How Accurate Is Accurate Enough?

Anonymous authors
Paper under double-blind review

Abstract

How accurate must a numerical approximation be within a learning system? Primitive error alone cannot answer this question: errors of the same magnitude can have very different consequences for losses, predictions, and gradients at different learning states. We study this question through the learning objective itself. The objective weights classwise numerical errors nonuniformly according to the current state, so the importance of an error depends not only on its magnitude but also on the class it affects and the weight that class receives. For softmax cross-entropy, we characterize this coupling between class weights and errors and derive the exact extrema of the signed loss change over pairings of fixed non-target probability and score-error multisets, with the target probability and target score error held fixed. Building on this structure, we establish finite-error guarantees that propagate primitive error to losses, probabilities, predictions, and feature gradients, then invert these guarantees to obtain a certified primitive tolerance for the current state under prescribed learning-level error requirements. We give a complete instantiation of the framework in high-dimensional von Mises–Fisher learning. Controlled interventions and a large collection of saved learning states show that identical primitive error can produce substantially different learning consequences, while certified numerical tolerances vary by orders of magnitude across states under the same learning-level requirements. These results show that the adequacy of a numerical approximation must be assessed in relation to the current learning state and the quantity to be preserved; numerical accuracy should itself be treated as part of the learning objective.

1 Introduction

Modern machine learning relies on approximate numerical computations, including special functions, normalization constants, integrals, and finite-precision operations (Higham, 2002; Micikevicius et al., 2018). Numerical implementations are commonly assessed by how closely they reproduce a mathematical function. Learning objectives, however, use these values to form scores, probabilities, losses, and gradients. The accuracy required of an intermediate computation therefore depends on how its error affects these downstream quantities (Rice, 1966; Blanchard et al., 2021).

The same primitive error can have different consequences across learning states. High confidence in the target class can attenuate the effect of a perturbation on the loss or gradient, while competing classes can make the learning signal more sensitive to the same error. Differentiation introduces another source of variation: even when the underlying numerical problem is unchanged, the coordinate scale determines how its error enters the gradient. The error of an approximation and the accuracy required by the current learning state are thus distinct quantities.

We study the second quantity by asking: **given a learning-level error budget, what primitive error allowance can be certified at the current state?** Starting from requirements on loss, probabilities, prediction stability, or feature gradients, we work backward to the largest allowance accepted by a specified certificate and error family. This state-conditioned tolerance depends on the learning state, the numerical realization, and the downstream requirement.

Our framework propagates primitive errors through the learning objective and inverts the resulting finite-error guarantees. The analysis retains primitive geometry, differentiation scale, and the current

Anonymous code: `anonymous.4open.science/r/state-conditioned-tolerance/`.

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

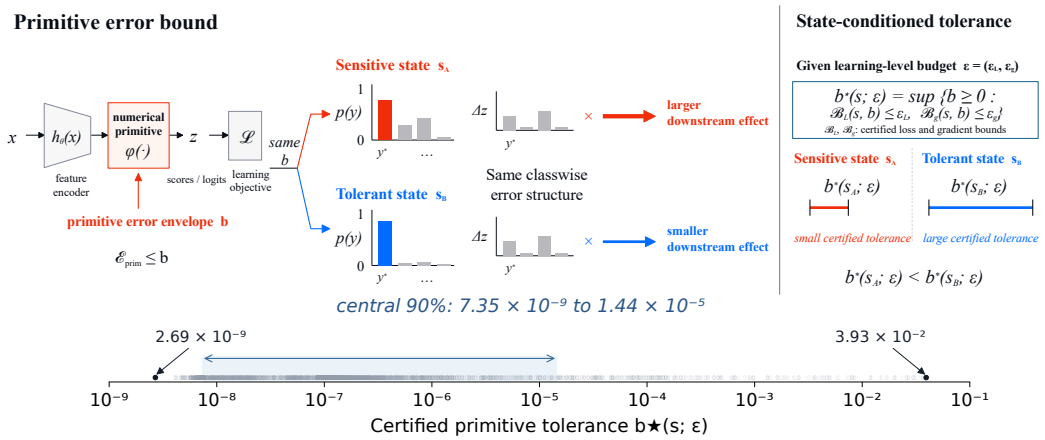

Figure 1: From primitive error bounds to state-conditioned numerical tolerance. A primitive error bound propagates through the learning objective. The same classwise error structure can have different downstream effects in different learning states. Inverting certified loss and gradient bounds under budgets ϵ_L and ϵ_g yields the certified primitive tolerance $b^\star(s;\boldsymbol{\epsilon})$. The upper panels are schematic. The lower panel shows saved-state tolerances under the joint budget $\epsilon_L = \epsilon_g = 10^{-5}$.

probability structure. For softmax cross-entropy, we identify a class-weight–error coupling: even with the target terms and the multisets of non-target probabilities and score errors fixed, changing their classwise pairing can change the loss error. The classical rearrangement inequality gives the exact extrema of the signed loss change. A first-order expansion further describes local certificate sensitivity; tolerances at finite budgets are obtained by nonlinear inversion.

Saved learning states exhibit substantial tolerance variation under common learning-level requirements, including within groups of fixed dimension and temperature. Controlled probability rearrangements isolate the classwise weighting mechanism. Holding reference target probability, entropy, exact top-two margin, geometry, and primitive errors fixed, changing the assignment of non-target probabilities to classes changes actual loss and gradient errors. These aggregate statistics therefore do not fully characterize the interaction between objective weights and numerical error.

Goal-oriented error estimation relates numerical discretization error to a selected quantity of interest and uses this relation to guide adaptive computation (Becker & Rannacher, 2001; Giles & Süli, 2002). We specialize this perspective to numerical primitives within learning objectives. Current probability weights and class structure yield computable finite-error guarantees that can be inverted into certified primitive tolerances. The cross-entropy coupling result identifies an additional structure of these objectives: the pairing of probability weights and score errors matters even when both multisets are fixed.

High-dimensional von Mises–Fisher learning provides a rigorous instantiation (Mardia & Jupp, 1999; Banerjee et al., 2005). Its special-function errors admit explicit bounds, allowing us to follow the full chain from a primitive allowance to a learning-level guarantee and its inverse.

Our contributions are:

1. We characterize class-weight–error coupling in softmax cross-entropy. With the target terms fixed, changing the pairing of fixed non-target probability and score-error multisets can change the loss error. Proposition 3 gives the exact extrema of the signed loss change.
2. We derive fixed-state loss, probability, prediction, and feature-gradient guarantees and characterize their local sensitivity. Inverting these guarantees yields a certified primitive tolerance under prescribed learning-level budgets.
3. We establish orders-of-magnitude variation in certified tolerances across saved states and isolate classwise weighting through controlled probability rearrangements. At fixed probability summaries, geometry, and primitive errors, class assignments change downstream errors and, under sufficiently strict budgets, numerical adequacy.

2 Problem Setup and Instantiation

2.1 Numerical adequacy at a fixed learning state

A learning state s specifies the inputs of a numerical primitive, the quantities held fixed, and the differentiation coordinate. A realization $\mathcal{N}$ replaces exact primitive values and derivatives inside an objective. For a specified downstream quantity G, let $E^G_{\mathcal{N}}(s)$ denote the induced absolute error, probability distance, or vector-norm error. Given a budget ϵ_G, numerical adequacy means $E^G_{\mathcal{N}}(s) \le \epsilon_G$.

A certificate $B^G_{\mathcal{N}}(s)$ supplies a computable sufficient condition,

$$E^G_{\mathcal{N}}(s) \le B^G_{\mathcal{N}}(s) \le \epsilon_G. \tag{1}$$

The certificate is computed from approximate outputs and valid primitive error allowances. Acceptance guarantees adequacy; otherwise, the bound leaves adequacy unresolved. Inverting the certificate therefore yields the *largest allowance certified by the chosen bound*, which may be smaller than the allowance that the actual learning error would permit.

2.2 A paired primitive for vMF learning

For feature dimension $p \ge 5$, let $\nu = p/2 - 1$. Define

$$\Phi_\nu(x) = \log I_\nu(x) - \nu \log x, \qquad R_\nu(x) = I_{\nu+1}(x)/I_\nu(x), \qquad \Phi'_\nu = R_\nu. \tag{2}$$

The ratio is central to directional inference and vMF parameter estimation (Sra, 2012). The functions use their continuous extensions at $x = 0$. We consider a differentiable paired realization (F, A) with $F' = A$ and $A(0) = 0$. Pairing makes the approximate objective and its feature derivative refer to the same mathematical surrogate.

Let $\mu_j \in \mathbb{R}^p$ be unit class directions, $k_j \ge 0$ concentrations, c_j finite shared class biases, $\tau > 0$ a temperature, and $f \in \mathbb{R}^p$ a feature of arbitrary norm. With finitely many classes $C \ge 2$, define

$$v_j = k_j \mu_j + f/\tau, \qquad r_j = \|v_j\|, \tag{3}$$

$$q_j = \Phi_\nu(r_j) - \Phi_\nu(k_j) + c_j, \qquad \widehat{q}_j = F(r_j) - F(k_j) + c_j. \tag{4}$$

Write $P = \mathrm{softmax}(q)$, $\pi = \widehat{P} = \mathrm{softmax}(\widehat{q})$, and $L_y = \mathrm{LSE}(q) - q_y$, with $\widehat{L}_y$ defined analogously. The log-partition and log-sum-exp structure connects probabilities to derivatives (Wainwright & Jordan, 2008; Boyd & Vandenberghe, 2004). Class statistics, biases, and temperature remain fixed in the Euclidean feature derivative ∇_f.

These normalizer differences occur in probabilistic contrastive objectives (Du et al., 2024; He et al., 2025). Supervised probabilistic vMF embeddings provide another instance, with an additional ratio-dependent attraction term (Scott et al., 2021); its propagation is given in Appendix G.

