A SHARP PHASE TRANSITION FOR MLE EXISTENCE IN MULTINOMIAL LOGISTIC REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper studies when multinomial logistic regression admits a finite maximum-likelihood estimator. Under a Gaussian feature model with softmax class probabilities and a fixed number of classes, we establish a sharp phase transition for MLE existence. As the sample size and feature dimension grow proportionally, the probability that the MLE is finite converges to zero or one depending on whether the sample-to-dimension ratio lies below or above a critical threshold. We characterize this threshold through a matrix Gaussian-distance optimization whose dimension depends only on the number of classes, rather than on the feature dimension. The result allows for general feature covariance and signal strength, and applies to models fitted both with and without an intercept. Although fitting an intercept adds only finitely many parameters, it can strictly increase the asymptotic existence threshold, even when the true intercept is zero. Our characterization recovers the binary logistic threshold and yields explicit thresholds when labels are uniform and independent of the features. Simulations support the predicted thresholds across signal strengths and true intercepts, and show that estimating an intercept can lead to a higher threshold.

1 INTRODUCTION

Multinomial logistic regression is a standard model for multiclass classification, but its likelihood need not attain a maximum at finite parameters, even when the number of observations exceeds the number of features. In that case, continued loss reduction can accompany diverging coefficients rather than convergence to a finite estimate. Distinguishing this failure of attainment from a large but finite estimate is therefore necessary for interpreting a softmax model fitted by maximum likelihood. We ask how many observations per feature are needed for a finite MLE, and how this requirement depends on the class probabilities and regression coefficients. We answer this question for Gaussian features when sample size and feature dimension grow proportionally, with a fixed number of classes.

Classical results relate finite-MLE existence to separation. With a fixed baseline class and a full-rank design augmented by a constant column, the MLE fails to exist if and only if there is a nonzero parameter direction along which the linear predictor of every observed class increases at least as fast as those of its competitors (Silvapulle, 1981; Albert & Anderson, 1984; Santner & Duffy, 1986). In the binary setting, Candès & Sur (2020) turned this sample-level criterion into a sharp threshold in the sample-to-dimension ratio for Gaussian logistic regression, depending on signal strength and the true intercept. Related binary results establish existence transitions for elliptical covariates (Tang & Ye, 2020) and finite-sample guarantees for MLE existence and predictive risk (Chardon et al., 2024). More recently, Asgari et al. (2025, Proposition 2) studied multinomial logistic regression with zero true intercept, no fitted intercept, and a positive-definite limiting signal Gram matrix. Their result determines whether an MLE of bounded norm exists with high probability, but leaves open whether a finite MLE can exist whose norm grows with dimension. We establish a threshold for finite existence itself by proving separation below it and ruling out separation above it.

Our contributions. We establish a sharp phase transition for finite-MLE existence in multinomial logistic regression with a fixed number of classes and Gaussian features. As the sample size and feature dimension grow proportionally, the probability of a finite MLE tends to zero below a critical sample-to-dimension ratio and to one above it, without imposing a dimension-independent bound

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

on the estimator norm (Theorems 2.3 and 2.5). We characterize the threshold by a matrix Gaussian-distance optimization whose dimension depends only on the number of classes; the true parameters enter through the fixed intercept and the limiting covariance of the log-odds. The theory allows general positive-definite feature covariance and covers both models fitted with and without an intercept, including a nonzero true intercept in either case. We show that estimating an intercept can strictly increase the threshold even when the true intercept is zero (Proposition 3.1), recover the binary logistic threshold, and obtain explicit thresholds under the null multinomial model (Section 3.1). Simulations for three and four classes support the predicted transitions across signal strengths and true intercepts, and a paired comparison checks the effect of fitting an intercept under the same data law (Section 4).

The multinomial setting introduces complications absent in the binary problem. Pairwise predictor differences must come from a common set of class coefficients, so separately constructed binary separators need not be compatible. Nonexistence also includes weak separation, where the observed class ties with competitors. Our argument keeps the class predictors together and controls their adjustments through a matrix second-moment constraint. This retains the dependence between class comparisons that separate binary or scalar constraints would lose. Below the threshold, we turn an approximate separator into an exact one; above it, we construct a certificate excluding every separating direction.

Related work establishes phase transitions for the existence of unregularized M-estimators in single-index models (Koriyama & Bellec, 2025), for one-vs-one and one-vs-all separability in multiclass Gaussian mixtures (Kini & Thrampoulidis, 2021), and for multiclass logistic MLE existence under a Gaussian-mixture model (Loureiro et al., 2021). In the model studied by Loureiro et al. (2021), features are Gaussian conditional on class; in the present paper, features are marginally Gaussian and labels follow a conditional softmax model. Other multiclass results study stochastic convex-hull criteria for separability (De Loera & Hogan, 2020) and interpolation and benign overfitting under sufficient overparameterization (Wang et al., 2021). High-dimensional inference has been developed for binary logistic regression (Sur & Candès, 2019). For multinomial logistic regression, Tan & Bellec (2023) establish asymptotic normality and chi-square inference for null covariates under assumptions that include finite-MLE existence. Complementing this work, we characterize finite-MLE existence through a sharp phase transition under the Gaussian multinomial logistic model studied here. Conditional-inference approaches to diverging parameter estimates have also been developed for multinomial and logistic regression (Kolassa, 2016) and proportional hazards regression (Kolassa & Zhang, 2023).

2 Model and Phase Transitions

Notation. We write $\boldsymbol{I}_m$ for the $m \times m$ identity matrix and $\mathbf{1}_m$ for the m-dimensional vector of ones. For symmetric matrices, $\succeq \mathbf{0}$ and $\succ \mathbf{0}$ denote positive semidefiniteness and positive definiteness, respectively. We denote by $\mathbb{S}_+^m$ the cone of symmetric positive-semidefinite $m \times m$ matrices and by tr the matrix trace. All matrix square roots are taken to be symmetric positive semidefinite.

2.1 Multinomial Logistic Model

Fix the number of classes $K \geq 2$ and write $q = K - 1$. We observe independent pairs $(\boldsymbol{x}_i, Y_i)$, $i = 1, \ldots, n$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the feature vector and $Y_i \in \{0, \ldots, q\}$ is the corresponding class label. We take class 0 as the baseline and set its linear predictor to $\eta_{i0} = 0$. For the remaining q classes, let $\boldsymbol{\Theta}_0 \in \mathbb{R}^{d \times q}$ and $\boldsymbol{b}_0 \in \mathbb{R}^q$ denote the true coefficient matrix and intercept vector, respectively. The corresponding linear predictors are given by

$$\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iq})^\top = \boldsymbol{\Theta}_0^\top \boldsymbol{x}_i + \boldsymbol{b}_0.$$

The labels follow the softmax law

$$\mathbb{P}(Y_i = j \mid \boldsymbol{x}_i) = \frac{e^{\eta_{ij}}}{\sum_{k=0}^{q} e^{\eta_{ik}}}, \qquad j = 0, \ldots, q. \tag{1}$$

Thus, η_{ij} is the log-odds of class j relative to class 0. Fixing a baseline removes the invariance to adding a common predictor to all classes; it imposes no ordering or equal-probability assumption.

For candidate parameters $\boldsymbol{\Theta} \in \mathbb{R}^{d\times q}$ and $\boldsymbol{b} \in \mathbb{R}^q$, write $\boldsymbol{v} = \boldsymbol{\Theta}^\top \boldsymbol{x} + \boldsymbol{b}$ and set $v_0 = 0$. The negative log-likelihood for label y is

$$\ell_y(\boldsymbol{v}) = \log\left(1 + \sum_{j=1}^{q} e^{v_j}\right) - v_y.$$

The model is fitted by minimizing the empirical average negative log-likelihood

$$L_n(\boldsymbol{\Theta}, \boldsymbol{b}) = \frac{1}{n}\sum_{i=1}^{n} \ell_{Y_i}\left(\boldsymbol{\Theta}^\top \boldsymbol{x}_i + \boldsymbol{b}\right), \qquad (\boldsymbol{\Theta}, \boldsymbol{b}) \in \mathbb{R}^{d\times q} \times \mathbb{R}^q. \tag{2}$$

Throughout the paper, we say that the MLE exists if the likelihood attains its maximum at finite parameter values in the fitted model under consideration, or equivalently, if the empirical loss attains its minimum at finite parameters. The fitted model contains dq slope coefficients and q intercepts. We measure the sample size relative to the feature dimension through the ratio n/d. The feature covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, the true coefficient matrix $\boldsymbol{\Theta}_0 \in \mathbb{R}^{d\times q}$, and the induced log-odds covariance $\boldsymbol{S} = \boldsymbol{\Theta}_0^\top \boldsymbol{\Sigma} \boldsymbol{\Theta}_0 \in \mathbb{R}^{q\times q}$ may all depend on the feature dimension d; we suppress this dependence in the notation.

Assumption 2.1. The number of classes K and the true intercept vector $\boldsymbol{b}_0 \in \mathbb{R}^q$ are fixed. For each $d \geq q$, the feature vectors $\boldsymbol{x}_i$ are iid $N(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \succ \boldsymbol{0}$ is deterministic. The labels follow (1), and the true coefficient matrices satisfy

$$\boldsymbol{S} := \boldsymbol{\Theta}_0^\top \boldsymbol{\Sigma} \boldsymbol{\Theta}_0 \longrightarrow \boldsymbol{S}_\star \succeq \boldsymbol{0}.$$

The convergence is in operator norm, with no requirement on the rate. We impose no uniform upper bound or strictly positive lower bound on the eigenvalues of $\boldsymbol{\Sigma}$.

Under the Gaussian design, $n \leq d$ implies that the observations can be strictly separated almost surely, so a finite MLE does not exist, whether or not an intercept is fitted. We therefore focus on the nontrivial regime $n/d \to \alpha > 1$.

The true log-odds vector $\boldsymbol{\Theta}_0^\top \boldsymbol{x}_i + \boldsymbol{b}_0$ has distribution $N(\boldsymbol{b}_0, \boldsymbol{S})$. As the theorem below shows, the threshold depends on the feature covariance and true coefficients only through the limiting log-odds covariance $\boldsymbol{S}_\star$.

2.2 Phase transition with a fitted intercept

Throughout this subsection, we estimate both the coefficients and the intercept, whether or not the true intercept $\boldsymbol{b}_0$ is zero. Our main result characterizes the critical sample-to-dimension threshold, denoted by $\alpha_\star(\boldsymbol{S}, \boldsymbol{b}_0)$. For a sequence $\boldsymbol{S} \to \boldsymbol{S}_\star$, the threshold is evaluated at the limiting log-odds covariance $\boldsymbol{S}_\star$.

The threshold formula is expressed in terms of decision cones and weighted distances. For each label $y \in \{0, \ldots, q\}$, define the *decision cone*

$$\mathcal{C}_y = \{\boldsymbol{v} \in \mathbb{R}^q : v_y \geq v_j \text{ for every } j \in \{0, \ldots, q\}\}, \qquad v_0 = 0.$$

Membership $\boldsymbol{v} \in \mathcal{C}_y$ means that class y has maximal fitted probability, with ties allowed. Applied to a predictor change $\Delta\boldsymbol{v}$, it means that the predictor for class y increases at least as fast as every competitor; its current fitted probability need not be maximal.

The classical separation criterion connects these cones to finite-MLE existence (Silvapulle, 1981; Albert & Anderson, 1984; Santner & Duffy, 1986). Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ have rows $\boldsymbol{x}_i^\top$ and write $\tilde{\boldsymbol{X}} = [\boldsymbol{X}, \boldsymbol{1}_n]$ for the augmented design.

Proposition 2.2. *If $\tilde{\boldsymbol{X}}$ has rank $d+1$, then the MLE exists if and only if there is no pair $(\boldsymbol{B}, \boldsymbol{a}) \in \mathbb{R}^{d\times q} \times \mathbb{R}^q$, other than $(\boldsymbol{0}, \boldsymbol{0})$, such that*

$$\boldsymbol{B}^\top \boldsymbol{x}_i + \boldsymbol{a} \in \mathcal{C}_{Y_i}, \qquad i = 1, \ldots, n.$$

When it exists, the MLE is unique.

The proof of Proposition 2.2 is provided in Appendix A.3. The inequalities defining the decision cones allow ties, so the criterion includes both complete and quasi-complete separation. It is deterministic and does not require a correctly specified model.

To state the phase-transition threshold, we first define a weighted squared distance. For $\boldsymbol{\Lambda} \in \mathbb{S}_+^q$ and a nonempty closed convex set $\mathcal{C} \subseteq \mathbb{R}^q$, let

$$\operatorname{dist}_{\boldsymbol{\Lambda}}^2(\boldsymbol{v}, \mathcal{C}) := \inf_{\boldsymbol{c} \in \mathcal{C}} (\boldsymbol{v} - \boldsymbol{c})^\top \boldsymbol{\Lambda} (\boldsymbol{v} - \boldsymbol{c}).$$

For fixed $\boldsymbol{S} \succeq \mathbf{0}$ and $\boldsymbol{b}_0 \in \mathbb{R}^q$, let $\boldsymbol{z}_0, \boldsymbol{z}_1 \sim N(\mathbf{0}, \boldsymbol{I}_q)$ be independent, set $\boldsymbol{\eta} = \boldsymbol{b}_0 + \boldsymbol{S}^{1/2}\boldsymbol{z}_0$, and draw Y from the softmax law (1) at $\boldsymbol{\eta}$, with $\eta_0 = 0$. The auxiliary vector $\boldsymbol{z}_1$ is independent of $(\boldsymbol{z}_0, Y)$. All expectations below use this joint law.

Theorem 2.3. *For each fixed $\boldsymbol{S} \succeq \mathbf{0}$ and $\boldsymbol{b}_0 \in \mathbb{R}^q$, define*

$$\alpha_\star(\boldsymbol{S}, \boldsymbol{b}_0) := \sup_{\substack{\boldsymbol{H} \in \mathbb{S}_+^q,\ \operatorname{tr} \boldsymbol{H} = 1 \\ \boldsymbol{M} \in \mathbb{R}^{q \times q},\ \boldsymbol{a} \in \mathbb{R}^q}} \inf_{\substack{\boldsymbol{\Lambda} \in \mathbb{S}_+^q \\ \operatorname{tr} \boldsymbol{\Lambda} = 1}} \frac{\operatorname{tr}(\boldsymbol{\Lambda} \boldsymbol{H}^2)}{\mathbb{E} \operatorname{dist}_{\boldsymbol{\Lambda}}^2(\boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{H}\boldsymbol{z}_1 + \boldsymbol{a}, \mathcal{C}_Y)}. \tag{3}$$

The threshold satisfies $1 < \alpha_\star(\boldsymbol{S}, \boldsymbol{b}_0) < \infty$. Under Assumption 2.1, if $n, d \to \infty$ with $n/d \to \alpha > 1$ and $\alpha \neq \alpha_\star(\boldsymbol{S}_\star, \boldsymbol{b}_0)$, then

$$\mathbb{P}\{\textit{MLE exists}\} \longrightarrow \begin{cases} 0, & 1 < \alpha < \alpha_\star(\boldsymbol{S}_\star, \boldsymbol{b}_0), \\ 1, & \alpha > \alpha_\star(\boldsymbol{S}_\star, \boldsymbol{b}_0). \end{cases} \tag{4}$$

In (3), a ratio with zero denominator, including $0/0$, is defined as $+\infty$. The proof of Theorem 2.3 is provided in Appendix G.3. The formula for $\alpha_\star$ in (3) considers candidate predictor changes of the form $\Delta\boldsymbol{v} = \boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{H}\boldsymbol{z}_1 + \boldsymbol{a}$. The three terms represent the signal coordinates, independent Gaussian variation, and the fitted intercept direction, respectively. For a candidate change, the denominator measures the expected weighted squared adjustment needed to reach the observed label's decision cone $\mathcal{C}_Y$. The numerator is the expected $\boldsymbol{\Lambda}$-weighted squared size of $\boldsymbol{H}\boldsymbol{z}_1$. The matrix $\boldsymbol{\Lambda}$ weights combinations of class comparisons; the inner infimum selects the weighting that gives the smallest ratio of Gaussian variation to required adjustment. The outer supremum searches over candidate changes. Using matrix weightings keeps the comparisons between classes coupled. We derive the formula in Appendix B.

Theorem 2.3 identifies $\alpha_\star(\boldsymbol{S}_\star, \boldsymbol{b}_0)$ as the critical number of observations per feature for the existence of a finite MLE. Although the fitted model has $q(d+1)$ parameters, the optimization in (3) involves only matrices and vectors whose dimensions depend on q, not on d. The theorem makes no assertion at the threshold. For $n/d \to \alpha \leq 1$, nonexistence follows from interpolation when $\alpha < 1$ and from Corollary 2.6 when $\alpha = 1$.

2.3 Phase transition without a fitted intercept

Throughout this subsection, we fit without an intercept, whether or not the true intercept $\boldsymbol{b}_0$ is zero. We keep the same data-generating model as in Section 2.2 and estimate only $\boldsymbol{\Theta}$ by minimizing $L_n(\boldsymbol{\Theta}, \mathbf{0})$. When $\boldsymbol{b}_0 \neq \mathbf{0}$, the fitted model is misspecified.

Corollary 2.4. *If $\boldsymbol{X}$ has rank d, then the MLE with the fitted intercept fixed at zero exists if and only if there is no nonzero $\boldsymbol{B} \in \mathbb{R}^{d \times q}$ satisfying*

$$\boldsymbol{B}^\top \boldsymbol{x}_i \in \mathcal{C}_{Y_i}, \qquad i = 1, \ldots, n.$$

When it exists, the MLE is unique.

The proof applies the argument of Proposition 2.2 to the design $\boldsymbol{X}$; see Appendix A.3.

The resulting threshold formula has the same form as in Section 2.2, with the intercept direction constrained to $\boldsymbol{a} = \mathbf{0}$. We denote this threshold by $\alpha_\star^0(\boldsymbol{S}, \boldsymbol{b}_0)$, where the superscript 0 indicates that the fitted intercept is fixed at zero; $\boldsymbol{b}_0$ remains the true intercept.

Theorem 2.5. *For each fixed $\boldsymbol{S} \succeq \mathbf{0}$ and $\boldsymbol{b}_0 \in \mathbb{R}^q$, using the Gaussian vectors and label law defined above, define*

$$\boxed{\alpha_\star^0(\boldsymbol{S}, \boldsymbol{b}_0) = \sup_{\substack{\boldsymbol{H} \in \mathbb{S}_+^q,\ \operatorname{tr} \boldsymbol{H} = 1 \\ \boldsymbol{M} \in \mathbb{R}^{q \times q}}} \inf_{\substack{\boldsymbol{\Lambda} \in \mathbb{S}_+^q \\ \operatorname{tr} \boldsymbol{\Lambda} = 1}} \frac{\operatorname{tr}(\boldsymbol{\Lambda} \boldsymbol{H}^2)}{\mathbb{E}\, \operatorname{dist}_{\boldsymbol{\Lambda}}^2(\boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{H}\boldsymbol{z}_1, \mathcal{C}_Y)}.} \tag{5}$$

Under Assumption 2.1, fit without an intercept by minimizing $L_n(\boldsymbol{\Theta}, \mathbf{0})$ over $\boldsymbol{\Theta} \in \mathbb{R}^{d \times q}$, keeping the true intercept $\boldsymbol{b}_0$ unchanged. Then $1 < \alpha_\star^0(\boldsymbol{S}_\star, \boldsymbol{b}_0) < \infty$, and if $n, d \to \infty$ with $n/d \to \alpha > 1$ and $\alpha \neq \alpha_\star^0(\boldsymbol{S}_\star, \boldsymbol{b}_0)$,

$$\mathbb{P}\{\textit{MLE exists}\} \longrightarrow \begin{cases} 0, & 1 < \alpha < \alpha_\star^0(\boldsymbol{S}_\star, \boldsymbol{b}_0), \\ 1, & \alpha > \alpha_\star^0(\boldsymbol{S}_\star, \boldsymbol{b}_0). \end{cases}$$

As in (3), a zero denominator gives $+\infty$, including $0/0$. The proof of Theorem 2.5 is provided in Appendix G.3. Only the fitted direction $\boldsymbol{a}$ is removed; the true intercept $\boldsymbol{b}_0$ remains in the label law.

For every fixed $(\boldsymbol{S}, \boldsymbol{b}_0)$,

$$\alpha_\star^0(\boldsymbol{S}, \boldsymbol{b}_0) \le \alpha_\star(\boldsymbol{S}, \boldsymbol{b}_0).$$

Setting $\boldsymbol{a} = \mathbf{0}$ in the fitted-intercept formula recovers every candidate and inner value in the no-intercept formula, so its supremum cannot be smaller. At zero true intercept, the thresholds coincide in the binary model (Section 3.1) and when labels are equally likely and independent of the features (Corollary I.3). They need not coincide for general multiclass signals: Proposition 3.1 gives positive-definite three-class covariances with $\alpha_\star^0(\boldsymbol{S}, \mathbf{0}) < \alpha_\star(\boldsymbol{S}, \mathbf{0})$.

Neither theorem specifies the probability of MLE existence at its critical ratio. The endpoint $n/d \to 1$ has a separate conclusion.

Corollary 2.6. *Under Assumption 2.1, for either fit, with or without an intercept, if $n, d \to \infty$ with $n/d \to 1$, then $\mathbb{P}\{\textit{MLE exists}\} \to 0$.*

The sample-inclusion proof is given in Appendix G.4.

Proof strategy. The finite-sample criteria in Proposition 2.2 and Corollary 2.4 reduce the problem to constructing or excluding a separating direction. We first express population feasibility through a matrix second-moment constraint and derive the threshold by duality (Appendix B). The main difficulty is then passing from this population condition to exact constraints on every sample row.

Below the threshold, approximate message passing (AMP) produces a coefficient matrix whose class comparisons nearly satisfy the separation inequalities on average. To obtain a separator, we adjust the coefficients until every comparison is satisfied, including those that must hold as exact ties. Bounds that hold for all small sets of rows let us choose which observations need correction after running AMP. Slack in the population constraints ensures that the adjustment preserves the other comparisons (Appendices D and E).

Above the threshold, we construct a certificate orthogonal to every fitted predictor change but with negative inner product with any nonzero separator. These two properties rule out separation. We obtain the certificate from a population minimizer through a Gaussian construction and correction (Appendix F). Both constructions succeed with probability tending to one; Appendix G.3 completes the two main proofs.

3 Interpreting the Phase-Transition Threshold

Theorems 2.3 and 2.5 characterize finite-MLE existence through the thresholds defined in (3) and (5). We now examine how these thresholds recover familiar benchmarks, how estimating an intercept changes them, and which true parameters determine them.

3.1 Binary and Null Benchmarks

For $K = 2$, write $S = \gamma^2$, with $\gamma \ge 0$, and use the signed label $\widetilde{Y} = 2Y - 1$. Let Z_0, Z_1 be the scalar entries of $\boldsymbol{z}_0, \boldsymbol{z}_1$; then Z_1 is independent of $(Z_0, \widetilde{Y})$ and $\mathbb{P}(\widetilde{Y} = 1 \mid Z_0) = (1 + e^{-(b_0 + \gamma Z_0)})^{-1}$.

Here $q = 1$, so the trace constraints in (3) force $H = \Lambda = 1$. For $y \in \{0, 1\}$, let $\widetilde{y} = 2y - 1$. Then $\mathcal{C}_y = \{v \in \mathbb{R} : \widetilde{y}v \geq 0\}$ and $\mathrm{dist}_1^2(v, \mathcal{C}_y) = (-\widetilde{y}v)_+^2$. Set $Z = \widetilde{Y}Z_1$, which is standard Gaussian and independent of $(Z_0, \widetilde{Y})$. With $t_0 = -a$ and $t_1 = -M$, the threshold formula becomes

$$\boxed{\alpha_\star(\gamma^2, b_0) = \left[\inf_{t_0, t_1 \in \mathbb{R}} \mathbb{E}\big(t_0\widetilde{Y} + t_1\widetilde{Y}Z_0 - Z\big)_+^2\right]^{-1} = \frac{1}{h_{\mathrm{MLE}}(b_0, \gamma)},} \tag{6}$$

where $t_+ = \max\{t, 0\}$ and h_{MLE} is the threshold for binary logistic regression of Candès & Sur (2020, Theorem 1). Their ratio $\kappa = d/n$ is the reciprocal of our $\alpha = n/d$, so their existence condition $\kappa < h_{\mathrm{MLE}}$ agrees with $\alpha > \alpha_\star(\gamma^2, b_0)$. Fixing the fitted intercept to zero restricts the infimum to $t_0 = 0$. When the true intercept is also zero, symmetry gives

$$\alpha_\star(\gamma^2, 0) = \alpha_\star^0(\gamma^2, 0).$$

The reduction and symmetry argument are given in Appendix I.1.

The multinomial null model provides benchmarks with more classes. When labels are uniform and independent of the features, the two fitting conventions have the same threshold, with values

$$\alpha_\star(\mathbf{0}, \mathbf{0}) = \alpha_\star^0(\mathbf{0}, \mathbf{0}) = \begin{cases} 2, & K = 2, \\ 12/5, & K = 3, \\ 2.758012\ldots, & K = 4. \end{cases}$$

The binary value recovers Cover's random-dichotomy threshold (Cover, 1965); Corollary I.3 gives the general null formula. These values also serve as checks for the numerical calculations in Appendix L.

3.2 Estimating an intercept can change the threshold

Although estimating an intercept adds only $K - 1$ parameters, it enlarges the set of separating directions and can change the limiting threshold. The inequality $\alpha_\star(\boldsymbol{S}, \mathbf{0}) \geq \alpha_\star^0(\boldsymbol{S}, \mathbf{0})$ from Section 2.3 can therefore be strict, even though it is an equality in the binary and null benchmarks above. The following example compares the two fits under the same data law.

Proposition 3.1. *For $K = 3$ and $\varepsilon > 0$, define*

$$\boldsymbol{S}_\varepsilon = \begin{pmatrix} 100 + \varepsilon & 100 \\ 100 & 100 + \varepsilon \end{pmatrix}.$$

There exists $\varepsilon_0 > 0$ such that, for every fixed $0 < \varepsilon < \varepsilon_0$,

$$\alpha_\star(\boldsymbol{S}_\varepsilon, \mathbf{0}) > \alpha_\star^0(\boldsymbol{S}_\varepsilon, \mathbf{0}).$$

In particular, zero true intercept does not in general make the two fitting conventions equivalent, even when $\boldsymbol{S}$ is positive definite.

The proof of Proposition 3.1 is provided in Appendix J. For ratios strictly between the two thresholds, the fit without an intercept has a finite MLE with probability tending to one, whereas the fit with an intercept has no finite MLE with probability tending to one. Both conclusions concern the same data-generating distribution with zero true intercept.

The mechanism is visible in a rank-one reference model with $\boldsymbol{x} \sim N(\mathbf{0}, \boldsymbol{I}_d)$ and true predictors $(0, \gamma x_1, \gamma x_1)$, where $\gamma = 10$. Grouping classes 1 and 2 gives

$$\log \frac{\mathbb{P}(Y \in \{1, 2\} \mid \boldsymbol{x})}{\mathbb{P}(Y = 0 \mid \boldsymbol{x})} = \log 2 + \gamma x_1.$$

The $\log 2$ term appears because the pooled group contains two classes with the same predictor. It acts as a true intercept in the grouped binary problem, where fitting an intercept strictly raises the threshold. The proof in Appendix J shows that this gap persists in the three-class model and for small positive-definite perturbations.

For fixed K and fitting convention, the threshold is determined by $(\boldsymbol{S}, \boldsymbol{b}_0)$, the covariance and mean of the Gaussian log-odds. These parameters encode the joint distribution of the class comparisons. In general, the eigenvalues of $\boldsymbol{S}$ and the norm of $\boldsymbol{b}_0$ alone do not determine the threshold (Proposition I.2).

4 NUMERICAL EXPERIMENTS

We compare predicted thresholds with empirical MLE-existence frequencies in three experiments. The first varies signal strength at zero true intercept for three and four classes (Figure 1); the second varies the true intercept at fixed signal covariance for the same class counts (Figure 2). Both fit an intercept. The third uses a three-class model to compare fits with and without an intercept on the same data (Figure 3).

Experimental design. For each of the first two experiments, we generate $n = 2000$ independent observations with $K \in \{3, 4\}$ classes, Gaussian features $\boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{I}_d)$, and labels drawn from the conditional softmax model. For each target $\boldsymbol{S}$, we place its symmetric square root $\boldsymbol{S}^{1/2}$ in the first q rows of $\boldsymbol{\Theta}_0$ and set the remaining rows to zero, so $\boldsymbol{\Theta}_0^\top \boldsymbol{\Theta}_0 = \boldsymbol{S}$. For target ratios $\alpha = 2, 2.1, \ldots, 4$, we choose d as the nearest integer to $2000/\alpha$ and plot the actual ratio n/d. Each heatmap cell reports the fraction of 100 independent repetitions with a finite MLE; all outcomes are classified. The true predictor distribution is held fixed as d varies. Data generation and the dimension-search procedure are described in Appendix L.3; separation-certificate checks are given in Appendix L.5.

Computing the thresholds. We evaluate the applicable theoretical threshold, $\alpha_\star$ or $\alpha_\star^0$, from (3) or (5), independently of the simulated data. The null thresholds are available in closed form. Other values require numerical integration and matrix optimization. Appendix K describes the calculations and numerical checks.

4.1 EFFECT OF SIGNAL STRENGTH

We set $\boldsymbol{b}_0 = \mathbf{0}$ and vary the signal strength through

$$\boldsymbol{S} = \gamma^2\big(0.1\boldsymbol{I}_{K-1} + 0.9\mathbf{1}_{K-1}\mathbf{1}_{K-1}^\top\big), \qquad \gamma = 0, 0.1, \ldots, 2.$$

For $\gamma > 0$, the nonbaseline true linear predictors have variance γ^2 and pairwise correlation 0.9, while the feature coordinates are independent.

Figure 1 shows how the transition threshold changes with signal strength. As γ increases, the computed thresholds increase, and the empirical transitions move to larger n/d. At $\gamma = 0$, the thresholds are 2.4 for three classes and approximately 2.758 for four classes (Corollary I.3). The four-class threshold is larger throughout the displayed grid.

4.2 EFFECT OF THE TRUE INTERCEPT

We retain the sampling and fitting procedure of Section 4.1, fix the signal covariance at its $\gamma = 1$ value, and vary the true intercept:

$$\boldsymbol{S} = 0.1\boldsymbol{I}_{K-1} + 0.9\mathbf{1}_{K-1}\mathbf{1}_{K-1}^\top, \qquad \boldsymbol{b}_0 = (b, 0, \ldots, 0)^\top, \qquad b = 0, 0.1, \ldots, 1.5.$$

Figure 2 shows that changing the true intercept can shift the transition even when the signal covariance is fixed. As b increases from 0 to 1.5, the computed threshold rises from 2.6272 to 3.2128 for three classes and from 2.9652 to 3.5218 for four classes. The empirical transitions follow the same pattern. At $b = 0$, these curves recover the $\gamma = 1$ values in Figure 1. The numerical calculations are detailed in Appendix K.4.

4.3 EFFECT OF FITTING AN INTERCEPT

To illustrate the intercept effect underlying Proposition 3.1, we use the rank-one reference model from its proof, with true predictors $(0, 10x_1, 10x_1)$ and zero true intercept. At each $n = 1000, 2000, \ldots, 8000$, we run 100 paired repetitions. In each repetition, we draw a Gaussian design at the largest dimension and use prefixes of its columns as d varies; the labels therefore stay fixed across dimensions. We fit both models to these same observations. For each fit, we bracket the change from a finite to a nonexistent MLE between adjacent dimensions and use the midpoint of that bracket in $\alpha = n/d$. The theoretical thresholds are 9.493 without a fitted intercept and 10.128 with one.

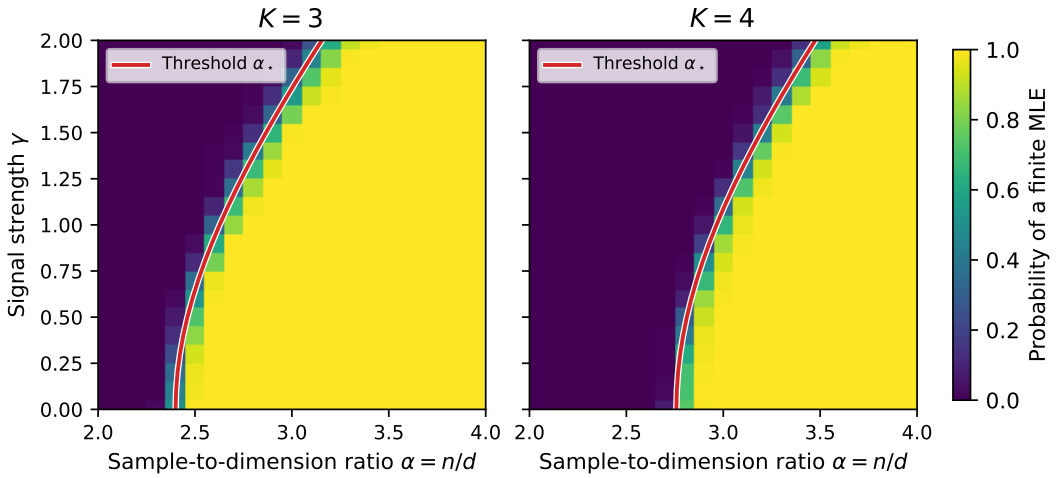

Figure 1: **MLE existence as signal strength varies**, for $K = 3$ (left) and $K = 4$ (right), with zero true intercept. The axes show $\alpha = n/d$ and γ. Color gives the fraction of 100 repetitions with a finite MLE: dark purple denotes none and yellow denotes all 100. The red line in each panel shows the threshold $\alpha_\star(\boldsymbol{S}, \mathbf{0})$, computed independently of the simulated data. Both panels use $n = 2000$ and an estimated intercept; the covariance family is specified in Section 4.1.

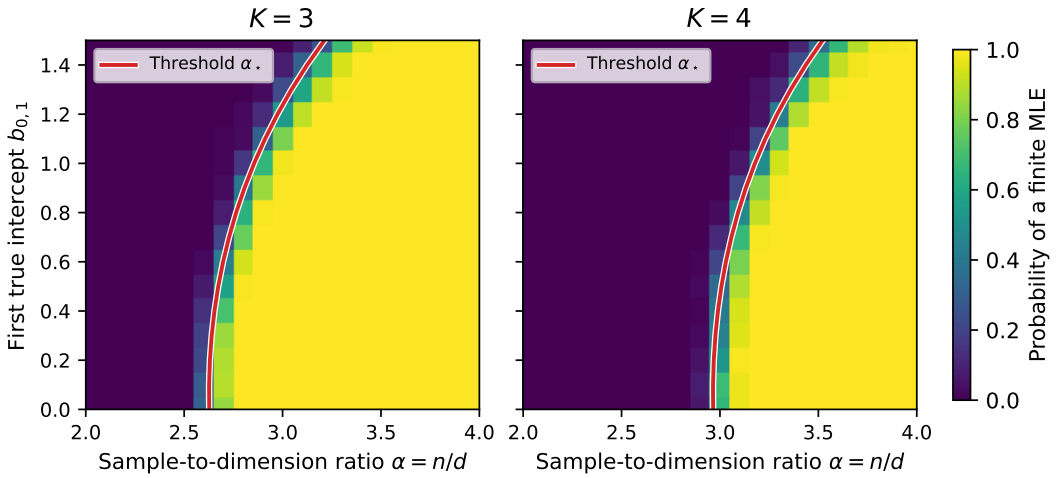

Figure 2: **MLE existence as the true intercept varies**, for $K = 3$ (left) and $K = 4$ (right). The axes show $\alpha = n/d$ and $b_{0,1} = b$; the remaining true intercepts are zero. Color gives the fraction of 100 repetitions with a finite MLE: dark purple denotes none and yellow denotes all 100. The red line in each panel shows the threshold $\alpha_\star(\boldsymbol{S}, \boldsymbol{b}_0)$, computed independently of the simulated data. Both panels use $n = 2000$, an estimated intercept, and the fixed covariance specified in Section 4.2.

Figure 3 shows the rightward shift caused by fitting an intercept. In panel (b), the *threshold shift* is the midpoint crossing with a fitted intercept minus that without one, computed on the same dataset and averaged over repetitions. The mean shifts at $n = 7000$ and 8000 are 0.645 and 0.654, with approximate 95% bootstrap intervals $[0.607, 0.682]$ and $[0.616, 0.694]$, respectively. Both are consistent with the predicted increase 0.636, although the estimates fluctuate across sample sizes. The transitions also sharpen as n increases (Appendix L.2). This experiment concerns the rank-one reference; the proposition establishes the strict inequality for the positive-definite family $\boldsymbol{S}_\varepsilon$ with sufficiently small $\varepsilon > 0$.

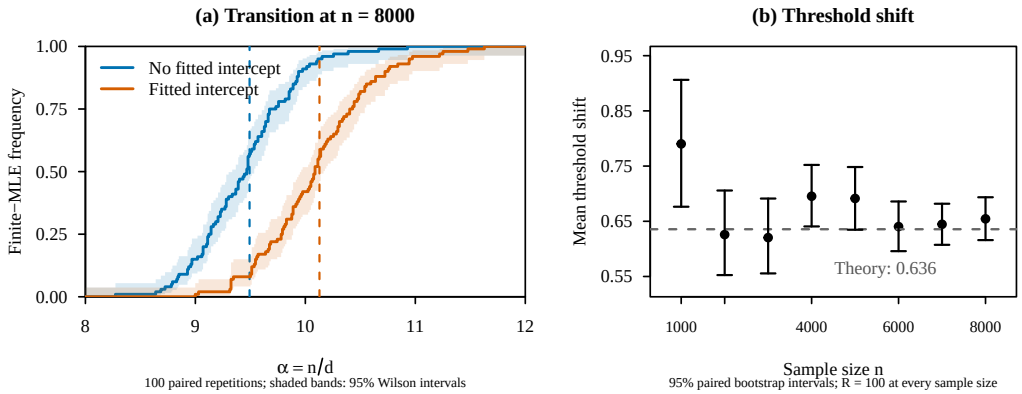

Figure 3: Intercept comparison under the same data law with zero true intercept. (a) Finite-MLE frequencies at $n = 8000$; bands are pointwise 95% Wilson intervals and dashed lines mark the theoretical thresholds. (b) Mean threshold shifts from fitting an intercept, with 95% bootstrap intervals; R is the number of pairs. The dashed line is the asymptotic increase 0.636, not a finite-sample mean prediction.

5 Discussion

We establish a sharp phase transition for finite-MLE existence in multinomial logistic regression with Gaussian features and a fixed number of classes. A finite-dimensional matrix optimization characterizes the threshold for models fitted with and without an intercept, allowing any fixed true intercept. The characterization recovers the binary threshold in Candès & Sur (2020) and gives explicit values when labels are uniform and independent of the features. It also shows that estimating an intercept can strictly increase the existence threshold, even when the true intercept is zero.

An important next step is to determine how much of this transition survives beyond Gaussian features. Does a sharp threshold still exist, and which aspects of the feature distribution determine it? Binary results for elliptical covariates (Tang & Ye, 2020) offer a starting point. Another question is whether finite-MLE existence has a sharp transition in other multi-index models fitted by maximum likelihood. Describing the MLE's norm and prediction error just above the threshold would also connect existence to statistical performance.

AI use statement

Generative AI tools were used to aid and polish the writing and to assist in proving mathematical claims, including developing proof arguments and reviewing and revising proofs. They also supported literature searches, LaTeX preparation, and numerical code and experiment development. The authors take responsibility for the final mathematical claims, proofs, text, code, and numerical results.

Reproducibility statement

The appendix provides proofs of the theoretical results. The two main phase-transition theorems are also formalized in Lean; the supplementary material includes the formalization and instructions for checking it, along with simulation code, saved results, and reproduction instructions.

References

Adelin Albert and John A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

Dennis Amelunxen, Martin Lotz, Michael B. McCoy, and Joel A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

Kiana Asgari, Andrea Montanari, and Basil Saeed. Local minima of the empirical risk in high dimension: General theorems and convex examples. *arXiv preprint arXiv:2502.01953*, 2025.

Raphaël Berthier, Andrea Montanari, and Phan-Minh Nguyen. State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA*, 9(1):33–79, 2020.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Burak Çakmak, Yue M. Lu, and Manfred Opper. A convergence analysis of approximate message passing with non-separable functions and applications to multi-class classification. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pp. 747–752. IEEE, 2024. Proposition 1 and Appendix F appear in arXiv:2402.08676v1.

Emmanuel J. Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27–42, 2020.

Hugo Chardon, Matthieu Lerasle, and Jaouad Mourtada. Finite-sample performance of the maximum likelihood estimator in logistic regression. *arXiv preprint arXiv:2411.02137*, 2024.

Thomas M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3): 326–334, 1965.

Kenneth R. Davidson and Stanisław J. Szarek. Local operator theory, random matrices and Banach spaces. In William B. Johnson and Joram Lindenstrauss (eds.), *Handbook of the Geometry of Banach Spaces*, volume 1, pp. 317–366. North-Holland, Amsterdam, 2001.

Jesús A. De Loera and Thomas Hogan. Stochastic Tverberg theorems with applications in multiclass logistic regression, separability, and centerpoints of data. *SIAM Journal on Mathematics of Data Science*, 2(4):1151–1166, 2020.

Jack Edmonds. Minimum partition of a matroid into independent subsets. *Journal of Research of the National Bureau of Standards, Section B: Mathematics and Mathematical Physics*, 69B(1–2): 67–72, 1965.

Alan J. Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4):263–265, 1952.

Qi Huangfu and J. A. Julian Hall. Parallelizing the dual revised simplex method. *Mathematical Programming Computation*, 10(1):119–142, 2018.

Zakhar Kabluchko and Dmitry Zaporozhets. Absorption probabilities for Gaussian polytopes and regular spherical simplices. *Advances in Applied Probability*, 52(2):588–616, 2020.

Ganesh Ramachandra Kini and Christos Thrampoulidis. Phase transitions for one-vs-one and one-vs-all linear separability in multiclass Gaussian mixtures. In *ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4020–4024. IEEE, 2021.

John E. Kolassa. Inference in the presence of likelihood monotonicity for polytomous and logistic regression. *Advances in Pure Mathematics*, 6(5):331–341, 2016.

John E. Kolassa and Juan Zhang. Inference in the presence of likelihood monotonicity for proportional hazards regression. *Statistica Neerlandica*, 77(3):322–339, 2023.

Takuya Koriyama and Pierre C. Bellec. Phase transitions for the existence of unregularized M-estimators in single index models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31519–31540. PMLR, 2025.

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. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10144–10157, 2021.

Thomas J. Santner and Diane E. Duffy. A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 73(3): 755–758, 1986.

Mervyn J. Silvapulle. On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(3):310–313, 1981.

Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

Kai Tan and Pierre C. Bellec. Multinomial logistic regression: Asymptotic normality on null covariates in high-dimensions. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70892–70925, 2023.

Wenpin Tang and Yuting Ye. The existence of maximum likelihood estimate in high-dimensional binary response generalized linear models. *Electronic Journal of Statistics*, 14(2):4028–4053, 2020.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24164–24179, 2021.

A PROOF ROADMAP AND FINITE-SAMPLE FOUNDATIONS

Appendices A and B establish the finite-sample criterion and population threshold. For a first reading, follow the overview in Appendix C, the separator and polar-certificate constructions in Appendices E and F, and the main proofs in Appendix G.3. The shared approximation and repair lemmas in Appendix D can be read as needed.

We initially take $\boldsymbol{S} \succeq \mathbf{0}$ and the true intercept $\boldsymbol{b}_0$ fixed, writing $\alpha_\star = \alpha_\star(\boldsymbol{S}, \boldsymbol{b}_0)$ in the fitted-intercept explanations. Fixing the direction $\boldsymbol{a} = \mathbf{0}$ gives the corresponding formulation without a fitted intercept. The final proof treats both fitting conventions together and allows $\boldsymbol{S} \to \boldsymbol{S}_\star$.

A.1 NOTATION AND BASIC IDENTITIES

Write $\mathbb{S}^m_{++}$ for the positive-definite cone, $\boldsymbol{A}^\dagger$ for the Moore–Penrose inverse, $\mathbb{O}(m)$ for the orthogonal group, and $\|\boldsymbol{v}\|^2_{\boldsymbol{A}} = \boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v}$. Norms without subscripts are Euclidean; $\|\cdot\|_F$, $\|\cdot\|_{\mathrm{op}}$, and $\|\cdot\|_*$ are Frobenius, operator, and nuclear norms. Derivatives use operator norms; matrix inner products are Frobenius. Vector inequalities are componentwise. Write $[m] = \{1, \dots, m\}$. For $B \subseteq [n]$, $\boldsymbol{A}_B$ is the row submatrix in increasing order, and rows are identified with their transposes in geometric expressions. For a subspace $U \leq \mathbb{R}^m$, $\boldsymbol{P}_U$ is its orthogonal projector. For a closed convex cone $\mathcal{C}$, $\mathcal{C}^\circ = \{\boldsymbol{z} : \boldsymbol{z}^\top \boldsymbol{c} \leq 0 \text{ for all } \boldsymbol{c} \in \mathcal{C}\}$; $\Pi_{\mathcal{C}}$ and $\mathrm{relint}(\mathcal{C})$ denote projection and relative interior. As in the main text, $\mathbf{1}_m$ denotes the all-ones vector; $\mathbb{1}_E$ denotes the indicator of an event E.

