Multiclass Classification with Rare Useful Features: Fundamental Limits and the Optimality of Diversity Pursuit Higher Criticism Thresholding

Gabriel Moryoussef
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
gmoryous@andrew.cmu.edu

Alon Kipnis
Efi Arazi School of Computer Science
Reichman University
Herzliya, Israel
alon.kipnis@runi.ac.il

Abstract

We consider classification in high-dimensional feature spaces where only a vanishing fraction of features are useful and the number of classes may be large. While this rare-and-weak useful-feature regime is well understood in the binary setting, its multiclass version raises new questions: a feature can be useful for class discrimination without separating any specific pair of classes. We focus on a rare feature diversity model, in which most features are useless, but a small fraction promote discrimination among classes in a way that can be detected by a per-feature diversity test, such as one-way ANOVA or the Kruskal–Wallis test. For this setting, we propose Diversity Pursuit Higher Criticism Thresholding (DP-HCT), a feature-screening procedure that applies a higher-criticism threshold to the collection of p-values obtained from per-feature diversity tests. We justify DP-HCT under a Gaussian multiclass Rare-and-Weak Useful Feature model with class-mean contrasts of fixed magnitude in random directions, extending the binary model of Donoho and Jin (2008). When the number of classes is fixed, our model recovers the binary phase transition. When the number of classes grows logarithmically with the number of features, however, we derive a new and strictly harder phase transition, driven by the growing degrees of freedom and their effect on the useful-feature tail rate. We show that whenever classification beyond chance is possible, DP-HCT with ANOVA, followed by a simple classifier based on similarity to class means, achieves asymptotically perfect accuracy. Experiments on real multiclass classification benchmarks demonstrate substantial improvements, and numerical simulations confirm that the derived phase-transition curve accurately predicts the finite-sample behavior of DP-HCT in the growing-class regime.

1 Introduction

Modern classification problems often involve high-dimensional representations, where the number of measured features p is large but the discriminating information is concentrated in a small subset of coordinates. This sparse-useful-feature structure arises naturally in genomics, text, tabular data, and learned representations, and it motivates feature-screening methods that identify informative coordinates before classification. Such identification is also known to be crucial when data are scarce or the noise level is high [Ingster et al., 2009, Donoho and Jin, 2009, Fan and Fan, 2008, Azar and Nadler, 2024]. Understanding when classification is possible in this rare-and-weak useful-feature regime is important for assessing sample complexity, the value of feature engineering, and the role of dimensionality reduction or feature screening.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

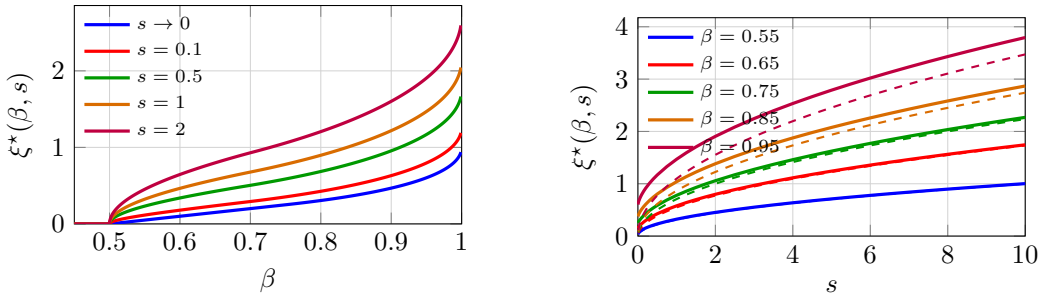


Figure 1: Multiclass classification boundary $\xi^\star(\beta, s)$ of (20). Left: boundary as a function of the feature-rarity exponent β, for several values of the class-growth parameter s. Right: boundary as a function of s, for several values of β. Dashed lines represent the large s approximation $\sqrt{2sr}$. The case $s \to 0$ corresponds to the binary Rare-and-Weak Useful Feature boundary [Donoho and Jin, 2008, Jin, 2009]. Larger s shifts the boundary upward, indicating the larger signal strength r required for nontrivial classification.

In the binary case, rare-and-weak useful-feature models provide a sharp theoretical framework for understanding this phenomenon [Donoho and Jin, 2008]. They characterize regimes in which classification is possible despite individual signals being weak [Jin, 2009, Donoho and Jin, 2009, Loureiro et al., 2021]. However, multiclass classification introduces qualitatively new issues. A feature may be useful not because it separates one class from another, but because it induces a pattern of variation across many classes [Bickel et al., 2010, Abramovich and Pensky, 2019, Thrampoulidis et al., 2020, Jin and Wang, 2016]. This raises a basic question: how does the number of classes affect the boundary between possible and impossible classification?

We address this question through a Gaussian multiclass Rare-and-Weak Useful Feature model in which class centroids are largely identical, except in a small fraction of the coordinates which individually exhibit statistically detectable diversity signal across classes. This model is designed to isolate the effect of the number of classes on classification in high-dimensions. Our main contribution is a phase-transition analysis for this model that separates classification with accuracy tending to one from classification with accuracy tending to chance level. When the number of classes is fixed, the classification boundary agrees with the binary rare/weak theory [Donoho and Jin, 2008, 2009, Jin, 2009, Donoho and Jin, 2015]. In contrast, when the number of classes grows logarithmically with the number of features, a new phase transition emerges, capturing the additional difficulty introduced by many classes.

The achievability side of the theory is established by a simple diversity-based screening procedure that computes a diversity score for each feature individually, then applies the higher-criticism threshold for selecting features [Donoho and Jin, 2008]. We call this screening procedure Diversity Pursuit Higher Criticism Thresholding (DP-HCT). Under the rare-and-weak useful feature model, DP-HCT followed by a simple centroid-similarity classifier asymptotically achieves perfect classification throughout the region where nontrivial classification is possible. Complementing this theoretical characterization of the classification boundary, experiments on several real-world multiclass benchmarks show performance gains from DP-HCT.

Contributions

1. **A fundamental limit for multiclass classification with rare and weak useful features.** We introduce a multiclass Rare-and-Weak Useful Feature model in which the number of classes grows logarithmically with the number of features p. The model is governed by three calibration parameters: β, which controls feature rarity; s, which controls the growth rate of the number of classes; and r, which controls signal strength. Our main result establishes a sharp phase transition as $p \to \infty$: there exists a function $\xi^\star(\beta, s)$ such that if $r < \xi^\star(\beta, s)$, no classifier achieves accuracy beyond chance asymptotically, whereas if $r > \xi^\star(\beta, s)$, classification with accuracy tending to one is possible. The function $\xi^\star(\beta, s)$ is illustrated in Fig. 1.

2. **Constructive achievability via Diversity Pursuit Higher Criticism Thresholding.** We prove that this fundamental limit is attained by a simple diversity-based screening procedure. Specifically, when $r > \xi^\star(\beta, s)$, applying higher-criticism thresholding to ANOVA p-values, followed by a centroid-similarity classifier on the selected features, achieves classification accuracy tending to one.
3. **Numerical validation and empirical evidence on multiclass benchmarks.** We run large-scale Monte Carlo experiments across the model's defining parameters. The simulations show that HC thresholding, despite not using the model parameters, closely tracks the oracle threshold and yields empirical accuracy consistent with the predicted phase-transition curve $\xi^\star(\beta, s)$. We also compare classification post-DP-HCT to alternative screening rules, including Bonferroni and Benjamini–Hochberg FDR [Benjamini and Hochberg, 1995]. Finally, we evaluate classification post-DP-HCT on five real multiclass benchmarks: gene expression (Heart-Circadian, Worm-Bacteria), metabolomics (Blood-Diet, Zebrafish-BPF) and miRNA expression (Biliary-miRNA). Across these datasets, DP-HCT substantially improves the kernel SVM and logistic-regression baselines, providing proof-of-concept evidence for diversity-based HC screening in multiclass problems.

Significance of the contributions compared to previous work. The phase transition expression we derive provides an accurate description of the effect of the number of classes on the possibility of successful classification. By using this expression as a pivot, one can assess the sample complexity of classification problems or the amount of feature engineering needed to achieve successful classification. Similar attempts in related models are provided in Abramovich et al. [2021], Thrampoulidis et al. [2020]. To the best of our knowledge, our setting is the first to incorporate the number of classes in an asymptotic phase transition analysis.

Classification under a rare-and-weak useful-feature setting has been studied by Donoho and Jin [2008, 2009], who also proposed using higher criticism for feature screening. The impossibility of classification in such a model was established in Jin [2009]. Our work is best seen as an extension of this result to the multiclass setting with an increasing number of classes. Our phase transition curve accounts for the change in the full useful-feature tail rate caused by the added degrees of freedom, an effect absent from the binary and fixed-class settings. Other rare-and-weak useful-feature settings can be found in Azar and Nadler [2024], Verzelen and Arias-Castro [2017] and references therein. More closely related to the multiclass setting, the works Jin and Wang [2016], Jin et al. [2017], Ma et al. [2026] proposed using higher criticism to identify useful features for clustering. Their setting can also be seen as a diversity-seeking situation among several groups, for which higher criticism has provided promising empirical results. The associated theoretical analysis, however, does not incorporate growth in the number of groups. We believe that ideas from our work can refine the theoretical understanding of clustering as well.

Analysis Overview The contributions combine several techniques. We derive the useful-feature tail exponent directly from the noncentral chi-square moment-generating function, using a Chernoff bound and an exponential tilting adjusted to the increase in the degrees of freedom. This analysis extends the fixed-degree chi-square approximations in Kipnis [2025], Donoho and Jin [2004]. Concentration of the within-class variance transfers this exponent to the exact ANOVA F-test p-values. We distinguish two regimes: $C = o(\log p)$, including fixed C, which recovers the binary boundary; and $C - 1 \sim 2s \log p$ with fixed $s > 0$, which yields the growing-class boundary.

When C is fixed (or grows as $o(\log p)$), the analysis recovers the classical Gaussian boundary [Donoho and Jin, 2004, 2008]; when $C \sim 2s \log p$, it produces a strictly different boundary. The new boundary represents an increase in the difficulty of classification due to the increased number of classes.

The impossibility region follows from a truncated likelihood-ratio second-moment argument for the joint training and test experiment. Above the boundary, HC attains a polynomially large score, which guarantees polynomially many useful selections and a favorable ratio of squared useful selections to total selections. Uniform threshold bounds transfer these properties to the data-chosen threshold.

Organization Section 2 introduces DP-HCT for feature screening. In Section 3, we specify the multiclass Rare-and-Weak Useful Feature framework. In Section 4, we derive the asymptotic phase transition for successful classification and we show that DP-HCT based on ANOVA and a centroid-similarity classifier achieves the theoretical detectability region. Section 5 reports simulation

Algorithm 1 DP-HCT: Diversity Pursuit with Higher Criticism Thresholding.

Require: training data $\{(X_i, Y_i)\}_{i=1}^n$; diversity test D;

1: **for** $j = 1, \ldots, p$ **do**
2: compute D_j on $\{(X_{i,j}, Y_i)\}_{i=1}^n$; obtain p-value π_j.
3: **end for**
4: sort: $\pi_{(1)} \le \pi_{(2)} \le \cdots \le \pi_{(p)}$.
5: for $j = 1, \ldots, \lfloor \gamma_0 p \rfloor$, set $\mathrm{HC}_{p,j} \leftarrow \sqrt{p}\,(j/p - \pi_{(j)})/\sqrt{j/p(1-j/p)}$.
6: $j^\star \leftarrow \arg\max_{1 \le j \le \lfloor \gamma_0 p \rfloor} \mathrm{HC}_{p,j}$;
7: $\hat{\mathcal{J}} \leftarrow \{j \in [p] : \pi_j \le \pi_{(j^\star)}\}$.

Ensure: selected set $\hat{\mathcal{J}}$

and real-data results. Additional simulation results and full proofs are provided in the Supplementary Material.

2 Diversity Pursuit with HC Thresholding

We now introduce the DP-HCT feature-screening procedure. Subsequent sections justify this procedure theoretically and empirically.

2.1 Feature screening

We have a training dataset $\{(X_i, Y_i)\}_{i=1}^n$ with $X_i \in \mathbb{R}^p$ and $Y_i \in [C] := \{1, \ldots, C\}$. For $j \in [p]$ and $c \in [C]$, let $P_{c,j}$ denote the class-conditional marginal distribution of feature j. A feature j is *useless* (a.k.a. *null*) if

$$H_0^{(j)} : P_{1,j} = P_{2,j} = \cdots = P_{C,j}, \tag{1}$$

and *useful* otherwise. Our goal is to improve classification performance by removing useless features. We propose the following DP-HCTscreening procedure, which decomposes into three steps.

(i) Per-feature diversity testing. For each j, compute a test statistic D_j of (1) on $\{(X_{i,j}, Y_i)\}_{i=1}^n$ and convert it to a p-value $\pi_j \in [0, 1]$.

(ii) Higher Criticism Thresholding. Let $\pi_{(1)} \le \cdots \le \pi_{(p)}$ denote the order statistics of $\{\pi_j\}_{j=1}^p$. The Higher Criticism Thresholding (HCT) procedure selects a set of indices $\hat{\mathcal{J}} \subseteq [p]$ whose p-values are below a threshold $\pi_{(j^\star)}$, where $j^\star$ is the index that maximizes the Higher Criticism statistic. Specifically, with

$$\mathrm{HC}_{p,j} \triangleq \sqrt{p} \cdot \frac{j/p - \pi_{(j)}}{\sqrt{t_j(1-t_j)}}, \qquad j = 1, \ldots, p,$$

where $t_j := j/p$. The HC statistic and the HCT are defined as in Donoho and Jin [2008, 2009, 2015]:

$$\mathrm{HC}_p := \mathrm{HC}_{p,j^\star}, \qquad j^\star := \arg \max_{1 \le j \le \lfloor p\gamma_0 \rfloor} \mathrm{HC}_{p,j}, \tag{2}$$

where γ_0 is a tunable parameter that reflects prior knowledge about the proportion of useful features; we used $\gamma_0 = 1/2$ in all evaluations.

(iii) Selection. The higher criticism-selected set is

$$\hat{\mathcal{J}} := \{j \in [p] : \pi_j \le \pi_{(j^\star)}\}.$$

The selection procedure (i)-(iii) is summarized in Algorithm 1.

Given the selected set $\hat{\mathcal{J}}$, we define the feature-screened data as

$$X_i[\hat{\mathcal{J}}] := (X_{i,j})_{j \in \hat{\mathcal{J}}} \in \mathbb{R}^{|\hat{\mathcal{J}}|}. \tag{3}$$

We train a multiclass classifier $\hat{h}$ on $\{(X_i[\hat{\mathcal{J}}], Y_i)\}_{i=1}^n$. For a test point X_{n+1}, predict $\hat{Y}_{n+1} := \hat{h}(X_{n+1}[\hat{\mathcal{J}}])$.

2.2 Diversity Statistics

Any per-feature test of (1) that produces a null-calibrated p-value fits the template, and different tests can even be applied to different features. Two canonical tests that we explain in detail and apply in practice are the classical Analysis of Variance (ANOVA) F-test and the Kruskal–Wallis test [Kruskal and Wallis, 1952]. Aside from these tests, other procedures can be used as diversity statistics, such as the Friedman test and permutation-based tests [Lehmann and Romano, 2006, Ch. 7]. Post-hoc tests such as Tukey's method [Lehmann and Romano, 2006, Ch. 9] may be useful in extensions of the framework to hierarchical features.

ANOVA F-test. Let $n_c := |\{i : Y_i = c\}|$ be the number of observations in class c and assume $n_c \geq 2$. At feature j, denote the class sample mean by

$$\hat{\mu}_{c,j} := \frac{1}{n_c} \sum_{i\,:\,Y_i=c} X_{i,j}, \quad c = 1, \ldots, C, \tag{4}$$

Define the within- and between-class sums of squares as

$$\mathrm{SS}_{\text{between}}(j) = \sum_{c=1}^{C} n_c \big(\hat{\mu}_{c,j} - n^{-1} \sum_{a=1}^{C} n_a \hat{\mu}_{a,j}\big)^2, \quad \mathrm{SS}_{\text{within}}(j) = \sum_{c=1}^{C} \sum_{i\,:\,Y_i=c} \big(X_{i,j} - \hat{\mu}_{c,j}\big)^2, \tag{5}$$

and

$$f_j := \frac{\mathrm{SS}_{\text{between}}(j)/(C-1)}{\mathrm{SS}_{\text{within}}(j)/(n-C)}. \tag{6}$$

With $F_{C-1,n-C}$ denoting the F distribution with $(C-1, n-C)$ degrees of freedom, p-values are provided by

$$\pi_j := \mathbb{P}\{F_{C-1,n-C} \geq f_j\}, \qquad j = 1, \ldots, p. \tag{7}$$

This test is appropriate under $H_0^{(j)}$ of (1) when the class-conditional marginals are Gaussian, i.e., $P_{c,j} = \mathcal{N}(\mu_c^{(j)}, \sigma^2)$, with an unknown common variance σ^2.

Kruskal–Wallis Test. This test is often seen as the rank-based analogue of the ANOVA F-test. Rank-based diversity is invariant to monotone transformations and requires no parametric assumption. Therefore, the Kruskal–Wallis (KW) test is particularly useful when class-conditional marginals are heavy-tailed, skewed, or contaminated [Lehmann and Romano, 2006, Ch. 7].

Let $R_{i,j}$ be the rank of $X_{i,j}$ in $\{X_{1,j}, \ldots, X_{n,j}\}$ and $\bar{R}_{c,j} := n_c^{-1} \sum_{i\,:\,Y_i=c} R_{i,j}$ be the class mean rank at feature j. The KW statistic of the j-th feature is defined as

$$\mathrm{KW}_j := \frac{12}{n(n+1)} \sum_{c=1}^{C} n_c \Big(\bar{R}_{c,j} - \tfrac{n+1}{2}\Big)^2. \tag{8}$$

As $n \to \infty$ under $H_0^{(j)}$ with continuous marginals, KW_j converges to a χ^2_{C-1} distribution. Thus, (asymptotic) p-values are provided by

$$\pi_j := \mathbb{P}\{\chi^2_{C-1} \geq \mathrm{KW}_j\}, \qquad j = 1, \ldots, p. \tag{9}$$

3 Multiclass Rare-and-Weak Useful Feature Model

3.1 Overview

We consider a model with C classes, where only a vanishing fraction of features are useful for inter-class discrimination. We call a feature *useful* if it promotes discrimination among classes in a way that can be detected by a per-feature diversity test, such as (7) or (9). We introduce a single parameter τ that controls the strength of the useful class-mean contrasts. To enforce rarity, we suppose that each feature belongs to the set of useful features $\mathcal{J} \subset [p]$ with probability ϵ, for some small number $\epsilon \in (0, 1)$. The resulting Multiclass Rare-and-Weak Useful Feature model is denoted by $\mathrm{MCRW}(p, C, \epsilon, \tau)$.

3.2 Multiclass Gaussian data

Suppose that, conditional on $Y_i = c$,

$$X_i \mid Y_i = c \sim \mathcal{N}\big(\mu_c,\, \sigma^2 I_p\big), \qquad \mu_c = (\mu_c^{(1)}, \ldots, \mu_c^{(p)}) \in \mathbb{R}^p. \tag{10}$$

Here μ_c is the class-c centroid and $\sigma^2 > 0$ is a shared noise-scaling parameter. For simplicity, we assume that $n_c = n/C$ is an integer, and that the design is balanced with n_c observations per class. Unbalanced designs are interesting in some contexts [Krawczyk, 2016], but we leave their analysis as future work.

3.3 Rare useful features

Inter-class diversity in the generation process (10) is encoded by the class means $\mu_1, \ldots, \mu_C$. Let $\mathcal{J} \subset [p]$ denote the coordinates of features that promote inter-class separation in the following sense. For $j \notin \mathcal{J}$, all class means have the same j-th coordinate value, whereas for $j \in \mathcal{J}$, some class means may have j-th coordinates that differ from the others. The set of useful features $\mathcal{J}$ is assumed to be random, with each coordinate belonging to $\mathcal{J}$ with probability ϵ independently of the others.

Deterministic constructions of $\mathcal{J}$ that directly control $|\mathcal{J}|$ are also possible [Stoepker et al., 2024]; they are slightly more difficult to analyze, yet behave similarly under the asymptotic regime of interest introduced below.

