

Softmax Attention on Gaussian Mixtures

Linear When It Can, Selective When It Must

Simon Gabet¹ Etienne Boursier¹ Claire Boyer^{1,2}

¹Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405 Orsay, France

²Institut Universitaire de France

Correspondence: simon.gabet@universite-paris-saclay.fr

ABSTRACT

Softmax attention, at the heart of Transformers, has demonstrated remarkable capabilities. Yet its underlying mechanisms remain only partially understood. Recent theoretical work studies Gaussian prompts, where the infinite-prompt limit reduces softmax attention to a linear map, but also removes the query-dependent selection that distinguishes it from linear attention. This work studies the infinite-prompt limit of softmax attention on Gaussian mixtures, which retain the tractability of Gaussian data while introducing latent structure, multimodality, and nonlinear dependencies. We show that softmax attention can represent and learn, via gradient-based methods, optimal solutions to a range of statistical tasks, including supervised classification and denoising. Our results highlight two complementary capabilities of softmax attention: it can recover linear tasks as effectively as its simpler linear counterpart, while also exploiting query-dependent context selection to solve nonlinear tasks beyond the reach of linear attention.

Keywords softmax attention · Gaussian mixtures · large-prompt limit · optimization landscape · denoising · in-context learning

1 Introduction

Softmax attention lies at the core of modern Transformer models. Yet, despite its remarkable empirical success, its theoretical behavior remains poorly understood, even in highly simplified settings. Theoretical accounts of attention often replace softmax by a linear kernel or study distributions for which the population softmax operator itself becomes linear. These reductions lead to precise learning-dynamics results, but they remove the very mechanism that makes attention selective: which part of the prompt is relevant may depend on the query.

We study attention in the large-prompt regime. In this work, a prompt of length L is a sequence of L tokens drawn independently from a distribution μ on $\mathbb{R}^d$. Softmax attention over these tokens depends on them only through their empirical measure, and converges, as $L \rightarrow \infty$, to an operator acting directly on μ (Section 2). In this large-prompt limit, an attention layer becomes a map that takes a probability measure, the prompt, and returns a function of the query. This viewpoint removes finite-sample fluctuations and exposes the geometry of the operator itself; concentration results relate it to its finite-prompt counterpart for sub-Gaussian tokens (Vuckovic et al., 2020; Boursier and Boyer, 2026; Bohbot et al., 2026).

The choice of μ then determines what can be observed. When μ is Gaussian (Castin et al., 2025; Boursier and Boyer, 2026), the limiting operator is an affine function of the query: the analysis is tractable, but selection has disappeared. Gaussian mixtures are the natural next step. They retain closed-form computations while introducing latent structure, multimodality and, crucially, a reason for a query to attend to one part of

the context rather than another. They thus offer a setting in which one can ask precisely what the softmax nonlinearity brings.

Contributions. We answer this question through one structural identity and two statistical tasks that play complementary roles.

- **Softmax attention as a gated mixture of experts.** On any Gaussian mixture, a softmax head acts as a query-dependent combination of affine experts; with shared covariances, it splits exactly into a global linear map and a softmax gate that selects, for each query, the relevant components (Proposition 2.1). Linear attention, by contrast, weighs the components by fixed proportions, blind to the query. This identity is the lens of the paper: it tells us when the gate is superfluous, and when it does the essential work.
- **When the gate is redundant: learning LDA.** In linear discriminant analysis (LDA), the optimal classifier is linear, and a single head can represent it exactly. The question is then one of optimization: does a redundant, nonconvex parameterization remain trainable? In effective (reparameterized) coordinates, the logistic risk has no spurious critical points (Theorem 3.1); in the original attention parameters, every nonoptimal critical point is a strict saddle as soon as the class means are not symmetric (Theorem 3.3). The landscape is benign but not coercive: near-optimal directions escape to infinity (Proposition 3.4). We therefore turn to the dynamics, and prove exponential convergence of gradient flow near the regular manifold of optimal parameters (Theorem 3.5) and, in the symmetric balanced case, global convergence on a Krylov-structured invariant subspace (Theorem 3.6).
- **When the gate is essential: denoising.** Optimal denoising of a Gaussian mixture requires two operations at once: shrinking the observation, and deciding which cluster it comes from. The optimal denoiser is therefore itself a gated affine map. Two softmax heads reproduce it exactly; a single head generally cannot, and no linear-attention predictor, whatever its number of heads, comes close to the optimal error (Theorem 4.1). Iterating the same map changes its meaning: below a sharp separation threshold, it collapses every input to the midpoint between the component centers; above it, it clusters (Proposition 4.3). Finally, with residual connections, a single set of parameters denoises optimally in context, for any cluster locations and proportions (Proposition 4.4): the weights store what is shared across tasks, while attention infers from the prompt what is specific to each one.

Related work. The measure formulation of attention was developed by Vuckovic et al. (2020) and used to study deep attention dynamics and clustering (Sander et al., 2022; Geshkovski et al., 2023; Castin et al., 2025). Recent work analyzes training in the large-prompt Gaussian regime (Boursier and Boyer, 2026; Maulen-Soto and Boyer, 2026; Goel et al., 2026); our mixture identity is the first step beyond the linear operator resulting from Gaussian prompt. Transformer classification has been studied through linear attention and algorithm emulation (Frei and Vardi, 2025; Shen et al., 2025; Zhang and Cao, 2026). Our result instead learns the Bayes boundary directly with a single softmax layer and exposes the geometry of its population risk in the attention parameters. For denoising, Smart et al. (2025) connect one-layer attention to associative-memory retrieval when the context is clean; Rosu et al. (2025) and Li et al. (2026) study attention-based denoising for diffusion. We treat noisy contexts and nondegenerate mixtures with standard dot-product softmax, give exact Bayes constructions, and prove a strict separation from linear attention. Concurrent work studies a more general two-stage empirical-Bayes viewpoint using RBF attention (Smart et al., 2026). A complementary statistical separation between softmax and linear attention appears in Duranthon et al. (2026).

2 Population Attention on Mixtures

Attention as an operator on measures. We represent a prompt by a probability measure μ on $\mathbb{R}^d$, absorb the query and key matrices into a single interaction matrix $U \in \mathbb{R}^{d \times d}$, and denote by $V \in \mathbb{R}^{d \times d}$ the value matrix. For a query $z \in \mathbb{R}^d$, the attention operator reads

$$T_{U,V}[\mu](z) = \frac{\int e^{(z')^\top U z} V z' d\mu(z')}{\int e^{(z')^\top U z} d\mu(z')}. \quad (1)$$

For a finite prompt $z_1, \dots, z_L$ with empirical measure $\hat{\mu}_L = L^{-1} \sum_{i=1}^L \delta_{z_i}$, the quantity $T_{U,V}[\hat{\mu}_L](z)$ is exactly softmax dot-product attention over the tokens. When the tokens are drawn i.i.d. from a sub-Gaussian measure μ , the law of large numbers, applied to the numerator and the denominator, gives $T_{U,V}[\hat{\mu}_L](z) \rightarrow T_{U,V}[\mu](z)$ almost surely for every query z ; quantitative versions are given in [Boursier and Boyer \(2026\)](#); [Bohbot et al. \(2026\)](#). All our results concern the limiting operator $T_{U,V}[\mu]$.

From a Gaussian to a mixture. If μ is Gaussian, centered at m and of covariance matrix Γ , the attention operator is

$$T_{U,V}[\mu](z) = V(m + \Gamma U z),$$

an affine function of the query, in which the softmax leaves no trace. On a mixture, the same computation applies component by component, and the components then compete through the normalization. Throughout, Gaussian covariances are positive definite unless a degenerate case is stated explicitly.

Proposition 2.1 (Softmax-gated affine experts). *Consider a mixture $\mu = \sum_{k=1}^M \pi_k \mathcal{N}(m_k, \Gamma_k)$ of M Gaussian components. The infinite-prompt attention layer parameterized by U and V gives for any $z \in \mathbb{R}^d$,*

$$T_{U,V}[\mu](z) = V \sum_{k=1}^M \alpha_k(z) (m_k + \Gamma_k U z), \quad (2)$$

with

$$\alpha_k(z) = \text{softmax}_k \left(\log \pi_k + m_k^\top U z + \frac{1}{2} (U z)^\top \Gamma_k U z \right).$$

If $\Gamma_k = \Gamma$ for all k , the common quadratic offsets in the softmax function cancels and

$$T_{U,V}[\mu](z) = V \Gamma U z + V \sum_k \alpha_k(z) m_k, \quad (3)$$

with $\alpha_k(z) = \text{softmax}_k(\log \pi_k + m_k^\top U z)$.

The proof of Proposition 2.1 is given in Appendix A.1. Equation (3) is the organizing principle of the paper: the first term performs global linear prediction; the second selects components and returns a query-dependent barycenter.

Attention with generalized scores. More generally, attention on any mixture $\sum_k \pi_k \mu_k$ is a query-dependent convex combination of the component-wise attention maps, with gates proportional to the alignment to the query (see the supplement). Indeed, for $T_a[\mu](z) := \int a(z, z') z' d\mu(z') / \int a(z, z') d\mu(z')$, we have

$$T_a[\mu](z) = \sum_{k=1}^M r_k(z) T_a[\mu_k](z)$$

with $r_k(z) := \pi_k \int_{\mathbb{R}^d} a(z, z') d\mu_k(z') / \int_{\mathbb{R}^d} a(z, z') d\mu(z')$ and $a > 0$ an arbitrary alignment function (e.g., exponential dot product for softmax scores), see Appendix A.2 for details. Normalized attention, such as softmax attention, can therefore perform *context selection*: in regions where $r_k(z)$ is close to one, the operator essentially selects the context μ_k ; allowing to adaptively select which cluster is relevant to a given query. This behavior contrasts with classical linear attention, defined by $T_{\text{lin}}[\mu](z) := \int_{\mathbb{R}^d} \langle U z, z' \rangle V z' d\mu(z')$. Indeed, $T_{\text{lin}}[\mu](z) = \sum_k \pi_k T_{\text{lin}}[\mu_k](z)$, so linear attention combines the different contexts with fixed mixture proportions π_k , independently of the query, and cannot separate the contribution of the different clusters.

3 Attention-based Linear Discriminant Analysis

We now turn to classification. Under linear discriminant analysis (LDA), the Bayes score is affine, and hence does not require the nonlinear selection mechanism of softmax attention. The question is instead whether a single softmax head can represent this score and whether gradient-based training can recover it despite this redundant and nonconvex parameterization.

3.1 From attention to an affine-sigmoid predictor

Let $Y \in \{-1, 1\}$ with $\mathbb{P}(Y = 1) = \pi_+$, and assume the classical LDA model, with input features $X \in \mathbb{R}^d$ such that

$$X \mid Y = c \sim \mathcal{N}(m_c, \Gamma), \quad c \in \{-1, 1\}, \quad (4)$$

where the two classes have different means and share a positive-definite covariance matrix. The Bayes classifier is $\text{sign}(f^*)$, where the Bayes score $f^*(x) = w^*{}^\top x + b^*$ is affine, with

$$\begin{aligned} w^* &= \Gamma^{-1}(m_1 - m_{-1}), \\ b^* &= -\frac{1}{2}(m_1^\top \Gamma^{-1} m_1 - m_{-1}^\top \Gamma^{-1} m_{-1}) + \log \frac{\pi_+}{1 - \pi_+}. \end{aligned} \quad (5)$$

We feed augmented context tokens $(X, Y) \in \mathbb{R}^{d+1}$ to one attention head and query it at $(x, 0)$; only the last output coordinate is therefore used to store the prediction $f_\theta(x)$. Writing the relevant blocks of U, V interacting with X or Y (see Appendix B.1) as $\theta = (U_{11}, u_{12}, v_{21}, v_{22})$, Proposition 2.1 yields for μ the Gaussian mixture described in (4):

$$\begin{aligned} T_{U,V}[\mu]((x, 0))_{d+1} &= f_\theta(x) \\ &:= w^\top x + b + r\sigma(s_0 + s^\top x), \end{aligned} \quad (6)$$

where $\sigma(t) = (1 + e^{-t})^{-1}$ and

$$\begin{aligned} w &= U_{11}\Gamma v_{21}, & b &= v_{21}^\top m_{-1} - v_{22}, \\ r &= v_{21}^\top (m_1 - m_{-1}) + 2v_{22}, \\ s &= 2u_{12} + U_{11}(m_1 - m_{-1}), \\ s_0 &= \log \frac{\pi_+}{1 - \pi_+}. \end{aligned} \quad (7)$$

The exact derivation is given in Appendix B.1. Thus a softmax head reduces, in effective coordinates, to an affine predictor plus a single sigmoidal correction. In particular, one head can realize the Bayes score exactly; an explicit construction is given in Appendix B.2. At the effective level, this happens through two distinct mechanisms: either $r = 0$ suppresses the nonlinear term, or $s = 0$ makes it constant and absorbable into the bias b . Their preimages in the original attention parameters need not have the same geometry—and, depending on the class means and priors, both mechanisms need not be simultaneously available. This distinction is central to the optimization analysis below.