Our experiments instantiate (F, A) by the stable second-order pair $(F_{2,\nu}, A_{2,\nu})$. It satisfies $F'_{2,\nu} = A_{2,\nu}$ and admits both a uniform ratio allowance $(4\nu^3)^{-1}$ and a tighter state-dependent allowance. The construction, stable endpoint formula, and full special-function proofs are in Appendices D, E, and H. The analysis below applies to any paired realization with valid error allowances.

3 From Primitive Error to Learning Consequences

3.1 Score allowances and a geometry-only contract

Suppose a finite integrable allowance satisfies $|R_\nu(x) - A(x)| \le e(x)$ on every integration path between k_j and r_j. Pairing implies

$$|q_j - \widehat{q}_j| \le d_j := \int_{\min(k_j, r_j)}^{\max(k_j, r_j)} e(x)\, dx, \qquad |R_\nu(r_j) - A(r_j)| \le a_j. \tag{5}$$

Here a_j is an unscaled ratio allowance. Let $d = \max_j d_j$.

Proposition 1 (Fixed-state contrastive contract)**.** *For the state and paired realization in Section 2,*

$$|\widehat{L}_y - L_y| \le d_y + d, \qquad \mathrm{TV}(P, \pi) \le \tanh(d/2), \tag{6}$$

$$\|\nabla_f \widehat{L}_y - \nabla_f L_y\| \le \frac{2\tanh(d/2) + \max_j a_j + a_y}{\tau}. \tag{7}$$

If the approximate top-two score margin $\widehat{m}$ exceeds $2d$, the exact and approximate predictions agree. The same sufficient condition applied to the exact margin is valid for reference-side evaluation.

The bounds have no explicit class-count multiplier. They retain endpoint geometry and the derivative scale $1/\tau$, but largely discard the current probability structure. Proofs appear in Appendix F.

3.2 Retaining the current probability structure

Define

$$S_j^{\pm} = \sum_{i \ne j} \pi_i \{\exp[\pm(d_i + d_j)] - 1\}, \qquad \ell_j = \frac{\pi_j}{1 + S_j^+}, \quad u_j = \min\left\{1, \frac{\pi_j}{1 + S_j^-}\right\}. \tag{8}$$

The denominators remain positive because the class-j self-term has no error. These intervals are computed from the approximate probabilities and classwise score allowances; the loss bound below also uses the target identity.

Proposition 2 (Objective-conditioned contract)**.** *The following quantities bound loss error, probability TV, and feature-gradient error, respectively:*

$$B_L = \max\{\log(1 + S_y^+), -\log(1 + S_y^-)\}, \tag{9}$$

$$B_P = \min\left\{\tanh(d/2), \frac{1}{2}\sum_j \max(\pi_j - \ell_j, u_j - \pi_j)\right\}, \tag{10}$$

$$B_g = \frac{2B_P + \sum_{j \ne y} \pi_j a_j + (1 - \pi_y) a_y}{\tau}. \tag{11}$$

Each is no larger than the corresponding bound in Proposition 1.

Class-weight–error coupling in cross-entropy. For finite reference logits q, let $\widehat{q} = q + e$ and $P = \mathrm{softmax}(q)$. The signed loss change satisfies the exact identity

$$\Delta L := L_y(\widehat{q}) - L_y(q) = \log \sum_j P_j \exp(e_j) - e_y. \tag{12}$$

Proposition 3 (Class-weight–error coupling)**.** *Fix $t = P_y \in (0, 1)$ and $e_y \in \mathbb{R}$. Let $p_1 \le \cdots \le p_m$ be positive non-target probabilities summing to $1 - t$, and $a_1 \le \cdots \le a_m$ finite non-target score errors, with $m = C - 1$. Over all pairings of these two multisets, the signed loss change has attained extrema*

$$D_- = \log\left[t + \sum_{i=1}^m p_i \exp(a_{m+1-i} - e_y)\right], \tag{13}$$

$$D_+ = \log\left[t + \sum_{i=1}^m p_i \exp(a_i - e_y)\right]. \tag{14}$$

Consequently $\max_\sigma |\Delta L_\sigma| = \max\{|D_-|, |D_+|\}$.

These extrema follow from the classical rearrangement inequality: matching larger probabilities with larger signed errors maximizes the signed loss change, while reverse matching minimizes it. The smallest absolute error can occur at an intermediate pairing. The result applies to any softmax cross-entropy objective; Appendix F.5 gives the proof. Gradient-error norms also depend on class directions.

To see the gradient mechanism, put $h_j = R_\nu(r_j)v_j/r_j$ and $\widehat{h}_j = A(r_j)v_j/r_j$, using zero at $r_j = 0$. Then

$$\tau(\nabla_f \widehat{L}_y - \nabla_f L_y) = \sum_j (\pi_j - P_j)h_j + \sum_j (\pi_j - \mathbf{1}_{j=y})(\widehat{h}_j - h_j). \tag{15}$$

The first term captures probability redistribution; the second weights each primitive derivative error by the current objective, with weights π_j for non-target classes and $1 - \pi_y$ for the target. Each error is weighted according to its role in the current objective. Changing the correspondence between probability weights and classwise errors can therefore change loss and gradient errors while preserving target probability and both multisets. For gradients, class directions also determine how these weighted errors combine. Appendix F.4 gives the proof and a linear-in-C implementation.

Primitive accuracy and learning adequacy. Common-shift cancellation illustrates why ranking primitive errors can give a different ordering from ranking learning errors. Consider two paired perturbations of the exact potential in a two-class example:

A: ratio error ε; score shifts $(\varepsilon/2, 5\varepsilon/6)$; CE changes.
B: ratio error 2ε; score shifts $(2\varepsilon, 2\varepsilon)$; CE is unchanged.

The larger error cancels as a common score shift, whereas the smaller, nonuniform error changes the loss. Appendix B.5 gives the paired construction and a budget separating their adequacy.

4 From Learning Budgets to State-Conditioned Tolerance

4.1 Inverting a fixed-state contract

Let $g_j = |r_j - k_j|$ and consider a scalar primitive allowance $b \geq 0$ that is valid along every endpoint integration path. At a fixed state, set

$$d_j(b) = bg_j, \qquad a_j(b) = b. \tag{16}$$

Fix the numerical realization $\mathcal{N}$, the propagation certificate $\mathcal{B}$, and the allowance family $\mathcal{E}_s(b)$ in Eq. (16). Keep the approximate scores and probabilities produced by $\mathcal{N}$ fixed while varying this allowance. Substitution into Proposition 2 gives $B_L(s, b)$, $B_P(s, b)$, and $B_g(s, b)$.

For any selected set of downstream quantities $\mathcal{G} \subseteq \{L, P, g\}$ and positive budgets $\boldsymbol{\epsilon}$, define

$$b_{\star,\mathcal{N},\mathcal{B},\mathcal{E}}(s; \boldsymbol{\epsilon}) = \sup\{b \geq 0 : B_G(s, b) \leq \epsilon_G \text{ for all } G \in \mathcal{G}\}. \tag{17}$$

With $\mathcal{N}$, $\mathcal{B}$, and $\mathcal{E}$ fixed, we abbreviate this threshold as $b_\star(s; \boldsymbol{\epsilon})$. It is the largest uniform primitive allowance accepted by this certificate for the current numerical outputs, providing post-evaluation certification of the chosen realization. A valid primitive envelope $b \leq b_\star$ certifies the requested learning quantities. The primitive allowance must hold over the full endpoint integration paths.

Proposition 4 (Monotone inversion)**.** *At a finite fixed state, the functions $B_L(s, b)$, $B_P(s, b)$, and $B_g(s, b)$ are continuous and nondecreasing on $b \geq 0$, and vanish at zero. Their simultaneous accepted set is an interval starting at zero. Any finite threshold can therefore be found by bracketed scalar inversion without access to the exact learning outputs.*

For prediction stability, add $2b \max_j g_j < \widehat{m}$. The strict inequality specifies the open acceptance boundary of the margin test. Probabilities and losses can be computed in the log domain during inversion. Appendix B proves the monotonicity and gives the numerical protocol.

4.2 First-order state sensitivity

The first-order expansion describes local behavior as $b \to 0$ at a fixed state. When probabilities are highly saturated and local sensitivity is close to zero, higher-order terms can substantially affect the

inverse at finite budgets. We use the full nonlinear bound for all reported tolerances. Define

$$c_L(s) = \sum_{j \neq y} \pi_j (g_j + g_y), \tag{18}$$

$$c_P(s) = \min\left\{ \frac{\max_j g_j}{2}, \sum_j \pi_j (1 - \pi_j) g_j \right\}, \tag{19}$$

$$c_g(s) = \frac{2c_P(s) + 2(1 - \pi_y)}{\tau}. \tag{20}$$

Proposition 5 (First-order sensitivity)**.** *At a fixed finite state, as $b \downarrow 0$,*

$$B_L(s,b) = c_L(s)b + O(b^2), \quad B_P(s,b) = c_P(s)b + O(b^2), \quad B_g(s,b) = c_g(s)b + O(b^2). \tag{21}$$

Thus, in the small-budget regime, the joint loss–gradient threshold is approximated by

$$b_\star(s; \epsilon_L, \epsilon_g) \approx \min\left\{ \frac{\epsilon_L}{c_L(s)}, \frac{\epsilon_g}{c_g(s)} \right\}, \tag{22}$$

with an inactive zero-sensitivity constraint omitted.

Target probability enters the gradient sensitivity through $1 - \pi_y$, while probability TV retains the full distribution and endpoint geometry. Temperature appears explicitly through the differentiation scale. Feature norm constrains the endpoint geometry through $g_j \leq \|f\|/\tau$. The top-two margin controls prediction stability, whereas target-loss sensitivity also depends on the identity of the dominant class.

These coefficients separate two ways in which a class can matter. In c_L, an endpoint gap is weighted by the probability assigned to a competing class. A large gap for a nearly inactive competitor contributes little, while the same gap matters more when that competitor carries appreciable probability. The target gap is multiplied by the total non-target mass. Thus high target confidence can relax the required accuracy without changing the primitive geometry.