3.4 Spherically symmetric diversity model of useful features

Let

$$\mathcal{V} := \mathbf{1}^\perp = \left\{u \in \mathbb{R}^C : \mathbf{1}^\top u = 0\right\}, \qquad P = I_C - \frac{1}{C}\mathbf{1}\mathbf{1}^\top.$$

For each $j \in \mathcal{J}$, draw independently

$$U^{(j)} \sim \mathrm{Unif}\{u \in \mathcal{V} : \|u\|_2 = 1\},$$

and set

$$v^{(j)} := \sqrt{C}\, U^{(j)}, \qquad \mu^{(j)} := \tau v^{(j)} = \big(\mu_1^{(j)}, \ldots, \mu_C^{(j)}\big)^\top. \tag{11}$$

For $j \notin \mathcal{J}$, all class means are equal in coordinate j, equivalently $P\mu^{(j)} = 0$. We say that this construction is spherically symmetric because, for each $j \in \mathcal{J}$,

$$\mathbf{1}^\top \mu^{(j)} = 0, \qquad \|\mu^{(j)}\|_2^2 = \tau^2 C.$$

The spherical contrast prior is a symmetry-preserving way to extend the binary Rare-and-Weak Useful Feature model of Donoho and Jin to multiple classes. When $C = 2$, the centered contrast space $\mathbf{1}^\perp$ is one-dimensional, and the uniform distribution on its unit sphere reduces to a random sign. For $C > 2$, the same construction gives the rotationally invariant prior on the centered class-contrast space. Such spherical priors are common in high-dimensional inference and random matrix theory because they preserve the symmetries of the null model and often lead to analytically tractable, least-favorable alternatives; see, for example, Baik et al. [2005], Perry et al. [2018], and El Alaoui et al. [2020]. Other multiclass sparse-feature constructions are possible, including the high-dimensional normal model of Abramovich and Pensky [2019], but it is unclear whether these can yield an interesting asymptotic behavior.

We denote the data-generation model defined so far as $\mathrm{MCRW}(p, C, \epsilon, \tau)$. We keep the dependence of the model on σ^2 and n implicit, since these parameters do not directly affect the asymptotic behavior.

3.5 The ideal threshold

For simplicity of the analysis, but also due to asymptotic optimality under $\mathrm{MCRW}(p, C, \epsilon, \tau)$, we consider a maximum-inner-product classification rule over the selected features:

$$\hat{h}(x; \mathcal{J}) := \arg\max_{c \in [C]} \big(x[\mathcal{J}]\big)^\top \hat{\mu}_c[\mathcal{J}], \tag{12}$$

where $\hat{\mu}_c = (\hat{\mu}_{c,1}, \ldots, \hat{\mu}_{c,p})$ collects the class sample means (4) and $x[\mathcal{J}]$ is coordinate selection operation (3). For a general threshold $t \in [0, 1]$, define the selected set of features based on ANOVA p-values as

$$\hat{\mathcal{J}}(t) := \{j \in [p] : \pi_j \leq t\}. \tag{13}$$

Notice that under this construction, the expected misclassification error under $\mathrm{MCRW}(p, C, \epsilon, \tau)$ is a function of the threshold t. We denote this function by

$$e(t) := \mathbb{E}_{\mathrm{MCRW}(p,C,\epsilon,\tau)} \left[\mathbf{1}\{\hat{h}(X; \hat{\mathcal{J}}(t)) \neq Y\} \right].$$

The ideal threshold, denoted by $t^\star$, is the one that minimizes $e(t)$:

$$t^\star := \arg\min_{t \in [0,1]} e(t). \tag{14}$$

It is not clear how to compute $e(t)$, as even the case of $C = 2$ is challenging and was not fully solved in Donoho and Jin [2009]. However, we can approximate $e(t)$, and thus $t^\star$, relatively straightforwardly using Monte Carlo simulations, as demonstrated in Fig. 3 below.

The ideal threshold $t^\star$ depends on the latent model parameters ϵ and τ, hence it is not accessible to the data analyst. We hope that HCT approximates the ideal threshold, at least in the parts of the parameter space above and nearby the detection threshold, where non-trivial classification is challenging but possible.

3.6 Model calibration for asymptotic analysis

We consider the asymptotic limit as $p \to \infty$, where the model parameters are calibrated to p through fixed parameters (β, r, s). The rarity level is

$$\epsilon = p^{-\beta}, \qquad \beta \in (1/2, 1). \tag{15}$$

The signal amplitude is calibrated by

$$\tau^2 = \frac{\sigma^2}{n} 2r \log p, \qquad r > 0. \tag{16}$$

Notice that under the spherical contrast construction, every useful feature has ANOVA non-centrality

$$\lambda_j = \frac{n_c}{\sigma^2} \mu^{(j)\top} P \mu^{(j)} = \frac{n/C}{\sigma^2} C\tau^2 = 2r \log p, \qquad j \in \mathcal{J}. \tag{17}$$

Thus, r is the standardized ANOVA-visible rare and weak signal strength [Donoho and Jin, 2004].

For the number of classes, we focus on the logarithmic multiclass regime

$$C \sim 2s \log p, \qquad s > 0. \tag{18}$$

We also discuss the fixed- or slowly-growing-class regime $C = o(\log p)$ as a comparison case. For the ANOVA F-test analysis, we assume

$$\log p = o(n), \tag{19}$$

so that the normalized within-class variance denominator of the F statistic concentrates at σ^2. Consequently, $(C - 1)f_j$ has the same polynomial-tail exponent as a non-central chi-square random variable with $C - 1$ degrees of freedom and non-centrality parameter λ_j.

4 Main Results: Classification Phase Transition and Optimality of DP-HCT

We now state the main theoretical consequences for DP-HCT under the $\mathrm{MCRW}(p, C, \epsilon, \tau)$ framework of Section 3. Throughout, π_j denotes the p-values obtained from the one-way ANOVA F-test (7).

4.1 The classification boundary

For $s > 0$, let $x_s(q) \geq s$ solve

$$g_s(x_s(q)) = q, \qquad g_s(x) = x - s - s\log(x/s),$$

and set $x_0(q) = q$. Define

$$\Psi_{r,s}(q) = J_{r,s}(x_s(q)), \qquad J_{r,s}(x) = \sup_{0 \leq \theta < 1} \left\{ \theta x + s\log(1-\theta) - \frac{r\theta}{1-\theta} \right\}.$$

For fixed (β, s), the signal-strength boundary function is defined by

$$\xi^\star(\beta, s) := \inf \left\{ r > 0 : \max_{0 \leq q \leq 1} \left(\frac{1}{2} - \beta + \frac{q}{2} - \Psi_{r,s}(q) \right) > 0 \right\}. \tag{20}$$

The maximum in (20) has an explicit optimizer, given in Section H of the Supplementary Material.

The boundary function $\xi^\star(\beta, s)$ is strictly increasing in s and has the following limiting forms.

- As $s \to 0$, $\xi^\star(\beta, s) \to \rho^\star(\beta)$, where $\rho^\star(\beta)$ is the Ingster-Donoho-Jin rare signal detection boundary function

$$\rho^\star(\beta) := \begin{cases} \beta - 1/2, & 1/2 < \beta < 3/4, \\ (1 - \sqrt{1-\beta})^2, & 3/4 \leq \beta < 1. \end{cases}$$

  This is also the boundary for $C = o(\log p)$, fixed C in particular.
- As $s \to \infty$, $\xi^\star(\beta, s) \sim \sqrt{2s\,\rho^\star(\beta)}$. This is an iterated limit of the fixed-s boundary. The ANOVA numerator has null standard deviation $\sqrt{2(C-1)}$, so its standardized non-centrality is $\lambda/\sqrt{2(C-1)} \sim r\sqrt{\log p/s}$. Thus additional degrees of freedom reduce per-feature separation. The large-s approximation is shown as dashed reference lines in Fig. 1 (right).

4.2 Impossibility below the boundary

Theorem 1 (Impossibility below the boundary). *Under* $\mathrm{MCRW}(p, C, \epsilon, \tau)$ *with calibrations* (15)–(19) *and a uniformly distributed independent test label, if* $r < \xi^\star(\beta, s)$*, then every measurable classifier* $\hat{h}$ *satisfies*

$$\mathbb{P}\{\hat{h}(X_{n+1}) = Y_{n+1}\} \leq \frac{1}{C} + o(1).$$

Consequently,

$$\sup_{\hat{h}} \mathbb{P}\{\hat{h}(X_{n+1}) = Y_{n+1}\} = \frac{1}{C} + o(1).$$

The proof is in Section F of the Supplementary Material. It uses a truncated likelihood-ratio second-moment argument to establish vanishing total variation for the joint training and test experiment; see Butucea and Ingster [2013] for a related use of this method.

4.3 DP-HCT succeeds above the boundary

Theorem 2 (DP-HCT succeeds above the boundary). *Under* $\mathrm{MCRW}(p, C, \epsilon, \tau)$ *with calibrations* (15)–(19) *and* $n = p^{o(1)}$*, let* $\hat{\mathcal{J}}$ *be the selected set of features based on ANOVA p-values* (7) *and threshold* (13)*, and let* $\hat{h}$ *be the maximum-inner-product classifier* (12) *with* $\tilde{\mathcal{J}} = \hat{\mathcal{J}}$*. If* $r > \xi^\star(\beta, s)$*, then, for an independent test point* $(X_{n+1}, Y_{n+1}) \sim \mathrm{MCRW}(p, C, \epsilon, \tau)$*,*

$$\mathbb{P}\{\hat{h}(X_{n+1}; \hat{\mathcal{J}}) \neq Y_{n+1}\} = o(1).$$

We leave for future work the study of DP-HCT optimality under other sample-size regimes, including polynomial growth $n = p^{\theta + o(1)}$ with fixed $\theta \in (0, 1)$.

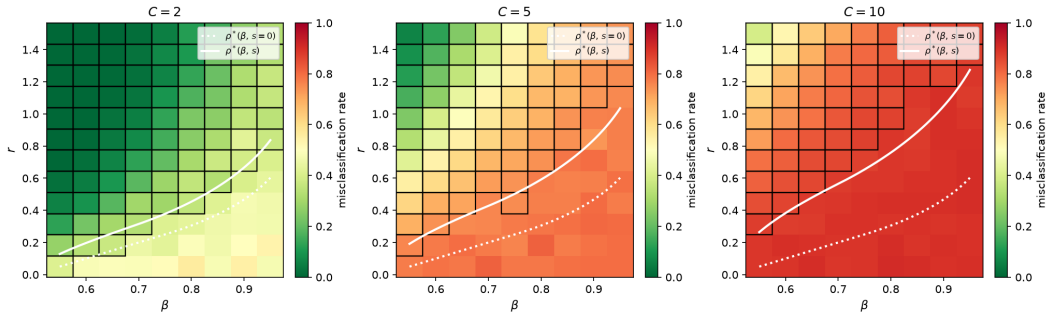


Figure 2: **DP-HCT misclassification phase transition**. Misclassification rate (color) over a grid of (β, r) values for $C \in \{2, 5, 10\}$, under the Gaussian multiclass Rare-and-Weak Useful Feature model with $p = 2{,}000$, $n_{\text{Monte}} = 150$. The x-axis is the sparsity exponent β and the y-axis is the calibrated signal strength r. Configurations in which the misclassification rate is above chance are marked with black frames. Solid white: the theoretical phase transition curve (Theorem 2). Dotted white: the phase transition curve corresponding to a fixed number of classes, derived in Donoho and Jin [2008].

5 Experiments

5.1 Verification of successful classification region

Setup. We simulate from the $\text{MCRW}(p, C, \epsilon, \tau)$ model with $p = 2{,}000$ and ϵ, τ, and C as in, respectively, (15), (16), and (18). We use $n_{\text{Monte}} = 150$ independent problem instances per (β, r) grid point. After feature screening, classification is performed with the centroid-similarity classifier of (12).

Misclassification phase transition. Fig. 2 shows the misclassification rate over a grid of (β, r) values for $C \in \{2, 5, 10\}$; configurations in which the misclassification rate is significantly above chance are marked with black frames. As C increases, the boundary rises, reflecting the increased difficulty of multiclass discrimination; two theoretical phase-transition boundaries are overlaid on every heatmap: (1) the boundary $r = \xi^\star(\beta, s)$ of (20) (solid) and (2) the boundary $r = \rho^\star(\beta)$ (dotted). The boundary (1), established in Theorems 1 and 2, assumes that C increaes with p. The boundary (2), derived in Donoho and Jin [2008], corresponds to a fixed C. The figure shows that the above-chance region better follows boundary (1), even for small C. Additional figures characterizing the ideal threshold and per-C side-by-side comparisons are collected in the Supplementary Material, Section A.

Comparison with other thresholds. Fig. 3 compares four feature-selection rules for data sampled from the $\text{MCRW}(p, C, \epsilon, \tau)$. The *ideal threshold* of (14), which achieves the best performance achievable by thresholding the ANOVA p-values, but is not a function of the data. We compare it with DP-HCT, Benjamini-Hochberg's False-Discovery Rate at level 0.10 [Benjamini and Hochberg, 1995], and selection after Bonferroni's correction with $\alpha = 0.05$.

5.2 Real data benchmarks

The main theoretical justification for DP-HCT is its role in a classification pipeline that attains the multiclass Rare-and-Weak Useful Feature phase transition. To complement this theoretical characterization, we evaluate classification following DP-HCT against baseline methods on five public biological multiclass benchmarks with $p > n$, spanning gene expression, metabolomics, and miRNA expression; see Section B.1 of the Supplementary Material for preprocessing details.

- Heart-Circadian (GSE10045; Bray et al., 2008): $n = 130$ mouse-heart expression profiles, $C = 8$ times of day, $p = 46{,}120$ probes.
- Worm-Bacteria (GSE15923; Coolon et al., 2009): $n = 72$ *C. elegans* expression profiles, $C = 4$ bacterial diets, $p = 672$ features.
- Blood-Diet (ST002019; Ferguson, 2022, Metabolomics Workbench): $n = 70$ blood LC–MS profiles from 35 participants, $C = 4$ diet $\times$ study-phase groups, $p = 736$ metabolite features.

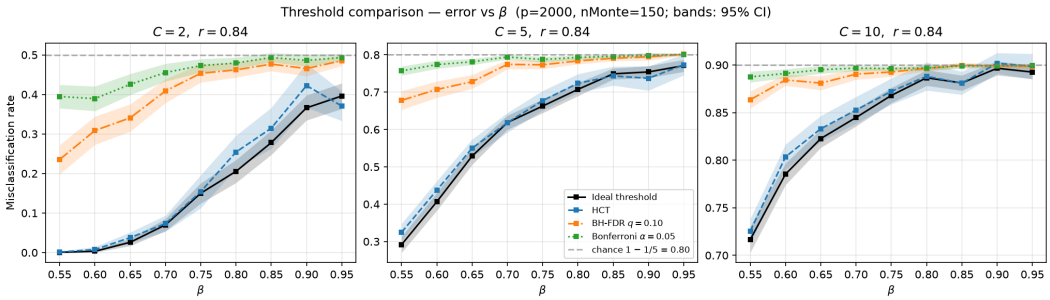


Figure 3: **Threshold comparison:** classification error vs. β for a fixed r value and several value of C ($p = 2{,}000$, $n_{\text{Monte}} = 150$). Solid black: ideal threshold of (14); dashed blue: DP-HCT; dash-dot orange: Benjamini-Hochberg's False-Discovery Rate (BH-FDR) at level 0.10; dotted green: Bonferroni ($\alpha = 0.05$). Dashed grey: chance-level error $1 - 1/C$. It can be seen that DP-HCT classification performance closely follows the ideal threshold's performance.

Table 1: CV accuracy (mean ± SEM, %) for classification post-DP-HCT feature screening with ANOVA F ("F") or KW ("KW") tests; results are averaged over 10-fold, unit stratified cross-validation (7-fold for Heart-Circadian). "# Sel." is the mean number of features retained by DP-HCT.

		Baseline (all features)		**DP-HCT-selected features**		
Dataset	$n/p/C$	SVM	LoR	SVM	LoR	# Sel.
Heart-Circadian (F)	130/46120/8	14.7±2.4	32.4±4.5	**74.8±6.1**	**76.3±4.9**	460±47
Worm-Bacteria (KW)	72/672/4	24.2±4.2	71.5±5.1	**53.4±7.7**	**80.5±5.4**	56±5
Blood-Diet (F)	70/736/4	47.5±5.0	60.8±7.1	**78.8±4.1**	**82.9±3.9**	108±7
Zebrafish-BPF (KW)	62/117/4	64.8±5.9	59.8±5.7	**72.9±7.7**	**73.9±7.0**	38±3
Biliary-miRNA (KW)	36/803/3	71.7±6.1	83.3±5.0	**90.0±4.4**	**96.7±2.2**	62±38

- Zebrafish-BPF (ST003804; Guo, 2025, Metabolomics Workbench): $n = 62$ zebrafish-liver LC–MS profiles, $C = 4$ exposure levels of bisphenol F, $p = 117$ metabolite features.
- Biliary-miRNA (GSE140001; Loeffler et al., 2020): $n = 36$ microdissected tissue samples from 12 patients, $C = 3$ tissue types, $p = 803$ miRNAs.

For the diversity test we use either ANOVA F (7) or KW (9), depending on which one obtained higher classification accuracy. We evaluate two classifiers before and after feature screening: a radial-basis-function SVM (DP-HCT-SVM) and ℓ_2-regularized multinomial logistic regression (DP-HCT-LoR). All baselines apply the same classifiers to the full, unselected feature matrix; see Section B of the Supplementary Material for full protocol details. Table 1 reports mean cross-validated accuracy for both classifiers pre- and post-DP-HCT, and the number of features retained ("# Sel.").

We also stress-test the pipeline, either by adding independent noise to every entry, scaled to each feature's own standard deviation, or by appending many pure-noise (useless) features. With added noise, DP-HCT-LoR substantially improves over the baseline at small noise levels in most cases. The improvement also holds up when a substantial number of pure-noise features is appended; the details and the results are in Section B.4 of the Supplementary Material.

6 Conclusion

DP-HCT extends higher-criticism feature screening to multiclass problems in which useful features are rare and individually weak. The theory identifies a multiclass phase transition that depends explicitly on the growth of the number of classes, and shows that diversity-based HC screening attains this boundary with a simple centroid-similarity classifier. Simulations confirm the predicted transition, while real-data benchmarks and stress tests show that the same screening principle can substantially improve classification when irrelevant or noisy coordinates dominate the feature space.

References

Felix Abramovich and Marianna Pensky. Classification with many classes: Challenges and pluses. *Journal of Multivariate Analysis*, 174:104536, 2019. doi: 10.1016/j.jmva.2019.104536.

Felix Abramovich, Vadim Grinshtein, and Tomer Levy. Multiclass classification by sparse multinomial logistic regression. *IEEE Transactions on Information Theory*, 67(7):4637–4646, 2021.

Eyar Azar and Boaz Nadler. Semi-supervised sparse gaussian classification: Provable benefits of unlabeled data. *Advances in Neural Information Processing Systems*, 37:20132–20169, 2024.

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005. doi: 10.1214/009117905000000233.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300, 1995.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Hierarchical selection of variables in sparse high-dimensional regression. In *Borrowing strength: theory powering applications–a Festschrift for Lawrence D. Brown*, volume 6, pages 56–70. Institute of Mathematical Statistics, 2010.

Molly S. Bray, Chad A. Shaw, Michael W. S. Moore, Rebeca A. P. Garcia, Marcelo M. Zanquetta, David J. Durgan, Woo Jung Jeong, Ju-Yun Tsai, Heiko Bugger, Dongfang Zhang, et al. Disruption of the circadian clock within the cardiomyocyte influences myocardial contractile function, metabolism, and gene expression. *American Journal of Physiology-Heart and Circulatory Physiology*, 294(2):H1036–H1047, 2008.

Cristina Butucea and Yuri I. Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688, 2013.

Joseph D. Coolon, Kenneth L. Jones, Timothy C. Todd, Bryan C. Carr, and Michael A. Herman. Caenorhabditis elegans genomic response to soil bacteria predicts environment-specific genetic effects on life history traits. *PLoS Genetics*, 5(6):e1000503, 2009.

David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32(3):962–994, 2004.

David Donoho and Jiashun Jin. Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 105(39): 14790–14795, 2008.

David Donoho and Jiashun Jin. Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical Science*, 30(1):1–25, 2015.

David L. Donoho and Jiashun Jin. Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4449–4470, 2009.

Ahmed El Alaoui, Florent Krzakala, and Michael I. Jordan. Fundamental limits of detection in the spiked Wigner model. *The Annals of Statistics*, 48(2):863–885, 2020. doi: 10.1214/19-AOS1826.

Jianqing Fan and Yingying Fan. High dimensional classification using features annealed independence rules. *Annals of statistics*, 36(6):2605, 2008.

Jane Ferguson. TIPs metabolomics (Metabolomics Workbench study ST002019, project PR001282). NIH Common Fund National Metabolomics Data Repository, Metabolomics Workbench, 2022. Vanderbilt University Medical Center. https://doi.org/10.21228/M80Q51.

Qiuyuan Guo. Metabolomics analysis of zebrafish liver under long-term BPF exposure (Metabolomics Workbench study ST003804, project PR002378). NIH Common Fund National Metabolomics Data Repository, Metabolomics Workbench, 2025. Washington University in St. Louis. https://doi.org/10.21228/M87Z5M.

Yuri I Ingster, Christophe Pouet, and Alexandre B Tsybakov. Classification of sparse high-dimensional vectors. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4427–4448, 2009.