3.2 A benign but noncoercive optimization landscape

We train the score with the population logistic risk¹

$$\begin{aligned} \mathcal{R}(\theta) &:= \mathbb{E} \left[\log(1 + e^{-Y f_\theta(X)}) \right] \\ &= \tilde{\mathcal{R}}(\Phi(\theta)), \end{aligned} \quad (8)$$

where $\Phi(\theta) = (w, b, r, s)$ is defined in (7), and

$$\tilde{\mathcal{R}}(w, b, r, s) := \mathbb{E} \left[\log(1 + e^{-Y g_{w,b,r,s}(X)}) \right],$$

with $g_{w,b,r,s}(x) := w^\top x + b + r\sigma(s_0 + s^\top x)$.

Thus $\mathcal{R}$ is the risk in the original attention parameters, whereas $\tilde{\mathcal{R}}$ is expressed in effective coordinates. The logistic loss is important here: its unique population minimizer over measurable scores is the Bayes score f^* . Since the attention model realizes this score, the unrestricted optimum is attained in both parameterizations; we denote its value by $\mathcal{R}^*$.

¹From now on, we only consider the relevant parameters θ , since the remaining blocks of U, V are neither updated during training nor affect the predicted value.

The effective landscape. The reduced model is not jointly convex in (w, b, r, s) . However, fixing s leaves a convex problem in (w, b, r) that still contains the Bayes predictor, through the choice $(w, b, r) = (w^*, b^*, 0)$. This simple observation explains why stationarity is enough to guarantee optimality.

Theorem 3.1 (Effective landscape). *All critical points of $\tilde{\mathcal{R}}$ are global minimizers, and*

$$\begin{aligned}\mathcal{E}^* &:= \arg \min \tilde{\mathcal{R}} = \text{Crit}(\tilde{\mathcal{R}}) = \mathcal{E}_1 \cup \mathcal{E}_2, \\ \mathcal{E}_1 &= \{(w^*, b^*, 0, s) : s \in \mathbb{R}^d\}, \\ \mathcal{E}_2 &= \{(w^*, b^* - r\sigma(s_0), r, 0) : r \in \mathbb{R}\}.\end{aligned}$$

These two affine spaces recover the representation mechanisms identified above: $\mathcal{E}_1$ removes the sigmoid, while $\mathcal{E}_2$ makes it constant and absorbs its contribution into the bias. They represent the same predictor but leave different effective parameters unconstrained.

Returning to attention parameters. Does this favorable landscape survive the return to the attention parameters? The chain rule

$$\nabla \mathcal{R}(\theta) = D\Phi(\theta)^* \nabla \tilde{\mathcal{R}}(\Phi(\theta))$$

shows the possible obstruction: a singular differential can annihilate a nonzero effective gradient. The following result identifies exactly where this obstruction is avoided.

Lemma 3.2 (Regularity of the reparameterization). *For any $\theta \in \mathbb{R}^{d \times d} \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$, the adjoint $D\Phi(\theta)^*$ is injective if and only if $m_1 \neq -m_{-1}$ and $v_{21} \neq \mathbf{0}$.*

Consequently, under asymmetric means, every critical point with $v_{21} \neq \mathbf{0}$ is globally optimal. The exceptional set $\{v_{21} = \mathbf{0}\}$ requires a separate curvature argument: its nonoptimal critical points admit a negative-curvature direction. Together, these facts yield the following characterization.

Theorem 3.3 (Landscape in attention parameters). *The minimizer set of $\mathcal{R}$ is $\Theta^* = \Theta_1 \cup \Theta_2$, where*

$$\Theta_1 = \Phi^{-1}(\mathcal{E}_1), \quad \Theta_2 = \Phi^{-1}(\mathcal{E}_2).$$

If $m_1 \neq -m_{-1}$, every critical point of $\mathcal{R}$ is either a global minimizer or a strict saddle. In particular, all local minima are global.

Why this does not establish global convergence. The preceding theorem controls finite critical points, but not parameter escape. In fact, the effective risk admits diverging minimizing sequences, as shown by the following construction.

Proposition 3.4 (Minimizing sequences at infinity). *For every $c \in \mathbb{R}^d$, the sequence*

$$\zeta_R = (w^* - \sigma'(s_0)c, b^* - \sigma(s_0)R, R, c/R)$$

satisfies $g_{\zeta_R} \rightarrow f^*$ in $L^1(\mu_X)$ and $\tilde{\mathcal{R}}(\zeta_R) \rightarrow \mathcal{R}^*$ as $R \rightarrow \infty$, where $g_{w,b,r,s}(x)$ is defined in (8) and μ_X is the query distribution.

As the sigmoid amplitude grows and its direction shrinks, its constant and linear contributions are canceled by the affine part. The predictor approaches Bayes optimality while the parameters diverge. This does not prove divergence of gradient flow, but shows why the landscape alone cannot supply a boundedness argument. Proofs of the landscape and noncompactness results are given in Appendix B.5.

3.3 Gradient-flow convergence

We now study the gradient flow $\dot{\theta}_t = -\nabla \mathcal{R}(\theta_t)$. The two Bayes families Θ_1 and Θ_2 have different local geometries, so their common optimality does not imply the same convergence behavior.

Local convergence near the regular component. Assume first that $m_1 \neq -m_{-1}$, and define $\Theta_1^{\text{reg}} = \{\theta \in \Theta_1 : s(\theta) \neq \mathbf{0}\}$. At every $\theta^* \in \Theta_1^{\text{reg}}$, the Hessian satisfies

$$\ker \nabla^2 \mathcal{R}(\theta^*) = T_{\theta^*} \Theta_1^{\text{reg}}.$$

Thus the only flat directions are tangent to the minimizer manifold; the risk has strictly positive curvature in the normal directions. This Morse–Bott structure provides the local stability needed for the following convergence result.

Theorem 3.5 (Local convergence to the regular Bayes manifold). *Assume $m_1 \neq -m_{-1}$ and let $\theta^* \in \Theta_1$ satisfy $s(\theta^*) \neq \mathbf{0}$. If initialized sufficiently close to θ^* , gradient flow exists for all $t \geq 0$, remains in a neighborhood of Θ_1 , and converges exponentially to a point $\theta_\infty \in \Theta_1$. In particular, $f_{\theta_\infty} = f^*$ almost surely.*

The proof (see Appendix B.7) controls both the decay normal to the manifold and the displacement along it, giving convergence to an optimal parameter.

Why the constant-gate component is different. For every θ^* in Θ_2 with $r(\theta^*) \neq 0$ perturbing the gate direction can be compensated to first order by changing the affine coefficient. Under the same asymmetry assumption, the resulting Hessian kernel has d extra dimensions:

$$\dim \ker \nabla^2 \mathcal{R}(\theta^*) - \dim T_{\theta^*} \Theta_2 = d.$$

These are flat directions that are transverse to the local minimizer manifold. Thus the preceding Morse–Bott argument does not extend to Θ_2 . The calculation is given in Appendix B.8. This degeneracy alone determines neither convergence nor its rate. The slower decay in Figure 1 is an empirical observation consistent with the previous discussion (see Appendix B.13 for a description of the numerical protocol).

Global convergence on a symmetric invariant subspace. The centrally symmetric case $m_1 = -m_{-1}$ falls outside the regularity argument for Φ . When the classes are also balanced, symmetry instead allows us to construct an invariant subspace on which the predictor remains linear throughout training. Assume $m_1 = -m_{-1} =: m$ and $\pi_+ = 1/2$, and define the Krylov space

$$K = \text{span}\{m, \Gamma m, \dots, \Gamma^{d-1} m\}.$$

With P_K the orthogonal projector onto K , let

$$\mathcal{I}_K = \{u_{12} = \mathbf{0}, v_{22} = 0, v_{21} \in K^\perp, U_{11} P_K = \mathbf{0}\}.$$

Orthogonality of v_{21} to m alone is not generally preserved by the flow, which repeatedly applies Γ . The Krylov space ensures that $K^\perp$ is Γ -invariant, so these constraints close under the dynamics.

Theorem 3.6 (Global convergence on the invariant subspace). *The subspace $\mathcal{I}_K$ is invariant under gradient flow. If $U_{11}(0) = \mathbf{0}$, $u_{12}(0) = \mathbf{0}$, $v_{22}(0) = 0$ and $v_{21}(0) \in K^\perp \setminus \{\mathbf{0}\}$, then θ_t converges exponentially to a limit $\theta_\infty \in \mathcal{I}_K$ satisfying $f_{\theta_\infty} = f^*$.*

On $\mathcal{I}_K$, the bias and sigmoid term vanish, leaving $f_{\theta_t}(x) = \langle U_{11}(t) \Gamma v_{21}(t), x \rangle$. The proof combines coercivity of the linear logistic risk with the conserved balance

$$\|U_{11}(t)\|_F^2 - \|v_{21}(t)\|^2 = -\|v_{21}(0)\|^2.$$

The former bounds the effective predictor; the latter prevents its induced gradient dynamics from degenerating. Further estimates control the factors and yield their exponential convergence; see Appendix B.12.

From the initialization point of view, setting $U_{11}(0) = \mathbf{0}$, $u_{12}(0) = \mathbf{0}$, and $v_{22}(0) = 0$ does not raise any difficulty. The only non-trivial condition is to choose $v_{21}(0) \in K^\perp \setminus \{\mathbf{0}\}$, requiring the knowledge of Γ and m . However, from a practical perspective, this assumption is mild in high-dimensional settings when $\dim K \ll d$, since an isotropic random initialization typically has only a fraction of order $\dim(K)/d$ of its mass in K . In particular, when $\Gamma = \gamma^2 \mathbf{I}_d$, one has $\dim K = 1$ and $d(v_{21}(0)/\|v_{21}(0)\|, K^\perp) = O_{\mathbb{P}}(\frac{1}{\sqrt{d}})$. Our experiments in dimension $d = 20$ show rapid convergence to near-Bayes risk without knowledge of m nor Γ (Appendix B.13, Figure 5).

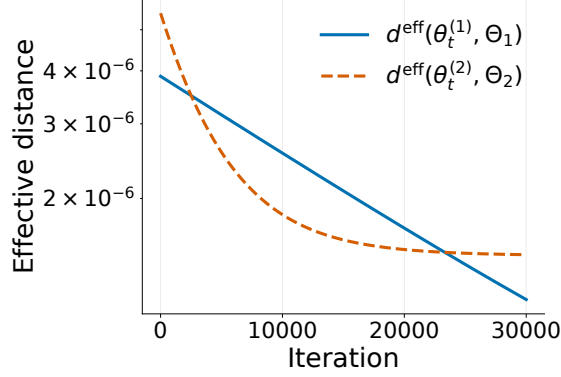


Figure 1: Population gradient descent near the two Bayes manifolds (see table 3 for initialization details). The effective distance (euclidean distance between the effective parameters and $\mathcal{E}_1$ or $\mathcal{E}_2$) to the regular component Θ_1 is linear on a log scale, while convergence toward the flatter component Θ_2 is markedly slower, matching the Morse–Bott characterization.

4 Denoising: Selection Becomes Essential

Classification used only the affine part of the mixture-attention decomposition. Denoising reveals the role of its nonlinear part. Denoising a sample from a Gaussian mixture requires both shrinking the observation within each component and inferring which component generated it. The latter is a genuinely query-dependent selection problem, and it is precisely where softmax attention differs from linear attention.

4.1 The Bayes denoiser as a gated affine map

Let $X \sim \sum_{k=1}^M \pi_k \mathcal{N}(m_k, \Gamma)$, consider a noisy version of it

$$Z = X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \xi^2 I_d), \quad (9)$$

where the noise ε is independent of X , and set $S = \Gamma + \xi^2 I_d$. Under squared loss $\mathcal{R}_{\text{sq}}(f) := \mathbb{E}[\|X - f(Z)\|^2]$, with a slight abuse, f^* now denotes the optimal reconstruction of X from Z (see Appendix C.1):

$$f^*(z) = \mathbb{E}[X \mid Z = z] = \Gamma S^{-1} z + \xi^2 S^{-1} \sum_k \beta_k(z) m_k, \quad (10)$$

where

$$\beta_k(z) = \text{softmax}_k(\log \pi_k + m_k^\top S^{-1} z - \tfrac{1}{2} m_k^\top S^{-1} m_k).$$

Equivalently, if C denotes the latent component label, then $\beta_k(z) = \mathbb{P}(C = k \mid Z = z)$. Thus the first term in (10) performs within-component Gaussian shrinkage, while the second returns the posterior barycenter of the component means. In the high-noise regime the latter approaches the prior mean $\sum_k \pi_k m_k$; in the low-noise regime the full denoiser approaches the identity.

Denoting by μ_Z the distribution of Z defined by (9), this decomposition is strikingly close to the shared-covariance attention formula (3):

$$T_{U,V}[\mu_Z](z) = V S U z + V \sum_k \alpha_k(z) m_k,$$

$$\text{with } \alpha_k(z) = \text{softmax}_k(\log \pi_k + m_k^\top U z).$$

Both maps combine a linear transform of the observation with a softmax-weighted barycenter. Two obstructions remain, however. First, the posterior weights β_k contain the offsets $-\frac{1}{2} m_k^\top S^{-1} m_k$, which standard dot-product attention cannot encode. Second, within a single attention head, the same matrices (U, V) govern both the linear and gated terms.