The gradient coefficient also contains the probability-redistribution term c_P. Two states with the same target probability can distribute the remaining mass differently across endpoint gaps, and therefore have different gradient sensitivities. The explicit factor $1/\tau$ converts this effect into the chosen feature coordinate. Changing temperature with f fixed also changes geometry and probabilities; the clean coordinate control instead holds f/τ fixed. Taken together, these terms explain why target confidence, temperature, or feature norm alone cannot determine numerical tolerance. The inverse contract combines their effects under the requested learning budget.

5 Experiments

We sample 5,120 query-level learning states from feature records at existing training checkpoints and saved training batches, retaining the complete class statistics and numerical geometry associated with each query. High-precision reference calculations and the full reproducibility protocol are described in Appendices B.4 and I.2.

5.1 Different learning consequences of the same primitive error

We first separate two mechanisms by controlled intervention. For coordinate conditioning, keep $u = f/\tau$, class statistics, and biases fixed while setting $f_\tau = \tau u$. All scalar vMF inputs and numerical errors remain identical. The feature-gradient discrepancy nevertheless scales as $1/\tau$, crossing a fixed absolute gradient budget (Fig. 2). This change follows directly from chain-rule scaling in the chosen feature coordinate: absolute gradient error varies with coordinate scale while the underlying numerical problem remains fixed.

For objective conditioning, add the same target-class bias to exact and approximate scores, keeping primitive geometry, temperature, the score-error vector, and primitive allowances fixed. Changing target probability changes how the loss weights the same numerical perturbation. The objective-conditioned bound follows this change, whereas the geometry-only bound remains fixed. The two controls isolate distinct mechanisms: coordinate scaling changes the units of an absolute gradient budget, while objective conditioning changes the weights applied to fixed classwise errors. Figure 2 compares their effects; Appendix C defines the interventions.

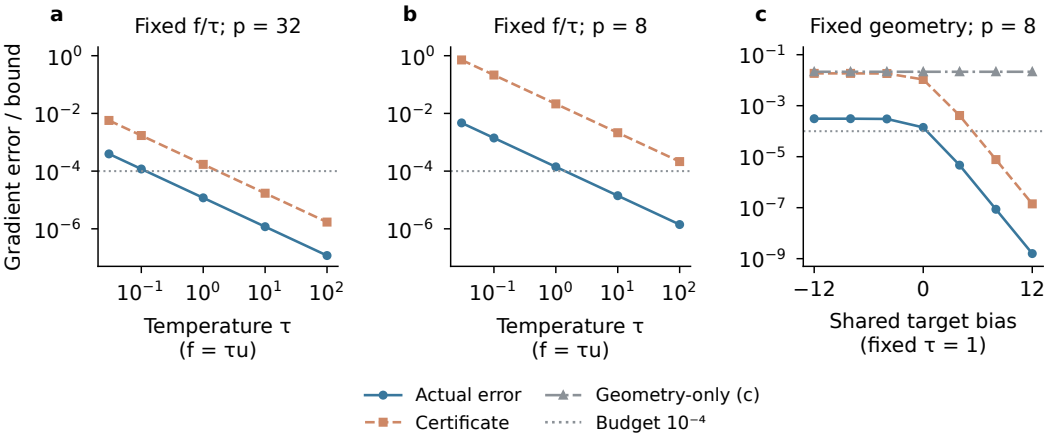

Figure 2: Learning consequences change while primitive error stays fixed. **a,b,** At fixed f/τ, changing temperature rescales feature-gradient error without changing the scalar numerical problem. **c,** A shared target-class bias changes objective weights at fixed geometry and temperature. Blue shows actual error; orange shows the geometry-only certificate in a,b and the objective-conditioned certificate in c. The gray curve in c is the unchanged geometry-only bound; dotted lines mark a common gradient budget.

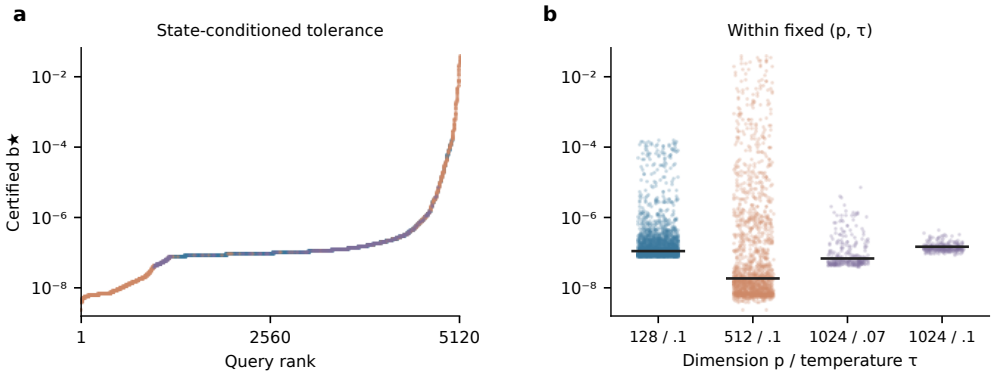

Figure 3: Query-level tolerance varies under a common learning-error budget. **a,** Ordered joint tolerances at budget 10^{-5}. **b,** The same queries grouped by dimension and temperature; black bars denote medians. Colors indicate dimension: blue 128, orange 512, and purple 1024.

5.2 Tolerance distributions in saved states

We compare numerical requirements across saved queries under common absolute loss and feature-gradient error budgets. At the joint budget 10^{-5}, the central 90% of certified primitive tolerances spans 3.29 orders of magnitude and the full range spans 7.16. At the looser budget 10^{-4}, the corresponding spans are 3.29 and 6.24 orders of magnitude. Substantial variation remains within the fixed $p = 512$, $\tau = 0.1$ group. These descriptive query-level distributions show heterogeneity at both budgets in the saved collection; Figure 3 displays the distributions and Table 1 reports their quantiles.

Under a prespecified paired perturbation family, detected actual gradient-response thresholds at budget 10^{-5} also span multiple orders of magnitude: 6.35 overall and 5.70 within the fixed $p = 512$, $\tau = 0.1$ group. The sufficient certificate is typically more conservative along this fixed perturbation family; Appendix I.3 describes the response check.

Table 1: Distribution of certified primitive tolerance under loss and joint budgets. Columns report the 5th percentile, median, 95th percentile, and full logarithmic span $\log_{10}(b_{\max}/b_{\min})$.

Budget	Quantity	5%	Median	95%	Full span
10^{-4}	Loss	4.86×10^{-7}	9.58×10^{-6}	2.40×10^{-3}	5.74
10^{-4}	Joint	7.35×10^{-8}	1.06×10^{-6}	1.44×10^{-4}	6.24
10^{-5}	Loss	4.86×10^{-8}	9.58×10^{-7}	2.86×10^{-4}	6.65
10^{-5}	Joint	7.35×10^{-9}	1.06×10^{-7}	1.44×10^{-5}	7.16

5.3 How do error structure and objective weights jointly determine tolerance?

The first-order expansion in Eqs. (18)–(22) characterizes local structure as $b \to 0$. At the joint budget 10^{-5}, the first-order prediction differs from the nonlinear tolerance by at most 1% for 4,989 of the 5,120 queries (97.44%), where relative deviation is measured against the nonlinear tolerance. In highly saturated probability states, however, local sensitivity approaches zero and higher-order terms can become important before the linear inverse reaches the prescribed budget. Figure 4a compares the first-order and nonlinear inverses, with an inset around agreement. All reported tolerances are obtained from the full nonlinear inversion.

The first-order expansion shows that sensitivity depends on the joint action of probability weights, endpoint geometry, and differentiation scale. To directly test the role of classwise probability–error correspondence, we construct controlled probability distributions on the saved query geometries, set the target probability to 0.5, and apply five prespecified rearrangements of the non-target probabilities. Across rearrangements, the probability multiset, entropy, exact top-two margin, and underlying numerical errors remain fixed; only the assignment of probability weights to classes changes. Shared additive biases in the exact and approximate logits realize these assignments without changing the classwise numerical errors.

Actual loss and gradient errors vary with the class assignment at fixed aggregate statistics (Fig. 4b), as predicted by Proposition 3 and the gradient decomposition in Eq. (15). Figure 4c places these changes on the budget scale: all five assignments satisfy joint budgets of 10^{-4} and 10^{-5}, whereas 1,001 queries change adequacy decisions across assignments at 10^{-7}. Classwise correspondence thus affects actual learning errors, with its impact on adequacy depending on the required error scale. Appendix C.1 gives the rearrangement definitions and complete budget results.

6 Discussion and Scope

The framework connects primitive error to the accuracy of specified learning quantities. Forward propagation gives finite-error guarantees, and inversion turns a prescribed budget into a certified allowance at the current state. The vMF instance makes this chain explicit through a paired special-function realization. Appendix A relates the analysis to numerical precision, inexact optimization, and probabilistic representation learning.

What the inverse guarantee means. The inverted threshold provides a state-dependent sufficient guarantee for the current numerical outputs, chosen certificate, and primitive-error family. Measured response thresholds also vary substantially across states and are typically larger than the certified allowances along the fixed perturbation family. Classwise correlations and cancellation can keep actual errors within budget even when the certificate is inconclusive. Each evaluator is assessed through its own primitive bounds and approximate probabilities.

Specifying learning-error budgets. We treat learning-level error budgets as external specifications determined by the application. A loss-error budget may reflect a numerical stability requirement, while a gradient-error budget may be set relative to the current gradient norm, the optimization step size, or an allowed update perturbation. Our analysis addresses propagation and inversion for a given budget; deriving such budgets automatically from long-term optimization or task-level requirements remains a direction for future work.