All random variables may be realized on a common sufficiently rich probability space. Write $\|\boldsymbol{x}\|_{L^k} = (\mathbb{E}\|\boldsymbol{x}\|^k)^{1/k}$ and $\boldsymbol{x}_n = O_{L^k}(a_n)$ when $\|\boldsymbol{x}_n\|_{L^k} = O(a_n)$. Throughout, $\kappa = \alpha^{-1}$ and $\kappa_\star = \alpha_\star^{-1}$; boundary arguments are retained when comparing fitting conventions. Undecorated $c, C > 0$ may change from line to line, with relevant dependence stated locally.

We use Stein's identity in the following form. If $\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{I}_m)$ is independent of T and $\boldsymbol{g}(\boldsymbol{z}, t) \in \mathbb{R}^r$ has a weak derivative in $\boldsymbol{z}$, with $\boldsymbol{g}(\boldsymbol{z}, T)$ and $D_{\boldsymbol{z}}\boldsymbol{g}(\boldsymbol{z}, T)$ square-integrable, then

$$\mathbb{E}[\boldsymbol{g}(\boldsymbol{z}, T)\boldsymbol{z}^\top] = \mathbb{E}[D_{\boldsymbol{z}}\boldsymbol{g}(\boldsymbol{z}, T)]. \tag{7}$$

Condition on T and integrate the Gaussian density by parts, then approximate in the Gaussian Sobolev norm. For $\boldsymbol{g}(\boldsymbol{z}, t) = \boldsymbol{f}(\boldsymbol{H}\boldsymbol{z}, t)$ the derivative is with respect to $\boldsymbol{z}$; $\boldsymbol{H}$ need not be invertible.

Finally, for each decision cone choose facet normals $\boldsymbol{a}_{y,1}, \dots, \boldsymbol{a}_{y,q} \in \mathbb{R}^q$ such that

$$\mathcal{C}_y = \{\boldsymbol{v} \in \mathbb{R}^q : \boldsymbol{a}_{y,\ell}^\top \boldsymbol{v} \geq 0,\ \ell \in [q]\}. \tag{8}$$

A.2 SEPARATION AND FINITE-MLE EXISTENCE

A direction $(\boldsymbol{B}, \boldsymbol{a})$ changes observation i's logits by $\Delta\boldsymbol{v}_i = \boldsymbol{B}^\top \boldsymbol{x}_i + \boldsymbol{a}$, with baseline change zero. We call a nonzero pair a separating direction if

$$\Delta\boldsymbol{v}_i \in \mathcal{C}_{Y_i} \qquad (i \in [n]). \tag{9}$$

The cone inequalities allow ties and therefore include complete and quasi-complete separation. Under full column rank of $\tilde{\boldsymbol{X}}$, at least one comparison must be strict; otherwise all predictor changes vanish and so does the parameter direction.

For nonsingular Gaussian features, $\mathrm{rank}\,\tilde{\boldsymbol{X}} = \min(n, d+1)$ almost surely. If $n \leq d+1$, full row rank allows scores $\boldsymbol{v}_i = -\mathbf{1}_q$ for $Y_i = 0$ and $\boldsymbol{v}_i = \boldsymbol{e}_{Y_i}$ otherwise. Each observed class exceeds every competitor by one, so scaling these scores by t gives loss $\log(1 + qe^{-t}) \to 0$, unattainable at finite parameters. When $n/d \to \alpha > 1$, full column rank holds eventually, and the following criterion applies.

A.3 PROOFS OF THE FINITE-SAMPLE CRITERIA

Proof of Proposition 2.2. Separation precludes attainment. Along a separating direction, every observed-minus-competing logit difference is nondecreasing, and at least one increases strictly by full column rank. Equivalently, the derivative of a row loss is $\sum_{j=0}^q p_{i,j}(t)(\Delta v_{i,j} - \Delta v_{i,Y_i}) \leq 0$, with a

strict inequality for at least one row because all fitted probabilities are positive. The loss therefore decreases from every finite parameter, excluding a finite minimizer.

Failure of attainment produces a separator. If no minimizer exists, a minimizing sequence $(\boldsymbol{\Theta}_m, \boldsymbol{b}_m)$ has norms $r_m \to \infty$; a bounded subsequence would attain the infimum by continuity. Along a subsequence its normalization converges to a unit pair $(\boldsymbol{B}, \boldsymbol{a})$. For $\boldsymbol{w}_i^{(m)} = \boldsymbol{\Theta}_m^\top \boldsymbol{x}_i + \boldsymbol{b}_m$, with $w_{i0}^{(m)} = 0$,

$$w_{i,j}^{(m)} - w_{i,Y_i}^{(m)} \le \ell_{Y_i}(\boldsymbol{w}_i^{(m)}) \le n L_n(\boldsymbol{\Theta}_m, \boldsymbol{b}_m) = O(1).$$

The first inequality follows from the log-sum-exp formula, and the second from nonnegative row losses. Divide by r_m and pass to the limit: $\Delta v_{i,j} \le \Delta v_{i,Y_i}$ for every i, j. Thus the nonzero limit pair satisfies (9). Finally, Lemma A.1, applied to the augmented design, gives uniqueness whenever a minimizer exists. □

For a fit without an intercept, write $L_n(\boldsymbol{\Theta}) = L_n(\boldsymbol{\Theta}, \boldsymbol{0})$. The following statements concern the fitted design only and do not require correct specification of the label model.

Lemma A.1 (Strict convexity)**.** *If the design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ has rank d, then $\boldsymbol{\Theta} \mapsto L_n(\boldsymbol{\Theta})$ is strictly convex on $\mathbb{R}^{d \times q}$.*

Proof. For finite $\boldsymbol{v} \in \mathbb{R}^q$, define

$$p_j(\boldsymbol{v}) = \frac{e^{v_j}}{\sum_{\ell=0}^{q} e^{v_\ell}}, \qquad j \in \{0, \dots, q\},$$

and put $\boldsymbol{p}(\boldsymbol{v}) = (p_1(\boldsymbol{v}), \dots, p_q(\boldsymbol{v}))$. The Hessian in baseline coordinates is

$$\nabla^2 \ell_y(\boldsymbol{v}) = \operatorname{diag}(\boldsymbol{p}(\boldsymbol{v})) - \boldsymbol{p}(\boldsymbol{v})\boldsymbol{p}(\boldsymbol{v})^\top.$$

For $\boldsymbol{a} \in \mathbb{R}^q \setminus \{\boldsymbol{0}\}$, this quadratic form is the variance of a random variable that takes value a_j on class $j \in [q]$ and value 0 on the baseline class. Every class probability is strictly positive, so the variance is positive. Hence $\nabla^2 \ell_y(\boldsymbol{v}) \succ \boldsymbol{0}$.

If $\boldsymbol{\Theta} \ne \boldsymbol{\Theta}'$, full column rank gives $\boldsymbol{X}\boldsymbol{\Theta} \ne \boldsymbol{X}\boldsymbol{\Theta}'$. Thus at least one row loss is strictly convex along the segment from $\boldsymbol{\Theta}$ to $\boldsymbol{\Theta}'$, and the others are convex. Their average is therefore strictly convex. □

Proof of Corollary 2.4. Apply the proof of Proposition 2.2 with $\boldsymbol{X}$ in place of $\tilde{\boldsymbol{X}}$. It uses only full column rank of the fitted design; Lemma A.1 also gives uniqueness. □

If $n \le d$ and $\boldsymbol{X}$ has full row rank, solve $\boldsymbol{X}\boldsymbol{B} = \boldsymbol{T}$ with $\boldsymbol{t}_i \in \operatorname{int} \mathcal{C}_{Y_i}$ for every row. Scaling $\boldsymbol{B}$ then drives the loss to zero, which no finite parameter attains. This proves nonexistence under Gaussian design whenever $n \le d$, including the regime $n/d \to \alpha < 1$.

Lemma A.2 (Strong conic alternative)**.** *Let $Y_1, \dots, Y_n \in \{0, 1, \dots, q\}$ be class labels, let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ have full column rank, and put $\mathcal{K} = \prod_{i=1}^n \mathcal{C}_{Y_i}$. Then*

$$\{\boldsymbol{X}\boldsymbol{B} : \boldsymbol{B} \in \mathbb{R}^{d \times q}\} \cap \mathcal{K} = \{\boldsymbol{0}\} \tag{10}$$

if and only if there is a matrix $\boldsymbol{Q} \in \mathbb{R}^{n \times q}$, with rows $\boldsymbol{q}_i^\top$, such that

$$\boldsymbol{X}^\top \boldsymbol{Q} = \boldsymbol{0}, \qquad \boldsymbol{q}_i \in \operatorname{int}(\mathcal{C}_{Y_i}^\circ) \quad (i \in [n]). \tag{11}$$

Proof. Let $\mathcal{S} = \{\boldsymbol{X}\boldsymbol{B} : \boldsymbol{B} \in \mathbb{R}^{d \times q}\}$ and choose $\boldsymbol{H} \in \operatorname{int}(-\mathcal{K}^\circ)$. The base $\mathcal{B} = \{\boldsymbol{C} \in \mathcal{K} : \langle \boldsymbol{H}, \boldsymbol{C} \rangle = 1\}$ is compact because the product cone is pointed. If (10) holds, strongly separate $\mathcal{B}$ from the subspace $\mathcal{S}$. Since $\mathcal{S}$ is unbounded in both directions, the separating vector belongs to $\mathcal{S}^\perp$; strict negativity on the compact base puts it in $\operatorname{int} \mathcal{K}^\circ$. This is exactly (11). Conversely, taking the inner product with such a $\boldsymbol{Q}$ shows that a nonzero element of $\mathcal{S} \cap \mathcal{K}$ is impossible. □

A.4 WHITENING AND SIGNAL COORDINATES

Whiten the features and coefficients by $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x} \sim N(\mathbf{0}, \boldsymbol{I}_d)$ and $\widetilde{\boldsymbol{\Theta}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Theta}$. This invertible change preserves every predictor and the likelihood; no uniform eigenvalue bounds on $\boldsymbol{\Sigma}$ are needed. Write $\boldsymbol{S} = \widetilde{\boldsymbol{\Theta}}_0^\top \widetilde{\boldsymbol{\Theta}}_0$ and assume $d \geq q$. We construct a deterministic orthogonal matrix $\boldsymbol{O} = [\boldsymbol{U}_0, \boldsymbol{U}_\perp]$ satisfying

$$\boldsymbol{U}_0^\top \boldsymbol{U}_0 = \boldsymbol{I}_q, \qquad \boldsymbol{U}_0^\top \widetilde{\boldsymbol{\Theta}}_0 = \boldsymbol{S}^{1/2}, \qquad \boldsymbol{U}_\perp^\top \widetilde{\boldsymbol{\Theta}}_0 = \mathbf{0}. \tag{12}$$

We use the following equal-Gram fact: if $\boldsymbol{R}_1, \boldsymbol{R}_2 \in \mathbb{R}^{d\times q}$ satisfy $\boldsymbol{R}_1^\top \boldsymbol{R}_1 = \boldsymbol{R}_2^\top \boldsymbol{R}_2$, then $\boldsymbol{R}_2 = \boldsymbol{O}\boldsymbol{R}_1$ for some orthogonal $\boldsymbol{O}$. Indeed, $\boldsymbol{R}_1\boldsymbol{v} \mapsto \boldsymbol{R}_2\boldsymbol{v}$ is a well-defined isometry of their ranges, since $\|\boldsymbol{R}_1\boldsymbol{v}\|^2 = \|\boldsymbol{R}_2\boldsymbol{v}\|^2$; extend orthonormal bases to obtain $\boldsymbol{O}$. Apply this fact to $\boldsymbol{R}_1 = [\boldsymbol{S}^{1/2}; \mathbf{0}]$ and $\boldsymbol{R}_2 = \widetilde{\boldsymbol{\Theta}}_0$. The first q columns of $\boldsymbol{O}$ give

$$\boldsymbol{U}_0^\top \boldsymbol{U}_0 = \boldsymbol{I}_q, \qquad \boldsymbol{U}_0 \boldsymbol{S}^{1/2} = \widetilde{\boldsymbol{\Theta}}_0, \tag{13}$$

and the remaining columns give (12). This argument includes singular or zero $\boldsymbol{S}$ and the empty residual block when $d = q$.

The rotated features $(\boldsymbol{z}_0, \boldsymbol{z}_\perp) = \boldsymbol{O}^\top \boldsymbol{z}$ are independent standard Gaussians of dimensions q and $d-q$. The true logits are $\boldsymbol{b}_0 + \boldsymbol{S}^{1/2}\boldsymbol{z}_0$, so $\boldsymbol{z}_\perp$ is independent of $(\boldsymbol{z}_0, Y)$. A fitted slope direction remains unrestricted: write

$$\boldsymbol{O}^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{B} = \begin{pmatrix} \boldsymbol{M}^\top \\ \boldsymbol{A} \end{pmatrix}, \qquad \Delta\boldsymbol{v} = \boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{A}^\top \boldsymbol{z}_\perp + \boldsymbol{a}, \quad \boldsymbol{A} \in \mathbb{R}^{(d-q)\times q}.$$

For a sample, collect the two feature blocks into $\boldsymbol{Z}_0$ and $\boldsymbol{G}$. Then $\boldsymbol{G}$ is iid standard Gaussian independently of $\boldsymbol{Z}_0$ and the labels, and

$$[\boldsymbol{Z}_0, \boldsymbol{G}] = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{O}, \qquad \Delta\boldsymbol{V} = \boldsymbol{X}\boldsymbol{B} + \mathbf{1}_n\boldsymbol{a}^\top = \boldsymbol{Z}_0\boldsymbol{M}^\top + \boldsymbol{G}\boldsymbol{A} + \mathbf{1}_n\boldsymbol{a}^\top.$$

Thus the transformed design has the same column space as $\boldsymbol{X}$, and one common Gaussian matrix acts on all class columns. Separation requires every row of $\Delta\boldsymbol{V}$ to lie in its decision cone.

B POPULATION FEASIBILITY AND CHARACTERIZATION OF THE BOUNDARY

The Gaussian coordinates now allow us to express the boundary through a population correction. Until the extension in Appendix G, fix K, $\boldsymbol{b}_0$, and the exact whitened Gram matrix $\widetilde{\boldsymbol{\Theta}}_0^\top \widetilde{\boldsymbol{\Theta}}_0 = \boldsymbol{S} \succeq \mathbf{0}$ independently of d. We first explain the correction and its matrix second-moment constraint, then prove the duality that eliminates it from the boundary formula. The two fitting conventions differ only in whether the intercept direction is free or fixed to zero.

B.1 POPULATION FEASIBILITY AND THE BOUNDARY FORMULA

For a fixed direction, the residual contribution has the law $\boldsymbol{H}\boldsymbol{z}_1$, where $\boldsymbol{H} = (\boldsymbol{A}^\top\boldsymbol{A})^{1/2}$ and $\boldsymbol{z}_1 \sim N(\mathbf{0}, \boldsymbol{I}_q)$ is independent of $(\boldsymbol{z}_0, Y)$. Write $\Delta\boldsymbol{v} = \boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{H}\boldsymbol{z}_1 + \boldsymbol{a}$. The true parameters enter through $\mathbb{P}(Y = y \mid \boldsymbol{z}_0) = \pi_y(\boldsymbol{b}_0 + \boldsymbol{S}^{1/2}\boldsymbol{z}_0) > 0$; $\boldsymbol{a}$ is a fitted direction, free even when $\boldsymbol{b}_0 = \mathbf{0}$.

A fixed nonzero direction cannot satisfy $\Delta\boldsymbol{v} \in \mathcal{C}_Y$ almost surely: positivity of every conditional class probability would force $\Delta\boldsymbol{v} \in \bigcap_y \mathcal{C}_y = \{\mathbf{0}\}$ and hence all coefficients to vanish. A sample separator can adapt to the labels. We represent that adaptation by a square-integrable correction $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{z}_0, \boldsymbol{z}_1, Y)$, with a matrix second-moment constraint. The correction is an auxiliary population variable; the eventual sample separator still uses one common coefficient matrix.

Definition B.1 (Population feasibility with an intercept)**.** For $\alpha > 1$, a triple $(\boldsymbol{H}, \boldsymbol{M}, \boldsymbol{a}) \in \mathbb{S}_+^q \times \mathbb{R}^{q\times q} \times \mathbb{R}^q$ is α-feasible if there is a square-integrable $\mathbb{R}^q$-valued correction $\boldsymbol{u}$ (meaning $\mathbb{E}\|\boldsymbol{u}\|^2 < \infty$) such that

$$\boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{H}\boldsymbol{z}_1 + \boldsymbol{a} + \boldsymbol{u} \in \mathcal{C}_Y \quad \text{a.s.}, \qquad \mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top] \preceq \alpha^{-1}\boldsymbol{H}^2. \tag{14}$$

Here $\boldsymbol{U} \preceq \boldsymbol{V}$ means that $\boldsymbol{V} - \boldsymbol{U}$ is positive semidefinite. The witness is nontrivial when $(\boldsymbol{H}, \boldsymbol{M}, \boldsymbol{a}) \neq (\mathbf{0}, \mathbf{0}, \mathbf{0})$.

The matrix constraint controls every contrast: $\mathbb{E}(\boldsymbol{w}^\top \boldsymbol{u})^2 \le \alpha^{-1}\|\boldsymbol{H}\boldsymbol{w}\|^2$ for all $\boldsymbol{w}$. Its factor α^{-1} is the limiting residual-feature fraction $(d-q)/n$. A scalar bound on $\mathbb{E}\|\boldsymbol{u}\|^2$ would control only the sum of these costs. For example, in the binary null model, let $Z \sim N(0,1)$ be independent of a uniform binary label Y. Then $(H, M, a) = (1, 0, 0)$ and $U = -Z\mathbb{1}_{\{Z \notin \mathcal{C}_Y\}}$ give $Z + U \in \mathcal{C}_Y$ and $\mathbb{E}U^2 = 1/2$, hence feasibility for $1 < \alpha \le 2$. The matching boundary is proved in Appendix I.4.

Eliminating the correction. For every $\boldsymbol{\Lambda} \succeq \boldsymbol{0}$, using $\Delta\boldsymbol{v} + \boldsymbol{u}$ as a cone point gives $\operatorname{dist}^2_{\boldsymbol{\Lambda}}(\Delta\boldsymbol{v}, \mathcal{C}_Y) \le \boldsymbol{u}^\top \boldsymbol{\Lambda}\boldsymbol{u}$. Thus feasibility implies

$$\mathbb{E}\operatorname{dist}^2_{\boldsymbol{\Lambda}}(\boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{H}\boldsymbol{z}_1 + \boldsymbol{a}, \mathcal{C}_Y) \le \alpha^{-1}\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{H}^2) \quad \text{for every } \boldsymbol{\Lambda} \succeq \boldsymbol{0}. \tag{15}$$

Lemma B.3 proves the converse: these inequalities yield one correction satisfying all matrix directions simultaneously. For a fixed triple, the largest allowed ratio is therefore the infimum of the trace-to-distance ratios over $\boldsymbol{\Lambda}$; optimizing the triple gives (3). A zero denominator imposes no restriction, explaining the $+\infty$ convention, including $0/0$. Homogeneity removes the two redundant scales. Appendix B.3 proves these normalizations and the following characterization.

Proposition B.2. *Fix $\boldsymbol{S} \succeq \boldsymbol{0}$ and $\boldsymbol{b}_0 \in \mathbb{R}^q$, and let $\alpha_\star(\boldsymbol{S}, \boldsymbol{b}_0)$ be defined by (3). Then*

$$\alpha_\star(\boldsymbol{S}, \boldsymbol{b}_0) = \sup\{\alpha > 1 : \textit{a nontrivial feasible triple exists}\}. \tag{16}$$

Here the supremum is set to 1 if the set is empty. Feasibility is understood in the sense of Definition B.1. A nontrivial feasible triple exists for every $1 < \alpha < \alpha_\star(\boldsymbol{S}, \boldsymbol{b}_0)$, and none exists for $\alpha > \alpha_\star(\boldsymbol{S}, \boldsymbol{b}_0)$.

Feasibility is downward closed in α, since decreasing α relaxes the moment bound. Below the boundary we may therefore choose a witness at a larger ratio and obtain strict moment slack. Appendix C explains how that slack enters the sample constructions.

B.2 Matrix second-moment duality

For a square-integrable random vector $\boldsymbol{v} \in \mathbb{R}^q$ and random cone $\mathcal{C} = \mathcal{C}_Y$, we characterize the existence of a correction using the weighted distance in (3).

Define the second-moment upper set

$$\mathcal{Q}(\boldsymbol{v}, \mathcal{C}) = \left\{ \boldsymbol{B} \in \mathbb{S}^q_+ : \begin{array}{l} \text{there exists } \boldsymbol{c} \in L^2(\Omega; \mathbb{R}^q) \text{ such that} \\ \boldsymbol{c} \in \mathcal{C} \text{ a.s. and } \mathbb{E}[(\boldsymbol{c}-\boldsymbol{v})(\boldsymbol{c}-\boldsymbol{v})^\top] \preceq \boldsymbol{B} \end{array} \right\}.$$

Lemma B.3 (Semidefinite support representation)**.** *The set $\mathcal{Q}(\boldsymbol{v}, \mathcal{C})$ is closed, convex, and upward-closed in the positive-semidefinite order. For every $\boldsymbol{B} \in \mathbb{S}^q_+$,*

$$\boldsymbol{B} \in \mathcal{Q}(\boldsymbol{v}, \mathcal{C}) \iff \operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{B}) \ge \mathbb{E}\operatorname{dist}^2_{\boldsymbol{\Lambda}}(\boldsymbol{v}, \mathcal{C}) \quad \textit{for every } \boldsymbol{\Lambda} \in \mathbb{S}^q_+. \tag{17}$$

Proof. Upward closure is immediate. If selections $\boldsymbol{c}_j \in \mathcal{C}$ have moment matrices $\boldsymbol{Q}_j \preceq \boldsymbol{B}_j$, their convex combination $\boldsymbol{c} = \theta\boldsymbol{c}_1 + (1-\theta)\boldsymbol{c}_2$ belongs to $\mathcal{C}$ and satisfies

$$\mathbb{E}[(\boldsymbol{c}-\boldsymbol{v})(\boldsymbol{c}-\boldsymbol{v})^\top] = \theta\boldsymbol{Q}_1 + (1-\theta)\boldsymbol{Q}_2 - \theta(1-\theta)\mathbb{E}[(\boldsymbol{c}_1-\boldsymbol{c}_2)(\boldsymbol{c}_1-\boldsymbol{c}_2)^\top] \preceq \theta\boldsymbol{B}_1 + (1-\theta)\boldsymbol{B}_2.$$

For closedness, let $\boldsymbol{B}_m \to \boldsymbol{B}$ and choose corrections $\boldsymbol{u}_m$ with $\boldsymbol{v} + \boldsymbol{u}_m \in \mathcal{C}$ almost surely and $\mathbb{E}[\boldsymbol{u}_m\boldsymbol{u}_m^\top] \preceq \boldsymbol{B}_m$. Their squared L^2 norms are bounded by $\operatorname{tr}\boldsymbol{B}_m$, so a subsequence converges weakly to $\boldsymbol{u}$. The L^2 selections of $\mathcal{C}$ form a closed convex set: convexity is pointwise, and norm closedness follows by taking an almost-surely convergent subsequence and using closedness of each cone. This set is therefore weakly closed, giving $\boldsymbol{v}+\boldsymbol{u} \in \mathcal{C}$ almost surely. Weak lower semicontinuity applied to each scalar projection yields

$$\mathbb{E}(\boldsymbol{e}^\top\boldsymbol{u})^2 \le \liminf_m \mathbb{E}(\boldsymbol{e}^\top\boldsymbol{u}_m)^2 \le \boldsymbol{e}^\top\boldsymbol{B}\boldsymbol{e} \qquad (\boldsymbol{e} \in \mathbb{R}^q),$$

so $\boldsymbol{B} \in \mathcal{Q}(\boldsymbol{v}, \mathcal{C})$.

For $\boldsymbol{\Lambda} \succeq \boldsymbol{0}$, we claim

$$\inf_{\substack{\boldsymbol{c} \in \mathcal{C} \text{ a.s.} \\ \boldsymbol{c}-\boldsymbol{v} \in L^2}} \mathbb{E}(\boldsymbol{c}-\boldsymbol{v})^\top\boldsymbol{\Lambda}(\boldsymbol{c}-\boldsymbol{v}) = \mathbb{E}\operatorname{dist}^2_{\boldsymbol{\Lambda}}(\boldsymbol{v}, \mathcal{C}). \tag{18}$$

For each of the finitely many labels y, choose a dense sequence $(\boldsymbol{c}_{y,j})_{j\ge 0}$ in $\mathcal{C}_y$ starting at zero. Let $\boldsymbol{c}_N$ be the first minimizer among $\boldsymbol{c}_{Y,0},\dots,\boldsymbol{c}_{Y,N}$ of the quadratic cost, and denote that minimum by d_N. Finite comparisons of continuous costs make $\boldsymbol{c}_N$ jointly measurable in $(\boldsymbol{\Lambda},\boldsymbol{v},Y)$; its finite range makes it an L^2 selection. Density and continuity give

$$d_N \downarrow \operatorname{dist}^2_{\boldsymbol{\Lambda}}(\boldsymbol{v},\mathcal{C}), \qquad 0 \le d_N \le \boldsymbol{v}^\top \boldsymbol{\Lambda}\boldsymbol{v} \le \|\boldsymbol{\Lambda}\|_{\mathrm{op}}\|\boldsymbol{v}\|^2.$$

The limit is measurable, and dominated convergence gives convergence of the expected costs. Every admissible selection has cost at least the pointwise infimum, while $\boldsymbol{c}_N$ approaches it in expectation. This proves (18) for every PSD metric, including singular and zero $\boldsymbol{\Lambda}$.

If $\boldsymbol{B} \in \mathcal{Q}$, a feasible selection immediately gives the inequalities in (17). Conversely, if $\boldsymbol{B} \notin \mathcal{Q}$, strong separation gives a symmetric $\boldsymbol{\Lambda}$ with

$$\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{B}) < \inf_{\boldsymbol{Q}\in\mathcal{Q}} \operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{Q}).$$

Here upward closure forces $\boldsymbol{\Lambda} \succeq \boldsymbol{0}$: otherwise adding $t\boldsymbol{e}\boldsymbol{e}^\top$ in a negative quadratic direction makes the infimum $-\infty$. The infimum is then $\mathbb{E}\operatorname{dist}^2_{\boldsymbol{\Lambda}}(\boldsymbol{v},\mathcal{C})$ by the definition of $\mathcal{Q}$ and (18), contradicting the proposed inequalities. □

B.3 Population boundary with a fitted intercept

Using Lemma B.3, we now prove Proposition B.2.

Proof of Proposition B.2. A nontrivial feasible triple has $\boldsymbol{H} \neq \boldsymbol{0}$: otherwise the moment bound gives $\boldsymbol{u} = \boldsymbol{0}$, and positivity of every conditional class probability forces $\boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{a} \in \bigcap_y \mathcal{C}_y = \{\boldsymbol{0}\}$ almost surely, hence $\boldsymbol{M} = \boldsymbol{a} = \boldsymbol{0}$. Since each $\mathcal{C}_y$ is a cone, dividing $(\boldsymbol{H},\boldsymbol{M},\boldsymbol{a},\boldsymbol{u})$ by $\operatorname{tr}\boldsymbol{H} > 0$ preserves both feasibility constraints. Thus nontrivial feasibility is equivalent to feasibility with $\operatorname{tr}\boldsymbol{H} = 1$.

Fix a triple with $\boldsymbol{H} \in \mathbb{S}^q_+$ and $\operatorname{tr}\boldsymbol{H} = 1$, and put $\Delta\boldsymbol{v} = \boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{H}\boldsymbol{z}_1 + \boldsymbol{a} \in L^2$. For every $\alpha > 1$, apply Lemma B.3 with moment bound $\alpha^{-1}\boldsymbol{H}^2$. It gives the exact equivalence

$$\begin{aligned} &(\boldsymbol{H},\boldsymbol{M},\boldsymbol{a}) \text{ is } \alpha\text{-feasible} \\ &\quad\Longleftrightarrow\quad \mathbb{E}\operatorname{dist}^2_{\boldsymbol{\Lambda}}(\Delta\boldsymbol{v},\mathcal{C}_Y) \le \alpha^{-1}\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{H}^2) \quad \text{for every } \boldsymbol{\Lambda} \succeq \boldsymbol{0}. \end{aligned}$$

In particular, the converse supplies one selection $\boldsymbol{c} \in \mathcal{C}_Y$ and hence one correction $\boldsymbol{u} = \boldsymbol{c} - \Delta\boldsymbol{v} \in L^2$ satisfying the matrix bound simultaneously in all directions. Conditional expectation given $(\boldsymbol{z}_0,\boldsymbol{z}_1,Y)$ makes this correction measurable in these variables and preserves both constraints.

For this fixed candidate, each weighting gives a bound on the feasible ratio. Write $d_{\boldsymbol{\Lambda}} = \mathbb{E}\operatorname{dist}^2_{\boldsymbol{\Lambda}}(\Delta\boldsymbol{v},\mathcal{C}_Y)$ and $n_{\boldsymbol{\Lambda}} = \operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{H}^2)$; both are finite and nonnegative, since $\boldsymbol{0} \in \mathcal{C}_Y$ and $\Delta\boldsymbol{v} \in L^2$. If $d_{\boldsymbol{\Lambda}} > 0$, its inequality is equivalent to $\alpha \le n_{\boldsymbol{\Lambda}}/d_{\boldsymbol{\Lambda}}$. If $d_{\boldsymbol{\Lambda}} = 0$, the inequality imposes no restriction, including when $n_{\boldsymbol{\Lambda}} = 0$; this gives exactly the stated $+\infty$ convention, including $0/0$. The zero matrix also imposes no restriction. Every nonzero PSD matrix has positive trace, and both $n_{\boldsymbol{\Lambda}}$ and $d_{\boldsymbol{\Lambda}}$ scale linearly with $\boldsymbol{\Lambda}$. We may therefore restrict to $\operatorname{tr}\boldsymbol{\Lambda} = 1$ and define

$$\beta(\boldsymbol{H},\boldsymbol{M},\boldsymbol{a}) := \inf_{\substack{\boldsymbol{\Lambda}\succeq\boldsymbol{0}\\ \operatorname{tr}\boldsymbol{\Lambda}=1}} \frac{n_{\boldsymbol{\Lambda}}}{d_{\boldsymbol{\Lambda}}}.$$

The triple is thus α-feasible exactly when $1 < \alpha \le \beta(\boldsymbol{H},\boldsymbol{M},\boldsymbol{a})$.

By the distance definition (3), $\alpha_\star$ is the supremum of β over these normalized triples. For $(\boldsymbol{H},\boldsymbol{M},\boldsymbol{a}) = (\boldsymbol{I}_q/q,\boldsymbol{0},\boldsymbol{0})$, comparison with $\boldsymbol{0} \in \mathcal{C}_Y$ gives

$$d_{\boldsymbol{\Lambda}} \le q^{-2}\mathbb{E}[\boldsymbol{z}_1^\top\boldsymbol{\Lambda}\boldsymbol{z}_1] = q^{-2}\operatorname{tr}\boldsymbol{\Lambda} = n_{\boldsymbol{\Lambda}},$$

so this triple has $\beta \ge 1$ and $\alpha_\star \ge 1$. Every $1 < \alpha < \alpha_\star$ has a nontrivial triple with $\beta > \alpha$ and hence a feasible correction; no $\alpha > \alpha_\star$ is feasible. If $\alpha_\star = 1$, the feasible set is empty and has the stipulated value one. This proves (16), without requiring attainment at the boundary. □

B.4 Population boundary without a fitted intercept

Retain the Gaussian vectors and label law of Appendix B.1, with any fixed true intercept $\boldsymbol{b}_0$, and constrain the fitted intercept direction to $\boldsymbol{a} = \mathbf{0}$.

Definition B.4 (Population feasibility)**.** For $\alpha > 1$, a pair $(\boldsymbol{H}, \boldsymbol{M}) \in \mathbb{S}_+^q \times \mathbb{R}^{q\times q}$ is called α-feasible if there exists a square-integrable $\mathbb{R}^q$-valued random vector $\boldsymbol{u}$ such that

$$\boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{H}\boldsymbol{z}_1 + \boldsymbol{u} \in \mathcal{C}_Y \quad \text{almost surely}, \qquad \mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top] \preceq \alpha^{-1}\boldsymbol{H}^2. \tag{19}$$

The witness is nontrivial when $(\boldsymbol{H}, \boldsymbol{M}) \neq (\mathbf{0}, \mathbf{0})$.

Conditional expectation given $(\boldsymbol{z}_0, \boldsymbol{z}_1, Y)$ preserves feasibility and decreases the second-moment matrix, so $\boldsymbol{u}$ may be chosen measurable in these variables. Feasibility is downward closed in $\alpha > 1$. The boundary defined by (5) has the population characterization

$$\alpha_\star^0(\boldsymbol{S}, \boldsymbol{b}_0) = \sup\{\alpha > 1 : \text{a nontrivial } \alpha\text{-feasible pair exists}\}, \tag{20}$$

where the supremum is set to 1 if the set is empty. The following theorem proves this identity. For this model write $\alpha_\star = \alpha_\star^0(\boldsymbol{S}, \boldsymbol{b}_0)$; see Section 2.3 for comparison with the fitted-intercept boundary.

Theorem B.5. *Under the population construction above, the boundary defined by* (5) *satisfies* (20)*. If Y is uniform and independent of the Gaussian features (the null model), the outer optimization in* (5) *can be restricted to $\boldsymbol{M} = \mathbf{0}$.*

Proof of Theorem B.5. With the same true intercept $\boldsymbol{b}_0$ in the label law and candidate domain restricted to $\boldsymbol{a} = \mathbf{0}$, the argument of Appendix B.3 gives (20).

Null-model specialization. In the null model Y is independent of the dummy $\boldsymbol{z}_0$. For fixed $(\boldsymbol{H}, \boldsymbol{M}, \boldsymbol{\Lambda})$ with $\operatorname{tr} \boldsymbol{H} = 1$, the map $\boldsymbol{a} \mapsto \operatorname{dist}_{\boldsymbol{\Lambda}}^2(\boldsymbol{a}, \mathcal{C}_Y)$ is convex. Conditional Jensen, given $(\boldsymbol{z}_1, Y)$, therefore gives

$$\mathbb{E}\operatorname{dist}_{\boldsymbol{\Lambda}}^2(\boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{H}\boldsymbol{z}_1, \mathcal{C}_Y) \geq \mathbb{E}\operatorname{dist}_{\boldsymbol{\Lambda}}^2(\boldsymbol{H}\boldsymbol{z}_1, \mathcal{C}_Y).$$

The numerator does not depend on $\boldsymbol{M}$, so no $\boldsymbol{M} \neq \mathbf{0}$ can increase the inner ratio. Hence the outer supremum is unchanged when restricted to $\boldsymbol{M} = \mathbf{0}$, which proves the stated rank-zero specialization. □

B.5 Nondegeneracy of both boundaries

Proposition B.6. *For every fixed $K \geq 2$, $\boldsymbol{S} \succeq \mathbf{0}$, and $\boldsymbol{b}_0 \in \mathbb{R}^q$, the population boundaries for the two fitting conventions satisfy*

$$1 < \alpha_\star^0(\boldsymbol{S}, \boldsymbol{b}_0) < \infty, \qquad 1 < \alpha_\star(\boldsymbol{S}, \boldsymbol{b}_0) < \infty.$$

Proof of Proposition B.6. Write $\alpha_\star$ for either distance-formula boundary. Let $p_y(\boldsymbol{z}_0) = \mathbb{P}(Y = y \mid \boldsymbol{z}_0)$ and

$$w(\boldsymbol{z}_0) = \prod_{y=0}^{q} p_y(\boldsymbol{z}_0), \qquad \overline{w} = \mathbb{E}w(\boldsymbol{z}_0).$$

The finite softmax probabilities are strictly positive, and $0 < w \leq p_y \leq 1$. Therefore $0 < \overline{w} \leq 1$.

A lower bound strictly larger than one. Fix any $\boldsymbol{\Lambda} \succeq 0$ and $\boldsymbol{v} \in \mathbb{R}^q$. Choose a class $y_\star$ maximizing the score of $\boldsymbol{v}$, so $\boldsymbol{v} \in \mathcal{C}_{y_\star}$ and $\operatorname{dist}_{\boldsymbol{\Lambda}}^2(\boldsymbol{v}, \mathcal{C}_{y_\star}) = 0$. For every other class, zero is a permissible comparison point, giving $\operatorname{dist}_{\boldsymbol{\Lambda}}^2(\boldsymbol{v}, \mathcal{C}_y) \leq \boldsymbol{v}^\top\boldsymbol{\Lambda}\boldsymbol{v}$. Consequently

$$\sum_y p_y(\boldsymbol{z}_0)\operatorname{dist}_{\boldsymbol{\Lambda}}^2(\boldsymbol{v}, \mathcal{C}_y) \leq (1 - p_{y_\star}(\boldsymbol{z}_0))\boldsymbol{v}^\top\boldsymbol{\Lambda}\boldsymbol{v} \leq (1 - w(\boldsymbol{z}_0))\boldsymbol{v}^\top\boldsymbol{\Lambda}\boldsymbol{v}.$$

This is a pointwise finite-sum inequality; no measurable maximizing-class selection is needed. Choose the normalized candidate $\boldsymbol{H} = q^{-1}\boldsymbol{I}_q$, $\boldsymbol{M} = 0$, and $\boldsymbol{a} = 0$, admissible for both fitting conventions. For every trace-one PSD weight, independence of $\boldsymbol{z}_0, \boldsymbol{z}_1$ and the Gaussian covariance identity imply

$$\mathbb{E}\operatorname{dist}_{\boldsymbol{\Lambda}}^2(q^{-1}\boldsymbol{z}_1, \mathcal{C}_Y) \leq q^{-2}(1 - \overline{w}), \qquad \operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{H}^2) = q^{-2}.$$

Put $r = 1 - \overline{w}/2 \in [1/2, 1)$. If the denominator is positive, the corresponding ratio is at least $r^{-1} > 1$; if it vanishes, the stipulated value is $+\infty$ and the same bound holds. Taking the infimum over weights and then the supremum over candidates gives $\alpha_\star \geq r^{-1} > 1$ directly from the distance formula.

A finite upper bound. For $\boldsymbol{v} \in \mathbb{R}^q$, if $v_i \geq 0$, every $\boldsymbol{c} \in \mathcal{C}_0$ has $c_i \leq 0$, so $v_i^2 \leq \operatorname{dist}^2(\boldsymbol{v}, \mathcal{C}_0)$. If $v_i \leq 0$, every $\boldsymbol{c} \in \mathcal{C}_i$ has $c_i \geq 0$, so $v_i^2 \leq \operatorname{dist}^2(\boldsymbol{v}, \mathcal{C}_i)$. Hence, because $q \geq 1$,

$$\|\boldsymbol{v}\|^2 \leq q \operatorname{dist}^2(\boldsymbol{v}, \mathcal{C}_0) + \sum_{i=1}^{q} \operatorname{dist}^2(\boldsymbol{v}, \mathcal{C}_i) \leq q \sum_{y=0}^{q} \operatorname{dist}^2(\boldsymbol{v}, \mathcal{C}_y).$$

For $\Delta\boldsymbol{v} = \boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{H}\boldsymbol{z}_1 + \boldsymbol{a}$, conditioning first on $(\boldsymbol{z}_0, \boldsymbol{z}_1)$ and then integrating over the independent $\boldsymbol{z}_1$ gives

$$\begin{aligned}
\mathbb{E} \operatorname{dist}^2(\Delta\boldsymbol{v}, \mathcal{C}_Y) &\geq q^{-1}\mathbb{E}\big[w(\boldsymbol{z}_0)\|\Delta\boldsymbol{v}\|^2\big] \\
&= q^{-1}\mathbb{E}\big[w(\boldsymbol{z}_0)\{\operatorname{tr}(\boldsymbol{H}^2) + \|\boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{a}\|^2\}\big] \\
&\geq q^{-1}\overline{w}\, \operatorname{tr}(\boldsymbol{H}^2).
\end{aligned}$$

All terms are integrable: $w \leq 1$, the affine Gaussian predictor is square-integrable, and comparison with zero bounds each squared distance by $\|\Delta\boldsymbol{v}\|^2$.

This also bounds the distance formula directly, without requiring its population characterization. For every candidate with $\operatorname{tr} \boldsymbol{H} = 1$, $\operatorname{tr}(\boldsymbol{H}^2) > 0$. Test its inner infimum with $\boldsymbol{\Lambda} = \boldsymbol{I}_q/q$. The numerator is $q^{-1} \operatorname{tr}(\boldsymbol{H}^2)$ and, by homogeneity in the weight, the denominator is at least $q^{-2}\overline{w} \operatorname{tr}(\boldsymbol{H}^2) > 0$. Consequently each candidate value is at most $q/\overline{w}$, and taking the outer supremum gives

$$\alpha_\star \leq q/\overline{w} < \infty.$$

The bound is independent of $\boldsymbol{M}$ and $\boldsymbol{a}$, so it applies to both fitting conventions, for every fixed $\boldsymbol{S} \succeq \boldsymbol{0}$ and $\boldsymbol{b}_0$. □

C Proof strategy for the phase transition

Population feasibility allows a correction depending on each realization; a sample separator must use one common coefficient matrix. The following constructions bridge this difference. We describe the fitted-intercept model; omitting the intercept removes its direction and its balance constraint.

C.1 Below the boundary: constructing a separator

For $1 < \alpha < \alpha_\star$, choose a witness feasible at $\alpha' \in (\alpha, \alpha_\star)$. With $\xi = (\boldsymbol{z}_0, Y)$ and $\boldsymbol{\mu}(\xi) = \boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{a}$, it satisfies

$$\boldsymbol{H}\boldsymbol{z}_1 + \boldsymbol{\mu}(\xi) + \boldsymbol{u} \in \mathcal{C}_Y, \qquad \mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top] \preceq (\alpha^{-1} - \delta)\boldsymbol{H}^2, \qquad \delta = \alpha^{-1} - (\alpha')^{-1} > 0.$$

The moment slack is on $\operatorname{range}(\boldsymbol{H})$; $\boldsymbol{H}$ may be singular. The goal is a matrix $\boldsymbol{A} \in \mathbb{R}^{(d-q)\times q}$ with $\boldsymbol{\mu}(\xi_i) + \boldsymbol{A}^\top \boldsymbol{g}_i \in \mathcal{C}_{Y_i}$ for every row and at least one strict comparison.

Lemma E.1 first mixes the witness with its conditional mean given ξ. Comparisons with zero conditional mean remain exact ties; the others acquire positive conditional margins, while the moment bound retains slack. We then choose buffers: every positive target margin is at least $h/2$, and the fraction of rows with an untied comparison but no positive buffer tends to $p(h) \to 0$ as $h \downarrow 0$. The AMP construction produces one coefficient matrix with average squared distance $O(h^4)$ from these buffered targets (Lemmas D.1 and D.2).

Average error does not ensure every comparison. The repair first enforces all ties and makes the rows with zero buffers strict, then holds those rows fixed while repairing the remaining inequalities. Uniform row-inverse bounds apply simultaneously to all small row sets, including sets selected from the same data used by AMP. The error-to-squared-buffer ratio is $O(h^2)$, so a sufficiently small fixed h permits exact repair. A strict comparison occurs with probability tending to one, giving a nonzero separator and hence nonexistence of the MLE. Appendix E gives the proof.

C.2 Above the boundary: ruling out separation

For $\alpha > \alpha_\star$, the finite-sample criterion instead asks for a matrix $\boldsymbol{P}$ with rows $\boldsymbol{p}_i \in \operatorname{int}(\mathcal{C}_{Y_i}^\circ)$ and $\tilde{\boldsymbol{X}}^\top \boldsymbol{P} = 0$. Such a matrix rules out every nonzero separator: the sum of its row pairings with a separating predictor is both zero by balance and strictly negative by cone polarity and full column rank.

Write $\boldsymbol{T} = [\boldsymbol{Z}_0, \mathbf{1}_n]$, $\boldsymbol{P}_T = \boldsymbol{I}_n - \boldsymbol{T}(\boldsymbol{T}^\top \boldsymbol{T})^{-1}\boldsymbol{T}^\top$, and $s_n = n - d - 1$. For $\boldsymbol{\Gamma}$ independent of the signal and labels, the following subspaces have the same conditional law:

$$\mathcal{N}_n = \ker([\boldsymbol{T}, \boldsymbol{G}]^\top), \qquad \operatorname{col}(\boldsymbol{P}_T \boldsymbol{\Gamma}), \quad \boldsymbol{\Gamma} \in \mathbb{R}^{n \times s_n} \text{ iid Gaussian.}$$

Both lie in $\operatorname{col}(\boldsymbol{T})^\perp$. The available dimension fraction is $s_n/n \to \rho = 1 - \alpha^{-1}$. Lemma F.3 supplies a population polar vector $\boldsymbol{p} = \boldsymbol{H}_\circ \boldsymbol{z} + \boldsymbol{w}$ with exact side balance and $\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top] \preceq (\rho - \delta)\boldsymbol{H}_\circ^2$ for some $\delta > 0$. It is the softmax gradient at a population minimizer; stationarity supplies the balance and moment identities.