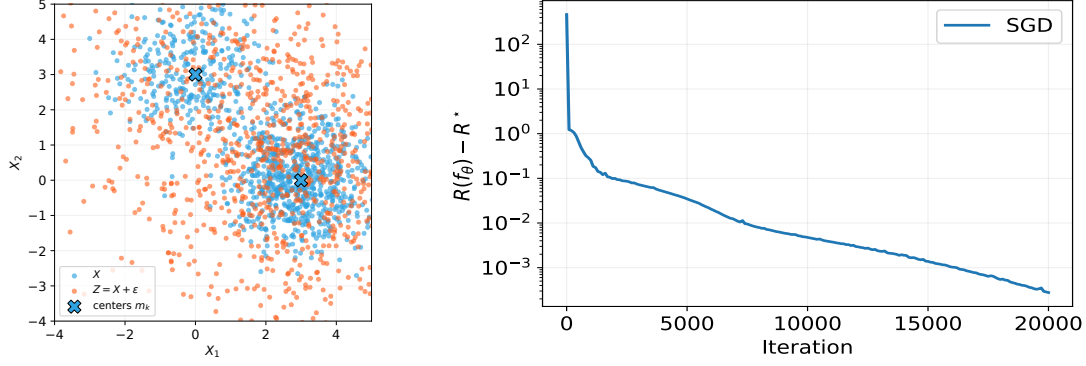


Figure 2: Learning the two-head Bayes denoiser in $d = 2$. Left: clean samples X and noisy observations $Z = X + \varepsilon$ for $m_1 = (3, 0)$, $m_2 = (0, 3)$, $\gamma = 1$, and $\xi = 2$. Right: excess population risk along stochastic gradient descent with step size 2×10^{-3} .

To isolate the second obstruction, we henceforth consider a mixture of 2 components in dimension $d \geq 2$, with $\Gamma = \gamma^2 I_d$, distinct means satisfying $\|m_1\| = \|m_2\|$, and arbitrary positive mixture weights. The equal-norm assumption cancels out the offsets in the barycenter coefficients, giving

$$f^*(z) = \frac{\gamma^2}{\gamma^2 + \xi^2} z + \frac{\xi^2}{\gamma^2 + \xi^2} \sum_{k=1}^2 \beta_k(z) m_k, \quad (11)$$

with $\beta_k(z) = \text{softmax}_k \left(\log \pi_k + \frac{m_k^\top z}{\gamma^2 + \xi^2} \right)$.

4.2 Two heads are sufficient, and softmax is necessary

Although Equation (11) has exactly the form of an attention map, one head is generally unable to match its two terms simultaneously.

Theorem 4.1 (Exact representation and a softmax separation). *Assume $\gamma > 0$. Two softmax heads realize the Bayes denoiser, i.e., there exist $(U_h, V_h)_{h=1,2}$ such that for all z ,*

$$T_{U_1, V_1}[\mu_Z](z) + T_{U_2, V_2}[\mu_Z](z) = f^*(z).$$

A single head suffices if and only if $\gamma^2 = \xi^2$.

Moreover, writing $\mathcal{F}_{\text{sm}}$ and $\mathcal{F}_{\text{lin}}$ for finite multi-head softmax-attention and multi-layer linear-attention predictors, respectively,

$$\inf_{f \in \mathcal{F}_{\text{sm}}} \{ \mathcal{R}_{\text{sq}}(f) - \mathcal{R}_{\text{sq}}(f^*) \} = 0,$$

$$\inf_{f \in \mathcal{F}_{\text{lin}}} \{ \mathcal{R}_{\text{sq}}(f) - \mathcal{R}_{\text{sq}}(f^*) \} > 0.$$

A single head couples the two terms of f^* through the same value matrix. To see this, take $U = S^{-1}$, which already reproduces the gate exactly ($\alpha_k = \beta_k$). The head then reads

$$T_{U, V}[\mu_Z](z) = Vz + V \sum_{k=1}^2 \beta_k(z) m_k,$$

whereas f^* multiplies z by $\gamma^2/(\gamma^2 + \xi^2)$ and the barycenter by $\xi^2/(\gamma^2 + \xi^2)$. The same matrix V would have to apply both factors, which is possible only when they coincide, that is, when $\gamma^2 = \xi^2$; then $V = \frac{1}{2}I_d$ gives f^* exactly. The proof shows that this tension persists for every choice of U (Appendix C.5).

An extra attention head path decouples the two factors. For non-collinear means, each value matrix keeps one mean and cancels the other, while the two linear parts add up to the required shrinkage; opposite means admit a similar construction (Appendix C.6).

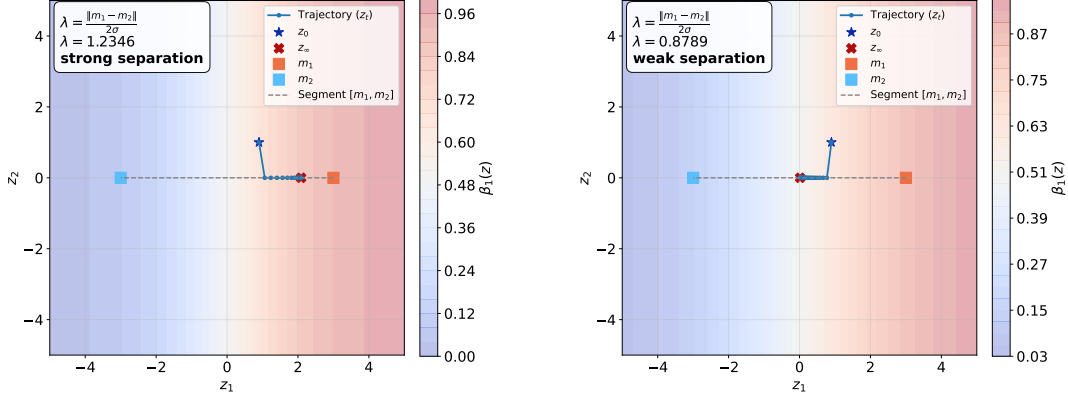


Figure 3: Trajectories of the exact denoiser under strong separation (left) and weak separation (right). The background shows the conditional probability of belonging to cluster 1; blue dots trace the iterates and the red cross marks their limit. Here $m_1 = -m_2 = (0, 3)$ and $\gamma = 0$.

Remark 4.2 (Degenerate mixtures). *When $\gamma = 0$, the clean distribution consists of two point masses, and f^* reduces to their posterior barycenter. Two heads still suffice, and the strict separation from linear attention persists. The single-head obstruction, however, depends on the geometry: it holds for non-collinear means in dimension two, whereas opposite means admit a single-head realization. We keep these complementary cases in Appendix C.3. They will also help isolate the clustering dynamics of repeated denoising below.*

Linear vs. softmax attention. Theorem 4.1 tells us that no network composed of linear attention layers can approximate the Bayes risk. Although the setting and task considered differ, this result aligns with the findings of Duranthon et al. (2026), providing another example where softmax attention is provably more expressive than its linear counterpart. The key reason is that the Bayes-optimal denoiser is inherently nonlinear: it maps each input z to an adaptive average of the mixture centroids m_1 and m_2 .

Can the exact denoiser be learned? Unlike in LDA, the sigmoidal term cannot be suppressed without losing the target itself, so the optimization problem remains genuinely nonlinear and nonconvex. We therefore examine it numerically at the population level. In dimension $d = 2$, we optimize the excess risk $\mathbb{E}_{Z \sim \mu_Z} \|f(Z) - f^*(Z)\|^2$ using online stochastic gradient descent, with gradient norm clipped at (10) to prevent numerical instabilities during the first optimization steps. Throughout training, we follow the evolution of the excess risk, estimated by Gauss–Hermite quadrature within each Gaussian component (Figure 2). The excess risk decreases close to zero in the reported experiment, suggesting that the Bayes predictor can be approached by direct optimization. This is numerical evidence for learnability.

4.3 Iterating the denoiser produces a clustering transition

While a single denoising step aims at recovering a clean estimate from a corrupted query, repeatedly applying the same denoiser defines a dynamical system whose long-term behavior can reveal the geometry encoded by the data distribution. In particular, Smart et al. (2025) show that a denoising step can be interpreted as a gradient descent step associated with a dense associative memory network making the study of its iterates natural from a dynamical perspective. Consider the degenerate balanced mixture $X \in \{m_1, m_2\}$, so that X emanates from a mixture of Dirac masses, with $\gamma = 0$ and $\pi_1 = \pi_2 = 1/2$, and consider the dynamics $z_{n+1} = f^*(z_n)$. One attention layer encoding f^* is obviously Bayes optimal, but iteration progressively erases information about the initial query. We show that this loss of information has a precise interpretation: depending on the separation of the mixture, the denoiser either collapses all inputs to their common midpoint or partitions the space into two basins of attraction.

Let $\lambda = \|m_1 - m_2\|/(2\xi)$ so that λ^2 can be interpreted as a signal-to-noise ratio. After projecting onto the line joining the means, the fixed-point equation reduces to $u = \tanh(\lambda^2 u)$, yielding a transition at $\lambda = 1$.

Proposition 4.3 (Denoising-to-clustering transition). *If $\lambda \leq 1$ (referred to as weak separation), the iterated optimal denoiser dynamics $z_{n+1} = f^*(z_n)$ converges to*

$$z_n \xrightarrow{n \rightarrow +\infty} (m_1 + m_2)/2.$$

If $\lambda > 1$ (referred to as strong separation), almost every initialization converges to one of two stable fixed points $m_{\lambda,1}, m_{\lambda,2}$, being such that $m_{\lambda,k} \rightarrow m_k$ as $\lambda \rightarrow \infty$.

The proof is given in Appendix C.4. Figure 3 displays the two regimes described by Proposition 4.3. Under weak separation the dynamic cannot distinguish the components and every trajectory collapses to the midpoint. Above the threshold, two stable attractors emerge and approach the component means as the separation grows. Repeated passages through the same layer therefore changes the statistical role of the same map: a one-step estimator becomes, under recurrence, a clustering mechanism.

4.4 Residual attention denoises in-context

The previous analyses considered denoising for a fixed mixture, with optimal attention parameters tailored to the mixture means. In-context denoising asks for a single architecture that adapts to different mixtures using only a prompt of noisy observations.

A task is therefore now specified by $\tau = (m_1, m_2, \pi_1)$, with clean and noisy distributions

$$\begin{aligned}\mu_X^\tau &= \pi_1 \mathcal{N}(m_1, \gamma^2 I_d) + (1 - \pi_1) \mathcal{N}(m_2, \gamma^2 I_d), \\ \mu_Z^\tau &= \pi_1 \mathcal{N}(m_1, S) + (1 - \pi_1) \mathcal{N}(m_2, S),\end{aligned}$$

where $S = (\gamma^2 + \xi^2)I_d$. The means and mixture weights vary across tasks, while γ^2 and ξ^2 remain fixed. For each task, a new parameter τ is sampled according to an admissible distribution (i.e., with $m_1 \neq m_2$, $\|m_1\| = \|m_2\|$, and $\pi_1 \in (0, 1)$ almost surely). The layer receives an infinite prompt $Z_1, \dots, Z_L, \dots$ drawn independently from μ_Z^τ and an independent noisy query $Z = X + \varepsilon$ from the same task. Its goal is to reconstruct the query's clean counterpart X . The predictor is evaluated through the in-context risk

$$\mathcal{R}_{\text{ICL}}(f) = \mathbb{E}_\tau \mathbb{E}[\|X - f[\mu_Z^\tau](Z)\|^2].$$

The attention weights are shared across tasks; adaptation occurs through the context argument μ_Z^τ . We therefore seek one set of weights satisfying $\hat{f}[\mu_Z^\tau](z) = \mathbb{E}[X \mid Z = z, \tau]$ for every admissible task τ , without retraining when its means or proportions change.

To solve this problem, the attention mechanism must infer from the context both the barycentric combination of the component centroids and the linear contraction required to reproduce the Bayes-optimal denoiser f^* in (10). Unlike in the previous setting, however, the attention weights can no longer depend on the task parameters (m_1, m_2, π_1) , as these vary across tasks. We therefore add a residual connection allowing the network to adjust the linear term and the barycentric mean separately. We define $T_{U,V}^{\text{res}}[\mu](z) = z + T_{U,V}[\mu](z)$ and combine two such residual heads through a linear output map.

Proposition 4.4 (Exact in-context adaptation). *For every admissible prompt law (such that $m_1 \neq m_2$, $\|m_1\| = \|m_2\|$, and $\pi_1 \in (0, 1)$ almost surely), the architecture*

$$\hat{f}[\mu](z) = W_{\text{out}} \begin{pmatrix} T_{U_1, V_1}^{\text{res}}[\mu](z) \\ T_{U_2, V_2}^{\text{res}}[\mu](z) \end{pmatrix}$$

equals its Bayes denoiser with the task-independent parameters

$$\begin{aligned}U_1 &= S^{-1}, & V_1 &= \frac{\xi^2}{\gamma^2 + \xi^2} I_d, \\ U_2 &= V_2 = 0, & W_{\text{out}} &= \left(I_d, -\frac{2\xi^2}{\gamma^2 + \xi^2} I_d \right).\end{aligned}$$

The proof is given in Appendix C.8. The task-dependent means and weights never enter the layer parameters. The first head reads them through the context distribution and returns their weighted barycenter; the second head provide copies of the query, allowing the output map to adjust the linear shrinkage independently. In fact, the parameters encode invariants shared across tasks, while attention extracts the latent structure specific to the current prompt. This proposal shows that a single-layer softmax attention architecture with residual connections can perform in-context denoising. Interestingly, it also highlights how residual connections can enhance the expressivity of attention-based architectures; in contrast to neural networks, where their primary role is to facilitate optimization.

