincreasing. Thus $\|G_S(W_T)\|_F \le \sqrt{2\log K/(\alpha T)}$. On the event in Proposition F.3,

$$\begin{aligned}\mathrm{psCE}_m(W_T) \le \min\Bigg\{2,\ B\left[\sqrt{\frac{2\log K}{\alpha T}} + \|\widetilde{D}_S(W_T)\|_F + e_q(W_T) + \rho_n(\delta)\right] \\ + \frac{2c_\mu}{3}\bar{d}(W_T)\Bigg\}.\end{aligned} \tag{F.5}$$

In Equation (F.5), the statistical allowance has no parameter-radius growth with T. Nevertheless, the marginal correction and representation contribution remain explicit and may be positive after optimization. Continuous inputs and finite M give a feature model that does not represent all functions on the input support. We evaluate these certificates for this continuous-input setting in Appendix G.3.

G Selected simulations and feature-certificate studies

We report four study families that isolate the mechanisms in the main results: calibrated-estimator fluctuations, representation at stationarity, finite-sample certificates under a known input marginal,

and full-class approximation with frozen neural features. We specify the fluctuation construction in Appendix G.1, the representation controls in Appendix G.2, and the continuous-input and neural-feature studies in Appendices G.3 and G.4. The separate supplement contains complementary audit methods, real-data comparisons, and numerical checks. The editable plotting tables were reconstructed from numerical summaries in vector figures. They are derived plotting data, not original run-level records. The original records are not included in this source package.

G.1 Empirical fluctuations at exact calibration

We simulate the construction underlying Theorem 1. For each odd class count K, let J be uniform on $\{1, \dots, K\}$, and define the probability vector

$$p_j = \frac{1}{2}\left(e_j + \frac{\mathbf{1}}{K}\right), \qquad j = 1, \dots, K.$$

Conditional on $J = j$, the one-hot label Y has mean p_j. Thus the predictor p_J is exactly calibrated and has strictly positive coordinates. We use this model to measure empirical fluctuations in the absence of population miscalibration.

The protocol uses the following fixed choices:

- Class counts $K \in \{3, 5, 9, 17\}$ and sample sizes $n \in \{2K^2, 8K^2, 32K^2\}$.
- For every (K, n) pair, 1000 independent multinomial count arrays representing n i.i.d. observations of (J, Y).
- Projection budgets $m \in \{1, (K-1)/2, K-1\}$, with duplicate values removed. The same count array is reused across these budgets.

To specify the recorded statistic, let $N_{\ell k}$ be the number of observations with $J = \ell$ and $Y = e_k$, and let $N_\ell = \sum_{k=1}^{K} N_{\ell k}$. The count matrix $(N_{\ell k})$ has K rows and K columns. Define the scalar residual sums and direction amplitude by

$$S_{k\ell} = N_{\ell k} - N_\ell p_{\ell,k}, \qquad b = \sqrt{\frac{m}{K-1}}.$$

For each output coordinate k, we order the indices other than k increasingly and pair consecutive indices. On each pair (ℓ, ℓ'), the direction has coefficients $b\,\mathrm{sgn}(S_{k\ell} - S_{k\ell'})$ and its negative; its kth coefficient is zero. These directions satisfy the prescribed cube and Euclidean-ball constraints. With the identity outer function, the resulting feasible-witness score is

$$\frac{b}{2n} \sum_{k=1}^{K} \sum_{(\ell,\ell') \text{ paired for } k} |S_{k\ell} - S_{k\ell'}|.$$

This statistic lower-bounds the full empirical calibration metric. It is not a numerical solution of the full supremum.

The design produces 12,000 independent count arrays and 33,000 scores. These counts differ because $K = 3$ has two distinct budgets, whereas each other class count has three. Scores from one array are dependent across budgets, and their $\sqrt{m}$ dependence is algebraic. The original protocol describes saved count arrays, per-sample seeds, score normalizations, and feasibility checks; these experiment-level outputs are not included in the source package.

Each point in Figure 4 averages 1000 independent count arrays. Error bars are pointwise approximate 95% normal-approximation Monte Carlo intervals for the mean feasible-witness score. The observed means, normalized by $\sqrt{mK/n}$, range from 0.148 to 0.190. The analytical curves compare these means with the construction-specific expected-score lower bound and its universal simplification $1/12$ from Proposition A.5.