Jiashun Jin. Impossibility of successful classification when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 106(22):8859–8864, 2009.

Jiashun Jin and Wanjie Wang. Influential features pca for high dimensional clustering. *The Annals of Statistics*, pages 2323–2359, 2016.

Jiashun Jin, Zheng Tracy Ke, and Wanjie Wang. Phase transitions for high dimensional clustering and related problems. *The Annals of Statistics*, 45(5):2151–2189, 2017.

Alon Kipnis. Unification of rare and weak multiple testing models using moderate deviations analysis and log-chisquared p-values. *Statistica Sinica*, 35:1155–1181, 2025.

Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in artificial intelligence*, 5(4):221–232, 2016.

William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.

Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

Markus A. Loeffler, Jing Hu, Marcus Kirchner, Xiaoyu Wei, Yue Xiao, Thomas Albrecht, Carolina De La Torre, Carsten Sticht, et al. miRNA profiling of biliary intraepithelial neoplasia reveals stepwise tumorigenesis in distal cholangiocarcinoma via the miR-451a/ATF2 axis. *The Journal of Pathology*, 252(3):239–251, 2020.

Bruno Loureiro, Gabriele Sicuro, Cédric Gerbelot, Alessandro Pacco, Florent Krzakala, and Lenka Zdeborová. Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. *Advances in Neural Information Processing Systems*, 34:10144–10157, 2021.

Chen Ma, Wanjie Wang, and Shuhao Fan. i-if-learn: Iterative feature selection and unsupervised learning for high-dimensional complex data. *arXiv preprint arXiv:2603.24025*, 2026.

Amelia Perry, Alexander S. Wein, Afonso S. Bandeira, and Ankur Moitra. Optimality and sub-optimality of PCA I: Spiked random matrix models. *The Annals of Statistics*, 46(5):2416–2451, 2018. doi: 10.1214/17-AOS1625.

Ivo V Stoepker, Rui M Castro, Ery Arias-Castro, and Edwin van den Heuvel. Anomaly detection for a large number of streams: A permutation-based higher criticism approach. *Journal of the American Statistical Association*, 119(545):461–474, 2024.

Manish Sud, Eoin Fahy, Dawn Cotter, Kenan Azam, Ilango Vadivelu, Charles Burant, Arthur Edison, Oliver Fiehn, Richard Higashi, K. Sreekumaran Nair, Susan Sumner, and Shankar Subramaniam. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Research*, 44(D1):D463–D470, 2016. doi: 10.1093/nar/gkv1042.

Christos Thrampoulidis, Samet Oymak, and Mahdi Soltanolkotabi. Theoretical insights into multiclass classification: A high-dimensional asymptotic view. *Advances in Neural Information Processing Systems*, 33:8907–8920, 2020.

Nicolas Verzelen and Ery Arias-Castro. Detection and feature selection in sparse mixture models. *Annals of Statistics*, 45(5):1920–1950, 2017. doi: 10.1214/16-AOS1513. URL https://hal.inrae.fr/hal-02624460.

Supplementary Material: Contents

A Synthetic-Data Simulations: Additional Figures **14**

B Real-Data Experimental Details **15**

B.1 Datasets . . . 15

B.2 Cross-validation protocol . . . 16

B.3 DP-HCT pipeline . . . 16

B.4 Stress-test protocol . . . 16

C Notation and auxiliary facts **18**

C.1 Notation . . . 18

C.2 Model and objects . . . 18

C.3 Chi-squared tail: quantitative expansion . . . 18

D ANOVA F-statistic and reduction to the chi-square numerator **21**

E Moderate-deviation exponent of ANOVA p-values **23**

E.1 Noncentral numerator tails . . . 23

E.2 Central numerator tails and quantiles . . . 23

E.3 Transfer to exact F p-values . . . 24

E.4 Uniformity and interpretation . . . 24

E.5 Local expansion at the useful-feature mean . . . 24

F Proof of Theorem 1 (Impossibility below the boundary) **25**

F.1 Reduction to the ANOVA p-value experiment . . . 26

F.2 One-feature likelihood ratios . . . 26

F.3 A likelihood-ratio upper bound . . . 27

F.4 Truncation . . . 27

F.5 Truncated second moment . . . 28

F.6 Conclusion: total variation and classification . . . 29

G HC growth and HCT adaptivity **29**

G.1 Count lemmas . . . 30

G.2 Pointwise HC growth . . . 32

G.3 Adaptivity of the HCT threshold . . . 33

H Boundary optimization **33**

I Proof of Theorem 2 (Misclassification rate of DP-HCT) **34**

I.1 Proof roadmap . . . 35

I.2 Reduction to pairwise margins . . . 35

I.3 Conditional Gaussianity of pairwise margins . . . 35

I.4 Count consequences of HCT adaptivity 36
I.5 Overshoot control 36
I.6 Spherical pairwise diversity 37
I.7 One-sided stability of selected signal noise 39
I.8 SNR divergence and conclusion 41

A Synthetic-Data Simulations: Additional Figures

All figures in this section use the Gaussian multiclass Rare-and-Weak Useful Feature model $\mathrm{MCRW}(p, C, \epsilon, \tau)$ with $n_{\mathrm{Monte}} = 150$ Monte Carlo draws per (β, r) grid point, unless stated otherwise. The ideal threshold for $\mathrm{MCRW}(p, C, \epsilon, \tau)$, HCT, BH-FDR ($q = 0.10$), and Bonferroni ($\alpha = 0.05$) selectors are all followed by a centroid similarity classifier. Phase boundaries overlaid on heatmaps are the multiclass boundary $r = \xi^\star(\beta, s)$ (solid) and the fixed-number-of-classes boundary of Donoho and Jin [2008] (dotted).

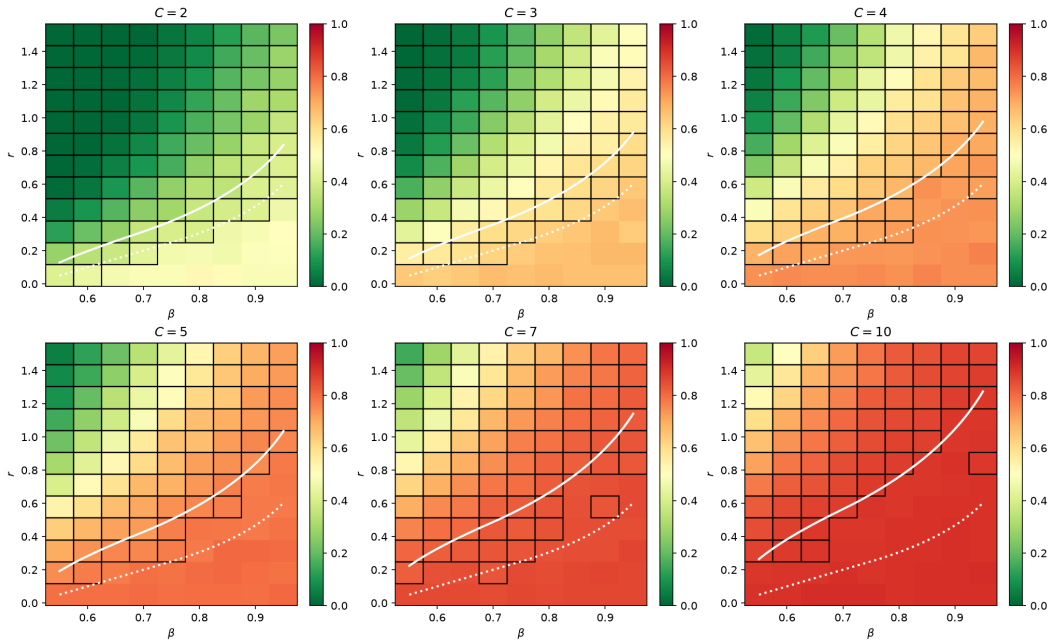


Figure 4: **Ideal-threshold misclassification rate heatmaps** for $C \in \{2, 3, 4, 5, 7, 10\}$. Color encodes misclassification rate. Solid white: the multiclass boundary $r = \xi^\star(\beta, s)$ of (20). Dotted white: the fixed-class Donoho–Jin boundary $r = \rho^\star(\beta)$. The transition from high to low error closely tracks the multiclass boundary; as C increases, the boundary rises, reflecting the harder multiclass discrimination problem.

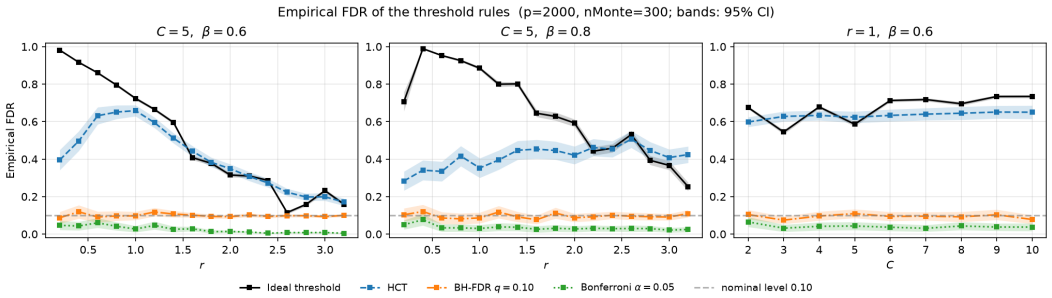


Figure 5: **Empirical false-discovery rate of the threshold rules.** The empirical FDR is the average, over Monte Carlo draws, of the proportion of selected features that are null. *Left*: FDR vs. r for $C = 5$, $\beta = 0.6$. *Middle*: FDR vs. r for $C = 5$, $\beta = 0.8$. *Right*: FDR vs. the number of classes $C \in \{2, \ldots, 10\}$ for $r = 1$, $\beta = 0.6$. Solid black: ideal threshold of (14); dashed blue: DP-HCT; dash-dot orange: Benjamini–Hochberg's False-Discovery Rate ($q = 0.10$); dotted green: Bonferroni ($\alpha = 0.05$); dashed grey: the nominal level 0.10. Shaded bands are 95% confidence intervals of the Monte Carlo mean ($p = 2{,}000$, $n_{\text{Monte}} = 300$). Draws in which DP-HCT does not reject the global null (HC statistic below 1.9) select no features and count as an FDR of zero. The figure shows that DP-HCT closely tracks the ideal threshold in the regime where detection is possible. These two rules are tuned for classification rather than for error control, and therefore admit many false discoveries. BH attains its nominal level and Bonferroni is conservative.

B Real-Data Experimental Details

B.1 Datasets

The 5 datasets are listed in Table 2.

Table 2: Five real-data benchmarks, their sizes, and additional relevant information. "Units" is the number of independent patients, participants, animals or arrays (equal to n when samples are independent).

Dataset	Description	n	p	C	Class sizes	Units
Heart-Circadian (GSE10045) [Bray et al., 2008]	Cardiac expression in wild-type and transgenic mice sampled every three hours. Classes are Zeitgeber times; genotypes and cardiac chambers are pooled. Units are animals.	130	46,120	8	19/17/16/ 16/16/16/ 16/14	66
Worm-Bacteria (GSE15923) [Coolon et al., 2009]	Two-color expression arrays from *C. elegans* grown on four bacterial environments, which define the classes. Units are arrays, each contributing two channel samples.	72	672	4	18/18/18/ 18	38
Blood-Diet (ST002019) [Ferguson, 2022]	Blood LC–MS metabolomics. Classes combine diet (pre/post) and study phase (1/2). Units are participants, each sampled at two visits.	70	736	4	18/18/17/ 17	35
Zebrafish-BPF (ST003804) [Guo, 2025]	Zebrafish liver LC–MS metabolomics after long-term bisphenol F exposure. Classes are four exposure groups; samples are independent fish.	62	117	4	16/16/15/ 15	62
Biliary-miRNA (GSE140001) [Loeffler et al., 2020]	miRNA expression in distal cholangiocarcinoma. Classes are three tissue types; each patient contributes one sample of each type and is a unit.	36	803	3	12/12/12	12

Preprocessing. We use the matrices provided by the repositories (GEO series matrices, Metabolomics Workbench data tables [Sud et al., 2016]). Features with fewer than 95% finite values are removed, remaining missing values are imputed by the feature median, $\log_2(x + 1)$ is applied when the matrix is non-negative with maximum above 100, and constant features are dropped. The values of p reported below are counted after these steps. No other preprocessing is applied beyond per-fold standardization inside the classifiers.

B.2 Cross-validation protocol

Classification accuracy results use stratified CV with $K = 10$ folds ($K = 7$ for Heart-Circadian, the largest number of folds that leaves at least one unit of every class in each fold). Table 3 reports averages over five CV runs with different seeds. Stress-test results use $K = 10$ folds with three repeats ($K = 7$ with a single repeat for Heart-Circadian, whose stress runs are the slowest).

We use stratified group folds when a patient, participant, animal or array contributes several samples, so that all samples of a unit are in the same fold and no unit appears in both a training and a test fold. Otherwise the repeated samples of a unit leak across the split and inflate the accuracy of both the baseline and the screened classifier. The units are listed in Table 2.

B.3 DP-HCT pipeline

We executes the following steps within each fold on training data only.

1. **Diversity p-values.** Compute the ANOVA F or Kruskal–Wallis statistic per feature on the training split, and convert it to a p-value π_j from the $F_{C-1,n-C}$ or χ^2_{C-1} reference distribution. p-values are clipped to $[10^{-300},\, 1 - 10^{-15}]$. The lower clip exists only to keep exact zeros out of the HC denominator and must lie below the smallest p-value the data produce.
2. **HC thresholding (HCT).** Obtain $j^\star$ of (2) and set $\hat{\mathcal{J}} = \{j : \pi_j \le \pi_{(j^\star)}\}$.
3. **Classifiers.** Two plug-in classifiers are evaluated on $\hat{\mathcal{J}}$:
   - *SVM*: RBF-kernel SVC (`sklearn`, `random_state=0`), features standardized with `StandardScaler`.
   - *Logistic*: ℓ_2-penalized multinomial logistic regression (`saga` solver).

Table 3: CV accuracy (mean ± SEM, %) for the five datasets, averaged over five CV runs with different seeds (the SEM is the mean within-run SEM across folds). The diversity test is given next to each dataset (F = ANOVA F, KW = Kruskal–Wallis); raw p-values; HC with $\gamma_0 = 1/2$. "Frac. sel." = mean fraction of features selected by HC.

Dataset	Method	SVM	LoR	Frac. sel.
Heart-Circadian (F)	Baseline	13.6±1.9	29.9±3.5	—
	DP-HCT	**69.3±5.5**	**75.2±4.7**	1.0%
Worm-Bacteria (KW)	Baseline	22.9±4.2	71.5±5.5	—
	DP-HCT	**55.2±6.1**	**80.3±5.6**	8.9%
Blood-Diet (F)	Baseline	41.0±5.0	58.2±5.3	—
	DP-HCT	**76.9±5.4**	**80.1±4.0**	14.6%
Zebrafish-BPF (KW)	Baseline	62.0±5.9	60.8±6.2	—
	DP-HCT	**70.1±5.0**	**71.3±5.5**	33.7%
Biliary-miRNA (KW)	Baseline	65.0±5.1	83.0±5.0	—
	DP-HCT	**91.0±4.5**	**98.7±1.1**	8.5%

B.4 Stress-test protocol

We perturb each of the five datasets by one of:

- *Noise injection* (σ sweep): independent $\mathcal{N}(0, \sigma^2)$ noise is added per fold to every element of X_{train} and X_{test} for $\sigma \in \{0, 0.5, 1.0, 1.5, 2.0\}$.
- *Nuisance feature inflation* (p_{extra} sweep): $p_{\text{extra}} \in \{0,\ 10^3,\ 2.5{\times}10^3,\ 5{\times}10^3,\ 10^4,\ 2.5{\times}10^4,\ 5{\times}10^4\}$ additional $\mathcal{N}(0, 1)$ columns are concatenated per fold.

Both perturbations occur after the train/test split and before feature selection, so HC is never exposed to test-fold data. Both perturbations are regenerated independently for each fold, as explained below in more detail. The results are illustrated in Fig. 6.

Noise injection. $X_{ij} \leftarrow X_{ij} + \sigma\, \hat{s}_j\, \varepsilon_{ij}$ with ε_{ij} i.i.d. $\mathcal{N}(0, 1)$ drawn fresh per fold, where $\hat{s}_j$ is the standard deviation of feature j on the training fold. Grid: $\sigma \in \{0.0,\ 0.5,\ 1.0,\ 1.5,\ 2.0\}$. The noise is scaled relative to each feature's own spread because the panels differ in units (the median feature

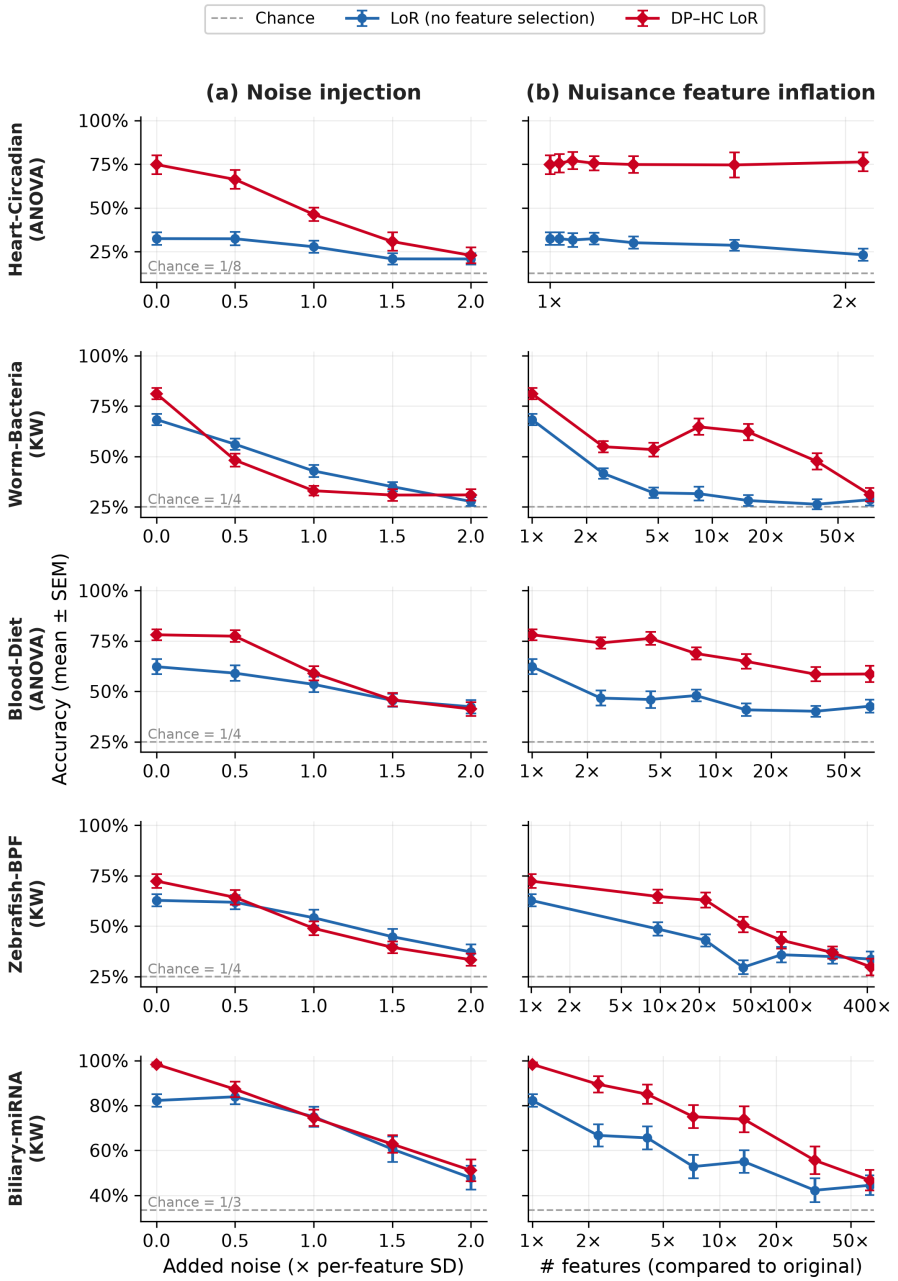


Figure 6: Classification with LoR pre- and post-feature screening by DP-HCT with added noise (left column) or nuisance features (right column). p-values are obtained either from ANOVA F (ANOVA) or Kruskal–Wallis (KW) as indicated for each panel; HC with $\gamma_0 = 1/2$; error bars indicate $\pm$ SEM over the pooled CV folds (10 folds $\times$ 3 repeats; 7 folds for Heart-Circadian); the uniform chance level $1/C$ is shown as a dashed horizontal line.

standard deviation ranges from 0.3 on Blood-Diet to 511 on Worm-Bacteria). Scales are estimated on the training fold only to prevent test information leakage.