5 Conclusion

Gaussian mixtures provide a tractable setting in which context selection has a concrete statistical role. For classification, the Bayes score is affine, yet the nonlinear attention parameterization creates a rich optimization geometry. For denoising, query-dependent selection becomes essential, and residual connections allow the same parameters to serve different mixture distributions. Together, these results connect specific features of attention architectures to the statistical operations they perform. Extending this understanding to finite prompts and explaining how training discovers such adaptive solutions are natural next steps.

AI use statement

In this work, we used generative AI for the following tasks: literature review, including identifying articles relevant to our research; idea exploration, through prompts aimed at investigating some of our initial leads and assessing their potential relevance; calculation and proof assistance; programming assistance; and language polishing. All AI-assisted work was reviewed and verified by the authors. In particular, we independently reviewed the cited articles, further investigated the ideas we found relevant, and checked AI-assisted calculations and proofs ourselves. The problem studied and our research strategies were developed through discussions among the authors, rather than generated by AI. We take full responsibility for the final content of this work, including all text, claims, calculations, and other artifacts produced with the assistance of generative AI.

Acknowledgment

This work was supported by the French Ministry of Higher Education, Research and Space (MESRE) through an AMX doctoral fellowship awarded by École Polytechnique.

References

- Augustin Banyaga and David E. Hurtubise. A proof of the morse–bott lemma. *Expositiones Mathematicae*, 22(4):365–373, 2004.
- Léa Bohbot, Cyril Letrouit, Gabriel Peyré, and François-Xavier Vialard. Token sample complexity of attention. In *Forty-third International Conference on Machine Learning*, 2026.
- Etienne Boursier and Claire Boyer. Softmax as linear attention in the large-prompt regime: a measure-based perspective. In *Forty-third International Conference on Machine Learning*, 2026.
- Valère Castin, Pierre Ablin, José A. Carrillo, and Gabriel Peyré. A unified perspective on the dynamics of deep transformers. *Foundations of Computational Mathematics*, 2025.
- Odilon Duranthon, Pierre Marion, Claire Boyer, Bruno Loureiro, and Lenka Zdeborová. Statistical advantage of softmax attention: Insights from single-location regression. In *International Conference on Learning Representations*, 2026.
- Spencer Frei and Gal Vardi. Trained transformer classifiers generalize and exhibit benign overfitting in-context. In *International Conference on Learning Representations*, 2025.

- Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. In *Advances in Neural Information Processing Systems*, volume 36, pages 57026–57037, 2023.
- Gautam Goel, Mahdi Soltanolkotabi, and Peter Bartlett. Training dynamics of softmax self-attention: Fast global convergence via preconditioning. *arXiv preprint arXiv:2603.01514*, 2026.
- Victor Guillemin and Alan Pollack. *Differential Topology*. Prentice-Hall, 1974.
- Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.
- Hongkang Li, Hancheng Min, and Rene Vidal. Transformers learn the optimal DDPM denoiser for multi-token GMMs, 2026.
- Rodrigo Mauleen-Soto and Claire Boyer. Attention-based PCA, 2026.
- Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
- Quentin Rebjock and Nicolas Boumal. Fast convergence to non-isolated minima: Four equivalent conditions for C^2 functions. *Mathematical Programming*, 213(1):151–199, 2025.
- Paul Rosu, Lawrence Carin, and Xiang Cheng. From softmax to score: Transformers can effectively implement in-context denoising steps. In *Advances in Neural Information Processing Systems*, volume 38, 2025.
- Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 3515–3530. PMLR, 2022.
- Wei Shen, Ruida Zhou, Jing Yang, and Cong Shen. On the training convergence of transformers for in-context classification of gaussian mixtures. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 54732–54771. PMLR, 2025.
- Matthew Smart, Alberto Bietti, and Anirvan M. Sengupta. In-context denoising with one-layer transformers: Connections between attention and associative memory retrieval. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 55950–55971. PMLR, 2025.
- Matthew Smart, Soumya Ganguly, Nilava Metya, Alexandre V. Morozov, and Anirvan M. Sengupta. Attention as in-context empirical bayes: A two-stage view via particle dynamics, 2026.
- James Vuckovic, Aristide Baratin, and Remi Tachet des Combes. A mathematical theory of attention, 2020.
- Chenyang Zhang and Yuan Cao. Transformers efficiently perform in-context logistic regression via normalized gradient descent. In *Forty-third International Conference on Machine Learning*, 2026.

Appendix

Table of Contents

A Proofs and Complements for Section 2	14
A.1 Gaussian-mixture attention (proof of Proposition 2.1)	14
A.2 General mixtures (extension of Proposition 2.1)	14
B Proofs and Complements for Section 3	14
B.1 Effective LDA predictor (derivation of Equations (6)–(7))	15
B.2 Bayes realizability (preparation for Theorems 3.1 and 3.3)	16
B.3 Population logistic-risk minimizer (preparation for Theorem 3.1)	17
B.4 Differential identities (tools for the LDA landscape and convergence proofs)	18
B.5 Optimization landscape (proofs of Theorems 3.1 and 3.3, Lemma 3.2, and Proposition 3.4)	24
B.6 Normal curvature (preparation for Theorem 3.5)	30
B.7 Local gradient-flow convergence (proof of Theorem 3.5)	33
B.8 Constant-gate geometry (complement to Theorem 3.5)	37
B.9 Reduced-risk estimates and conserved balance (preparation for Theorem 3.6)	38
B.10 Feasibility of the Krylov initialization (complement to Theorem 3.6)	40
B.11 Invariance of the Krylov subspace (preparation for Theorem 3.6)	42
B.12 Global convergence on the invariant subspace (proof of Theorem 3.6)	43
B.13 LDA experiments (illustrations of Theorems 3.5 and 3.6)	45
C Proofs and Complements for Section 4	47
C.1 Bayes denoiser (derivation of Equation (10))	47
C.2 Expressivity and risk separation (complete proof of Theorem 4.1)	48
C.3 Degenerate mixtures (complement to Theorem 4.1)	48
C.4 Denoising-to-clustering transition (proof of Proposition 4.3)	49
C.5 Single-head necessity (Theorem 4.1, single-head claim)	50
C.6 Two-head construction (Theorem 4.1, exact representation)	51
C.7 Strict risk gap (Theorem 4.1, separation claim)	53
C.8 Residual in-context denoising (proof of Proposition 4.4)	54

Where to find each main-text proof. The table below locates every numbered result in the main text. Titles distinguish complete proofs, preparatory arguments, and complements. For results assembled from several lemmas, the indicated location contains an explicitly named proof that identifies the ingredients.

Main-text result	Proof location
Proposition 2.1	Appendix A.1.
Theorem 3.1	Appendix B.5.2, using the minimizers identified in Appendix B.5.1.
Lemma 3.2	Appendix B.5.3.
Theorem 3.3	Appendix B.5.4, using the minimizer preimage in Appendix B.5.1.
Proposition 3.4	Appendix B.5.5.
Theorem 3.5	Appendix B.7.
Theorem 3.6	Appendix B.12.
Theorem 4.1	Appendix C.2; detailed arguments in Appendices C.5, C.6, and C.7.
Proposition 4.3	Appendix C.4.
Proposition 4.4	Appendix C.8.

A Proofs and Complements for Section 2

A.1 Gaussian-mixture attention (proof of Proposition 2.1)

Proof. Set $u := Uz$. Then

$$T_{U,V}[\mu](z) = V \frac{\mathbb{E}[Xe^{\langle u, X \rangle}]}{\mathbb{E}[e^{\langle u, X \rangle}]}.$$

It is well known that the moment-generating function of a Gaussian vector is given by $\mathbb{E}[e^{\langle u, X \rangle}] = \exp(\langle u, m \rangle + \frac{1}{2}u^\top \Gamma u)$. By differentiation, we obtain $\mathbb{E}[Xe^{\langle u, X \rangle}] = (m + \Gamma u) \exp(\langle u, m \rangle + \frac{1}{2}u^\top \Gamma u)$

Thus, for a Gaussian mixture $X \sim \mu = \sum \pi_k \mu_k$, one has

$$\mathbb{E}[e^{\langle u, X \rangle}] = \sum_{k=1}^M \pi_k \exp\left(\langle u, m_k \rangle + \frac{1}{2}u^\top \Gamma_k u\right),$$

and

$$\mathbb{E}[Xe^{\langle u, X \rangle}] = \sum_{k=1}^M \pi_k (m_k + \Gamma_k u) \exp\left(\langle u, m_k \rangle + \frac{1}{2}u^\top \Gamma_k u\right).$$

Therefore,

$$T_{U,V}[\mu](z) = V \sum_{k=1}^M \alpha_k(z) (m_k + \Gamma_k u),$$

where $\alpha_k(z) = \frac{\pi_k \exp\left(\langle u, m_k \rangle + \frac{1}{2}u^\top \Gamma_k u\right)}{\sum_{\ell=1}^M \pi_\ell \exp\left(\langle u, m_\ell \rangle + \frac{1}{2}u^\top \Gamma_\ell u\right)}$ which concludes the proof. $\square$

A.2 General mixtures (extension of Proposition 2.1)

For a positive alignment function a , define the normalized attention map

$$T_a[\mu](z) = \frac{\int a(z, z') z' d\mu(z')}{\int a(z, z') d\mu(z')}.$$

Proposition A.1 (Attention on a general mixture). *Let $\mu = \sum_{k=1}^M \pi_k \mu_k$. Then, for every $z \in \mathbb{R}^d$,*

$$T_a[\mu](z) = \sum_{k=1}^M r_k(z) T_a[\mu_k](z), \quad r_k(z) = \pi_k \frac{\int a(z, z') d\mu_k(z')}{\int a(z, z') d\mu(z')}.$$

Proof. This proposition follows directly from the fact that

$$\int f d\mu = \int f \sum_k \pi_k d\mu_k.$$

$\square$

B Proofs and Complements for Section 3

This appendix develops the representation, landscape, and convergence arguments underlying Section 3. We first derive the attention predictor and identify the population target of logistic training. We then separate two questions: which parameters minimize the risk, and whether gradient flow approaches such parameters. The distinction matters because a favorable finite critical-point structure does not control trajectories at infinity.

The landscape argument is developed in Appendix B.5; the subsequent analysis establishes local convergence near the regular Bayes manifold and global convergence for a prescribed initialization in a symmetric invariant subspace. The auxiliary calculations are collected before the arguments that use them. Throughout, $\tilde{\mathcal{R}}$ denotes the effective LDA risk and $\mathcal{R} = \tilde{\mathcal{R}} \circ \Phi$ its expression in attention parameters.

B.1 Effective LDA predictor (derivation of Equations (6)–(7))

For parameter matrices $U, V \in \mathbb{R}^{(d+1) \times (d+1)}$, we consider the block structure:

$$U^\top = \left[\begin{array}{c|c} U_{11} & u_{12} \\ \hline u_{21}^\top & u_{22} \end{array} \right], \quad V = \left[\begin{array}{c|c} V_{11} & v_{12} \\ \hline v_{21}^\top & v_{22} \end{array} \right], \quad (12)$$

where $U_{11}, V_{11} \in \mathbb{R}^{d \times d}$, $u_{12}, u_{21}, v_{12}, v_{21} \in \mathbb{R}^d$ are $U_{22}, v_{22} \in \mathbb{R}$.

Equation (6) is a direct consequence of the following lemma.

Lemma B.1. *Let μ_{LDA} denote the joint law defined by Equation (4). For every $(x, y) \in \mathbb{R}^d \times \{-1, 0, 1\}$, one has*

$$T_{U,V}[\mu_{\text{LDA}}](x, y) = V \begin{pmatrix} \Gamma(U_{11}^\top x + y u_{21}) + (1 - \alpha(x, y)) m_{-1} + \alpha(x, y) m_1 \\ 2\alpha(x, y) - 1 \end{pmatrix},$$

where

$$\alpha(x, y) = \sigma \left(\log \frac{\pi_+}{1 - \pi_+} + 2(u_{12}^\top x + y u_{22}) + \langle U_{11}^\top x + y u_{21}, m_1 - m_{-1} \rangle \right),$$

with $\sigma(t) := \frac{1}{1 + e^{-t}}$ the sigmoid function.