G.2 Feature representation and stationary calibration

The first control separates calibration from individual label noise. Write the input as $X = (Z, U)$, where $Z \in \{0, 1\}$ is a coarse cell with probabilities $P(Z = 0) = 0.45$ and $P(Z = 1) = 0.55$, and U

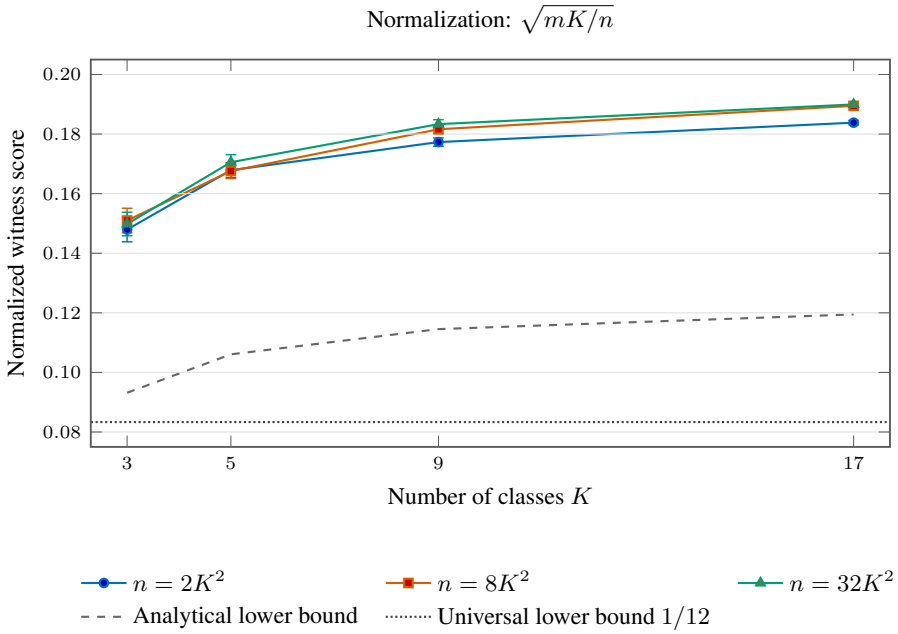

Figure 4: Empirical fluctuations at exact calibration. The normalized mean feasible-witness scores exceed the construction-specific lower bound on the expected normalized score and its universal simplification $1/12$. Error bars are pointwise approximate 95% Monte Carlo intervals for the mean score.

Table 2: Conditional label laws for the noisy two-cell control. The latent sign U takes values -1 and 1 with equal probability.

K	Z	c_Z	d_Z
3	0	$(.65, .25, .10)$	$(.10, -.08, -.02)$
	1	$(.15, .30, .55)$	$(-.06, .08, -.02)$
5	0	$(.45, .20, .15, .12, .08)$	$(.08, -.05, .025, -.035, -.02)$
	1	$(.08, .12, .15, .25, .40)$	$(-.025, .03, -.035, .07, -.04)$

is an independent uniform sign. The feature map has values

$$\varphi(0, U) = (1, 0)^\top, \qquad \varphi(1, U) = (0, 1)^\top.$$

Consequently, the parameter matrix $W \in \mathbb{R}^{K\times 2}$ assigns one logit vector to each cell and cannot distinguish its two latent subclasses. Conditional label probabilities are

$$\eta(Z, U) = c_Z + U d_Z,$$

where the probability vectors $c_Z \in \mathbb{R}^K$ and zero-sum perturbations $d_Z \in \mathbb{R}^K$ are specified in Table 2. Every coordinate of $c_Z + U d_Z$ is strictly positive.

For the label laws in Table 2, the restricted-model population minimizer is $p^*(Z) = c_Z$. At any parameter value, every centered witness is represented by the two-column coefficient matrix

$$A = \left[Qw(p_W(0)),\ Qw(p_W(1))\right] \in \mathbb{R}^{K\times 2}, \qquad \|A\|_F \le \sqrt{2K},$$

where $p_W(z)$ denotes the prediction in cell z. The model therefore represents all prediction-group indicators despite omitting information relevant to the conditional label law.

The training and evaluation protocol is:

- Ordinary full-batch cross-entropy GD starts from $W_0 = 0$, uses step size $\alpha = 1$, and runs for 4096 updates. The unit-norm features place this step size within the descent range.

Table 3: Full population psCE_1 after 4096 empirical GD updates in the two-cell control. Entries give mean $\pm$ sample standard deviation over 20 independent training datasets.

n	$K = 3$	$K = 5$
128	0.103762 ± 0.036771	0.147416 ± 0.047865
512	0.051620 ± 0.016446	0.072639 ± 0.027027
2048	0.022799 ± 0.013036	0.033979 ± 0.011154
8192	0.013356 ± 0.008120	0.021970 ± 0.007257

- Each $K \in \{3, 5\}$ and $n \in \{128, 512, 2048, 8192\}$ receives 20 independent training datasets. Joint multinomial counts reproduce the i.i.d.-sample gradients exactly. Saved checkpoints are $0, 1, 4, 16, 64, 256, 1024, 4096$.
- Two additional population-gradient trajectories isolate optimization. Exact finite-support integration gives population diagnostics. The audit uses the finite active-set support formula for the two distinct predictions and budgets $m \in \{0, 0.25, 1, 2, K\}$, after removing duplicates.