Nuisance feature inflation. $X \leftarrow [X, G]$ where G has i.i.d. $\mathcal{N}(0,1)$ entries. Grid: $p_{\text{extra}} \in \{0,\ 10^3,\ 2.5\times10^3,\ 5\times10^3,\ 10^4,\ 2.5\times10^4,\ 5\times10^4\}$. Because the grid is absolute while p varies across panels, the resulting feature ratio $(p + p_{\text{extra}})/p$ depends on the size of the dataset.

C Notation and auxiliary facts

We collect the notation used throughout the proofs and record several auxiliary facts about the chi-squared distribution that will be invoked repeatedly.

C.1 Notation

Throughout, $p \to \infty$ indexes the asymptotic regime, and all probabilistic statements are understood in this limit unless stated otherwise. We write $a_p = O(b_p)$ if $|a_p|/|b_p|$ is eventually bounded, $a_p = o(b_p)$ if $a_p/b_p \to 0$, and $a_p \asymp b_p$ if both $a_p = O(b_p)$ and $b_p = O(a_p)$. Probabilistic versions $O_{\mathbb{P}}$, $o_{\mathbb{P}}$ are defined analogously in probability. For a set Q, we write $a_p(q) = o(b_p(q))$ *uniformly in* $q \in Q$ if $\sup_{q\in Q} |a_p(q)/b_p(q)| \to 0$.

We use C, K, c, c' to denote positive constants that may change from line to line but do not depend on p unless explicitly indicated.

C.2 Model and objects

We work under the Gaussian multiclass rare/weak model of Section 3.2. Thus

$$X_i \mid Y_i = c \sim \mathcal{N}(\mu_c, \sigma^2 I_p), \qquad n_c = n/C,$$

and the set of useful features $\mathcal{J} \subset [p]$ has rarity $\epsilon_p = p^{-\beta}$, $\beta \in (1/2, 1)$. On $\mathcal{J}$, the spherical homogeneous contrast construction (11)–(17) gives $\lambda_j = 2r \log p$ for $j \in \mathcal{J}$ and $\lambda_j = 0$ for $j \notin \mathcal{J}$. Class-c sample means are $\hat{\mu}_{c,j}$ as in (4), and

$$\hat{\mu}_{\cdot,j} := (\hat{\mu}_{1,j}, \dots, \hat{\mu}_{C,j})^\top$$

is the vector of those means at feature j. The model means are the superscript coordinates $\mu_c^{(j)}$ of $\mu^{(j)} = (\mu_1^{(j)}, \dots, \mu_C^{(j)})^\top$.

For the numerator-based analysis, define

$$T_j := \frac{\mathrm{SS}_{\text{between}}(j)}{\sigma^2}, \qquad \pi_j^\chi := \overline{F}_{\chi^2_k}(T_j), \qquad k := C - 1,$$

and

$$Y_j^\chi := -2 \log \pi_j^\chi = G_k(T_j), \qquad G_k(t) := -2 \log \overline{F}_{\chi^2_k}(t).$$

C.3 Chi-squared tail: quantitative expansion

The following one-sided tail expansion records useful auxiliary estimates. The exact exponent in Theorem 7 is proved separately by exponential tilting.

Lemma 3 (Chi-squared tail expansion). *Let $\overline{F}_k(t) := \mathbb{P}(\chi^2_k \geq t)$ denote the survival function of a central chi-squared random variable with $k \geq 1$ degrees of freedom, and define*

$$G_k(t) := -2 \log \bar{F}_k(t), \qquad c_k := (k-2)\log 2 + 2 \log \Gamma(k/2).$$

Fix $\eta \in (0,1)$. Uniformly over integers $k \geq 1$ and real numbers $t \to \infty$ satisfying

$$\frac{k}{t} \leq 1 - \eta,$$

we have

$$\overline{F}_k(t) = \frac{t^{(k-2)/2} e^{-t/2}}{2^{(k-2)/2}\Gamma(k/2)} \left[1 + \frac{k-2}{t} + O_\eta\left(\frac{(k+1)^2}{t^2}\right)\right]. \tag{21}$$

Consequently,

$$G_k(t) = t - (k-2)\log t + c_k + r_k(t), \tag{22}$$

where

$$r_k(t) = -2\log\left[1 + \frac{k-2}{t} + O_\eta\left(\frac{(k+1)^2}{t^2}\right)\right]. \tag{23}$$

In particular, if additionally $k/t = o(1)$, then

$$r_k(t) = -\frac{2(k-2)}{t} + O\left(\frac{(k+1)^2}{t^2}\right). \tag{24}$$

Moreover, in the same uniform regime,

$$G_k'(t) = 1 - \frac{k-2}{t} + O_\eta\left(\frac{(k+1)^2}{t^2}\right), \tag{25}$$

and

$$G_k''(t) = \frac{k-2}{t^2} + O_\eta\left(\frac{(k+1)^2}{t^3}\right). \tag{26}$$

Finally, as $k \to \infty$,

$$c_k = k\log k - k - \log k + O(1). \tag{27}$$

Proof. Write

$$s = \frac{k}{2}, \qquad x = \frac{t}{2}.$$

Then

$$\overline{F}_k(t) = \frac{\Gamma(s,x)}{\Gamma(s)},$$

where

$$\Gamma(s,x) = \int_x^\infty u^{s-1}e^{-u}\,du$$

is the upper incomplete gamma function.

We first prove the tail expansion. Integrating by parts once gives

$$\Gamma(s,x) = x^{s-1}e^{-x} + (s-1)\int_x^\infty u^{s-2}e^{-u}\,du.$$

Integrating by parts a second time gives

$$\int_x^\infty u^{s-2}e^{-u}\,du = x^{s-2}e^{-x} + (s-2)\int_x^\infty u^{s-3}e^{-u}\,du.$$

Therefore

$$\Gamma(s,x) = x^{s-1}e^{-x}\left[1 + \frac{s-1}{x} + \frac{(s-1)(s-2)}{x^2}\,\frac{\int_x^\infty u^{s-3}e^{-u}\,du}{x^{s-3}e^{-x}}\right].$$

We now bound the remainder. Since $k/t \le 1-\eta$, we have

$$\frac{s}{x} = \frac{k}{t} \le 1 - \eta.$$

For $u \ge x$,

$$u^{s-3}e^{-u} = x^{s-3}e^{-x}\left(\frac{u}{x}\right)^{s-3}e^{-(u-x)}.$$

Using the change of variables $u = x + v$,

$$\frac{\int_x^\infty u^{s-3}e^{-u}\,du}{x^{s-3}e^{-x}} = \int_0^\infty \left(1 + \frac{v}{x}\right)^{s-3}e^{-v}\,dv.$$

For $v \geq 0$ and $s/x \leq 1-\eta$,

$$\left(1+\frac{v}{x}\right)^{s-3} \leq \exp\left(\frac{sv}{x}\right) \leq \exp((1-\eta)v).$$

Hence

$$\frac{\int_x^\infty u^{s-3}e^{-u}\,du}{x^{s-3}e^{-x}} \leq \int_0^\infty e^{-\eta v}\,dv = \frac{1}{\eta}.$$

Substituting this remainder bound into the expansion of $\Gamma(s,x)$ gives

$$\Gamma(s,x) = x^{s-1}e^{-x}\left[1+\frac{s-1}{x}+O_\eta\left(\frac{(s+1)^2}{x^2}\right)\right],$$

uniformly in the stated regime. Dividing by $\Gamma(s)$ and substituting $s=k/2$, $x=t/2$, we obtain

$$\overline{F}_k(t) = \frac{(t/2)^{k/2-1}e^{-t/2}}{\Gamma(k/2)}\left[1+\frac{k-2}{t}+O_\eta\left(\frac{(k+1)^2}{t^2}\right)\right].$$

This proves (21).

Taking logarithms of (21),

$$\log\overline{F}_k(t) = \frac{k-2}{2}\log t - \frac{t}{2} - \frac{k-2}{2}\log 2 - \log\Gamma(k/2) + \log\left[1+\frac{k-2}{t}+O_\eta\left(\frac{(k+1)^2}{t^2}\right)\right].$$

Multiplying by -2 yields

$$G_k(t) = t - (k-2)\log t + (k-2)\log 2 + 2\log\Gamma(k/2) - 2\log\left[1+\frac{k-2}{t}+O_\eta\left(\frac{(k+1)^2}{t^2}\right)\right].$$

Since

$$c_k = (k-2)\log 2 + 2\log\Gamma(k/2),$$

this proves (22)–(23). If $k/t = o(1)$, then the logarithm may be expanded as $\log(1+u) = u+O(u^2)$, giving (24).

It remains to prove the derivative expansions. We use the hazard representation rather than differentiating the remainder in (21). The chi-squared density is

$$f_k(t) = \frac{1}{2^{k/2}\Gamma(k/2)}t^{k/2-1}e^{-t/2}.$$

Since $\bar{F}_k'(t) = -f_k(t)$,

$$G_k'(t) = -2\frac{\bar{F}_k'(t)}{\bar{F}_k(t)} = 2\frac{f_k(t)}{\bar{F}_k(t)}.$$

But (21) is equivalently

$$\bar{F}_k(t) = 2f_k(t)\left[1+\frac{k-2}{t}+O_\eta\left(\frac{(k+1)^2}{t^2}\right)\right].$$

Therefore

$$G_k'(t) = \left[1+\frac{k-2}{t}+O_\eta\left(\frac{(k+1)^2}{t^2}\right)\right]^{-1}.$$

Expanding the reciprocal gives

$$G_k'(t) = 1-\frac{k-2}{t}+O_\eta\left(\frac{(k+1)^2}{t^2}\right),$$

which proves (25).

For the second derivative, set $a=(k-2)/2$ and use the exact integral

$$H_k(t) := \frac{\overline{F}_k(t)}{2f_k(t)} = \frac{1}{2}\int_0^\infty (1+v/t)^a e^{-v/2}\,dv.$$

In the stated regime this integral is bounded above and below by positive constants depending only on η, and $H_k(t) = 1 + O_\eta((k+1)/t)$. Differentiation under the integral is justified by an exponentially decaying dominating function, because $a/t \le (1-\eta)/2$. Consequently,

$$H_k'(t) = -\frac{a}{2t^2}\int_0^\infty v(1+v/t)^{a-1}e^{-v/2}\,dv = -\frac{k-2}{t^2} + O_\eta\left(\frac{(k+1)^2}{t^3}\right).$$

For the remainder, apply the mean value theorem to $(1+v/t)^{a-1} - 1$ and integrate the bound $|a-1|(v/t)\exp((a-1)_+ v/t)$. Since $G_k' = H_k^{-1}$, it follows that

$$G_k''(t) = -H_k'(t)H_k(t)^{-2} = \frac{k-2}{t^2} + O_\eta\left(\frac{(k+1)^2}{t^3}\right),$$

proving (26).

Finally, Stirling's formula gives

$$\log\Gamma(z) = z\log z - z - \frac{1}{2}\log z + O(1), \qquad z \to \infty.$$

Taking $z = k/2$,

$$2\log\Gamma(k/2) = k\log(k/2) - k - \log(k/2) + O(1).$$

Thus

$$2\log\Gamma(k/2) = k\log k - k\log 2 - k - \log k + O(1).$$

Adding $(k-2)\log 2 = k\log 2 + O(1)$ yields

$$c_k = k\log k - k - \log k + O(1),$$

as claimed. □

Lemma 4 (Gaussian moderate-deviation bound). *Let $Z \sim \mathcal{N}(0,1)$ and $x_p \to \infty$. Then*

$$-\log\mathbb{P}(Z \ge x_p) \;=\; \frac{x_p^2}{2} + \log(x_p\sqrt{2\pi}) + O(x_p^{-2}).$$

In particular, if $x_p^2/(2\log p) \to \ell \in [0,\infty)$, then $\mathbb{P}(Z \ge x_p) = p^{-\ell+o(1)}$.

The proof is simply Mills' ratio.

D ANOVA F-statistic and reduction to the chi-square numerator

We derive the exact distribution of the one-way ANOVA F-statistic under the Gaussian model and explain why, on the moderate-deviation scale used in Theorem 7, it suffices to analyze the noncentral chi-square numerator.

Let

$$k := C - 1, \qquad m := n - C.$$

For feature j, define

$$T_j := \frac{\mathrm{SS}_{\mathrm{between}}(j)}{\sigma^2}, \qquad U_j := \frac{\mathrm{SS}_{\mathrm{within}}(j)}{\sigma^2}.$$

Thus

$$f_j = \frac{T_j/k}{U_j/m}.$$

Lemma 5 (Exact ANOVA distribution). *Under model (10), for each $j \in [p]$,*

$$T_j \sim \chi^2_k(\lambda_j), \qquad U_j \sim \chi^2_m, \qquad T_j \perp U_j,$$

where

$$\lambda_j = \frac{n_c}{\sigma^2}\mu^{(j)\top}P\mu^{(j)}, \qquad \mu^{(j)} := (\mu_1^{(j)},\dots,\mu_C^{(j)})^\top.$$

Consequently,

$$f_j \sim F_{k,m}(\lambda_j).$$

For $j \notin \mathcal{J}$, $\lambda_j = 0$, so $f_j \sim F_{k,m}$ and the F-test p-value is uniform under $H_0^{(j)}$. For $j \in \mathcal{J}$, under the spherical homogeneous calibration (17),

$$\lambda_j = 2r\log p.$$

Proof. The vector of class sample means at feature j satisfies

$$\hat{\mu}_{\cdot,j} = (\hat{\mu}_{1,j}, \dots, \hat{\mu}_{C,j})^\top \sim \mathcal{N}\left(\mu^{(j)}, \frac{\sigma^2}{n_c} I_C\right).$$

Since P is the orthogonal projection onto $\mathbf{1}^\perp$,

$$T_j = \frac{n_c}{\sigma^2} \hat{\mu}_{\cdot,j}^\top P \hat{\mu}_{\cdot,j}$$

is a noncentral chi-square random variable with $k = C - 1$ degrees of freedom and non-centrality

$$\lambda_j = \frac{n_c}{\sigma^2} \mu^{(j)\top} P \mu^{(j)}.$$

The within-class residual sum of squares is independent of the class sample means in the Gaussian one-way ANOVA model, and

$$U_j = \frac{\mathrm{SS}_{\mathrm{within}}(j)}{\sigma^2} \sim \chi^2_{n-C}.$$

Therefore $T_j \perp U_j$, and

$$f_j = \frac{T_j/k}{U_j/m} \sim F_{k,m}(\lambda_j).$$

Finally, if $j \in \mathcal{J}$, then by (17),

$$\lambda_j = 2r \log p.$$

If $j \notin \mathcal{J}$, then $P\mu^{(j)} = 0$, so $\lambda_j = 0$. □

Lemma 6 (Reduction from F p-values to the chi-square numerator). *Let π_j^F be the ANOVA F-test p-value at the feature j, as defined in (7) and let*

$$\pi_j^\chi := \mathbb{P}\{\chi_k^2 \geq T_j\}$$

be the oracle chi-square numerator p-value. Suppose

$$\log p = o(m), \qquad m = n - C.$$

Assume also $k/(2\log p) \to s \geq 0$, $\lambda_p/(2\log p) \to r > 0$, and use the continuous exponent $\Psi_{r,s}$ derived directly in Section E of the Supplementary Material. If, for q in a compact subset of $(0,1)$,

$$\mathbb{P}\big(\pi_j^\chi \leq p^{-q} \mid j \in \mathcal{J}\big) = p^{-\Psi_p(q)+o(1)}$$

uniformly in q, then

$$\mathbb{P}\big(\pi_j^F \leq p^{-q} \mid j \in \mathcal{J}\big) = p^{-\Psi_p(q)+o(1)}$$

uniformly in q.

Proof. Write

$$f_j = \frac{T_j/k}{R_j/m}, \qquad R_j \sim \chi^2_m, \qquad R_j \perp T_j.$$

Put $L = \log p$ and $\eta_p = (L/m)^{1/4}$, so that $\eta_p \to 0$. The chi-square concentration inequality gives

$$a_p := \mathbb{P}\{|R_j/m - 1| > \eta_p\} \leq 2\exp\{-c\eta_p^2 m\} = 2\exp\{-c\sqrt{mL}\} = p^{-\omega(1)}. \tag{28}$$

On the event $|R_j/m - 1| \leq \eta_p$,

$$\{T_j \geq 2Lx(1+\eta_p)\} \subseteq \{kf_j \geq 2Lx\} \subseteq \{T_j \geq 2Lx(1-\eta_p)\}.$$

Adding the exceptional denominator event gives the corresponding probability bounds. The central and noncentral numerator rates derived directly in Section E of the Supplementary Material are continuous in x, and the moving thresholds converge to x, so both bounds have the same exponent uniformly on compact q-intervals bounded away from zero. The additive error a_p is negligible compared with every fixed polynomial tail. Applying the same bracketing to the central quantiles transfers the numerator quantile to the central-F quantile and proves the claimed uniform expansion. □

E Moderate-deviation exponent of ANOVA p-values

For $x \geq 0$, define

$$J_{r,s}(x) = \sup_{0 \leq \theta < 1} \left\{ \theta x + s \log(1-\theta) - \frac{r\theta}{1-\theta} \right\}. \tag{29}$$

This upper-tail rate equals zero for $x \leq s + r$. For $x > s + r$, put $d = \sqrt{s^2 + 4rx}$ and $y = 2x/(s+d) > 1$; then

$$J_{r,s}(x) = x + r - d - s \log y. \tag{30}$$

For $s = 0$, this is $(\sqrt{x} - \sqrt{r})_+^2$. For $s > 0$, let $g_s(x) = x - s - s\log(x/s)$ on $x \geq s$, and let $x_s(q)$ be its increasing inverse; set $x_0(q) = q$. Write $L = \log p$. The p-value exponent is $\Psi_{r,s}(q) = J_{r,s}(x_s(q))$.

Theorem 7 (Uniform p-value tail exponent). *Suppose $k/(2L) \to s \geq 0$, $\lambda_p/(2L) \to r > 0$, and $m/L \to \infty$, with integer $k, m \geq 1$. Let $T \sim \chi^2_k(\lambda_p)$ and $R \sim \chi^2_m$ be independent. If π is the central-$F_{k,m}$ upper-tail p-value of $(T/k)/(R/m)$, then, for every fixed $Q < \infty$,*

$$\mathbb{P}(\pi \leq p^{-q}) = p^{-\Psi_p(q) + o(1)}, \qquad \Psi_p(q) := \Psi_{r,s}(q), \quad \textit{uniformly for } q \in [0, Q]. \tag{31}$$

The same conclusion holds for the numerator p-value $\pi^\chi = \mathbb{P}(\chi^2_k \geq T)$.

For later use, define the boundary exponent (the HC growth exponent under null-variance normalization)

$$E_p(q) := \frac{1}{2} - \beta + \frac{q}{2} - \Psi_p(q). \tag{32}$$

E.1 Noncentral numerator tails

Let $W_p = T/(2L)$. For $\theta < 1$,

$$\Lambda_p(\theta) := L^{-1} \log \mathbb{E} e^{L\theta W_p} = -\frac{k}{2L} \log(1-\theta) + \frac{\lambda_p}{2L} \frac{\theta}{1-\theta} \longrightarrow \Lambda(\theta) := -s\log(1-\theta) + \frac{r\theta}{1-\theta}.$$

The first two derivatives converge at each fixed $\theta < 1$. Chernoff's bound gives

$$\liminf_p -L^{-1} \log \mathbb{P}(W_p \geq x) \geq J_{r,s}(x).$$

For the matching lower bound when $x > s + r$, choose $\varepsilon > 0$ and the unique $\theta_\varepsilon \in (0,1)$ satisfying $\Lambda'(\theta_\varepsilon) = x + \varepsilon$; existence follows from $\Lambda'(0) = s + r$ and $\Lambda'(\theta) \uparrow \infty$. Tilt the law by

$$\frac{d\mathbb{P}_{\theta_\varepsilon}}{d\mathbb{P}} = \exp\{L\theta_\varepsilon W_p - L\Lambda_p(\theta_\varepsilon)\}.$$

Under this law, the mean of W_p tends to $x + \varepsilon$ and its variance is $\Lambda_p''(\theta_\varepsilon)/L \to 0$. Thus $A_\varepsilon = \{x \leq W_p \leq x + 2\varepsilon\}$ has tilted probability tending to one, and

$$\mathbb{P}(W_p \geq x) \geq e^{-L\theta_\varepsilon(x+2\varepsilon) + L\Lambda_p(\theta_\varepsilon)} \Pr_{\theta_\varepsilon}(A_\varepsilon).$$

Taking logarithms, then $p \to \infty$ and $\varepsilon \to 0$, proves the matching exponent. Below $s + r$, $W_p \to s + r$ in probability because its variance is $O(L^{-1})$, so the upper-tail exponent is zero. At $x = s + r$, bound the tail below by its value at $x + \varepsilon$ and let $\varepsilon \to 0$. This proves the rate at every $x \geq 0$.

The optimizer above the mean satisfies $x = s/(1-\theta) + r/(1-\theta)^2$. Writing $y = (1-\theta)^{-1}$ gives $x = sy + ry^2$, whose positive root is $2x/(s + \sqrt{s^2 + 4rx})$. Substitution proves (30). These formulas also show continuity of the rate.