Proof. For $z = (x, y)$ and $z' = (x', y')$, we then have

$$z^\top U^\top z' = x^\top U_{11} x' + x^\top u_{12} y' + y u_{21}^\top x' + y u_{22} y'.$$

Hence,

$$z^\top U^\top z' = \langle s(x, y), x' \rangle + t(x, y) y',$$

where the linear quantities are explicitly given by

$$s(x, y) := U_{11}^\top x + y u_{21} \in \mathbb{R}^d, \quad t(x, y) := u_{12}^\top x + y u_{22} \in \mathbb{R}.$$

Denominator. For any $s \in \mathbb{R}^d$ and any $c \in \{-1, 1\}$, the moment-generating function of a Gaussian vector gives

$$\mathbb{E}[e^{\langle s, X \rangle} \mid Y = c] = \exp \left(\langle s, m_c \rangle + \frac{1}{2} s^\top \Gamma s \right).$$

Therefore,

$$\begin{aligned} \int \exp(z^\top U^\top z) d\mu_{\text{LDA}}(z') &= \mathbb{E}_{(X,Y) \sim \mu_{\text{LDA}}} [\exp(\langle s(x, y), X \rangle + t(x, y) Y)] \\ &= \sum_{c \in \{-1, 1\}} \pi_c \exp(c t(x, y)) \mathbb{E}[e^{\langle s(x, y), X \rangle} \mid Y = c] \\ &= \exp \left(\frac{1}{2} s(x, y)^\top \Gamma s(x, y) \right) \left((1 - \pi_+) \exp(-t(x, y) + \langle s(x, y), m_{-1} \rangle) \right. \\ &\quad \left. + \pi_+ \exp(t(x, y) + \langle s(x, y), m_1 \rangle) \right). \end{aligned}$$

Numerator. Similarly, for any $c \in \{-1, 1\}$,

$$\mathbb{E}[X e^{\langle s, X \rangle} \mid Y = c] = (m_c + \Gamma s) \exp \left(\langle s, m_c \rangle + \frac{1}{2} s^\top \Gamma s \right).$$

Therefore,

$$\begin{aligned} \int \exp(z^\top U^\top z') z' d\mu_{\text{LDA}}(z') &= \sum_{c \in \{-1, 1\}} \pi_c \exp(c t(x, y)) \mathbb{E} \left[\begin{pmatrix} X \\ c \end{pmatrix} e^{\langle s(x, y), X \rangle} \mid Y = c \right] \\ &= \exp \left(\frac{1}{2} s^\top \Gamma s \right) \left[(1 - \pi_+) \exp(-t + \langle s, m_{-1} \rangle) \begin{pmatrix} m_{-1} + \Gamma s \\ -1 \end{pmatrix} \right. \\ &\quad \left. + \pi_+ \exp(t + \langle s, m_1 \rangle) \begin{pmatrix} m_1 + \Gamma s \\ 1 \end{pmatrix} \right], \end{aligned}$$

where we wrote $s = s(x, y)$ and $t = t(x, y)$ in the last inequality. Dividing numerator and denominator, we obtain

$$\frac{\int \exp(\langle Uz, z' \rangle) z' d\mu(z')}{\int \exp(\langle Uz, z' \rangle) d\mu(z')} = \left(\frac{\Gamma s(x, y) + (1 - \alpha(x, y)) m_{-1} + \alpha(x, y) m_1}{2\alpha(x, y) - 1} \right),$$

where

$$\alpha(x, y) := \frac{\pi_+ \exp(t(x, y) + \langle s(x, y), m_1 \rangle)}{(1 - \pi_+) \exp(-t(x, y) + \langle s(x, y), m_{-1} \rangle) + \pi_+ \exp(t(x, y) + \langle s(x, y), m_1 \rangle)}.$$

Equivalently,

$$\alpha(x, y) = \sigma \left(\log \frac{\pi_+}{1 - \pi_+} + 2t(x, y) + \langle s(x, y), m_1 - m_{-1} \rangle \right),$$

which allows to conclude by plugging the values of $s(x, y)$ and $t(x, y)$. $\square$

B.2 Bayes realizability (preparation for Theorems 3.1 and 3.3)

We recall the form of the optimal predictor in the LDA model, which is a classical result in statistical learning theory, see e.g., [Hastie et al. \(2009\)](#).

Proposition B.2 (Bayes classifier in the LDA model). *Under the model (4), the Bayes classifier is $g^*(x) = \text{sign}(f^*(x))$, where*

$$f^*(x) = \langle w^*, x \rangle + b^*, \quad w^* = \Gamma^{-1}(m_1 - m_{-1}),$$

and

$$b^* = -\frac{1}{2} (m_1^\top \Gamma^{-1} m_1 - m_{-1}^\top \Gamma^{-1} m_{-1}) + \log \frac{\pi_+}{1 - \pi_+}.$$

Lemma B.3 (Existence of Bayes-realizing attention parameters). *There exists an attention parameter θ^* such that $f_{\theta^*}(x) = f^*(x)$ for every $x \in \mathbb{R}^d$. Consequently, the attention class attains the Bayes logistic risk.*

Proof of Lemma B.3. We recall that, in the LDA model, the Bayes score is affine and can be written as

$$f^*(x) = \langle w^*, x \rangle + b^*.$$

Equation (6) gives f_θ in terms of the effective parameters:

$$f_\theta(x) = \underbrace{v_{21}^\top \Gamma U_{11}^\top x + v_{21}^\top m_{-1} - v_{22}}_{\text{linear part}} + \underbrace{\alpha(x) (v_{21}^\top (m_1 - m_{-1}) + 2v_{22})}_{\text{nonlinear part}},$$

where

$$\alpha(x) = \sigma \left(\log \frac{\pi_+}{1 - \pi_+} + (2u_{12} + U_{11}(m_1 - m_{-1}))^\top x \right).$$

We start by imposing the constraint

$$2u_{12} + U_{11}(m_1 - m_{-1}) = 0, \quad \text{that is,} \quad u_{12} = -\frac{1}{2} U_{11}(m_1 - m_{-1}). \quad (\text{C})$$

With this choice, the argument of the sigmoid no longer depends on x , and therefore

$$\alpha(x) = \sigma \left(\log \frac{\pi_+}{1 - \pi_+} + (2u_{12} + U_{11}(m_1 - m_{-1}))^\top x \right) = \sigma \left(\log \frac{\pi_+}{1 - \pi_+} \right) = \pi_+.$$

Hence f_θ becomes affine:

$$f_\theta(x) = v_{21}^\top \Gamma U_{11}^\top x + v_{21}^\top \bar{m}_{\pi_+} + (2\pi_+ - 1) v_{22}, \quad \bar{m}_{\pi_+} := (1 - \pi_+) m_{-1} + \pi_+ m_1.$$

Case 1. Assume first that $\pi_+ \neq \frac{1}{2}$. Then, the coefficient of v_{22} is nonzero, so the intercept can be adjusted freely. More precisely, given any $w^* \in \mathbb{R}^d$ and $b^* \in \mathbb{R}$, one may choose

$$U_{11} = I_d, \quad u_{12} = -\frac{1}{2}(m_1 - m_{-1}), \quad v_{21} = \Gamma^{-1}w^*,$$

and then define

$$v_{22} = \frac{b^* - (\Gamma^{-1}w^*)^\top \bar{m}_{\pi_+}}{2\pi_+ - 1}.$$

With this choice, we obtain

$$f_\theta(x) = \langle w^*, x \rangle + b^*.$$

Case 2. Assume now that $\pi_+ = \frac{1}{2}$. In that case, still under constraint (C)

$$\alpha(x) \equiv \frac{1}{2}, \quad \bar{m}_{\pi_+} = \frac{m_{-1} + m_1}{2},$$

and the term involving v_{22} disappears:

$$f_\theta(x) = v_{21}^\top \Gamma U_{11}^\top x + v_{21}^\top \frac{m_{-1} + m_1}{2}.$$

Case 2(a). If $m_1 \neq -m_{-1}$, that is, if $\bar{m} := \frac{m_{-1} + m_1}{2} \neq 0$, one may choose $v_{21} \neq 0$ so that $v_{21}^\top \bar{m} = b^*$ and then choose U_{11} so that $U_{11} \Gamma v_{21} = w^*$. For instance, say $v_{21} = \frac{b^*}{\|\bar{m}\|^2} \bar{m}$ (if $b^* = 0$, one may choose any $v_{21} \in \text{Span}(\bar{m})^\perp \setminus \{0\}$), and define $a := \Gamma v_{21} \neq 0$. One may take $U_{11} = \frac{w^* a^\top}{\|a\|^2}$ so that we have

$$f_\theta(x) = \langle w^*, x \rangle + b^*.$$

Case 2(b). Finally, if $\pi_+ = \frac{1}{2}$ and $m_1 = -m_{-1}$, then $\bar{m} = 0$ and

$$f_\theta(x) = v_{21}^\top \Gamma U_{11}^\top x.$$

In the symmetric balanced case considered here, the target intercept is

$$b^* = -\frac{1}{2} \left(m_1^\top \Gamma^{-1} m_1 - m_{-1}^\top \Gamma^{-1} m_{-1} \right) + \log \frac{\pi_+}{1 - \pi_+} = 0,$$

so there is no obstruction: it is enough to choose

$$U_{11} \Gamma v_{21} = w^*,$$

for instance $U_{11} = I_d$ and $v_{21} = \Gamma^{-1}w^*$.

Remark. This choice of parameters is not unique. Another possible strategy, when $m_1 \neq -m_{-1}$, is to first cancel the amplitude in front of the sigmoid term and then adjust the remaining coefficients accordingly. $\square$

B.3 Population logistic-risk minimizer (preparation for Theorem 3.1)

Before studying critical points in parameter space, we identify the predictor that training seeks to recover. The logistic loss is particularly well suited to this purpose: its unrestricted population minimizer is the Bayes predictor, which the attention model can represent exactly. Thus the optimal risk is the same in the effective model, in the original attention model, and over all measurable scores.

Lemma B.4. *Over all measurable scores, the population logistic risk is uniquely minimized almost surely by the Bayes predictor f^* defined in Equation (5).*

Proof. We want to minimize $\mathcal{R}_{\text{LDA}}(f)$ over the set of measurable functions.

$$\mathcal{R}_{\text{LDA}}(f) = \mathbb{E}_{(X,Y) \sim \mu_{\text{LDA}}} [\ell_{\log}(Y, f(X))] = \mathbb{E}_X [\ell_{\log}(1, f(X))\eta(X) + \ell_{\log}(-1, f(X))(1 - \eta(X))]$$

We now define $g_X(t) := \ell_{\log}(1, t)\eta(X) + \ell_{\log}(-1, t)(1 - \eta(X))$. Using that $\ell'(y, t) = -\frac{y}{1+e^{yt}}$, we obtain $g'_X(t) = \sigma(t) - \eta(X)$. Using the strict convexity and coercivity of g_X (as long as $\eta(X) \notin \{0, 1\}$ which is almost sure), we obtain by differentiation that :

$$t(X) \in \arg \min g_X \iff t(X) = \log \left(\frac{\eta(X)}{1 - \eta(X)} \right)$$

Finally :

$$\mathcal{R}_{\text{LDA}}(f) = \mathbb{E}_X [g_X(f(X))] \geq \mathbb{E}_X (g_X(t(X))) = \mathcal{R}_{\text{LDA}}(f^*)$$

where the inequality is an equality if and only if

$$f(X) = \log \left(\frac{\eta(X)}{1 - \eta(X)} \right) \quad \text{a.s.}$$

Moreover, this quantity is well-defined since $0 < \eta(X) < 1$ almost surely in the LDA setting. $\square$

Remark B.5. *This property is specific to the logistic loss in the present setting. Indeed, if one considers instead the quadratic loss, the corresponding population minimizer is*

$$f_{\text{sq}}(x) = \mathbb{E}[Y \mid X = x] = 2\eta(x) - 1.$$

Although f_{sq} yields the same Bayes classifier after thresholding at 0, there is in general no reason for f_{sq} to belong to the class $\mathcal{F} := \{f_\theta, \theta \in \Theta\}$. In other words, even if $\mathcal{F}$ contains the logistic target $f_{\log}$, it does not necessarily contain the quadratic target f_{sq} . Hence, minimizing the quadratic risk over $\mathcal{F}$ does not in general allow one to recover the global minimizer of the quadratic loss, thus the optimal classifier.

B.4 Differential identities (tools for the LDA landscape and convergence proofs)

The arguments below use three kinds of calculations: derivatives of the logistic risk, the Hessian at a Bayes realization, and the differential of the map Φ . We collect them here to keep the subsequent landscape argument focused on their geometric consequences. Readers interested first in the structure of the critical set may proceed to Appendix B.5 and return to these identities as needed.

The first Gaussian identity will also be used to prove coercivity of the linear risk in the symmetric setting. It quantifies the contribution of observations on the wrong side of a candidate linear decision boundary.