Figure 2a shows calibration falling below 10^{-14} under population GD, while expected individual residual norms remain 1.0980 and 1.4504 for $K = 3, 5$. Bayes-excess loss can remain positive because the features omit U (Figure 5a). Table 3 reports the full population psCE_1 after empirical training, as mean $\pm$ sample standard deviation across the 20 datasets. We evaluate the fitted predictors by summing over the known finite distribution, rather than estimating their calibration from an audit sample. Figure 5b displays the same decrease with sample size.

For the converse control, we use the three-point affine construction in Appendix B.1. Its population minimizer p_* is stationary but has the positive witness correlation $1/150$.

To compute the full metric, parameterize the prediction line by

$$p(s) = \big(s, (1-s)/2, (1-s)/2\big), \qquad s \in [0, 1].$$

For a direction $a \in \mathbb{R}^3$, the scalar projection has slope $a_1 - (a_2 + a_3)/2$. Its maximum absolute value under the cube and squared-norm constraints is the corresponding support function at $(1, -1/2, -1/2)^\top$. A single feasible direction attains all maximal pairwise distances, reducing each coordinate supremum to a three-node bounded-Lipschitz linear program. At $m = 1$, the maximal slope is $\sqrt{3/2}$, and

$$\mathrm{psCE}_1(p_*) = 2(0.1/3)(0.15 + 0.25)\sqrt{3/2} = \frac{2\sqrt{6}}{150}.$$

The numerical linear program agrees with this value to 10^{-10}; it yields the positive stationary limit in Figure 2b. Population GD starts from zero and runs for 16384 updates with step size one; the bound $\|\varphi(x)\|_2^2 \le 2$ places this step within the descent range. The final gradient norm is 3.46×10^{-17}, while the full metric is 0.0326598632371.

The two-point audit is solver-free. The three-point linear program requires successful SciPy HiGHS termination with primal and dual feasibility tolerances 10^{-9} and records constraint violations. Every saved checkpoint is checked against the stationarity certificate, the residual upper bound, nonnegativity of the metrics, and nonincreasing cross-entropy. These float64 checks use no formal rounding-error certification. Logarithmic plots display values below 10^{-14} at 10^{-14}.

For the separate finite-cell concentration calculation, we derive a bound that holds uniformly over predictors constant on the two predetermined cells in Supplement Section S5.1. At $\delta = 0.05$ and $n = 8192$, its deviation radii are 0.08358785 for $K = 3$ and 0.09846083 for $K = 5$. The 95% event applies per dataset, simultaneously over cell-constant predictors, checkpoints, and budgets; it does not assert joint coverage across datasets. The corresponding numerical records appear in Supplement Section S6.2; Figure 2 reports the known population metrics.

G.3 Finite-sample certificates with a known input marginal

The controls in Appendix G.2 separate calibration from label noise and illustrate the role of exact feature representation. We now examine a finite feature model on a continuous input distribution, where

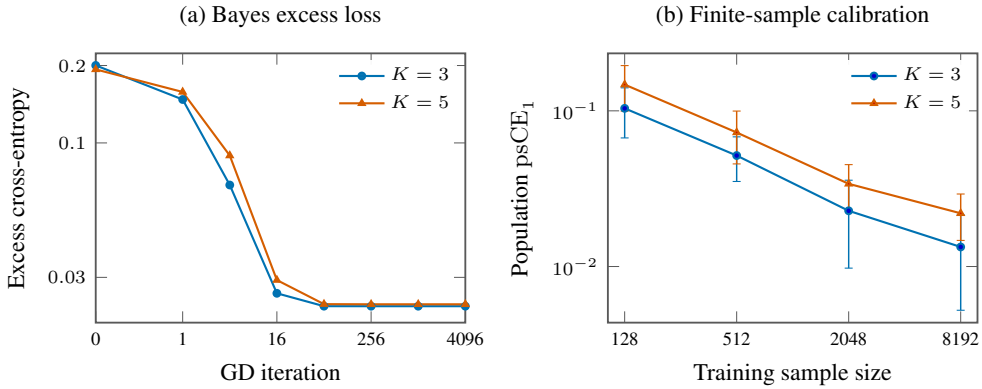

Figure 5: Additional two-cell controls. (a) Positive Bayes-excess loss persists because the features omit the latent subclass. (b) Full population calibration after empirical training improves with sample size; error bars show one sample standard deviation over 20 datasets.

the features approximate witnesses imperfectly. This study evaluates the optimization–representation decomposition of Theorem 3 using the known-marginal specialization in Appendix F. Knowledge of the input marginal permits concentration of a finite label-moment matrix in place of the general parameter-class deviation. The reported numbers use this refinement rather than a direct evaluation of the generic Δ_n term in Theorem 3.