E.2 Central numerator tails and quantiles

For $T_0 \sim \chi^2_k$ and $s > 0$, the preceding argument with $r = 0$ gives upper-tail rate zero for $x \leq s$ and $g_s(x)$ for $x > s$. When $s = 0$, for any fixed $x > 0$ and $0 < \theta < 1$, Chernoff's bound gives $\liminf -L^{-1} \log \mathbb{P}(T_0 \geq 2Lx) \geq \theta x$. Letting $\theta \uparrow 1$ gives the lower bound x on the exponent. Since $k \geq 1$, T_0 stochastically dominates a squared standard normal, whose tail at $2Lx$ is $p^{-x+o(1)}$. This gives the opposite bound. Thus the rate is $x = g_0(x)$ in this case.

For every fixed $q > 0$, the central chi-square upper quantile at probability p^{-q}, divided by $2L$, tends to $x_s(q)$. Indeed, evaluate the tail at $x_s(q) - \varepsilon$ and $x_s(q) + \varepsilon$, choosing the lower point above s when $s > 0$; strict monotonicity of g_s brackets the quantile. Applying the noncentral rate to these same brackets proves the numerator p-value exponent for fixed $q > 0$.

E.3 Transfer to exact F p-values

Let $D = R/m$, $\eta_p = (L/m)^{1/4} \to 0$, and $a_p = \mathbb{P}(|D - 1| > \eta_p)$. By (28),

$$a_p \le 2e^{-c\sqrt{mL}} = p^{-\omega(1)}.$$

For every fixed $x > 0$, the variable $V = T/D = kF$ satisfies

$$\mathbb{P}(T \ge 2Lx(1 + \eta_p)) - a_p \le \mathbb{P}(V \ge 2Lx) \le \mathbb{P}(T \ge 2Lx(1 - \eta_p)) + a_p.$$

The moving numerator thresholds have the same rate as x, by bracketing them between fixed $x \pm \varepsilon$ and using continuity. The additive error is negligible relative to every fixed polynomial tail. These statements hold for both central and noncentral numerators. Consequently, if $c_p(q)$ is the central-$F_{k,m}$ upper quantile at p^{-q}, then $kc_p(q)/(2L) \to x_s(q)$ for each fixed $q > 0$. Applying the noncentral bounds at this moving quantile yields (31) pointwise for $q > 0$. At $q = 0$, the event is $\{\pi \le 1\}$, with probability one, so its exponent is zero. We do not claim convergence of the scaled quantile to $x_s(0)$ at $q = 0$ when $s > 0$.

E.4 Uniformity and interpretation

The functions $f_p(q) = -L^{-1} \log \mathbb{P}(\pi \le p^{-q})$ are nondecreasing, and their pointwise limit $\Psi_{r,s}$ is continuous and nondecreasing. On $[0, Q]$, choose a finite partition so that the oscillation of the limit on each subinterval is at most ε. Pointwise convergence at its endpoints and monotonicity bound $|f_p - \Psi_{r,s}|$ by 2ε everywhere for all large p. This proves uniformity; the numerator argument is identical. All useful features have the same non-centrality, so feature-wise uniformity follows as well. This completes the proof of Theorem 7.

In particular, for $s > 0$,

$$\Psi_{r,s}(q) = 0 \quad \Longleftrightarrow \quad q \le g_s(s + r) = r - s \log(1 + r/s).$$

Local Gaussian approximations can describe fluctuations near this location, but do not determine the full tail exponent at fixed q. The tail-equivalence condition in Kipnis [2025, Eq. (2.5)] must be verified for a proposed shifted-square approximation; our direct calculation instead supplies the exact rate needed here.

E.5 Local expansion at the useful-feature mean

For $s > 0$, define

$$\bar{\mu} = r - s \log(1 + r/s), \qquad \bar{\nu}^2 = \left(\frac{r}{r + s}\right)^2 \frac{s + 2r}{2\bar{\mu}}.$$

These parameters describe a local expansion, not the full polynomial-tail rate. Indeed, $g_s'(s + r) = r/(s + r) > 0$, so as $q \to \bar{\mu}$ from above,

$$x_s(q) - (s + r) = \frac{s + r}{r}(q - \bar{\mu}) + O((q - \bar{\mu})^2).$$

The cumulant function satisfies $\Lambda''(0) = s + 2r$. Consequently,

$$J_{r,s}(s + r + h) = \frac{h^2}{2(s + 2r)} + O(h^3), \qquad h \to 0,$$

and therefore

$$\Psi_{r,s}(q) = \frac{(s + r)^2}{2r^2(s + 2r)}(q - \bar{\mu})^2 + O((q - \bar{\mu})^3) = \frac{(\sqrt{q} - \sqrt{\bar{\mu}})^2}{\bar{\nu}^2} + O((q - \bar{\mu})^3).$$

Below $\bar{\mu}$, both upper-tail expressions are zero when the positive part is used. The displayed expansion holds locally as $q \to \bar{\mu}$; it does not justify replacing $\Psi_{r,s}$ at other fixed thresholds.

F Proof of Theorem 1 (Impossibility below the boundary)

We prove the impossibility result under the spherical homogeneous contrast model. Let

$$\mathcal{V} = \mathbf{1}^{\perp} = \left\{ v \in \mathbb{R}^C : \mathbf{1}^\top v = 0 \right\}, \qquad P = I_C - \frac{1}{C}\mathbf{1}\mathbf{1}^\top.$$

Each feature is randomly set as useful independently with probability

$$\epsilon_p := \epsilon = p^{-\beta}, \qquad \beta \in (1/2, 1).$$

If $j \notin \mathcal{J}$, then $\mu^{(j)} = 0$. If $j \in \mathcal{J}$, then

$$\mu^{(j)} = a_p U_j, \qquad U_j \sim \mathrm{Unif}\left\{ u \in \mathcal{V} : \|u\|_2 = 1 \right\},$$

independently over j, where

$$a_p = \frac{\sigma}{\sqrt{n_c}} \sqrt{2r \log p}.$$

Thus every useful feature has ANOVA non-centrality

$$\lambda_j = \frac{n_c}{\sigma^2} \mu^{(j)\top} P \mu^{(j)} = 2r \log p.$$

For feature j, define the normalized ANOVA statistic

$$T_j = \frac{n_c}{\sigma^2} \hat{\mu}_{\cdot,j}^\top P \hat{\mu}_{\cdot,j},$$

and its null-calibrated p-value

$$\pi_j = \mathbb{P}\{\chi^2_{C-1} \geq T_j\}.$$

Let

$$Q_j = -\log_p \pi_j.$$

By Theorem 7, under a useful feature,

$$\mathbb{P}_{\mathrm{sig}}(Q_j \geq q) = p^{-\Psi_p(q) + o(1)}$$

uniformly on compact q-intervals, where

$$\Psi_p(q) := \Psi_{r,s}(q).$$

Here we have used the equivalence between $Q_j \geq q$ and $-2 \log \pi_j \geq 2q \log p$.

Recall the boundary exponent from (32),

$$E_p(q) = \frac{1}{2} - \beta + \frac{q}{2} - \Psi_p(q).$$

For the exact exponent, the below-boundary assumption is equivalently that the positive-exponent gap is strict. Hence there exist constants $\eta > 0$ and $a > 0$ such that

$$\sup_{0 \leq q \leq 1+\eta} E_p(q) \leq -a \tag{33}$$

for all sufficiently large p. We prove the result under the equivalent strict-below-boundary formulation (33).

Let $P_0^{(p)}$ and $P_1^{(p)}$ denote the null and rare/weak alternative laws of the training data $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$. The ancillary residual components have the same law under both models, so it is enough to control the likelihood ratio of the independent feature scores. We first prove

$$\mathrm{TV}(P_1^{(p)}, P_0^{(p)}) \to 0$$

for this training experiment. The additional test-observation argument is given after the truncated-moment calculation.

F.1 Reduction to the ANOVA p-value experiment

For each feature j, define

$$W_j = \frac{\sqrt{n_c}}{\sigma} P\hat{\mu}_{\cdot,j} \in \mathcal{V}.$$

Under the null,

$$W_j \sim N_{\mathcal{V}}(0, I_{\mathcal{V}}).$$

Under a useful feature,

$$W_j \mid U_j \sim N_{\mathcal{V}}(\theta_p U_j, I_{\mathcal{V}}), \qquad \theta_p = \sqrt{2r \log p},$$

where

$$U_j \sim \mathrm{Unif}\left\{u \in \mathcal{V} : \|u\|_2 = 1\right\}.$$

The within-class residuals and the grand mean across classes have the same distribution under the null and alternative and are independent of W_j: the mean contrast lies in $\mathcal{V}$, and the Gaussian projections onto $\mathcal{V}$ and its orthogonal complement are independent. Thus the feature-wise likelihood ratio depends only on W_j.

Under the spherical prior, the useful-feature density is

$$f_1(w) = \int_{\{u \in \mathcal{V}: \|u\|_2 = 1\}} \phi_{\mathcal{V}}(w - \theta_p u)\, d\sigma(u),$$

where $d\sigma$ is the uniform probability measure on the unit sphere in $\mathcal{V}$. Since $d\sigma$ is rotationally invariant, $f_1(w)$ depends only on $\|w\|_2$. The null density

$$f_0(w) = \phi_{\mathcal{V}}(w)$$

also depends only on $\|w\|_2$. Hence $f_1(w)/f_0(w)$ depends only on $\|w\|_2^2$. But

$$\|W_j\|_2^2 = \frac{n_c}{\sigma^2} \hat{\mu}_{\cdot,j}^\top P \hat{\mu}_{\cdot,j} = T_j.$$

Therefore the full likelihood ratio depends on feature j only through T_j, equivalently through π_j, equivalently through $Q_j = -\log_p \pi_j$. Thus we may work with the independent scores

$$Q_1, \ldots, Q_p.$$

F.2 One-feature likelihood ratios

Under the null, $\pi_j \sim \mathrm{Unif}(0, 1)$. Hence

$$\mathbb{P}_0(Q_j \geq q) = p^{-q},$$

and the null density of Q_j is

$$h_0(q) = (\log p) p^{-q}, \qquad q \geq 0.$$

Let h_{sig} denote the density of Q_j conditional on $j \in \mathcal{J}$, and define

$$\ell_p(q) = \frac{h_{\mathrm{sig}}(q)}{h_0(q)}.$$

This is the likelihood ratio of a useful feature versus a null feature at score q.

The rare/weak one-feature likelihood ratio is

$$L_j(q) = 1 - \epsilon_p + \epsilon_p \ell_p(q).$$

Since features are independent,

$$L_p = \frac{dP_1^{(p)}}{dP_0^{(p)}} = \prod_{j=1}^{p} L_j(Q_j).$$

F.3 A likelihood-ratio upper bound

We need a pointwise upper bound on $\ell_p(q)$. Let $T \sim \chi^2_k$ under the null and $T \sim \chi^2_k(\lambda_p)$ under a useful feature, with $k = C - 1$ and $\lambda_p = 2r \log p$. Let f_0 and f_1 be their densities, and define

$$R_p(t) = \frac{f_1(t)}{f_0(t)}.$$

The noncentral chi-square density has the Poisson mixture representation

$$f_1(t) = e^{-\lambda_p/2} \sum_{m=0}^{\infty} \frac{(\lambda_p/2)^m}{m!} f_{k+2m}(t).$$

Moreover,

$$\frac{f_{k+2m}(t)}{f_k(t)} = \frac{\Gamma(k/2)}{2^m \Gamma(k/2+m)} t^m.$$

Therefore

$$R_p(t) = e^{-\lambda_p/2} \sum_{m=0}^{\infty} a_m t^m,$$

where all coefficients satisfy $a_m \geq 0$. Hence $R_p(t)$ is nondecreasing in t.

The map $t \mapsto Q = -\log_p S_0(t)$, where $S_0(t) = \mathbb{P}(\chi^2_k \geq t)$, is increasing. Therefore $q \mapsto \ell_p(q)$ is nondecreasing. Hence, for any q,

$$\begin{aligned}
\mathbb{P}_{\text{sig}}(Q \geq q) &= \int_q^{\infty} \ell_p(u) h_0(u)\, du \\
&\geq \ell_p(q) \int_q^{\infty} h_0(u)\, du \\
&= \ell_p(q) \mathbb{P}_0(Q \geq q).
\end{aligned}$$

Thus

$$\ell_p(q) \leq \frac{\mathbb{P}_{\text{sig}}(Q \geq q)}{\mathbb{P}_0(Q \geq q)}.$$

Using Theorem 7 and the exact null tail,

$$\mathbb{P}_{\text{sig}}(Q \geq q) = p^{-\Psi_p(q)+o(1)}, \qquad \mathbb{P}_0(Q \geq q) = p^{-q},$$

we obtain

$$\ell_p(q) \leq p^{q-\Psi_p(q)+o(1)} \tag{34}$$

uniformly for $q \in [0, 1+\eta]$.

F.4 Truncation

Define

$$A_p = \left\{ \max_{1 \leq j \leq p} Q_j \leq 1 + \eta \right\}.$$

We first show that A_p is typical under both $P_0^{(p)}$ and $P_1^{(p)}$.

Under the null,

$$\mathbb{P}_0(Q_j > 1 + \eta) = p^{-(1+\eta)}.$$

By the union bound,

$$\mathbb{P}_0(A_p^c) \leq p \cdot p^{-(1+\eta)} = p^{-\eta} \to 0.$$

Under the alternative,

$$\begin{aligned}
\mathbb{P}_1(Q_j > 1 + \eta) &= (1 - \epsilon_p)\mathbb{P}_0(Q_j > 1 + \eta) + \epsilon_p \mathbb{P}_{\text{sig}}(Q_j > 1 + \eta) \\
&\leq p^{-(1+\eta)} + p^{-\beta - \Psi_p(1+\eta)+o(1)}.
\end{aligned}$$

Therefore,

$$\mathbb{P}_1(A_p^c) \leq p^{-\eta} + p^{1-\beta-\Psi_p(1+\eta)+o(1)}.$$

Since

$$E_p(1+\eta) = 1 - \beta + \frac{\eta}{2} - \Psi_p(1+\eta),$$

and by (33) $E_p(1+\eta) \le -a$, we have

$$1 - \beta - \Psi_p(1+\eta) \le -a - \frac{\eta}{2} < 0.$$

Thus

$$\mathbb{P}_1(A_p^c) \to 0.$$

Equivalently, since $L_p = dP_1^{(p)}/dP_0^{(p)}$,

$$\mathbb{E}_0[L_p \mathbf{1}_{A_p}] = \mathbb{P}_1(A_p) \to 1. \tag{35}$$

F.5 Truncated second moment

Because

$$L_p = \prod_{j=1}^{p} L_j(Q_j), \qquad A_p = \bigcap_{j=1}^{p} \{Q_j \le 1+\eta\},$$

we have, under $P_0^{(p)}$,

$$\mathbb{E}_0[L_p^2 \mathbf{1}_{A_p}] = \left(\mathbb{E}_0[L_1(Q)^2 \mathbf{1}_{\{Q \le 1+\eta\}}]\right)^p.$$

Now

$$L_1(q) = 1 - \epsilon_p + \epsilon_p \ell_p(q),$$

so

$$L_1(q)^2 = (1-\epsilon_p)^2 + 2\epsilon_p(1-\epsilon_p)\ell_p(q) + \epsilon_p^2 \ell_p(q)^2.$$

Using

$$\mathbb{P}_0(Q \le 1+\eta) \le 1$$

and

$$\mathbb{E}_0[\ell_p(Q)\mathbf{1}_{\{Q \le 1+\eta\}}] \le \mathbb{E}_0[\ell_p(Q)] = 1,$$

we get

$$\mathbb{E}_0[L_1(Q)^2 \mathbf{1}_{\{Q \le 1+\eta\}}] \le 1 + \epsilon_p^2 \mathbb{E}_0[\ell_p(Q)^2 \mathbf{1}_{\{Q \le 1+\eta\}}].$$

Moreover,

$$\begin{aligned}
\mathbb{E}_0[\ell_p(Q)^2 \mathbf{1}_{\{Q \le 1+\eta\}}] &= \int_0^{1+\eta} \ell_p(q)^2 h_0(q)\, dq \\
&\le \int_0^{1+\eta} p^{2q - 2\Psi_p(q) + o(1)} (\log p) p^{-q}\, dq \\
&= (\log p) \int_0^{1+\eta} p^{q - 2\Psi_p(q) + o(1)}\, dq.
\end{aligned}$$

The $o(1)$ term is uniform on the compact interval $[0, 1+\eta]$, so

$$\mathbb{E}_0[\ell_p(Q)^2 \mathbf{1}_{\{Q \le 1+\eta\}}] \le p^{\sup_{0 \le q \le 1+\eta}\{q - 2\Psi_p(q)\} + o(1)}.$$

Since $\epsilon_p = p^{-\beta}$,

$$\begin{aligned}
p\epsilon_p^2 \mathbb{E}_0[\ell_p(Q)^2 \mathbf{1}_{\{Q \le 1+\eta\}}] &\le p^{1 - 2\beta + \sup_{0 \le q \le 1+\eta}\{q - 2\Psi_p(q)\} + o(1)} \\
&= p^{2 \sup_{0 \le q \le 1+\eta} E_p(q) + o(1)} \\
&\le p^{-2a + o(1)} \to 0.
\end{aligned}$$

Therefore,

$$\epsilon_p^2 \mathbb{E}_0[\ell_p(Q)^2 \mathbf{1}_{\{Q \le 1+\eta\}}] = o(p^{-1}),$$

and hence

$$\mathbb{E}_0[L_1(Q)^2 \mathbf{1}_{\{Q \le 1+\eta\}}] \le 1 + o(p^{-1}).$$

Consequently,

$$\mathbb{E}_0[L_p^2 \mathbf{1}_{A_p}] \le (1 + o(p^{-1}))^p = 1 + o(1).$$

For the matching lower bound, Cauchy–Schwarz gives

$$\mathbb{E}_0[L_p^2 \mathbf{1}_{A_p}] \ge \left(\mathbb{E}_0[L_p \mathbf{1}_{A_p}]\right)^2.$$

By (35), the right-hand side tends to 1. Thus

$$\mathbb{E}_0[L_p^2 \mathbf{1}_{A_p}] \to 1. \tag{36}$$

F.6 Conclusion: total variation and classification

We now show that the truncated moment bounds imply total variation convergence. For any measurable event B,

$$P_1^{(p)}(B) - P_0^{(p)}(B) = \mathbb{E}_0[L_p \mathbf{1}_B] - P_0^{(p)}(B).$$

Decompose

$$\mathbb{E}_0[L_p \mathbf{1}_B] = \mathbb{E}_0[L_p \mathbf{1}_{A_p} \mathbf{1}_B] + \mathbb{E}_0[L_p \mathbf{1}_{A_p^c} \mathbf{1}_B].$$

Then

$$\begin{aligned} |P_1^{(p)}(B) - P_0^{(p)}(B)| \le \big|\mathbb{E}_0[(L_p \mathbf{1}_{A_p} - 1)\mathbf{1}_B]\big| \\ + \mathbb{E}_0[L_p \mathbf{1}_{A_p^c}] + P_0^{(p)}(A_p^c). \end{aligned}$$

The second term equals

$$\mathbb{E}_0[L_p \mathbf{1}_{A_p^c}] = P_1^{(p)}(A_p^c) \to 0,$$

and the third term also tends to zero. For the first term, Cauchy–Schwarz gives

$$\big|\mathbb{E}_0[(L_p \mathbf{1}_{A_p} - 1)\mathbf{1}_B]\big| \le \big(\mathbb{E}_0[(L_p \mathbf{1}_{A_p} - 1)^2]\big)^{1/2}.$$

But

$$\begin{aligned} \mathbb{E}_0[(L_p \mathbf{1}_{A_p} - 1)^2] &= \mathbb{E}_0[L_p^2 \mathbf{1}_{A_p}] - 2\mathbb{E}_0[L_p \mathbf{1}_{A_p}] + 1 \\ &\to 1 - 2 + 1 = 0, \end{aligned}$$

by (35) and (36). Hence

$$\sup_B |P_1^{(p)}(B) - P_0^{(p)}(B)| \to 0.$$

Therefore

$$\mathrm{TV}(P_1^{(p)}, P_0^{(p)}) \to 0.$$

The training-score calculation above proves indistinguishability for the training experiment. To include the test observation, augment the experiment by one independent observation Z_c from each class, conditional on the same latent means. The augmented design has class size $n_c + 1$ and non-centrality

$$\lambda_p^+ = 2r \log p \, (1 + 1/n_c).$$

Because $n_c \to \infty$, its normalized non-centrality has the same limit r; the preceding truncated-moment proof therefore gives total variation $o(1)$ for the augmented experiment as well. Now draw Y uniformly from $[C]$, independently of the augmented observations, and evaluate $\hat{h}(\mathcal{D}_n, Z_Y)$. Under the alternative this has exactly the law of the original test pair. Under the null, Y is independent of the observed data, so the correctness probability is $1/C$. Applying total variation to the correctness event gives, uniformly over classifiers,

$$\mathbb{P}_1\{\hat{h}(\mathcal{D}_n, Z_Y) = Y\} \le \frac{1}{C} + o(1).$$

A constant classifier attains $1/C$, completing the proof of Theorem 1.