Lemma B.6 (Positive part of a Gaussian). *Let $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma > 0$, and let $X_+ := \max(X, 0)$. If φ and Φ denote respectively the density and distribution function of a standard normal random variable, then*

$$\mathbb{E}[X_+] = \mu\Phi\left(\frac{\mu}{\sigma}\right) + \sigma\varphi\left(\frac{\mu}{\sigma}\right).$$

Proof. Let $c = -\mu/\sigma$ and $Z = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$.

$$\begin{aligned} \mathbb{E}[X_+] &= \int_0^\infty x \frac{1}{\sigma} \varphi\left(\frac{x - \mu}{\sigma}\right) dx \\ &= \int_c^\infty (\mu + \sigma z) \varphi(z) dz \\ &= \mu \int_c^\infty \varphi(z) dz + \sigma \int_c^\infty z \varphi(z) dz \\ &= \mu(1 - \Phi(c)) + \sigma[-\varphi(z)]_c^\infty, \quad \text{since } \varphi'(z) = -z\varphi(z), \\ &= \mu\Phi\left(\frac{\mu}{\sigma}\right) + \sigma\varphi\left(\frac{\mu}{\sigma}\right), \end{aligned}$$

using $1 - \Phi(-u) = \Phi(u)$ and $\varphi(-u) = \varphi(u)$. $\square$

From prediction error to risk derivatives. Conditioning on the query turns the derivative of the logistic loss into the discrepancy between the model probability $\sigma(f_\theta(X))$ and the true posterior $\eta(X)$. This identity is the common starting point for the gradient and Hessian calculations.

Lemma B.7 (A useful identity for logistic loss). *For the logistic loss, conditioning on X gives*

$$\mathbb{E} [\ell'(Y, f_\theta(X)) \mid X] = \sigma(f_\theta(X)) - \eta(X)$$

Proof. For the logistic loss we have $\ell'(y, t) = -\frac{y}{1+e^{yt}}$. Since $Y \in \{-1, 1\}$ and $\eta(X) = \mathbb{P}(Y = 1 \mid X)$, conditioning on X yields

$$\begin{aligned} \mathbb{E} [\ell'(Y, f_\theta(X)) \mid X] &= -\eta(X) \frac{1}{1+e^{f_\theta(X)}} + (1-\eta(X)) \frac{1}{1+e^{-f_\theta(X)}} \\ &= -\eta(X)(1-\sigma(f_\theta(X))) + (1-\eta(X))\sigma(f_\theta(X)) \\ &= \sigma(f_\theta(X)) - \eta(X). \end{aligned}$$

□

Lemma B.8 (First and second variations of the model). *Set*

$$\Delta m := m_1 - m_{-1}, \quad z_\theta(x) := s_0 + s(\theta)^\top x = s_0 + (2u_{12} + U_{11}\Delta m)^\top x.$$

Recall that

$$f_\theta(x) = v_{21}^\top \Gamma U_{11}^\top x + v_{21}^\top m_{-1} - v_{22} + (v_{21}^\top \Delta m + 2v_{22})\sigma(z_\theta(x)).$$

Let

$$h = (\delta v_{21}, \delta v_{22}, \delta U_{11}, \delta u_{12}).$$

Then

$$\begin{aligned} Df_\theta(x)[h] &= \delta v_{21}^\top \Gamma U_{11}^\top x + v_{21}^\top \Gamma \delta U_{11}^\top x + \delta v_{21}^\top m_{-1} - \delta v_{22} \\ &\quad + (\delta v_{21}^\top \Delta m + 2\delta v_{22})\sigma(z_\theta(x)) \\ &\quad + (v_{21}^\top \Delta m + 2v_{22})\sigma'(z_\theta(x))(2\delta u_{12} + \delta U_{11}\Delta m)^\top x. \end{aligned}$$

Moreover,

$$\begin{aligned} D^2 f_\theta(x)[h, h] &= 2\delta v_{21}^\top \Gamma \delta U_{11}^\top x \\ &\quad + 2(\delta v_{21}^\top \Delta m + 2\delta v_{22})\sigma'(z_\theta(x))(2\delta u_{12} + \delta U_{11}\Delta m)^\top x \\ &\quad + (v_{21}^\top \Delta m + 2v_{22})\sigma''(z_\theta(x)) \left[(2\delta u_{12} + \delta U_{11}\Delta m)^\top x \right]^2. \end{aligned}$$

Proof. Direct computations

□

Lemma B.9. *For every $(w, b, r, s) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$, the function $\tilde{\mathcal{R}}$ is differentiable and its gradients are given by*

$$\begin{cases} \nabla_w \tilde{\mathcal{R}}(w, b, r, s) = \mathbb{E} \left[(\sigma(g_{w,b,r,s}(X)) - \eta(X)) X \right], \\ \partial_b \tilde{\mathcal{R}}(w, b, r, s) = \mathbb{E} \left[\sigma(g_{w,b,r,s}(X)) - \eta(X) \right], \\ \partial_r \tilde{\mathcal{R}}(w, b, r, s) = \mathbb{E} \left[(\sigma(g_{w,b,r,s}(X)) - \eta(X)) \sigma(s_0 + s^\top X) \right], \\ \nabla_s \tilde{\mathcal{R}}(w, b, r, s) = r \mathbb{E} \left[(\sigma(g_{w,b,r,s}(X)) - \eta(X)) \sigma(s_0 + s^\top X) (1 - \sigma(s_0 + s^\top X)) X \right]. \end{cases}$$

This can also be written as

$$\nabla_\zeta \tilde{\mathcal{R}}(\zeta) = \mathbb{E} \left[(\sigma(g_\zeta(X)) - \eta(X)) \nabla g_\zeta(X) \right], \quad \zeta = (w, b, r, s).$$

Proof. Let $g(X) = g_{w,b,r,s}(X) = w^\top x + b + r \sigma(s_0 + s^\top x)$. One has,

$$\nabla \tilde{\mathcal{R}}(w, b, r, s) = \mathbb{E} \left[\ell'(Y, g(X)) \nabla g(X) \right],$$

Using Lemma B.7,

$$\nabla \tilde{\mathcal{R}}(w, b, r, s) = \mathbb{E} \left[(\sigma(g(X)) - \eta(X)) \nabla g(X) \right].$$

The formulas for the partial gradients then follow by differentiating $g_{w,b,r,s}$ with respect to (w, b, r, s) (where we use the fact that $\sigma' = \sigma(1 - \sigma)$ for the gradient with respect to s). $\square$

Lemma B.10 (Explicit gradients of $\mathcal{R}$). *The gradients of $\mathcal{R}$ are given by*

$$\begin{cases} \nabla_{U_{11}} \mathcal{R}(\theta) = \mathbb{E} \left[(\sigma(f_\theta(X)) - \eta(X)) (X(\Gamma^\top v_{21})^\top + (v_{21}^\top \Delta m + 2v_{22}) \sigma'(s_0 + (2u_{12} + U_{11} \Delta m)^\top X) X \Delta m^\top) \right], \\ \nabla_{u_{12}} \mathcal{R}(\theta) = 2(v_{21}^\top \Delta m + 2v_{22}) \mathbb{E} \left[(\sigma(f_\theta(X)) - \eta(X)) \sigma'(s_0 + (2u_{12} + U_{11} \Delta m)^\top X) X \right], \\ \nabla_{v_{21}} \mathcal{R}(\theta) = \mathbb{E} \left[(\sigma(f_\theta(X)) - \eta(X)) (\Gamma U_{11}^\top X + m_{-1} + \Delta m \sigma(s_0 + (2u_{12} + U_{11} \Delta m)^\top X)) \right], \\ \partial_{v_{22}} \mathcal{R}(\theta) = \mathbb{E} \left[(\sigma(f_\theta(X)) - \eta(X)) (-1 + 2\sigma(s_0 + (2u_{12} + U_{11} \Delta m)^\top X)) \right]. \end{cases}$$

This can be written as

$$\begin{cases} \nabla_{U_{11}} \mathcal{R}(\theta) = \nabla_w \tilde{R}(\zeta) (\Gamma^\top v_{21})^\top + \nabla_s \tilde{R}(\zeta) \Delta m^\top, \\ \nabla_{u_{12}} \mathcal{R}(\theta) = 2 \nabla_s \tilde{R}(\zeta), \\ \nabla_{v_{21}} \mathcal{R}(\theta) = \Gamma U_{11}^\top \nabla_w \tilde{R}(\zeta) + \nabla_b \tilde{R}(\zeta) m_{-1} + \nabla_r \tilde{R}(\zeta) \Delta m, \\ \partial_{v_{22}} \mathcal{R}(\theta) = -\nabla_b \tilde{R}(\zeta) + 2 \nabla_r \tilde{R}(\zeta). \end{cases}$$

Proof. For the logistic loss, Lemma B.7 gives

$$\mathbb{E} [\ell'(Y, f_\theta(X)) | X] = \sigma(f_\theta(X)) - \eta(X)$$

Hence, for any parameter component α ,

$$\nabla_\alpha \mathcal{R}(\theta) = \mathbb{E} [(\sigma(f_\theta(X)) - \eta(X)) \nabla_\alpha f_\theta(X)].$$

The derivatives of f_θ are

$$\begin{aligned} \nabla_{U_{11}} f_\theta(x) &= x(\Gamma^\top v_{21})^\top + (v_{21}^\top \Delta m + 2v_{22}) \sigma'(s_0 + (2u_{12} + U_{11} \Delta m)^\top x) x \Delta m^\top, \\ \nabla_{u_{12}} f_\theta(x) &= 2(v_{21}^\top \Delta m + 2v_{22}) \sigma'(s_0 + (2u_{12} + U_{11} \Delta m)^\top x) x, \\ \nabla_{v_{21}} f_\theta(x) &= \Gamma U_{11}^\top x + m_{-1} + \Delta m \sigma(s_0 + (2u_{12} + U_{11} \Delta m)^\top x), \end{aligned}$$

and

$$\partial_{v_{22}} f_\theta(x) = -1 + 2\sigma(s_0 + (2u_{12} + U_{11} \Delta m)^\top x).$$

Plugging these identities into the previous formula gives the result. $\square$

Lemma B.11 (Second-order expansion of the risk). *Let*

$$\mathcal{R}(\theta) = \mathbb{E} [\ell(Y, f_\theta(X))],$$

where ℓ is the logistic loss, and recall that $\eta(x) = \mathbb{P}(Y = 1 | X = x)$. Then, for any direction h , one has

$$D^2 \mathcal{R}(\theta)[h, h] = \mathbb{E} \left[\sigma(f_\theta(X)) (1 - \sigma(f_\theta(X))) (Df_\theta(X)[h])^2 + (\sigma(f_\theta(X)) - \eta(X)) D^2 f_\theta(X)[h, h] \right].$$

Proof. The first and second variations of $\mathcal{R}$ are given by the chain rule:

$$D\mathcal{R}(\theta)[h] = \mathbb{E} \left[\ell'(Y, f_\theta(X)) Df_\theta(X)[h] \right],$$

and

$$D^2 \mathcal{R}(\theta)[h, h] = \mathbb{E} \left[\ell''(Y, f_\theta(X)) (Df_\theta(X)[h])^2 + \ell'(Y, f_\theta(X)) D^2 f_\theta(X)[h, h] \right].$$

Lemma B.7 gives

$$\mathbb{E} [\ell'(Y, f_\theta(X)) | X] = \sigma(f_\theta(X)) - \eta(X),$$

so

$$\mathbb{E} [\ell''(Y, f_\theta(X)) | X] = \sigma(f_\theta(X)) (1 - \sigma(f_\theta(X))).$$

Substituting these identities in the previous expression gives the result. $\square$

Curvature at a Bayes realization. At an optimal predictor, the probability discrepancy vanishes. The Hessian therefore reduces to a weighted second moment of the first variation of the predictor. On $\mathcal{E}_1$, this reveals exactly which perturbations are invisible to second order and explains why the condition $s \neq 0$ is needed for the local convergence argument.

Proposition B.12. *Let $\mathcal{E}_1$ be as defined in Lemma B.17, and let*

$$\zeta^* = (w^*, b^*, 0, s^*) \in \mathcal{E}_1, \quad s^* \in \mathbb{R}^d,$$

where w^* and b^* are the coefficients of the Bayes score $f^*(x) = \langle w^*, x \rangle + b^*$. Then the Hessian of $\tilde{\mathcal{R}}$ at ζ^* is given by

$$\nabla^2 \tilde{\mathcal{R}}(\zeta^*) = \mathbb{E} \left[\eta(X)(1 - \eta(X)) \nabla g_{\zeta^*}(X) \nabla g_{\zeta^*}(X)^\top \right],$$

where

$$\nabla g_{\zeta^*}(X) = \begin{pmatrix} X \\ 1 \\ \sigma(s_0 + s^{*\top} X) \\ 0 \end{pmatrix}.$$

Equivalently, for every

$$\delta\zeta = (\delta w, \delta b, \delta r, \delta s), \quad \delta\zeta' = (\delta w', \delta b', \delta r', \delta s'),$$

one has

$$D^2 \tilde{\mathcal{R}}(\zeta^*)(\delta\zeta, \delta\zeta') = \mathbb{E} \left[\eta(X)(1 - \eta(X)) (\delta w^\top X + \delta b + \delta r \sigma(s_0 + s^{*\top} X)) (\delta w'^\top X + \delta b' + \delta r' \sigma(s_0 + s^{*\top} X)) \right].$$

In particular, the Hessian does not depend on the components δs and $\delta s'$.