We generate three independent datasets, each containing $n = 32768$ i.i.d. observations. The input is uniform on $[0, 1]$, there are $K = 3$ classes, and the conditional class-probability vector is

$$\eta(x) = \mathrm{softmax}\begin{pmatrix} 0.7\sin(2\pi x) + 0.4\sin(40\pi x) \\ 0.5\cos(2\pi x) - 0.3\sin(40\pi x) \\ -0.2\sin(2\pi x) \end{pmatrix}.$$

For $M \in \{9, 17, 33\}$, the feature vector $\varphi(x) \in \mathbb{R}^M$ consists of the continuous piecewise-linear hat functions on the equally spaced vertices $v_j = (j-1)/(M-1)$, $j = 1, \ldots, M$. The jth function equals one at v_j, zero at every other vertex, and is linear between adjacent vertices. These nonnegative features sum to one and satisfy $\|\varphi(x)\|_2 \le 1$. We fit all three feature maps to each dataset, giving nine models. Their logits are $W\varphi(x)$ for $W \in \mathbb{R}^{3\times M}$.

We use unregularized full-batch cross-entropy GD from zero, with step size $\alpha = 2$ and $T = 4096$ updates. Evaluation uses projection budget $m = 1$ and $q = 4096$ midpoint quadrature nodes per feature interval. The total failure probability 0.05 is divided across the nine prespecified feature-map/sample pairs. A single label-moment event for each pair covers every parameter value and hence every training iterate. This remains valid when the three feature maps share a dataset.

To identify the terms being measured, we define the coefficient budget and adjacent-vertex logit differences by

$$B = \sqrt{3M}, \qquad d_j(W) = \|W_{\cdot,j+1} - W_{\cdot,j}\|_2, \qquad \bar{d}(W) = \frac{1}{M-1}\sum_{j=1}^{M-1} d_j(W).$$

The scalar $\bar{d}(W)$ controls witness interpolation error. Let $P_{n,X}$ denote averaging over the training inputs. The observable input-marginal correction is the $3 \times M$ matrix

$$D_S(W) = P_X(p_W\varphi^\top) - P_{n,X}(p_W\varphi^\top).$$

Its quadrature approximation $\widetilde{D}_S(W)$ uses the known uniform marginal. The scalar allowance

$$e_q(W) = \frac{\bar{d}(W)/2 + \sqrt{2}}{4q}$$

bounds the Frobenius error from midpoint discretization in real arithmetic. The scalar label-moment deviation for a failure probability $u \in (0, 1)$ is

$$\rho_n(u) = \sqrt{\frac{2}{3n}} + \sqrt{\frac{\log(1/u)}{n}}.$$

Table 4: Known-marginal population certificates after 4096 updates. Entries are means over three independent datasets. "Other terms" combines input-marginal correction, label-moment concentration, and quadrature.

M	Gradient	Approximation	Other terms	Observed	Finite-T
9	$< 5 \times 10^{-9}$	0.138	0.104	0.242	0.327
17	$< 5 \times 10^{-9}$	0.078	0.143	0.221	0.338
33	$< 5 \times 10^{-9}$	0.145	0.199	0.344	0.507

The label-moment deviation $\rho_n(u)$ is distinct from the observed residual RMS $\rho_S(W)$ in Theorem 3.

Applying Proposition F.3 with $K = 3$, $m = 1$, and failure probability $0.05/9$ gives the population upper bound

$$\begin{aligned}\mathrm{psCE}_1(W) \le \min\big\{2,\, B\|G_S(W)\|_F + B\|\widetilde{D}_S(W)\|_F \\ + B\rho_n(0.05/9) + Be_q(W) + \bar{d}(W)/3\big\}.\end{aligned} \tag{G.1}$$

In Equation (G.1), $G_S(W) = \nabla\widehat{L}(W)$ is the empirical gradient. The five terms in the bound represent optimization, observable input-marginal correction, label-moment concentration, quadrature, and witness approximation. The conditional law η is used to generate observations and evaluate the separate residual and Bayes-excess comparisons; it is not an input to this certificate.

Table 4 reports the components of Equation (G.1). The observed certificate uses the final gradient; the finite-T certificate replaces its norm by $\sqrt{2\log K/(\alpha T)}$. Every final certificate, including this latter version, lies below the corresponding lower bound on the mean individual label residual $P(\|Y - p_W\|_1)$, which exceeds 1.26 in all nine models. The largest observed optimization contribution is 4.45×10^{-9}, so representation and statistical terms determine the final certificate values.