Apply AMP at dimension fraction ρ, project by $\boldsymbol{P}_T$ to enforce sample balance, and repair within the same projected design. Projection costs only $O(h^2)$ in root-mean-square, so the buffered repair still applies. The resulting matrix has every row in its polar-cone interior and every column in $\operatorname{col}(\boldsymbol{P}_T \boldsymbol{\Gamma})$. The conditional subspace law transfers its existence probability to $\mathcal{N}_n$, proving MLE existence with probability tending to one. Appendices F.3 and G.3 establish this construction and transfer, including measurability.

D Gaussian approximation and exact repair

Proofs of the AMP, buffered-repair, and tensor-Hall tools are collected in Appendix H.

The two sample constructions need the same passage from a population witness to one coefficient matrix. An average approximation error alone does not suffice: separation and its exclusion certificate both require constraints on every row. We first construct an approximate matrix, then correct its row violations while preserving the required equalities. Lemma D.1 supplies the Gaussian construction, and Lemma D.2 applies it to buffered cone constraints. Lemma D.4 enforces the required equalities and repairs rows without buffers; Lemma D.5 then repairs the remaining inequalities while preserving those rows.

Both repair steps require bounded right inverses for the sample row maps. For the separator, Lemma D.8 derives the tensor-Hall condition from the population second-moment gap, and Theorem D.6 and Lemma D.7 give the required uniform bounds. For the polar certificate, projection removes the side coordinates, and Lemma F.4 supplies the corresponding bounds for the projected design. Appendices E and F assemble these tools into the separator theorem Theorem E.3 and the polar-certificate theorem Theorem F.5.

D.1 Gaussian construction by approximate message passing

We construct a sample coefficient matrix whose loss, ridge penalty, and optional balance penalty approach a population upper bound. The construction uses proximal vector approximate message passing (AMP), as in Çakmak et al. (2024). We give the population optimality, finite-depth state evolution, and contraction arguments separately, then combine them in the proof of Lemma D.1. The population optimality conditions specify the AMP denoiser and its stationary covariance. State evolution describes any fixed number of iterations; contraction then makes the limiting approximation error small. For fixed regularization parameters, we choose a finite iteration count and a positive perturbation before taking the sample dimension to infinity. The final step removes the auxiliary randomness, leaving coefficients measurable in the data.

The side vector ξ collects the label and signal information, while the rows of $\boldsymbol{G}$ contain the independent Gaussian features. In the applications below, $\ell(\boldsymbol{v}; \xi)$ is a smooth penalty for violating cone constraints; the later repair lemmas make those constraints exact. We use $\boldsymbol{A}$ for the population Gaussian factor and $\boldsymbol{W}$ for the sample coefficient matrix. Vector norms are Euclidean, and matrix norms marked F are Frobenius.

Lemma D.1. Loss and side information. *Fix integers $r \geq 1$, $p, s \geq 0$ and constants $\kappa, \lambda > 0$, $\tau \geq 0$. Let $\xi \in \mathbb{R}^s$ be a side vector and let $\ell(\boldsymbol{v}; \xi) \geq 0$ be convex and C^3 in $\boldsymbol{v} \in \mathbb{R}^r$. The loss and its first three predictor derivatives are jointly measurable; the Hessian and third derivative are uniformly bounded. Assume that $\ell(\mathbf{0}; \xi)$ and $\|\nabla \ell(\mathbf{0}; \xi)\|$ are bounded almost surely. Let $\boldsymbol{b}(\xi) \in \mathbb{R}^p$ have all finite moments; $p = 0$ means that the balance term is absent. The vector $\boldsymbol{b}(\xi)$ specifies the side coordinates whose empirical cross-moments with the constructed predictors will be penalized; setting $\tau = 0$ removes this penalty.*

Population problem. *For $\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{I}_r)$ independent of ξ, define*

$$J_\star = \min_{\substack{\boldsymbol{A} \in \mathbb{R}^{r \times r},\, \boldsymbol{u} \in L^2(\boldsymbol{z}, \xi; \mathbb{R}^r) \\ \mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top] \preceq \kappa \boldsymbol{A}\boldsymbol{A}^\top}} \left\{ \mathbb{E}\ell(\boldsymbol{A}\boldsymbol{z} + \boldsymbol{u}; \xi) + \frac{\lambda}{2}\|\boldsymbol{A}\|_F^2 + \frac{\tau}{2}\|\mathbb{E}[\boldsymbol{b}(\boldsymbol{A}\boldsymbol{z} + \boldsymbol{u})^\top]\|_F^2 \right\}.$$

The minimum is attained. Here $\boldsymbol{A}\boldsymbol{z}$ is the Gaussian contribution and $\boldsymbol{u}$ is its population correction; the matrix constraint limits the correction's second moment.

Sample conclusion. *Let $m = m_n$ be positive integers with $m/n \to \kappa$, and let $\boldsymbol{G} \in \mathbb{R}^{n \times m}$ have iid standard Gaussian entries, independently of iid copies ξ_i of ξ. Write $\boldsymbol{g}_i^\top$ for row i of $\boldsymbol{G}$ and $\boldsymbol{b}_i = \boldsymbol{b}(\xi_i)$. For every $\varepsilon > 0$, there are coefficients $\boldsymbol{W} = \boldsymbol{W}_n \in \mathbb{R}^{m \times r}$ measurable in the design and side sample such that, with probability tending to one,*

$$\frac{1}{n}\sum_i \ell(\boldsymbol{W}^\top \boldsymbol{g}_i; \xi_i) + \frac{\lambda}{2}\|\boldsymbol{W}\|_F^2 + \frac{\tau}{2}\left\|\frac{1}{n}\sum_i \boldsymbol{b}_i(\boldsymbol{W}^\top \boldsymbol{g}_i)^\top\right\|_F^2 \leq J_\star + \varepsilon. \tag{21}$$

Equivalently, for every fixed $\varepsilon > 0$ there is such a sequence $(\boldsymbol{W}_n)$ for which, for every $\eta > 0$, the probability of (21) is at least $1 - \eta$ for all sufficiently large n. All population parameters, including λ, are fixed in this assertion. The predictors use one common coefficient matrix; no statement about the distribution of a sample optimizer is needed.

D.2 Buffered approximation and exact repair

We now specialize the approximation to class comparisons. A comparison is *active* if it must remain an equality; the other comparisons must have positive margins. A *buffer* is a positive target margin that leaves room for the predictor to move during repair. We first enforce the equalities and make the rows without positive buffers strict. We then hold those rows fixed while correcting the buffered rows. Uniform right-inverse bounds allow both corrections to use the same Gaussian design as the approximation.

The notation below expresses a population predictor as $\boldsymbol{\mu}(\xi) + \boldsymbol{D}(\boldsymbol{z} + \boldsymbol{w})$. In the separator application, $\boldsymbol{\mu}(\xi) = \boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{a}$ is the signal and intercept contribution, $\boldsymbol{D}$ maps the residual coordinates into the class predictors, and $\boldsymbol{w}$ is the correction in these residual coordinates. To connect this notation to $\boldsymbol{H}\boldsymbol{z}_1 + \boldsymbol{u}$ in Appendix C.1, retain all q Gaussian coordinates and set

$$r = q, \qquad \boldsymbol{D} = \boldsymbol{H}, \qquad \boldsymbol{z} = \boldsymbol{z}_1, \qquad \boldsymbol{w} = \boldsymbol{H}^\dagger \boldsymbol{u}.$$

Let $\boldsymbol{P} = \boldsymbol{H}^\dagger \boldsymbol{H}$, the orthogonal projector onto $\mathrm{range}(\boldsymbol{H})$. The moment bound forces $\boldsymbol{u} \in \mathrm{range}(\boldsymbol{H})$ almost surely. Therefore

$$\boldsymbol{H}\boldsymbol{z}_1 + \boldsymbol{u} = \boldsymbol{H}(\boldsymbol{z}_1 + \boldsymbol{w}), \qquad \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top] \preceq (\kappa - \delta)\boldsymbol{P} \preceq (\kappa - \delta)\boldsymbol{I}_q.$$

Here $\kappa - \delta \geq 0$, as follows from the assumed moment inequality and $\boldsymbol{H} \neq 0$. This representation covers singular $\boldsymbol{H}$ without changing the latent dimension. The side vectors in the two applications are $(\boldsymbol{z}_0, Y)$ and $(\boldsymbol{z}_0, Y, U)$, respectively; discrete labels are represented by real coordinates. The projected certificate application uses the finite-cone version stated below.

Fix integers $q, r \geq 1$ and a nonzero matrix $\boldsymbol{D} \in \mathbb{R}^{q \times r}$. Let $\xi \in \mathbb{R}^s$ be a finite-dimensional side vector determining a class $Y(\xi) \in \{0, 1, \ldots, q\}$, a square-integrable vector $\boldsymbol{\mu}(\xi) \in \mathbb{R}^q$, and a measurable set $J(\xi) \subseteq [q]$. Let $\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{I}_r)$ be independent of ξ. For $\kappa \in (0, 1)$, $\delta > 0$, a correction $\boldsymbol{w} = \boldsymbol{w}(\boldsymbol{z}, \xi) \in L^2(\Omega; \mathbb{R}^r)$, and measurable margins $\mathrm{gap}_\ell(\xi) > 0$, $\ell \in [q] \setminus J(\xi)$, consider the conditions

$$\boldsymbol{a}_{Y(\xi),\ell}^\top \{\boldsymbol{\mu}(\xi) + \boldsymbol{D}(\boldsymbol{z} + \boldsymbol{w})\} = 0, \qquad \ell \in J(\xi), \tag{22}$$

$$\boldsymbol{a}_{Y(\xi),\ell}^\top \{\boldsymbol{\mu}(\xi) + \boldsymbol{D}(\boldsymbol{z} + \boldsymbol{w})\} \geq \mathrm{gap}_\ell(\xi) > 0, \qquad \ell \in [q] \setminus J(\xi), \tag{23}$$

$$\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top] \preceq (\kappa - \delta)\boldsymbol{I}_r. \tag{24}$$

These conditions require exact ties, positive margins on the remaining comparisons, and a positive gap δ in the second-moment bound, respectively. Each result below specifies which conditions it requires.

For $\boldsymbol{v} \in \mathbb{R}^r$ and $\ell \in [q]$, write

$$s_\ell(\boldsymbol{v}, \xi) := \boldsymbol{a}_{Y(\xi),\ell}^\top \{\boldsymbol{\mu}(\xi) + \boldsymbol{D}\boldsymbol{v}\}.$$

The active equations act through the subspace

$$N_{\mathrm{act}}(\xi) := \mathrm{span}\{\boldsymbol{D}^\top \boldsymbol{a}_{Y(\xi),\ell} : \ell \in J(\xi)\} \subseteq \mathbb{R}^r.$$

For any measurable nonnegative buffer vector $\boldsymbol{b}(\xi) = (b_\ell(\xi))_{\ell\in[q]}$ that vanishes on $J(\xi)$, define the shifted latent face

$$F_{\boldsymbol{b}}(\xi) := \{\boldsymbol{v} \in \mathbb{R}^r : s_\ell(\boldsymbol{v}, \xi) = 0 \ (\ell \in J(\xi)), s_\ell(\boldsymbol{v}, \xi) \geq b_\ell(\xi) \ (\ell \in [q] \setminus J(\xi))\}.$$

Write $F_{\boldsymbol{0}}(\xi)$ for the choice with every buffer equal to zero.

For iid side observations $\xi_1, \ldots, \xi_n$, write $Y_i = Y(\xi_i)$, $J_i = J(\xi_i)$, and $N_{\mathrm{act},i} = N_{\mathrm{act}}(\xi_i)$.

The next lemma applies Lemma D.1 to smoothed squared comparison violations. It produces one coefficient matrix with small average squared distance to the buffered constraints. This is still an approximate solution: a small average error does not imply that every observation satisfies its constraints. The optional balance condition is used for the polar certificate in Appendix F, where the sample vectors must also be orthogonal to the side coordinates.

Lemma D.2 (Buffered approximation with optional balance)**.** Population assumptions and sample design. *Under (22)–(24), let $\boldsymbol{G}_{\mathrm{main}} \in \mathbb{R}^{n\times m_0}$ have iid standard Gaussian entries, independently of the side sample, with $m_0/n \to \kappa_0 > \kappa - \delta$. Write $\boldsymbol{g}_{\mathrm{main},i}^\top$ for row i of $\boldsymbol{G}_{\mathrm{main}}$. The inequality $\kappa_0 > \kappa - \delta$ leaves a positive gap between the available feature ratio and the population correction's second-moment bound.*

Optional balance condition. *Optionally, fix $p \geq 1$ and a measurable $\boldsymbol{b}_{\mathrm{bal}}(\xi) \in \mathbb{R}^p$ having all finite moments and satisfying*

$$\mathbb{E}[\boldsymbol{b}_{\mathrm{bal}}(\xi)(\boldsymbol{z} + \boldsymbol{w}(\boldsymbol{z}, \xi))^\top] = \boldsymbol{0}. \tag{25}$$

Without balance take $p = 0$ and an empty feature vector.

Approximate sample solution. *For every fixed $0 < h \leq 1$, there are measurable buffers $\boldsymbol{b}_h = (b_{h,\ell})_{\ell\in[q]}$ and coefficient matrices $\boldsymbol{W}^{(0)} \in \mathbb{R}^{m_0\times r}$ measurable in the main block and side sample such that, writing $b_{i\ell} = b_{h,\ell}(\xi_i)$ and $\boldsymbol{b}_i = (b_{i\ell})_{\ell\in[q]}$, with probability tending to one,*

$$\|\boldsymbol{W}^{(0)}\|_F^2 \leq r + Ch^4, \tag{26}$$

$$\frac{1}{n}\sum_i \mathrm{dist}^2((\boldsymbol{W}^{(0)})^\top \boldsymbol{g}_{\mathrm{main},i}, F_{\boldsymbol{b}_i}(\xi_i)) \leq Ch^4, \tag{27}$$

$$\left\| \frac{1}{n}\sum_i \boldsymbol{b}_{\mathrm{bal}}(\xi_i)((\boldsymbol{W}^{(0)})^\top \boldsymbol{g}_{\mathrm{main},i})^\top \right\|_F \leq Ch^2 \quad (p \geq 1). \tag{28}$$

The buffers vanish on active facets. On every inactive facet they are either zero or at least $h/2$, and satisfy

$$0 \leq b_{i\ell} \leq \mathrm{gap}_\ell(\xi_i)/2. \tag{29}$$

The constant C is independent of h. The fraction of rows having an inactive facet with zero buffer converges in probability to a deterministic $p(h)$, where $p(h) \to 0$ as $h \downarrow 0$. The scale h remains fixed throughout each sample-size limit.

The root-mean-square approximation error is $O(h^2)$, while every positive buffer is at least $h/2$. For any fixed $c > 0$, (27) therefore bounds the fraction of rows at distance greater than ch from their buffered target by $O(h^2)$. Rows with a zero inactive buffer form a separate fraction converging to $p(h)$; the repair procedure handles them first. Both this fraction and the ratio of squared error to squared buffer can be made as small as needed by choosing one fixed h.

Remark D.3 (Approximation for finite cone families)**.** The same lemma holds when the winner normals $\boldsymbol{a}_{Y,\ell}$ are replaced by the facet normals of any fixed finite family of full-dimensional pointed simplicial cones. Its proof uses only finitely many constraint matrices, so both derivative and Hoffman bounds remain uniform. For Theorem F.5, take $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{D} = \boldsymbol{H}$, $J = \varnothing$, and $\boldsymbol{a}_{Y,\ell} = \boldsymbol{d}_{Y,\ell}$.

Deterministic exact repair of buffered constraints.

The next two lemmas repair a common coefficient matrix in two stages. First, enforce every required equality and make the comparisons strict on the exceptional rows that lack buffers. Then hold those rows fixed while repairing the remaining inequalities. Both lemmas are deterministic and apply also to projected designs.

In their statements, A_i is the affine space of required equalities, C_i imposes the additional nonnegative comparisons, and F_i is a buffered target set. The right-inverse assumptions mean that prescribed changes in the indicated row coordinates can be achieved by a small change to the common coefficient matrix. The Gaussian estimates in the next subsection verify these assumptions in our applications.

Lemma D.4 (Enforcing equalities and repairing rows without buffers)**.** Design and constraints. *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ satisfy $\|\boldsymbol{X}\|_{\mathrm{op}} \le M\sqrt{n}$, and write $\boldsymbol{e}_i(\boldsymbol{W}) = (\boldsymbol{X}\boldsymbol{W})_{i\cdot}^{\top}$ for $\boldsymbol{W} \in \mathbb{R}^{p\times r}$. For each i, let $A_i \subseteq \mathbb{R}^r$ be a nonempty affine space with normal space N_i, let $C_i = \{\boldsymbol{v} \in A_i : s_{i\ell}(\boldsymbol{v}) \ge 0 \text{ for all } \ell\}$ for finitely many affine functions, and let $F_i \subseteq C_i$ be nonempty, closed and convex. Fix a set $L \subseteq [n]$ and, for each $i \in L$, a point $\boldsymbol{v}_i^{\circ} \in C_i$ with all its displayed inequalities strict. The set L contains the rows to be made strictly feasible in this step. Set $U_i = \mathbb{R}^r$ on L and $U_i = N_i$ elsewhere: we control the entire row on L, and only the equality constraints on the other rows.*

Ability to correct the rows. *Suppose the row map $\boldsymbol{D} \mapsto (\boldsymbol{P}_{U_i}\boldsymbol{e}_i(\boldsymbol{D}))_i$ is onto $\bigoplus_i U_i$ and has a right inverse of norm at most $K/\sqrt{n}$.*

Conclusion. *If $h > 0$, $\eta \ge 0$, and $\boldsymbol{W}_0$ satisfies $\sum_i \mathrm{dist}^2(\boldsymbol{e}_i(\boldsymbol{W}_0), F_i) \le n\eta^2$, there is $\boldsymbol{W}_1$ such that all $\boldsymbol{e}_i(\boldsymbol{W}_1)$ belong to A_i, every row in L satisfies its inequalities strictly, and*

$$\|\boldsymbol{W}_1 - \boldsymbol{W}_0\|_F \le \sqrt{2}K(\eta + h^2), \qquad \sum_i \mathrm{dist}^2(\boldsymbol{e}_i(\boldsymbol{W}_1), F_i) \le Cn(\eta^2 + h^4), \tag{30}$$

where $C = 2 + 4M^2K^2$.

Lemma D.5 (Repairing buffered inequalities while preserving fixed rows)**.** Constraints to preserve. *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ satisfy $\|\boldsymbol{X}\|_{\mathrm{op}} \le M\sqrt{n}$, and write $\boldsymbol{e}_i(\boldsymbol{W}) = (\boldsymbol{X}\boldsymbol{W})_{i\cdot}^{\top}$ for $\boldsymbol{W} \in \mathbb{R}^{p\times r}$. For every i, let $A_i \subseteq \mathbb{R}^r$ be a nonempty affine subspace with normal space N_i. Let $L, S \subseteq [n]$ be disjoint sets of locked and strong rows. Fix $\boldsymbol{t}_i \in A_i$ for $i \in L$, and suppose the affine space*

$$\mathcal{A}_{\mathrm{fix}} = \{\boldsymbol{W} : \boldsymbol{e}_i(\boldsymbol{W}) \in A_i \ (i \in [n]), \quad \boldsymbol{e}_i(\boldsymbol{W}) = \boldsymbol{t}_i \ (i \in L)\}$$

is nonempty. Thus rows in L must remain fixed, and every row must continue to satisfy its equality constraints. The set S contains the rows whose buffered inequalities remain to be repaired.

Buffered targets and initial error. *On each $i \in S$, let $s_{i\ell}(\boldsymbol{v}) = \boldsymbol{a}_{i\ell}^{\top}\boldsymbol{v} + c_{i\ell}$ be finitely many affine facet functions with $\|\boldsymbol{a}_{i\ell}\| \le L_a$, where $L_a \ge 1$. Fix $h > 0$ and $b_{i\ell} \ge h/2$, and suppose*

$$F_i = \{\boldsymbol{v} \in A_i : s_{i\ell}(\boldsymbol{v}) \ge b_{i\ell} \text{ for all } \ell\}$$

is nonempty. Put

$$H_i = \{\boldsymbol{v} \in A_i : s_{i\ell}(\boldsymbol{v}) \ge b_{i\ell}/2 + h/8 \text{ for all } \ell\}.$$

Suppose $\boldsymbol{W}_1 \in \mathcal{A}_{\mathrm{fix}}$ and $\eta \ge 0$ satisfy

$$\sum_{i\in S} \mathrm{dist}^2(\boldsymbol{e}_i(\boldsymbol{W}_1), F_i) \le n\eta^2. \tag{31}$$

Ability to correct small sets of rows. *For $B \subseteq S$, define $U_i(B) = \mathbb{R}^r$ on $L \cup B$ and $U_i(B) = N_i$ elsewhere, and let*

$$\mathcal{T}_B\boldsymbol{D} = \big(\boldsymbol{P}_{U_i(B)}\boldsymbol{e}_i(\boldsymbol{D})\big)_{i\in[n]}.$$

Assume that, for every $B \subseteq S$ with $|L \cup B| \le \rho n$, this row map is onto $\bigoplus_i U_i(B)$ and has a right inverse of norm at most $K/\sqrt{n}$, where $K > 0$.

Conclusion. *If*

$$|L| + C_B n\eta^2/h^2 \le \rho n, \qquad C_B := 128L_a^2(1+4M^2K^2), \tag{32}$$

then there is $\boldsymbol{W}_\star \in \mathcal{A}_{\mathrm{fix}}$ *such that*

$$\boldsymbol{e}_i(\boldsymbol{W}_\star) \in H_i \quad (i \in S), \qquad \|\boldsymbol{W}_\star - \boldsymbol{W}_1\|_F \le 2K\eta.$$

In particular, every displayed margin on S *is at least* $3h/8 > 0$*, all rows in* L *are unchanged, and all equality constraints are preserved.*

D.3 A QUANTITATIVE TENSOR-HALL ESCAPE LEMMA

Repairing the linear predictors requires solving many rowwise equations with one shared coefficient matrix. The following result gives a right inverse with controlled norm when the equations leave enough feature directions in every contrast subspace. This directional counting condition is the tensor-Hall condition. We then derive this condition from the population second-moment gap.

For an integer $n \ge 1$ and subspaces $U_1, \dots, U_n \le \mathbb{R}^r$, write $\boldsymbol{x}_i^\top$ for row i of $\boldsymbol{X} \in \mathbb{R}^{n\times r}$ and define

$$\mathcal{E}(U_1, \dots, U_n) := \{\boldsymbol{X} \in \mathbb{R}^{n\times r} : \boldsymbol{x}_i \in U_i \text{ for every } i \in [n]\}.$$

For an integer $p \ge 1$, if $\boldsymbol{G} = (\boldsymbol{g}_1^\top; \dots; \boldsymbol{g}_n^\top) \in \mathbb{R}^{n\times p}$, define

$$\mathcal{A}_{\boldsymbol{G}} : \mathbb{R}^{p\times r} \longrightarrow \mathcal{E}(U_1, \dots, U_n), \qquad \mathcal{A}_{\boldsymbol{G}}(\boldsymbol{W}) := \big(\boldsymbol{P}_{U_i}\boldsymbol{W}^\top \boldsymbol{g}_i\big)_{i\in[n]}.$$

The vectors on the right are stacked as transposed rows. Under the natural Hilbert-space identification of $\bigoplus_{i=1}^n U_i$ with $\mathcal{E}(U_1, \dots, U_n)$, its adjoint is $\mathcal{A}_{\boldsymbol{G}}^* \boldsymbol{X} = \boldsymbol{G}^\top \boldsymbol{X}$. Set

$$s_{\mathrm{row}}(\mathcal{A}_{\boldsymbol{G}}) := \inf_{\substack{\boldsymbol{X}\in\mathcal{E}(U_1,\dots,U_n)\\ \|\boldsymbol{X}\|_F=1}} \|\boldsymbol{G}^\top \boldsymbol{X}\|_F.$$

If $\mathcal{E}(U_1, \dots, U_n) = \{\boldsymbol{0}\}$, this infimum is understood to be $+\infty$.

The quantity $s_{\mathrm{row}}(\mathcal{A}_{\boldsymbol{G}})$ is the smallest singular value of the adjoint row map. A lower bound of order $\sqrt{n}$ says, equivalently, that any admissible collection of active-row residuals has a coefficient correction with norm $O(n^{-1/2})$ times the residual norm. Later, we *upgrade* a row by replacing U_i with $\mathbb{R}^r$: the correction must then control the entire row, so that row can be *locked* and left unchanged. The uniformity over upgrade sets will allow the violated rows to be chosen by a convex optimization problem.

Theorem D.6. *Fix an integer* $r \ge 1$*, a finite family* $\mathcal{U}$ *of subspaces of* $\mathbb{R}^r$*,* $\varepsilon > 0$*, and* $0 < \underline{\kappa} \le \overline{\kappa} < \infty$*. For each integer* $n \ge 1$*, let* $m = m_n \ge 1$ *be an integer, let* $U_1, \dots, U_n \in \mathcal{U}$*, and let* $\boldsymbol{G} = (\boldsymbol{g}_1^\top; \dots; \boldsymbol{g}_n^\top) \in \mathbb{R}^{n\times m}$ *have iid standard Gaussian entries. Suppose*

$$\underline{\kappa} \le m/n \le \overline{\kappa}$$

and

$$\sum_{i=1}^n \dim(U_i \cap S) \le (m - \varepsilon n)\dim S \qquad \textit{for every } S \le \mathbb{R}^r. \tag{33}$$

There are $c, C > 0$ *such that, for all sufficiently large* n*,*

$$\mathbb{P}\{s_{\mathrm{row}}(\mathcal{A}_{\boldsymbol{G}}) \ge c\sqrt{n}\} \ge 1 - Ce^{-cn}. \tag{34}$$

In particular, on this event every collection of desired row changes $\boldsymbol{r}_i \in U_i$ *can be realized by one coefficient correction* $\Delta\boldsymbol{W} \in \mathbb{R}^{m\times r}$ *satisfying*

$$\boldsymbol{P}_{U_i}(\Delta\boldsymbol{W})^\top \boldsymbol{g}_i = \boldsymbol{r}_i \quad (i \in [n]), \qquad \|\Delta\boldsymbol{W}\|_F \le \frac{1}{c\sqrt{n}}\left(\sum_i \|\boldsymbol{r}_i\|^2\right)^{1/2}.$$

Thus $\mathcal{A}_{\boldsymbol{G}}$ *is surjective and has the bounded right inverse required by the repair lemmas.*

Lemma D.7 (Uniform control after prescribing a small set of entire rows)**.** *Under the hypotheses of Theorem D.6, there are* $\rho, c, C > 0$ *such that the following holds. For* $I \subseteq [n]$*, replace* U_i *by* $\mathbb{R}^r$ *on*

I and leave it unchanged off I, and denote the new operator by $\mathcal{A}_{\boldsymbol{G}}^{I}$. This replacement allows the entire output on rows in I to be prescribed, rather than only its projection onto U_i. Then

$$\mathbb{P}\left\{\inf_{\substack{I\subseteq[n]\\|I|\le\rho n}} s_{\mathrm{row}}(\mathcal{A}_{\boldsymbol{G}}^{I}) \ge c\sqrt{n}\right\} \ge 1 - Ce^{-cn}.$$

The event holds simultaneously for all $|I| \le \rho n$, so the rows in I may be selected after observing the Gaussian design. This is the uniform control needed to repair whichever rows violate the constraints.

The second-moment gap implies tensor Hall.

Recall the active-equation subspace $N_{\mathrm{act}}(\xi)$ defined in the shared notation. Under (22), the affine active equations determine $\boldsymbol{P}_{N_{\mathrm{act}}(\xi)}(\boldsymbol{z}+\boldsymbol{w})$ as a measurable function of ξ; denote it by $\boldsymbol{h}_{\mathrm{act}}(\xi)$. The following observation is the link between the population second-moment bound and Theorem D.6.

For the face-restricted construction, set κ_0 and choose positive integers $m_0 = m_0(n)$ such that

$$\kappa_0 := \kappa - \delta/4, \qquad m_0(n)/n \longrightarrow \kappa_0. \tag{35}$$

Lemma D.8 (The second-moment gap leaves room to solve the sample equalities)**.** Population bound on equality constraints. *Under (22) and (24), for every subspace $S \le \mathbb{R}^r$,*

$$\mathbb{E}\dim\{N_{\mathrm{act}}(\xi)\cap S\} \le (\kappa-\delta)\dim S. \tag{36}$$

Sample consequence. *Use $N_{\mathrm{act},i} = N_{\mathrm{act}}(\xi_i)$ from the shared notation and κ_0, m_0 from (35). Then there exists a deterministic $\varepsilon_0 > 0$ such that, with probability tending to one, simultaneously for every $S \le \mathbb{R}^r$,*

$$\sum_{i=1}^{n} \dim(N_{\mathrm{act},i}\cap S) \le (m_0 - \varepsilon_0 n)\dim S. \tag{37}$$

This is precisely the directional counting condition in Theorem D.6, with $m = m_0$ and $U_i = N_{\mathrm{act},i}$. It provides the link from the population second-moment gap to the sample right inverses used in the repair lemmas.

E Below the boundary: constructing a separator

We now prove the separator construction explained in Appendix C.1. First we strengthen the population margins while preserving tied comparisons. We then apply the approximation and repair tools from Appendix D to obtain an exact sample separator and hence nonexistence of the finite MLE below the boundary.

E.1 Creating positive margins while preserving ties

A singular second-moment bound confines corrections to the accessible range. Conditional averaging identifies the facet equalities that must be preserved and creates positive margins on the others.

We state the lemma in an abstract form used in the trapping theorem. Fix $r \in [q]$, let $\boldsymbol{D} \in \mathbb{R}^{q\times r}$ be nonzero, let $\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{I}_r)$, and let ξ be side information independent of $\boldsymbol{z}$. The side information contains the class $Y = Y(\xi) \in \{0, 1, \dots, q\}$ and a square-integrable measurable vector $\boldsymbol{\mu}(\xi) \in \mathbb{R}^q$. Let the target cones have the facet representation (8).

Lemma E.1 (Conditional-face strictification)**.** *Assume that a square-integrable measurable correction $\boldsymbol{w} = \boldsymbol{w}(\boldsymbol{z}, \xi) \in L^2(\Omega; \mathbb{R}^r)$ satisfies*

$$\boldsymbol{\mu}(\xi) + \boldsymbol{D}\{\boldsymbol{z} + \boldsymbol{w}(\boldsymbol{z}, \xi)\} \in \mathcal{C}_{Y(\xi)} \quad \text{a.s.}$$

and, for some $\delta > 0$ and $\kappa \in (0, 1)$,

$$\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top] \preceq (\kappa - \delta)\boldsymbol{I}_r.$$

Then there exist a correction $\boldsymbol{w}_\circ = \boldsymbol{w}_\circ(\boldsymbol{z}, \xi) \in L^2(\Omega; \mathbb{R}^r)$, a number $\delta_\circ > 0$, a measurable random subset $J = J(\xi) \subseteq [q]$, and measurable margins $\mathrm{gap}_\ell(\xi) > 0$ for $\ell \in [q] \setminus J(\xi)$ such that

$$\mathbb{E}[\boldsymbol{w}_\circ\boldsymbol{w}_\circ^\top] \preceq (\kappa - \delta_\circ)\boldsymbol{I}_r, \tag{38}$$

while, almost surely,

$$\begin{aligned} \boldsymbol{a}_{Y,\ell}^\top\{\boldsymbol{\mu}(\xi) + \boldsymbol{D}(\boldsymbol{z} + \boldsymbol{w}_\circ)\} &= 0, && \ell \in J(\xi), \\ \boldsymbol{a}_{Y,\ell}^\top\{\boldsymbol{\mu}(\xi) + \boldsymbol{D}(\boldsymbol{z} + \boldsymbol{w}_\circ)\} &\geq \mathrm{gap}_\ell(\xi) > 0, && \ell \in [q] \setminus J(\xi), \end{aligned} \tag{39}$$

Moreover, with positive probability $J(\xi) \neq [q]$.

Proof. Put $\boldsymbol{e} = \boldsymbol{z} + \boldsymbol{w}$, $\bar{\boldsymbol{e}} = \mathbb{E}[\boldsymbol{e} \mid \xi]$, and

$$s_\ell(\xi) = \mathbb{E}[\boldsymbol{a}_{Y,\ell}^\top(\boldsymbol{\mu} + \boldsymbol{D}\boldsymbol{e}) \mid \xi] = \boldsymbol{a}_{Y,\ell}^\top(\boldsymbol{\mu} + \boldsymbol{D}\bar{\boldsymbol{e}}), \qquad J(\xi) = \{\ell : s_\ell(\xi) = 0\}.$$

All slacks are integrable and nonnegative. Choose versions of these finitely many conditional expectations on a common probability-one set. Nonnegativity and the tower property give

$$\boldsymbol{a}_{Y,\ell}^\top(\boldsymbol{\mu} + \boldsymbol{D}\boldsymbol{e}) = 0 \quad \text{a.s. on } \{\ell \in J(\xi)\}. \tag{40}$$

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

$$\boldsymbol{e}_t = (1 - t)\boldsymbol{e} + t\bar{\boldsymbol{e}}, \qquad \boldsymbol{w}_t = \boldsymbol{e}_t - \boldsymbol{z}. \tag{41}$$

The affine facet slacks at $\boldsymbol{e}_t$ are zero on $J(\xi)$ and at least $ts_\ell(\xi) > 0$ off $J(\xi)$. Since conditional expectation is an L^2 contraction, $\boldsymbol{w}_t \to \boldsymbol{w}$ in L^2, so

$$\|\mathbb{E}[\boldsymbol{w}_t\boldsymbol{w}_t^\top] - \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top]\|_{\mathrm{op}} \to 0.$$

Choose $t > 0$ so this difference is at most $\delta/2$. Then $\boldsymbol{w}_\circ = \boldsymbol{w}_t$, $\delta_\circ = \delta/2$, and $\mathrm{gap}_\ell = ts_\ell$ satisfy all asserted relations.

Finally, if $J(\xi) = [q]$ almost surely, every facet slack vanishes. Pointedness gives $\boldsymbol{D}(\boldsymbol{z} + \boldsymbol{w}) = -\boldsymbol{\mu}$. Choose a nonzero row $\boldsymbol{v}^\top$ of $\boldsymbol{D}$, say row i. Then $\boldsymbol{v}^\top\boldsymbol{w} = -\boldsymbol{v}^\top\boldsymbol{z} - \mu_i(\xi)$, and independence gives

$$\mathbb{E}[(\boldsymbol{v}^\top\boldsymbol{w})^2] = \|\boldsymbol{v}\|^2 + \mathbb{E}[\mu_i(\xi)^2] \geq \|\boldsymbol{v}\|^2.$$

This contradicts $\mathbb{E}[(\boldsymbol{v}^\top\boldsymbol{w})^2] \leq (\kappa - \delta)\|\boldsymbol{v}\|^2$ because $\kappa < 1$ and $\boldsymbol{v} \neq 0$. Hence $J(\xi) \neq [q]$ with positive probability. □

Remark E.2 (Cone-family form)**.** The proof of Lemma E.1 uses only the finite facet representation, convexity, and pointedness of the target cones. As written with exactly q facet indices, it therefore applies verbatim to any fixed finite family of full-dimensional pointed simplicial polyhedral cones after labeling their q facets and replacing the normals $\boldsymbol{a}_{Y,\ell}$ by that family's facet normals. This is the form used in Theorem F.5.

E.2 Face-restricted separator construction

The theorem below constructs coefficients in the residual coordinates used by the shared tools. In the multinomial application, take $\boldsymbol{\mu}(\xi_i) = \boldsymbol{M}\boldsymbol{z}_{0,i} + \boldsymbol{a}$ and $m_n = d - q$. Its matrix $\boldsymbol{W}_n \in \mathbb{R}^{(d-q)\times r}$ gives the sample slope matrix $\boldsymbol{A} = \boldsymbol{W}_n\boldsymbol{D}^\top$ from Appendix C.1, since

$$\boldsymbol{\mu}(\xi_i) + \boldsymbol{D}\boldsymbol{W}_n^\top\boldsymbol{g}_i = \boldsymbol{M}\boldsymbol{z}_{0,i} + \boldsymbol{A}^\top\boldsymbol{g}_i + \boldsymbol{a}.$$

Thus the theorem enforces the desired class comparisons using one common sample coefficient matrix.

Theorem E.3. *Under the population setup and assumptions (22)–(24), assume additionally*

$$\mathbb{P}\{J(\xi) \neq [q]\} > 0. \tag{42}$$

Let $(m_n)_{n\geq 1}$ be a sequence of positive integers with $m_n/n \to \kappa$. For each integer $n \geq 1$, let $\xi_1, \ldots, \xi_n$ be iid copies of ξ, write $Y_i = Y(\xi_i)$ and $J_i = J(\xi_i)$, and, independently, let $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n$ be iid $N(\boldsymbol{0}, \boldsymbol{I}_{m_n})$ vectors. Then, with probability tending to one, there is $\boldsymbol{W}_n \in \mathbb{R}^{m_n\times r}$ such that, for every $i \in [n]$,

$$\boldsymbol{a}_{Y_i,\ell}^\top\{\boldsymbol{\mu}(\xi_i) + \boldsymbol{D}\boldsymbol{W}_n^\top\boldsymbol{g}_i\} = 0, \qquad \ell \in J_i, \tag{43}$$

$$\boldsymbol{a}_{Y_i,\ell}^\top\{\boldsymbol{\mu}(\xi_i) + \boldsymbol{D}\boldsymbol{W}_n^\top\boldsymbol{g}_i\} > 0, \qquad \ell \in [q] \setminus J_i. \tag{44}$$

Moreover, at least one of the latter strict inequalities is present. Consequently, in the multinomial application these linear predictors define a nonzero complete or quasi-complete separator.

Proof. Uniform design bounds. Use the first m_0 columns of the design as $\boldsymbol{G}_{\mathrm{main}}$, where $m_0/n \to \kappa_0 < \kappa$ is specified in (35). Lemmas D.7 and D.8 give fixed $c_0, \rho > 0$ and an event of probability tending to one on which replacing any at most ρn active spaces $N_{\mathrm{act},i}$ by $\mathbb{R}^r$ yields an onto row map with

$$\|\mathcal{A}^{\dagger}_{\boldsymbol{G}_{\mathrm{main}}}\|_{\mathrm{op}} \le (c_0\sqrt{n})^{-1}. \tag{45}$$

Intersect this event with the Gaussian operator bound

$$\|\boldsymbol{G}_{\mathrm{main}}\|_{\mathrm{op}} \le C\sqrt{n}. \tag{46}$$

Uniformity permits row sets chosen from the same design as the AMP candidate.

Initialization. Lemma D.2 gives coefficients $\boldsymbol{W}^{(0)}$, buffers $\boldsymbol{b}_i$, and a weak set L with $|L|/n \to p(h)$. Strong rows have an inactive facet and all inactive buffers positive. Choose h with $p(h) < \mathbb{P}\{J \neq [q]\}/2$; the law of large numbers then gives a positive fraction of strong rows. Apply Lemma D.4 with $\boldsymbol{X} = \boldsymbol{G}_{\mathrm{main}}$, the active affine spaces A_i, $N_i = N_{\mathrm{act},i}$, $C_i = F_{\boldsymbol{0}}(\xi_i)$, and $F_i = F_{\boldsymbol{b}_i}(\xi_i)$. Strict target points on weak rows exist by the population margins (use half of each positive margin). The uniform inverse and (27) yield $\boldsymbol{W}_1$ with all active equations exact, weak rows strict, and total squared distance to F_i at most Cnh^4.

Repair and nontriviality. Apply Lemma D.5 with weak rows locked at their current values and S the strong rows. The outer targets are F_i, and the inner targets have buffers $b_{i\ell}/2 + h/8$. The facet normals $\boldsymbol{D}^{\top}\boldsymbol{a}_{Y_i,\ell}$ have a fixed norm bound; (45)–(46) supply the operator hypotheses. Since $\eta^2 \le Ch^4$, the size condition (32) holds with probability tending to one if also $p(h) < \rho/4$ and $C_B C h^2 < \rho/4$. The inverse constants are fixed before h, which remains fixed as $n \to \infty$. Repair preserves the active equations and weak rows and gives strong rows margins at least $3h/8$. Extend the coefficients by zero on unused design columns. The positive fraction of strong rows ensures at least one strict comparison.

The inverse event is measurable: for each of the finitely many row types and subsets, its adjoint lower bound defines a closed set of designs. The AMP, norm, and empirical-mean events are measurable too. Their intersection has probability tending to one and is contained in the measurable separation event. Only finite samplewise choices of strict targets and repaired coefficients were needed. □

F Above the boundary: excluding separation

We now prove the exclusion-certificate construction explained in Appendix C.2. We first establish population attainment and construct a balanced polar witness. The sample construction must preserve two properties: each row lies in the interior of its polar cone, and the certificate is orthogonal to the side coordinates. Projection enforces the second property but can disrupt the first. The approximation and repair tools from Appendix D restore cone membership within the projected design, so both properties hold simultaneously. The conditional Gaussian comparison in Appendix G.3 then transfers the resulting existence probability to the kernel required by the finite-sample criterion.

F.1 Coercivity and attainment of the population loss

Fix $\alpha > 1$, $\boldsymbol{t} \in L^2(\Omega;\mathbb{R}^p)$, and $Y \in \{0,\dots,q\}$. Let $\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{I}_q)$ be independent of $(\boldsymbol{t}, Y)$. For unrestricted factors $\boldsymbol{A} \in \mathbb{R}^{q\times q}$ and $\boldsymbol{C} \in \mathbb{R}^{q\times p}$, define

$$\mathcal{R}_\alpha(\boldsymbol{A},\boldsymbol{C}) = \inf_{\mathbb{E}[\boldsymbol{u}\boldsymbol{u}^{\top}] \preceq \alpha^{-1}\boldsymbol{A}\boldsymbol{A}^{\top}} \mathbb{E}\ell_Y(\boldsymbol{A}\boldsymbol{z} + \boldsymbol{C}\boldsymbol{t} + \boldsymbol{u}), \qquad \boldsymbol{u} \in L^2. \tag{47}$$

A feasible triple is a nontrivial cone witness if $(\boldsymbol{A},\boldsymbol{C}) \neq (\boldsymbol{0},\boldsymbol{0})$ and $\boldsymbol{A}\boldsymbol{z} + \boldsymbol{C}\boldsymbol{t} + \boldsymbol{u} \in \mathcal{C}_Y$ almost surely. The two models use $\boldsymbol{t} = \boldsymbol{z}_0$ or $\boldsymbol{t} = (\boldsymbol{z}_0^{\top}, 1)^{\top}$. For PSD factors this is the unregularized functional motivated by Asgari et al. (2025, equation (43)); we prove the needed attainment directly.

Lemma F.1. *There is no nontrivial cone witness at ratio α if and only if $\mathcal{R}_\alpha$ is level-bounded in $(\boldsymbol{A},\boldsymbol{C})$. In that case there is $c_\alpha > 0$ such that*

$$\mathcal{R}_\alpha(\boldsymbol{A},\boldsymbol{C}) \ge c_\alpha(\|\boldsymbol{A}\|_F^2 + \|\boldsymbol{C}\|_F^2)^{1/2}, \tag{48}$$

and the joint minimum is attained with a correction measurable in $(\boldsymbol{z},\boldsymbol{t},Y)$. The same statement holds if the factors are restricted to $\boldsymbol{A} \succeq \boldsymbol{0}$, with witnesses restricted accordingly.

Proof. Put $h_y(\boldsymbol{v}) = \sum_{j=0}^{q} \max\{v_j - v_y, 0\}$, with $v_0 = 0$. This nonnegative convex Lipschitz function is positively homogeneous, vanishes exactly on $\mathcal{C}_y$, and satisfies $h_y \le K\ell_y$. On the feasible set with $\|\boldsymbol{A}\|_F^2 + \|\boldsymbol{C}\|_F^2 = 1$, the coefficients are compact and the trace constraint bounds $\boldsymbol{u}$ in L^2. The moment constraint is weakly closed: for every fixed $\boldsymbol{e}$, $\mathbb{E}(\boldsymbol{e}^\top \boldsymbol{u})^2$ is weakly lower semicontinuous and $\boldsymbol{e}^\top \boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{e}$ is continuous in the coefficients. The predictor map is bounded linear, so its expected h_Y is convex and norm-continuous, hence weakly lower semicontinuous. It therefore attains a minimum $\eta > 0$ on this normalized set when no nontrivial cone witness exists. Homogeneity and $h_y \le K\ell_y$ give (48) with $c_\alpha = \eta/K$. If the normalized set is empty, every feasible coefficient is zero and any positive constant suffices.