Proof. Let $\zeta^* = (w^*, b^*, 0, s^*) \in \mathcal{E}_1$ and write $g_\zeta(X) = g_{w,b,r,s}(X)$. By differentiation under the expectation and the chain rule,

$$\nabla^2 \tilde{\mathcal{R}}(\zeta) = \mathbb{E} \left[\ell''(Y, g_\zeta(X)) \nabla g_\zeta(X) \nabla g_\zeta(X)^\top + \ell'(Y, g_\zeta(X)) \nabla^2 g_\zeta(X) \right].$$

By Lemma B.7,

$$\mathbb{E}[\ell'(Y, g_\zeta(X)) \mid X] = \sigma(g_\zeta(X)) - \eta(X).$$

Since $\zeta^* \in \mathcal{E}_1$, we have $g_{\zeta^*}(X) = f^*(X)$ and hence $\sigma(g_{\zeta^*}(X)) = \sigma(f^*(X)) = \eta(X)$. Therefore, $\mathbb{E}[\ell'(Y, g_{\zeta^*}(X)) \mid X] = 0$, and the second term vanishes:

$$\mathbb{E}[\ell'(Y, g_{\zeta^*}(X)) \nabla^2 g_{\zeta^*}(X)] = 0.$$

Next, as in Lemma B.11, we get

$$\mathbb{E}[\ell''(Y, g_\zeta(X)) \mid X] = \sigma(g_\zeta(X))(1 - \sigma(g_\zeta(X))).$$

Thus, since $g_{\zeta^*}(X) = f^*(X)$ and $\sigma(f^*(X)) = \eta(X)$, we get

$$\ell''(Y, g_{\zeta^*}(X)) = \eta(X)(1 - \eta(X)).$$

Hence

$$\nabla^2 \tilde{\mathcal{R}}(\zeta^*) = \mathbb{E} \left[\eta(X)(1 - \eta(X)) \nabla g_{\zeta^*}(X) \nabla g_{\zeta^*}(X)^\top \right].$$

Finally, since $\zeta^* \in \mathcal{E}_1$, we have $r^* = 0$, so

$$\nabla_s g_{\zeta^*}(X) = r^* \sigma(s_0 + s^{*\top} X)(1 - \sigma(s_0 + s^{*\top} X))X = 0.$$

Therefore,

$$\nabla g_{\zeta^*}(X) = \begin{pmatrix} X \\ 1 \\ \sigma(s_0 + s^{*\top} X) \\ 0 \end{pmatrix},$$

□

Lemma B.13. *Let*

$$\zeta^* = (w^*, b^*, 0, s^*) \in \mathcal{E}_1.$$

Then

$$\ker \nabla^2 \tilde{\mathcal{R}}(\zeta^*) = \begin{cases} \{(0, 0, 0, \delta s) : \delta s \in \mathbb{R}^d\}, & \text{if } s^* \neq 0, \\ \{(0, -\delta r \sigma(s_0), \delta r, \delta s) : (\delta r, \delta s) \in \mathbb{R} \times \mathbb{R}^d\}, & \text{if } s^* = 0. \end{cases}$$

Proof. Recall that for every $\delta\zeta = (\delta w, \delta b, \delta r, \delta s)$ and $\delta\zeta' = (\delta w', \delta b', \delta r', \delta s')$, we have

$$D^2 \tilde{\mathcal{R}}(\zeta^*)(\delta\zeta, \delta\zeta') = \mathbb{E} \left[\eta(X)(1 - \eta(X)) (\delta w^\top X + \delta b + \delta r \sigma(s_0 + s^{*\top} X)) (\delta w'^\top X + \delta b' + \delta r' \sigma(s_0 + s^{*\top} X)) \right].$$

Hence, if $\delta\zeta \in \ker \nabla^2 \tilde{\mathcal{R}}(\zeta^*)$, then

$$D^2 \tilde{\mathcal{R}}(\zeta^*)(\delta\zeta, \delta\zeta) = 0,$$

so that

$$\mathbb{E} \left[\eta(X)(1 - \eta(X)) (\delta w^\top X + \delta b + \delta r \sigma(s_0 + s^{*\top} X))^2 \right] = 0.$$

Since $\eta(X)(1 - \eta(X)) > 0$ almost surely, this implies

$$\delta w^\top X + \delta b + \delta r \sigma(s_0 + s^{*\top} X) = 0 \quad \text{a.s.}$$

If $s^* \neq 0$, since X is a non-degenerate Gaussian mixture, and $x \mapsto \sigma(s_0 + s^{*\top} x)$ is not affine when $s^* \neq 0$, this forces

$$\delta w = 0, \quad \delta b = 0, \quad \delta r = 0.$$

Conversely, every vector of the form $(0, 0, 0, \delta s)$ clearly belongs to the kernel, since the Hessian does not depend on δs . Therefore

$$\ker \nabla^2 \tilde{\mathcal{R}}(\zeta^*) = \{(0, 0, 0, \delta s) : \delta s \in \mathbb{R}^d\}.$$

If $s^* = 0$, then

$$\sigma(s_0 + s^{*\top} x) = \sigma(s_0),$$

and the condition becomes $\delta w^\top X + \delta b + \delta r \sigma(s_0) = 0$ almost surely. Thus, we obtain

$$\delta w = 0, \quad \delta b = -\delta r \sigma(s_0).$$

Conversely, every vector of the form

$$(0, -\delta r \sigma(s_0), \delta r, \delta s)$$

belongs to the kernel. This proves the second case. $\square$

Returning to attention parameters. The effective coordinates expose the predictor, but training evolves the original matrix and vector parameters. The differential of Φ connects these two descriptions. In particular, the chain rule below identifies the rank condition needed to transfer stationarity from the original risk to the effective risk.

Lemma B.14. *Let*

$$\theta = (U_{11}, u_{12}, v_{21}, v_{22}) \in \Theta, \quad \delta\theta = (\delta U_{11}, \delta u_{12}, \delta v_{21}, \delta v_{22}) \in \Theta.$$

Recall that

$$\Phi : \Theta \rightarrow \mathbb{R}^d \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d, \quad \Phi(\theta) = (w(\theta), b(\theta), r(\theta), s(\theta)),$$

with

$$w = U_{11} \Gamma v_{21}, \quad b = v_{21}^\top m_{-1} - v_{22}, \quad r = v_{21}^\top (m_1 - m_{-1}) + 2v_{22}, \quad s = 2u_{12} + U_{11}(m_1 - m_{-1}).$$

Then

$$D\Phi(\theta)[\delta\theta] = (\delta w, \delta b, \delta r, \delta s),$$

where

$$\begin{aligned}\delta w &= \delta U_{11} \Gamma v_{21} + U_{11} \Gamma \delta v_{21}, \\ \delta b &= \delta v_{21}^\top m_{-1} - \delta v_{22}, \\ \delta r &= \delta v_{21}^\top (m_1 - m_{-1}) + 2 \delta v_{22},\end{aligned}$$

and

$$\delta s = 2 \delta u_{12} + \delta U_{11} (m_1 - m_{-1}).$$

Proof. Direct computation. □

Lemma B.15.

$$\nabla \mathcal{R}(\theta) = D\Phi(\theta)^* (\nabla \tilde{\mathcal{R}}(\Phi(\theta))).$$

The singular set. When $v_{21} = 0$, the differential of Φ loses rank, so stationarity of the original risk need not imply stationarity of the effective risk. The next lemma identifies an effective descent direction that survives at these exceptional critical points. It is the key input to the strict-saddle construction in Appendix B.5.

Lemma B.16 (Non-vanishing of the reduced w -gradient). *Let $\theta_{\text{sp}} \in \text{Crit}(\mathcal{R}) \cap \{v_{21} = 0\}$. Assume that $m_1 + m_{-1} \neq 0$. Then*

$$\nabla_w \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) \neq 0.$$

Proof. We argue by contradiction, assume that $\nabla_w \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) = 0$.

Since θ_{sp} is a critical point of $\mathcal{R}$, using the expression of $\nabla \mathcal{R}$ in terms of $\nabla \tilde{\mathcal{R}}$ in lemma B.10 the criticality identities give

$$2\partial_r \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) - \partial_b \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) = 0,$$

and

$$\Gamma U_{11}^\top \nabla_w \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) + m_{-1} \partial_b \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) + \Delta m \partial_r \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) = 0.$$

Using $\nabla_w \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) = 0$, we obtain

$$2m_{-1} \partial_r \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) + \Delta m \partial_r \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) = 0.$$

This becomes

$$(m_1 + m_{-1}) \partial_r \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) = 0.$$

By the assumption $m_1 + m_{-1} \neq 0$, we get $\partial_r \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) = 0$. Therefore,

$$\partial_b \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) = 0.$$

Now, since $v_{21} = 0$ at θ_{sp} , we have

$$f_{\theta_{\text{sp}}}(x) = -v_{22} + 2v_{22}\sigma(s_0 + s(\theta_{\text{sp}})^\top x).$$

Hence

$$\begin{aligned}\mathbb{E}[(\sigma(f_{\theta_{\text{sp}}}(X)) - \eta(X))f_{\theta_{\text{sp}}}(X)] &= -v_{22}\mathbb{E}[\sigma(f_{\theta_{\text{sp}}}(X)) - \eta(X)] \\ &\quad + 2v_{22}\mathbb{E}[(\sigma(f_{\theta_{\text{sp}}}(X)) - \eta(X)) \\ &\quad \quad \quad \cdot \sigma(s_0 + s(\theta_{\text{sp}})^\top X)] \\ &= v_{22}(-\partial_b \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) + 2\partial_r \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}}))) \\ &= 0.\end{aligned}$$

Moreover, since

$$f^*(x) = w^{\star\top} x + b^*,$$

we also have

$$\mathbb{E}[(\sigma(f_{\theta_{\text{sp}}}(X)) - \eta(X))f^*(X)] = w^{\star\top} \nabla_w \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) + b^* \partial_b \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) = 0.$$

Using $\eta(x) = \sigma(f^*(x))$, we obtain

$$\mathbb{E} \left[(\sigma(f_{\theta_{\text{sp}}}(X)) - \sigma(f^*(X))) (f_{\theta_{\text{sp}}}(X) - f^*(X)) \right] = 0.$$

Since σ is strictly increasing, the integrand is nonnegative and vanishes if and only if

$$f_{\theta_{\text{sp}}}(X) = f^*(X) \quad \text{a.s.}$$

Moreover no function of the form $x \mapsto c_1 + c_2 \sigma(s_0 + s^\top x)$ coincides almost surely with f^* so that last equality is impossible. Thus, our assumption must be false and

$$\nabla_w \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) \neq 0.$$

□

B.5 Optimization landscape (proofs of Theorems 3.1 and 3.3, Lemma 3.2, and Proposition 3.4)

We now prove Theorems 3.1 and 3.3, Lemma 3.2, and Proposition 3.4. The argument proceeds from the predictor to its parameterization: we identify all effective Bayes realizations, show that every effective stationary point is optimal, and determine what changes under the attention factorization. We then address a separate issue—loss of compactness—which explains why these landscape results alone do not establish global convergence.

Each main-text result has an explicitly identified proof below. The proof of Theorem 3.1 combines the characterization of the effective minimizers with effective stationarity. The proof of Theorem 3.3 combines the preimage description with the regular-point and strict-saddle lemmas. The proofs of Lemma 3.2 and Proposition 3.4 are given directly, after restating their claims for use elsewhere in the appendix.

Recall that $g_{w,b,r,s}(x) = w^\top x + b + r\sigma(s_0 + s^\top x)$ and $\mathcal{R} = \tilde{\mathcal{R}} \circ \Phi$. The coordinates (w, b) describe the affine part, while r and s control the amplitude and direction of the sigmoid. This decomposition makes the two ways of representing an affine Bayes score transparent.

B.5.1 Bayes minimizers (Theorems 3.1 and 3.3, minimizer claims)

Two mechanisms for attaining the Bayes risk. By Lemma B.4, a parameter is optimal exactly when its predictor equals f^* . This can happen by setting $r = 0$, which removes the sigmoid, or by setting $s = 0$, which makes it constant and allows its contribution to be absorbed into the bias. The following lemma shows that these mechanisms exhaust the effective minimizers.