The comparison between $M = 17$ and $M = 33$ is particularly informative: both the approximation contribution and the remaining terms increase on the finer mesh. A larger feature set therefore does not automatically improve the complete certificate. The population RMS approximation upper bounds from Equation (F.3), at budget $\sqrt{3M}$, have means 0.135, 0.079, and 0.153 for $M = 9, 17, 33$, respectively. These bounds concern the RMS approximation profile; the approximation column of Table 4 instead reports the interpolation contribution $\bar{d}(W)/3$.

Figure 3 displays the same tradeoff throughout training. The underlying real-arithmetic event gives 95% simultaneous coverage for all nine models and their iterates. The recorded evaluations use float64 without formal rounding-error certification. The original numerical verification reports that independent replay reproduces a saved 64-step checkpoint and that doubling the quadrature resolution changes one recorded population moment by 8.3×10^{-11}, within its discretization allowance. We derive the certificate and control its quadrature error in Appendix F. The separate residual and Bayes-excess integration checks are recorded in Supplement Section S7.1.

G.4 Full-class approximation with frozen neural features

The known-marginal study in Appendix G.3 uses explicit interpolation features. We also evaluate whether the anchor operator of Proposition D.1 yields a useful bound for learned features. The CIFAR-10 study has two distinct stages: empirical witness approximation on the head-training sample, followed by a retrospective calculation of generalization on the official test set. Keeping these stages separate identifies the statistical cost that an empirical certificate alone does not measure.

The fixed data and training design is as follows:

- A label-independent permutation with seed 20270916 splits the 50000 training images into 20000 auxiliary encoder images, a disjoint 20000-image head pool, and 10000 validation images. The 10000 official test images supply the later audit.
- The encoder uses three 3×3 convolutional layers with $16, 32, 64$ channels, batch normalization, and ReLU activations; the last two layers have stride two. Average pooling and a linear map produce a 128-dimensional ReLU representation. Auxiliary-only training uses

seed 20270917, 12 epochs, batch size 256, Adam step size 10^{-3}, weight decay 5×10^{-4}, and random horizontal flips.

- We freeze the encoder in evaluation mode, subtract the auxiliary embedding mean, and divide by the maximum of the centered norm and 10^{-12}. Appending an intercept and dividing by $\sqrt{2}$ gives a feature map $\varphi : \mathcal{X} \to \mathbb{R}^{129}$ with $\|\varphi(x)\|_2 \le 1$.
- Seeds $0, 1, 2, 3, 4$ select 16000-image subsets without replacement from the head pool. Each unscaled head uses unregularized full-batch cross-entropy GD from zero, step size two, and 2048 updates. The five heads share their encoder and have overlapping training subsets.

We evaluate the final heads at $m = 1$. For each head, deterministic farthest-point selection supplies $M \in \{16, 32, 64\}$ prediction anchors. We fit the assignment indicators by ridge regression with penalties $10^{-6}, 10^{-4}, 10^{-2}, 1$ and an unpenalized intercept. For each operator, the set Γ in Proposition D.1 contains 81 logarithmically spaced Young-inequality parameters from 10^{-2} to 10^{2}. The five heads, three anchor counts, and four ridge penalties give 60 head–operator configurations. This grid records refinements made after exploratory analysis. Because every operator certificate is a deterministic sample inequality, choosing the smallest valid empirical bound within this grid requires no confidence correction.

Each row of the main-text Table 1 uses 64 anchors and ridge penalty 10^{-4}, which minimize that head's reported empirical calibration bound. The approximation columns give lower and upper bounds on $\varepsilon_{1,S}(W; B)$ at the displayed coefficient budget B. The sample-certificate column uses the observed-gradient bound

$$B\|G_S(W)\|_F + \rho_S(W)\, e,$$

where e is the corresponding full-class approximation upper bound. The residual comparator is the empirical mean $P_n(\|Y - p_W\|_1)$.

The full-class empirical certificate improves on the residual comparator for every head. At the same time, the positive approximation lower bounds show that representation error remains at these finite coefficient budgets. These observations support the decomposition in Theorem 3: a small gradient and an informative certificate can coexist with positive approximation error at the selected finite budgets. We obtain the approximation lower bounds in Table 1 by testing whether a finite coefficient budget can represent one explicit witness. At $m = 1$, we choose the feasible vector witness with coordinates

$$w_k(p) = |p_k - 1/2| - 1/2, \qquad k = 1, \dots, K.$$

For coordinate k, the direction is $a = e_k$ and the outer function is $\phi(t) = |t - 1/2| - 1/2$ on $[-1, 1]$. Let $\Phi \in \mathbb{R}^{n \times d}$ have rows $\varphi(X_i)^\top$, and let $T \in \mathbb{R}^{n \times K}$ have rows $(Qw(p_i))^\top$, where $p_i = p_W(X_i)$. Here T is a witness-evaluation matrix, distinct from the training horizon. For any nonzero matrix $U \in \mathbb{R}^{n \times K}$, Frobenius duality gives

$$\varepsilon_{1,S}(W; B) \ge \frac{[\langle U, T\rangle_F - B\|\Phi^\top U\|_F]_+}{\sqrt{n}\|U\|_F}.$$