The feasible sublevel with loss at most $\log K$ is nonempty and weakly compact by this bound and the same closure argument. The softmax loss has bounded gradient, so its expectation is convex and norm-continuous on L^2, hence weakly lower semicontinuous. It attains the global minimum on this sublevel. Conditional Jensen and L^2 contraction allow replacement of $\boldsymbol{u}$ by $\mathbb{E}[\boldsymbol{u} \mid \boldsymbol{z}, \boldsymbol{t}, Y]$. These arguments do not require convexity of the feasible set and remain valid after restricting the factor to the closed PSD cone. Conversely, scaling a nontrivial cone witness by any $s \ge 0$ gives an unbounded coefficient ray with loss at most $\log K$, contradicting level-boundedness. □

F.2 A BALANCED POPULATION POLAR WITNESS

Here $\boldsymbol{t}$ collects the signal coordinates and, when fitted, the constant intercept coordinate. Thus $\boldsymbol{C}\boldsymbol{t}$ combines the terms $\boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{a}$ from Appendix C.2. The vector $\boldsymbol{p}$ denotes a polar certificate, rather than a predictor change. We obtain it as a softmax gradient at a population minimizer; the stationarity identities below supply its balance and second-moment properties.

Lemma F.2 (Nondegenerate softmax gradients)**.** *Suppose $\mathbb{P}\{Y = y\} > 0$ for every class. For any almost surely finite predictor $\boldsymbol{v} \in \mathbb{R}^q$, its softmax gradient $\boldsymbol{p} = \nabla \ell_Y(\boldsymbol{v})$ satisfies $\mathbb{E}[\boldsymbol{p}\boldsymbol{p}^\top] \succ \boldsymbol{0}$.*

Proof. For nonzero $\boldsymbol{e} \in \mathbb{R}^q$, append $e_0 = 0$ and choose a class y maximizing $e_0, \dots, e_q$. At least one of these coordinates is smaller. On the positive-probability event $\{Y = y\}$,

$$\boldsymbol{e}^\top \boldsymbol{p} = \sum_{j=0}^{q} \pi_j(\boldsymbol{v})(e_j - e_y) < 0,$$

because all fitted probabilities are positive. Thus $\mathbb{E}(\boldsymbol{e}^\top \boldsymbol{p})^2 > 0$; boundedness of the gradient ensures finiteness. □

Lemma F.3 (Direct balanced population polar witness)**.** *Use the side vector $\boldsymbol{t}$ and boundary $\alpha_\star$ of either model in Appendix G.3. If $0 < \kappa < \bar{\kappa} < \alpha_\star^{-1}$, there are $\boldsymbol{H}_\circ \in \mathbb{S}_{++}^q$, a standard Gaussian $\boldsymbol{z}_\circ$ independent of $(\boldsymbol{t}, Y)$, and $\boldsymbol{w} \in L^2$ such that*

$$\boldsymbol{p} = \boldsymbol{H}_\circ \boldsymbol{z}_\circ + \boldsymbol{w} \in \operatorname{int}(\mathcal{C}_Y^\circ) \quad a.s., \qquad \mathbb{E}[\boldsymbol{t}\boldsymbol{p}^\top] = \boldsymbol{0}, \qquad \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top] = (1 - \bar{\kappa})\boldsymbol{H}_\circ^2.$$

Proof. We obtain a population minimizer, show that its Gaussian factor is invertible, derive its stationarity identities, and extract the balanced polar witness from its gradient. Invertibility is needed to obtain a strictly feasible direction for the multiplier argument.

Let $\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{I}_q)$ be independent of $(\boldsymbol{t}, Y)$. We use an unrestricted square factor $\boldsymbol{A}$ and minimize

$$\mathbb{E}\ell_Y(\boldsymbol{A}\boldsymbol{z} + \boldsymbol{C}\boldsymbol{t} + \boldsymbol{u}) \quad \text{over } \boldsymbol{A} \in \mathbb{R}^{q\times q},\ \boldsymbol{C},\ \boldsymbol{u} \in L^2, \qquad \mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top] \preceq \bar{\kappa}\boldsymbol{A}\boldsymbol{A}^\top. \tag{49}$$

To apply Lemma F.1 at $\alpha = \bar{\kappa}^{-1}$, we exclude a nontrivial cone witness for the unrestricted factor.

Suppose a feasible triple $(\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{u})$ also satisfies $\boldsymbol{A}\boldsymbol{z} + \boldsymbol{C}\boldsymbol{t} + \boldsymbol{u} \in \mathcal{C}_Y$ almost surely. For every $\boldsymbol{\Lambda} \succeq \boldsymbol{0}$, pointwise comparison with this cone point gives

$$\mathbb{E}\operatorname{dist}_{\boldsymbol{\Lambda}}^2(\boldsymbol{A}\boldsymbol{z} + \boldsymbol{C}\boldsymbol{t}, \mathcal{C}_Y) \le \operatorname{tr}\{\boldsymbol{\Lambda}\mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top]\} \le \bar{\kappa}\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{A}\boldsymbol{A}^\top).$$

Put $\boldsymbol{H} = (\boldsymbol{A}\boldsymbol{A}^\top)^{1/2}$. Since $\boldsymbol{A}\boldsymbol{z}$ and $\boldsymbol{H}\boldsymbol{z}$ are centered Gaussian vectors with the same covariance, both independent of $(\boldsymbol{t}, Y)$, their displayed distance expectations agree. This comparison also holds when $\boldsymbol{A}$ is singular; no transformation of $\boldsymbol{u}$ is needed. If $\boldsymbol{A} \neq \boldsymbol{0}$, then $\operatorname{tr}\boldsymbol{H} > 0$. Dividing $(\boldsymbol{H}, \boldsymbol{C})$

by $\operatorname{tr} \boldsymbol{H}$ and using the homogeneity of squared distance to a cone yields a normalized boundary candidate with value at least $\bar{\kappa}^{-1}$. This contradicts $\bar{\kappa}^{-1} > \alpha_\star$.

If $\boldsymbol{A} = \boldsymbol{0}$, moment domination gives $\boldsymbol{u} = \boldsymbol{0}$ almost surely. Write $\boldsymbol{C}\boldsymbol{t} = \boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{a}$, with $\boldsymbol{a} = \boldsymbol{0}$ when the intercept is not fitted. Every conditional class probability is positive, so $\boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{a} \in \mathcal{C}_Y$ almost surely implies $\boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{a} \in \bigcap_y \mathcal{C}_y = \{\boldsymbol{0}\}$ almost surely. Consequently

$$0 = \mathbb{E}\|\boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{a}\|^2 = \|\boldsymbol{M}\|_F^2 + \|\boldsymbol{a}\|^2,$$

and all coefficient blocks vanish. Thus no nontrivial cone witness exists, and Lemma F.1 gives a minimizer of (49), measurable in $(\boldsymbol{z}, \boldsymbol{t}, Y)$.

Suppose a minimizing $\boldsymbol{A}$ is singular, and choose a unit $\boldsymbol{f} \in \ker \boldsymbol{A}$. Replacing $\boldsymbol{u}$ by $\mathbb{E}[\boldsymbol{u} \mid \boldsymbol{A}\boldsymbol{z}, \boldsymbol{t}, Y]$ preserves feasibility and cannot increase the loss, so it gives another minimizer with the same $\boldsymbol{A}$. Both this correction and $\boldsymbol{p} = \nabla \ell_Y(\boldsymbol{A}\boldsymbol{z} + \boldsymbol{C}\boldsymbol{t} + \boldsymbol{u})$ are independent of $Z_f = \boldsymbol{f}^\top \boldsymbol{z}$, since $\boldsymbol{A}\boldsymbol{f} = \boldsymbol{0}$. Moment domination confines $\boldsymbol{u}$ to $\operatorname{range}(\boldsymbol{A})$ almost surely. The scalar L^2-span $\mathcal{H}_{\boldsymbol{u}}$ of its coordinates therefore has dimension at most $\operatorname{rank}(\boldsymbol{A}) < q$. By Lemma F.2, the coordinates of $\boldsymbol{p}$ are linearly independent, so some deterministic unit $\boldsymbol{e}$ has

$$R = P_{\mathcal{H}_{\boldsymbol{u}}^\perp}(\boldsymbol{e}^\top \boldsymbol{p}) \neq 0, \qquad B = -\sqrt{\bar{\kappa}}\, R / \|R\|_{L^2}.$$

Then $\mathbb{E}[\boldsymbol{u}B] = \boldsymbol{0}$, $\mathbb{E}B^2 = \bar{\kappa}$, and B is independent of Z_f. For $\varepsilon > 0$, set $\boldsymbol{A}_\varepsilon = \boldsymbol{A} + \varepsilon \boldsymbol{e}\boldsymbol{f}^\top$ and $\boldsymbol{u}_\varepsilon = \boldsymbol{u} + \varepsilon \boldsymbol{e} B$. The identities

$$\boldsymbol{A}_\varepsilon \boldsymbol{A}_\varepsilon^\top = \boldsymbol{A}\boldsymbol{A}^\top + \varepsilon^2 \boldsymbol{e}\boldsymbol{e}^\top, \qquad \mathbb{E}[\boldsymbol{u}_\varepsilon \boldsymbol{u}_\varepsilon^\top] = \mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top] + \bar{\kappa}\varepsilon^2 \boldsymbol{e}\boldsymbol{e}^\top$$

show feasibility. The loss derivative at zero is

$$\mathbb{E}[(\boldsymbol{e}^\top \boldsymbol{p})(Z_f + B)] = -\sqrt{\bar{\kappa}}\, \|R\|_{L^2} < 0.$$

The bounded softmax Hessian justifies this derivative and its $O(\varepsilon^2)$ remainder. This contradicts minimality, proving that $\boldsymbol{A}$ is invertible.

Fix the minimizing correction $\boldsymbol{u}$ and its gradient $\boldsymbol{p} = \nabla \ell_Y(\boldsymbol{A}\boldsymbol{z} + \boldsymbol{C}\boldsymbol{t} + \boldsymbol{u})$. As in the AMP population reduction, keep $\boldsymbol{C}$ fixed and vary $\boldsymbol{A}$ freely while restricting the correction to $\boldsymbol{u} + a\boldsymbol{p} + \boldsymbol{B}\boldsymbol{u}$, with $a \in \mathbb{R}$ and $\boldsymbol{B} \in \mathbb{R}^{q \times q}$. This is a finite-dimensional problem with the same minimizing point. For a direction $(\Delta\boldsymbol{A}, a, \boldsymbol{B})$, set $\boldsymbol{h} = a\boldsymbol{p} + \boldsymbol{B}\boldsymbol{u}$ and $\Delta\boldsymbol{v} = \Delta\boldsymbol{A}\boldsymbol{z} + \boldsymbol{h}$. The objective F and moment constraint $\boldsymbol{\Psi}$ have derivatives

$$DF[\Delta\boldsymbol{A}, \boldsymbol{h}] = \mathbb{E}[\boldsymbol{p}^\top \Delta\boldsymbol{v}],$$
$$D\boldsymbol{\Psi}[\Delta\boldsymbol{A}, \boldsymbol{h}] = \mathbb{E}[\boldsymbol{h}\boldsymbol{u}^\top + \boldsymbol{u}\boldsymbol{h}^\top] - \bar{\kappa}(\Delta\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{A}\Delta\boldsymbol{A}^\top).$$

The bounded softmax Hessian bounds the objective remainder by $(H/2)\|\Delta\boldsymbol{v}\|_{L^2}^2$, and the constraint remainder has operator norm at most $\|\boldsymbol{h}\|_{L^2}^2 + \bar{\kappa}\|\Delta\boldsymbol{A}\|_F^2$. These bounds are quadratic in the finite parameters. Since $\boldsymbol{A}$ is invertible, the coefficient direction $\Delta\boldsymbol{A} = \boldsymbol{A}$ satisfies

$$D\boldsymbol{\Psi}[\boldsymbol{A}, \boldsymbol{0}] = -2\bar{\kappa}\boldsymbol{A}\boldsymbol{A}^\top \prec \boldsymbol{0}.$$

Thus Lemma H.1 supplies $\boldsymbol{V} \succeq \boldsymbol{0}$ and complementarity. Stationarity in a and $\boldsymbol{B}$ gives

$$\mathbb{E}[(\boldsymbol{p} + \boldsymbol{V}\boldsymbol{u})^\top \boldsymbol{p}] = 0, \qquad \mathbb{E}[(\boldsymbol{p} + \boldsymbol{V}\boldsymbol{u})^\top \boldsymbol{B}\boldsymbol{u}] = 0 \quad (\boldsymbol{B} \in \mathbb{R}^{q \times q}).$$

Take $\boldsymbol{B} = \boldsymbol{V}$ and add these equalities to obtain $\mathbb{E}\|\boldsymbol{p} + \boldsymbol{V}\boldsymbol{u}\|^2 = 0$. Stationarity in $\boldsymbol{A}$ gives the Gaussian cross-moment identity below. Finally, varying any entry of $\boldsymbol{C}$ leaves the moment constraint unchanged. The derivative along each such line is zero, giving exact side balance:

$$\boldsymbol{p} + \boldsymbol{V}\boldsymbol{u} = \boldsymbol{0}, \qquad \mathbb{E}[\boldsymbol{p}\boldsymbol{t}^\top] = \boldsymbol{0}, \qquad \mathbb{E}[\boldsymbol{p}\boldsymbol{z}^\top] = \bar{\kappa}\boldsymbol{V}\boldsymbol{A}, \qquad (50)$$
$$\operatorname{tr}\{\boldsymbol{V}(\bar{\kappa}\boldsymbol{A}\boldsymbol{A}^\top - \mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top])\} = 0.$$

All derivative pairings are integrable because $\boldsymbol{p}$ is bounded and $\boldsymbol{z}$, $\boldsymbol{t}$, and $\boldsymbol{u}$ are square-integrable.

Gradient nondegeneracy and $\boldsymbol{p} = -\boldsymbol{V}\boldsymbol{u}$ force $\boldsymbol{V} \succ \boldsymbol{0}$. Complementarity then gives $\mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top] = \bar{\kappa}\boldsymbol{A}\boldsymbol{A}^\top$. Therefore, with the invertible matrix $\boldsymbol{F} = \boldsymbol{V}\boldsymbol{A}$,

$$\mathbb{E}[\boldsymbol{p}\boldsymbol{p}^\top] = \bar{\kappa}\boldsymbol{F}\boldsymbol{F}^\top, \qquad \mathbb{E}[\boldsymbol{p}\boldsymbol{z}^\top] = \bar{\kappa}\boldsymbol{F}, \qquad \mathbb{E}[\boldsymbol{t}\boldsymbol{p}^\top] = \boldsymbol{0}.$$

Every fitted softmax probability is positive, so $\boldsymbol{p}$ is a strictly positive combination of the generators of $\mathcal{C}_Y^\circ$ and is already in its interior. Set $\boldsymbol{w} = \boldsymbol{p} - \boldsymbol{F}\boldsymbol{z}$. The preceding identities yield

$$\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top] = (1 - \bar{\kappa})\boldsymbol{F}\boldsymbol{F}^\top.$$

Finally write $\boldsymbol{F} = \boldsymbol{H}_\circ \boldsymbol{O}$ with $\boldsymbol{H}_\circ = (\boldsymbol{F}\boldsymbol{F}^\top)^{1/2} \succ \boldsymbol{0}$ and $\boldsymbol{O}$ orthogonal, and put $\boldsymbol{z}_\circ = \boldsymbol{O}\boldsymbol{z}$. This preserves the Gaussian law and side independence, while $\boldsymbol{p} = \boldsymbol{H}_\circ \boldsymbol{z}_\circ + \boldsymbol{w}$. Since $\boldsymbol{p}$ is bounded, $\boldsymbol{w} \in L^2$, completing the proof. □

F.3 PROJECTED POLAR-CERTIFICATE CONSTRUCTION

We now pass from the population polar witness to a sample certificate. The theorem below uses the normalized form $\boldsymbol{H}(\boldsymbol{z}+\boldsymbol{w})$. To apply it to the witness $\boldsymbol{p}=\boldsymbol{H}_\circ \boldsymbol{z}_\circ+\boldsymbol{w}$ just constructed, set $\rho=1-\kappa$, $\delta=\bar{\kappa}-\kappa>0$, and normalize the correction:

$$\boldsymbol{w}_{\mathrm{lat}}=\boldsymbol{H}_\circ^{-1}\boldsymbol{w}, \qquad \boldsymbol{p}=\boldsymbol{H}_\circ(\boldsymbol{z}_\circ+\boldsymbol{w}_{\mathrm{lat}}), \qquad \mathbb{E}[\boldsymbol{w}_{\mathrm{lat}}\boldsymbol{w}_{\mathrm{lat}}^\top]=(\rho-\delta)\boldsymbol{I}_q.$$

Use $\boldsymbol{H}=\boldsymbol{H}_\circ$ in the theorem and write $\boldsymbol{w}$ for the normalized correction $\boldsymbol{w}_{\mathrm{lat}}$. Thus $\boldsymbol{H}$ maps latent vectors into certificate vectors. Below, $\boldsymbol{Q}$ is the sample certificate and $\boldsymbol{P}_0$ projects away the side coordinates; they correspond to $\boldsymbol{P}$ and $\boldsymbol{P}_T$ in Appendix C.2.

We treat Gaussian side coordinates with or without a constant column in one argument. Fix an integer $a\ge 0$ and $\varepsilon\in\{0,1\}$, and set

$$\boldsymbol{t}=\begin{cases}\boldsymbol{z}_0, & \varepsilon=0,\\ (\boldsymbol{z}_0^\top,1)^\top, & \varepsilon=1,\end{cases} \qquad \boldsymbol{z}_0\sim N(\boldsymbol{0},\boldsymbol{I}_a), \qquad p=a+\varepsilon.$$

For $n\ge\max\{1,p\}$, sample iid copies and let $\boldsymbol{T}=(\boldsymbol{t}_1^\top;\dots;\boldsymbol{t}_n^\top)$ and $\boldsymbol{P}_0=\boldsymbol{I}_n-\boldsymbol{T}(\boldsymbol{T}^\top\boldsymbol{T})^{-1}\boldsymbol{T}^\top$. When $p=0$, all side matrices are empty and $\boldsymbol{P}_0=\boldsymbol{I}_n$. When $p>0$ and $n\ge p$, $\boldsymbol{T}$ has full column rank almost surely: a $p\times p$ minor is a nonzero polynomial in the Gaussian entries. For $\varepsilon=1$, this is seen by taking Gaussian row parts $\boldsymbol{0},\boldsymbol{e}_1,\dots,\boldsymbol{e}_a$; when $a=0$, the minor equals 1. In particular, the constant-only case has $\boldsymbol{P}_0=\boldsymbol{I}_n-\boldsymbol{1}_n\boldsymbol{1}_n^\top/n$.

Lemma F.4 (Projected sparse-row right inverses)**.** *With the preceding side matrix, fix $\rho_0\in(0,1)$ and positive integers $m_0=m_0(n)$ with $m_0/n\to\rho_0$. Independently of $\boldsymbol{T}$, let $\boldsymbol{\Gamma}_0\in\mathbb{R}^{n\times m_0}$ have iid standard Gaussian entries, and put $\boldsymbol{X}_0=\boldsymbol{P}_0\boldsymbol{\Gamma}_0$. There are deterministic $\theta,c,C>0$ such that, with probability tending to one, $\|\boldsymbol{X}_0\|_{\mathrm{op}}\le C\sqrt{n}$ and*

$$\inf_{\substack{B\subseteq[n]\\ 1\le|B|\le\theta n}} \sigma_{\min}((\boldsymbol{X}_0)_B)\ge c\sqrt{n}. \tag{51}$$

Proof. For $p>0$, put $\boldsymbol{U}=\boldsymbol{T}(\boldsymbol{T}^\top\boldsymbol{T})^{-1/2}$. The law of large numbers gives $\lambda_{\min}(\boldsymbol{T}^\top\boldsymbol{T})\ge cn$ and $\sum_i\|\boldsymbol{t}_i\|^4\le Cn$ with probability tending to one. On this event, Cauchy's inequality gives, uniformly over B,

$$\|\boldsymbol{U}_B\|_{\mathrm{op}}^2\le\frac{C}{n}\sum_{i\in B}\|\boldsymbol{t}_i\|^2\le C\sqrt{|B|/n}.$$

For $p=0$, take $\boldsymbol{U}$ empty and the left side zero. Apply Lemma D.7 with scalar output $r=1$ and the baseline spaces $U_i=\{0\}$. Since $m_0/n\to\rho_0>0$, the strict Hall condition holds with a fixed gap, say $\rho_0/4$, for all sufficiently large n. The lemma gives constants $\theta_0,c_0>0$ such that, simultaneously for $|B|\le\theta_0 n$,

$$\sigma_{\min}((\boldsymbol{\Gamma}_0)_B)\ge c_0\sqrt{n}$$

with probability tending to one. Indeed, for scalar output the upgraded residual space consists exactly of vectors supported on B, and its adjoint is $(\boldsymbol{\Gamma}_0)_B^\top$. The Gaussian operator-norm estimate used in Theorem D.6 also gives $\|\boldsymbol{\Gamma}_0\|_{\mathrm{op}}\le C_0\sqrt{n}$ on an event of probability tending to one.

For every such B,

$$(\boldsymbol{X}_0)_B=(\boldsymbol{\Gamma}_0)_B-\boldsymbol{U}_B\boldsymbol{U}^\top\boldsymbol{\Gamma}_0, \qquad \|\boldsymbol{U}_B\boldsymbol{U}^\top\boldsymbol{\Gamma}_0\|_{\mathrm{op}}\le C_1(|B|/n)^{1/4}\sqrt{n},$$

because $\|\boldsymbol{U}^\top\|_{\mathrm{op}}\le 1$ and the previous bound on $\|\boldsymbol{U}_B\|_{\mathrm{op}}^2$ applies uniformly in B. Choose a fixed $0<\theta\le\theta_0$ with $C_1\theta^{1/4}\le c_0/2$. The reverse triangle inequality, applied to the transposed row operators on each unit vector, now yields (51) with $c=c_0/2$. Finally, $\|\boldsymbol{X}_0\|_{\mathrm{op}}\le\|\boldsymbol{\Gamma}_0\|_{\mathrm{op}}\le C_0\sqrt{n}$.

□

Theorem F.5. *Use the side vector $\boldsymbol{t}$ above, with $a\ge 0$ and $\varepsilon\in\{0,1\}$. Let Y take values in a fixed finite set $\mathcal{Y}$, with any conditional distribution given $\boldsymbol{z}_0$, and let U be independent uniform on $(0,1)$. Fix $\rho\in(0,1)$ and full-dimensional pointed simplicial polyhedral cones $D_y\subset\mathbb{R}^q$, $y\in\mathcal{Y}$. Suppose, for standard Gaussian $\boldsymbol{z}\in\mathbb{R}^q$ independent of $(\boldsymbol{z}_0,Y,U)$, there are $\boldsymbol{H}\in\mathbb{S}_{++}^q$, $\delta>0$, and $\boldsymbol{w}=\boldsymbol{w}(\boldsymbol{z},\boldsymbol{z}_0,Y,U)\in L^2$ such that*

$$\boldsymbol{H}(\boldsymbol{z}+\boldsymbol{w})\in\operatorname{int}D_Y \quad a.s., \qquad \mathbb{E}[\boldsymbol{t}\{\boldsymbol{H}(\boldsymbol{z}+\boldsymbol{w})\}^\top]=\boldsymbol{0}, \qquad \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top]\preceq(\rho-\delta)\boldsymbol{I}_q.$$

For positive integers s_n with $s_n/n \to \rho$ and $n \geq \max\{1, p\}$, sample iid $(\boldsymbol{z}_{0,i}, Y_i, U_i)$ and form $\boldsymbol{T}$ and $\boldsymbol{P}_0$ as above. Independently, let $\boldsymbol{\Gamma} \in \mathbb{R}^{n \times s_n}$ have iid standard Gaussian entries. Then, with probability tending to one, there is $\boldsymbol{A} \in \mathbb{R}^{s_n \times q}$ such that

$$\boldsymbol{Q} = \boldsymbol{P}_0 \boldsymbol{\Gamma} \boldsymbol{A}, \qquad \boldsymbol{q}_i \in \operatorname{int} D_{Y_i} \quad (i \in [n]), \tag{52}$$

where $\boldsymbol{q}_i^\top$ is row i of $\boldsymbol{Q}$. In particular, $\boldsymbol{T}^\top \boldsymbol{Q} = \boldsymbol{0}$; when $\varepsilon = 1$, this includes $\boldsymbol{1}_n^\top \boldsymbol{Q} = \boldsymbol{0}$.

Proof. Balanced approximate interior certificates. For each $y \in \mathcal{Y}$, label the q facets of D_y by $\ell \in [q]$ and choose normalized inward facet normals $\boldsymbol{d}_{y,\ell}$. For a vector of facet offsets $\boldsymbol{b} = (b_\ell)_{\ell \in [q]} \in \mathbb{R}^q$, write

$$D_y(\boldsymbol{b}) = \{\boldsymbol{v} \in \mathbb{R}^q : \boldsymbol{d}_{y,\ell}^\top \boldsymbol{v} \geq b_\ell \text{ for every } \ell \in [q]\}.$$

Apply Lemma E.1 after the invertible change of coordinates by $\boldsymbol{H}$. Because the starting certificate is already interior, the active set is empty. We obtain positive conditional margins $\operatorname{gap}_\ell(\boldsymbol{z}_0, Y, U)$ and, after choosing the L^2 change sufficiently small, an output $\boldsymbol{w}_\circ$ satisfying $\mathbb{E}[\boldsymbol{w}_\circ \boldsymbol{w}_\circ^\top] \preceq (\rho - \delta/2)\boldsymbol{I}_q$. To verify that the change preserves balance, write $\boldsymbol{e} = \boldsymbol{z} + \boldsymbol{w}$ for the original latent vector and $\xi = (\boldsymbol{z}_0, Y, U)$. Then

$$\mathbb{E}\big[\boldsymbol{t}\{\boldsymbol{H}\mathbb{E}(\boldsymbol{e} \mid \xi)\}^\top\big] = \mathbb{E}[\boldsymbol{t}(\boldsymbol{H}\boldsymbol{e})^\top] = \boldsymbol{0},$$

so the convex combination $\boldsymbol{e}_\circ = \boldsymbol{z} + \boldsymbol{w}_\circ$ used in Lemma E.1 satisfies $\mathbb{E}[\boldsymbol{t}(\boldsymbol{H}\boldsymbol{e}_\circ)^\top] = \boldsymbol{0}$. Since $\boldsymbol{H}$ is invertible, it also satisfies $\mathbb{E}[\boldsymbol{t}\boldsymbol{e}_\circ^\top] = \boldsymbol{0}$, which is the latent-coordinate balance required by Lemma D.2. Relabel $\boldsymbol{w}_\circ$ as $\boldsymbol{w}$ from now on. We sample iid copies of U on an independent product extension. Since the event in (52) depends only on $(\boldsymbol{T}, (Y_i)_{i \in [n]}, \boldsymbol{\Gamma})$, averaging out these auxiliary copies preserves its probability. The strict covariance bound after this fixed mixing is below ρ. We therefore use the prescribed matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{n \times s_n}$ throughout and apply the AMP approximation at $s_n/n \to \rho$.

Apply Lemma D.2 with the optional balance feature $\boldsymbol{b}_{\mathrm{bal}}(\xi) = \boldsymbol{t}$, empty when $p = 0$. After multiplying its latent vectors by $\boldsymbol{H}$, it gives, for each fixed sufficiently small $h > 0$, inward facet-buffer vectors $\boldsymbol{b}_i = (b_{i\ell})_{\ell \in [q]}$, and $\boldsymbol{W}^{(0)} \in \mathbb{R}^{s_n \times q}$ such that, with $\widetilde{\boldsymbol{Q}} = \boldsymbol{\Gamma}\boldsymbol{W}^{(0)}\boldsymbol{H}$,

$$\frac{1}{n}\sum_{i=1}^n \operatorname{dist}^2(\widetilde{\boldsymbol{q}}_i, D_{Y_i}(\boldsymbol{b}_i)) \leq Ch^4, \tag{53}$$

$$\left\| \frac{1}{n} \boldsymbol{T}^\top \widetilde{\boldsymbol{Q}} \right\|_F \leq Ch^2. \tag{54}$$

Every positive buffer is at least $h/2$, and the fraction of rows having a zero buffer converges to $p(h)$, where $p(h) \to 0$ as $h \downarrow 0$. Fixed invertible $\boldsymbol{H}$ only changes the constants in these statements.

Projection. Projecting $\widetilde{\boldsymbol{Q}}$ by $\boldsymbol{P}_0$ makes its balance with the side coordinates exact, but may move rows outside their target cones. We first bound this change; the subsequent repair restores cone membership within the projected design. Write

$$\boldsymbol{U} = \boldsymbol{T}(\boldsymbol{T}^\top \boldsymbol{T})^{-1/2}, \qquad \boldsymbol{P}_0 = \boldsymbol{I}_n - \boldsymbol{U}\boldsymbol{U}^\top, \qquad \boldsymbol{B}_n = \boldsymbol{T}^\top \widetilde{\boldsymbol{Q}}/n, \qquad \boldsymbol{C}_{T,n} = \boldsymbol{T}^\top \boldsymbol{T}/n.$$

For $p = 0$, $\boldsymbol{U}$, $\boldsymbol{B}_n$, and $\boldsymbol{C}_{T,n}$ have the corresponding zero-dimensional shapes, while $\boldsymbol{P}_0 = \boldsymbol{I}_n$; the balance terms and projection cost are zero. The exact projection identity

$$\frac{1}{n}\|(\boldsymbol{I}_n - \boldsymbol{P}_0)\widetilde{\boldsymbol{Q}}\|_F^2 = \operatorname{tr}(\boldsymbol{B}_n^\top \boldsymbol{C}_{T,n}^{-1} \boldsymbol{B}_n) \tag{55}$$

together with (53)–(54) shows that $\boldsymbol{Q}^{\mathrm{base}} = \boldsymbol{P}_0 \widetilde{\boldsymbol{Q}}$ has average squared distance at most Ch^4 from the buffered cones, with probability tending to one. Indeed, when $p \geq 1$, the law of large numbers gives $\boldsymbol{C}_{T,n} \to \boldsymbol{I}_p$, so $\|\boldsymbol{C}_{T,n}^{-1}\|_{\mathrm{op}} \leq 2$ with probability tending to one; for $p = 0$ the projection cost is zero.

Exact interior repair. Put $\boldsymbol{X} = \boldsymbol{P}_0 \boldsymbol{\Gamma}$ and use the uniform inverse and operator bounds of Lemma F.4, valid for row sets selected from this same design. Let L be the zero-buffer rows, so $|L|/n \to p(h)$, and fix $\boldsymbol{v}_y^\circ \in \operatorname{int} D_y$. In Lemma D.4, take $\boldsymbol{W}_0 = \boldsymbol{W}^{(0)}\boldsymbol{H}$, $A_i = \mathbb{R}^q$, $N_i = \{0\}$, $C_i = D_{Y_i}$, and

$F_i = D_{Y_i}(\boldsymbol{b}_i)$. The row map is restriction to L, whose inverse is supplied by (51). Projection's error bound gives an initialized $\boldsymbol{A}_1$ with strict weak rows and

$$\sum_i \operatorname{dist}^2((\boldsymbol{X}\boldsymbol{A}_1)_{i\cdot}^\top, D_{Y_i}(\boldsymbol{b}_i)) \le Cnh^4. \tag{56}$$

Apply Lemma D.5, locking L and using $S = [n] \setminus L$, with inner facet bounds $b_{i\ell}/2 + h/8$. The normals have norm one, and every required inverse is a restriction to $L \cup B$. Since $\eta^2 \le Ch^4$, choose fixed h after the inverse constants so that $p(h) < \theta/4$ and $C_B C h^2 < \theta/4$. Then $|L|/n < \theta/2$ with probability tending to one and (32) holds. The repaired $\boldsymbol{A}$ leaves weak rows strict and gives all other rows margins at least $3h/8$. Every score remains in $\operatorname{col}(\boldsymbol{P}_0\boldsymbol{\Gamma})$, preserving exact side balance. This proves (52). □

G Extensions and completion of the main proofs

The separator and exclusion-certificate constructions have now been proved. Two steps remain: transfer the projected polar certificate to the actual sample kernel, and allow the signal Gram matrix to converge rather than remain fixed. We prove the Gram comparison first. Appendix G.3 then establishes the conditional Gaussian comparison and assembles both main theorems. The ratio-one corollary follows as a consequence.

G.1 Comparison with the limiting signal covariance

Proposition G.1 couples the original sample between smaller and larger exact-$\boldsymbol{S}_\star$ reference samples of sizes $n - o(n)$ and $n + o(n)$, with inclusion errors tending to zero. Feature rescaling preserves MLE existence. Full column rank makes existence monotone under adding observations: a separator of a larger sample restricts to a nonzero separator of its full-rank subset. Consequently,

$$\mathbb{P}(\text{MLE in smaller reference}) - o(1) \le \mathbb{P}(\text{MLE in original}) \le \mathbb{P}(\text{MLE in larger reference}) + o(1).$$

Both reference ratios tend to α, so the fixed-Gram theorem squeezes the original probability to the same limit on either strict side of the boundary, without a rate assumption on $\boldsymbol{S} \to \boldsymbol{S}_\star$.

G.2 Proof of the sample comparison

Nearby reference laws with exactly the limiting signal Gram matrix permit finite-mixture couplings that change sample sizes by $o(n)$.

Fix $K \ge 2$, put $q = K-1$, and fix $\boldsymbol{b}_0 \in \mathbb{R}^q$. Choose either a freely fitted intercept or an intercept fixed to zero, and use that fitting convention throughout. For $\boldsymbol{T} \in \mathbb{S}_+^q$ and $d \ge q$, let $F_{m,d}(\boldsymbol{T}, \boldsymbol{b}_0)$ be the probability that a finite MLE exists from m independent observations with standard Gaussian features in $\mathbb{R}^d$ and true coefficient matrix $[\boldsymbol{T}^{1/2}; \boldsymbol{0}_{(d-q)\times q}]$, with true intercept $\boldsymbol{b}_0$. This probability is the same for every true coefficient matrix with Gram matrix $\boldsymbol{T}$, by the equal-Gram fact in Appendix A.4 and Gaussian rotational invariance, including singular $\boldsymbol{T}$.

Proposition G.1. *Suppose $\boldsymbol{S}, \boldsymbol{S}_\star \in \mathbb{S}_+^q$, $\boldsymbol{S} \to \boldsymbol{S}_\star$, and $n, d \to \infty$ with $n/d \to \alpha > 1$. There are deterministic positive integers $m_d \le n \le N_d$ with $m_d/n \to 1$ and $N_d/n \to 1$, and a constant $C < \infty$, such that, for all sufficiently large d,*

$$F_{m_d,d}(\boldsymbol{S}_\star, \boldsymbol{b}_0) - Cn^{-1/3} \le F_{n,d}(\boldsymbol{S}, \boldsymbol{b}_0) \le F_{N_d,d}(\boldsymbol{S}_\star, \boldsymbol{b}_0) + Cn^{-1/3}. \tag{57}$$

No rate of convergence of $\boldsymbol{S}$ is required.

Proof. Set

$$\boldsymbol{A}_d = \boldsymbol{S}^{1/2}, \qquad \boldsymbol{A} = \boldsymbol{S}_\star^{1/2}, \qquad \varepsilon_d = \|\boldsymbol{A}_d - \boldsymbol{A}\|_{\mathrm{op}} + d^{-1}.$$

Write $\boldsymbol{x} = (\boldsymbol{z}, \boldsymbol{g})$ for the first q and remaining feature coordinates. Let P_d denote the law of one observation in the model defining $F_{n,d}(\boldsymbol{S}, \boldsymbol{b}_0)$. Thus $\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{I}_q)$, $\boldsymbol{g}$ is an independent standard Gaussian vector, and the conditional class probabilities are $\pi_y(\boldsymbol{b}_0 + \boldsymbol{A}_d\boldsymbol{z})$, where

$$\pi_y(\boldsymbol{v}) = \frac{e^{v_y}}{\sum_{j=0}^q e^{v_j}}, \qquad v_0 = 0.$$

Continuity of the PSD square root gives $\varepsilon_d \to 0$; hence $0 < \varepsilon_d < 1/2$ eventually. Put $v_d = 1-\varepsilon_d$ and choose $M = \|\boldsymbol{A}\|_{\mathrm{op}}+1$; eventually both $\|\boldsymbol{A}_d\|_{\mathrm{op}}$ and $\|\boldsymbol{A}\|_{\mathrm{op}}$ are at most M. For each $\boldsymbol{T} \in \{\boldsymbol{A}, \boldsymbol{A}_d\}$, define $Q_d(\boldsymbol{T})$ by

$$\boldsymbol{z} \sim N(\boldsymbol{0}, v_d \boldsymbol{I}_q), \qquad \mathbb{P}_{Q_d(\boldsymbol{T})}(Y = y \mid \boldsymbol{z}) = \pi_y\left(\boldsymbol{b}_0 + \frac{\boldsymbol{T}\boldsymbol{z}}{\sqrt{v_d}}\right),$$

with the same independent standard Gaussian $\boldsymbol{g}$. Rescaling $\boldsymbol{z}$ by $1/\sqrt{v_d}$ transforms this law into the standard-feature model with signal factor $\boldsymbol{T}$. The transformation is invertible and preserves the likelihood and the chosen intercept restriction. Hence its finite-MLE probability is $F_{m,d}(\boldsymbol{S}_\star, \boldsymbol{b}_0)$ for $\boldsymbol{T} = \boldsymbol{A}$, and $F_{m,d}(\boldsymbol{S}, \boldsymbol{b}_0)$ for $\boldsymbol{T} = \boldsymbol{A}_d$.

Write $P_d(\boldsymbol{T})$ for the corresponding standard-feature observation law, so the original law is $P_d(\boldsymbol{A}_d)$. For both ordered pairs $(\boldsymbol{T}, \boldsymbol{T}') = (\boldsymbol{A}, \boldsymbol{A}_d)$ and $(\boldsymbol{A}_d, \boldsymbol{A})$,

$$\|\boldsymbol{T}/\sqrt{1-\varepsilon_d} - \boldsymbol{T}'\|_{\mathrm{op}} \le (1+2M)\varepsilon_d.$$

Here $|(1-\varepsilon)^{-1/2} - 1| \le 2\varepsilon$ for $0 \le \varepsilon \le 1/2$, by rationalizing the square root. Every $\log \pi_y$ is globally 2-Lipschitz, so the conditional log-class probabilities differ by at most $C_A \varepsilon_d r$, where $r = \|\boldsymbol{z}\|$ and $C_A = 2 + 4M$. The Gaussian log-density ratio is

$$-\frac{q}{2}\log(1-\varepsilon_d) - \frac{\varepsilon_d}{2(1-\varepsilon_d)} r^2 \le q\varepsilon_d - \varepsilon_d r^2/3;$$

the remaining feature coordinates cancel. Completing the square gives

$$\log \frac{dQ_d(\boldsymbol{T})}{dP_d(\boldsymbol{T}')} \le \varepsilon_d\left(q - \frac{r^2}{3} + C_A r\right) \le \left(q + \frac{3}{4}C_A^2\right)\varepsilon_d.$$

Thus, with $c_d = \exp\{(q + 3C_A^2/4)\varepsilon_d\} \to 1$,

$$Q_d(\boldsymbol{A}) \le c_d P_d(\boldsymbol{A}_d), \qquad Q_d(\boldsymbol{A}_d) \le c_d P_d(\boldsymbol{A}). \tag{58}$$

The same smaller-variance comparison therefore supplies both probability inequalities.

We use a finite mixture to convert domination into sample inclusion. For probability laws $R \le cP$, with $c > 1$, put $p = 1/c$ and $S = (P - pR)/(1-p)$. This is a probability law and $P = pR + (1-p)S$. For k observations, independently draw flags $B_i \sim \mathrm{Bernoulli}(p)$, observations $Z_i \sim R$ and $V_i \sim S$, and set $W_i = Z_i$ if $B_i = 1$, and $W_i = V_i$ otherwise. The W_i are iid with law P. Conditional on any flag pattern with at least m successes, its first m success coordinates are iid with law R and form a subset of the P sample. The case $c = 1$ uses $R = P$ directly.

Set

$$m_d = \left\lfloor n/c_d - n^{2/3} \right\rfloor, \qquad N_d = \left\lceil c_d n + n^{2/3} \right\rceil.$$

Apply the finite mixture once with $(R, P) = (Q_d(\boldsymbol{A}), P_d(\boldsymbol{A}_d))$ and $k = n$, and once with $(R, P) = (Q_d(\boldsymbol{A}_d), P_d(\boldsymbol{A}))$ and $k = N_d$. Their success counts have variance at most $n/4$ and $N_d/4$, respectively. Their means exceed the required counts by at least $n^{2/3}$ and $n^{2/3}/c_d$, so Chebyshev's inequality bounds either failure probability by $O(n^{-1/3})$. For each successful flag pattern the selected sample has exactly the target iid law; averaging over the finitely many patterns therefore gives the same additive failure bound for every event increasing under the corresponding sample inclusion.

The three sample sizes have ratios to n tending to one and hence ratios to d tending to $\alpha > 1$. Their Gaussian designs, with a constant column when an intercept is fitted, have full column rank almost surely for all sufficiently large d. Finite-MLE existence is increasing under inclusion of such samples: a nonzero separator of a larger sample would separate its full-rank subset. Applying this implication on the two coupling events gives (57), with their failure probabilities as the additive errors. □

In Appendix G.3, we first prove the transitions under the exact-Gram condition. Only then do we apply Proposition G.1: both reference sample-to-dimension ratios tend to α, so their probability limits give the same limits for arbitrary-rate $\boldsymbol{S} \to \boldsymbol{S}_\star \succeq \boldsymbol{0}$ away from the relevant boundary. The exact-Gram argument uses (13) and includes singular and zero limits; Corollary I.3 evaluates the latter when $\boldsymbol{b}_0 = \boldsymbol{0}$.

Remark G.2 (Convergent true intercepts)**.** We keep the true intercept fixed in Assumption 2.1 for simplicity. More generally, with q fixed, both transitions also hold for true intercepts $\boldsymbol{b}_{0,d} \in \mathbb{R}^q$ satisfying $\|\boldsymbol{b}_{0,d} - \boldsymbol{b}_\star\| \to 0$, with respective boundaries $\alpha_\star(\boldsymbol{S}_\star, \boldsymbol{b}_\star)$ and $\alpha^0_\star(\boldsymbol{S}_\star, \boldsymbol{b}_\star)$ and no convergence-rate requirement. Indeed, in the proof of Proposition G.1, use $\boldsymbol{b}_{0,d}$ for the current signal factor $\boldsymbol{A}_d$ and $\boldsymbol{b}_\star$ for the limiting factor $\boldsymbol{A}$, in both corresponding standard-feature and smaller-variance laws. The 2-Lipschitz bound for $\log \pi_y$ adds at most $2\|\boldsymbol{b}_{0,d} - \boldsymbol{b}_\star\|$ to each upper bound on a log-density ratio. Replacing c_d by $c_d \exp\{2\|\boldsymbol{b}_{0,d} - \boldsymbol{b}_\star\|\} \to 1$ therefore gives the same sample comparison and transition.

G.3 Proofs of the main phase-transition theorems

We use one proof for the two fitting conventions, with any fixed true intercept $\boldsymbol{b}_0 \in \mathbb{R}^q$ in both cases. Write $\varepsilon = 0$ when the fitted intercept is fixed to zero (Theorem 2.5), and $\varepsilon = 1$ when it is estimated (Theorem 2.3). In both cases, $Y \mid \boldsymbol{z}_0$ follows the softmax law at $\boldsymbol{b}_0 + \boldsymbol{S}^{1/2}\boldsymbol{z}_0$. Thus ε specifies the fitted model only. Initially impose the exact Gram condition $\boldsymbol{S} = \boldsymbol{S}_\star$ and whiten the features as in Appendix A.4. In either model retain q Gaussian signal coordinates, even when $\boldsymbol{S}$ is singular, and set

$$(\boldsymbol{t}, \boldsymbol{C}) = \begin{cases} (\boldsymbol{z}_0, \boldsymbol{M}), & \varepsilon = 0, \\ ((\boldsymbol{z}_0^\top, 1)^\top, [\boldsymbol{M}, \boldsymbol{a}]), & \varepsilon = 1. \end{cases}$$

Within this proof, $\alpha_\star$ abbreviates $\alpha^0_\star(\boldsymbol{S}, \boldsymbol{b}_0)$ for $\varepsilon = 0$ and $\alpha_\star(\boldsymbol{S}, \boldsymbol{b}_0)$ for $\varepsilon = 1$. Every conditional class probability is positive and continuous in $\boldsymbol{z}_0$ for either choice. The constructions below use these properties and residual Gaussian independence, not equality of the true and fitted intercepts.