G HC growth and HCT adaptivity

We analyze the HCT rule of Algorithm 1. For a threshold $t \in [0, 1]$, write

$$\mathrm{TP}(t) = \sum_{j \in \mathcal{J}} \mathbf{1}\{\pi_j \le t\}, \qquad \mathrm{FP}(t) = \sum_{j \notin \mathcal{J}} \mathbf{1}\{\pi_j \le t\}, \qquad N(t) = \mathrm{TP}(t) + \mathrm{FP}(t).$$

On the admissible set $\mathcal{A}_p = \{t : 1 \le N(t) \le \lfloor \gamma_0 p \rfloor\}$, define

$$H_p(t) = \frac{N(t) - pt}{\sqrt{N(t)(1 - N(t)/p)}}. \tag{37}$$

For pointwise probability statements set $H_p(t) = 0$ outside $\mathcal{A}_p$. The exact F p-values are continuous, so ties have probability zero and $H_p(\pi_{(j)}) = \mathrm{HC}_{p,j}$. Between consecutive p-values the count and denominator are constant, while the numerator decreases with t. Thus the algorithm maximizes (37)

over $\mathcal{A}_p$. Write $\hat{t} = \pi_{(j^\star)}$ for its chosen threshold, with deterministic tie-breaking if several scores attain the maximum.

For $q \in (0,1)$ define

$$A_p(q) = 1 - \beta - \Psi_p(q), \qquad B(q) = 1 - q, \qquad E_p(q) = A_p(q) - B(q)/2,$$

and the HC growth exponent for the empirical denominator in (37)

$$\widetilde{E}_p(q) = A_p(q) - \tfrac{1}{2}\max\{A_p(q), B(q)\} = \min\{E_p(q), A_p(q)/2\}. \tag{38}$$

Here $A_p(q)$ and $B(q)$ are the exponents of the expected useful and null selection counts. The denominator in (37) uses the observed fraction $N(t)/p$, rather than the null probability t. Where $E_p(q) > 0$, its exponent is $\frac{1}{2}\max\{A_p(q), B(q)\}$; subtracting this from the numerator exponent $A_p(q)$ gives $\widetilde{E}_p(q)$. Since $B(q) > 0$, $\widetilde{E}_p(q) > 0$ if and only if $E_p(q) > 0$. Thus HC grows as a positive power of p for the same parameter values that define the boundary in (20), although its growth rate and maximizing threshold can differ from those for the null denominator. The algorithm scans the ordered observed p-values; the deterministic grids below are used only to prove uniform bounds.

Theorem 8 (HC growth and HCT adaptivity). *Assume the spherical homogeneous model, feature-wise independence, the calibrations of Theorem 7, $\beta \in (1/2, 1)$, and fixed $\gamma_0 \in (0,1)$. Then:*

(a) *For every fixed $q \in (0,1)$ with $E_p(q) > 0$,*

$$H_p(p^{-q}) = p^{\widetilde{E}_p(q) + o_{\mathbb{P}}(1)}. \tag{39}$$

For arbitrary fixed $q \in (0,1)$, its positive part satisfies

$$[H_p(p^{-q})]_+ = O_{\mathbb{P}}\big(p^{\max\{\widetilde{E}_p(q), 0\} + o(1)}\big). \tag{40}$$

At a nonadmissible threshold set this positive part to zero; any fixed $q \in (0,1)$ is admissible with probability tending to one.

(b) *If $\sup_{0<q<1} E_p(q) > 0$, there is a fixed $d > 0$ such that*

$$\mathbb{P}\left\{\frac{\mathrm{TP}(\hat{t})^2}{N(\hat{t})} \geq p^d\right\} \longrightarrow 1. \tag{41}$$

In particular, after decreasing d if necessary, $\mathrm{TP}(\hat{t}) \geq p^d$ and $\mathrm{TP}(\hat{t})^2/(\mathrm{FP}(\hat{t}) + 1) \geq p^d$ with probability tending to one.

G.1 Count lemmas

In the following two lemmas, $\mathcal{Q}_p \subset (0,1)$ is any deterministic grid of polynomial cardinality. It is used for analysis, not to define HCT. Write $\mathrm{TP}_q = \mathrm{TP}(p^{-q})$ and $\mathrm{FP}_q = \mathrm{FP}(p^{-q})$ for exponent thresholds.

Lemma 9 (False positives). *Assume $\{\pi_j : j \in \mathcal{J}^c\}$ are i.i.d. $\mathrm{Uniform}[0,1]$ and the Bernoulli signal-set model of Section 3. Then, for every compact $Q \Subset (0,1)$ and every polynomial grid $\mathcal{Q}_p \subset Q$ with $|\mathcal{Q}_p| \leq p^a$,*

$$\mathrm{FP}_q = p^{1-q}(1 + o(1))$$

uniformly over $q \in \mathcal{Q}_p$, with probability $1 - o(1)$. Moreover,

$$\mathrm{FP}_q - p^{1-q} = O_{\mathbb{P}}(p^{(1-q)/2})$$

for each fixed $q \in Q$.

Proof. Unconditionally, FP_q is binomial with parameters p and $(1 - \epsilon_p)p^{-q}$. Its mean differs from p^{1-q} by $p^{1-\beta-q}$, which is $o(p^{(1-q)/2})$ since $\beta > 1/2$. Bernstein's inequality gives relative concentration uniformly on the stated grid because its smallest mean is polynomial in p. The variance bound and the negligible mean difference give the fixed-q statement. □

Lemma 10 (True positives). *Assume the conditions of Theorem 7, feature-wise independence, and condition on a signal set of size $|\mathcal{J}| = p^{1-\beta}(1+o(1))$, an event of probability $1-o(1)$. Then, uniformly over the deterministic grid,*

$$\mathbb{E}[\mathrm{TP}_q] = p^{1-\beta-\Psi_p(q)+o(1)}.$$

Furthermore, whenever $1-\beta-\Psi_p(q) > 0$ is bounded away from zero on a subset of the grid, the corresponding counts concentrate relatively uniformly:

$$\mathrm{TP}_q = p^{1-\beta-\Psi_p(q)+o_{\mathbb{P}}(1)}.$$

For all grid points, the upper bound

$$\mathrm{TP}_q = O_{\mathbb{P}}\big(p^{\max\{1-\beta-\Psi_p(q),0\}+o(1)}\big)$$

holds uniformly over $q \in \mathcal{Q}_p$.

Proof. Condition on the signal set $\mathcal{J}$. For $q \in \mathcal{Q}_p$, write

$$\mathrm{TP}_q = \sum_{j\in\mathcal{J}} X_j(q), \qquad X_j(q) := \mathbf{1}\{\pi_j \le p^{-q}\}.$$

By feature-wise independence, for each fixed q the variables $\{X_j(q) : j \in \mathcal{J}\}$ are independent Bernoulli variables. By Theorem 7,

$$\mathbb{P}(X_j(q)=1) = \mathbb{P}(\pi_j \le p^{-q} \mid j \in \mathcal{J}) = p^{-\Psi_p(q)+o(1)}$$

uniformly over $q \in \mathcal{Q}_p$. Hence, on the event $|\mathcal{J}| = p^{1-\beta}(1+o(1))$,

$$m_p(q) := \mathbb{E}[\mathrm{TP}_q \mid \mathcal{J}] = |\mathcal{J}|p^{-\Psi_p(q)+o(1)} = p^{1-\beta-\Psi_p(q)+o(1)}$$

uniformly over $q \in \mathcal{Q}_p$. This proves the expectation claim.

We next prove the relative concentration statement. Let $\mathcal{A}_p \subseteq \mathcal{Q}_p$ be a subset on which

$$1-\beta-\Psi_p(q) \ge \eta$$

for some fixed $\eta > 0$ and all large p. Then, uniformly over $q \in \mathcal{A}_p$,

$$m_p(q) \ge p^{\eta/2}$$

for all sufficiently large p. Bernstein's inequality for sums of independent Bernoulli variables gives, for any fixed $\varepsilon > 0$,

$$\mathbb{P}\left(|\mathrm{TP}_q - m_p(q)| \ge \varepsilon m_p(q) \,|\, \mathcal{J}\right) \le 2\exp\{-c_\varepsilon m_p(q)\}$$

for a constant $c_\varepsilon > 0$. Therefore

$$\mathbb{P}\left(\exists q \in \mathcal{A}_p : |\mathrm{TP}_q - m_p(q)| \ge \varepsilon m_p(q) \,|\, \mathcal{J}\right) \le 2|\mathcal{Q}_p|\exp\{-c_\varepsilon p^{\eta/2}\}.$$

Since $|\mathcal{Q}_p| \le p^a$, the right-hand side is $o(1)$. Hence

$$\sup_{q\in\mathcal{A}_p}\left|\frac{\mathrm{TP}_q}{m_p(q)} - 1\right| = o_{\mathbb{P}}(1).$$

Combining this with $m_p(q) = p^{1-\beta-\Psi_p(q)+o(1)}$ yields

$$\mathrm{TP}_q = p^{1-\beta-\Psi_p(q)+o_{\mathbb{P}}(1)}$$

uniformly over $q \in \mathcal{A}_p$. This proves the relative concentration statement.

It remains to prove the uniform upper bound over all grid points. Again by Bernstein's inequality, for every $t > 0$,

$$\mathbb{P}\left(\mathrm{TP}_q - m_p(q) \ge t \,|\, \mathcal{J}\right) \le \exp\left(-\frac{t^2}{2m_p(q)+2t/3}\right).$$

Take

$$t = K\{\sqrt{m_p(q)\log p} + \log p\}.$$

For K sufficiently large,

$$\mathbb{P}\left(\mathrm{TP}_q \geq m_p(q) + K\{\sqrt{m_p(q)\log p} + \log p\} \,\middle|\, \mathcal{J}\right) \leq p^{-a-2}.$$

A union bound over $|\mathcal{Q}_p| \leq p^a$ therefore gives, with probability $1 - o(1)$,

$$\mathrm{TP}_q \leq m_p(q) + K\sqrt{m_p(q)\log p} + K\log p$$

simultaneously for all $q \in \mathcal{Q}_p$.

Since

$$m_p(q) = p^{1-\beta-\Psi_p(q)+o(1)},$$

the preceding display implies

$$\mathrm{TP}_q = O_{\mathbb{P}}\left(p^{\max\{1-\beta-\Psi_p(q),0\}+o(1)}\right)$$

uniformly over $q \in \mathcal{Q}_p$. Indeed, when $1-\beta-\Psi_p(q) > 0$, the mean term dominates on the polynomial scale, while when $1 - \beta - \Psi_p(q) \leq 0$, all three terms are at most $p^{o(1)}$. This proves the lemma. □

Lemma 11 (Auxiliary grid and uniform counts). *Conditional on $\mathcal{J}$, put $m_J = |\mathcal{J}|$, $a_p(t) = \mathbb{P}(\pi_j \leq t \mid j \in \mathcal{J})$ and $m_T(t) = m_J a_p(t)$. There is a deterministic grid $\mathcal{T}_p \subset [0,1]$ of cardinality $O(p^4)$ such that, with probability $1 - o(1)$, each interval between successive grid points contains at most one observed p-value. Furthermore, with probability $1 - o(1)$, simultaneously for every $t \in [0,1]$,*

$$|\mathrm{FP}(t) - (p - m_J)t| \leq K\{\sqrt{pt\log p} + \log p\}, \tag{42}$$

$$|\mathrm{TP}(t) - m_T(t)| \leq K\{\sqrt{m_T(t)\log p} + \log p\}. \tag{43}$$

For each fixed $d > 0$, these bounds imply $m_T(t)/\mathrm{TP}(t) = 1 + o(1)$ uniformly over thresholds with $\mathrm{TP}(t) \geq p^d$.

Proof. The useful-feature F p-value distribution a_p is continuous. The function $G_p(t) = (t + a_p(t))/2$ is therefore continuous and strictly increasing from zero to one. Set $M_p = \lceil p^4 \rceil$ and

$$u_\ell = G_p^{-1}(\ell/M_p), \quad 0 \leq \ell \leq M_p, \qquad \mathcal{T}_p = \{u_0, \ldots, u_{M_p}\}.$$

Each cell $(u_{\ell-1}, u_\ell]$ has probability at most $2/M_p$ under either a null or a useful p-value distribution. Conditional on $\mathcal{J}$, two different features are independent, so the probability that their p-values share a cell is at most $2/M_p$. A union bound gives

$$\mathbb{P}(\text{some cell contains at least two p-values} \mid \mathcal{J}) \leq p(p-1)/M_p = o(1).$$

At each deterministic grid point, apply binomial Bernstein bounds to FP and TP with deviation parameter $K_0 \log p$. Taking K_0 sufficiently large makes a union bound over $O(p^4)$ points valid. This proves the two displayed bounds on $\mathcal{T}_p$. For an arbitrary threshold, move to its lower adjacent grid point. On the no-collision event each count changes by at most one; its mean changes by at most $2p/M_p$. Both changes are absorbed by the $K \log p$ terms, proving uniformity on $[0,1]$. Finally (43) implies $\mathrm{TP}(t) \leq 2m_T(t) + K' \log p$. If $\mathrm{TP}(t) \geq p^d$, then $m_T(t)$ is polynomially large, and dividing the same bound by $m_T(t)$ proves relative concentration uniformly in this region. The grid depends on model distributions but is used only in the proof; the algorithm does not construct it. □

G.2 Pointwise HC growth

Fix $q \in (0,1)$, put $t = p^{-q}$, and condition on the high-probability event $m_J = p^{1-\beta}(1 + o(1))$. The null count has mean $p^{B(q)}(1 + o(1))$ and fluctuation $O_{\mathbb{P}}(p^{B(q)/2})$. If $E_p(q) > 0$, then $A_p(q) > B(q)/2 > 0$, so the useful count concentrates:

$$\mathrm{TP}(t) = p^{A_p(q)+o_{\mathbb{P}}(1)}.$$

The identity

$$N(t) - pt = \mathrm{TP}(t) + \{\mathrm{FP}(t) - (p - m_J)t\} - m_J t$$

shows that the numerator has exponent $A_p(q)$. Indeed, the null fluctuation has smaller exponent, and $\Psi_p(q) < q$ makes $m_J t$ negligible. Moreover,

$$N(t) = p^{\max\{A_p(q), B(q)\} + o_{\mathbb{P}}(1)}, \qquad N(t)/p \to 0.$$

Substituting in (37) proves (39). If $E_p(q) \le 0$, then $A_p(q) \le B(q)/2 < B(q)$; the null count dominates $N(t)$, while the positive numerator is at most $p^{B(q)/2 + o_{\mathbb{P}}(1)}$ by the count bounds. The positive part of HC is thus at most subpolynomial, proving (40). Since $B(q) > 0$ and $A_p(q) \le 1 - \beta < 1$, each such fixed threshold is admissible with probability tending to one.

G.3 Adaptivity of the HCT threshold

Choose a fixed $q_0 \in (0, 1)$ with $E_p(q_0) > 0$, which exists above the boundary by continuity, including when the maximum over $[0, 1]$ is achieved at an endpoint. Part (a) gives some $a > 0$ such that $H_p(p^{-q_0}) \ge p^a$ with probability $1 - o(1)$. Let $j_0 = N(p^{-q_0})$. With high probability this is an admissible rank, and $\pi_{(j_0)} \le p^{-q_0}$. The denominator depends only on the count, so

$$H_p(\hat{t}) \ge H_p(\pi_{(j_0)}) \ge H_p(p^{-q_0}) \ge p^a.$$

On the uniform event in Lemma 11, consider any admissible t with $H_p(t) > 0$. Then $pt < N(t)$, and dropping the nonpositive missing-null term $-m_J t$ gives

$$\begin{aligned} H_p(t) &\le \frac{\mathrm{TP}(t) + K\{\sqrt{pt \log p} + \log p\}}{\sqrt{N(t)(1 - N(t)/p)}} \\ &\le \frac{\mathrm{TP}(t)}{\sqrt{(1 - \gamma_0) N(t)}} + K_{\gamma_0} \log p. \end{aligned}$$

Applying this inequality at $\hat{t}$ yields $\mathrm{TP}(\hat{t})/\sqrt{N(\hat{t})} \ge c p^a$ with high probability. Squaring and reducing the exponent proves (41). Finally $\mathrm{TP} \le N$ and $\mathrm{FP} + 1 \le 2N$ imply the stated count consequences. This argument covers the entire rank scan, including thresholds below $1/p$, without asserting that the selected exponent belongs to $(0, 1)$. This proves Theorem 8. □

H Boundary optimization

Proposition 12 (Explicit maximizer and boundary properties). *Write* $\Delta(\beta, r, s) = \max_{0 \le q \le 1} E_p(q)$. *For* $s \ge 0$, *let* $g_0(x) = x$, $x_0(1) = 1$, *and define*

$$x_1 = x_s(1), \qquad x_\dagger = 2r + \sqrt{s^2 + 4r^2}, \qquad x_* = \min\{x_1, x_\dagger\}.$$

Then

$$\Delta(\beta, r, s) = \frac{1}{2} - \beta + \frac{1}{2} g_s(x_*) - J_{r,s}(x_*). \tag{44}$$

For every $\beta \in (1/2, 1)$ *and fixed* $s \ge 0$, Δ *crosses zero at a unique positive* $r = \xi^\star(\beta, s)$, *with strict signs on either side. The boundary is strictly increasing in* s *and*

$$\xi^\star(\beta, 0) = \rho^\star(\beta), \quad \lim_{s \to 0} \xi^\star(\beta, s) = \rho^\star(\beta), \quad \xi^\star(\beta, s) \sim \sqrt{2 s \rho^\star(\beta)} \quad (s \to \infty).$$

Proof. Substitute $q = g_s(x)$ so that $s \le x \le x_1$. Below $s + r$, $J = 0$ and g_s is increasing. Above $s + r$, the envelope theorem in (29), or direct differentiation of (30), gives

$$\frac{d}{dx}\left(\tfrac{1}{2} g_s(x) - J_{r,s}(x)\right) = \frac{\sqrt{s^2 + 4rx} - x}{2x}.$$

This derivative changes sign once, at $x = x_\dagger > s + r$. The same formula holds for $s = 0$ above $x = r$, where $x_\dagger = 4r$. This proves (44).

For fixed s, as $r \to 0$ the useful rate tends to the null rate and $\Delta \to 1/2 - \beta < 0$. Once $s + r \ge x_1$, $\Delta = 1 - \beta > 0$. Continuity follows from the explicit formulas. At any r where $\Delta \le 0$, the optimizer lies strictly above $s + r$ (otherwise $x_1 \le s + r$ and $\Delta = 1 - \beta$). In that region, at fixed x, $\partial_r J = -(y - 1) < 0$. Increasing r thus strictly increases the optimized value until it becomes

positive; if the same x enters the zero-rate region, its rate has still strictly decreased. Monotonicity thereafter prevents a second crossing. This proves the existence, uniqueness and signs.

At $s = 0$, if $r \le 1/4$, the maximum is $\Delta = 1/2 - \beta + r$. If $1/4 < r < 1$, it is $\Delta = 1 - \beta - (1 - \sqrt{r})^2$; for $r \ge 1$ it is $1 - \beta$. Solving for its zero gives precisely the fixed-class boundary $\rho^\star(\beta)$ stated in Section 4.1 of the main paper.

For strict dependence on s, fix r, q in the positive-rate region and write $z = x_s(q)/s$ and $y = 2x_s(q)/(s + \sqrt{s^2 + 4rx_s(q)})$. Then $z > y > 1$. Implicit differentiation of $g_s(x_s(q)) = q$ gives

$$\partial_s x_s(q) = A(z), \qquad A(u) = \frac{\log u}{1 - 1/u}, \qquad A'(u) = \frac{u - 1 - \log u}{(u-1)^2} > 0.$$

The envelope theorem now yields

$$\partial_s \Psi_{r,s}(q) = (1 - 1/y)\{A(z) - A(y)\} > 0.$$

In the zero-rate region the rate is zero; its onset $g_s(s + r) = r - s\log(1 + r/s)$ decreases with s. Thus Ψ is nondecreasing in s everywhere and strictly increases once positive. At a boundary point every maximizing q has positive rate, by the optimizer argument above. If $s_2 > s_1$, evaluating at $r = \xi^\star(\beta, s_1)$ gives $E_{s_2}(q) \le E_{s_1}(q) \le 0$, with strict inequality at every zero of E_{s_1}. Compactness implies $\max_q E_{s_2}(q) < 0$, and hence $\xi^\star(\beta, s_2) > \xi^\star(\beta, s_1)$. The same conclusion with $s_1 = 0$ follows by continuity at zero and strict increase on positive s (apply it first between $s_2/2$ and s_2).