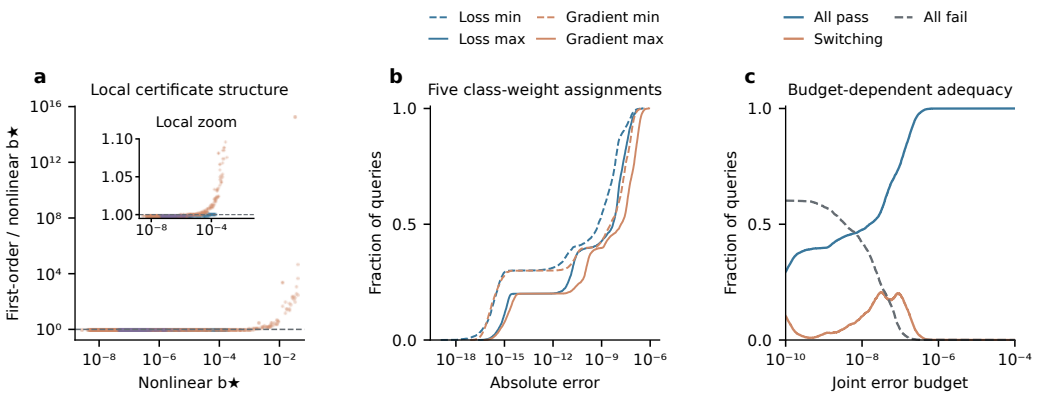

Figure 4: Local sensitivity and classwise correspondence shape numerical requirements. **a,** Ratio of first-order to nonlinear tolerance at budget 10^{-5}; the inset magnifies the region near one. Colors denote dimension as in Fig. 3. **b,** Cumulative distributions of minimum and maximum absolute errors over five fixed assignments of non-target probabilities. Target probability, entropy, exact top-two margin, geometry, and primitive errors are held fixed. **c,** Fractions of queries for which all assignments satisfy the joint budget, assignments yield different decisions, or all assignments fail.

From feature gradients to parameter gradients. For the same differentiable encoder $f = h_\theta(x)$ at a fixed input and parameter state, let $J_\theta = \partial h_\theta(x)/\partial\theta$. When the numerical discrepancy enters through the loss as a function of f, the shared encoder Jacobian propagates the feature-gradient error to parameter space:

$$\nabla_\theta L = J_\theta^\top \nabla_f L, \qquad \|\Delta\nabla_\theta L\|_2 \le \|J_\theta\|_{2\to 2}\, \|\Delta\nabla_f L\|_2.$$

Thus the feature-gradient certificate controls the error passed to the encoder, with parameter-gradient sensitivity determined by its Jacobian at the current state.

Mathematical and execution scope. The theory characterizes numerical quantities at fixed learning states in specified differentiation coordinates. Propagating these guarantees to long-term optimization trajectories and task-level performance is a direction for further study. The certificates control the mathematical error of the analytic surrogate. Full floating-point execution guarantees would additionally require explicit allowances for rounding, endpoint construction, transcendental evaluation, and numerical reductions. High-precision comparisons assess the implementation over the tested range.

State-conditioned certificates also support coarse-to-fine analysis: an inexpensive bound is evaluated first, followed by a tighter bound when needed to resolve adequacy. This procedure gives the same decisions as always evaluating the tighter bound (Appendix B.3).

Beyond the vMF instance. The paired vMF primitive supplies one explicit setting in which both primitive allowances and their learning-level propagation are available. Applying the interface to a different numerical primitive requires its own valid allowances and propagation analysis. More broadly, the framework connects requirements on underlying numerical errors to the quantities that a learning objective needs to protect.

Reproducibility Statement

The appendix gives the primitive and learning-contract proofs, tolerance inversion, controlled interventions, and shared-geometry reference protocol. The source materials provide the saved-query data, numerical settings, and verification records. The analysis covers 5,120 queries, with queries from shared training runs treated as dependent observations. Mathematical guarantees and floating-point evaluation are distinguished throughout.

AI USE STATEMENT

AI tools were used to assist with manuscript language polishing, code drafting, mathematical derivations, and figure design. The authors take full responsibility for all content.

REFERENCES

D. E. Amos. Computation of modified Bessel functions and their ratios. *Mathematics of Computation*, 28(125):239–251, 1974. ISSN 0025-5718. doi: 10.1090/s0025-5718-1974-0333287-7.

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6(46): 1345–1382, 2005.

Atılım Güneş Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–43, 2018.

Roland Becker and Rolf Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica*, 10:1–102, 2001. ISSN 1474-0508. doi: 10.1017/s0962492901000010.

Dimitri P. Bertsekas and John N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000. ISSN 1095-7189. doi: 10.1137/s1052623497331063.

Pierre Blanchard, Desmond J Higham, and Nicholas J Higham. Accurately computing the log-sum-exp and softmax functions. *IMA Journal of Numerical Analysis*, 41(4):2311–2330, 2021. ISSN 1464-3642. doi: 10.1093/imanum/draa038.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. ISBN 9780511804441. doi: 10.1017/cbo9780511804441.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009. doi: 10.1109/cvpr.2009.5206848.

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2):37–75, 2014. ISSN 1436-4646. doi: 10.1007/s10107-013-0677-5.

Chaoqun Du, Yulin Wang, Shiji Song, and Gao Huang. Probabilistic contrastive learning for long-tailed visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9): 5890–5904, 2024. ISSN 1939-3539. doi: 10.1109/tpami.2024.3369102.

Amparo Gil, Javier Segura, and Nico M. Temme. *Numerical Methods for Special Functions*. Society for Industrial and Applied Mathematics, 2007. ISBN 9780898717822. doi: 10.1137/1.9780898717822.

Michael B. Giles and Endre Süli. Adjoint methods for PDEs: a posteriori error analysis and postprocessing by duality. *Acta Numerica*, 11:145–236, 2002. ISSN 1474-0508. doi: 10.1017/s096249290200003x.

David Goldberg. What every computer scientist should know about floating-point arithmetic. *ACM Computing Surveys*, 23(1):5–48, 1991. ISSN 1557-7341. doi: 10.1145/103162.103163.

Andreas Griewank and Andrea Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation, Second Edition*. Society for Industrial and Applied Mathematics, 2008. ISBN 9780898717761. doi: 10.1137/1.9780898717761.

Yina He, Lei Peng, Yongcun Zhang, Juanjuan Weng, Shaozi Li, and Zhiming Luo. Long-tailed out-of-distribution detection: Prioritizing attention to tail. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(3):3446–3454, 2025. ISSN 2159-5399. doi: 10.1609/aaai.v39i3.32357.

Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms: Second Edition*. Society for Industrial and Applied Mathematics, 2002. ISBN 9780898718027. doi: 10.1137/1.9780898718027.

Kurt Hornik and Bettina Grün. Amos-type bounds for modified Bessel function ratios. *Journal of Mathematical Analysis and Applications*, 408(1):91–101, 2013. ISSN 0022-247X. doi: 10.1016/j.jmaa.2013.05.070.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. Wiley, 1999. ISBN 9780470316979. doi: 10.1002/9780470316979.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations*, 2018.

F. W. J. Olver. The asymptotic expansion of Bessel functions of large order. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 247(930): 328–368, 1954. ISSN 2054-0272. doi: 10.1098/rsta.1954.0021.

John R. Rice. A theory of condition. *SIAM Journal on Numerical Analysis*, 3(2):287–310, 1966. ISSN 1095-7170. doi: 10.1137/0703023.

Diego Ruiz-Antolín and Javier Segura. A new type of sharp bounds for ratios of modified Bessel functions. *Journal of Mathematical Analysis and Applications*, 443(2):1232–1246, 2016. ISSN 0022-247X. doi: 10.1016/j.jmaa.2016.06.011.

Tyler R. Scott, Andrew C. Gallagher, and Michael C. Mozer. von Mises-Fisher loss: An exploration of embedding geometries for supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10612–10622, October 2021.

Javier Segura. Simple bounds with best possible accuracy for ratios of modified Bessel functions. *Journal of Mathematical Analysis and Applications*, 526(1):127211, 2023. ISSN 0022-247X. doi: 10.1016/j.jmaa.2023.127211.

Suvrit Sra. A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of $I_s(x)$. *Computational Statistics*, 27(1):177–190, 2012. ISSN 1613-9658. doi: 10.1007/s00180-011-0232-x.

Nico M. Temme. *Special Functions: An Introduction to the Classical Functions of Mathematical Physics*. Wiley, 1996. ISBN 9781118032572. doi: 10.1002/9781118032572.

The mpmath development team. *mpmath: a Python library for arbitrary-precision floating-point arithmetic*, 2026.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8769–8778, June 2018.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008. ISSN 1935-8245. doi: 10.1561/2200000001.

How Accurate Is Accurate Enough?

Supplementary Material

CONTENTS

A Related Work . . . 13

B Tolerance Inversion, Sensitivity, and Refinement . . . 13

B.1 Monotonicity of the inverse contract . . . 13

B.2 Derivation of the first-order coefficients . . . 13

B.3 Decision equivalence under certificate refinement . . . 14

B.4 Saved states and inversion protocol . . . 14

B.5 A paired counterexample to primitive-only ranking . . . 14

C Conditioning Controls and Class-Weight Rearrangement . . . 14

C.1 Class-weight rearrangement at fixed target probability . . . 15

D The Paired vMF Instance . . . 16

E Mathematical Foundation of the Analytic Primitive . . . 16

E.1 Riccati identity and a preliminary ratio bound . . . 16

E.2 Refined uniform certificate . . . 18

F Proofs of the Learning-Level Contracts . . . 19

F.1 Scores and cross-entropy . . . 19

F.2 Probability TV and prediction stability . . . 19

F.3 Feature-gradient propagation . . . 20

F.4 Objective-conditioned intervals . . . 20

F.5 Proof of class-weight–error coupling . . . 20

G A Second vMF Objective . . . 21

H Numerical Realization and Certificate Construction . . . 21

H.1 State-dependent residual envelope . . . 21

H.2 Score certificates from pointwise envelopes . . . 22

H.3 Stable endpoint realization . . . 23

I Reference Evaluation and Numerical Scope . . . 23

I.1 Saved-state provenance . . . 23

I.2 High-precision reference at shared geometry . . . 24

I.3 Supplementary actual-response check under a fixed perturbation family . . . 24

I.4 Stable realization and the floating-point boundary . . . 24

A Related Work

Numerical precision in learning. Mixed-precision training uses lower-precision arithmetic while preserving training performance through techniques such as loss scaling and higher-precision weight updates (Micikevicius et al., 2018). Special-function algorithms provide tools for accurate evaluation across parameter regimes (Gil et al., 2007), while rounding-error analysis of softmax and log-sum-exp explains how equivalent formulas differ in floating-point stability (Blanchard et al., 2021). These perspectives motivate careful numerical realization inside learning systems. We study the complementary question of how much primitive error a particular learning state can tolerate under a specified downstream requirement, with tolerance depending on the objective weights, geometry, and differentiation scale.