Use the signal and residual coordinates from Appendix A.4, denoting the whitened features again by $\boldsymbol{x}_i$ and writing $\boldsymbol{g}_i = \boldsymbol{z}_{\perp,i}$. Thus $\boldsymbol{x}_i = \boldsymbol{U}_0\boldsymbol{z}_{0,i} + \boldsymbol{U}_\perp \boldsymbol{g}_i$, and the matrix $\boldsymbol{G}$ with rows $\boldsymbol{g}_i^\top$ is iid standard Gaussian and independent of $\boldsymbol{Z}_0$ and the labels. For $\boldsymbol{D} \in \mathbb{R}^{q\times r}$ and $\boldsymbol{W} \in \mathbb{R}^{(d-q)\times r}$, the coefficient construction

$$\boldsymbol{B}(\boldsymbol{M}, \boldsymbol{D}, \boldsymbol{W}) = \boldsymbol{U}_0\boldsymbol{M}^\top + \boldsymbol{U}_\perp \boldsymbol{W}\boldsymbol{D}^\top, \tag{59}$$

$$\boldsymbol{B}(\boldsymbol{M}, \boldsymbol{D}, \boldsymbol{W})^\top \boldsymbol{x}_i = \boldsymbol{M}\boldsymbol{z}_{0,i} + \boldsymbol{D}\boldsymbol{W}^\top \boldsymbol{g}_i \tag{60}$$

turns residual-coordinate scores into a common fitted slope; when $\varepsilon = 1$, its fitted intercept adds $\boldsymbol{a}$ to the score.

Lemma G.3. *Fix integers $q, n \geq 1$, $0 \leq p < n$, and $0 \leq k \leq n - p$, and put $a = n - p$ and $r = a - k$. Let $\boldsymbol{T} \in \mathbb{R}^{n\times p}$ have full column rank almost surely, and let $\boldsymbol{y} = (y_1, \ldots, y_n)$ take values in a finite set $\mathcal{Y}^n$. Independently of $(\boldsymbol{T}, \boldsymbol{y})$, let $\boldsymbol{G} \in \mathbb{R}^{n\times k}$ and $\boldsymbol{\Gamma} \in \mathbb{R}^{n\times r}$ be independent standard Gaussian matrices. Set*

$$\boldsymbol{P}_0 = \boldsymbol{I}_n - \boldsymbol{T}(\boldsymbol{T}^\top\boldsymbol{T})^{-1}\boldsymbol{T}^\top, \qquad \mathcal{N} = \ker([\boldsymbol{T}, \boldsymbol{G}]^\top), \qquad \mathcal{S} = \operatorname{col}(\boldsymbol{P}_0\boldsymbol{\Gamma}).$$

For $p = 0$, take $\boldsymbol{P}_0 = \boldsymbol{I}_n$; matrices with zero columns have zero column space. Almost surely, $\boldsymbol{P}_0\boldsymbol{G}$ has rank k, $\boldsymbol{P}_0\boldsymbol{\Gamma}$ has rank r, and $[\boldsymbol{T}, \boldsymbol{G}]$ has rank $p + k$. The orthogonal projectors onto $\mathcal{N}$ and $\mathcal{S}$ have the same conditional law given $(\boldsymbol{T}, \boldsymbol{y})$.

For fixed open sets $O_y \subseteq \mathbb{R}^q$, $y \in \mathcal{Y}$, the event that a subspace admits a matrix whose columns lie in it and whose ith row lies in O_{y_i} is measurable as a function of its orthogonal projector and $\boldsymbol{y}$. Consequently, the two subspaces above have the same conditional probability of admitting such a matrix.

Proof. Conditional subspace laws. Fix full-rank $\boldsymbol{T}$ and the labels, and choose an orthonormal frame $\boldsymbol{B}$ for $\operatorname{col}(\boldsymbol{T})^\perp$. The matrices $\boldsymbol{B}^\top\boldsymbol{G}$ and $\boldsymbol{B}^\top\boldsymbol{\Gamma}$ are standard Gaussian of sizes $a\times k$ and $a\times r$ and have full column rank almost surely: each successive column avoids the proper span of its predecessors. This also gives the asserted rank of $[\boldsymbol{T}, \boldsymbol{G}]$. Their transpose-kernel and column-space projectors both have rank r and orthogonally invariant laws. These laws coincide, since Haar averaging makes $\mathbb{E}h(OPO^\top)$ independent of the rank-r projector P for every bounded Borel h; integrating against any invariant law gives the same expectation. Conjugation by $\boldsymbol{B}$ therefore equates the two projector laws for each fixed side value.

On the full-rank event the projectors are

$$\boldsymbol{P}_\mathcal{N} = \boldsymbol{P}_0 - \boldsymbol{P}_0\boldsymbol{G}(\boldsymbol{G}^\top\boldsymbol{P}_0\boldsymbol{G})^{-1}\boldsymbol{G}^\top\boldsymbol{P}_0,$$
$$\boldsymbol{P}_\mathcal{S} = \boldsymbol{P}_0\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^\top\boldsymbol{P}_0\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^\top\boldsymbol{P}_0.$$

Zero-column terms are zero. These formulas have measurable extensions off the full-rank event, since inversion is continuous on its domain. Fubini thus integrates the fixed-side equality to equality of joint, and hence conditional, laws without a measurable choice of $\boldsymbol{B}$. The argument includes zero-dimensional subspaces.

The certificate event. For fixed labels, the set

$$\{\boldsymbol{P} : \exists \boldsymbol{A} \in \mathbb{R}^{n\times q},\ (\boldsymbol{P}\boldsymbol{A})_{i\cdot}^{\top} \in O_{y_i}\ (i \in [n])\}$$

is open: it is a union over $\boldsymbol{A}$ of finite intersections of open sets. There are finitely many label vectors, giving joint measurability. For an orthogonal projector this is exactly the desired event, since $\boldsymbol{P}\boldsymbol{Q} = \boldsymbol{Q}$ precisely when every column of $\boldsymbol{Q}$ is in its range. Apply equality of the conditional projector laws to this event. □

Proof of Theorems 2.3 and 2.5. Boundary and finite-sample reduction. Use the model indexed by $\varepsilon \in \{0, 1\}$ above. Appendix B.3, Theorem B.5, and Proposition B.6 identify the boundary with the supremum of population-feasible ratios and give $1 < \alpha_\star < \infty$. The fitted design $\boldsymbol{X}_\varepsilon$ is $\boldsymbol{X}$ or $[\boldsymbol{X}, \mathbf{1}_n]$. It has full column rank eventually almost surely: a maximal minor is a nonzero Gaussian polynomial (with an intercept, use feature rows $0, \boldsymbol{e}_1, \ldots, \boldsymbol{e}_d$). The finite-sample criteria and Lemma A.2 reduce the two conclusions to a nonzero separator or an interior polar certificate in $\ker(\boldsymbol{X}_\varepsilon^{\top})$. For $\varepsilon = 1$ the latter satisfies

$$\boldsymbol{X}^{\top}\boldsymbol{Q} = 0, \qquad \mathbf{1}_n^{\top}\boldsymbol{Q} = 0, \qquad \boldsymbol{q}_i \in \operatorname{int}(\mathcal{C}_{Y_i}^{\circ}) \quad (i \in [n]). \tag{61}$$

Lemma A.1 gives uniqueness when the MLE exists.

Below the boundary. If $1 < \alpha < \alpha_\star$, choose $\alpha' \in (\alpha, \alpha_\star)$. With $\kappa = \alpha^{-1}$ and $\delta = \kappa - (\alpha')^{-1} > 0$, population feasibility supplies $\boldsymbol{H} \neq 0$ and

$$\boldsymbol{H}\boldsymbol{z}_1 + \boldsymbol{C}\boldsymbol{t} + \boldsymbol{u} \in \mathcal{C}_Y \text{ a.s.}, \qquad \mathbb{E}[\boldsymbol{u}\boldsymbol{u}^{\top}] \preceq (\kappa - \delta)\boldsymbol{H}^2. \tag{62}$$

Use the shared-tools representation $r = q$, $\boldsymbol{D} = \boldsymbol{H}$, $\boldsymbol{z} = \boldsymbol{z}_1$, $\boldsymbol{w} = \boldsymbol{H}^{\dagger}\boldsymbol{u}$. Its support and congruence argument gives $\boldsymbol{u} = \boldsymbol{D}\boldsymbol{w}$ and $\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^{\top}] \preceq (\kappa - \delta)\boldsymbol{I}_q$, including singular $\boldsymbol{H}$. Conditional expectation makes $\boldsymbol{w}$ measurable in $(\boldsymbol{z}, \boldsymbol{z}_0, Y)$ without changing these properties. Lemma E.1 supplies the conditional margins; Theorem E.3, with $m_n = d - q$ and $\boldsymbol{\mu}(\xi) = \boldsymbol{C}\boldsymbol{t}$, supplies one coefficient matrix satisfying all sample comparisons, at least one strictly, with probability tending to one. Equations (59)–(60) convert it into a nonzero fitted direction. The finite-sample criterion therefore gives MLE nonexistence.

Above the boundary. If $\alpha > \alpha_\star$, choose $\kappa = \alpha^{-1} < \bar{\kappa} < \alpha_\star^{-1}$ and put $\rho = 1 - \kappa$, $\delta = \bar{\kappa} - \kappa$. Lemma F.3 gives $\boldsymbol{H} \succ 0$ and $\boldsymbol{p} = \boldsymbol{H}\boldsymbol{z} + \boldsymbol{w} \in \operatorname{int}(\mathcal{C}_Y^{\circ})$ almost surely, with $\mathbb{E}[\boldsymbol{t}\boldsymbol{p}^{\top}] = 0$ and $\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^{\top}] = (\rho - \delta)\boldsymbol{H}^2$. Normalize the correction by $\boldsymbol{H}^{-1}$ and apply Theorem F.5 with $a = q$, $s_n = n - d - \varepsilon$, and $s_n/n \to \rho$. It produces a certificate in $\operatorname{col}(\boldsymbol{P}_0\boldsymbol{\Gamma})$ with probability tending to one, where $\boldsymbol{P}_0$ projects away $\boldsymbol{T} = \boldsymbol{Z}_0$ or $[\boldsymbol{Z}_0, \mathbf{1}_n]$.

The side matrix has rank $q + \varepsilon$ almost surely by the same Gaussian minor argument. Conditional on its signal coordinates and labels, apply Lemma G.3 with

$$a = n - q - \varepsilon, \qquad k = d - q, \qquad r = a - k = s_n = n - d - \varepsilon, \qquad O_y = \operatorname{int}(\mathcal{C}_y^{\circ}).$$

These dimensions are valid eventually since $\alpha > 1$. The residual matrix $\boldsymbol{G}$ is independent Gaussian, and whitening identifies $\ker([\boldsymbol{T}, \boldsymbol{G}]^{\top}) = \ker(\boldsymbol{X}_\varepsilon^{\top})$. The lemma equates the conditional certificate probabilities for this kernel and $\operatorname{col}(\boldsymbol{P}_0\boldsymbol{\Gamma})$; averaging proves the required probability limit. Exact mean balance is imposed only when $\varepsilon = 1$, independently of the true intercept $\boldsymbol{b}_0$. The finite-sample alternative now gives MLE existence.

Converging Gram matrices. The preceding arguments treat exact $\boldsymbol{S}$. Whitening preserves MLE existence, and Proposition G.1 transfers both strict-side limits to $\boldsymbol{S} \to \boldsymbol{S}_\star$, with each boundary evaluated at $\boldsymbol{S}_\star$. No conclusion is asserted at critical equality. □

Section 3.1 identifies the $K = 2$ specialization with the Candès–Sur boundary directly from population feasibility.

G.4 Nonexistence near one observation per feature

Proof of Corollary 2.6. Write $\alpha_\star$ for the boundary of the fitted model and choose $1 < c < \alpha_\star$. Append independent observations from the same d-dependent model until the sample size is $N_d = \lceil cd \rceil$; eventually $N_d \geq n$. The transition theorem gives nonexistence for this enlarged sample with probability tending to one. If the original design has full column rank, a nonzero separator of the enlarged sample restricts to a nonzero separator of the original. Consequently,

$$\mathbb{P}\{\text{original MLE exists}\} \leq \mathbb{P}\{\text{enlarged-sample MLE exists}\} \longrightarrow 0.$$

Here the design includes the constant column precisely when an intercept is fitted. If the original design does not have full column rank, it has full row rank almost surely, and interpolation already gives nonexistence. □

H Proofs of shared technical tools

H.1 Approximate message passing

H.1.1 Population and proximal preliminaries

The first lemma supplies the population multiplier using separation in a finite-dimensional space. The next two lemmas give the regularity and Gaussian integration-by-parts identities needed by AMP.

Lemma H.1. *Let $\mathcal{H}$ be a real Hilbert space, let $F : \mathcal{H} \to \mathbb{R}$ and $\Psi : \mathcal{H} \to \mathbb{S}^k$ be Fréchet differentiable at $x_\star$, and suppose that $x_\star$ locally minimizes F subject to $\Psi(x) \preceq \mathbf{0}$. Here $\mathbb{S}^k$ is the finite-dimensional space of symmetric $k \times k$ matrices, with trace inner product. If some $d \in \mathcal{H}$ satisfies $D\Psi(x_\star)[d] \prec \mathbf{0}$, then there exists $\boldsymbol{V} \succeq \mathbf{0}$ such that*

$$DF(x_\star)[h] + \frac{1}{2}\operatorname{tr}\{\boldsymbol{V} D\Psi(x_\star)[h]\} = 0 \quad (h \in \mathcal{H}), \qquad \operatorname{tr}\{\boldsymbol{V}\Psi(x_\star)\} = 0.$$

Proof. Write $\Psi_\star = \Psi(x_\star)$ and consider the open convex set

$$\mathcal{O} = \{(DF(x_\star)[h] + a,\ \Psi_\star + D\Psi(x_\star)[h] + \boldsymbol{B}) : h \in \mathcal{H},\ a > 0,\ \boldsymbol{B} \succ 0\} \subset \mathbb{R} \times \mathbb{S}^k.$$

It excludes the origin: otherwise some h would satisfy $DF(x_\star)[h] < 0$ and $\Psi_\star + D\Psi(x_\star)[h] \prec 0$. For small $t > 0$, differentiability would then give

$$\Psi(x_\star + th) = (1-t)\Psi_\star + t(\Psi_\star + D\Psi(x_\star)[h]) + o(t) \prec 0, \qquad F(x_\star + th) < F(x_\star),$$

contradicting local optimality; the matrix remainder is in operator norm. Finite-dimensional separation supplies a nonzero functional $(\beta, \boldsymbol{W})$ nonnegative on $\mathcal{O}$. Varying $a > 0$ and $\boldsymbol{B} \succ 0$ gives $\beta \geq 0$ and $\boldsymbol{W} \succeq 0$; varying h through all real multiples gives

$$\beta DF(x_\star)[h] + \operatorname{tr}\{\boldsymbol{W} D\Psi(x_\star)[h]\} = 0.$$

Setting $h = 0$ and sending $a, \boldsymbol{B}$ to zero yields $\operatorname{tr}(\boldsymbol{W}\Psi_\star) \geq 0$. Feasibility gives the reverse inequality, so this trace is zero. If $\beta = 0$, the nonzero PSD matrix $\boldsymbol{W}$ would have strictly negative pairing with $D\Psi(x_\star)[d] \prec 0$, contradicting stationarity. Thus $\beta > 0$ and $\boldsymbol{V} = 2\boldsymbol{W}/\beta$ proves the claim. □

Lemma H.2. *Fix $k \geq 1$. Let $\boldsymbol{V} \succ \mathbf{0}$, let $\xi \in \mathbb{R}^s$, and let $L(\boldsymbol{x}; \xi)$ be convex and C^3 in $\boldsymbol{x} \in \mathbb{R}^k$. Assume that L and its first three predictor derivatives are jointly measurable, with $\|\nabla^2 L\|_{\mathrm{op}} \leq H_0$ and $\|\nabla^3 L\|_{\mathrm{op}} \leq H_1$ uniformly. Write $\mu = \lambda_{\min}(\boldsymbol{V}) > 0$ and $K = \|\boldsymbol{V}\|_{\mathrm{op}}/\mu$. The maps*

$$\boldsymbol{m}(\boldsymbol{g}; \xi) = \arg\min_{\boldsymbol{x}} \{L(\boldsymbol{x}; \xi) + \tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{g})^\top \boldsymbol{V}(\boldsymbol{x} - \boldsymbol{g})\}, \qquad \boldsymbol{f}(\boldsymbol{g}; \xi) = \boldsymbol{m}(\boldsymbol{g}; \xi) - \boldsymbol{g}$$

are well defined, jointly measurable, and C^2 in $\boldsymbol{g}$. They satisfy

$$\nabla L(\boldsymbol{m}(\boldsymbol{g}; \xi); \xi) + \boldsymbol{V}\{\boldsymbol{m}(\boldsymbol{g}; \xi) - \boldsymbol{g}\} = \mathbf{0}, \tag{63}$$

$$\|\boldsymbol{m}(\boldsymbol{g}; \xi) - \boldsymbol{m}(\boldsymbol{h}; \xi)\| \leq K\|\boldsymbol{g} - \boldsymbol{h}\|, \qquad \|\boldsymbol{m}(\mathbf{0}; \xi)\| \leq \mu^{-1}\|\nabla L(\mathbf{0}; \xi)\|, \tag{64}$$

$$D\boldsymbol{f}(\boldsymbol{g}; \xi) = \{\boldsymbol{V} + \nabla^2 L(\boldsymbol{m}(\boldsymbol{g}; \xi); \xi)\}^{-1}\boldsymbol{V} - \boldsymbol{I}_k, \tag{65}$$

$$\|D\boldsymbol{f}(\boldsymbol{g}; \xi)\|_{\mathrm{op}} \leq K + 1, \qquad \|D^2 \boldsymbol{f}(\boldsymbol{g}; \xi)\|_{\mathrm{op}} \leq H_1 K^2/\mu. \tag{66}$$

If $\nabla L(\mathbf{0}; \xi)$ has all finite moments, then so does $\boldsymbol{f}(\mathbf{0}; \xi)$. For every fixed matrix $\boldsymbol{R}$ and an independent standard Gaussian $\boldsymbol{z}$, $\boldsymbol{f}(\boldsymbol{R}\boldsymbol{z}; \xi)$ has all finite moments.

Proof. The supporting affine bound for L and $\boldsymbol{V} \succeq \mu \boldsymbol{I}_k$ make the proximal objective coercive and strictly convex. Its unique minimizer satisfies (63). Subtracting this equation at two inputs and using gradient monotonicity with $\boldsymbol{d} = \boldsymbol{m}(\boldsymbol{g};\xi) - \boldsymbol{m}(\boldsymbol{h};\xi)$ gives

$$\mu\|\boldsymbol{d}\|^2 \le \boldsymbol{d}^\top \boldsymbol{V}(\boldsymbol{g}-\boldsymbol{h}) \le \|\boldsymbol{V}\|_{\mathrm{op}}\|\boldsymbol{d}\|\|\boldsymbol{g}-\boldsymbol{h}\|.$$

At input zero, the same argument gives $\mu\|\boldsymbol{m}(0;\xi)\| \le \|\nabla L(0;\xi)\|$. These are (64).

The Borel map $(\boldsymbol{x},\xi) \mapsto (\boldsymbol{x} + \boldsymbol{V}^{-1}\nabla L(\boldsymbol{x};\xi),\xi)$ is bijective by existence and uniqueness. The measurable-inverse theorem on standard Borel spaces makes its inverse $(\boldsymbol{g},\xi) \mapsto (\boldsymbol{m}(\boldsymbol{g};\xi),\xi)$ measurable. For fixed ξ, the implicit function theorem applies because $\boldsymbol{V} + \nabla^2 L \succeq \mu \boldsymbol{I}_k$; it gives C^2 regularity and (65). A second differentiation yields

$$(\boldsymbol{V} + \nabla^2 L) D^2\boldsymbol{m}[\boldsymbol{a},\boldsymbol{b}] = -\nabla^3 L[D\boldsymbol{m}\,\boldsymbol{a}, D\boldsymbol{m}\,\boldsymbol{b}].$$

The inverse norm bound μ^{-1} and $\|D\boldsymbol{m}\|_{\mathrm{op}} \le K$ give (66); the formulas also give joint measurability of the derivatives. Finally, $\|\boldsymbol{f}(\boldsymbol{g};\xi)\| \le \mu^{-1}\|\nabla L(0;\xi)\| + (K+1)\|\boldsymbol{g}\|$ proves the moment assertions. □

Lemma H.3. *Let $\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{I}_r)$ be independent of ξ, and let $\boldsymbol{R} \in \mathbb{R}^{k\times r}$ be deterministic. Suppose that $\boldsymbol{f}(\boldsymbol{g};\xi)$ is jointly measurable, is continuously differentiable in $\boldsymbol{g}$, has a uniformly bounded predictor Jacobian, and satisfies $\mathbb{E}\|\boldsymbol{f}(\boldsymbol{0};\xi)\|^2 < \infty$. Its Jacobian is assumed jointly measurable. Then all entries below are integrable and*

$$\mathbb{E}[\boldsymbol{f}(\boldsymbol{R}\boldsymbol{z};\xi)\boldsymbol{z}^\top] = \mathbb{E}[D\boldsymbol{f}(\boldsymbol{R}\boldsymbol{z};\xi)]\boldsymbol{R}. \tag{67}$$

Proof. If L_f bounds the Jacobian, then $\|\boldsymbol{f}(\boldsymbol{R}\boldsymbol{z};\xi)\| \le \|\boldsymbol{f}(\boldsymbol{0};\xi)\| + L_f\|\boldsymbol{R}\|_{\mathrm{op}}\|\boldsymbol{z}\|$. Thus the function and its Gaussian derivative are square-integrable. Apply (7) conditionally on ξ and use the chain rule $D_{\boldsymbol{z}}\boldsymbol{f}(\boldsymbol{R}\boldsymbol{z};\xi) = D\boldsymbol{f}(\boldsymbol{R}\boldsymbol{z};\xi)\boldsymbol{R}$. The growth bound and Cauchy–Schwarz justify Fubini and all pairings, including singular $\boldsymbol{R}$. □

H.1.2 Finite-depth state evolution

We give the finite-depth statement separately from contraction. The time-dependent version below accommodates the fixed perturbations used to make the finite history covariance positive definite. All derivatives are with respect to the predictor, and all displayed matrix errors use the Frobenius norm. The proof expands the Gaussian revelation argument in Çakmak et al. (2024, Proposition 1 and Appendix F). We use state evolution only for positive definite finite histories. A fixed perturbation, following the idea in Berthier et al. (2020, Section 5.4), supplies this condition in the construction below.

Proposition H.4. *Fix integers $k \ge 1$, $s \ge 0$, $T \ge 1$, and $\alpha > 0$. Let ξ take values in $\mathbb{R}^s$. For $1 \le t \le T$, let $f_t : \mathbb{R}^k \times \mathbb{R}^s \to \mathbb{R}^k$ and its first two predictor derivatives be jointly measurable. Suppose that f_t is C^2 in its predictor, that $Df_t, D^2 f_t$ are uniformly bounded, and that $\mathbb{E}\|f_t(0;\xi)\|^p < \infty$ for every finite $p \ge 1$. Let $C^{1,1} \succeq 0$. Define a centered Gaussian process $(g^1,\dots,g^{T+1})$, independent of ξ, by its block covariances*

$$\begin{aligned} C^{1,t} &= C^{t,1} = 0 \quad (t>1), \\ C^{s+1,t+1} &= \alpha\mathbb{E}[f_s(g^s;\xi)f_t(g^t;\xi)^\top], \qquad 1 \le s,t \le T, \\ Q_t &= \mathbb{E}Df_t(g^t;\xi). \end{aligned} \tag{68}$$

Here the same side mark is used at every time in the expectation. The construction is recursive: the covariance up to time $t+1$ uses only the Gaussian process up to time t. Its new lower-right block is a second-moment matrix, so these covariance matrices are consistent and positive semidefinite. Assume that the full block covariance $C^{1:T+1,1:T+1}$ is positive definite.

Let $n, m \to \infty$ with $\alpha_n = n/m \to \alpha$. Let $X \in \mathbb{R}^{n\times m}$ have independent $N(0,1/m)$ entries. Independently of X, take iid marks ξ_i and iid rows of $\Omega^1 \in \mathbb{R}^{m\times k}$ with law $N(0, C^{1,1})$; the marks and initialization are also independent. Set $F^0 = 0$ and

$$\Gamma^t = X\Omega^t - F^{t-1}, \qquad (F^t)_{i\cdot} = f_t((\Gamma^t)_{i\cdot}^\top;\xi_i)^\top, \qquad \Omega^{t+1} = X^\top F^t - \alpha_n \Omega^t Q_t^\top. \tag{69}$$

There is a coupling, preserving the law of these inputs and iterates, with Gaussian arrays $\widehat{\Gamma}^{1:T}$ and $\widehat{\Omega}^{1:T+1}$ such that: the row histories of each array are iid with the corresponding covariance blocks $C^{s,t}$; the two Gaussian arrays are independent of one another and of the marks; and

$$\max_{1\le t\le T}\frac{\|\Gamma^t-\widehat{\Gamma}^t\|_F}{\sqrt{n}}+\max_{1\le t\le T+1}\frac{\|\Omega^t-\widehat{\Omega}^t\|_F}{\sqrt{m}}\longrightarrow 0\quad \textit{in probability}. \tag{70}$$

In particular, $m^{-1}(\Omega^s)^\top\Omega^t\to C^{s,t}$ in probability for every fixed pair of times.

When $f_t=f$ and $C=\alpha\mathbb{E}[f(g;\xi)f(g;\xi)^\top]$ for $g\sim N(0,C)$ independent of ξ, the population recursion with $C^{1,1}=C$ has $C^{t,t}=C$ and $Q_t=Q:=\mathbb{E}Df(g;\xi)$ at every time. Its full history covariance need not be invertible. We use this stationary population process to choose an iteration count, and apply the proposition to the fixed perturbation in Lemma H.7. Every iteration count is fixed before the sample-size limit.

Exact conditioning on an adaptive transcript. The following identity is the point at which independence must be used. The query coefficients may depend on previous responses from the same matrix. They are not assumed independent of that matrix.

We first give the fixed-projection calculation, including the conditional-kernel statement needed in the adaptive induction.

Lemma H.5. *Let $g\sim N(0,I_d)$ and let $P\in\mathbb{R}^{d\times d}$ satisfy $P^\top=P$ and $P^2=P$. Put $Q=I_d-P$. Then Pg and Qg are independent, and a version of the conditional law of g given $Pg=u$ is the measurable kernel*

$$K(u,\cdot)=\mathrm{Law}(u+Q\widetilde{g}),\qquad \widetilde{g}\sim N(0,I_d).$$

The equality of conditional measures holds for almost every u under the law of Pg, with a single exceptional set for all Borel test sets. The same statement holds with both g and $\widetilde{g}$ having covariance $\sigma^2 I_d$, for $\sigma>0$.

Proof. The jointly Gaussian vectors Pg, Qg have zero cross-covariance $PQ=0$, so they are independent, including for singular projections. The continuous map $(u,z)\mapsto u+Qz$ defines a Borel kernel, and $g=Pg+Qg$ gives, for bounded Borel φ,

$$\mathbb{E}\varphi(Pg,g)=\int\!\!\int\varphi(u,u+Qz)\,d\gamma(z)\,d\gamma_P(u),$$

where γ,γ_P are the laws of g, Pg. Uniqueness of disintegration gives the asserted almost-everywhere identity of conditional measures. Scaling proves the last assertion. □

For nested projectors $P_0P_1=P_0$, orthogonality of P_1-P_0 and $I-P_1$ gives the corresponding residual update

$$\mathcal{L}((I-P_0)g\mid(P_1-P_0)g=u)=\mathcal{L}(u+(I-P_1)\widetilde{g}).$$

The formulas also hold for Borel families $P(e)$ indexed by an independent random element E, when E is retained in the conditioning: integrate the displayed joint-law identity against its law and apply Tonelli and disintegration. Adaptive queries require the joint-law induction below, because their projectors depend on earlier responses.

Lemma H.6. *Let X have the Gaussian law above, independently of an initial random element E in a standard Borel space. Reveal a fixed finite sequence of right queries XU and left queries $X^\top V$. Each query matrix must be a measurable function of E and the previously revealed responses. The transcript contains these variables and no additional information about X.*

At any point concatenate the right and left query matrices as $R\in\mathbb{R}^{m\times a}$ and $L\in\mathbb{R}^{n\times b}$, and their responses as $Y=XR$ and $Z=X^\top L$. Empty blocks are allowed. Write $R^\dagger, L^\dagger$ for the Moore–Penrose inverses, $P_R=RR^\dagger$, and $P_L=LL^\dagger$, with $P^\perp=I-P$. Almost surely, the regular conditional law of X given the transcript is the law of

$$M+P_L^\perp\widetilde{X}P_R^\perp,\qquad M=YR^\dagger+(L^\dagger)^\top Z^\top P_R^\perp, \tag{71}$$

where $\widetilde{X}$ has independent $N(0,1/m)$ entries independently of the transcript. This is an equality of conditional laws, not an assertion that the original matrix is independent of its transcript. It holds without any rank assumption on R or L.

Proof. We first identify the Gaussian law for fixed linear constraints. To justify using this law with adaptive queries, we then reveal one scalar response at a time and verify the joint law after every reveal. This keeps track of the same underlying matrix throughout.

For fixed compatible constraints $AR = Y$, $A^\top L = Z$, the map $A \mapsto P_L^\perp A P_R^\perp$ is the Frobenius orthogonal projector onto $\mathcal{K} = \{A : AR = 0,\ A^\top L = 0\}$. For any compatible X, the displayed mean is $M = XP_R + P_L X P_R^\perp = X - P_L^\perp X P_R^\perp$, the unique least-norm solution. Responses and M determine one another. Lemma H.5, applied after vectorization, therefore gives (71) for fixed queries, with residual covariance $m^{-1}(P_R^\perp \otimes P_L^\perp)$. We now prove that adaptive revelation preserves this kernel, including when the constraint rank changes.

Choose a fixed vectorization preserving the Frobenius inner product, and write $x = \mathrm{vec}(X) \in \mathbb{R}^{nm}$ and $\sigma = m^{-1/2}$. Reveal the entries of each response block in a fixed deterministic order, keeping the entire query matrix fixed until that block is complete. If u_j is a column of a right-query matrix and v_j a column of a left-query matrix, their individual entries are scalar linear observations with coefficient vectors

$$\mathrm{vec}(e_i u_j^\top), \qquad \mathrm{vec}(v_j e_\ell^\top),$$

respectively. Each coefficient is a measurable function of E and the preceding responses. It therefore suffices to treat a finite sequence of predictable scalar queries $y_j = a_j(H_{j-1})^\top x$, where $H_j = (E, y_1, \dots, y_j)$.

Initially set $\mu_0 = 0$ and $Q_0 = I_{nm}$. At a history with current mean μ and symmetric idempotent residual projector Q, put

$$v = Qa, \qquad s = v^\top v, \qquad w = \begin{cases} v/s, & s > 0, \\ 0, & s = 0, \end{cases} \qquad D = wv^\top.$$

After observing y, set

$$Q' = Q - D, \qquad \mu' = \mu + (y - a^\top \mu)w. \tag{72}$$

The update has no additional factor of σ. All these maps are Borel, including at $s = 0$. Direct multiplication gives

$$D^\top = D^2 = D, \qquad QD = DQ = D, \qquad (Q')^\top = (Q')^2 = Q', \qquad Q'a = 0.$$

Moreover, for every vector z,

$$Q'z = z \iff Qz = z \text{ and } a^\top z = 0. \tag{73}$$

Thus each new residual space is the old residual space intersected with the new query's kernel. The stored mean and projector can be recomputed from E and the preceding scalar responses by (72); storing them adds no information to the transcript.

For fixed history h, suppose the current sampling kernel is $K_h = \mathcal{L}(\mu(h) + \sigma Q(h)g)$, with g standard Gaussian. Under this kernel the response is

$$y = a^\top \mu + \sigma v^\top g, \qquad \nu_h := \mathcal{L}(a^\top \mu + \sigma v^\top g).$$

The matrices D and Q' satisfy $D(Q')^\top = 0$. Consequently Dg and $Q'g$ are independent by the Gaussian projection calculation above. The scalar observation and Dg determine one another on their supports:

$$Dg = w(a^\top Qg), \qquad a^\top Dg = a^\top Qg.$$

Since $Q = D + Q'$, the actual mean update is $\mu' = \mu + \sigma Dg$. If $s = 0$, then $v = D = 0$ and $\nu_h = \delta_{a^\top \mu}$; the query has a deterministic response and leaves the kernel unchanged. The same Borel update formula may be used for inconsistent responses, which have zero probability.

For any nonnegative Borel function φ, independence and this reconstruction prove the exact identity

$$\begin{aligned} &\int \varphi(h, a^\top \mu + \sigma v^\top g, \mu + \sigma Q g)\, d\gamma_{nm}(g) \\ &\qquad = \int\int \varphi(h, y, \mu'(h, y) + \sigma Q'(h) g')\, d\gamma_{nm}(g')\, d\nu_h(y). \end{aligned} \tag{74}$$

Here γ_{nm} is the standard Gaussian law, and every parameter in the formula is a measurable function of the old history. The initial joint law factors because E and x are independent. If the joint law

of (H_{j-1}, x) is $\mathcal{L}(H_{j-1})(dh)K_h(dx)$, integrate (74) against its history marginal. Tonelli's theorem gives the corresponding factorization for (H_j, x), with kernel $\mathcal{L}(\mu_j + \sigma Q_j g')$. This proves the finite induction for one fixed hidden vector x queried throughout the program. Uniqueness of disintegration gives an equality of conditional measures at each time. Intersecting the finitely many full-measure sets gives a single exceptional set for the prescribed transcript.

Along every actual response history, finite induction also gives

$$\mu_j = (I - Q_j)x, \qquad \operatorname{range}(Q_j) = \bigcap_{i \le j} \ker\big(a_i(H_{i-1})^\top\big).$$

For the mean identity, insert $\mu = (I-Q)x$ into (72); then $y - a^\top \mu = a^\top Q x$ and the correction is Dx. The residual-space identity follows from (73). Hence μ_j is the unique least-Euclidean-norm vector satisfying all revealed scalar constraints. At a completed query block, these constraints are precisely $AR = Y$ and $A^\top L = Z$. The Frobenius projection and least-norm identities proved at the start of this proof therefore identify the recursively constructed kernel with (71). This also explains why rank changes cause no measurability problem: the scalar recursion constructs the Borel kernel, and the pseudoinverses are used only to identify its displayed matrix form on compatible transcripts. □

For the AMP program, put $E = (\Omega^1, \xi_1, \dots, \xi_n)$ and reveal, in order, $Y_t = X\Omega^t$ followed by $Z_t = X^\top F^t$. The complete-block history before Y_t is

$$R_{t-1} = [\Omega^1, \dots, \Omega^{t-1}], \qquad Y_{t-1}^{\mathrm{all}} = [\Gamma^1 + F^0, \dots, \Gamma^{t-1} + F^{t-2}],$$
$$L_{t-1} = [F^1, \dots, F^{t-1}], \qquad Z_{t-1}^{\mathrm{all}} = [\Omega^2 + \alpha_n \Omega^1 Q_1^\top, \dots, \Omega^t + \alpha_n \Omega^{t-1} Q_{t-1}^\top].$$

More precisely, the sigma fields immediately before and after the right query are

$$\mathcal{F}_t^- = \sigma(E, Y_1, Z_1, \dots, Y_{t-1}, Z_{t-1}), \qquad \mathcal{F}_t^+ = \mathcal{F}_t^- \vee \sigma(Y_t).$$

The right query Ω^t is $\mathcal{F}_t^-$-measurable. After Y_t is revealed, F^t is $\mathcal{F}_t^+$-measurable and is a valid left query; the next complete history is $\mathcal{F}_{t+1}^- = \mathcal{F}_t^+ \vee \sigma(Z_t)$. Thus the preceding lemma applies to the actual recursion.

In particular, with $U_t = P_{R_{t-1}}^\perp \Omega^t$, the right-query innovation has conditional law

$$P_{L_{t-1}}^\perp G_t \Delta_{n,t}^{1/2}, \qquad \Delta_{n,t} = m^{-1} U_t^\top U_t, \tag{75}$$

for an independent standard Gaussian $n \times k$ matrix G_t. After that revelation, with $V_t = P_{L_{t-1}}^\perp F^t$, the left-query innovation has conditional law

$$P_{R_t}^\perp H_t \Sigma_{n,t}^{1/2}, \qquad \Sigma_{n,t} = m^{-1} V_t^\top V_t, \tag{76}$$

for an independent standard Gaussian $m \times k$ matrix H_t. Neither innovation has independent rows before the left projection is removed. Conditional on the appropriate history, however,

$$\mathbb{E}[\|P_{L_{t-1}} G_t \Delta_{n,t}^{1/2}\|_F^2 \mid \mathcal{F}_t^-] = \operatorname{rank}(L_{t-1}) \operatorname{tr} \Delta_{n,t},$$
$$\mathbb{E}[\|P_{R_t} H_t \Sigma_{n,t}^{1/2}\|_F^2 \mid \mathcal{F}_t^+] = \operatorname{rank}(R_t) \operatorname{tr} \Sigma_{n,t}. \tag{77}$$

The ranks are at most $(t-1)k$ and tk. Tightness of the innovation covariances therefore makes these discarded projections negligible after division by $\sqrt{n}$ and $\sqrt{m}$, respectively. Explicitly, conditional Markov inequality and truncation give

$$\mathbb{P}\{\|P_{L_{t-1}} G_t \Delta_{n,t}^{1/2}\|_F > \delta\sqrt{n}\} \le \mathbb{P}\{\operatorname{tr} \Delta_{n,t} > B\} + \frac{(t-1)kB}{n\delta^2}.$$

First choose B by tightness and then send n to infinity. The left-query estimate is identical with tk, m, and $\Sigma_{n,t}$.

Proof of Proposition H.4. The full history covariance is positive definite by hypothesis. At each time we first couple the right response to Gaussian reference rows, transfer their moments through the denoiser, and then couple the left response. The two conditional means require separate Onsager

cancellations. An expression $o_P(\sqrt{n})$ below means a matrix whose Frobenius norm divided by $\sqrt{n}$ tends to zero; the same convention is used with m.

The induction maintains joint Gaussian row couplings for the coefficient history and cavity history, and the resulting empirical cross-moment limits. To construct the coupling, first draw E and mutually independent standard Gaussian matrices $G_1, \dots, G_T$ and $H_1, \dots, H_T$, all independent of E. Use G_t to sample the right-query kernel and H_t to sample the left-query kernel. The reference cavities use only the G matrices, through regressions with deterministic limiting coefficients; the reference coefficients use only Ω^1 and the H matrices. Thus the two reference histories are independent of one another and of the marks. This independence follows from the construction of their joint law, not merely from the marginal conditional laws of the two innovations. Finally sample X from its conditional kernel given the completed transcript, using fresh randomness conditionally independent of the drivers given that transcript. Sequential disintegration shows that the resulting (E, X) and all responses have their original joint law, with the responses equal to the stated queries of X almost surely. In this construction the auxiliary drivers are not added to the transcript on which the original matrix law is conditioned.

To initialize the coupling, take $\widehat{\Omega}^1 = \Omega^1$. The Gaussian law of large numbers gives $m^{-1}(\Omega^1)^\top \Omega^1 \to C^{1,1}$. Conditional on Ω^1, $X\Omega^1$ has independent Gaussian rows with this empirical covariance. Using the same standard normal rows and continuity of the positive semidefinite square root couples it to iid $N(0, C^{1,1})$ rows with $o_P(\sqrt{n})$ error, independently of the initialization and marks on the reference side.

For $t \geq 2$, abbreviate $R = R_{t-1}$, $L = L_{t-1}$ and put

$$K = C^{1:t-1,1:t-1}, \quad c = C^{1:t-1,t}, \quad A = K^{-1}c, \quad \Delta = C^{t,t} - c^\top K^{-1} c.$$

The coefficient-history coupling implies $R^\dagger \Omega^t \to A$ and $m^{-1} U_t^\top U_t \to \Delta$ in probability. All empirical inverses used here exist on events of probability tending to one, since their limiting Gram matrices are positive definite. Outside these events the exact pseudoinverse formulas remain defined.

Let A_s denote the $k \times k$ blocks of A, let $e \in \mathbb{R}^{(t-1)k \times k}$ be the last block column of the identity, and put $\mathsf{S}A = (A_2^\top, \dots, A_{t-1}^\top, 0)^\top$. Write $H = C^{2:t,2:t}$. Since $C^{1,t} = 0$, the normal equations for A give $A_1 = 0$ and the block identity

$$H(e - \mathsf{S}A) = e\Delta. \tag{78}$$

Indeed its first $t-2$ blocks are the regression normal equations, and its last block is the Schur-complement formula for Δ.

There is also an exact sample identity $(Z_{t-1}^{\mathrm{all}})^\top U_t = e U_t^\top U_t$: all previous coefficient columns are orthogonal to U_t, and only the last Ω^t block contributes. Substituting this in the conditional mean (71) gives

$$\begin{aligned} M\Omega^t - F^{t-1} = [\Gamma^1, \dots, \Gamma^{t-1}] R^\dagger \Omega^t \\ + L\mathsf{S}(R^\dagger \Omega^t) + L(L^\top L)^{-1} e U_t^\top U_t - Le. \end{aligned}$$

The cavity-history coupling and rowwise law of large numbers give $n^{-1} L^\top L \to H/\alpha$ and $\|L\|_F/\sqrt{n} = O_P(1)$. Thus the second line is $o_P(\sqrt{n})$ by (78) and $m/n \to 1/\alpha$. The right conditional mean is therefore $\sum_{s<t} \Gamma^s A_s + o_P(\sqrt{n})$.

Equations (75)–(77), the convergence of $\Delta_{n,t}$, and square-root continuity now give

$$\Gamma^t = \sum_{s<t} \Gamma^s A_s + G_t \Delta^{1/2} + o_P(\sqrt{n})$$

on the sequential coupling. Replacing previous cavities by their reference versions constructs the next Gaussian reference row with exactly the required block covariance. The new G_t is independent of the existing reference variables and marks.

The Lipschitz property gives the corresponding $o_P(\sqrt{n})$ coupling of F^t to $f_t(\widehat{\Gamma}^t; \xi)$. The latter rows have all finite moments, since $\|f_t(g;\xi)\| \leq \|f_t(0;\xi)\| + L\|g\|$. The iid law of large numbers and Cauchy–Schwarz therefore give, jointly for the finitely many times,

$$\begin{aligned} n^{-1}(F^s)^\top F^t &\longrightarrow C^{s+1,t+1}/\alpha, \\ n^{-1}(\Gamma^s)^\top F^t &\longrightarrow C^{s,t} Q_t^\top. \end{aligned} \tag{79}$$

For the second identity apply Gaussian integration by parts to the joint Gaussian reference row, conditionally on its independent mark, and then integrate the mark. Bounded Df_t and the displayed growth bound justify every integral. The transpose is essential: the Jacobian acts on column-valued denoisers.

Now set $R = R_t$, $K = C^{1:t,1:t}$, $c = C^{1:t,t+1}$, and let $e_t \in \mathbb{R}^{tk\times k}$ be the last block column of the identity. Every column of Z^{all}_{t-1} lies in the range of R_t. Hence the second term in the transpose of the conditional mean (71), after the right query, vanishes, and

$$M^\top F^t - \alpha_n \Omega^t Q_t^\top = R\left\{(R^\top R/m)^{-1}(Y_t^{\mathrm{all}})^\top F^t/m - \alpha_n e_t Q_t^\top\right\},$$

where $Y_t^{\mathrm{all}} = [\Gamma^1 + F^0, \dots, \Gamma^t + F^{t-1}]$. The two identities in (79) give

$$(Y_t^{\mathrm{all}})^\top F^t/m \longrightarrow \alpha K e_t Q_t^\top + c.$$

Consequently the braces converge to $K^{-1}c$; this is the Onsager cancellation. No matrix factors have been commuted. The residual covariance in (76) converges to

$$C^{t+1,t+1} - c^\top K^{-1} c.$$

To see this, expand the regression residual of F^t on $F^1, \dots, F^{t-1}$ using (79). Its limiting Gram is $C^{2:t,2:t}/\alpha$; the initial block of c is zero, so the same Schur complement is obtained with K. Finite-rank projection removal and an independent Gaussian H_t therefore construct $\widehat{\Omega}^{t+1}$ by Gaussian regression on $\widehat{\Omega}^{1:t}$, with $o_P(\sqrt{m})$ error. This reference coefficient history is independent of the cavity reference history and marks. Its iid Gaussian law gives the coefficient Gram limits needed for the next right query. Thus both row couplings and their empirical moment limits are available at the next time, closing the induction for the prescribed finite horizon.

□

Lemma H.7. *Fix a finite horizon and denoisers satisfying the regularity and moment assumptions of Proposition H.4, allowing the unperturbed history covariance to be singular. For $\epsilon > 0$, replace the initial covariance by $C^{1,1} + \epsilon^2 I_k$ and use the augmented marks $(\xi, v_1, \dots, v_T)$, where the v_t are independent standard Gaussian vectors, independent of ξ. Set*

$$f_t^\epsilon(g; \xi, v_1, \dots, v_T) = f_t(g; \xi) + \epsilon v_t.$$

The perturbed finite-history covariance C^ϵ is positive definite, and $C^\epsilon \to C$ as $\epsilon \downarrow 0$. The perturbed Onsager matrices also converge to their unperturbed values.

Proof. The reference Gaussian history is independent of the entire augmented mark. Its covariance has initial block $C^{1,1} + \epsilon^2 I_k$, zero off-diagonal initial blocks, and remaining block

$$\alpha \mathbb{E}[\mathbf{f}^\epsilon (\mathbf{f}^\epsilon)^\top] + \alpha\epsilon^2 I_{Tk}, \qquad \mathbf{f}^\epsilon = \begin{pmatrix} f_1(g^{1,\epsilon}; \xi) \\ \vdots \\ f_T(g^{T,\epsilon}; \xi) \end{pmatrix}.$$

Both diagonal blocks are positive definite. This verifies every nondegeneracy condition in Proposition H.4.