As $s \to 0$, $x_s(q) \to q$ uniformly on $[0, 1]$: $x_s(q) - q = s + s\log(x_s(q)/s)$ lies between s and its value at $q = 1$, which tends to zero. The explicit formula for J then gives uniform convergence of $\Psi_{r,s}$ to $(\sqrt{q} - \sqrt{r})_+^2$ for every fixed $r > 0$. Strict signs on each side of the classical boundary bracket the roots and prove their convergence.

Finally let $s \to \infty$ and set $r = a\sqrt{2s}$ for fixed $a > 0$. Taylor expansion of g_s yields, uniformly for $q \in [0, 1]$,

$$x_s(q) = s + \sqrt{2sq} + O(1).$$

For $x = x_s(q) > s + r$, the optimizer θ in (29) is $O(s^{-1/2})$ uniformly. Expanding at zero,

$$\theta x - \Lambda(\theta) = \theta(x - s - r) - \tfrac{1}{2}(s + 2r)\theta^2 + O((s + r)\theta^3).$$

Optimization, with the zero-rate case included, gives

$$\Psi_{a\sqrt{2s},s}(q) = \frac{(x_s(q) - s - r)_+^2}{2(s + 2r)} + o(1) = (\sqrt{q} - a)_+^2 + o(1)$$

uniformly on $[0, 1]$. The limiting boundary functional has its unique zero at $a^2 = \rho^\star(\beta)$, so bracketing again proves the claimed large-s equivalence. This is a limit of the fixed-s boundary, not a theorem for arbitrary joint sequences $s = s_p \to \infty$. □

I Proof of Theorem 2 (Misclassification rate of DP-HCT)

We prove the classification theorem under the spherical homogeneous contrast model. The proof analyzes the inner-product margin after HCT feature selection. Fix a true class $c^\star \in [C]$ and a competing class $c \ne c^\star$. For a useful feature j, the population contribution to the pairwise margin is

$$\mu_{c^\star}^{(j)}\big(\mu_{c^\star}^{(j)} - \mu_c^{(j)}\big) = \tau^2 v_{c^\star}^{(j)}\big(v_{c^\star}^{(j)} - v_c^{(j)}\big). \tag{45}$$

The spherical symmetry of $v^{(j)}$ gives positive average separation for every fixed class pair. Indeed,

$$\mathbb{E}\left[v_{c^\star}^{(j)}\big(v_{c^\star}^{(j)} - v_c^{(j)}\big)\right] = \frac{C}{C-1}. \tag{46}$$

Therefore, once HCT selects polynomially many true coordinates, the aggregate spherical contrast over the selected features produces a positive margin of order $\tau^2 \mathrm{TP}$, where

$$\mathrm{TP} := |\hat{\mathcal{J}} \cap \mathcal{J}|.$$

The rest of the proof shows that this positive population margin dominates both the training-noise terms introduced by estimating the centroids on the selected features and the variance contributed by selected null features.

I.1 Proof roadmap

Fix the true test label $c^\star$ and a competitor $c \neq c^\star$. The proof has seven steps.

(I.1) **Pairwise reduction.** Misclassification is contained in the union of the pairwise margin failures $\{\Delta_c \leq 0\}$.

(I.2) **Conditional Gaussianity.** Given the training data, latent means, and true test label, each margin Δ_c is Gaussian with conditional mean $m_c(\mathcal{D})$ and variance $v_c(\mathcal{D})$.

(I.3) **HCT count consequences.** HCT adaptivity makes $\mathrm{TP}^2/(\mathrm{TP}+\mathrm{FP})$ polynomially large. Consequently, for some fixed $d > 0$,

$$\mathrm{TP} \geq p^d, \qquad \frac{\mathrm{TP}^2}{\mathrm{FP}+1} \geq p^d$$

with probability $1 - o(1)$. We do not need, and generally do not have, $\mathrm{FP}/\mathrm{TP} \to 0$.

(I.4) **Overshoot control.** Selected null and selected signal ANOVA statistics are $O(\log p)$ uniformly. This keeps empirical selected contrasts on the rare-and-weak scale.

(I.5) **Spherical pairwise diversity.** For every fixed pair $c^\star \neq c$, spherical symmetry implies positive average pairwise separation:

$$\mathbb{E}\left[v_{c^\star}^{(j)}\left(v_{c^\star}^{(j)} - v_c^{(j)}\right)\right] = \frac{C}{C-1}.$$

Since HCT selects polynomially many true positives, these pairwise separations concentrate over $\hat{\mathcal{J}} \cap \mathcal{J}$, giving an aggregate population margin of order $\tau^2\mathrm{TP}$ uniformly over class pairs.

(I.6) **No-sample-splitting stability.** Although the same data are used for screening and centroid estimation, the selection-induced signal-noise term is bounded below by $-o_{\mathbb{P}}(\tau^2\mathrm{TP})$. The reason is that ANOVA selection is radial in the projected class-mean vector and does not create an aggregate negative drift large enough to cancel the spherical signal margin.

(I.7) **SNR divergence.** The conditional mean is at least order $\tau^2\mathrm{TP}$, while the conditional variance is bounded by a true-signal term plus a selected-null term. The first is dominated by polynomial growth of TP; the second is dominated by $\mathrm{TP}^2/(\mathrm{FP}+1) \to \infty$.

I.2 Reduction to pairwise margins

Recall the score classifier (12), $\hat{h}(x) = \arg\max_{a\in[C]} V_a(x)$ with

$$V_a(x) := \langle \hat{\mu}_a[\hat{\mathcal{J}}], x[\hat{\mathcal{J}}] \rangle.$$

For a test point X_{n+1} with true label $c^\star$, define the pairwise margin against competitor $c \neq c^\star$ by

$$\Delta_c := V_{c^\star}(X_{n+1}) - V_c(X_{n+1}) = \sum_{j\in\hat{\mathcal{J}}} \left(\hat{\mu}_{c^\star,j} - \hat{\mu}_{c,j}\right) X_{n+1,j}. \tag{47}$$

Then

$$\mathbb{P}\{\hat{h}(X_{n+1}) \neq c^\star\} \leq \sum_{c\neq c^\star} \mathbb{P}\{\Delta_c \leq 0\}. \tag{48}$$

We will prove a super-polynomial conditional error bound on one event that holds simultaneously for all class pairs with probability $1 - o(1)$. The complement of that event is counted once when removing the conditioning.

I.3 Conditional Gaussianity of pairwise margins

For this section, let $\mathcal{D}$ denote the sigma-field generated by the training data, the latent signal set, all latent class means, and the fixed true test label $c^\star$. This enlarged conditioning is only a proof device; the classifier observes the training data and test feature vector. Define

$$a_j(\mathcal{D}) := \hat{\mu}_{c^\star,j} - \hat{\mu}_{c,j}.$$

Lemma 13 (Conditional Gaussianity). *Conditional on $\mathcal{D}$, for each $c \neq c^\star$,*

$$\Delta_c \mid \mathcal{D} \sim \mathcal{N}\big(m_c(\mathcal{D}), v_c(\mathcal{D})\big), \tag{49}$$

where

$$m_c(\mathcal{D}) = \sum_{j \in \hat{\mathcal{J}}} a_j(\mathcal{D}) \mu_{c^\star}^{(j)}, \tag{50}$$

$$v_c(\mathcal{D}) = \sigma^2 \sum_{j \in \hat{\mathcal{J}}} a_j(\mathcal{D})^2. \tag{51}$$

Proof. Given $\mathcal{D}$, the coefficients $a_j(\mathcal{D})$ are fixed. Write $X_{n+1,j} = \mu_{c^\star}^{(j)} + \varepsilon_j^\star$, where the test noises $\varepsilon_j^\star$ are independent $\mathcal{N}(0, \sigma^2)$ and independent of $\mathcal{D}$. Substitution in (47) gives a deterministic mean plus a centered Gaussian sum with variance $\sigma^2 \sum_j a_j^2$. □

By Mills' bound, on the event $m_c(\mathcal{D}) > 0$,

$$\mathbb{P}(\Delta_c \leq 0 \mid \mathcal{D}) \leq \exp\left\{-\frac{m_c(\mathcal{D})^2}{2 v_c(\mathcal{D})}\right\}. \tag{52}$$

Thus the task is to prove $m_c(\mathcal{D})^2 / v_c(\mathcal{D}) \to \infty$ with high probability.

I.4 Count consequences of HCT adaptivity

Let $\hat{t} = \pi_{(j^\star)}$ be the HCT threshold of Algorithm 1, and write

$$\mathrm{TP} = |\hat{\mathcal{J}} \cap \mathcal{J}|, \qquad \mathrm{FP} = |\hat{\mathcal{J}} \cap \mathcal{J}^c|, \qquad N = \mathrm{TP} + \mathrm{FP}.$$

Lemma 14 (Selected counts). *If $\sup_{0<q<1} E_p(q) > 0$, there is a fixed $d > 0$ such that, with probability $1 - o(1)$,*

$$\mathrm{TP} \geq p^d, \tag{53}$$

$$\frac{\mathrm{TP}^2}{N} \geq p^d, \tag{54}$$

$$\frac{\mathrm{TP}^2}{\mathrm{FP} + 1} \geq p^d. \tag{55}$$

Proof. Theorem 8(b) proves the middle inequality with some positive exponent. Since $\mathrm{TP} \leq N$, $\mathrm{TP} \geq \mathrm{TP}^2/N$. Also $\mathrm{FP} + 1 \leq N + 1 \leq 2N$. These two inequalities prove the remaining conclusions after slightly decreasing the exponent to absorb constants. No restriction on a random selected exponent $-\log \hat{t} / \log p$ is required. □

I.5 Overshoot control

The next lemmas say that selected empirical ANOVA contrasts remain on the $\log p$ scale. This is not a weakness assumption; it is exactly what the rare-and-weak calibration $\lambda_j \asymp \log p$ predicts.

Lemma 15 (Null overshoot). *For any $B > 0$, there exists $M_0 < \infty$ such that*

$$\mathbb{P}\left(\max_{j \in \mathcal{J}^c} T_j > M_0 \log p\right) \leq p^{-B}.$$

Consequently, with probability $1 - o(1)$,

$$\max_{j \in \hat{\mathcal{J}} \cap \mathcal{J}^c} T_j \leq M_0 \log p.$$

Proof. For $j \in \mathcal{J}^c$, $T_j \sim \chi^2_k$. The standard Chernoff bound gives, for $x > k$,

$$\mathbb{P}(\chi^2_k \geq x) \leq \exp\left\{-\frac{1}{2}\big[x - k - k\log(x/k)\big]\right\}.$$

Since $k = O(\log p)$ in the regimes considered, taking $x = M_0 \log p$ with M_0 large enough gives probability at most $p^{-(B+2)}$. A union bound over at most p null coordinates proves the claim. □

Lemma 16 (Signal overshoot). *For any $B > 0$, there exists $M_1 < \infty$ such that*

$$\mathbb{P}\left(\max_{j\in\mathcal{J}} T_j > M_1 \log p\right) \leq p^{-B}.$$

Proof. For $j \in \mathcal{J}, T_j \sim \chi^2_k(\lambda_j)$ with $k = O(\log p)$, and under the spherical homogeneous calibration (11)–(17),

$$\lambda_j = 2r \log p$$

exactly. A noncentral chi-squared Chernoff bound states that, for $u > 0$,

$$\mathbb{P}\{\chi^2_k(\lambda) \geq k + \lambda + 2\sqrt{(k+2\lambda)u} + 2u\} \leq e^{-u}.$$

Taking $u = (B+2)\log p$ shows that $T_j \leq M_1 \log p$ with probability at least $1 - p^{-(B+2)}$, uniformly over $j \in \mathcal{J}$, for M_1 large enough. Since $|\mathcal{J}| \leq p$, a union bound proves the result. □

Corollary 17 (Selected empirical contrast bound). *With probability $1 - o(1)$, for every selected feature $j \in \hat{\mathcal{J}}$ and every pair $a, b \in [C]$,*

$$\big(\hat{\mu}_{a,j} - \hat{\mu}_{b,j}\big)^2 \leq \frac{4\sigma^2}{n_c} M \log p, \tag{56}$$

where $M := \max\{M_0, M_1\}$.

Proof. The deterministic contrast bound

$$(\hat{\mu}_{a,j} - \hat{\mu}_{b,j})^2 \leq \frac{4\sigma^2}{n_c} T_j$$

follows from Cauchy–Schwarz: $\hat{\mu}_{a,j} - \hat{\mu}_{b,j} = (e_a - e_b)^\top P\hat{\mu}_{\cdot,j}$ and $\|e_a - e_b\|^2 = 2$, while $T_j = (n_c/\sigma^2)\|P\hat{\mu}_{\cdot,j}\|^2$. Combining this deterministic inequality with Lemmas 15 and 16 gives (56). □

I.6 Spherical pairwise diversity

We next record the pairwise diversity properties of the spherical contrast directions. These properties are used to lower-bound the population part of the pairwise inner-product margin and to control the signal part of its conditional variance.

For $c^\star \neq c$, define

$$B_j^{(c^\star,c)} := v_{c^\star}^{(j)}\big(v_{c^\star}^{(j)} - v_c^{(j)}\big), \qquad D_j^{(c^\star,c)} := \big(v_{c^\star}^{(j)} - v_c^{(j)}\big)^2.$$

The variables $B_j^{(c^\star,c)}$ control the population pairwise margin, whereas $D_j^{(c^\star,c)}$ controls the signal part of the conditional variance.

Lemma 18 (Spherical pairwise diversity). *Assume the spherical homogeneous contrast model (11)–(17). Let*

$$\mathrm{TP} := |\hat{\mathcal{J}} \cap \mathcal{J}|.$$

On the event $\mathrm{TP} \geq p^d$ for some fixed $d > 0$, for every fixed pair $c^\star \neq c$,

$$\sum_{j\in\hat{\mathcal{J}}\cap\mathcal{J}} v_{c^\star}^{(j)}\big(v_{c^\star}^{(j)} - v_c^{(j)}\big) = \frac{C}{C-1}\,\mathrm{TP} + o_{\mathbb{P}}(\mathrm{TP}),$$

and

$$\sum_{j\in\hat{\mathcal{J}}\cap\mathcal{J}} \big(v_{c^\star}^{(j)} - v_c^{(j)}\big)^2 = \frac{2C}{C-1}\,\mathrm{TP} + o_{\mathbb{P}}(\mathrm{TP}).$$

In particular, with probability $1 - o(1)$*, uniformly over all* $c^\star \neq c$*,*

$$\sum_{j \in \hat{\mathcal{J}} \cap \mathcal{J}} v_{c^\star}^{(j)} \left(v_{c^\star}^{(j)} - v_c^{(j)} \right) \geq \frac{1}{2} \, \mathrm{TP},$$

and

$$\sum_{j \in \hat{\mathcal{J}} \cap \mathcal{J}} \left(v_{c^\star}^{(j)} - v_c^{(j)} \right)^2 \leq 4 \, \mathrm{TP}.$$

Moreover, for every fixed $c^\star$*,*

$$\sum_{j \in \hat{\mathcal{J}} \cap \mathcal{J}} \left(v_{c^\star}^{(j)} \right)^2 = \mathrm{TP} + o_{\mathbb{P}}(\mathrm{TP}),$$

and hence, with probability $1 - o(1)$*,*

$$\sum_{j \in \hat{\mathcal{J}} \cap \mathcal{J}} \left(v_{c^\star}^{(j)} \right)^2 \leq 2 \, \mathrm{TP}$$

uniformly over $c^\star \in [C]$*.*

Proof. Let

$$U \sim \mathrm{Unif} \left\{ u \in \mathbf{1}^\perp : \|u\|_2 = 1 \right\}, \qquad v = \sqrt{C} \, U.$$

By rotational symmetry on the $(C-1)$-dimensional subspace $\mathbf{1}^\perp$,

$$\mathbb{E}[UU^\top] = \frac{1}{C-1} P.$$

Therefore,

$$\mathbb{E}[vv^\top] = \frac{C}{C-1} P.$$

Since $P_{aa} = 1 - 1/C$ and $P_{ab} = -1/C$ for $a \neq b$, we get

$$\mathbb{E}[v_a^2] = \frac{C}{C-1} \left(1 - \frac{1}{C} \right) = 1,$$

and, for $a \neq b$,

$$\mathbb{E}[v_a v_b] = \frac{C}{C-1} \left(-\frac{1}{C} \right) = -\frac{1}{C-1}.$$

Thus, for $c^\star \neq c$,

$$\mathbb{E}\left[v_{c^\star} \left(v_{c^\star} - v_c \right) \right] = \mathbb{E}[v_{c^\star}^2] - \mathbb{E}[v_{c^\star} v_c] = 1 + \frac{1}{C-1} = \frac{C}{C-1},$$

and

$$\mathbb{E}\left[\left(v_{c^\star} - v_c \right)^2 \right] = \mathbb{E}[v_{c^\star}^2] + \mathbb{E}[v_c^2] - 2\mathbb{E}[v_{c^\star} v_c] = 2 + \frac{2}{C-1} = \frac{2C}{C-1}.$$

It remains to justify that these averages still concentrate after HCT selection. Conditional on $j \in \mathcal{J}$, the statistic used for selection is the ANOVA statistic. Under the spherical homogeneous calibration, its non-centrality is

$$\lambda_j = 2r \log p$$

for every direction $v^{(j)}$. Consequently, the entire conditional distribution of a useful feature's p-value depends on its spherical direction only through $\|v^{(j)}\|_2^2 = C$, which is fixed. By independence across features, the vector of useful directions is independent of the full vector of p-values, conditional on the signal set. This also applies to the adaptive HCT selection, which is a measurable function of that p-value vector. Hence, conditional on the selected true-positive set $\hat{\mathcal{J}} \cap \mathcal{J}$ and on its size TP, the directions

$$\{ v^{(j)} : j \in \hat{\mathcal{J}} \cap \mathcal{J} \}$$

remain iid draws from the spherical law.

For each fixed pair $c^\star \neq c$, the variables

$$B_j^{(c^\star,c)} = v_{c^\star}^{(j)}\big(v_{c^\star}^{(j)} - v_c^{(j)}\big)$$

are iid with finite moments of all orders and mean $C/(C-1)$. Therefore, conditional on $\mathrm{TP} \geq p^d$,

$$\frac{1}{\mathrm{TP}} \sum_{j \in \hat{\mathcal{J}} \cap \mathcal{J}} B_j^{(c^\star,c)} = \frac{C}{C-1} + o_{\mathbb{P}}(1).$$

Similarly,

$$\frac{1}{\mathrm{TP}} \sum_{j \in \hat{\mathcal{J}} \cap \mathcal{J}} D_j^{(c^\star,c)} = \frac{2C}{C-1} + o_{\mathbb{P}}(1),$$

and

$$\frac{1}{\mathrm{TP}} \sum_{j \in \hat{\mathcal{J}} \cap \mathcal{J}} \big(v_{c^\star}^{(j)}\big)^2 = 1 + o_{\mathbb{P}}(1).$$

The number of class pairs is at most C^2. In the regimes considered in Theorem 2, $C = O(\log p)$, while Lemma 14 gives that TP is polynomially large with probability $1 - o(1)$. Each displayed summand is bounded in absolute value by $4C$. Hoeffding's inequality bounds deviations of their averages by $2\exp\{-c_\varepsilon \mathrm{TP}/C^2\}$ for every fixed $\varepsilon > 0$. A union bound over $O(C^2)$ class pairs upgrades the fixed-pair convergence to uniform convergence over all $c^\star \neq c$. The displayed high-probability bounds follow immediately. □

I.7 One-sided stability of selected signal noise

The delicate point in the proof is that $\hat{\mathcal{J}}$ is selected using the same training noise that appears in the empirical centroids. We first record a symmetrization identity that controls this dependence, and then use it to prove the one-sided stability bound.

Lemma 19 (Symmetrization). *Let $Z_j \mid \mu^{(j)} \sim \mathcal{N}(0, (\sigma^2/n_c) I_C)$, let $P = I_C - C^{-1}\mathbf{1}\mathbf{1}^\top$ denote the orthogonal projection onto $\mathbf{1}^\perp$, and let*

$$A_j(q) := \big\{\, \|P\mu^{(j)} + PZ_j\|^2 \geq t_q \,\big\}$$

for a deterministic threshold $t_q > 0$. Then there exists a scalar $b_{j,q} \geq 0$ such that

$$\mathbb{E}\big[\, PZ_j \mathbf{1}_{A_j(q)} \mid \mu^{(j)} \,\big] = b_{j,q}\, P\mu^{(j)}. \tag{57}$$

Proof. All expectations are conditional on $\mu^{(j)}$; we suppress this from the notation. Write $W := PZ_j$ and $m := P\mu^{(j)}$, both vectors in the subspace $\mathcal{V} := \mathbf{1}^\perp$. Then W is a centered Gaussian on $\mathcal{V}$ with covariance $(\sigma^2/n_c)P$, and $A_j(q) = \{\|m + W\|^2 \geq t_q\}$ depends on W only through $\|m + W\|^2$.