Inexact optimization and goal-oriented error analysis. Gradient methods with errors and inexact-oracle methods relate controlled function or gradient inaccuracies to optimization behavior (Bertsekas & Tsitsiklis, 2000; Devolder et al., 2014). Goal-oriented error estimation uses sensitivity and duality to relate local errors to a selected output and guide adaptive computation (Becker & Rannacher, 2001; Giles & Süli, 2002). Our contribution lies within this broader objective-dependent view: for a numerical primitive embedded in a probabilistic learning objective, we construct computable finite-error guarantees for losses, probabilities, predictions, and feature gradients, and invert them to obtain a sufficient primitive tolerance at the current state. This fixed-state analysis provides primitive-derived error bounds of the kind that inexact optimization takes as inputs; propagating them along an optimization trajectory requires further assumptions.

von Mises–Fisher primitives in probabilistic representation learning. ProCo models class-conditional features with von Mises–Fisher distributions for probabilistic contrastive learning (Du et al., 2024); probabilistic vMF embeddings also support supervised classification and retrieval (Scott et al., 2021). Their normalizing constants involve modified Bessel functions, and differentiation introduces Bessel ratios, connecting probabilistic scores and gradients to special-function evaluation. High-dimensional vMF learning provides an explicit instance in which primitive approximation errors can be rigorously bounded and propagated to the learning quantities of interest.

B Tolerance Inversion, Sensitivity, and Refinement

B.1 Monotonicity of the inverse contract

Fix finite approximate logits and hence $0 < \pi_j < 1$, endpoint gaps $g_j \geq 0$, a target, and $\tau > 0$. With $d_j = bg_j$, each S_j^+ is continuous and nondecreasing in b, while S_j^- is continuous and nonincreasing. Consequently $\log(1 + S_y^+)$ and $-\log(1 + S_y^-)$ are nondecreasing. The lower probability endpoint decreases and the upper endpoint increases; clipping the latter at one preserves validity and monotonicity. Both candidate TV bounds are nondecreasing, so their minimum is nondecreasing as well. Finally, the ratio contribution to B_g is $2b(1 - \pi_y)/\tau$. These facts prove Proposition 4. All bounds vanish at zero. Their sublevel sets are closed intervals starting at zero, possibly unbounded if the corresponding constraint is inactive. Intersecting the selected sublevel sets gives Eq. (17). The strict prediction-margin condition has an open upper boundary.

B.2 Derivation of the first-order coefficients

Let $t_j = \sum_{i \neq j} \pi_i(g_i + g_j)$. Taylor expansion gives $S_j^{\pm} = \pm bt_j + O(b^2)$, and hence

$$\log(1 + S_y^+) = bt_y + O(b^2), \qquad -\log(1 + S_y^-) = bt_y + O(b^2).$$

This yields $c_L = t_y$. For sufficiently small b at a fixed finite state, the probability clipping is inactive, and both coordinate deviations have expansion $b\pi_j t_j + O(b^2)$. Their half-sum has coefficient

$$\frac{1}{2} \sum_j \pi_j t_j = \sum_j \pi_j(1 - \pi_j)g_j.$$

The other TV bound, $\tanh(b \max_j g_j/2)$, has coefficient $\max_j g_j/2$. Taking the minimum gives c_P, including the case where the coefficients coincide. The unscaled ratio contribution is exactly

$2b(1-\pi_y)$, yielding c_g. These expansions prove Proposition 5. A positive first-order coefficient gives an inverse threshold $\epsilon/c + O(\epsilon^2)$ as its budget decreases; simultaneous small budgets select the active constraint. When a coefficient vanishes, the full nonlinear bound determines whether higher-order terms constrain the tolerance or the constraint remains inactive.

B.3 Decision equivalence under certificate refinement

For the same numerical realization, let $\mathbf{B}^{\mathrm{cheap}}$ be an inexpensive certificate and $\mathbf{B}^{\mathrm{tight}}$ a tighter certificate, satisfying

$$\mathbf{E} \le \mathbf{B}^{\mathrm{tight}} \le \mathbf{B}^{\mathrm{cheap}}.$$

The cascade first tests $\mathbf{B}^{\mathrm{cheap}} \le \boldsymbol{\epsilon}$ and computes $\mathbf{B}^{\mathrm{tight}}$ only when the first test is inconclusive. Both stages certify the same evaluator outputs through the same downstream propagation, using successively tighter primitive envelopes.

The ordering above ensures that acceptance by the inexpensive bound implies acceptance by the tighter bound. Otherwise, the cascade evaluates the tighter bound directly. The cascade and the policy that always evaluates the tighter bound therefore return the same decision at every state.

B.4 Saved states and inversion protocol

The saved-query analysis uses the same complete class banks and numerical geometry as the reference comparisons in Appendix I. At fixed approximate probabilities, we set $d_j = b|r_j - k_j|$ and $a_j = b$ and invert the loss and gradient certificates by bracketed bisection. Stable log-domain reductions evaluate the probability intervals over wide allowance ranges. The first-order prediction uses the same states and budgets without fitted coefficients.

The nonlinear inverse remains well defined when saturated probabilities make the first-order coefficients small. In this regime, higher-order terms can restrict the accepted allowance far earlier than a linear extrapolation predicts. Figure 4a shows both the agreement of the inverses for most queries and their large departures in highly saturated states.

B.5 A paired counterexample to primitive-only ranking

Take $p = 8$, $\tau = 1$, $f = e_1$, two class directions equal to e_1, and concentrations $(1, 2)$, so the query radii are $(2, 3)$. Let the target be the first class and $\varepsilon = 10^{-4}$. Define paired potentials using the exact potential Φ:

$$F^{\mathrm{A}}(x) = \Phi(x) + \varepsilon H(x), \qquad F^{\mathrm{B}}(x) = \Phi(x) + 2\varepsilon G(x),$$

where

$$H(x) = \begin{cases} x^2/6, & 0 \le x \le 3, \\ x - 3/2, & x \ge 3, \end{cases} \qquad G(x) = \begin{cases} x^2/2, & 0 \le x \le 1, \\ x - 1/2, & x \ge 1. \end{cases}$$

Both potentials are continuously differentiable and use their derivatives as ratio realizations. Their derivative errors vanish at zero and attain global maxima ε and 2ε. At the query, score errors are $(\varepsilon/2, 5\varepsilon/6)$ for A and $(2\varepsilon, 2\varepsilon)$ for B. B leaves cross-entropy exactly unchanged by common-shift invariance; A has loss error approximately 1.7568×10^{-5}. Thus B meets a 10^{-5} loss budget while A does not, despite A having a smaller maximum primitive error.

Common-shift cancellation thus allows a larger primitive error to produce a smaller loss error.

C Conditioning Controls and Class-Weight Rearrangement

The coordinate controls in Figure 2a,b fix $u = f/\tau$, class directions, concentrations, and biases, then set $f_\tau = \tau u$. Every scalar vMF input and error is unchanged within each dimension. The feature-gradient error and its certificate scale as $1/\tau$, so the control isolates the effect of the differentiation coordinate on an absolute budget.

The objective control in Figure 2c adds the same target-class bias to exact and approximate logits. Geometry, temperature, score errors, ratio errors, and primitive allowances remain fixed. Changing only the softmax weights reduces the discrepancy and the objective-conditioned certificate as target probability approaches one; the geometry-only bound remains fixed.

C.1 Class-weight rearrangement at fixed target probability

For each of the 5,120 saved states, let P be its exact reference probability vector, y its target, and $j_1, \ldots, j_{C-1}$ its non-target classes in their stored order. Define $w_a = P_{j_a}/(1 - P_y)$. Each prescribed permutation σ produces

$$P_y^{(\sigma)} = \tfrac{1}{2}, \qquad P_{j_a}^{(\sigma)} = \tfrac{1}{2} w_{\sigma(a)}.$$

The target probability 0.5 provides a common nonsaturated control condition. The classwise weighting mechanism in Eq. (15) applies more generally across target probabilities.

The five permutations are identity, a cyclic shift by one position, a cyclic shift by $\lfloor (C-1)/2 \rfloor$ positions, reversal, and a pseudorandom permutation generated with seed 3407. They are fixed before comparing downstream errors.

A query-specific additive class bias realizes each vector: if q_j is the original exact logit, add $\beta_j^{(\sigma)} = \log P_j^{(\sigma)} - q_j$ to both exact and approximate logits. With $e_j = \widehat{q}_j - q_j$, their new probabilities are $P^{(\sigma)}$ and $\pi_j^{(\sigma)} \propto P_j^{(\sigma)} \exp(e_j)$. Treating the added biases as constants during feature differentiation preserves the exact and approximate class-gradient vectors. Thus the intervention changes only the correspondence between probability weights and the fixed classwise errors and directions. Dimension, temperature, feature norm, endpoints, class directions, score errors, ratio errors, and primitive envelopes remain unchanged.