For continuity, induct on the horizon. Couple a finite reference history of covariance K^ϵ and its unperturbed counterpart using $z(K^\epsilon)^{1/2}$ and $zK^{1/2}$, with one standard Gaussian row z independent of the marks. Continuity of the positive semidefinite square root gives convergence in L^2 whenever $K^\epsilon \to K$. Uniform Lipschitz continuity of the denoisers and their square-integrable values at zero transfer this convergence to their outputs. Cauchy–Schwarz then gives convergence of every second-moment block in the next covariance. The additional term $\alpha\epsilon^2 I$ tends to zero, completing the induction. The uniform bound on $D^2 f_t$ similarly transfers L^2 convergence to the expectations of Df_t, proving the Onsager assertion. Only fixed-dimensional covariance matrices are compared; no inverse of a singular limiting covariance occurs. □

H.1.3 CONTRACTION OF SUCCESSIVE ITERATES

Lemma H.8. *Fix $k \geq 1$. Let $L(\cdot;\xi)$ be a convex C^3 loss on $\mathbb{R}^k$, with jointly measurable gradient and uniformly bounded Hessian and third derivative, and let $\boldsymbol{V} \succ \mathbf{0}$. Define*

$$\boldsymbol{m}(\boldsymbol{g};\xi) = \arg\min_{\boldsymbol{x}} \{L(\boldsymbol{x};\xi) + \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{g})^\top \boldsymbol{V}(\boldsymbol{x}-\boldsymbol{g})\}, \qquad \boldsymbol{f}(\boldsymbol{g};\xi) = \boldsymbol{m}(\boldsymbol{g};\xi) - \boldsymbol{g},$$

and suppose $\mathbb{E}\|\boldsymbol{f}(\mathbf{0};\xi)\|^2 < \infty$. Let $\alpha, \lambda_0 > 0$ and $\boldsymbol{C} \succeq \mathbf{0}$ satisfy

$$\boldsymbol{C} = \alpha \mathbb{E}[\boldsymbol{f}(\boldsymbol{g};\xi)\boldsymbol{f}(\boldsymbol{g};\xi)^\top], \qquad \boldsymbol{V} = \lambda_0 \boldsymbol{I}_k - \alpha \boldsymbol{V} \mathbb{E}[D\boldsymbol{f}(\boldsymbol{g};\xi)], \qquad \boldsymbol{g} \sim N(\mathbf{0}, \boldsymbol{C}), \quad \boldsymbol{g} \perp \xi. \tag{80}$$

For $\mathbf{0} \preceq \boldsymbol{S} \preceq \boldsymbol{C}$, set

$$\Phi(\boldsymbol{S}) = \alpha \mathbb{E}[\boldsymbol{f}(\boldsymbol{g}_1;\xi)\boldsymbol{f}(\boldsymbol{g}_2;\xi)^\top], \qquad \begin{pmatrix}\boldsymbol{g}_1 \\ \boldsymbol{g}_2\end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix}\boldsymbol{C} & \boldsymbol{S} \\ \boldsymbol{S} & \boldsymbol{C}\end{pmatrix}\right),$$

where the Gaussian pair is independent of the single, shared mark ξ. Put $\boldsymbol{S}_1 = \mathbf{0}$ and $\boldsymbol{S}_{t+1} = \Phi(\boldsymbol{S}_t)$ for $t \geq 1$. Then this recursion is well defined, $\mathbf{0} \preceq \boldsymbol{S}_t \preceq \boldsymbol{C}$, and

$$\operatorname{tr}\{\boldsymbol{V}(\boldsymbol{C} - \boldsymbol{S}_t)\} \leq c^{t-1} \operatorname{tr}(\boldsymbol{V}\boldsymbol{C}), \qquad c := 1 - \frac{\lambda_0}{\lambda_{\max}(\boldsymbol{V})} \in [0, 1). \tag{81}$$

Here $c^0 = 1$, including when $c = 0$. In particular, if a centered Gaussian process $(\boldsymbol{g}^t)$ has marginal covariance $\boldsymbol{C}$ and adjacent cross-covariance $\boldsymbol{S}_t$, then

$$\mathbb{E}\|\boldsymbol{g}^{t+1} - \boldsymbol{g}^t\|^2 = 2\operatorname{tr}(\boldsymbol{C} - \boldsymbol{S}_t) \leq \frac{2\operatorname{tr}(\boldsymbol{V}\boldsymbol{C})}{\lambda_{\min}(\boldsymbol{V})} c^{t-1}. \tag{82}$$

Proof. By Lemma H.2, the residual $\boldsymbol{f}$ is C^2 with bounded first and second derivatives. Write $\boldsymbol{J} = D\boldsymbol{f}$ and $\boldsymbol{H} = \nabla^2 L(\boldsymbol{m}(\boldsymbol{g};\xi);\xi) \succeq \mathbf{0}$. Differentiating the proximal optimality equation gives $\boldsymbol{J} = (\boldsymbol{H} + \boldsymbol{V})^{-1}\boldsymbol{V} - \boldsymbol{I}_k$, hence

$$-\boldsymbol{V}\boldsymbol{J} - \boldsymbol{J}^\top \boldsymbol{V}\boldsymbol{J} = (\boldsymbol{I}_k + \boldsymbol{J})^\top \boldsymbol{H}(\boldsymbol{I}_k + \boldsymbol{J}) \succeq \mathbf{0}.$$

Taking expectations and using the ridge equation in (80) yields

$$\alpha \mathbb{E}[\boldsymbol{J}^\top \boldsymbol{V}\boldsymbol{J}] \preceq \boldsymbol{V} - \lambda_0 \boldsymbol{I}_k \preceq c\boldsymbol{V}, \qquad c = 1 - \frac{\lambda_0}{\lambda_{\max}(\boldsymbol{V})}. \tag{83}$$

The last inequality follows from $\boldsymbol{V} \preceq \lambda_{\max}(\boldsymbol{V})\boldsymbol{I}_k$. The leftmost matrix is positive semidefinite, so $\boldsymbol{V} \succeq \lambda_0 \boldsymbol{I}_k$ and $0 \leq c < 1$. No matrix factors have been commuted.

Fix $\mathbf{0} \preceq \boldsymbol{S} \preceq \boldsymbol{C}$ and realize the Gaussian pair as $\boldsymbol{g}_j = \boldsymbol{U} + \boldsymbol{Z}_j$, where $\boldsymbol{U} \sim N(\mathbf{0}, \boldsymbol{S})$ and $\boldsymbol{Z}_1, \boldsymbol{Z}_2 \sim N(\mathbf{0}, \boldsymbol{C} - \boldsymbol{S})$ are mutually independent and independent of ξ. Conditional on $(\boldsymbol{U}, \xi)$, the two residuals are independent with common mean $\boldsymbol{a} = \mathbb{E}[\boldsymbol{f}(\boldsymbol{g}_1;\xi) \mid \boldsymbol{U}, \xi]$. Thus the fixed-point identity gives

$$\Phi(\boldsymbol{S}) = \alpha \mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top] \succeq \mathbf{0}, \qquad \boldsymbol{C} - \Phi(\boldsymbol{S}) = \alpha \mathbb{E}[\operatorname{Cov}\{\boldsymbol{f}(\boldsymbol{g}_1;\xi) \mid \boldsymbol{U}, \xi\}] \succeq \mathbf{0}.$$

All terms are integrable by the Lipschitz bound on $\boldsymbol{f}$, the assumed second moment at zero, and conditional Jensen.

For fixed $(\boldsymbol{U}, \xi)$ and $\boldsymbol{v} \in \mathbb{R}^k$, apply Gaussian Poincaré (Boucheron et al., 2013, Theorem 3.20) to $h(\boldsymbol{z}) = \boldsymbol{v}^\top \boldsymbol{f}(\boldsymbol{U} + \boldsymbol{R}\boldsymbol{z};\xi)$, where $\boldsymbol{R} = (\boldsymbol{C} - \boldsymbol{S})^{1/2}$ and $\boldsymbol{z}$ is standard Gaussian. Its gradient is $\boldsymbol{R}^\top \boldsymbol{J}(\boldsymbol{U} + \boldsymbol{R}\boldsymbol{z};\xi)^\top \boldsymbol{v}$ and is bounded. Averaging the resulting variance bound over $(\boldsymbol{U}, \xi)$ and using $\boldsymbol{U} + \boldsymbol{R}\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{C})$ gives, for every quadratic form,

$$\mathbf{0} \preceq \boldsymbol{C} - \Phi(\boldsymbol{S}) \preceq \alpha \mathbb{E}[\boldsymbol{J}(\boldsymbol{g};\xi)(\boldsymbol{C} - \boldsymbol{S})\boldsymbol{J}(\boldsymbol{g};\xi)^\top]. \tag{84}$$

The square-root representation also covers singular $\boldsymbol{S}$ and $\boldsymbol{C} - \boldsymbol{S}$, without an inverse or a limiting argument.

The covariance bound shows that Φ maps $[\mathbf{0}, \boldsymbol{C}]$ into itself, so the recursion is well defined. Put $\boldsymbol{\Delta}_t = \boldsymbol{C} - \boldsymbol{S}_t \succeq \mathbf{0}$. By (84), cyclicity of trace, and (83),

$$\operatorname{tr}(\boldsymbol{V}\boldsymbol{\Delta}_{t+1}) \leq \operatorname{tr}\{\boldsymbol{\Delta}_t \alpha \mathbb{E}[\boldsymbol{J}^\top \boldsymbol{V}\boldsymbol{J}]\} \leq c \operatorname{tr}(\boldsymbol{V}\boldsymbol{\Delta}_t).$$

Since $\boldsymbol{\Delta}_1 = \boldsymbol{C}$, induction proves (81). Finally, $\operatorname{Cov}(\boldsymbol{g}^{t+1} - \boldsymbol{g}^t) = 2\boldsymbol{\Delta}_t$ and $\lambda_{\min}(\boldsymbol{V}) \operatorname{tr} \boldsymbol{\Delta}_t \leq \operatorname{tr}(\boldsymbol{V}\boldsymbol{\Delta}_t)$ give (82). □

H.1.4 PROOF OF THE CONSTRUCTION LEMMA

We first record the deterministic estimate that transfers a predictor approximation to the empirical loss and balance penalty.

Lemma H.9. *Let $n \geq 1$, let $\ell_i : \mathbb{R}^k \to \mathbb{R}$ be differentiable with H-Lipschitz gradient, $H \geq 0$, and let $\boldsymbol{p}_i, \boldsymbol{e}_i \in \mathbb{R}^k$ and $\boldsymbol{b}_i \in \mathbb{R}^p$. Set*

$$e_n^2 = \frac{1}{n}\sum_i \|\boldsymbol{e}_i\|^2, \qquad a_n^2 = \frac{1}{n}\sum_i \|\nabla \ell_i(\boldsymbol{0})\|^2, \qquad r_n^2 = \frac{1}{n}\sum_i \|\boldsymbol{p}_i\|^2, \qquad b_n^2 = \frac{1}{n}\sum_i \|\boldsymbol{b}_i\|^2,$$

where e_n, a_n, r_n, b_n are nonnegative. Write $\boldsymbol{B}(\boldsymbol{P}) = n^{-1}\sum_i \boldsymbol{b}_i \boldsymbol{p}_i^\top$ and let $\boldsymbol{E}$ have rows $\boldsymbol{e}_i^\top$. Then

$$\left|\frac{1}{n}\sum_i \{\ell_i(\boldsymbol{p}_i + \boldsymbol{e}_i) - \ell_i(\boldsymbol{p}_i)\}\right| \leq (a_n + H r_n) e_n + \frac{H}{2} e_n^2, \tag{85}$$

$$\|\boldsymbol{B}(\boldsymbol{P} + \boldsymbol{E}) - \boldsymbol{B}(\boldsymbol{P})\|_F \leq b_n e_n. \tag{86}$$

For $\tau \geq 0$, adding the balance penalty gives the upper bound

$$\begin{aligned}&\frac{1}{n}\sum_i \ell_i(\boldsymbol{p}_i + \boldsymbol{e}_i) + \frac{\tau}{2}\|\boldsymbol{B}(\boldsymbol{P} + \boldsymbol{E})\|_F^2 - \frac{1}{n}\sum_i \ell_i(\boldsymbol{p}_i) - \frac{\tau}{2}\|\boldsymbol{B}(\boldsymbol{P})\|_F^2 \\ &\qquad \leq \{a_n + H r_n + \tau \|\boldsymbol{B}(\boldsymbol{P})\|_F b_n\} e_n + \frac{H + \tau b_n^2}{2} e_n^2. \end{aligned} \tag{87}$$

Proof. The gradient bound and the integral along the segment from $\boldsymbol{p}_i$ to $\boldsymbol{p}_i + \boldsymbol{e}_i$ give

$$|\ell_i(\boldsymbol{p}_i + \boldsymbol{e}_i) - \ell_i(\boldsymbol{p}_i)| \leq \{\|\nabla \ell_i(\boldsymbol{0})\| + H\|\boldsymbol{p}_i\|\}\|\boldsymbol{e}_i\| + \frac{H}{2}\|\boldsymbol{e}_i\|^2.$$

Sum and apply Cauchy–Schwarz to obtain (85). For balance, the triangle inequality, the identity $\|\boldsymbol{b}_i \boldsymbol{e}_i^\top\|_F = \|\boldsymbol{b}_i\|\|\boldsymbol{e}_i\|$, and Cauchy–Schwarz prove (86). Finally use $\|\boldsymbol{B} + \Delta\|_F^2 - \|\boldsymbol{B}\|_F^2 \leq 2\|\boldsymbol{B}\|_F\|\Delta\|_F + \|\Delta\|_F^2$. □

Proof of Lemma D.1. Work on the product law of $(\boldsymbol{z}, \xi)$; all L^2 spaces below are real Hilbert spaces, and random variables are identified up to almost sure equality. Let H bound the loss Hessian, let c_0 bound $\ell(\boldsymbol{0}; \xi)$, and let c_1 bound $\|\nabla \ell(\boldsymbol{0}; \xi)\|$. The pointwise estimate $0 \leq \ell(\boldsymbol{v}; \xi) \leq c_0 + c_1\|\boldsymbol{v}\| + (H/2)\|\boldsymbol{v}\|^2$ first makes the loss integrable for every $\boldsymbol{v} \in L^2$. For any predictors $\boldsymbol{v}, \boldsymbol{h} \in L^2$, the pointwise Taylor bound and Cauchy–Schwarz give

$$\|\nabla \ell(\boldsymbol{v}; \xi)\|_{L^2} \leq c_1 + H\|\boldsymbol{v}\|_{L^2},$$

$$\left|\mathbb{E}\ell(\boldsymbol{v} + \boldsymbol{h}; \xi) - \mathbb{E}\ell(\boldsymbol{v}; \xi) - \mathbb{E}[\nabla \ell(\boldsymbol{v}; \xi)^\top \boldsymbol{h}]\right| \leq \frac{H}{2}\|\boldsymbol{h}\|_{L^2}^2.$$

In particular, the expected loss is finite, convex, and norm-continuous on all of L^2, and hence weakly lower semicontinuous. The balance map $\mathcal{B}(\boldsymbol{v}) = \mathbb{E}[\boldsymbol{b}\boldsymbol{v}^\top]$ is bounded linear, because

$$\|\mathcal{B}(\boldsymbol{v})\|_F \leq \|\boldsymbol{b}\|_{L^2}\|\boldsymbol{v}\|_{L^2}.$$

Only the second moment of $\boldsymbol{b}$ is needed in this step.

Let $J_0 = \mathbb{E}\ell(\boldsymbol{0}; \xi) < \infty$. Choose a feasible minimizing sequence $(\boldsymbol{A}_j, \boldsymbol{u}_j)$ with objective at most $J_0 + 1$. Nonnegativity of all three objective terms and the trace constraint give

$$\|\boldsymbol{A}_j\|_F^2 \leq \frac{2(J_0 + 1)}{\lambda}, \qquad \|\boldsymbol{u}_j\|_{L^2}^2 \leq \kappa \|\boldsymbol{A}_j\|_F^2.$$

Finite-dimensional compactness and weak sequential compactness of bounded sets in a Hilbert space give a subsequence such that $\boldsymbol{A}_j \to \boldsymbol{A}$ and $\boldsymbol{u}_j \rightharpoonup \boldsymbol{u}$ in L^2. For every deterministic $\boldsymbol{e} \in \mathbb{R}^r$,

$$\mathbb{E}(\boldsymbol{e}^\top \boldsymbol{u})^2 \leq \liminf_j \mathbb{E}(\boldsymbol{e}^\top \boldsymbol{u}_j)^2 \leq \kappa \boldsymbol{e}^\top \boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{e}.$$

Thus the limit is feasible. Also $\|(\boldsymbol{A}_j - \boldsymbol{A})\boldsymbol{z}\|_{L^2} = \|\boldsymbol{A}_j - \boldsymbol{A}\|_F$, so the predictors converge weakly to $\boldsymbol{A}\boldsymbol{z} + \boldsymbol{u}$.

The loss is weakly lower semicontinuous, and the balance matrices converge because their entries are continuous linear functionals on L^2. Together with convergence of the ridge term, this proves attainment without requiring convexity of the joint feasible set.

Fix a minimizer $(\boldsymbol{A}, \boldsymbol{u})$. If $\boldsymbol{e} \in \ker \boldsymbol{A}^\top$, feasibility gives $\mathbb{E}(\boldsymbol{e}^\top \boldsymbol{u})^2 = 0$. Applying this to a finite basis of the kernel shows that $\boldsymbol{u} \in \operatorname{range}(\boldsymbol{A})$ almost surely. We keep this correction on the original probability space. Its dependence on the reduced predictor $\boldsymbol{Rz}$ will follow from the optimality equations below.

If $\boldsymbol{A} = \boldsymbol{0}$, feasibility forces $\boldsymbol{u} = \boldsymbol{0}$ and $J_\star = J_0$. Taking $\boldsymbol{W} = \boldsymbol{0}$ proves the sample conclusion by the law of large numbers. Otherwise let $k = \operatorname{rank}(\boldsymbol{A})$, and choose $\boldsymbol{D} \in \mathbb{R}^{r \times k}$ with orthonormal columns spanning $\operatorname{range}(\boldsymbol{A})$. Set

$$\boldsymbol{R} = \boldsymbol{D}^\top \boldsymbol{A} \in \mathbb{R}^{k \times r}, \qquad \boldsymbol{C} = \boldsymbol{R}\boldsymbol{R}^\top \succ \boldsymbol{0}, \qquad \boldsymbol{u}_k = \boldsymbol{D}^\top \boldsymbol{u}.$$

Then $\boldsymbol{A} = \boldsymbol{D}\boldsymbol{R}$ and $\boldsymbol{u} = \boldsymbol{D}\boldsymbol{u}_k$ almost surely. The matrix $\boldsymbol{R}$ has full row rank, because its range is $\mathbb{R}^k$. We retain the original Gaussian vector $\boldsymbol{z} \in \mathbb{R}^r$ and the original probability space; no change of input coordinates is needed.

To check optimality after this reduction, take any reduced competitor $(\boldsymbol{S}, \boldsymbol{w})$ with $\boldsymbol{S} \in \mathbb{R}^{k \times r}$ and $\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top] \preceq \kappa \boldsymbol{S}\boldsymbol{S}^\top$, where $\boldsymbol{w}$ is any L^2 correction on the original probability space. Its embedding

$$\boldsymbol{A}' = \boldsymbol{D}\boldsymbol{S}, \qquad \boldsymbol{u}' = \boldsymbol{D}\boldsymbol{w}$$

is feasible in the original program and has predictor $\boldsymbol{D}(\boldsymbol{S}\boldsymbol{z} + \boldsymbol{w})$. Moreover, $\|\boldsymbol{A}'\|_F = \|\boldsymbol{S}\|_F$, and multiplication of a $p \times k$ balance matrix by $\boldsymbol{D}^\top$ preserves its Frobenius norm. Consequently $(\boldsymbol{R}, \boldsymbol{u}_k)$ is a global minimizer of the reduced program with loss $\ell_D(\boldsymbol{x}; \xi) = \ell(\boldsymbol{D}\boldsymbol{x}; \xi)$ and the same objective value $J_\star$. The matrix $\boldsymbol{S}$ in that program is unrestricted; in particular, all perturbations of the full-row-rank minimizer $\boldsymbol{R}$ are legitimate coefficient directions. We now suppress the subscript on $\boldsymbol{u}_k$ and write $\boldsymbol{v} = \boldsymbol{R}\boldsymbol{z} + \boldsymbol{u}$.

For coefficient matrices $\boldsymbol{S}$ and L^2 corrections $\boldsymbol{w}$ on the original probability space, write

$$\begin{aligned} F(\boldsymbol{S}, \boldsymbol{w}) &= \mathbb{E}\ell_D(\boldsymbol{S}\boldsymbol{z} + \boldsymbol{w}; \xi) + \frac{\lambda}{2}\|\boldsymbol{S}\|_F^2 + \frac{\tau}{2}\|\mathcal{B}(\boldsymbol{S}\boldsymbol{z} + \boldsymbol{w})\|_F^2, \\ \Psi(\boldsymbol{S}, \boldsymbol{w}) &= \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top] - \kappa \boldsymbol{S}\boldsymbol{S}^\top. \end{aligned}$$

Let $\boldsymbol{M} = \mathcal{B}(\boldsymbol{v})$ and define the deterministic tilted loss and its gradient at the minimizing predictor by

$$L(\boldsymbol{x}; \xi) = \ell_D(\boldsymbol{x}; \xi) + \tau \langle \boldsymbol{M}, \boldsymbol{b}(\xi)\boldsymbol{x}^\top \rangle_F, \qquad \boldsymbol{p} = \nabla \ell_D(\boldsymbol{v}; \xi) + \tau \boldsymbol{M}^\top \boldsymbol{b}(\xi).$$

Here $\boldsymbol{p} \in L^2$. The tilt preserves convexity and all predictor derivatives of order at least two, and $\|\nabla L(\boldsymbol{0}; \xi)\|$ has all finite moments. The value of L may be signed; it is not being substituted for the nonnegative objective F.

Fix the functions $\boldsymbol{p}$ and $\boldsymbol{u}$, and restrict the correction to $\boldsymbol{w} = \boldsymbol{u} + a\boldsymbol{p} + \boldsymbol{B}\boldsymbol{u}$, where $a \in \mathbb{R}$ and $\boldsymbol{B} \in \mathbb{R}^{k \times k}$. The resulting functions of $(\boldsymbol{S}, a, \boldsymbol{B})$ form a finite-dimensional optimization problem with minimizer $(\boldsymbol{R}, 0, \boldsymbol{0})$. For a direction $(\Delta\boldsymbol{R}, a, \boldsymbol{B})$, put $\boldsymbol{h} = a\boldsymbol{p} + \boldsymbol{B}\boldsymbol{u}$ and $\Delta\boldsymbol{v} = \Delta\boldsymbol{R}\boldsymbol{z} + \boldsymbol{h}$. Its derivatives are

$$\begin{aligned} DF(\boldsymbol{R}, \boldsymbol{u})[\Delta\boldsymbol{R}, \boldsymbol{h}] &= \mathbb{E}[\boldsymbol{p}^\top \Delta\boldsymbol{v}] + \lambda \langle \boldsymbol{R}, \Delta\boldsymbol{R} \rangle_F, \\ D\Psi(\boldsymbol{R}, \boldsymbol{u})[\Delta\boldsymbol{R}, \boldsymbol{h}] &= \mathbb{E}[\boldsymbol{h}\boldsymbol{u}^\top + \boldsymbol{u}\boldsymbol{h}^\top] - \kappa(\Delta\boldsymbol{R}\boldsymbol{R}^\top + \boldsymbol{R}\Delta\boldsymbol{R}^\top). \end{aligned}$$

These are Fréchet derivatives in the finite-dimensional parameters. Indeed, the absolute remainder in the first formula is at most

$$\frac{H + \tau\|\boldsymbol{b}\|_{L^2}^2}{2}\|\Delta\boldsymbol{v}\|_{L^2}^2 + \frac{\lambda}{2}\|\Delta\boldsymbol{R}\|_F^2,$$

and the matrix-operator-norm remainder in the second is at most $\|\boldsymbol{h}\|_{L^2}^2 + \kappa\|\Delta\boldsymbol{R}\|_F^2$. Also $\|\Delta\boldsymbol{v}\|_{L^2} \le \|\Delta\boldsymbol{R}\|_F + |a|\|\boldsymbol{p}\|_{L^2} + \|\boldsymbol{B}\|_{\mathrm{op}}\|\boldsymbol{u}\|_{L^2}$, so both remainders are quadratic in these parameters.

The finite-parameter direction $(\boldsymbol{R}, 0, \boldsymbol{0})$ satisfies

$$D\Psi(\boldsymbol{R}, \boldsymbol{u})[\boldsymbol{R}, \boldsymbol{0}] = -2\kappa\boldsymbol{R}\boldsymbol{R}^\top \prec \boldsymbol{0}.$$

Apply Lemma H.1 to obtain a symmetric $\boldsymbol{V} \succeq \boldsymbol{0}$. Stationarity in the correction parameters gives

$$\mathbb{E}[(\boldsymbol{p} + \boldsymbol{V}\boldsymbol{u})^\top \boldsymbol{p}] = 0, \qquad \mathbb{E}[(\boldsymbol{p} + \boldsymbol{V}\boldsymbol{u})^\top \boldsymbol{B}\boldsymbol{u}] = 0 \quad (\boldsymbol{B} \in \mathbb{R}^{k \times k}).$$

Take $\boldsymbol{B} = \boldsymbol{V}$ and add these equations. Then $\mathbb{E}\|\boldsymbol{p} + \boldsymbol{V}\boldsymbol{u}\|^2 = 0$, so $\boldsymbol{p} = -\boldsymbol{V}\boldsymbol{u}$ almost surely. Stationarity in each coefficient direction and complementarity give

$$\begin{aligned} \boldsymbol{p} := \nabla L(\boldsymbol{v};\xi) = -\boldsymbol{V}\boldsymbol{u}, \qquad & \mathbb{E}[\boldsymbol{p}\boldsymbol{z}^\top] + \lambda\boldsymbol{R} - \kappa\boldsymbol{V}\boldsymbol{R} = \boldsymbol{0}, \\ \boldsymbol{V} \succeq \boldsymbol{0}, \qquad & \operatorname{tr}\{\boldsymbol{V}(\kappa\boldsymbol{R}\boldsymbol{R}^\top - \mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top])\} = 0. \end{aligned} \tag{88}$$

All displayed expectations are finite by Cauchy–Schwarz.

If $\boldsymbol{e} \in \ker \boldsymbol{V}$, the first identity gives $\boldsymbol{e}^\top \boldsymbol{p} = 0$ almost surely, and multiplying coefficient stationarity by $\boldsymbol{e}^\top$ gives $\lambda \boldsymbol{e}^\top \boldsymbol{R} = 0$. Since $\lambda > 0$ and $\boldsymbol{R}$ has full row rank, $\boldsymbol{e} = \boldsymbol{0}$. Thus $\boldsymbol{V} \succ \boldsymbol{0}$. Let $\boldsymbol{S}_0 = \kappa\boldsymbol{R}\boldsymbol{R}^\top - \mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top] \succeq \boldsymbol{0}$. If $\mu = \lambda_{\min}(\boldsymbol{V}) > 0$, then complementarity yields $0 = \operatorname{tr}(\boldsymbol{V}\boldsymbol{S}_0) \geq \mu \operatorname{tr} \boldsymbol{S}_0 \geq 0$. Hence $\boldsymbol{S}_0 = \boldsymbol{0}$, proving

$$\mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top] = \kappa\boldsymbol{R}\boldsymbol{R}^\top. \tag{89}$$

Define the proximal residual

$$\boldsymbol{m}(\boldsymbol{g};\xi) = \arg\min_{\boldsymbol{x}}\{L(\boldsymbol{x};\xi) + \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{g})^\top \boldsymbol{V}(\boldsymbol{x}-\boldsymbol{g})\}, \qquad \boldsymbol{f}(\boldsymbol{g};\xi) = \boldsymbol{m}(\boldsymbol{g};\xi) - \boldsymbol{g}, \qquad \boldsymbol{C} = \boldsymbol{R}\boldsymbol{R}^\top.$$

By Lemma H.2, these maps are jointly measurable, $\boldsymbol{f}$ is C^2 with bounded first two predictor derivatives, and $\boldsymbol{f}(\boldsymbol{0};\xi)$ has all finite moments. The proximal first-order equation and (88) show that $\boldsymbol{u} = \boldsymbol{f}(\boldsymbol{R}\boldsymbol{z};\xi)$ almost surely. Apply Lemma H.3, whose integrability conditions follow from these bounds, to get

$$\mathbb{E}[\boldsymbol{u}\boldsymbol{z}^\top] = \mathbb{E}[D\boldsymbol{f}(\boldsymbol{R}\boldsymbol{z};\xi)]\boldsymbol{R}.$$

Substitution in coefficient stationarity gives $-\boldsymbol{V}\mathbb{E}D\boldsymbol{f}\,\boldsymbol{R} + \lambda\boldsymbol{R} - \kappa\boldsymbol{V}\boldsymbol{R} = \boldsymbol{0}$. Right multiplication by $\boldsymbol{R}^\top\boldsymbol{C}^{-1}$, together with (89), yields

$$\boldsymbol{C} = \alpha\mathbb{E}[\boldsymbol{f}\boldsymbol{f}^\top], \qquad \boldsymbol{V} = \lambda_0\boldsymbol{I}_k - \alpha\boldsymbol{V}\mathbb{E}D\boldsymbol{f}, \qquad \alpha = \kappa^{-1}, \quad \lambda_0 = \lambda/\kappa, \tag{90}$$

where the expectations use $\boldsymbol{g} \sim N(\boldsymbol{0}, \boldsymbol{C})$ independently of ξ. In particular, $\boldsymbol{Q} := \mathbb{E}D\boldsymbol{f} = \lambda\boldsymbol{V}^{-1} - \kappa\boldsymbol{I}_k$ is symmetric. The covariance equality concerns this auxiliary optimizer. A population separator used to bound its objective still need only satisfy the original second-moment inequality.

The stationary population process associated with $\boldsymbol{f}$ has marginal covariance $\boldsymbol{C}$ at every time. Its first cross-covariances are zero, and its recursion is

$$\mathbb{E}[\boldsymbol{g}^{t+1}(\boldsymbol{g}^{s+1})^\top] = \alpha\mathbb{E}[\boldsymbol{f}(\boldsymbol{g}^t;\xi)\boldsymbol{f}(\boldsymbol{g}^s;\xi)^\top].$$

The same mark occurs at both times. Its adjacent covariance is $\boldsymbol{S}_t$ from Lemma H.8, with $\boldsymbol{S}_1 = 0$. Choose $\theta \in (c, 1)$, where $c = 1 - \lambda_0/\lambda_{\max}(\boldsymbol{V})$. That lemma gives

$$\{2\operatorname{tr}(\boldsymbol{C} - \boldsymbol{S}_t)\}^{1/2} \leq C_0\theta^{t/2}, \qquad C_0 = \left\{\frac{2\operatorname{tr}(\boldsymbol{V}\boldsymbol{C})}{\lambda_{\min}(\boldsymbol{V})\theta}\right\}^{1/2}. \tag{91}$$

This is a population statement and requires no sample state evolution for a singular history.

Fix an iteration count T and a perturbation $\epsilon > 0$. Put $\boldsymbol{X} = \boldsymbol{G}/\sqrt{m}$ and $\alpha_n = n/m$. Initialize independent rows of $\boldsymbol{\Omega}^{1,\epsilon}$ with law $N(0, \boldsymbol{C} + \epsilon^2\boldsymbol{I}_k)$, independently of the design and marks. Take independent standard Gaussian matrices $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_{T+1}$, also independent of these inputs, and run

$$\begin{aligned} \boldsymbol{\Gamma}^{t,\epsilon} &= \boldsymbol{X}\boldsymbol{\Omega}^{t,\epsilon} - \boldsymbol{F}^{t-1,\epsilon}, \\ (\boldsymbol{F}^{t,\epsilon})_{i\cdot} &= \boldsymbol{f}((\boldsymbol{\Gamma}^{t,\epsilon})_{i\cdot}^\top;\xi_i)^\top + \epsilon(\boldsymbol{V}_t)_{i\cdot}, \\ \boldsymbol{\Omega}^{t+1,\epsilon} &= \boldsymbol{X}^\top\boldsymbol{F}^{t,\epsilon} - \alpha_n\boldsymbol{\Omega}^{t,\epsilon}(\boldsymbol{Q}_t^\epsilon)^\top, \qquad \boldsymbol{F}^{0,\epsilon} = 0. \end{aligned} \tag{92}$$

Here $\boldsymbol{Q}_t^\epsilon = \mathbb{E}D\boldsymbol{f}(\boldsymbol{g}^{t,\epsilon};\xi)$ uses the perturbed population recursion. The augmented row marks contain $(\xi_i, (\boldsymbol{V}_1)_{i\cdot}, \ldots, (\boldsymbol{V}_{T+1})_{i\cdot})$. The proximal bounds verify the regularity and moment assumptions of Proposition H.4; Lemma H.7 verifies positive definiteness. Apply state evolution with horizon $T+1$.

Let $\boldsymbol{P}^{T,\epsilon}$ have rows $\boldsymbol{m}((\boldsymbol{\Gamma}^{T,\epsilon})_{i\cdot}^\top;\xi_i)^\top$. Since $\boldsymbol{f} = \boldsymbol{m} - \mathrm{id}$, the first update gives the exact identity

$$\boldsymbol{E}^{T,\epsilon} := \boldsymbol{X}\boldsymbol{\Omega}^{T+1,\epsilon} - \boldsymbol{P}^{T,\epsilon} = \boldsymbol{\Gamma}^{T+1,\epsilon} - \boldsymbol{\Gamma}^{T,\epsilon} + \epsilon\boldsymbol{V}_T. \tag{93}$$

Write

$$d_T^\epsilon = \operatorname{tr}\boldsymbol{C}_\epsilon^{T,T} + \operatorname{tr}\boldsymbol{C}_\epsilon^{T+1,T+1} - 2\operatorname{tr}\boldsymbol{C}_\epsilon^{T,T+1}.$$

The joint cavity coupling, the law of large numbers, and the triangle inequality imply, for these fixed T, ϵ,

$$e_n^{T,\epsilon} := \frac{\|\boldsymbol{E}^{T,\epsilon}\|_F}{\sqrt{n}} \le \sqrt{d_T^\epsilon} + \epsilon\sqrt{k} + o_{\mathbb{P}}(1).$$

At fixed T, Lemma H.7 gives $d_T^\epsilon \to 2\,\mathrm{tr}(\boldsymbol{C} - \boldsymbol{S}_T)$ as $\epsilon \downarrow 0$.

Define the actual limiting predictor of the candidate by

$$\boldsymbol{y}_T^\epsilon = \boldsymbol{g}^{T+1,\epsilon} + \boldsymbol{f}(\boldsymbol{g}^{T,\epsilon};\xi) + \epsilon\boldsymbol{v}_T, \qquad J_T^\epsilon = \mathbb{E}\ell_D(\boldsymbol{y}_T^\epsilon;\xi) + \frac{\lambda}{2}\,\mathrm{tr}\,\boldsymbol{C}_\epsilon^{T+1,T+1} + \frac{\tau}{2}\|\mathbb{E}[\boldsymbol{b}(\xi)(\boldsymbol{y}_T^\epsilon)^\top]\|_F^2.$$

Here $\boldsymbol{v}_T$ is the Tth standard Gaussian block in the augmented mark, with the dependence prescribed by the population construction. For fixed $T, \epsilon > 0$, joint finite-depth state evolution and Lemma H.9 give, for $\boldsymbol{W}_{n,T,\epsilon} = m^{-1/2}\boldsymbol{\Omega}^{T+1,\epsilon}\boldsymbol{D}^\top$,

$$\frac{1}{n}\sum_i \ell(\boldsymbol{W}_{n,T,\epsilon}^\top \boldsymbol{g}_i;\xi_i) + \frac{\lambda}{2}\|\boldsymbol{W}_{n,T,\epsilon}\|_F^2 + \frac{\tau}{2}\left\|\frac{1}{n}\sum_i \boldsymbol{b}_i(\boldsymbol{W}_{n,T,\epsilon}^\top \boldsymbol{g}_i)^\top\right\|_F^2 = J_T^\epsilon + o_{\mathbb{P}}(1). \tag{94}$$

The predictor identity follows directly from the AMP cavity update. The loss has quadratic growth, the denoiser is uniformly Lipschitz, and its value at zero has a finite second moment. These facts give integrable reference loss and squared predictor norm. Each balance entry is integrable by Cauchy–Schwarz. Thus the ordinary law of large numbers applies to the reference rows, and their L^2 coupling transfers the displayed empirical quantities. The loss in this formula is the original loss; the tilted loss only defines the denoiser.

At fixed T, the finite Gaussian covariance history is continuous at $\epsilon = 0$ by Lemma H.7. Its symmetric square root couples every history to one fixed standard Gaussian vector. This proves $J_T^\epsilon \to J_T^0$. To justify passage of expectations, for either the loss or a balance entry $h(\boldsymbol{x};\xi)$ we have $|h(\boldsymbol{x};\xi)| \le a(\xi) + B\|\boldsymbol{x}\|^2$ with $\mathbb{E}a(\xi) < \infty$. On a bounded set of covariance roots, the predictor is bounded by $\|\boldsymbol{f}(0;\xi)\| + K\|\boldsymbol{z}\|$ for a fixed finite Gaussian source $\boldsymbol{z}$. Squaring this bound gives a single integrable dominator, including at singular limiting covariances. No continuity in ξ is required.

For the unperturbed history, $\mathrm{Cov}(\boldsymbol{g}^T) = \mathrm{Cov}(\boldsymbol{g}^{T+1}) = \boldsymbol{C}$ and their cross-covariance is $\boldsymbol{S}_T$. Equality of Gaussian laws gives

$$(\boldsymbol{g}^T, \boldsymbol{g}^{T+1}) \overset{d}{=} (\boldsymbol{S}_T^{1/2}\boldsymbol{z}_0 + (\boldsymbol{C} - \boldsymbol{S}_T)^{1/2}\boldsymbol{z}_1, \boldsymbol{S}_T^{1/2}\boldsymbol{z}_0 + (\boldsymbol{C} - \boldsymbol{S}_T)^{1/2}\boldsymbol{z}_2),$$

where $\boldsymbol{z}_0, \boldsymbol{z}_1, \boldsymbol{z}_2$ are independent standard Gaussian vectors, independent of the shared mark ξ. By Lemma H.8, $\mathrm{tr}(\boldsymbol{C} - \boldsymbol{S}_T) \to 0$. Since $\boldsymbol{C} - \boldsymbol{S}_T \succeq 0$, this also implies $\boldsymbol{S}_T \to \boldsymbol{C}$ in matrix norm. Continuity of the positive semidefinite square root and the same integrable-envelope argument now show that the loss and balance of $\boldsymbol{g}^{T+1} + \boldsymbol{f}(\boldsymbol{g}^T;\xi)$ converge to those of $\boldsymbol{g} + \boldsymbol{f}(\boldsymbol{g};\xi) = \boldsymbol{m}(\boldsymbol{g};\xi)$, where $\boldsymbol{g} \sim N(0, \boldsymbol{C})$ independently of ξ. The ridge term is already $\lambda\,\mathrm{tr}\,\boldsymbol{C}/2$ at every unperturbed time. Consequently $J_T^0 \to J_\star$.

Given accuracy $\varepsilon > 0$, first choose a finite T with $|J_T^0 - J_\star| < \varepsilon/3$, next choose a fixed $\epsilon > 0$ with $|J_T^\epsilon - J_T^0| < \varepsilon/3$, and finally let $n \to \infty$ in (94). This proves the required objective bound. All Gaussian-law transports and continuity statements allow singular zero-dither histories, while state evolution is used only for the fixed positive dither.

The candidate depends on independent initialization and dither arrays. Joint measurability of the finite recursion makes its failure event measurable in the data and all these auxiliary seeds. For the fixed T, ϵ, let $p_n \to 0$ be the joint failure probability. Fubini expresses p_n as the average of the conditional data failure probability over the seeds. Some deterministic seed therefore has failure probability at most p_n: otherwise its nonnegative excess over p_n would have positive integral. Fix such a seed for each n, including every dither block. The resulting coefficients are measurable in the design and side sample alone, as required. □

H.2 Buffered approximation and repair

H.2.1 Buffered approximation

Proof. Fix $0 < h \le 1$ and set

$$b_{h,\ell}(\xi) = \frac{h}{2}\mathbb{1}\{\mathrm{gap}_\ell(\xi) \ge h\} \quad (\ell \notin J(\xi)), \qquad b_{h,\ell}(\xi) = 0 \quad (\ell \in J(\xi)).$$

Write $\boldsymbol{b}_h = (b_{h,\ell})_{\ell\in[q]}$; these buffers satisfy (29). Choose a bounded measurable truncation $\widehat{\boldsymbol{\mu}}_h$ such that

$$\mathbb{E}\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_h\|^2 \le h^8, \tag{95}$$

and put

$$\widehat{s}_{h,\ell}(\boldsymbol{v},\xi) = \boldsymbol{a}_{Y,\ell}^\top(\widehat{\boldsymbol{\mu}}_h + \boldsymbol{D}\boldsymbol{v}), \qquad V_h(\boldsymbol{v};\xi) = \sum_{\ell\in J}\widehat{s}_{h,\ell}(\boldsymbol{v},\xi)^2 + \sum_{\ell\notin J}(b_{h,\ell} - \widehat{s}_{h,\ell}(\boldsymbol{v},\xi))_+^2.$$

Convolve V_h in $\boldsymbol{v}$ with $N(\boldsymbol{0}, \varsigma_h^2\boldsymbol{I}_r)$, where $0 < \varsigma_h^2 \le h^8$, to obtain ℓ_h. At each fixed h, this nonnegative convex loss is C^∞, with bounded Hessian and third derivative, and bounded loss and gradient at zero. Indeed, there are finitely many facet matrices, the truncation and buffers are bounded, and each summand is a squared affine residual or squared hinge. Jensen's inequality and the uniform quadratic upper bound give

$$V_h \le \ell_h \le V_h + C\varsigma_h^2. \tag{96}$$

The population witness satisfies the actual buffered constraints. Thus its only unsmoothed violations come from the truncated shift, and (95)–(96) imply $\mathbb{E}\ell_h(\boldsymbol{z}+\boldsymbol{w};\xi) \le Ch^8$.

For this fixed h, the side vector is finite dimensional, the loss and its predictor derivatives are jointly measurable, and the bounds just proved verify every loss assumption of Lemma D.1. Apply that lemma to $\boldsymbol{G}_{\mathrm{main}}$ at ratio κ_0 with $\lambda = h^4$, tolerance h^8, and $\tau = \mathbb{1}\{p \ge 1\}$. In particular, λ is a fixed positive constant throughout this application; it is not replaced by an n-dependent parameter inside state evolution. The competitor $\boldsymbol{A} = \boldsymbol{I}_r$, $\boldsymbol{u} = \boldsymbol{w}$ is feasible since $\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top] \preceq (\kappa-\delta)\boldsymbol{I}_r \prec \kappa_0\boldsymbol{I}_r$; its optional balance is zero by (25). For fixed h, the resulting measurable coefficient $\boldsymbol{W}_h$ therefore satisfies, with probability tending to one,

$$\frac{1}{n}\sum_i V_h(\boldsymbol{W}_h^\top \boldsymbol{g}_{\mathrm{main},i};\xi_i) + \frac{h^4}{2}\|\boldsymbol{W}_h\|_F^2 + \frac{\tau}{2}\left\|\frac{1}{n}\sum_i \boldsymbol{b}_{\mathrm{bal}}(\xi_i)(\boldsymbol{W}_h^\top \boldsymbol{g}_{\mathrm{main},i})^\top\right\|_F^2 \le \frac{rh^4}{2} + Ch^8. \tag{97}$$

All terms on the left are nonnegative. Hence the norm and optional balance bounds follow, and the mean squared actual facet violation is at most

$$\frac{2}{n}\sum_i V_h(\boldsymbol{W}_h^\top \boldsymbol{g}_{\mathrm{main},i};\xi_i) + \frac{C}{n}\sum_i \|\boldsymbol{\mu}(\xi_i) - \widehat{\boldsymbol{\mu}}_h(\xi_i)\|^2 \le Ch^4$$

on the additional law-of-large-numbers event that the last empirical shift error is at most $2h^8$. The actual faces are nonempty almost surely because the population witness belongs to them. Hoffman's bound (Hoffman, 1952), with a uniform constant for the finitely many matrices indexed by (Y, J), converts this violation bound into mean squared distance at most Ch^4.