Step 1: the conditional mean is parallel to m. If $m = 0$, then $A_j(q)$ depends only on $\|W\|^2$, so $\mathbb{E}[W\mathbf{1}_{A_j(q)}] = 0$ by symmetry of W, and the claim holds with $b_{j,q} = 0$. Assume $m \neq 0$.

Let R be any orthogonal transformation of $\mathcal{V}$ that fixes m, i.e., $Rm = m$. Then $\|m + RW\|^2 = \|R(m + W)\|^2 = \|m + W\|^2$, so $A_j(q)$ is R-invariant. Since the law of W on $\mathcal{V}$ is also R-invariant (rotational invariance of the Gaussian),

$$\mathbb{E}[W\mathbf{1}_{A_j(q)}] = \mathbb{E}[RW\mathbf{1}_{A_j(q)}] = R\,\mathbb{E}[W\mathbf{1}_{A_j(q)}].$$

Hence $v := \mathbb{E}[W\mathbf{1}_{A_j(q)}]$ is fixed by every orthogonal R on $\mathcal{V}$ with $Rm = m$. The stabilizer of m acts as the full orthogonal group on $m^\perp \cap \mathcal{V}$, so any vector fixed by all such R must lie in $\mathrm{span}(m)$. Therefore

$$\mathbb{E}[W\mathbf{1}_{A_j(q)}] = b_{j,q}\, m \qquad \text{for some scalar } b_{j,q} \in \mathbb{R}.$$

Step 2: the scalar is nonnegative. Decompose W along m and its orthogonal complement in $\mathcal{V}$:

$$W = X\hat{m} + W_\perp, \qquad \hat{m} := m/\|m\|, \quad X := \langle W, \hat{m}\rangle, \quad W_\perp \perp \hat{m}.$$

Under the Gaussian law, $X \sim \mathcal{N}(0, \sigma_*^2)$ with $\sigma_*^2 := \sigma^2/n_c$, and X is independent of $W_\perp$. Taking the inner product of Step 1 with $\hat{m}$ gives

$$b_{j,q} \|m\| = \mathbb{E}[X \mathbf{1}_{A_j(q)}],$$

so it suffices to show $\mathbb{E}[X \mathbf{1}_{A_j(q)}] \geq 0$.

Conditional on $W_\perp$, the event $A_j(q)$ becomes

$$\big\{(\|m\| + X)^2 + \|W_\perp\|^2 \geq t_q\big\}.$$

We consider two cases.

Case (i): $\|W_\perp\|^2 \geq t_q$. The event holds for all X, so the conditional contribution is $\mathbb{E}[X] = 0$.

Case (ii): $\|W_\perp\|^2 < t_q$. Setting $\rho := \sqrt{t_q - \|W_\perp\|^2} > 0$, the event becomes $\{|\|m\| + X| \geq \rho\}$, i.e.,

$$\{X \leq a\} \cup \{X \geq b\}, \qquad a := -\|m\| - \rho, \quad b := -\|m\| + \rho.$$

Note that $a + b = -2\|m\| \leq 0$, so $|a| \geq |b|$. Using the Gaussian truncation identities $\mathbb{E}[X \mathbf{1}_{\{X \leq a\}}] = -\sigma_*^2 \phi_{\sigma_*}(a)$ and $\mathbb{E}[X \mathbf{1}_{\{X \geq b\}}] = \sigma_*^2 \phi_{\sigma_*}(b)$, where ϕ_{σ_*} denotes the $\mathcal{N}(0, \sigma_*^2)$ density, the conditional contribution is

$$\sigma_*^2 \big(\phi_{\sigma_*}(b) - \phi_{\sigma_*}(a)\big) \geq 0,$$

because $|b| \leq |a|$ and ϕ_{σ_*} is decreasing in $|\cdot|$.

Both cases yield a nonnegative contribution, so integrating over $W_\perp$ gives $\mathbb{E}[X \mathbf{1}_{A_j(q)}] \geq 0$, hence $b_{j,q} \geq 0$. □

Remark. Geometrically, conditioning on the radial event $\{\|m + W\|^2 \geq t_q\}$ biases W along $+m$ rather than $-m$: reaching the threshold on the $+m$ side requires a smaller deviation than on the $-m$ side, while rotational symmetry kills all components of the conditional mean orthogonal to m.

For exact F p-values the threshold is random, but independent of the class-mean noise. Conditional on the residual sum of squares U_j, the event $\{\pi_j \leq p^{-q}\}$ is

$$\{\|P\mu^{(j)} + PZ_j\|^2 \geq (\sigma^2/n_c)\, k f_{k,m}^{-1}(1 - p^{-q})\, U_j/m\}.$$

Here $f_{k,m}^{-1}$ denotes the central F quantile function. Apply Lemma 19 at this threshold and average over U_j. Its conclusion remains valid with a nonnegative scalar depending only on the common signal norm, q, and the model parameters.

Lemma 20 (One-sided selected-signal stability). *For $a \neq b$ and $t \in [0, 1]$, let*

$$R_{a,b}(t) = \sum_{j \in \mathcal{J}} \mu_a^{(j)} (\bar{\varepsilon}_{a,j} - \bar{\varepsilon}_{b,j}) \mathbf{1}\{\pi_j \leq t\}, \qquad \bar{\varepsilon}_{a,j} = \hat{\mu}_{a,j} - \mu_a^{(j)}.$$

For every fixed $d > 0$, uniformly over all class pairs and all thresholds with $\mathrm{TP}(t) \geq p^d$,

$$R_{a,b}(t) \geq -o_{\mathbb{P}}(\tau^2 \mathrm{TP}(t)).$$

In particular, writing $R_c(t) = R_{c^\star,c}(t)$,

$$R_c(\hat{t}) \geq -o_{\mathbb{P}}(\tau^2 \mathrm{TP})$$

simultaneously for all ordered class pairs.

Proof. Condition on $\mathcal{J}$; expectations below average over both spherical directions and training observations, including residuals. For a fixed threshold t put

$$W_j^{a,b} = \mu_a^{(j)} (\bar{\varepsilon}_{a,j} - \bar{\varepsilon}_{b,j}), \qquad a_p(t) = \mathbb{P}(\pi_j \leq t \mid j \in \mathcal{J}), \qquad m_T(t) = |\mathcal{J}| a_p(t).$$

The extended symmetrization identity for exact F selection gives

$$\mathbb{E}[W_j^{a,b} \mathbf{1}\{\pi_j \leq t\} \mid \mu^{(j)}] = b_t \mu_a^{(j)} (\mu_a^{(j)} - \mu_b^{(j)}), \qquad b_t \geq 0.$$

The scalar is common to all directions because their norms are equal. Averaging over the spherical direction therefore yields

$$\mathbb{E}[W_j^{a,b} \mathbf{1}\{\pi_j \leq t\}] = b_t \tau^2 \frac{C}{C-1} \geq 0. \tag{58}$$

There is no additional selection indicator outside this expectation; nonnegativity is asserted after averaging over the direction.

Truncation and selected second moments. Put $L = \log p$, $\sigma_*^2 = \sigma^2/n_c$, and $\xi_j = PZ_j/\sigma_*$. Then $\|\xi_j\|^2 \sim \chi^2_{C-1}$ and $\sigma_*^2 = C\tau^2/(2rL)$. Choose a fixed M large enough that $B_j = \{\|\xi_j\|^2 \le ML\}$ satisfies $\mathbb{P}(B_j^c) \le p^{-12}$. Since $C = O(L)$, the chi-square tail bound permits this choice uniformly over features. On B_j, for every class pair,

$$|W_j^{a,b}| \le \tau\sqrt{C}\,\sqrt{2}\sigma_*\sqrt{ML} \le K_0\tau^2 C.$$

For $\widetilde{H}_j(t) = W_j^{a,b}\mathbf{1}_{B_j}\mathbf{1}\{\pi_j \le t\}$,

$$\mathbb{E}\widetilde{H}_j(t)^2 \le K_0^2\tau^4 C^2 a_p(t). \tag{59}$$

Also $\mathbb{E}(W_j^{a,b})^2 \le 2\tau^2 C\sigma_*^2$. Cauchy–Schwarz therefore bounds the lost first moment, uniformly in t, by $K\tau^2 Cp^{-6}$ (for all sufficiently large p). After summing over at most p features this is $o(\tau^2)$, and $\mathbb{P}(\bigcup_{j\in\mathcal{J}} B_j^c) = o(1)$.

Uniform concentration over deterministic thresholds. Use the auxiliary quantile grid $\mathcal{T}_p$ of Lemma 11. The truncated summands are independent across features conditional on $\mathcal{J}$. Bernstein's inequality, the selected second-moment bound, and a union bound over $O(p^4)$ grid points and $O(C^2)$ class pairs give, with probability $1 - o(1)$,

$$\sum_{j\in\mathcal{J}}(\widetilde{H}_j(t) - \mathbb{E}\widetilde{H}_j(t)) \ge -K_1\tau^2 C\{\sqrt{m_T(t)L} + L\}$$

simultaneously on the grid. The nonnegative untruncated mean in (58), the lost-mean bound, and the event that all B_j hold imply

$$R_{a,b}(t) \ge -K_1\tau^2 C\{\sqrt{m_T(t)L} + L\} - o(\tau^2)$$

at every grid point, simultaneously for all pairs.

Extension to all thresholds and adaptive selection. On the no-collision event of Lemma 11, moving any threshold to its lower adjacent grid point changes the sum by at most one truncated mark, whose absolute value is at most $K_0\tau^2 C$. This is absorbed by the $\tau^2 CL$ remainder. Since m_T is nondecreasing, the same lower bound, with an enlarged constant, holds simultaneously for every $t \in [0,1]$. On thresholds with $\mathrm{TP}(t) \ge p^d$, that lemma also gives $m_T(t) = \mathrm{TP}(t)(1 + o(1))$ uniformly. Dividing by $\tau^2\mathrm{TP}(t)$ bounds the negative relative error by

$$O\left(C\sqrt{\frac{L}{\mathrm{TP}(t)}} + \frac{CL}{\mathrm{TP}(t)}\right) + o(1) = o(1),$$

because $C = O(L)$ and $\mathrm{TP}(t) \ge p^d$. Lemma 14 puts the selected threshold in this region with probability $1 - o(1)$. Evaluation of the simultaneous bound at $\hat{t}$ proves the result without conditioning on the random threshold as though it were fixed or independent of the training noise. □

I.8 SNR divergence and conclusion

We now combine the preceding ingredients. Fix the true class $c^\star$ and a competitor $c \ne c^\star$. Recall from Lemma 13 that, conditional on the enlarged sigma-field $\mathcal{D}$, the pairwise margin Δ_c is Gaussian with mean

$$m_c(\mathcal{D}) = \sum_{j\in\hat{\mathcal{J}}} \widehat{d}_{c^\star c}(j)\mu_{c^\star}^{(j)}, \qquad \widehat{d}_{c^\star c}(j) := \hat{\mu}_{c^\star,j} - \hat{\mu}_{c,j},$$

and variance

$$v_c(\mathcal{D}) = \sigma^2 \sum_{j\in\hat{\mathcal{J}}} \widehat{d}_{c^\star c}(j)^2.$$

We show that

$$\frac{m_c(\mathcal{D})^2}{v_c(\mathcal{D})} \to \infty$$

in probability, uniformly over $c \ne c^\star$.

First decompose the conditional mean into its population part and its training-noise part:

$$m_c(\mathcal{D}) = \sum_{j\in\hat{\mathcal{J}}} \mu_{c^\star}^{(j)}\big(\mu_{c^\star}^{(j)} - \mu_c^{(j)}\big) + \sum_{j\in\hat{\mathcal{J}}} \mu_{c^\star}^{(j)}\big(\bar{\varepsilon}_{c^\star,j} - \bar{\varepsilon}_{c,j}\big).$$

Null coordinates have $\mu_{c^\star}^{(j)} = 0$, so only selected signals contribute. Therefore

$$m_c(\mathcal{D}) = \sum_{j\in\hat{\mathcal{J}}\cap\mathcal{J}} \mu_{c^\star}^{(j)}\big(\mu_{c^\star}^{(j)} - \mu_c^{(j)}\big) + R_c(\hat{t}).$$

Using $\mu^{(j)} = \tau v^{(j)}$, the population term is

$$\sum_{j\in\hat{\mathcal{J}}\cap\mathcal{J}} \mu_{c^\star}^{(j)}\big(\mu_{c^\star}^{(j)} - \mu_c^{(j)}\big) = \tau^2 \sum_{j\in\hat{\mathcal{J}}\cap\mathcal{J}} v_{c^\star}^{(j)}\big(v_{c^\star}^{(j)} - v_c^{(j)}\big).$$

By Lemma 18, with probability $1 - o(1)$, uniformly over all $c^\star \neq c$,

$$\sum_{j\in\hat{\mathcal{J}}\cap\mathcal{J}} v_{c^\star}^{(j)}\big(v_{c^\star}^{(j)} - v_c^{(j)}\big) \geq \frac{1}{2}\,\mathrm{TP}.$$

By Lemma 20,

$$R_c(\hat{t}) \geq -o_{\mathbb{P}}(\tau^2\mathrm{TP}).$$

Consequently,

$$m_c(\mathcal{D}) \geq \left(\frac{1}{2} - o_{\mathbb{P}}(1)\right)\tau^2\mathrm{TP} \gtrsim \tau^2\mathrm{TP}, \tag{60}$$

uniformly over $c \neq c^\star$.

We next bound the conditional variance. Write

$$\widehat{d}_{c^\star c}(j) = \big(\mu_{c^\star}^{(j)} - \mu_c^{(j)}\big) + \big(\bar{\varepsilon}_{c^\star,j} - \bar{\varepsilon}_{c,j}\big).$$

Using $(a+b)^2 \leq 2a^2 + 2b^2$,

$$\sum_{j\in\hat{\mathcal{J}}} \widehat{d}_{c^\star c}(j)^2 \leq 2\sum_{j\in\hat{\mathcal{J}}}\big(\mu_{c^\star}^{(j)} - \mu_c^{(j)}\big)^2 + 2\sum_{j\in\hat{\mathcal{J}}}\big(\bar{\varepsilon}_{c^\star,j} - \bar{\varepsilon}_{c,j}\big)^2.$$

For selected signals,

$$\mu_{c^\star}^{(j)} - \mu_c^{(j)} = \tau\big(v_{c^\star}^{(j)} - v_c^{(j)}\big),$$

and therefore Lemma 18 gives

$$\sum_{j\in\hat{\mathcal{J}}\cap\mathcal{J}}\big(\mu_{c^\star}^{(j)} - \mu_c^{(j)}\big)^2 = \tau^2 \sum_{j\in\hat{\mathcal{J}}\cap\mathcal{J}}\big(v_{c^\star}^{(j)} - v_c^{(j)}\big)^2 \leq 4\tau^2\mathrm{TP}$$

with probability $1 - o(1)$, uniformly over class pairs. For selected nulls, the population difference is zero.

It remains to control the empirical-noise part. By Corollary 17, selected empirical contrasts are on the rare-and-weak scale:

$$\sum_{j\in\hat{\mathcal{J}}\cap\mathcal{J}}\big(\bar{\varepsilon}_{c^\star,j} - \bar{\varepsilon}_{c,j}\big)^2 = O_{\mathbb{P}}\left(\frac{\sigma^2\log p}{n_c}\,\mathrm{TP}\right),$$

and

$$\sum_{j\in\hat{\mathcal{J}}\cap\mathcal{J}^c}\big(\bar{\varepsilon}_{c^\star,j} - \bar{\varepsilon}_{c,j}\big)^2 = O_{\mathbb{P}}\left(\frac{\sigma^2\log p}{n_c}\,\mathrm{FP}\right).$$

Since

$$\tau^2 = \frac{\sigma^2}{n}\, 2r \log p = \frac{\sigma^2}{n_c}\frac{2r \log p}{C},$$

and $C = O(\log p)$, the selected-signal empirical-noise contribution is

$$O_{\mathbb{P}}\left(\frac{\sigma^2 \log p}{n_c}\,\mathrm{TP}\right) = O_{\mathbb{P}}(C\tau^2 \mathrm{TP}).$$

Absorbing the harmless $C = O(\log p)$ factor into the subsequent SNR calculation, we obtain

$$v_c(\mathcal{D}) \le O_{\mathbb{P}}\big(\sigma^2 C\tau^2 \mathrm{TP}\big) + O_{\mathbb{P}}\left(\frac{\sigma^4 \log p}{n_c}\mathrm{FP}\right), \tag{61}$$

uniformly over $c \neq c^\star$.

Combining (60) and (61),

$$\frac{m_c(\mathcal{D})^2}{v_c(\mathcal{D})} \ge \frac{c\,\tau^4 \mathrm{TP}^2}{\sigma^2 C\tau^2 \mathrm{TP} + (\sigma^4 \log p/n_c)\mathrm{FP}}$$

for some constant $c > 0$, with probability $1 - o(1)$.

The first denominator term is negligible because

$$\frac{\tau^4 \mathrm{TP}^2}{\sigma^2 C\tau^2 \mathrm{TP}} = \frac{\tau^2 \mathrm{TP}}{\sigma^2 C} = \frac{2r \log p}{nC}\,\mathrm{TP}.$$

Since $n = p^{o(1)}$, $C = O(\log p)$, and Lemma 14 gives $\mathrm{TP} \ge p^\delta$ with probability $1 - o(1)$ for some $\delta > 0$, this ratio diverges.

For the selected-null denominator term, enlarge FP to FP $+ 1$ to avoid division by zero:

$$\frac{\tau^4 \mathrm{TP}^2}{(\sigma^4 \log p/n_c)(\mathrm{FP}+1)} = \frac{((\sigma^2/n)2r \log p)^2 \mathrm{TP}^2}{(\sigma^4 \log p/n_c)(\mathrm{FP}+1)} = \frac{4r^2 \log p}{nC}\frac{\mathrm{TP}^2}{\mathrm{FP}+1}.$$

By Lemma 14,

$$\frac{\mathrm{TP}^2}{\mathrm{FP}+1} \to \infty$$

polynomially, while $n = p^{o(1)}$ and $C = O(\log p)$. Hence this ratio also diverges.

Therefore, for some fixed $d' > 0$, there is an event $\mathcal{G}_p$ with $\mathbb{P}(\mathcal{G}_p) \to 1$ on which all ordered class pairs satisfy

$$m_c(\mathcal{D}) > 0, \qquad \frac{m_c(\mathcal{D})^2}{v_c(\mathcal{D})} \ge p^{d'}.$$

The polynomial lower bound, rather than mere divergence, follows from the selected-count exponents and $n = p^{o(1)}$. If a conditional variance is zero, the positive margin is deterministic. Otherwise conditional Gaussianity and a union bound over competitors give

$$\mathbb{P}\{\hat{h}(X_{n+1}) \neq c^\star\} \le \mathbb{P}(\mathcal{G}_p^c) + (C-1)e^{-p^{d'}/2} = o(1).$$

The exceptional event is counted once, not once per competitor. Uniformity over the true class gives the unconditional result and proves Theorem 2.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction explains the main addition of our paper and clearly states what we prove in the manuscript.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sections 3.2 and 4 specify the Gaussian, balanced-design and asymptotic setting. Unbalanced designs and other sample-size regimes are explicitly left for future work; these statements limit the scope of the theoretical guarantees.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Proofs are supplied in Appendices C–I. All assumptions are well stated in the theorems and proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 2, Section 5, and Appendices A–B describe the algorithm, simulation model, dataset accessions, preprocessing and evaluation protocols.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The public dataset sources and accession identifiers are given in Section 5.2 and Appendix B.1. The authors plan to make a code repository available.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Appendix B gives grouped cross-validation, fold counts, repeats, diversity tests, classifier families and the logistic-regression solver.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Tables 1 and 3 report means and SEMs, with the averaging procedure described in Appendix B; Figure 6 identifies SEMs over pooled folds. The additional simulation figures identify Monte Carlo confidence intervals where displayed; these summaries are not claims of independent-fold hypothesis tests.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The manuscript reports experiment sizes and repetition counts, but does not report hardware specifications, memory requirements, per-run execution times or total compute consumption.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The authors confirm that the research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper focuses on foundational classification theory and benchmark evaluation. It does not include a dedicated discussion of potential positive and negative societal impacts or deployment risks.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The manuscript does not release high-risk pretrained models, scraped datasets or other assets requiring the release safeguards described in this question.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The benchmark studies and repositories are credited in Section 5.2 and Appendix B.1, and the authors confirm that the applicable terms have been respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new dataset or pretrained-model release is described in the manuscript. The contribution is a feature-screening method and its theoretical and empirical analysis.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The reported experiments reanalyze existing public benchmark data; no new participant recruitment, crowdsourcing tasks, participant instructions or compensation are described for this work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This is a theory paper and this is not applicable.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: No LLM were used besides checking for typos in the spelling.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.