The non-target probability multiset, reference entropy, and exact top-two logit margin $\log[(1/2)/\max_{j \neq y} P_j^{(\sigma)}]$ are invariant across permutations. These invariants refer to the exact distribution; the approximate target probability is allowed to respond to the unchanged classwise numerical errors. This controlled modification isolates the weighting mechanism at each saved state.

Table 2 compares relative variation with absolute error ranges. Large fold changes can arise from very small minimum errors, whereas Table 3 measures the consequence for actual loss and gradient budgets. Together, the tables distinguish a change in learning error from a change in whether that error meets a specified requirement.

Table 2: Class assignments change learning errors at fixed probability summaries. Fold change is the largest error divided by the smallest error for each state; absolute range is their difference.

Summary	Loss	Feature gradient
Fold change: median	3.18	3.28
Fold change: maximum	1.06×10^{7}	1.16×10^{7}
Median absolute range	2.36×10^{-9}	9.09×10^{-9}
Maximum absolute range	1.78×10^{-7}	8.55×10^{-7}

Table 3: Actual numerical adequacy under class-weight rearrangement. Counts give the states for which the five assignments yield different adequacy decisions. The joint requirement applies the same budget to absolute loss error and Euclidean feature-gradient error.

Budget	Loss	Gradient	Joint
10^{-4}	0	0	0
10^{-5}	0	0	0
10^{-6}	0	0	0
10^{-7}	56	1014	1001
10^{-8}	1150	526	521
10^{-9}	571	148	138

The five assignments per query preserve the reference probability summaries while changing their correspondence with numerical errors. Their signed loss changes lie within the rearrangement extrema of Proposition 3. Reference evaluation uses the common geometry described in Appendix I.2.

D THE PAIRED VMF INSTANCE

Put $s(x) = \sqrt{\nu^2 + x^2}$. The realization used in the main experiments is

$$A_{2,\nu}(x) = \frac{x}{\nu + s(x)} - \frac{x}{2s(x)^2} + \frac{x(4\nu^2 - x^2)}{8s(x)^5}, \tag{23}$$

$$F_{2,\nu}(x) = s(x) - \nu \log(\nu + s(x)) - \frac{1}{2}\log s(x) + \frac{1}{8s(x)} - \frac{5\nu^2}{24s(x)^3}. \tag{24}$$

Direct differentiation gives $F'_{2,\nu} = A_{2,\nu}$ and $A_{2,\nu}(0) = 0$. Additive constants in the potential cancel from endpoint differences. The construction uses the modified-Bessel asymptotic structure (Olver, 1954; Temme, 1996); its role here is to provide a concrete paired realization with a provable allowance.

Theorem 6 (Uniform primitive allowance)**.** *For every $\nu \geq 10/9$ and $x \geq 0$,*

$$|R_\nu(x) - A_{2,\nu}(x)| \leq b_\nu, \qquad b_\nu = \frac{1}{4\nu^3}.$$

Appendix E proves this bound. Thus $a_j = b_\nu$ and $d_j = b_\nu|r_j - k_j|$ are valid inputs to the main-text contract.

Refined allowance interface. Let $m = 2\nu + 2$ and

$$D_\nu(t) = \nu^{-3}\left[(1 + (t/\nu)^2)^{-3/2} + \frac{1}{4\nu}(1 + (t/\nu)^2)^{-2}\right].$$

For $0 \leq \alpha \leq 1$, a valid pointwise envelope is

$$B_\alpha(x) = \frac{x}{m}\{D_\nu(0)\alpha^m + D_\nu(\alpha x)(1 - \alpha^m)\}.$$

Taking its minimum with b_ν, or with other valid choices of α, preserves validity. Appendix H.1 derives the envelope and Appendix H.2 gives valid interval bounds for score integration. These certificates all describe identical F_2/A_2 outputs. The stable endpoint formula is given in Appendix H.3.

Special-function context. Classical algorithms evaluate modified Bessel functions and their ratios through recurrence, asymptotics, and region-dependent numerical methods (Amos, 1974; Gil et al., 2007). Analytic ratio inequalities provide another source of usable allowances (Hornik & Grün, 2013; Ruiz-Antolín & Segura, 2016; Segura, 2023). Here these tools supply the primitive-level premise; the state-conditioned tolerance additionally depends on its propagation through the learning objective.

E MATHEMATICAL FOUNDATION OF THE ANALYTIC PRIMITIVE

This appendix proves the uniform ratio certificate for the analytic vMF primitive in Section 2. We first derive a preliminary ν^{-3} enclosure directly from the Riccati equation, then sharpen the barrier argument to obtain the factor-$1/4$ bound in Theorem 6. The proof supplies the uniform allowance used by the inexpensive certificate.

E.1 RICCATI IDENTITY AND A PRELIMINARY RATIO BOUND

Set

$$\varepsilon = \nu^{-1}, \qquad x = \nu z,$$

and define

$$H_\varepsilon(z) = \frac{I_{\nu+1}(\nu z)}{I_\nu(\nu z)}. \tag{25}$$

The adjacent-order modified-Bessel ratio satisfies the Riccati equation

$$\varepsilon H'_\varepsilon(z) = 1 - \frac{2 + \varepsilon}{z} H_\varepsilon(z) - H_\varepsilon(z)^2. \tag{26}$$

Let

$$S = \sqrt{1+z^2}$$

and define

$$\begin{aligned} r_0(z) &= \frac{z}{1+S}, \\ r_1(z) &= -\frac{z}{2S^2}, \\ r_2(z) &= \frac{z(4-z^2)}{8S^5}, \\ a_\varepsilon(z) &= r_0(z) + \varepsilon r_1(z) + \varepsilon^2 r_2(z). \end{aligned} \tag{27}$$

Substituting $x = \nu z$ shows that

$$a_\varepsilon(z) = A_{2,\nu}(x).$$

We evaluate how accurately a_ε satisfies the exact Riccati equation. Define the residual

$$\mathcal{E} = 1 - \frac{2+\varepsilon}{z} a_\varepsilon - a_\varepsilon^2 - \varepsilon a_\varepsilon'. \tag{28}$$

Direct simplification gives the exact identity

$$\begin{aligned} \mathcal{E} = & -\varepsilon^3 \frac{z^4 - 10z^2 + 4}{4(1+z^2)^{7/2}} \\ & -\varepsilon^4 \frac{z^2(z^2-4)^2}{64(1+z^2)^5}. \end{aligned} \tag{29}$$

For $u = z^2 \geq 0$,

$$|u^2 - 10u + 4| \leq 4(1+u)^2, \qquad u(u-4)^2 \leq 16(1+u)^3. \tag{30}$$

Applying these inequalities to Eq. (29) yields

$$|\mathcal{E}(z,\varepsilon)| \leq \varepsilon^3 \left(1 + \frac{\varepsilon}{4}\right). \tag{31}$$

Let

$$e(z) = H_\varepsilon(z) - a_\varepsilon(z).$$

Subtracting the candidate equation from Eq. (26) gives

$$\varepsilon e' = \mathcal{E} - Ke - e^2, \tag{32}$$

where

$$\begin{aligned} K &= \frac{2+\varepsilon}{z} + 2a_\varepsilon \\ &= \frac{2S}{z} + \frac{\varepsilon}{zS^2} + 2\varepsilon^2 r_2(z). \end{aligned} \tag{33}$$

When $0 < z \leq 2$, we have $r_2(z) \geq 0$. For $z \geq 2$,

$$-r_2(z) = \frac{z(z^2-4)}{8S^5} \leq \frac{1}{40}. \tag{34}$$

Therefore, whenever $0 < \varepsilon \leq 9/10$,

$$K \geq 2 - \frac{\varepsilon^2}{20} \geq \frac{3919}{2000}. \tag{35}$$

We first choose the constant barrier

$$b = \varepsilon^3.$$

At the upper boundary $e = b$, Eqs. (31) and (35) imply

$$\mathcal{E} - Kb - b^2 < 0.$$

Thus the vector field in Eq. (32) points inward at $e = b$.

At the lower boundary $e = -b$, the right-hand side of Eq. (32) is bounded below by

$$\varepsilon^3 \left(-\frac{49}{40} + \frac{3919}{2000} - \frac{729}{1000} \right) = \frac{11}{2000}\varepsilon^3 > 0. \tag{36}$$

Hence the vector field also points inward at $e = -b$.

It remains to initialize the barrier argument at the singular endpoint. The small-z expansions are

$$H_\varepsilon(z) = \frac{z}{2(1+\varepsilon)} + O(z^3) \tag{37}$$

and

$$a_\varepsilon(z) = \frac{z}{2}\left(1 - \varepsilon + \varepsilon^2\right) + O(z^3). \tag{38}$$

Subtracting gives

$$e(z) = -\frac{\varepsilon^3}{2(1+\varepsilon)} z + O(z^3). \tag{39}$$

Thus e begins strictly inside $[-b, b]$ for sufficiently small positive z.

If a first contact with the upper barrier existed, it would require $e' \geq 0$ at that point, contradicting the strict inward direction proved above. A first contact with the lower barrier would similarly require $e' \leq 0$ and contradict the opposite inward direction. Consequently,

$$|e(z)| < \varepsilon^3, \qquad z > 0. \tag{40}$$

Both H_ε and a_ε extend continuously to $z = 0$ with value zero, so the corresponding non-strict bound holds at the origin. Since

$$\varepsilon \leq \frac{9}{10} \qquad \Longleftrightarrow \qquad \nu \geq \frac{10}{9},$$

we obtain the preliminary all-axis enclosure

$$|R_\nu(x) - A_{2,\nu}(x)| \leq \nu^{-3}.$$

The next subsection sharpens this constant to the factor $1/4$ stated in Theorem 6.