Indeed, the dual pairing with $T - \Phi A^\top$ is at least $\langle U, T\rangle_F - B\|\Phi^\top U\|_F$ for every feasible coefficient matrix A, and Cauchy–Schwarz bounds this pairing by $\|U\|_F\|T - \Phi A^\top\|_F$. The calculation uses a computed least-squares residual U and retains its dual defect $\|\Phi^\top U\|_F$, which accounts for imperfect numerical orthogonality at the reported finite budget. At $m = 0$, the constant feature gives $\varepsilon_{0,S}(W; \sqrt{2K}) = 0$. The calculations use float64 with positive engineering padding; formal rounding-error certification would additionally certify the numerical endpoints.

For generalization, we retain two nonconstant operators with $M \in \{32, 64\}$ and ridge penalty 10^{-4}, plus one constant-operator baseline, for each head. This yields 15 predictor–operator candidates. The predictors, anchors, and operators use the original head-training inputs, while their fixed-operator errors and gradients are evaluated on the 10000 official test observations. The calculation allocates total failure probability 0.05 across the 15 candidates under the stated prospective sampling assumptions. We derive the full-class approximation transfer and gradient concentration bounds under the required independence conditions in Appendix E.

For this study, $K = 10$ and the feature dimension is $d = 129$. Specializing Equation (E.1) to $m = 1$, we write the population approximation profile as

$$\varepsilon_{1,P}(W; B) = \sup_{w \in \mathcal{W}_1} \inf_{\substack{A \in \mathbb{R}^{K \times d} \\ \|A\|_F \le B}} \left(P_X\left(\|Qw(p_W) - A\varphi\|_2^2\right)\right)^{1/2}.$$

The expectation averages the squared approximation error over a fresh population input. The matrix A may depend on the witness, as in the sample profile, and the same coefficient budget B applies to every witness.

Table 5: Retrospective generalization calculation at the final unscaled GD checkpoint, $m = 1$. Entries are means across five dependent heads.

Operator	B	Audit approximation bound	Population approximation bound	Population calibration bound
Constant baseline	4.472	0.916	0.974	0.761
Neural anchors, $M = 32$	16.088	0.909	1.341	1.302
Neural anchors, $M = 64$	16.151	0.846	1.285	1.266

In Table 5, the audit approximation bound controls the RMS error of the same fixed full-class operator on test inputs. Under the independent-audit assumptions of Proposition E.2, the population approximation bound controls $\varepsilon_{1,P}(W; B)$. The final column combines that representation bound with an audit-based population-gradient bound from Corollary E.3. The nonconstant operators reduce empirical approximation error, but their larger coefficient budgets and error ranges increase the generalization cost. Consequently, the constant operator yields the smallest reported population calibration bound.

The heads share the encoder, some training inputs, and previously inspected test data. The generalization comparison is therefore retrospective; prospective coverage requires an independent audit. Tables 1 and 5 use float64 with engineering padding, without formal rounding-error certification.

H NOTATION REFERENCE

We collect the notation by role. The symbols ϕ and φ denote, respectively, a scalar witness's outer function and the feature map. The support size, anchor count, and interpolation vertex count are each denoted by M in their separate constructions; none is the sample size n. Auxiliary variables introduced only within a proof retain their local definitions.

H.1 DATA AND PREDICTIONS

Symbol	Meaning
K, $[K]$, e_j	Number of classes; $\{1, \dots, K\}$; the jth standard basis vector.
X, Y, $\mathcal{X}$	Input, one-hot label $Y \in \{e_1, \dots, e_K\}$, and input space.
P, P_X, $\eta(x)$	Joint law, input marginal, and conditional class probabilities $\mathbb{E}[Y \mid X = x]$.
S, n, P_n	I.i.d. sample, sample size, and empirical averaging operator.
Δ^{K-1}	Probability simplex in $\mathbb{R}^K$.
g, p_g	Logit function and its softmax prediction; $p_W = \mathrm{softmax}(W\varphi)$ in the feature model.
x_s, M, π_s	Finite support points, their number, and positive masses $P_X(\{x_s\})$ in Theorem 2.
$p^{(j)}$, J, I_j, v_j	Distinct predictions, their number, equal-prediction groups, and group-indicator vectors in $\mathbb{R}^M$.