Let $p(h)$ be the probability that an inactive gap lies in $(0, h)$. The inactive gaps are positive almost surely and there are finitely many facets, so $p(h) \to 0$ as $h \downarrow 0$ by continuity of probability from above. For each fixed h, the law of large numbers gives convergence of the weak-row fraction to $p(h)$. Take $\boldsymbol{W}^{(0)} = \boldsymbol{W}_h$. All constants in the displayed approximation bounds come from the fixed facet matrices, the population witness and the fixed balance feature; the truncation and smoothing errors are bounded by h^8. Hence these constants are independent of h. The sample-size thresholds and the finite AMP parameters may depend on the chosen h, which does not change with n. □

H.2.2 Repairing rows without buffers

Proof. For every i, choose the nearest point $\boldsymbol{p}_i = \Pi_{F_i}(\boldsymbol{e}_i(\boldsymbol{W}_0))$. On L, set

$$\gamma_i = \min\{1, h^2/(1 + \|\boldsymbol{v}_i^\circ - \boldsymbol{p}_i\|)\}, \qquad \boldsymbol{t}_i = (1-\gamma_i)\boldsymbol{p}_i + \gamma_i\boldsymbol{v}_i^\circ;$$

outside L, set $\boldsymbol{t}_i = \boldsymbol{p}_i$. Every $\boldsymbol{t}_i$ belongs to A_i. Since $F_i \subseteq C_i$, the rows in L satisfy all inequalities strictly, and $\|\boldsymbol{t}_i - \boldsymbol{p}_i\| \le h^2$ there. Prescribe $\boldsymbol{r}_i = \boldsymbol{P}_{U_i}(\boldsymbol{t}_i - \boldsymbol{e}_i(\boldsymbol{W}_0))$. Orthogonal projection is a contraction, so $\sum_i \|\boldsymbol{r}_i\|^2 \le 2n(\eta^2 + h^4)$. The right inverse gives a correction $\Delta\boldsymbol{W}$ of norm at most $\sqrt{2}K\sqrt{\eta^2+h^4}$ with these projected residuals. Thus $\boldsymbol{W}_1 = \boldsymbol{W}_0 + \Delta\boldsymbol{W}$ has the asserted exact constraints. Finally, the distance inequality gives

$$\sum_i \mathrm{dist}^2(\boldsymbol{e}_i(\boldsymbol{W}_1), F_i) \le 2n\eta^2 + 2\|\boldsymbol{X}\Delta\boldsymbol{W}\|_F^2 \le 2n\eta^2 + 4M^2K^2n(\eta^2+h^4),$$

which proves (30). □

H.2.3 REPAIRING BUFFERED INEQUALITIES

Proof. We minimize a distance penalty while keeping the coefficient matrix close to its initial value. The buffers ensure that only a small set of rows can violate the inner targets within this neighborhood. If a minimizer still has a violation, the uniform right inverse gives a feasible direction that strictly lowers the objective.

If $S = \varnothing$ or $\eta = 0$, take $\boldsymbol{W}_\star = \boldsymbol{W}_1$. Otherwise define

$$\phi(\boldsymbol{W}) = \left\{ \frac{1}{n} \sum_{i \in S} \operatorname{dist}^2(\boldsymbol{e}_i(\boldsymbol{W}), H_i) \right\}^{1/2}.$$

This is the distance to the closed convex product $\prod_{i \in S} H_i$, composed with a linear map and divided by $\sqrt{n}$; it is therefore continuous and convex. Minimize

$$\phi(\boldsymbol{W}) + \frac{1}{2K} \|\boldsymbol{W} - \boldsymbol{W}_1\|_F \quad \text{over } \boldsymbol{W} \in \mathcal{A}_{\text{fix}}. \tag{98}$$

The objective is coercive on a nonempty closed affine space, so a minimizer $\boldsymbol{W}_\star$ exists. Since $F_i \subseteq H_i$, comparison with $\boldsymbol{W}_1$ gives

$$\|\boldsymbol{W}_\star - \boldsymbol{W}_1\|_F \le 2K\phi(\boldsymbol{W}_1) \le 2K\eta. \tag{99}$$

For every $\boldsymbol{W} \in \mathcal{A}_{\text{fix}}$ in this coefficient ball, the distance inequality and the operator bound imply

$$\begin{aligned} \sum_{i \in S} \operatorname{dist}^2(\boldsymbol{e}_i(\boldsymbol{W}), F_i) &\le 2 \sum_{i \in S} \operatorname{dist}^2(\boldsymbol{e}_i(\boldsymbol{W}_1), F_i) + 2\|\boldsymbol{X}(\boldsymbol{W} - \boldsymbol{W}_1)\|_F^2 \\ &\le 2(1 + 4M^2K^2)n\eta^2. \end{aligned}$$

Let $B(\boldsymbol{W}) = \{i \in S : \boldsymbol{e}_i(\boldsymbol{W}) \notin H_i\}$. These rows satisfy their active equations because $\boldsymbol{W} \in \mathcal{A}_{\text{fix}}$. On a row in $B(\boldsymbol{W})$ an inactive facet is below its inner bound. Its gap from the corresponding outer bound is at least $b_{i\ell}/2 - h/8 \ge h/8$; consequently

$$\operatorname{dist}(\boldsymbol{e}_i(\boldsymbol{W}), F_i) \ge h/(8L_a), \qquad |B(\boldsymbol{W})| \le C_B n\eta^2/h^2.$$

By (32), the uniform right inverse applies to $L \cup B(\boldsymbol{W})$ throughout this ball.

Suppose now that $\phi(\boldsymbol{W}_\star) > 0$ and abbreviate $B = B(\boldsymbol{W}_\star)$. Define a matrix $\boldsymbol{R} \in \mathbb{R}^{n \times r}$ by row transposes

$$\boldsymbol{r}_i = \boldsymbol{e}_i(\boldsymbol{W}_\star) - \Pi_{H_i}(\boldsymbol{e}_i(\boldsymbol{W}_\star)) \quad (i \in S), \qquad \boldsymbol{r}_i = \boldsymbol{0} \quad (i \notin S).$$

Its nonzero rows are precisely B, and $\|\boldsymbol{R}\|_F = \sqrt{n}\,\phi(\boldsymbol{W}_\star)$. Both the current vector and its projection belong to A_i, so $\boldsymbol{r}_i \in N_i^\perp$. Apply the right inverse of $\mathcal{T}_B$ to the right-hand side $-\boldsymbol{r}_i$ on B and zero elsewhere. It gives $\boldsymbol{D} \in \mathbb{R}^{p \times r}$ satisfying

$$\boldsymbol{e}_i(\boldsymbol{D}) = -\boldsymbol{r}_i \ (i \in B), \qquad \boldsymbol{e}_i(\boldsymbol{D}) = \boldsymbol{0} \ (i \in L), \qquad \boldsymbol{P}_{N_i} \boldsymbol{e}_i(\boldsymbol{D}) = \boldsymbol{0} \ (i \notin L \cup B),$$

and $\|\boldsymbol{D}\|_F \le K\phi(\boldsymbol{W}_\star)$. The relation $\boldsymbol{r}_i \in N_i^\perp$ also preserves the active equations on B, so $\boldsymbol{D}$ is a feasible direction in $\mathcal{A}_{\text{fix}}$. For $0 \le t \le 1$, the point on a violated row moves from its current score toward its chosen projection. Comparing to that same target point bounds its new distance by $(1-t)\|\boldsymbol{r}_i\|$. Every other strong row starts in H_i, so its new distance is at most $t\|\boldsymbol{e}_i(\boldsymbol{D})\|$. Summing these bounds gives

$$\phi(\boldsymbol{W}_\star + t\boldsymbol{D})^2 \le \{(1-t)^2 + M^2K^2t^2\}\phi(\boldsymbol{W}_\star)^2.$$

Choose $0 < t \le \{4(1+M^2K^2)\}^{-1}$. Elementary expansion shows $(1-t)^2 + M^2K^2t^2 \le (1 - 3t/4)^2$, whence $\phi(\boldsymbol{W}_\star + t\boldsymbol{D}) \le (1 - 3t/4)\phi(\boldsymbol{W}_\star)$. The triangle inequality bounds the increase in the norm penalty by $t\|\boldsymbol{D}\|_F/(2K) \le t\phi(\boldsymbol{W}_\star)/2$. Thus the objective decreases by at least $t\phi(\boldsymbol{W}_\star)/4 > 0$ along a feasible finite step, contradicting minimality. Therefore $\phi(\boldsymbol{W}_\star) = 0$; (99) gives the remaining assertion. □

H.3 TENSOR-HALL ESCAPE AND ITS CONSEQUENCES

We use the vector-matroid partition criterion of Edmonds (1965): for a finite multiset $\mathcal{V}$ of vectors and an integer $N_{\mathrm{col}} \geq 1$, the multiset can be partitioned into N_{col} linearly independent sets if and only if

$$|\mathcal{J}| \leq N_{\mathrm{col}} \dim \operatorname{span}(\mathcal{J}) \quad \text{for every submultiset } \mathcal{J} \subseteq \mathcal{V}.$$

The next two lemmas support the Gaussian row-map estimate. The first splits the constraint directions into linearly independent groups with a uniform singular-value bound. The second uses this grouping to bound how many points are needed to approximate all possible residual matrices. These are intermediate linear-algebra steps; their consequence for repairing sample constraints is stated in Theorem D.6.

Lemma H.10 (Partitioning constraint directions into uniformly independent groups)**.** *Fix an integer $r \geq 1$ and a finite family $\mathcal{U}$ of subspaces of $\mathbb{R}^r$. There is $c_{\mathcal{U}} > 0$ with the following property. For every integer $n \geq 1$, let $U_1, \dots, U_n \in \mathcal{U}$ and suppose that, for an integer $N_{\mathrm{col}} \geq 1$,*

$$\sum_{i=1}^{n} \dim(U_i \cap S) \leq N_{\mathrm{col}} \dim S \qquad \text{for every } S \leq \mathbb{R}^r.$$

The displayed condition limits the number of constraint directions inside every subspace S. For each $U \in \mathcal{U}$, choose one orthonormal basis of U and repeat it for every occurrence of that type. The resulting multiset can be partitioned into N_{col} sets $\mathcal{V}_1, \dots, \mathcal{V}_{N_{\mathrm{col}}}$, possibly empty, such that

$$\sigma_{\min}([\boldsymbol{v} : \boldsymbol{v} \in \mathcal{V}_a]) \geq c_{\mathcal{U}}$$

for every nonempty $\mathcal{V}_a$, where $[\boldsymbol{v} : \boldsymbol{v} \in \mathcal{V}_a]$ denotes the matrix having the vectors of $\mathcal{V}_a$ as columns, in any fixed order. The constant is uniform over the basis choices, multiplicities, and number of color classes.

Proof. For a submultiset $\mathcal{J}$ of the chosen basis vectors, put $S = \operatorname{span}(\mathcal{J})$. For one occurrence of U_i, at most $\dim(U_i \cap S)$ of its basis vectors can belong to $\mathcal{J}$. Hence

$$|\mathcal{J}| \leq \sum_{i=1}^{n} \dim(U_i \cap S) \leq N_{\mathrm{col}} \dim S,$$

and the matroid partition criterion gives N_{col} linearly independent color classes.

For uniform conditioning, the case where every subspace is zero is vacuous. Otherwise, suppose that there is a sequence of admissible configurations for which the best conditioning over all partitions tends to zero. Multiplicities and the number of colors may vary along this sequence. The family $\mathcal{U}$ is fixed and finite, and the orthonormal bases of each $U \in \mathcal{U}$ form a compact set. We may therefore pass to a subsequence along which every type's basis converges to an orthonormal basis of the same subspace.

Label a basis vector by its type U and its position in that basis. There are only $\sum_{U \in \mathcal{U}} \dim U$ such labels. An independent color contains at most r vectors and cannot contain two copies of the same label. Hence there are finitely many possible independent color patterns in the limiting configuration. The nonempty independent patterns have a positive minimum smallest singular value, denoted by c_*. For each member of the sequence, apply the matroid partition criterion to the limiting vectors with that member's multiplicities and number of colors; the subspace Hall inequalities are unchanged. Replace each limiting vector in this partition by the corresponding vector from the current bases. Continuity of singular values, uniformly over the finitely many patterns, makes every nonempty color's smallest singular value at least $c_*/2$ for all sufficiently late members. This contradicts the assumed failure of a uniform bound, regardless of how many times a pattern is repeated. □

Lemma H.11 (Covering residual matrices under a quadratic norm bound)**.** *Under the hypotheses of Lemma H.10, with N_{col} denoting the number of color classes there, put $\mathcal{E} = \mathcal{E}(U_1, \dots, U_n)$. For $\boldsymbol{H} \in \mathbb{S}^r_{++}$ and $R > 0$, define*

$$\mathcal{E}_{\boldsymbol{H}}(R) := \left\{ \boldsymbol{X} \in \mathcal{E} : \sum_{i=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{H} \boldsymbol{x}_i \leq R^2 \right\}.$$

Let $B_F = \{\boldsymbol{X} \in \mathcal{E} : \|\boldsymbol{X}\|_F \le 1\}$, and let $N(A, \eta B_F)$ denote the least number of translates of ηB_F needed to cover a set $A \subseteq \mathcal{E}$. Thus the covering number counts how many matrices are needed so that every matrix in A lies within Frobenius distance η of one of them. There is $C_{\mathcal{U}} < \infty$, depending only on $(r, \mathcal{U})$, such that, for every $\eta > 0$,

$$\log N(\mathcal{E}_{\boldsymbol{H}}(R), \eta B_F) \le C_{\mathcal{U}} n + \frac{N_{\mathrm{col}}}{2} \log\det\left(\boldsymbol{I}_r + \frac{C_{\mathcal{U}} R^2}{\eta^2} \boldsymbol{H}^{-1}\right), \tag{100}$$

Proof. Choose one orthonormal eigenbasis of each compression $\boldsymbol{P}_U \boldsymbol{H}|_U$, shared across all occurrences of U, and denote its positive eigenvalues on row i by $\lambda_{i,b}$. A finite maximal η-separated subset is an η-net with centers in the ellipsoid. Comparing the volumes of its disjoint radius-$\eta/2$ balls gives

$$N(\mathcal{E}_{\boldsymbol{H}}(R), \eta B_F) \le 2^{\dim \mathcal{E}} \prod_{i=1}^{n} \prod_{b=1}^{\dim U_i} \left(1 + \frac{4R^2}{\eta^2 \lambda_{i,b}}\right)^{1/2}. \tag{101}$$

Indeed, if ℓ_j are the semiaxes of the original ellipsoid, the union of these balls lies in the ellipsoid with semiaxes $2\sqrt{\ell_j^2 + (\eta/2)^2}$. To see this, write a point of the union as $a+b$, use $(a_j+b_j)^2 \le 2a_j^2 + 2b_j^2$, and sum after division by $\ell_j^2 + (\eta/2)^2$. The sum is at most 4. Haar-volume scaling under an invertible linear map gives the displayed bound. Compactness supplies a finite maximal separated set.

Partition the chosen eigenvectors by Lemma H.10. For a nonempty color with $k \le r$ vectors, write $\boldsymbol{V}$ for its column matrix, $\boldsymbol{A} = \boldsymbol{V}^\top \boldsymbol{H} \boldsymbol{V} \succ 0$, $d_j = A_{jj}$ and $\boldsymbol{D} = \mathrm{diag}(d_j)$. Cauchy's inequality in the $\boldsymbol{H}$ inner product gives $\boldsymbol{A} \preceq k\boldsymbol{D}$. Inverting this inequality gives $\boldsymbol{D}^{-1} \preceq k\boldsymbol{A}^{-1}$, hence, for $s \ge 0$,

$$\prod_j (1 + s/d_j) \le \det(\boldsymbol{I}_k + ks\boldsymbol{A}^{-1}) \le \det(\boldsymbol{I}_r + rsc_{\mathcal{U}}^{-2} \boldsymbol{H}^{-1}). \tag{102}$$

For the second inequality, use $\boldsymbol{V}^\top \boldsymbol{V} \succeq c_{\mathcal{U}}^2 \boldsymbol{I}_k$ and congruence by $\boldsymbol{A}^{-1/2}$, followed by Sylvester's determinant identity. This bounds the middle determinant by $\det(\boldsymbol{I}_r + ksc_{\mathcal{U}}^{-2} \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{V}^\top)$. The Schur complement of $\boldsymbol{A}$ in the Gram matrix of $[\boldsymbol{H}^{-1/2}, \boldsymbol{H}^{1/2}\boldsymbol{V}]$ gives $\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{V}^\top \preceq \boldsymbol{H}^{-1}$. Determinant monotonicity for positive definite matrices and $k \le r$ prove the claim. Empty colors contribute one.

Multiply (102) over at most N_{col} colors and take $s = 4R^2/\eta^2$ in (101). Since $\dim \mathcal{E} \le nr$, all fixed factors can be absorbed into $C_{\mathcal{U}}$, proving (100). □

H.3.1 Tensor-Hall escape

Proof. The result is immediate if $\mathcal{E}(U_1, \dots, U_n) = \{\boldsymbol{0}\}$. Otherwise (33), applied to a line contained in some U_i, implies $m - \varepsilon n > 0$. Put $N_{\mathrm{col}} = \lceil m - \varepsilon n \rceil$. Then

$$\sum_{i=1}^{n} \dim(U_i \cap S) \le N_{\mathrm{col}} \dim S, \qquad m - N_{\mathrm{col}} \ge \varepsilon n - 1 \ge \varepsilon n/2 \tag{103}$$

for large n.

Let $\mathcal{T} = \{\boldsymbol{X} \in \mathcal{E}(U_1, \dots, U_n) : \|\boldsymbol{X}\|_F = 1\}$. Fix a small constant $t > 0$ and put $C_G = 2\sqrt{2(\log 18 + 1)(1 + \overline{\kappa})}$. The Gaussian singular-value bound of Davidson & Szarek (2001, Theorem II.13), after rescaling and transposing when $m > n$, gives

$$\mathbb{P}\{\|\boldsymbol{G}\|_{\mathrm{op}} > \sqrt{n} + \sqrt{m} + u\} \le e^{-u^2/2}, \qquad u \ge 0.$$

Taking $u = \sqrt{2(n+m)}$ and using $m/n \le \overline{\kappa}$ yields

$$\|\boldsymbol{G}\|_{\mathrm{op}} \le C_G \sqrt{n} \tag{104}$$

outside an event of probability at most $e^{-(n+m)}$. Set $\eta = t/(4C_G)$.

We group matrices by their $r \times r$ Gram matrices. Choose a finite operator-norm t^2-net $\mathcal{K}_t$ of the compact set $\{\boldsymbol{S} \in \mathbb{S}_+^r : \mathrm{tr}\, \boldsymbol{S} = 1\}$, with net points in that set. For each $\boldsymbol{S} \in \mathcal{K}_t$, let

$$\mathcal{T}_{\boldsymbol{S}} = \{\boldsymbol{X} \in \mathcal{T} : \|\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{S}\|_{\mathrm{op}} \le t^2\},$$

and take a maximal η-separated subset $\mathcal{N}_{\boldsymbol{S}}$ of this patch in Frobenius norm; ignore empty patches. These are η-nets of their patches, and the number of patches depends only on (r, t), not on n. For $\boldsymbol{X} \in \mathcal{T}_{\boldsymbol{S}}$, the inequality $\boldsymbol{X}^\top \boldsymbol{X} \preceq \boldsymbol{S} + t^2 \boldsymbol{I}_r$ gives

$$\operatorname{tr}\left((\boldsymbol{S} + t^2 \boldsymbol{I}_r)^{-1} \boldsymbol{X}^\top \boldsymbol{X}\right) \le r.$$

Thus $\mathcal{T}_{\boldsymbol{S}} \subseteq \mathcal{E}_{\boldsymbol{H}}(\sqrt{r})$ with $\boldsymbol{H} = (\boldsymbol{S} + t^2 \boldsymbol{I}_r)^{-1}$. Apply Lemma H.11 at radius $\eta/3$. Each covering ball contains at most one point of $\mathcal{N}_{\boldsymbol{S}}$, so, writing the eigenvalues of $\boldsymbol{S}$ as $\lambda_1 \ge \cdots \ge \lambda_r \ge 0$, we obtain

$$\log |\mathcal{N}_{\boldsymbol{S}}| \le C_1 n + \frac{N_{\mathrm{col}}}{2} \sum_{j=1}^{r} \log\left(1 + \frac{C_1(\lambda_j + t^2)}{t^2}\right). \tag{105}$$

The constant C_1 depends only on the fixed parameters and is independent of t.

For a deterministic $\boldsymbol{X} \in \mathcal{T}_{\boldsymbol{S}}$, put $\boldsymbol{A} = \boldsymbol{X}^\top \boldsymbol{X}$. Each row of $\boldsymbol{G}^\top \boldsymbol{X}$ is an independent centered Gaussian vector with covariance $\boldsymbol{A}$. Diagonalizing $\boldsymbol{A}$ and integrating the one-dimensional Gaussian densities gives

$$\mathbb{E} \exp\{-\|\boldsymbol{G}^\top \boldsymbol{X}\|_F^2/(2t^2)\} = \det(\boldsymbol{I}_r + \boldsymbol{A}/t^2)^{-m/2}.$$

This identity includes zero eigenvalues. Markov's inequality yields

$$\mathbb{P}\{\|\boldsymbol{G}^\top \boldsymbol{X}\|_F \le 2t\sqrt{n}\} \le e^{2n} \det(\boldsymbol{I}_r + \boldsymbol{A}/t^2)^{-m/2}. \tag{106}$$

The Gram-patch condition implies $\boldsymbol{S} \preceq \boldsymbol{A} + t^2 \boldsymbol{I}_r$, hence

$$\boldsymbol{I}_r + \boldsymbol{S}/t^2 \preceq 2(\boldsymbol{I}_r + \boldsymbol{A}/t^2), \qquad \det(\boldsymbol{I}_r + \boldsymbol{A}/t^2) \ge 2^{-r} \det(\boldsymbol{I}_r + \boldsymbol{S}/t^2).$$

Meanwhile (105) implies

$$\log |\mathcal{N}_{\boldsymbol{S}}| \le C_2 n + \frac{N_{\mathrm{col}}}{2} \log\det(\boldsymbol{I}_r + \boldsymbol{S}/t^2).$$

Indeed, $1 + C_1(1 + x) \le (1 + C_1)(1 + x)$ for $x \ge 0$, and N_{col}/n is bounded. The constants here are independent of t. Combining these inequalities and taking a union bound over the patch net gives

$$\mathbb{P}\{\exists \boldsymbol{X} \in \mathcal{N}_{\boldsymbol{S}} : \|\boldsymbol{G}^\top \boldsymbol{X}\|_F \le 2t\sqrt{n}\} \le \exp\left\{C_3 n - \frac{m - N_{\mathrm{col}}}{2} \log\det(\boldsymbol{I}_r + \boldsymbol{S}/t^2)\right\}.$$

Since $\boldsymbol{S} \succeq 0$ and $\operatorname{tr} \boldsymbol{S} = 1$, its largest eigenvalue is at least $1/r$, so $\log\det(\boldsymbol{I}_r + \boldsymbol{S}/t^2) \ge \log(1 + 1/(rt^2))$. Choose a fixed $t > 0$ small enough that

$$\frac{\varepsilon}{4} \log(1 + 1/(rt^2)) > C_3 + 2.$$

After fixing this t, the finite union over $\mathcal{K}_t$ has probability at most $|\mathcal{K}_t| e^{-2n}$ by (103). Together with the operator-norm failure probability, this is at most Ae^{-n} for a fixed A. Finally, on (104), an η-net approximation changes $\|\boldsymbol{G}^\top \boldsymbol{X}\|_F$ by at most $C_G\sqrt{n}\,\eta = t\sqrt{n}/4$. Thus the lower bound $2t\sqrt{n}$ at the net points gives $\inf_{\boldsymbol{X} \in \mathcal{T}} \|\boldsymbol{G}^\top \boldsymbol{X}\|_F \ge (7t/4)\sqrt{n}$. This proves (34) after decreasing c. □

H.3.2 Uniform row control

Proof. An upgrade increases the left side of (33) by at most $|I| \dim S$. For $\rho \le \varepsilon/2$, every upgraded family therefore satisfies the theorem with gap $\varepsilon/2$ and finite type family $\mathcal{U} \cup \{\mathbb{R}^r\}$. The preceding proof gives a uniform failure bound Ae^{-n}, with a fixed singular-value constant. There are at most 2^n choices of I, so a union bound gives $Ae^{-(1-\log 2)n}$. Since $1 - \log 2 > 0$, this proves the claim.

□

H.3.3 Second-moment gap

Proof. Fix S and set $T(\xi) = N_{\mathrm{act}}(\xi) \cap S$. Projecting the active equations gives

$$\boldsymbol{P}_T \boldsymbol{w} = \boldsymbol{P}_T \boldsymbol{h}_{\mathrm{act}}(\xi) - \boldsymbol{P}_T \boldsymbol{z}.$$

Because $\boldsymbol{z}$ is standard Gaussian and independent of ξ,

$$\mathbb{E}[\|\boldsymbol{P}_T \boldsymbol{w}\|^2 \mid \xi] = \dim T + \|\boldsymbol{P}_T \boldsymbol{h}_{\mathrm{act}}(\xi)\|^2 \ge \dim T.$$

Since $T \subseteq S$, second-moment domination gives

$$\mathbb{E}\dim\{N_{\mathrm{act}}(\xi) \cap S\} \le \mathbb{E}\|\boldsymbol{P}_S \boldsymbol{w}\|^2 \le (\kappa - \delta)\dim S,$$

which proves (36).

The space $N_{\mathrm{act}}(\xi)$ takes values in a finite family $\mathcal{T}_{\mathrm{act}}$ because Y and J take finitely many values. For $t \in \mathcal{T}_{\mathrm{act}}$, let $\widehat{p}_t$ and p_t be the empirical and population probabilities of type t. Then, uniformly over S,

$$\left| \frac{1}{n}\sum_{i=1}^{n} \dim(N_{\mathrm{act},i} \cap S) - \mathbb{E}\dim\{N_{\mathrm{act}}(\xi) \cap S\} \right| \le r \sum_{t \in \mathcal{T}_{\mathrm{act}}} |\widehat{p}_t - p_t| = o_{\mathbb{P}}(1).$$

Together with (35), this gives (37), for example with any fixed $\varepsilon_0 < \delta/2$ for all sufficiently large n. □

I Binary reduction, parameter dependence, and null boundaries

I.1 Reduction to the binary logistic boundary

In this subsection, we identify one-dimensional parameters and Gaussian vectors with their scalar entries. For $K = 2$, we have $q = 1$, and the linear predictor $v = \boldsymbol{x}^\top \boldsymbol{\theta} + b$ is the log-odds of class 1 relative to class 0. With the signed label $\widetilde{Y} = 2Y - 1 \in \{-1, 1\}$,

$$\ell_Y(v) = \log(1 + e^v) - Yv = \log\bigl(1 + e^{-\widetilde{Y}v}\bigr),$$

which is the binary logistic loss in Candès & Sur (2020). Changing the label encoding changes neither the parameters nor the MLE-existence event, regardless of the class probabilities.

The decision cones also reduce to the binary comparison. For a label $y \in \{0, 1\}$, write $\widetilde{y} = 2y - 1$ and define $t_+ = \max\{t, 0\}$. Then $\mathcal{C}_y = \{c \in \mathbb{R} : \widetilde{y}c \ge 0\}$, and

$$\mathrm{dist}^2(v, \mathcal{C}_y) = \inf_{c \in \mathcal{C}_y} (v - c)^2 = \inf_{t \ge 0} (\widetilde{y}v - t)^2 = (-\widetilde{y}v)_+^2.$$

Thus the positive-part penalty is the squared distance to the permitted half-line. The multinomial distance enforces all class comparisons at once.

Evaluate the boundary at $S = \gamma^2$ with $\gamma \ge 0$. The population signal coordinate and label have joint distribution

$$Z_0 \sim N(0, 1), \qquad \mathbb{P}(\widetilde{Y} = 1 \mid Z_0) = \frac{1}{1 + \exp\{-(b_0 + \gamma Z_0)\}}.$$

When $q = 1$, the trace constraints in (3) force $H = \Lambda = 1$. Take $Z_1 \sim N(0, 1)$ independent of $(Z_0, \widetilde{Y})$ and set $t_0 = -a$, $t_1 = -M$, and $Z = \widetilde{Y} Z_1$. Gaussian sign symmetry makes Z standard Gaussian and independent of $(Z_0, \widetilde{Y})$. Substitution into the distance formula gives (6), where h_{MLE} is the function in Candès & Sur (2020, Theorem 1).

Their feature dimension is $p = d$, their true intercept is $\beta_0 = b_0$, and their signal strength is $\gamma_0 = \gamma$. Their ratio $\kappa = p/n$ equals $1/\alpha$, so their existence condition $\kappa < h_{\mathrm{MLE}}$ is precisely $\alpha > \alpha_\star(\gamma^2, b_0)$. We recover both their boundary and their probability limits whenever $\boldsymbol{\Theta}_0^\top \boldsymbol{\Sigma} \boldsymbol{\Theta}_0 \to \gamma^2 \ge 0$, with any fixed true intercept and nonsingular Gaussian feature covariance.

If the fitted intercept is fixed to zero, the same reduction fixes $t_0 = 0$ while leaving b_0 in the label law:

$$\frac{1}{\alpha_\star^0(\gamma^2, b_0)} = \inf_{t_1 \in \mathbb{R}} \mathbb{E}(t_1 \widetilde{Y} Z_0 - Z)_+^2.$$

This covers omission of a nonzero true intercept as well.

When $b_0 = 0$, $(\widetilde{Y}, Z_0)$ and $(-\widetilde{Y}, -Z_0)$ have the same law. This sign change leaves $\widetilde{Y} Z_0$ unchanged, so for each fixed t_1 the objective in (6) is even in t_0. Convexity in t_0 then makes $t_0 = 0$ a minimizer. Under this data law, estimating an intercept therefore gives the same boundary as fixing it to zero (Candès & Sur, 2020, Lemma 1 and Theorem 2):

$$\alpha_\star(\gamma^2, 0) = \alpha_\star^0(\gamma^2, 0).$$

Swapping the classes also gives $\alpha_\star(\gamma^2, b_0) = \alpha_\star(\gamma^2, -b_0)$. At zero signal and zero true intercept, the null calculation (Corollary I.3) gives $\alpha_\star(0,0) = 2$, or $h_{\mathrm{MLE}}(0,0) = 1/2$, recovering Cover's random-dichotomy threshold (Cover, 1965).

A nonzero true intercept can matter even when the coefficients vanish. For $\boldsymbol{\Theta}_0 = \mathbf{0}$ and $b_0 = \log 9$, we have $\mathbb{P}(Y = 1) = 0.9$ and $\mathbb{P}(Y = 0) = 0.1$. The separate zero-slope formula of Candès & Sur (2020, Lemma 1) gives a critical ratio of about 3.9 with a fitted intercept, compared with 2 for balanced labels.

I.2 Invariance under contrast coordinates

Lemma I.1 (Invariance under contrast coordinates)**.** *Let $\boldsymbol{T} \in \mathbb{R}^{q\times q}$ be invertible. For $\boldsymbol{v} \in \mathbb{R}^q$ and each class label $y \in \{0, 1, \ldots, q\}$, represent the same class linear predictors in new coordinates by*

$$\widetilde{\boldsymbol{v}} = \boldsymbol{T}\boldsymbol{v}, \qquad \widetilde{\mathcal{C}}_y = \boldsymbol{T}\mathcal{C}_y, \qquad \widetilde{\boldsymbol{\eta}} = \boldsymbol{T}\boldsymbol{\eta}.$$

The label Y is unchanged: its conditional probabilities at $\widetilde{\boldsymbol{\eta}}$ are the original softmax probabilities evaluated at $\boldsymbol{T}^{-1}\widetilde{\boldsymbol{\eta}}$. Thus the probability map and the cones are both expressed in the new coordinates. Then finite separation and population feasibility at every α are unchanged for either fitting choice. The corresponding critical ratios $\alpha^0_\star(\boldsymbol{S}, \boldsymbol{b}_0)$ and $\alpha_\star(\boldsymbol{S}, \boldsymbol{b}_0)$ are therefore unchanged when the linear predictors, cones, and label probabilities are transformed together. For the intercept model, transform the intercept direction by $\widetilde{\boldsymbol{a}} = \boldsymbol{T}\boldsymbol{a}$. The matrix distance formula is also unchanged: for every $\widetilde{\boldsymbol{\Lambda}} \in \mathbb{S}^q_+$, transform its metric by $\boldsymbol{\Lambda} = \boldsymbol{T}^\top\widetilde{\boldsymbol{\Lambda}}\boldsymbol{T}$.

Proof. At the finite-sample level, if $\boldsymbol{B}^\top\boldsymbol{x}_i \in \mathcal{C}_{Y_i}$, then $\widetilde{\boldsymbol{B}} = \boldsymbol{B}\boldsymbol{T}^\top$ satisfies $\widetilde{\boldsymbol{B}}^\top\boldsymbol{x}_i = \boldsymbol{T}\boldsymbol{B}^\top\boldsymbol{x}_i \in \widetilde{\mathcal{C}}_{Y_i}$; applying $\boldsymbol{T}^{-1}$ gives the converse.

For a population witness $(\boldsymbol{H}, \boldsymbol{M}, \boldsymbol{u})$, choose the left polar factorization $\boldsymbol{T}\boldsymbol{H} = \widetilde{\boldsymbol{H}}\boldsymbol{O}$, where $\widetilde{\boldsymbol{H}} = (\boldsymbol{T}\boldsymbol{H}^2\boldsymbol{T}^\top)^{1/2}$ and $\boldsymbol{O}$ is orthogonal, extending the partial isometry if $\boldsymbol{H}$ is singular. Set $\widetilde{\boldsymbol{z}}_1 = \boldsymbol{O}\boldsymbol{z}_1$, $\widetilde{\boldsymbol{M}} = \boldsymbol{T}\boldsymbol{M}$, and $\widetilde{\boldsymbol{u}} = \boldsymbol{T}\boldsymbol{u}$. Then $\widetilde{\boldsymbol{z}}_1$ is still standard Gaussian independent of the side variables, the linear predictor is transformed by $\boldsymbol{T}$, and

$$\mathbb{E}[\widetilde{\boldsymbol{u}}\widetilde{\boldsymbol{u}}^\top] \preceq \alpha^{-1}\boldsymbol{T}\boldsymbol{H}^2\boldsymbol{T}^\top = \alpha^{-1}\widetilde{\boldsymbol{H}}^2.$$

Applying the same construction to $\boldsymbol{T}^{-1}$ proves equivalence. For linear predictors with an intercept, additionally set $\widetilde{\boldsymbol{a}} = \boldsymbol{T}\boldsymbol{a}$ in both the sample and population constructions; the second-moment bound is unchanged.

Finally, let $\boldsymbol{a} \in \mathbb{R}^q$, let $\mathcal{C} \subseteq \mathbb{R}^q$ be a closed cone, and let $\widetilde{\boldsymbol{\Lambda}} \in \mathbb{S}^q_+$. Then

$$\mathrm{dist}^2_{\widetilde{\boldsymbol{\Lambda}}}(\boldsymbol{T}\boldsymbol{a}, \boldsymbol{T}\mathcal{C}) = \mathrm{dist}^2_{\boldsymbol{T}^\top\widetilde{\boldsymbol{\Lambda}}\boldsymbol{T}}(\boldsymbol{a}, \mathcal{C}), \qquad \mathrm{tr}(\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{H}}^2) = \mathrm{tr}(\boldsymbol{T}^\top\widetilde{\boldsymbol{\Lambda}}\boldsymbol{T}\boldsymbol{H}^2).$$

Congruence by $\boldsymbol{T}$ bijects the nonzero PSD weight rays. Since the ratio in (5) or (3) is homogeneous in its metric, imposing trace one before or after the congruence does not change the infimum. Dividing the transformed candidate coefficients by $\mathrm{tr}\,\widetilde{\boldsymbol{H}} > 0$ restores the factor's trace-one constraint; cone homogeneity leaves every ratio unchanged. □

I.3 Why the full parameter configuration matters

Throughout this subsection, coordinates and Euclidean norms are the baseline coordinates of Section 2.

Proposition I.2. *For every fixed $K \geq 3$, the map $\boldsymbol{S} \mapsto \alpha_\star(\boldsymbol{S}, \mathbf{0})$ over positive-definite matrices is not a function of the eigenvalues of $\boldsymbol{S}$ alone. At the fixed matrix $\boldsymbol{S}_0 = \boldsymbol{I}_q + \mathbf{1}_q\mathbf{1}_q^\top$, the map $\boldsymbol{b}_0 \mapsto \alpha_\star(\boldsymbol{S}_0, \boldsymbol{b}_0)$ is not a function of the Euclidean norm of $\boldsymbol{b}_0$ alone.*

Proof. We use class relabeling to obtain an invariance that any boundary depending only on covariance eigenvalues or the intercept norm would have to satisfy. In each case, this would make the boundary constant along a sequence for which explicit population witnesses force it to diverge.

Swapping the baseline class 0 with class 1 transforms a baseline linear predictor vector by

$$\boldsymbol{T}(v_1, v_2, \dots, v_q)^\top = (-v_1, v_2 - v_1, \dots, v_q - v_1)^\top.$$

For the label permutation π swapping 0 and 1, one has $\boldsymbol{T}\mathcal{C}_y = \mathcal{C}_{\pi(y)}$, and the softmax probabilities are permuted accordingly. By Lemma I.1, followed by an orthogonal rotation of the Gaussian signal coordinates to the canonical square-root representation, this gives

$$\alpha_\star(\boldsymbol{T}\boldsymbol{S}\boldsymbol{T}^\top, \boldsymbol{T}\boldsymbol{b}) = \alpha_\star(\boldsymbol{S}, \boldsymbol{b}).$$

Suppose first that the boundary at $\boldsymbol{b} = \boldsymbol{0}$ depended only on the eigenvalues of $\boldsymbol{S}$. It would then be invariant under every orthogonal conjugation. The singular values of $\boldsymbol{T}$ are σ, σ^{-1} and $q-2$ copies of 1, where

$$\sigma^2 = \frac{q + 1 + \sqrt{(q-1)(q+3)}}{2} > 1.$$

Choose a singular-value decomposition $\boldsymbol{T} = \boldsymbol{U}\,\mathrm{diag}(\sigma, \sigma^{-1}, 1, \dots, 1)\boldsymbol{V}^\top$, and set

$$\boldsymbol{D}_m = \mathrm{diag}(\sigma^{2m}, \sigma^{-2m}, 1, \dots, 1), \qquad m = 0, 1, \dots.$$

Since $\boldsymbol{T}\boldsymbol{V}\boldsymbol{D}_m\boldsymbol{V}^\top\boldsymbol{T}^\top = \boldsymbol{U}\boldsymbol{D}_{m+1}\boldsymbol{U}^\top$, the assumed orthogonal invariance and the actual class-relabeling invariance would give

$$\alpha_\star(\boldsymbol{D}_{m+1}, \boldsymbol{0}) = \alpha_\star(\boldsymbol{T}\boldsymbol{V}\boldsymbol{D}_m\boldsymbol{V}^\top\boldsymbol{T}^\top, \boldsymbol{0}) = \alpha_\star(\boldsymbol{V}\boldsymbol{D}_m\boldsymbol{V}^\top, \boldsymbol{0}) = \alpha_\star(\boldsymbol{D}_m, \boldsymbol{0}).$$

Thus all these values would equal the finite number $\alpha_\star(\boldsymbol{I}_q, \boldsymbol{0})$.

We show instead that $\alpha_\star(\boldsymbol{D}_m, \boldsymbol{0}) \to \infty$ by constructing population feasibility witnesses. Under the model with true linear predictor covariance $\boldsymbol{D}_m$, the nonbaseline linear predictors are

$$(\sigma^m z_{01}, \sigma^{-m} z_{02}, z_{03}, \dots, z_{0q}), \qquad \boldsymbol{z}_0 \sim N(\boldsymbol{0}, \boldsymbol{I}_q).$$

The conditional probability of class 1 tends to $\mathbb{1}_{\{z_{01}>0\}}$ almost surely. Coupling the class-1 indicators with a common uniform variable independent of $\boldsymbol{z}_0$ therefore gives $\epsilon_m \to \mathrm{sign}(z_{01})$ almost surely, where $\epsilon_m = 1$ if $Y = 1$ and $\epsilon_m = -1$ otherwise.

Let $\boldsymbol{e}_1$ be the first coordinate vector and $G = (\boldsymbol{z}_1)_1$ the first coordinate of the independent Gaussian vector $\boldsymbol{z}_1$. A vector $r\boldsymbol{e}_1$ belongs to $\mathcal{C}_Y$ exactly when $\epsilon_m r \geq 0$: all classes other than 1 have linear predictor zero, and ties are allowed. Write $x_+ = \max\{x, 0\}$. For $t \geq 0$, take

$$\boldsymbol{H} = \boldsymbol{e}_1\boldsymbol{e}_1^\top, \qquad \boldsymbol{M} = t\boldsymbol{e}_1\boldsymbol{e}_1^\top, \qquad \boldsymbol{a} = \boldsymbol{0}, \qquad \boldsymbol{u} = \boldsymbol{e}_1\epsilon_m[-\epsilon_m(G + tz_{01})]_+.$$

The corrected linear predictor lies in $\mathcal{C}_Y$ almost surely, and

$$\mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top] = f_m(t)\boldsymbol{H}^2, \qquad f_m(t) = \mathbb{E}[-\epsilon_m(G + tz_{01})]_+^2.$$

Conditional Gaussian symmetry and dominated convergence give, for fixed t,

$$f_m(t) = \mathbb{E}(G - t\epsilon_m z_{01})_+^2 \longrightarrow \mathbb{E}(G - t|z_{01}|)_+^2.$$

The domination is $(|G| + t|z_{01}|)^2$. The limiting expectation tends to zero as $t \to \infty$, now dominated by G_+^2. For any fixed $\alpha > 1$, choosing t and then all sufficiently large m makes $f_m(t) < \alpha^{-1}$. The displayed nontrivial witness is then α-feasible in (14). Hence $\alpha_\star(\boldsymbol{D}_m, \boldsymbol{0}) \to \infty$, contradicting the proposed spectral invariance.

For the intercept claim, direct calculation gives $\boldsymbol{T}\boldsymbol{S}_0\boldsymbol{T}^\top = \boldsymbol{S}_0$ and $\boldsymbol{T}\boldsymbol{e}_1 = -\boldsymbol{1}_q$. If $\alpha_\star(\boldsymbol{S}_0, \boldsymbol{b}) = f(\|\boldsymbol{b}\|_2)$, class relabeling would imply

$$f(r) = f(\sqrt{q}\, r) = f(q^{m/2} r), \qquad r > 0, \quad m = 0, 1, \dots.$$

On the other hand, keep $\boldsymbol{S}_0$ fixed and let $\boldsymbol{b}_R = R\boldsymbol{e}_1$ with $R \to \infty$. The conditional probability of class 1 tends to one. Use the same rank-one construction with $\boldsymbol{M} = \boldsymbol{0}$ and $\boldsymbol{a} = t\boldsymbol{e}_1$, replacing $G + tz_{01}$ by $G + t$ in the correction above. If $\epsilon_R = 1$ on class 1 and -1 otherwise, its second-moment cost satisfies

$$\mathbb{E}[-\epsilon_R(G + t)]_+^2 \longrightarrow \mathbb{E}(-G - t)_+^2 \longrightarrow 0,$$

where the limits are taken first in R and then in t. The same feasibility argument proves $\alpha_\star(\boldsymbol{S}_0, R\boldsymbol{e}_1) \to \infty$. This contradicts $f(q^{m/2} r) = f(r) < \infty$ for any fixed $r > 0$, since $q \geq 2$. Finiteness at every fixed parameter value follows from Proposition B.6.

These limits compare population boundaries across fixed models; they do not invoke a finite-sample theorem with diverging signal Gram matrices or intercepts. □

I.4 THE NULL BOUNDARY

In the null model, uniform labels are independent of the features. For this calculation, choose regular-simplex vertices $\boldsymbol{s}_0, \dots, \boldsymbol{s}_q \in \mathbb{R}^q$ with $\sum_{j=0}^q \boldsymbol{s}_j = \mathbf{0}$ and $\|\boldsymbol{s}_j - \boldsymbol{s}_k\| = 1$ for $j \neq k$. A vector $\boldsymbol{g} \in \mathbb{R}^q$ represents the centered class linear predictors $\boldsymbol{s}_j^\top \boldsymbol{g}$. Their decision cones are

$$\mathcal{K}_y = \{\boldsymbol{g} \in \mathbb{R}^q : \boldsymbol{s}_y^\top \boldsymbol{g} \geq \boldsymbol{s}_j^\top \boldsymbol{g} \text{ for every } j = 0, \dots, q\}.$$

The invertible map $\boldsymbol{T}\boldsymbol{g} = ((\boldsymbol{s}_j - \boldsymbol{s}_0)^\top \boldsymbol{g})_{j=1}^q$ sends $\mathcal{K}_y$ to the baseline cone $\mathcal{C}_y$. By Lemma I.1, working in these centered coordinates preserves both separation and the population boundary.

Set $\mathcal{D}_K = \mathcal{K}_0^\circ$, with statistical dimension

$$\delta(\mathcal{D}_K) = \mathbb{E}\|\Pi_{\mathcal{D}_K}(\boldsymbol{z})\|^2, \qquad \boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{I}_q),$$

where $\Pi_{\mathcal{D}_K}$ is Euclidean projection. Each $\mathcal{K}_y$ and its polar are full-dimensional pointed simplicial cones; the polar is generated by $\boldsymbol{s}_j - \boldsymbol{s}_y$, $j \neq y$. Relabeling the simplex vertices is an orthogonal transformation, so all these polar cones have the same statistical dimension.