E.2 Refined uniform certificate

The preliminary proof replaces the concentration-dependent residual by a single uniform upper bound. Retaining its denominators yields a sharper comparison. With the notation above,

$$|\mathcal{E}| \leq \varepsilon^3 \left[S^{-3} + \frac{\varepsilon}{4} S^{-4} \right]. \tag{41}$$

Moreover,

$$K \geq \kappa \frac{2S}{z}, \qquad \kappa = \frac{3919}{4000}. \tag{42}$$

Indeed, Eq. (33) and Eq. (34) give

$$K \geq \frac{2S}{z} - \frac{\varepsilon^2}{20},$$

while

$$\frac{2S}{z} \geq 2 \qquad \text{and} \qquad \varepsilon \leq \frac{9}{10}.$$

The relevant rational functions satisfy

$$\sup_{z \geq 0} \frac{z}{2S^4} = \frac{9}{32\sqrt{3}}, \qquad \sup_{z \geq 0} \frac{z}{2S^5} = \frac{8}{25\sqrt{5}}. \tag{43}$$

Combining Eqs. (41)–(43) gives

$$\frac{|\mathcal{E}|}{K\varepsilon^3} \leq \kappa^{-1} \left[\frac{9}{32\sqrt{3}} + \frac{9}{40} \frac{8}{25\sqrt{5}} \right] < \frac{1}{5}. \tag{44}$$

The final strict inequality already follows from the rational lower bounds

$$\sqrt{3} > \frac{1732}{1000}, \qquad \sqrt{5} > \frac{2236}{1000}.$$

We now choose the refined constant barrier

$$B = \frac{\varepsilon^3}{4}. \tag{45}$$

At $e = B$, Eq. (44) gives

$$\mathcal{E} - KB - B^2 < 0.$$

At $e = -B$, the right-hand side of Eq. (32) is bounded below by

$$K\varepsilon^3 \left(\frac{1}{4} - \frac{1}{5} - \frac{\varepsilon^3}{16K} \right). \tag{46}$$

Furthermore,

$$\frac{\varepsilon^3}{16K} \le \frac{1458}{62704} < \frac{1}{20}, \tag{47}$$

so Eq. (46) is strictly positive.

The same small-z initialization from Eq. (39) and the same first-contact argument used in Appendix E.1 therefore yield

$$|H_\varepsilon(z) - a_\varepsilon(z)| \le \frac{\varepsilon^3}{4}, \qquad z \ge 0. \tag{48}$$

Returning to $x = \nu z$ and $\varepsilon = \nu^{-1}$ gives

$$|R_\nu(x) - A_{2,\nu}(x)| \le \frac{1}{4\nu^3}, \qquad \nu \ge \frac{10}{9}, \quad x \ge 0, \tag{49}$$

which proves Theorem 6.

F Proofs of the Learning-Level Contracts

F.1 Scores and cross-entropy

Pairing gives

$$(q_j - \widehat{q}_j) = \int_{k_j}^{r_j} (R_\nu(x) - A(x))\, dx,$$

so the integrated primitive envelope bounds its absolute value by d_j. Since log-sum-exp is coordinatewise increasing and invariant to a common shift up to that shift,

$$\mathrm{LSE}(\widehat{q}) - d \le \mathrm{LSE}(q) \le \mathrm{LSE}(\widehat{q}) + d.$$

Adding the target-score error proves $|L_y - \widehat{L}_y| \le d + d_y$.

F.2 Probability TV and prediction stability

Put $e_j = \widehat{q}_j - q_j$, so $|e_j| \le d$. Under the exact distribution P, let the positive random variable X take values $\exp(e_j)$. Then $X \in [a, b] = [e^{-d}, e^d]$, and

$$\mathrm{TV}(P, \pi) = \frac{\mathbb{E}_P |X - \mathbb{E}_P X|}{2\mathbb{E}_P X}.$$

For $m = \mathbb{E}_P X$, convexity bounds the numerator by the chord joining the endpoints:

$$\mathbb{E}_P |X - m| \le \frac{2(b - m)(m - a)}{b - a}.$$

Maximizing $(b - m)(m - a)/(m(b - a))$ over $m \in [a, b]$ gives

$$\mathrm{TV}(P, \pi) \le \frac{\sqrt{b} - \sqrt{a}}{\sqrt{b} + \sqrt{a}} = \tanh(d/2).$$

For $d = 0$ the distributions coincide. If $j_\star$ is the approximate winner, then $q_{j_\star} - q_j \ge \widehat{q}_{j_\star} - \widehat{q}_j - 2d$. An approximate margin greater than $2d$ therefore certifies the winner. Interchanging exact and approximate scores proves the exact-margin version.

F.3 Feature-gradient propagation

All class statistics, biases, and τ are held fixed. Define

$$h_j = R_\nu(r_j)v_j/r_j, \quad \widehat{h}_j = A(r_j)v_j/r_j, \quad \nabla_f q_j = h_j/\tau, \quad \nabla_f \widehat{q}_j = \widehat{h}_j/\tau,$$

with continuous zero extensions when $r_j = 0$. Because $0 \le R_\nu \le 1$, $\|h_j\| \le 1$ and $\|\widehat{h}_j - h_j\| \le a_j$. Subtracting the two cross-entropy gradients gives Eq. (15). The probability-change term has norm at most $2\,\mathrm{TV}(P, \pi)$; the primitive-error term has norm at most

$$\sum_{j \ne y} \pi_j a_j + (1 - \pi_y)a_y \le \max_j a_j + a_y.$$

Together with the TV bound, this proves Proposition 1.

F.4 Objective-conditioned intervals

The exact probability can be written in terms of approximate quantities as

$$P_j = \frac{\pi_j}{D_j(e)}, \qquad D_j(e) = \pi_j + \sum_{i \ne j} \pi_i \exp(e_j - e_i).$$

The self-term has no perturbation. Since $|e_j - e_i| \le d_j + d_i$,

$$0 < 1 + S_j^- \le D_j(e) \le 1 + S_j^+.$$

This proves the main-text probability intervals, with the upper endpoint optionally clipped at one. Also $L_y - \widehat{L}_y = \log D_y(e)$, so its absolute value is at most B_L. The two loss endpoints are attained over the independent error box by choosing opposite target and non-target signs. They need not be jointly realizable by a single paired primitive.

The coordinate intervals give

$$\mathrm{TV}(P, \pi) \le \frac{1}{2} \sum_j \max(\pi_j - \ell_j, u_j - \pi_j).$$

Combining this with the uniform TV bound proves B_P. Substituting B_P and the probability-weighted primitive-error term into the gradient decomposition proves B_g.

These bounds dominate the base contract: the loss denominator has logarithm in $[-d_y - d, d_y + d]$, the TV bound explicitly takes a minimum with the base bound, and the weighted ratio term is at most $\max_j a_j + a_y$. This proves Proposition 2. The coordinate intervals need not be jointly attainable; the resulting TV and gradient bounds are sufficient rather than sharp in general.

Computation. Let $m_i^\pm = \mathrm{expm1}(\pm d_i)$. Then

$$S_j^\pm = m_j^\pm(1 - \pi_j) + (1 + m_j^\pm)\left(\sum_i \pi_i m_i^\pm - \pi_j m_j^\pm\right).$$

This reduces the mathematical construction to linear rather than quadratic work in class count. Log-domain exclusive sums are used for wide allowance ranges during inversion.

Fixed-state scope. A common target-class bias changes probabilities while leaving every primitive error and allowance fixed. As that bias tends to infinity, the exact and approximate target probabilities tend to one; the loss and feature-gradient errors and their objective-conditioned bounds tend to zero. Changing $f = \tau u$ at fixed u instead preserves probabilities and all primitive inputs, but changes the named feature-gradient scale by $1/\tau$. These are the distinct interventions in Section 5.1.

F.5 Proof of class-weight–error coupling

Subtracting $L_y(q) = \log \sum_j \exp(q_j) - q_y$ from $L_y(q + e)$ gives Eq. (12). Subtracting e_y inside the logarithm rewrites it as $\log[t + \sum_{j \ne y} P_j \exp(e_j - e_y)]$. For $p_i \le p_k$ and $z_i \le z_k$, the exchange identity

$$(p_i z_i + p_k z_k) - (p_i z_k + p_k z_i) = (p_k - p_i)(z_k - z_i) \ge 0$$

shows that matched ordering maximizes the weighted sum and reverse ordering minimizes it. Apply this classical rearrangement argument to $z_i = \exp(a_i - e_y)$. Adding t and taking the increasing logarithm preserves the extrema, proving Proposition 3. Every signed change lies between the two attained endpoints; its largest absolute value is therefore the larger endpoint magnitude. The finite set of pairings need not attain zero even when $D_- < 0 < D_+$. Common shifts of all errors leave the identity unchanged. For $C \geq 3$, if neither non-target multiset is constant, the two extrema differ. These extrema characterize signed loss changes. Ordering gradient-error norms would additionally require accounting for the class directions in the gradient decomposition.

G A Second vMF Objective

The propagation interface also applies to supervised probabilistic vMF embeddings (Scott et al., 2021), where the primitive enters both a log-normalizer difference and a ratio-dependent attraction term. Write $z = a\mu_z$, $w_j = k_j\mu_j$, and fix common unit samples s and inverse temperature $\beta > 0$. The objective is

$$L = \mathbb{E}_s \operatorname{LSE}_j\{\Phi_\nu(\|w_j + \beta s\|) - \Phi_\nu(k_j)\} - \beta R_\nu(a)R_\nu(k_y)\mu_z^\top \mu_y.$$

Replace Φ_ν, R_ν by a paired F, A with uniform ratio allowance b. The reverse triangle inequality gives $|\|w_j + \beta s\| - k_j| \leq \beta$, so the normalizer-difference score error is at most βb. Log-sum-exp therefore changes by at most βb for each fixed sample. Since $0 \leq R_\nu \leq 1$ and $|A - R_\nu| \leq b$,

$$|A(a)A(k_y) - R_\nu(a)R_\nu(k_y)| \leq 2b + b^2.$$

Consequently,

$$|\widehat{L} - L| \leq \beta(3b + b^2).$$

The bound holds for a common finite sample average and after expectation over a common distribution. Class-specific endpoint and ratio allowances can replace these uniform terms. The same primitive-allowance interface thus supports the distinct propagation algebra of the supervised vMF objective.