H.2 LOSS, GRADIENTS, AND MATRIX OPERATIONS

Symbol	Meaning
$\ell(z, y)$	Cross-entropy $\log \sum_j e^{z_j} - \langle y, z \rangle$.
L, $\widehat{L}$	Population and empirical cross-entropy risks.
$\nabla L(W)$, $G_S(W)$	Population parameter gradient and empirical gradient $\nabla \widehat{L}(W)$.
G_t	Empirical gradient $G_S(W_t)$ at GD iterate t.
$\langle A, B \rangle_F$, $\|A\|_F$	Frobenius inner product $\mathrm{tr}(A^\top B)$ and its norm.
Q, $\mathbf{1}$, I_K	Class-centering projection $I_K - \mathbf{1}\mathbf{1}^\top / K$, all-ones vector, and identity matrix.
$\lambda_{\max}(H)$, $\|H\|_{\mathrm{op}}$	Largest eigenvalue of a symmetric matrix and Euclidean induced operator norm.

H.3 CALIBRATION OBJECTS

Symbol	Meaning
m, a, ϕ	Projection budget, direction with $\|a\|_\infty \le 1$ and $\|a\|_2^2 \le m$, and bounded 1-Lipschitz outer function.
$\mathcal{H}_m$, h	Scalar witness class and a witness $h(p) = \phi(a^\top p)$.
$\mathcal{W}_m$, w	Vector witness class $\mathcal{H}_m^K$ and a vector of K separately chosen scalar witnesses.
$F_{g,w}$	Residual test $\langle Y - p_g(X), w(p_g(X)) \rangle$.
psCE_m, $\widehat{\mathrm{psCE}}_m$	Supremum of absolute population and empirical witness correlations; each lies in $[0, 2]$.
$\rho_S(W)$	Empirical root-mean-square residual $\left(P_n \|Y - p_W\|_2^2\right)^{1/2} \le \sqrt{2}$.
$\varepsilon_{m,S}(W; B)$	Worst empirical RMS error when each centered witness is approximated by $A\varphi$ with $\|A\|_F \le B$; see Equation (8).
$\varepsilon_{m,P}(W; B)$	Population counterpart of the approximation profile; see Equation (E.1).

H.4 Statistical constants and classes

Symbol	Meaning
μ, r	Effective projection budget $\min\{m, K\}$ and truncation $\min\{\mu, n\}$ in the statistical bounds.
$A_\mu(n)$, $\mathcal{C}_\mu(n)$	Witness-complexity factors $1 + \sqrt{r}\log(en/r)$ and $1 + 12\sqrt{r}\log(en/r)$; the added term is zero at $r = 0$.
$B_{m,K,n}$	Fixed-prediction complexity bound $\min\{2, \mathcal{C}_\mu(n)\sqrt{2K/n}\}$.
L_μ	Logit Lipschitz constant $1/\sqrt{2} + \min\{\sqrt{\mu}, \sqrt{2}\}$ for residual-witness tests.
c_μ	Logit Lipschitz constant $\min\{\sqrt{\mu}/2, 1/\sqrt{2}\}$ for scalar witnesses.
δ, R, $\Delta_n(R, \delta)$	Failure probability, deterministic parameter radius, and uniform deviation allowance in Equation (11).
$\mathcal{G}$, $\mathcal{G}_R$, $\mathcal{G}_R^0$	Deterministic logit class, finite-feature class $\{W\varphi : \|W\|_F \le R\}$, and its centered subclass with $QW = W$.
$\mathcal{F}_{m,\mathcal{G}}$	Residual-test class $\{F_{g,w} : g \in \mathcal{G},\ w \in \mathcal{W}_m\}$.
$\widehat{\mathfrak{R}}_S$, σ_i	Conditional empirical Rademacher complexity and independent uniform signs.
$N(\mathcal{C}, d, \epsilon)$	Number of radius-ϵ balls with centers in $\mathcal{C}$ needed to cover it under metric d.

H.5 Feature representation and training

Symbol	Meaning
φ, d, Λ	Fixed feature map, feature dimension, and deterministic bound $\|\varphi(x)\|_2^2 \le \Lambda$.
W, A, B	Logit parameter matrix, witness coefficient matrix, and coefficient budget $\|A\|_F \le B$.
Φ	Feature evaluation matrix: $M \times d$ on a finite support or $n \times d$ on a sample.
$\mathcal{T}$, $\Pi_{\mathcal{T}}$	Centered feature span $\{QA\varphi\}$ and its orthogonal projection under the support-weighted inner product.
$\langle u, v\rangle_\pi$, $\|u\|_\pi$	Finite-support inner product $\sum_s \pi_s u(x_s)^\top v(x_s)$ and its norm.
α, T, W_t	GD step size, deterministic training horizon, and iterate from $W_0 = 0$.
R_T	Deterministic trajectory radius $\sqrt{2\alpha T \log K}$.
(B, e)	Coefficient budget and approximation upper bound satisfying $\varepsilon_{m,S}(W; B) \le e$.