Corollary I.3. *For fixed $K \geq 2$, let the observations be independent, with $N(\mathbf{0}, \boldsymbol{I}_d)$ features and labels equally likely in $\{0, \dots, K-1\}$ and independent of the features. Then*

$$\boxed{\alpha_\star(\mathbf{0}, \mathbf{0}) = \alpha_\star^0(\mathbf{0}, \mathbf{0}) = \frac{K-1}{\delta(\mathcal{D}_K)}.} \tag{107}$$

For either fitted model, if $n, d \to \infty$ with $n/d \to \alpha > 1$ away from this boundary, the probability of a finite MLE tends to zero below it and to one above it. The same limits hold for softmax labels with $\boldsymbol{b}_0 = \mathbf{0}$ and $\boldsymbol{\Theta}_0^\top \boldsymbol{\Theta}_0 \to \mathbf{0}$.

Table 1: Population boundaries for equally likely labels independent of the features, common to fitting with and without an intercept.

Classes K	Critical ratio $\alpha_\star(\mathbf{0}, \mathbf{0})$
2	2.000000
3	2.400000
4	2.758012
5	3.088686
6	3.399635

Proof of Corollary I.3. Let Y be uniform on $\{0, \dots, q\}$ and independent of $\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{I}_q)$. Define

$$\boldsymbol{p} = \Pi_{\mathcal{K}_Y^\circ}(\boldsymbol{z}), \qquad f_0 = \frac{1}{q}\mathbb{E}\|\boldsymbol{p}\|^2 = \frac{\delta(\mathcal{D}_K)}{q}.$$

Moreau's decomposition (Amelunxen et al., 2014, Appendix B.3) gives $\boldsymbol{z} - \boldsymbol{p} = \Pi_{\mathcal{K}_Y}(\boldsymbol{z})$ and $\boldsymbol{p}^\top \boldsymbol{z} = \|\boldsymbol{p}\|^2$. Since both cones have nonempty interior, their Gaussian projections have positive second moments; hence $0 < f_0 < 1$.

The required projection identities are

$$\mathbb{E}\boldsymbol{p} = \mathbf{0}, \qquad \mathbb{E}[\boldsymbol{p}\boldsymbol{p}^\top] = \mathbb{E}[\boldsymbol{p}\boldsymbol{z}^\top] = f_0 \boldsymbol{I}_q. \tag{108}$$

Indeed, permuting simplex vertices acts orthogonally and preserves $(\boldsymbol{z}, Y)$ in law; projection commutes with this action. Identify $\mathbb{R}^q$ with the zero-sum subspace of $\mathbb{R}^K$ by $\boldsymbol{v} \mapsto (\sqrt{2}\boldsymbol{s}_j^\top \boldsymbol{v})_{j=0}^q$. The only permutation-invariant vector there is zero. Extending either moment matrix by zero on constant vectors gives a matrix invariant under every coordinate permutation, hence with common diagonal and off-diagonal entries. It is therefore scalar on the zero-sum subspace. Both traces are qf_0, by Moreau orthogonality, proving (108).

Population boundary. There is no signal block in this model. Equivalently, any independent dummy term $\boldsymbol{M}\boldsymbol{z}_0$ in the population criterion can be removed by conditioning on $(\boldsymbol{z}, Y)$: convexity preserves

cone membership and conditional Jensen can only decrease the correction's second-moment matrix. We therefore consider candidates

$$\Delta \boldsymbol{v} = \boldsymbol{H}\boldsymbol{z} + \boldsymbol{a} + \boldsymbol{u} \in \mathcal{K}_Y, \qquad \mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top] \preceq \kappa \boldsymbol{H}^2,$$

where $\boldsymbol{a}$ is unrestricted when an intercept is fitted and zero otherwise. Here $\kappa = \alpha^{-1}$. The choice $(\boldsymbol{H}, \boldsymbol{a}, \boldsymbol{u}) = (\boldsymbol{I}_q, \boldsymbol{0}, -\boldsymbol{p})$ is feasible at $\kappa = f_0$, by Moreau's decomposition and (108).

Conversely, suppose $\kappa < f_0$ and such a candidate is feasible. Couple it to $\boldsymbol{p}$ using the same $(\boldsymbol{z}, Y)$. The identities above give

$$\mathbb{E}[\boldsymbol{p}^\top(\boldsymbol{H}\boldsymbol{z} + \boldsymbol{a})] = f_0 \operatorname{tr} \boldsymbol{H}.$$

Let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_q$ be an orthonormal eigenbasis of $\boldsymbol{H} \succeq \boldsymbol{0}$, with eigenvalues $\lambda_j \geq 0$. Cauchy–Schwarz and the second-moment bound imply

$$\mathbb{E}[\boldsymbol{p}^\top \boldsymbol{u}] \geq -\sum_{j=1}^{q} \sqrt{\mathbb{E}[(\boldsymbol{v}_j^\top \boldsymbol{p})^2]\, \mathbb{E}[(\boldsymbol{v}_j^\top \boldsymbol{u})^2]} \geq -\sqrt{\kappa f_0} \sum_{j=1}^{q} \lambda_j.$$

If $\boldsymbol{H} \neq \boldsymbol{0}$, it follows that

$$\mathbb{E}[\boldsymbol{p}^\top \Delta \boldsymbol{v}] \geq (f_0 - \sqrt{\kappa f_0}) \operatorname{tr} \boldsymbol{H} > 0,$$

contradicting $\boldsymbol{p} \in \mathcal{K}_Y^\circ$ and $\Delta \boldsymbol{v} \in \mathcal{K}_Y$. If $\boldsymbol{H} = \boldsymbol{0}$, the bound forces $\boldsymbol{u} = \boldsymbol{0}$, and every label has positive probability, so $\boldsymbol{a} \in \bigcap_y \mathcal{K}_y = \{\boldsymbol{0}\}$. A retained dummy term similarly forces $\boldsymbol{M} = \boldsymbol{0}$. Thus no nontrivial candidate exists at $\kappa < f_0$, for either intercept convention, proving the population identity in (107).

Finite-sample transition. Apply Theorems 2.3 and 2.5 with $\boldsymbol{S}_\star = \boldsymbol{0}$ and $\boldsymbol{b}_0 = \boldsymbol{0}$, using the population identity (107). This gives both strict sides of the transition for zero coefficients and for every sequence with $\boldsymbol{\Theta}_0^\top \boldsymbol{\Theta}_0 \to \boldsymbol{0}$, under the corresponding fitting convention. □

I.5 Evaluation of the null constants

The polar cone $\mathcal{D}_K$ is generated by $\boldsymbol{s}_j - \boldsymbol{s}_0$, $j = 1, \dots, q$. Multiplying these generators by $\sqrt{2}$ leaves the cone unchanged and gives a Gram matrix with diagonal 2 and off-diagonal 1. Thus $\mathcal{D}_K$ is isometric to the cone $C_q(1)$ in the notation of Kabluchko & Zaporozhets (2020). For $m \geq 1$ and $r \in [-1/m, \infty)$, let $g_m(r)$ be the negative-orthant probability of a centered m-variate Gaussian vector whose covariance has diagonal $1 + r$ and off-diagonal r, and set $g_0(r) = 1$. Writing $\nu_\ell(\mathcal{C})$ for the ℓth conic intrinsic volume of $\mathcal{C}$, Proposition 1.3 of Kabluchko & Zaporozhets (2020) gives

$$\nu_\ell(C_q(1)) = \binom{q}{\ell} g_\ell\left(-\frac{1}{1+\ell}\right) g_{q-\ell}\left(\frac{1}{1+\ell}\right), \qquad \ell = 0, \dots, q. \tag{109}$$

The statistical dimension of a polyhedral cone is its intrinsic-volume mean (Amelunxen et al., 2014, Definition 2.2):

$$\delta(\mathcal{C}) = \sum_{\ell=0}^{q} \ell \nu_\ell(\mathcal{C}). \tag{110}$$

These formulas give the values in Table 1. For $K = 3$, $q = 2$, the values $g_1(-1/2) = g_1(1/2) = 1/2$, $g_2(1) = 1/3$, and $g_2(-1/3) = 1/6$ give

$$(\nu_0, \nu_1, \nu_2) = \left(\frac{1}{3}, \frac{1}{2}, \frac{1}{6}\right), \qquad \delta(\mathcal{D}_3) = \frac{1}{2} + 2 \cdot \frac{1}{6} = \frac{5}{6}.$$

Consequently $\alpha_\star(\boldsymbol{0}, \boldsymbol{0}) = \alpha_\star^0(\boldsymbol{0}, \boldsymbol{0}) = 12/5$, the three-class value used in the main text.

J The fitted intercept can change the boundary

The equality between $\alpha_\star(\boldsymbol{S}, \boldsymbol{0})$ and $\alpha_\star^0(\boldsymbol{S}, \boldsymbol{0})$ in the binary model does not extend to every positive-definite $\boldsymbol{S}$. We first establish the binary comparison needed for a three-class counterexample. Throughout this section, the feature covariance is $\boldsymbol{I}_d$ and the sample-to-dimension ratio is n/d.

Lemma J.1 (Binary comparison). *Let $Z \sim N(0,1)$, and let $B \in \{-1,1\}$ have conditional probability*

$$p(z) := \mathbb{P}(B = 1 \mid Z = z) = \frac{e^{b+\gamma z}}{1 + e^{b+\gamma z}}, \qquad b \in \mathbb{R}, \quad \gamma \geq 0.$$

Take $G \sim N(0,1)$ independent of (Z, B) and define

$$\begin{aligned} \rho(s) &= \mathbb{E}(G - s)_+^2, \\ F(a,t) &= \mathbb{E}\big[p(Z)\rho(a + tZ) + (1 - p(Z))\rho(-a - tZ)\big], \\ h_0 &= \min_{t \in \mathbb{R}} F(0,t), \qquad h_1 = \min_{a,t \in \mathbb{R}} F(a,t). \end{aligned} \tag{111}$$

For independent $N(\mathbf{0}, \boldsymbol{I}_d)$ features and binary labels generated by $b + \gamma x_1$, the finite-MLE boundaries are $1/h_0$ when the fitted intercept is fixed to zero and $1/h_1$ when it is estimated. Both minima are positive and attained. Moreover, $h_1 < h_0$ if $b > 0$, whereas $h_1 = h_0$ if $b = 0$.

Proof. On $|Z| \leq 1$, both $p(Z)$ and $1 - p(Z)$ are bounded below by a positive constant c. Since $\rho(s) + \rho(-s) = 1 + s^2$,

$$F(a,t) \geq c\,\mathbb{E}\big[\{1 + (a + tZ)^2\}\mathbb{1}_{\{|Z| \leq 1\}}\big] \geq c'(1 + a^2 + t^2)$$

for some $c' > 0$. Continuity therefore gives positive attained minima. To identify the boundaries, condition on (Z, B) and use symmetry of G to write

$$F(a,t) = \mathbb{E}\rho\big(B(a + tZ)\big) = \mathbb{E}[-B(G + a + tZ)]_+^2.$$

This is the scalar distance objective in Section 3.1. Consequently, Theorems 2.3 and 2.5 give

$$\alpha_\star^0(\gamma^2, b) = h_0^{-1}, \qquad \alpha_\star(\gamma^2, b) = h_1^{-1},$$

by restricting $a = 0$ or optimizing over a, respectively. Both specializations allow the same arbitrary true intercept b.

Finally, $\rho'(s) = -2\mathbb{E}(G - s)_+ < 0$. Differentiation under the expectation is justified locally by an integrable envelope proportional to $1 + |Z|$. Changing Z to $-Z$ in the second term of $\partial_a F(0,t)$ gives

$$\partial_a F(0,t) = \mathbb{E}\big[\{p(Z) + p(-Z) - 1\}\rho'(tZ)\big]. \tag{112}$$

If $b > 0$, then $p(z) > 1 - p(-z)$ for every z, so this derivative is strictly negative for every finite t. Moving to a small positive a at a minimizer of $F(0,t)$ proves $h_1 < h_0$. If $b = 0$, the derivative vanishes for every t, and convexity in a gives $F(a,t) \geq F(0,t)$, hence $h_1 = h_0$. □

Proof of Proposition 3.1. We first identify both transitions in a model with identical true linear predictors for classes 1 and 2, and then use Proposition G.1 to transfer their strict difference to positive-definite signal matrices. The key step is to show that, at the relevant sample ratios, every multinomial separator assigns identical predictors to classes 1 and 2 with probability tending to one.

Take true linear predictors

$$(\eta_0(\boldsymbol{x}), \eta_1(\boldsymbol{x}), \eta_2(\boldsymbol{x})) = (0, \gamma x_1, \gamma x_1), \qquad \boldsymbol{S}_0 = \gamma^2 \mathbf{1}\mathbf{1}^\top, \qquad \mathbf{1} = (1,1)^\top.$$

Every true intercept is zero. Group classes 1 and 2 by setting $B = 1$ for those labels and $B = -1$ for class 0. Then

$$\mathbb{P}(B = 1 \mid \boldsymbol{x}) = \frac{2e^{\gamma x_1}}{1 + 2e^{\gamma x_1}} = \frac{e^{\log 2 + \gamma x_1}}{1 + e^{\log 2 + \gamma x_1}}. \tag{113}$$

Thus the grouped binary model has true intercept $b = \log 2 > 0$. Writing $\pi_{\mathrm{grp}} = \mathbb{P}(B = 1)$, symmetry and $p(z) + p(-z) > 1$ imply $\pi_{\mathrm{grp}} > 1/2$. Conditional on the covariates and grouped labels, each label in the group is an independent fair choice between classes 1 and 2.

If $n/d \to \alpha > 2/\pi_{\mathrm{grp}}$, their number N satisfies $N/d \to \alpha\pi_{\mathrm{grp}} > 2$ in probability. Their feature distribution is absolutely continuous, hence in both linear and affine general position almost surely. Conditional on these features, Cover's counting formula (Cover, 1965) bounds the probability of an affine separator of their fair labels by

$$2^{1-N} \sum_{j=0}^{d} \binom{N-1}{j} \longrightarrow 0;$$

the homogeneous bound has upper index $d-1$. The limits follow from a binomial lower-tail bound. These bounds also exclude nonzero weak separators: a nonzero homogeneous or affine hyperplane contains at most $d-1$ or d selected points, respectively, and their original or augmented rows are linearly independent. A perturbation can therefore make their signed margins positive while preserving the other, already positive margins. A direction with every margin zero is excluded by full column rank. Hence every nonzero separating direction would yield a strict one.

Let v_1, v_2 be the linear predictor functions in a multinomial separating direction. Their difference weakly separates the class-1 observations from the class-2 observations. With probability tending to one, the preceding argument forces this difference to be the zero function, so $v_1 = v_2$, including their intercepts when these are fitted. The remaining inequalities are exactly those for the grouped binary labels. Conversely, at every sample-to-dimension ratio a grouped binary separator gives a multinomial separator by assigning the same binary linear predictor to both classes. The resulting ties between classes 1 and 2 are allowed. The fixed $\log 2$ offset in the grouped likelihood does not affect its recession directions and therefore does not affect this separation equivalence.

Let $A_0 = 1/h_0$ and $A_1 = 1/h_1$ for the binary law $b = \log 2$ in Lemma J.1. We have $A_1 > A_0$. Provided $A_0 > 2/\pi_{\mathrm{grp}}$, these are also the exact three-class thresholds at $\boldsymbol{S}_0$: below each boundary, grouped binary separation gives multinomial separation; above it, the forced equality $v_1 = v_2$ rules out every other multinomial separator.

To check that the preceding condition holds at the explicit finite signal strength $\gamma = 10$, it suffices to show $h_0 < 1/4$. Polar coordinates for (G, Z) give

$$\mathbb{E}\rho(|Z|) = \mathbb{E}(G - |Z|)_+^2 = \frac{2}{\pi}\int_0^{\pi/4} (1 - \sin 2\theta)\, d\theta = \frac{1}{2} - \frac{1}{\pi}.$$

The conditional probability that B disagrees with $\operatorname{sign}(Z)$ is at most $2e^{-\gamma|Z|}$, while

$$0 \le \rho(-u) - \rho(u) \le 4u(\mathbb{E}|G| + u), \qquad u \ge 0,$$

by integrating $|\rho'(s)| \le 2(\mathbb{E}|G| + |s|)$ over $[-u, u]$. Using the normal density bound $(2\pi)^{-1/2}$ therefore yields

$$\begin{aligned} h_0 \le F(0,1) &\le \frac{1}{2} - \frac{1}{\pi} + 8\mathbb{E}\big[e^{-\gamma|Z|}|Z|(\mathbb{E}|G| + |Z|)\big] \\ &\le \frac{1}{2} - \frac{1}{\pi} + \frac{16}{\pi\gamma^2} + \frac{32}{\sqrt{2\pi}\gamma^3} < \frac{1}{4} \qquad (\gamma = 10). \end{aligned} \tag{114}$$

Thus $4 < A_0 < A_1 < \infty$ and $2/\pi_{\mathrm{grp}} < 4$, as required. Consequently, Theorems 2.3 and 2.5 identify $A_0 = \alpha_\star^0(\boldsymbol{S}_0, \boldsymbol{0})$ and $A_1 = \alpha_\star(\boldsymbol{S}_0, \boldsymbol{0})$. Only the intermediate signal matrix $\boldsymbol{S}_0 = 100\mathbf{1}\mathbf{1}^\top$ is singular; the feature covariance has remained $\boldsymbol{I}_d$ throughout.

It remains to transfer the strict difference to $\boldsymbol{S}_\varepsilon = \boldsymbol{S}_0 + \varepsilon \boldsymbol{I}_2$. Choose $A_0 < a_- < a_+ < A_1$. We claim that, for all sufficiently small fixed $\varepsilon > 0$,

$$\alpha_\star^0(\boldsymbol{S}_\varepsilon, \boldsymbol{0}) \le a_- < a_+ \le \alpha_\star(\boldsymbol{S}_\varepsilon, \boldsymbol{0}). \tag{115}$$

This follows from Proposition G.1 by a diagonal argument, which avoids assuming continuity of the boundary at a singular matrix. For example, suppose there were $\varepsilon_j \downarrow 0$ with $\alpha_\star^0(\boldsymbol{S}_{\varepsilon_j}, \boldsymbol{0}) > a_-$. The no-intercept theorem lets us choose increasing dimensions d_j such that, with $n_j = \lfloor a_- d_j \rfloor$, the probability of a finite MLE under $\boldsymbol{S}_{\varepsilon_j}$ is less than $1/4$. Define a sequence $\boldsymbol{T}_d$ by $\boldsymbol{T}_{d_j} = \boldsymbol{S}_{\varepsilon_j}$ and $\boldsymbol{T}_d = \boldsymbol{S}_0$ at all other dimensions. Then $\boldsymbol{T}_d \to \boldsymbol{S}_0$, and Proposition G.1 applied with $n_d = \lfloor a_- d \rfloor$ forces that probability to tend to one at sample-to-dimension ratio $a_- > A_0$, a contradiction. The same argument at $a_+ < A_1$ rules out a sequence with $\alpha_\star(\boldsymbol{S}_{\varepsilon_j}, \boldsymbol{0}) < a_+$: choose the dimensions so the fitted-intercept MLE exists with probability greater than $3/4$, whereas the comparison forces this probability to tend to zero. This proves (115) and the proposition. □

K Numerical evaluation of the phase boundary

The reference curves are computed from the population boundary, independently of the simulated separation outcomes. Binary and rank-one cases reduce to one-dimensional quadrature; null cases have explicit formulas; the general multiclass cases require Gaussian integration and a matrix search. We describe these computations and the checks used for the plotted curves below.

K.1 Binary and null reference values

For binary log-odds $b + \gamma Z$, $Z \sim N(0,1)$, averaging over the auxiliary Gaussian in (111) gives

$$\begin{aligned} \rho(s) &= (1+s^2)\Phi(-s) - s\phi(s), \\ \rho'(s) &= 2\{s\Phi(-s) - \phi(s)\}, \qquad \rho''(s) = 2\Phi(-s), \end{aligned} \tag{116}$$

where ϕ and Φ are the standard normal density and distribution function. Summing over the label leaves

$$\begin{aligned} F(a,t) &= \int_{\mathbb{R}} \big[p(z)\rho(a+tz) + (1-p(z))\rho(-a-tz)\big]\phi(z)\,dz, \\ p(z) &= \{1+\exp[-(b+\gamma z)]\}^{-1}. \end{aligned} \tag{117}$$

The boundary is $1/\min_{a,t} F(a,t)$ with a fitted intercept and $1/\min_t F(0,t)$ without one. Both problems are convex. The code uses $\rho(-s)$ and reverses both optimization variables, giving the same minima.

Quadrature and acceptance checks. We minimize the integral by Gauss–Hermite quadrature and Newton's method with analytical derivatives. Successive quadrature refinements and an independent adaptive integral must agree within $10^{-10} + 10^{-9}\alpha_\star$; the adaptive gradient infinity norm must be at most 10^{-10}. Implementation settings are specified in `code/reporting.py`.

For the rank-one model $(0, 10Z, 10Z)$, grouping labels 1 and 2 gives $b = \log 2$ and $\gamma = 10$. After checking the reduction condition $1/h_0 > 2/\mathbb{P}(Y \in \{1,2\})$ from Appendix J, the boundaries are 9.492705191964 without a fitted intercept and 10.128319893985 with one. Independent integration checks differ by less than 2.35×10^{-12}; this agreement does not certify integration error.

Null values. For $K = 2, 3, 4$, Appendix I.4 gives the exact boundaries 2, $12/5$, and $3/[5/4 - 3\arcsin(1/3)/(2\pi)]$. The additional $K = 5, 6$ values in Table 1 use (109); multivariate-normal CDF evaluation (Virtanen et al., 2020) and independent one-dimensional quadrature (Kabluchko & Zaporozhets, 2020, Proposition 1.4(b)) check their rounding.

Reproduction commands, package versions, and saved-file descriptions are given in `code/README.md`.

K.2 Multiclass optimization

Population objective. Fix $(\boldsymbol{S}, \boldsymbol{b}_0)$ and the Gaussian vectors and label law of Theorem 2.3. For $\boldsymbol{H} \succeq \boldsymbol{0}$ with $\operatorname{tr} \boldsymbol{H} = 1$, define

$$t(\boldsymbol{H}) = \inf_{t \ge 0,\, \boldsymbol{M},\, \boldsymbol{a},\, \boldsymbol{u}} \left\{ t : \ \boldsymbol{M}\boldsymbol{z}_0 + \boldsymbol{H}\boldsymbol{z}_1 + \boldsymbol{a} + \boldsymbol{u} \in \mathcal{C}_Y \text{ a.s.}, \quad \mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top] \preceq t\boldsymbol{H}^2 \right\}, \tag{118}$$

where $\boldsymbol{M} \in \mathbb{R}^{q \times q}$, $\boldsymbol{a} \in \mathbb{R}^q$, and $\boldsymbol{u}$ is a square-integrable function of $(\boldsymbol{z}_0, \boldsymbol{z}_1, Y)$. Then

$$\alpha_\star(\boldsymbol{S}, \boldsymbol{b}_0) = \left[\inf_{\substack{\boldsymbol{H} \succeq \boldsymbol{0} \\ \operatorname{tr} \boldsymbol{H} = 1}} t(\boldsymbol{H}) \right]^{-1}. \tag{119}$$

For $\alpha_\star^0$, set $\boldsymbol{a} = \boldsymbol{0}$ and keep the same label law. The optimization at fixed $\boldsymbol{H}$ is convex; the outer search over $\boldsymbol{H}$ is generally nonconvex.

Gaussian integration. Choose N points $(\boldsymbol{z}_{0,j}, \boldsymbol{z}_{1,j})$ for the $2q$ standard Gaussian coordinates, with positive weights ω_j summing to one. Sum over labels exactly using $w_{jy} = \omega_j \mathbb{P}(Y = y \mid \boldsymbol{z}_{0,j})$. For corrections $\boldsymbol{u}_{jy} \in \mathbb{R}^q$, form $\boldsymbol{U} = [\sqrt{w_{jy}}\,\boldsymbol{u}_{jy}]_{j,y}$. A finite convex approximation minimizes $t \ge 0$ subject to

$$\boldsymbol{M}\boldsymbol{z}_{0,j} + \boldsymbol{H}\boldsymbol{z}_{1,j} + \boldsymbol{a} + \boldsymbol{u}_{jy} \in \mathcal{C}_y \quad \text{for every } (j,y), \qquad \begin{pmatrix} t\boldsymbol{H}^2 & \boldsymbol{U} \\ \boldsymbol{U}^\top & \boldsymbol{I}_{N(q+1)} \end{pmatrix} \succeq \boldsymbol{0}. \tag{120}$$

By the Schur complement, the block inequality is exactly $\sum_{j,y} w_{jy}\boldsymbol{u}_{jy}\boldsymbol{u}_{jy}^\top \preceq t\boldsymbol{H}^2$. This is a reference semidefinite formulation; implementations can use the small moment matrix without storing the identity block densely. The integration size N is distinct from the simulation sample size n. Points can be reused when changing the label weights or fitting convention.

Matrix search and singular factors. For three classes, the full trace-one domain is the disk

$$\boldsymbol{H}(s_1, s_2) = \frac{1}{2}\begin{pmatrix} 1+s_1 & s_2 \\ s_2 & 1-s_1 \end{pmatrix}, \qquad s_1^2 + s_2^2 \le 1. \tag{121}$$

The interior has rank two and the boundary circle has rank one. Both must be considered; requiring a positive minimum eigenvalue of $\boldsymbol{H}$ changes the optimization domain.

For a singular factor, write $\boldsymbol{H} = \boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^\top$ with $\boldsymbol{R} \in \mathbb{R}^{q\times r}$ orthonormal and $\boldsymbol{D}$ positive diagonal. The moment constraint forces $\boldsymbol{u} = \boldsymbol{R}\boldsymbol{w}$ and reduces to $\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top] \preceq t\boldsymbol{D}^2$. The range reduction does not restrict $\boldsymbol{M}$ or $\boldsymbol{a}$, but feasibility must hold beyond the integration points. Write $\mathcal{C}_y = \{\boldsymbol{v} : \boldsymbol{A}_y\boldsymbol{v} \ge \boldsymbol{0}\}$, where $\boldsymbol{A}_y$ has rows $(\boldsymbol{e}_y - \boldsymbol{e}_j)^\top$, $j \ne y$, with $\boldsymbol{e}_0 = \boldsymbol{0}$. For each class, impose

$$\boldsymbol{g}^\top\boldsymbol{M} = \boldsymbol{0}^\top, \qquad \boldsymbol{g}^\top\boldsymbol{a} \ge 0, \quad \boldsymbol{g} = \boldsymbol{A}_y^\top\boldsymbol{\lambda}, \quad \boldsymbol{\lambda} \ge \boldsymbol{0}, \quad (\boldsymbol{A}_y\boldsymbol{R})^\top\boldsymbol{\lambda} = \boldsymbol{0}. \tag{122}$$

It suffices to check the extreme rays of this polyhedral cone. By Farkas' lemma and the full support of $\boldsymbol{z}_0$, these conditions ensure that the affine predictor space meets $\mathcal{C}_y$ at every Gaussian input. They prevent spurious feasibility caused by a finite integration rule.

Scope of the numerical checks. There are three error sources: the fixed-$\boldsymbol{H}$ solve, the outer matrix search, and Gaussian integration. A primal–dual gap at fixed $\boldsymbol{H}$ controls only the first. We check the others by changing starting matrices and ranks, refining integration, and evaluating correction policies on fresh points. The curves below remain numerical approximations: these checks do not certify global optimality or population error. A certified population interval would require bounds on both the outer optimization and Gaussian integration.

K.3 Signal-strength curves in the multiclass heatmaps

For each of the 21 signal strengths in Figure 1, we sum the K label contributions exactly and use scrambled Sobol points for Gaussian integration. The common settings for both sets of multiclass curves are:

- Compare 2^{14} and 2^{16} points with independent seeds 71 and 193; report the average of the two finer evaluations.
- Evaluate the resulting correction policies on 2^{18} fresh points with seed 397; use 0.005 as the comparison tolerance.
- Check multiple starting matrices and lower-rank factors. Retain a metric eigenvalue floor of 10^{-5} and bounded log-Cholesky parameters, and record unsuccessful exploratory solves separately.

Zero-signal values use the explicit null constants. Gaussian points are reused across strengths, and neighboring solutions initialize the matrix searches. Weighted cone projections enumerate the faces and test membership, with metric-dependent operators cached. The plotted line connects the computed values.

For $K = 4$, ranks one, two, and three are explored; the strongest retained candidates have full rank. The maximum changes in the boundary are 0.002905 under integration refinement, 0.000379 across seeds, and 0.001971 on fresh points, all below 0.005.

The drivers and diagnostics for both multiclass figures are archived in `code/reproduction/boundary_methods.tar.gz`; extraction and execution instructions are in `code/reproduction/README.md`. The saved manifests identify inputs, source and data hashes, and numerical checks. Before plotting, `code/heatmap.py plot` checks the exported curve parameters against the simulation plans.

K.4 Nonzero-intercept curves

For Figure 2, use $K = 3, 4$, $\boldsymbol{S} = 0.1\boldsymbol{I}_{K-1} + 0.9\boldsymbol{1}_{K-1}\boldsymbol{1}_{K-1}^\top$, and $\boldsymbol{b}_0 = (b, 0, \ldots, 0)^\top$ for $b = 0, 0.1, \ldots, 1.5$. Only the conditional label weights change; cone projections and the fitted-intercept optimization remain the same. With the integration settings in Appendix K.3, the largest change across refinement, seed, and fresh-point checks over all 32 boundary points is less than 0.00218.

Multiple full-rank starts agree within 3.74×10^{-6}. Lower-rank candidates, reevaluated with finer integration, give boundaries below the full-rank candidate. At $b = 0$, both class counts agree with the independently computed $\gamma = 1$ boundaries in Figure 1. Search settings and numerical checks are retained with the boundary code.

L ADDITIONAL SIMULATION RESULTS

The experiments below check explicit threshold formulas, compare the two intercept-fitting choices as sample size increases, and illustrate why the matrix constraint is needed. Each simulation condition uses 100 independent repetitions. These finite-sample comparisons support the predictions of the theory; the experimental design and numerical checks follow the results.

L.1 VALIDATION AGAINST EXPLICIT THRESHOLDS

The zero-signal experiments check the explicit thresholds in Corollary I.3. For $K = 2, 3, 4$, we use $n = 2000$, zero true coefficients, and zero true intercept, so labels are uniform and independent of the Gaussian features. Both fitting choices have limiting thresholds 2, 2.4, and 2.758012, respectively. In Table 2, midpoint median crossings differ from these thresholds by at most 0.019. The four-class observations come from a separate campaign with a coarser dimension grid, so their close midpoint agreement should be read alongside the wider brackets. For binary labels, the exact finite-sample probabilities provide an additional reference: writing $F_{n-1}(m) = \mathbb{P}\{\mathrm{Bin}(n-1, 1/2) \le m\}$, Cover (1965) gives finite-MLE probabilities $1 - F_{n-1}(d-1)$ without a fitted intercept and $1 - F_{n-1}(d)$ with one. This finite-sample difference is compatible with their common limiting threshold 2.

The binary nonzero-signal experiments check the recovered formula in Section 3.1. They fix $S = 1$, fit an intercept, and vary $b_0 \in \{0, 0.75, 1.5\}$ at $n = 2000$. The independently computed boundaries are 2.278221, 2.421829, and 2.896391; the midpoint median crossings are 2.277, 2.420, and 2.867. These benchmarks validate the computation in a known special case, complementing the multiclass heatmaps in Section 4.

For each repetition, the *crossing* is the value of $\alpha = n/d$ where the sample changes between separation and finite-MLE existence. Across the 100 repetitions, the median crossing locates the transition, while the *transition width*, $Q_{0.9} - Q_{0.1}$, measures its spread; Q_p denotes the empirical p-quantile of the crossings. A smaller width means a sharper transition. Since we test a discrete dimension grid, each crossing is known only within an interval. The table therefore reports lower and upper bounds on both summaries, rather than single values. These bounds reflect grid resolution, not sampling confidence: for example, the first row places the median crossing in $[2.004, 2.007]$ and the transition width in $[0.107, 0.112]$. Appendix L.4 gives the precise calculation.

Table 2: Validation against explicit thresholds at $n = 2000$, with 100 repetitions per row. The boundary column gives the null or binary-formula reference. The median crossing locates the transition in $\alpha = n/d$; the transition width is the difference between the 90th and 10th percentiles of crossings across repetitions. Each reported interval bounds the corresponding summary on the tested dimension grid; it is not a sampling confidence interval.

K	$b_{0,1}$	Intercept	$\alpha_\star$	Median crossing	Transition width
Zero signal ($\gamma = 0$)					
2	0	Zero	2.000	$[2.004, 2.007]$	$[0.107, 0.112]$
2	0	Estimated	2.000	$[2.008, 2.011]$	$[0.109, 0.114]$
3	0	Zero	2.400	$[2.389, 2.393]$	$[0.115, 0.121]$
3	0	Estimated	2.400	$[2.395, 2.399]$	$[0.112, 0.119]$
4	0	Zero	2.758	$[2.751, 2.802]$	$[0.102, 0.200]$
4	0	Estimated	2.758	$[2.751, 2.802]$	$[0.102, 0.200]$
Binary signal ($\gamma = 1$)					
2	0	Estimated	2.278	$[2.275, 2.278]$	$[0.171, 0.177]$
2	0.75	Estimated	2.422	$[2.418, 2.422]$	$[0.179, 0.186]$
2	1.5	Estimated	2.896	$[2.865, 2.870]$	$[0.330, 0.339]$

Table 3: Rank-one experiment: theoretical boundaries and bounds on the empirical median crossing and transition width in $\alpha = n/d$. The width is the 90th minus the 10th percentile of crossings; smaller values indicate a sharper transition. Reported intervals retain dimension-grid uncertainty and two censored tail crossings at $n = 500$; they are not sampling confidence intervals. All repetitions are retained.

n	Repetitions	Fitted intercept	$\alpha_\star$	Median crossing	Transition width
500	100	No	9.493	$[10.000, 10.205]$	$[4.382, 4.808]$
1000	100	No	9.493	$[9.433, 9.524]$	$[2.777, 2.972]$
2000	100	No	9.493	$[9.478, 9.524]$	$[1.924, 2.018]$
3000	100	No	9.493	$[9.584, 9.616]$	$[1.644, 1.707]$
4000	100	No	9.493	$[9.592, 9.616]$	$[1.598, 1.645]$
5000	100	No	9.493	$[9.487, 9.506]$	$[1.412, 1.450]$
6000	100	No	9.493	$[9.463, 9.479]$	$[1.204, 1.235]$
7000	100	No	9.493	$[9.395, 9.409]$	$[1.116, 1.143]$
8000	100	No	9.493	$[9.456, 9.468]$	$[0.996, 1.020]$
500	100	Yes	10.128	$[10.638, 10.870]$	$[4.961, 5.449]$
1000	100	Yes	10.128	$[10.101, 10.205]$	$[3.496, 3.729]$
2000	100	Yes	10.128	$[10.050, 10.102]$	$[2.296, 2.404]$
3000	100	Yes	10.128	$[10.101, 10.136]$	$[1.595, 1.666]$
4000	100	Yes	10.128	$[10.230, 10.257]$	$[1.557, 1.613]$
5000	100	Yes	10.128	$[10.080, 10.102]$	$[1.623, 1.667]$
6000	100	Yes	10.128	$[10.118, 10.136]$	$[1.361, 1.397]$
7000	100	Yes	10.128	$[10.028, 10.044]$	$[1.192, 1.222]$
8000	100	Yes	10.128	$[10.075, 10.089]$	$[1.197, 1.223]$

L.2 Intercept effects and transition sharpening

Table 3 supplements the rank-one comparison in Section 4.3 with median-crossing and transition-width bounds. Each sample size uses 100 pairs; all crossings at $n \geq 1000$ are resolved and uncensored.

The separate dimension-grid bounds on the mean increase are $[0.630, 0.659]$ at $n = 7000$ and $[0.642, 0.667]$ at $n = 8000$, with endpoints rounded outward. These bounds average paired endpoint differences; they are not sampling confidence intervals. The bootstrap resamples whole pairs, as described in Appendix L.4.

Table 3 shows sharper transitions at larger sample sizes. Between $n = 1000$ and 8000, the midpoint 10–90% width decreases from approximately 2.87 to 1.01 without a fitted intercept, and from 3.61 to 1.21 with one. The intermediate widths need not decrease monotonically. This supports concentration near a threshold without establishing a convergence rate. The earlier $n = 500$ results retain two censored tail crossings; their unbounded endpoints remain in the table's quantile bounds.

L.3 Data generation and dimension search

For a prescribed $\boldsymbol{S}$ and $\boldsymbol{b}_0$, place a square root of $\boldsymbol{S}$ in the first $q = K - 1$ rows of $\boldsymbol{\Theta}_0$, with all remaining rows zero. Draw a Gaussian design with n rows and the largest required number of columns. Conditional-softmax labels use these signal coordinates and independent uniform random variables. In the rank-one model, both nonbaseline predictors are $10x_1$. All dimension prefixes therefore have the prescribed sampling law and the same true predictors. Repetitions and sample sizes use independent random streams. Within a repetition, signal or intercept conditions share the design and uniforms; the two fitting choices also share labels. These paired conditions are not independent replications.

Adding feature columns can preserve or create a separator. We exploit this monotonicity to binary-search the largest dimension with a finite MLE and the smallest dimension without one. A separator for the fit without an intercept is also valid when an intercept is fitted; a certificate excluding separators with an intercept also excludes those without one. These implications reuse checked decisions across fits. All grids have $d + 1 < n$, ensuring full column rank almost surely. The implementation may

choose a different baseline when forming margins; this invertible reparameterization preserves the separation event and both fitting spaces.

The heatmap dimensions are the distinct integers nearest $2000/\alpha$ for $\alpha = 2, 2.1, \ldots, 4$, with strengths $\gamma = 0, 0.1, \ldots, 2$ in Section 4.1. Section 4.2 instead fixes $\gamma = 1$ and varies $b_{0,1} = 0, 0.1, \ldots, 1.5$. The plots use the actual ratios $2000/d$. Except for the reused four-class zero-signal data, supplemental experiments search every integer dimension from $\lceil n/\alpha_{\max} \rceil$ to $\lfloor n/\alpha_{\min} \rfloor$. The ranges $[\alpha_{\min}, \alpha_{\max}]$ are $[6, 16]$ for the rank-one experiment, $[1.7, 4]$ for binary nonzero signals, $[1.7, 3]$ for the binary zero-signal experiment, $[1.8, 3.6]$ for its three-class counterpart. The added rank-one runs at $n = 3000, 4000, \ldots, 8000$ search every integer dimension from $\lfloor n/20 \rfloor$ through $\lceil n/4 \rceil$; the earlier runs retain their original grids. Rounding dimensions makes the actual endpoints differ slightly from these nominal ratios.

The plan seeds are listed in `code/README.md`; the new rank-one runs use seed 2026092104 at $n = 4000$; the original $n = 2000$ cohort uses seed 2026092002. The extension uses $2026092200+n/1000$ at $n = 3000, 5000, 6000, 7000, 8000$. Each frozen plan stores its derived repetition seeds and separate streams for signal features, residual features, and label uniforms.

L.4 Sampling, grid, and numerical uncertainty

For a trial with a largest nonseparating dimension d_- and a smallest separating dimension d_+, the crossing is bracketed by $[n/d_+, n/d_-]$. The heatmaps use their coarser dimension grid; resolved, uncensored supplemental crossings have $d_+ = d_- + 1$, except for the coarser four-class zero-signal grid. When an endpoint is outside the grid, we retain an unbounded endpoint rather than assign a finite crossing. In the rank-one experiment at $n = 500$, one omitted-intercept crossing lies below 6.024, and one fitted-intercept crossing lies above 15.625.

Let L_r and U_r denote the lower and upper endpoints in repetition r, using 0 or $+\infty$ for a missing lower or upper endpoint. For Q_p, the inverse empirical-CDF quantile over all repetitions in a group, the empirical median lies in $[Q_{0.5}(L), Q_{0.5}(U)]$, and the 10–90% width lies in

$$[\max\{0, Q_{0.9}(L) - Q_{0.1}(U)\},\ Q_{0.9}(U) - Q_{0.1}(L)]\,.$$

The tables report these bounds, including censored trials. They measure dimension-grid uncertainty, not Monte Carlo sampling error. Midpoint summaries are used only for groups whose crossings are all resolved and uncensored. In particular, a degenerate bootstrap interval caused by several median estimates falling in one grid cell does not imply exact threshold estimation.

Probability curves have pointwise 95% Wilson intervals. For the mean threshold shift, we resample whole pairs 10,000 times with seed 20260922 and use linearly interpolated bootstrap percentiles. Each pair contributes the fitted-intercept midpoint crossing minus the no-intercept midpoint crossing. These confidence intervals are approximate and pointwise. The separate grid bounds average the paired lower and upper endpoint differences.

All displayed heatmap cells and the $n = 8000$ probability curves use 100 classified outcomes, with no unknown outcomes. Sampling uncertainty remains: with 100 repetitions, the Monte Carlo standard error is 0.05 at probability $1/2$. No unsuccessful search is dropped from a denominator or classified using solver termination alone.

L.5 Separation certificates and efficient computation

Let $p = d$ and $\bar{\boldsymbol{x}}_i = \boldsymbol{x}_i$ without a fitted intercept, and let $p = d + 1$ and $\bar{\boldsymbol{x}}_i = (\boldsymbol{x}_i^\top, 1)^\top$ otherwise. Ratios remain n/d. For $\boldsymbol{B} \in \mathbb{R}^{p \times q}$, define the predictor changes $\Delta v_{i,0} = 0$ and $\Delta v_{i,j} = (\boldsymbol{B}^\top \bar{\boldsymbol{x}}_i)_j$ for $j = 1, \ldots, q$. Writing $\boldsymbol{\vartheta} = \mathrm{vec}(\boldsymbol{B})$, form the margin matrix $\boldsymbol{A} \in \mathbb{R}^{nq \times pq}$ so that

$$(\boldsymbol{A}\boldsymbol{\vartheta})_{i,j} = \Delta v_{i,y} - \Delta v_{i,j}, \qquad j \neq y,$$

where y is the observed class and j ranges over all competing classes, including class 0 when $y \neq 0$. The bounded separation problem is

$$\max_{\boldsymbol{\vartheta}}\ \mathbf{1}_{nq}^\top \boldsymbol{A}\boldsymbol{\vartheta} \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{\vartheta} \geq \mathbf{0}, \qquad \|\boldsymbol{\vartheta}\|_\infty \leq 1. \tag{123}$$

Full column rank implies that a nonzero separating direction has at least one strict comparison. Homogeneity therefore makes the optimum positive exactly when a separating direction exists. This

criterion includes tied comparisons, rather than requiring strict separation of every observed class from every competitor.

The CPU implementation uses screening followed, when necessary, by HiGHS linear programs (Huangfu & Hall, 2018) through SciPy. A query is classified only when its floating-point certificate passes the checks below, regardless of solver termination status. For the row-normalized matrix $\widetilde{\boldsymbol{A}}$ with m nonzero rows, the feasibility formulations seek

$$\widetilde{\boldsymbol{A}}\boldsymbol{\vartheta} \geq \mathbf{0}, \qquad \frac{(\widetilde{\boldsymbol{A}}^\top \mathbf{1}_m)^\top \boldsymbol{\vartheta}}{\|\widetilde{\boldsymbol{A}}^\top \mathbf{1}_m\|} = 1, \quad \text{or} \quad \widetilde{\boldsymbol{A}}^\top \boldsymbol{\lambda} = \mathbf{0}, \qquad \boldsymbol{\lambda} \geq \mathbf{1}_m.$$

The latter is a positive polar certificate excluding all nonzero separating directions (Lemma A.2).

HiGHS uses primal and dual feasibility tolerances 10^{-9} and interior-point optimality tolerance 10^{-10}. Acceptance checks require primal minimum margin at least -5×10^{-8} and normalization error at most 5×10^{-8}, or polar minimum weight at least $1 - 5 \times 10^{-8}$ and

$$\|\widetilde{\boldsymbol{A}}^\top \boldsymbol{\lambda}\|_\infty \leq \max\{5 \times 10^{-8},\, 5 \times 10^{-11} \|\boldsymbol{\lambda}\|_1\}. \tag{124}$$

These are floating-point numerical checks. Saved diagnostics record the tolerances and accepted residuals. The original runs retain these diagnostics rather than the coefficient and multiplier arrays. For the additional $n = 8000$ certificate computations, witness arrays are retained and independently checked against the original design matrices. The result audit reconstructs brackets and probabilities from accepted decisions; these checks do not constitute exact-arithmetic verification.

Solver provenance and computational settings are documented with the code.

For difficult queries, a bounded primal LP with coefficients in $[-1, 1]$ supplies candidate separating directions or polar multipliers from its dual solution. These candidates are accepted only after the same residual checks against the original matrix. All rank-one crossings at $n \geq 1000$ are resolved between adjacent dimensions; two tail crossings at $n = 500$ have unbounded endpoints.