H Numerical Realization and Certificate Construction

This appendix derives the state-dependent primitive envelope, constructs score allowances, and gives the stable endpoint representation used by the paired surrogate.

H.1 State-dependent residual envelope

Let

$$e(x) = R_\nu(x) - A_{2,\nu}(x)$$

denote the exact pointwise ratio error. Returning from the rescaled Riccati analysis of Appendix E to the original coordinate, e satisfies

$$e' + \left[\frac{2\nu + 1}{x} + R_\nu(x) + A_{2,\nu}(x)\right] e = \mathcal{E}(x), \tag{50}$$

where the residual admits the concentration-dependent bound summarized below.

Define

$$m = 2\nu + 2 \tag{51}$$

and

$$D_\nu(t) = \nu^{-3}\left[\left(1 + (t/\nu)^2\right)^{-3/2} + \frac{1}{4\nu}\left(1 + (t/\nu)^2\right)^{-2}\right]. \tag{52}$$

The coefficient multiplying e in Eq. (50) is bounded below by

$$\frac{2\nu + 1}{x}.$$

For the exact ratio, nonnegativity is immediate:

$$R_\nu(x) \geq 0.$$

For $A_{2,\nu}$, use the rescaled coordinate $z = x/\nu$ and the notation of Appendix E. When $z \le 2$,

$$r_2(z) \ge 0$$

and

$$r_0 + \varepsilon r_1 \ge r_0 \left(1 - \frac{\varepsilon}{S}\right). \tag{53}$$

When $z \ge 2$,

$$\frac{|r_2|}{r_0} \le \frac{1}{4S^2}, \qquad S \ge \sqrt{5}. \tag{54}$$

These inequalities imply

$$A_{2,\nu}(x) \ge 0$$

throughout the theorem domain. Hence the coefficient of e is indeed at least $(2\nu + 1)/x$.

Using the continuous endpoint condition

$$e(0) = 0$$

and variation of constants in Eq. (50) gives

$$|e(x)| \le x^{-(m-1)} \int_0^x t^{m-1} D_\nu(t)\, dt. \tag{55}$$

The function D_ν is decreasing. For any

$$\alpha \in [0, 1],$$

split the integral at αx. On the first interval,

$$D_\nu(t) \le D_\nu(0),$$

while on the second,

$$D_\nu(t) \le D_\nu(\alpha x).$$

Therefore

$$\begin{aligned}\int_0^x t^{m-1} D_\nu(t)\, dt \le D_\nu(0) \int_0^{\alpha x} t^{m-1}\, dt \\ + D_\nu(\alpha x) \int_{\alpha x}^x t^{m-1}\, dt.\end{aligned} \tag{56}$$

Evaluating the two monomial integrals and multiplying by $x^{-(m-1)}$ yields the pointwise bound

$$B_\alpha(x) = \frac{x}{m} \left\{ D_\nu(0)\alpha^m + D_\nu(\alpha x)\left(1 - \alpha^m\right) \right\}. \tag{57}$$

Thus, for every admissible α,

$$|R_\nu(x) - A_{2,\nu}(x)| \le B_\alpha(x). \tag{58}$$

Any minimum over valid choices remains a valid certificate. In particular, the implementation used in the experiments takes the minimum of the uniform quarter allowance and six state-dependent candidates with

$$\alpha = \exp(-c/m), \qquad c \in \{1, 2, 4, 8, 16, 32\}. \tag{59}$$

Each candidate is valid separately; their minimum tightens the certificate.

H.2 Score certificates from pointwise envelopes

For an endpoint interval

$$[l, h], \qquad 0 \le l \le h,$$

a valid pointwise envelope $B(x)$ on the ratio error yields

$$\left| \int_l^h \left[R_\nu(x) - A_{2,\nu}(x) \right] dx \right| \le \int_l^h B(x)\, dx. \tag{60}$$

For the state-dependent candidate $B_\alpha(x)$ in Eq. (57), a simple upper-sum construction is obtained by partitioning the endpoint interval into panels. On a panel $[l_p, h_p]$, we use

$$\frac{x}{m} \le \frac{h_p}{m}$$

and, because D_ν is decreasing,

$$D_\nu(\alpha x) \le D_\nu(\alpha l_p).$$

Hence

$$B_\alpha(x) \le \frac{h_p}{m}\left\{D_\nu(0)\alpha^m + D_\nu(\alpha l_p)\left(1-\alpha^m\right)\right\} \tag{61}$$

for every $x \in [l_p, h_p]$.

Multiplying the right-hand side by the panel width and summing over all panels therefore produces an exact-arithmetic upper sum for the score error. The experiments use 16 equal panels; any partition with valid pointwise envelopes preserves the guarantee.

H.3 Stable endpoint realization

The vMF objectives use potential differences

$$F_{2,\nu}(r) - F_{2,\nu}(k)$$

rather than isolated potential values. Direct subtraction of two nearly equal floating-point evaluations can lose substantial relative accuracy.

Let

$$d = r^2 - k^2, \qquad s_0 = s(k), \qquad s_1 = s(r),$$

and define the rationalized radius difference

$$\Delta_s = \frac{d}{s_1 + s_0}. \tag{62}$$

Instead of subtracting two separately evaluated potential values, compute

$$\begin{aligned}\widehat{q}(r,k) = \Delta_s - \nu\,\mathrm{log1p}\left(\frac{\Delta_s}{\nu + s_0}\right) - \frac{1}{2}\,\mathrm{log1p}\left(\frac{\Delta_s}{s_0}\right) \\ - \frac{\Delta_s}{8 s_1 s_0} + \frac{5\nu^2\Delta_s\left(s_1^2 + s_1 s_0 + s_0^2\right)}{24 s_1^3 s_0^3}.\end{aligned} \tag{63}$$

In exact arithmetic,

$$\widehat{q}(r,k) = F_{2,\nu}(r) - F_{2,\nu}(k). \tag{64}$$

Equation (63) is an algebraically equivalent evaluation of the same paired surrogate and preserves its exact-arithmetic certificates.

When the application supplies the squared-radius difference d directly, Eq. (62) avoids first constructing and subtracting two nearly identical rounded radii. With rounded endpoint radii as inputs, the attainable accuracy is also limited by information lost during their construction.

I Reference Evaluation and Numerical Scope

I.1 Saved-state provenance

The saved queries come from ProCo, PATT, and a vMF contrastive setting with its auxiliary classification term disabled (Du et al., 2024; He et al., 2025), with CIFAR, ImageNet-LT, and iNaturalist class banks (Krizhevsky, 2009; Deng et al., 2009; Van Horn et al., 2018). Each query retains its label, complete class statistics, prior, feature geometry, and temperature. Queries sharing a source batch or training run are not independent replications.

I.2 High-precision reference at shared geometry

Each radius is reconstructed once in FP64 and supplied unchanged to the Bessel reference, paired realization, and primitive envelopes. Scalar scores, ratios, and probabilities are evaluated with 80-digit `mpmath` arithmetic (The mpmath development team, 2026). Gradient coefficient differences are formed at high precision before contraction with the common FP64 direction vectors. This shared geometry isolates primitive approximation from differences in endpoint reconstruction.

Certificate inversion uses FP64 log-domain calculations, with the propagated bounds and inverse boundaries evaluated at high precision for the reference comparison. Stable formulas preserve small perturbations without subtracting nearly equal probabilities (Blanchard et al., 2021). No bound violation is observed in the saved-query comparison, and all predictions satisfying the margin certificate are preserved. The analytic proof establishes the mathematical guarantee; these comparisons assess its numerical realization.

I.3 Supplementary actual-response check under a fixed perturbation family

The perturbation family $F_b = F_2 - bH_\nu$ is fixed before observing the responses, with $H_\nu(x) = \sqrt{\nu^2 + x^2} - \nu$. Its derivative perturbation is bounded by b and vanishes at zero. We keep the saved geometry, biases, and targets fixed and recompute perturbed probabilities and feature gradients. Stable probability differences and shared direction vectors isolate the induced gradient discrepancy. The unchanged paired potential F_2 serves as the reference for this response comparison.

On the sampled endpoints, A_2 and H'_ν lie in $[0, 1]$. For $0 \le b \le 1$, the perturbed radial derivative therefore has magnitude at most one. Taking the unchanged F_2 realization as the approximate side of the gradient decomposition makes the same fixed-probability contract applicable, with $d_j = b|r_j - k_j|$ and $a_j = b$. This preserves the saved certified threshold while measuring the actual gradient difference independently.

We locate the first budget crossing resolved by a scan on $0 \le b \le 1$ and refine it locally, then compare it with the same query's gradient certificate. The resulting threshold is defined at the scan's resolution; a nonmonotone response may have narrower crossings between scan points.

Actual gradient-response thresholds span multiple orders of magnitude: 5.38 at budget 10^{-4} and 6.35 at 10^{-5}, with a span of 5.70 at 10^{-5} even within the fixed $p = 512$, $\tau = 0.1$ group. The sufficient certificate is typically more conservative for this fixed perturbation family; the median detected response threshold is 24.61 times the gradient certificate at 10^{-5}.

I.4 Stable realization and the floating-point boundary

Automatic differentiation through the paired potential agrees with its explicit derivative at the precision expected from the tested arithmetic (Griewank & Walther, 2008; Baydin et al., 2018). Near coincident endpoints, the stable expression in Appendix H.3 avoids the cancellation seen in direct subtraction. Those comparisons use identical stored squared-gap inputs and differentiation coordinates, separating realization error from Bessel-approximation error.

The mathematical allowance controls the exact-arithmetic surrogate. Floating-point endpoint construction, transcendental evaluation, reductions, and rounding introduce separate errors (Goldberg, 1991; Higham, 2002). Our numerical checks support the tested implementation range; a complete execution certificate would additionally require justified rounding allowances or outward-rounded computation. In particular, a probability rounded to zero is not automatically an exact zero in the contract.