H.6 Appendix and supplementary constructions

Anchor operator (Appendix D)

Symbol	Meaning
q_j, M, $s(i)$	Anchors in the probability simplex, anchor count, and assignment of sample i to an anchor.
Z, C	Assignment matrix $Z_{ij} = \mathbf{1}\{s(i) = j\}$ and anchor operator $C \in \mathbb{R}^{d\times M}$.
E, D, R	Assignment residual $Z - \Phi C$, its Gram matrix $E^\top E/n$, and coefficient Gram matrix $C^\top C$. Here R is a matrix, not the statistical radius.
V	Prediction-to-anchor displacement matrix $n^{-1}\sum_i (p_i - q_{s(i)})(p_i - q_{s(i)})^\top$.
$\sigma(H)$	Positive row-sum functional $\sum_j [(H\mathbf{1})_j]_+$, distinct from the assignment $s(i)$.
$\mathcal{L}(H)$	Anchor-difference matrix $\sum_{j<\ell}[-H_{j\ell}]_+(q_j - q_\ell)(q_j - q_\ell)^\top$.
Γ, γ	Finite positive grid and a Young-inequality parameter chosen from it.
B_C, e_C	Computable coefficient budget and approximation upper bound in Proposition D.1.
z, T_w, A_w	Scalar anchor values, the vector-witness matrix with rows $w(q_j)^\top$, and its feature coefficients $Q(CT_w)^\top$.

Independent-audit generalization (Appendix E)

Symbol	Meaning
N, P_N, J	Independent audit size, audit average, and number of prespecified predictor–operator pairs.
A_w, u, R	Fixed witness-to-coefficient map, empirical uniform squared-error bound, and range bound for those errors.
$\mathcal{U}(u, R, s)$	One-sided upper bound $\min\{R, u + Rs + \sqrt{(Rs)^2 + 2Rsu}\}$.
G, $\widehat{G}$	Population and independent-audit gradients.
$\widehat{V}$, $\widehat{v}$, $\widehat{q}$	Audit residual-weighted feature second moment, its largest eigenvalue, and audit residual second moment.
s, e_+, v_+, q_+	Confidence scale $\log(4J/\delta)/N$ and upper bounds for population approximation error, gradient second moment, and residual second moment.
γ_+	Gradient sampling allowance $\sqrt{2v_+ s} + 4\sqrt{2\Lambda}\, s/3$.

Interpolation with a known input marginal (Appendix F)

Symbol	Meaning
v_j, M, λ_j	Input vertices, their number, and piecewise-linear hat features; these v_j are input points, not group indicators.
W_j, $\bar{W}(x)$	Vertex logits $W_{\cdot j}$ and interpolated logits $\sum_j \lambda_j(x) W_j$.
$\xi_W(x)$, $\mathrm{osc}(W)$	Weighted vertex-to-interpolant logit distance and largest logit difference across a simplex.
$d_j(W)$, $\bar{d}(W)$	Adjacent-column logit differences and their average on $[0, 1]$.
$D_S(W)$, $P_{n,X}$	Input-marginal correction $(P_X - P_{n,X})(p_W \varphi^\top)$ and empirical input average.
$\widetilde{D}_S(W)$, $e_q(W)$	Numerical correction and quadrature error bound $\|\widetilde{D}_S - D_S\|_F \le e_q$.
$\rho_n(\delta)$	Label-moment allowance $\sqrt{2/(3n)} + \sqrt{\log(1/\delta)/n}$, distinct from residual RMS $\rho_S(W)$.

Reference-curvature trajectories (Supplement, Section S1)

These quantities occur in the separate supplementary trajectory analysis, not in Theorems 1–3.

Symbol	Meaning
Σ_S, s_S, L_S	Feature second moment $P_n(\varphi\varphi^\top)$, its largest eigenvalue, and smoothness constant $s_S/2$.
$\mathcal{V}$, $H_S(W)$	Centered parameter space $\{A : QA = A\}$ and empirical Hessian restricted to it.
U, γ_U, σ_U	Reference parameter, its gradient norm, and smallest restricted Hessian eigenvalue.
κ, τ_U, $\chi(x)$	Curvature variation constant $\sqrt{2\Lambda}$, test ratio $\kappa\gamma_U/\sigma_U$, and $(1-e^{-x})/x$ with $\chi(0)=1$.
d_U, R_U	Reference-to-minimizer bound $-\log(1-\tau_U)/\kappa$ and radius $\|U\|_F + d_U$.
κ_U, q_U, W^*	Curvature lower bound $\sigma_U(1-\tau_U)\chi(\kappa R_U)$, contraction factor $1-\alpha\kappa_U$, and centered empirical minimizer.
$R_{U,s}$, $q_{U,s}$, v_t	Radius from a saved checkpoint, its contraction factor, and scalar distance refinement.