Lemma B.17 (Effective minimizers). *The minimizer set of $\tilde{\mathcal{R}}$ is*

$$\begin{aligned} \mathcal{E}^* &= \mathcal{E}_1 \cup \mathcal{E}_2, \quad \mathcal{E}_1 = \{(w^*, b^*, 0, s) : s \in \mathbb{R}^d\}, \\ \mathcal{E}_2 &= \{(w^*, b^* - r\sigma(s_0), r, 0) : r \in \mathbb{R}\}. \end{aligned}$$

Proof of Lemma B.17. By Lemma B.4, the set of minimizers of $\tilde{\mathcal{R}}$ is given by

$$\mathcal{E}^* = \left\{ (w, b, r, s) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d : w^\top x + b + r\sigma(s_0 + s^\top x) = f^*(x) \quad \text{a.s.} \right\},$$

where the Bayes score is $f^*(x) = w^{*\top} x + b^*$, $w^* = \Gamma^{-1}(m_1 - m_{-1})$. Therefore, if $(w, b, r, s) \in \mathcal{E}^*$, then

$$w^\top x + b + r\sigma(s_0 + s^\top x) = w^{*\top} x + b^* \quad \text{a.s.}$$

This is equivalent to saying that one of the following two situations occurs:

- either

$$w = w^*, \quad s = 0, \quad b + r\sigma(s_0) = b^*,$$

that is,

$$w = w^*, \quad s = 0, \quad b = b^* - r\sigma(s_0).$$

We denote this set by

$$\mathcal{E}_2 = \left\{ (w^*, b^* - r\sigma(s_0), r, 0) : r \in \mathbb{R} \right\}.$$

• or

$$w = w^*, \quad b = b^*, \quad r = 0, \quad s \in \mathbb{R}^d \text{ arbitrary.}$$

We denote this set by

$$\mathcal{E}_1 = \{(w^*, b^*, 0, s) : s \in \mathbb{R}^d\}.$$

Hence,

$$\mathcal{E}^* = \mathcal{E}_1 \cup \mathcal{E}_2 = \{(w^*, b^*, 0, s) : s \in \mathbb{R}^d\} \cup \{(w^*, b^* - r \sigma(s_0), r, 0) : r \in \mathbb{R}\}.$$

□

The corresponding attention minimizers are obtained by taking preimages under Φ . This step describes the optimal set without assuming that every effective parameter is attainable. In particular, the existence of a branch in effective coordinates does not by itself guarantee that its preimage is nonempty; the realizability discussion in Appendix B.2 addresses this distinction.

Corollary B.18 (Minimizers in attention coordinates). *The minimizers of $\mathcal{R}$ satisfy $\Theta^* = \Theta_1 \cup \Theta_2$, where $\Theta_i = \Phi^{-1}(\mathcal{E}_i)$.*

Proof of Corollary B.18. Since $\mathcal{R} = \tilde{\mathcal{R}} \circ \Phi$, we have

$$\theta \in \operatorname{argmin} \mathcal{R} \iff \tilde{\mathcal{R}}(\Phi(\theta)) = \min \mathcal{R}.$$

Using $\min \mathcal{R} = \min \tilde{\mathcal{R}}$, this is equivalent to

$$\Phi(\theta) \in \operatorname{argmin} \tilde{\mathcal{R}} = \mathcal{E}^*.$$

Hence

$$\operatorname{argmin} \mathcal{R} = \Phi^{-1}(\mathcal{E}^*) = \underbrace{\Phi^{-1}(\mathcal{E}_1)}_{\Theta_1} \cup \underbrace{\Phi^{-1}(\mathcal{E}_2)}_{\Theta_2}.$$

□

Why effective stationarity implies optimality. The reduced risk is not jointly convex in (w, b, r, s) . Nevertheless, fixing s leaves a convex problem in (w, b, r) , and this restricted problem still contains the Bayes predictor through the choice $r = 0$. Thus stationarity in the affine coefficients and sigmoid amplitude already forces the risk to attain its global minimum.

B.5.2 Effective stationary points (completion of Theorem 3.1)

Lemma B.19 (No spurious effective critical points). $\operatorname{Crit}(\tilde{\mathcal{R}}) = \operatorname{argmin} \tilde{\mathcal{R}}$.

Proof of Lemma B.19. For each fixed $s \in \mathbb{R}^d$, we define the partial risk

$$\tilde{\mathcal{R}}_s(w, b, r) := \tilde{\mathcal{R}}(w, b, r, s) = \mathbb{E} \left[\ell_{\log}(Y, w^\top X + b + r \sigma(s_0 + s^\top X)) \right].$$

Then, $\tilde{\mathcal{R}}_s$ is convex as a function of $(w, b, r) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}$ (as the expectation of a convex function).

So its critical points coincide with its global minimizers. Moreover, $\min_{(w', b', r')} \tilde{\mathcal{R}}_s(w', b', r') \geq \mathcal{R}^*$ and the choice $(w, b, r) = (w^*, b^*, 0)$ yields $\mathcal{R}_s(w, b, r) = \mathcal{R}^*$. This leads to

$$\min_{(w', b', r')} \tilde{\mathcal{R}}_s(w', b', r') = \mathcal{R}^*.$$

Let now $(w, b, r, s) \in \operatorname{Crit}(\tilde{\mathcal{R}})$. Then, in particular,

$$\nabla_w \tilde{\mathcal{R}}(w, b, r, s) = 0, \quad \partial_b \tilde{\mathcal{R}}(w, b, r, s) = 0, \quad \partial_r \tilde{\mathcal{R}}(w, b, r, s) = 0.$$

Equivalently,

$$\nabla_{(w, b, r)} \tilde{\mathcal{R}}_s(w, b, r) = 0.$$

Hence (w, b, r) is a critical point of the convex function $\tilde{\mathcal{R}}_s$, and therefore a global minimizer of $\tilde{\mathcal{R}}_s$. It follows that

$$\tilde{\mathcal{R}}(w, b, r, s) = \tilde{\mathcal{R}}_s(w, b, r) = \min_{(w', b', r')} \tilde{\mathcal{R}}_s(w', b', r') = \mathcal{R}^*.$$

which shows that

$$\text{Crit}(\tilde{\mathcal{R}}) \subset \text{argmin } \tilde{\mathcal{R}}.$$

The reverse inclusion is immediate: since $\tilde{\mathcal{R}}$ is differentiable on an open set, every global minimizer is a critical point. We conclude that

$$\text{Crit}(\tilde{\mathcal{R}}) = \text{argmin } \tilde{\mathcal{R}}.$$

□

Proof of Theorem 3.1. Lemma B.17 identifies the minimizer set as $\mathcal{E}_1 \cup \mathcal{E}_2$, and Lemma B.19 shows that it coincides with $\text{Crit}(\tilde{\mathcal{R}})$. Together, these give both claims of the theorem. □

From effective to attention parameters. The favorable effective landscape does not automatically transfer to the attention parameters. The chain rule gives $\nabla \mathcal{R}(\theta) = D\Phi(\theta)^* \nabla \tilde{\mathcal{R}}(\Phi(\theta))$: a singular differential may annihilate a nonzero effective gradient. The next result identifies precisely where this obstruction is absent. Here, asymmetry means $m_1 \neq -m_{-1}$, rather than merely distinct class means.

B.5.3 Regularity of the parameter map (proof of Lemma 3.2)

Lemma B.20 (Regularity of the reparametrization). *The adjoint $D\Phi(\theta)^*$ is injective if and only if $m_1 \neq -m_{-1}$ and $v_{21} \neq 0$.*

Proof of Lemma 3.2. This is the statement restated in Lemma B.20. Let

$$\delta\theta = (\delta U_{11}, \delta u_{12}, \delta v_{21}, \delta v_{22}) \in \Theta,$$

and let

$$(\delta w, \delta b, \delta r, \delta s) = D\Phi(\theta)[\delta\theta].$$

By Lemma B.14, this means that

$$\delta w = \delta U_{11} \Gamma v_{21} + U_{11} \Gamma \delta v_{21}, \tag{13}$$

$$\delta b = \delta v_{21}^\top m_{-1} - \delta v_{22}, \tag{14}$$

$$\delta r = \delta v_{21}^\top (m_1 - m_{-1}) + 2\delta v_{22}, \tag{15}$$

$$\delta s = 2\delta u_{12} + \delta U_{11} (m_1 - m_{-1}). \tag{16}$$

Proof of the necessary conditions. Assume first that $m_1 = -m_{-1}$. Then $m_1 - m_{-1} = -2m_{-1}$, and combining (14) and (15) yields

$$\delta r = -2\delta v_{21}^\top m_{-1} + 2\delta v_{22} = -2\delta b.$$

Hence every vector in the image of $D\Phi(\theta)$ satisfies

$$2\delta b + \delta r = 0,$$

so $D\Phi(\theta)$ cannot be surjective.

Assume now that $v_{21} = 0$. Then (13) becomes

$$\delta w = U_{11} \Gamma \delta v_{21}.$$

Therefore $(\delta w, \delta b, \delta r)$ depend only on $(\delta v_{21}, \delta v_{22})$ through a linear map

$$h : \mathbb{R}^{d+1} \rightarrow \mathbb{R}^{d+2}.$$

Since $\dim(\mathbb{R}^{d+1}) < \dim(\mathbb{R}^{d+2})$, the map h cannot be surjective, and neither can $D\Phi(\theta)$.

This proves the necessity of the two conditions.

Proof of the sufficient conditions. Conversely, assume that

$$m_1 \neq -m_{-1} \quad \text{and} \quad v_{21} \neq 0.$$

Let us fix

$$(\delta w, \delta b, \delta r, \delta s) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$$

be arbitrary. We show that the system (13) - (16) can be solved.

First, using (14) and (15), we want to choose δv_{21} and δv_{22} so that

$$\delta b = \delta v_{21}^\top m_{-1} - \delta v_{22}, \quad \delta r = \delta v_{21}^\top (m_1 - m_{-1}) + 2\delta v_{22}.$$

Adding the second equation to twice the first one gives

$$\delta v_{21}^\top (m_1 + m_{-1}) = \delta r + 2\delta b.$$

Since $m_1 + m_{-1} \neq 0$, one can choose $\delta v_{21} \in \mathbb{R}^d$ such that

$$\delta v_{21}^\top (m_1 + m_{-1}) = \delta r + 2\delta b.$$

Then define

$$\delta v_{22} := \delta v_{21}^\top m_{-1} - \delta b.$$

With this choice, both (14) and (15) are satisfied.

Next, we solve (13). Since $v_{21} \neq 0$ and Γ is invertible, we have $\Gamma v_{21} \neq 0$. So there exists a matrix $\delta U_{11} \in \mathbb{R}^{d \times d}$ such that

$$\delta U_{11} \Gamma v_{21} = \delta w - U_{11} \Gamma \delta v_{21}.$$

With this choice, (13) holds.

Finally, once δU_{11} is fixed, we define

$$\delta u_{12} = \frac{1}{2} \left(\delta s - \delta U_{11} (m_1 - m_{-1}) \right),$$

which ensures (16).

This proves that every $(\delta w, \delta b, \delta r, \delta s)$ belongs to the image of $D\Phi(\theta)$, so $D\Phi(\theta)$ is surjective. $\square$

Away from the singular set $\{v_{21} = 0\}$, injectivity of the adjoint differential transfers stationarity to the effective risk. The preceding characterization then rules out any nonoptimal critical point in this regular region.

B.5.4 Critical points and strict saddles (completion of Theorem 3.3)

Lemma B.21 (Critical points away from the singular set). *If $m_1 \neq -m_{-1}$, then every critical point of $\mathcal{R}$ with $v_{21} \neq 0$ is globally minimizing.*

Proof of Lemma B.21. We assume that $m_1 \neq -m_{-1}$. Let $\theta \in \text{Crit}(\mathcal{R})$ and assume that $v_{21} \neq 0$. Since

$$\nabla \mathcal{R}(\theta) = D\Phi(\theta)^* \nabla \tilde{\mathcal{R}}(\Phi(\theta)),$$

we obtain

$$D\Phi(\theta)^* \nabla \tilde{\mathcal{R}}(\Phi(\theta)) = 0.$$

By Lemma B.20, $D\Phi(\theta)$ is surjective so $D\Phi(\theta)^*$ is injective, and then

$$\nabla \tilde{\mathcal{R}}(\Phi(\theta)) = 0.$$

By Lemma B.19, this implies that

$$\Phi(\theta) \in \text{argmin } \tilde{\mathcal{R}},$$

Since $\mathcal{R}(\theta) = \tilde{\mathcal{R}}(\Phi(\theta))$, we get

$$\mathcal{R}(\theta) = R^*,$$

so that $\theta \in \operatorname{argmin} \mathcal{R}$.

This shows that every critical point θ with $v_{21} \neq 0$ is a global minimizer of $\mathcal{R}$. Equivalently,

$$\operatorname{Crit}(\mathcal{R}) \subset \operatorname{argmin} \mathcal{R} \sqcup \left(\operatorname{Crit}(\mathcal{R}) \cap \{\theta \in \Theta : v_{21} = 0\} \right).$$

The reverse inclusion is immediate, which concludes the proof. $\square$

The exceptional critical points are strict saddles. It remains to examine $\{v_{21} = 0\}$, where the transfer argument fails. Lemma B.16 supplies a nonzero effective w -gradient at these points. By coupling variations of the two factors while keeping the gate fixed to first order, the proof turns this gradient into directions of opposite curvature. The attention parameterization therefore introduces saddles, but no additional local minima.

Lemma B.22 (Strict-saddle property). *Assume $m_1 \neq -m_{-1}$. Every nonoptimal critical point of $\mathcal{R}$ admits both a positive- and a negative-curvature direction. In particular, every critical point is either a global minimizer or a strict saddle.*

Proof of Lemma B.22. Let $\theta_{\text{sp}} \in \operatorname{Crit} \mathcal{R} \cap \{v_{21} = 0\}$. Then

$$f_{\theta_{\text{sp}}}(x) = -v_{22} + 2v_{22}\sigma(s_0 + s(\theta_{\text{sp}})^\top x).$$

We choose a direction h such that

$$\delta u_{12} = -\frac{1}{2}\delta U_{11}\Delta m, \quad \delta v_{22} = 0.$$

So that, at θ_{sp} , we get (using Lemma B.8)

$$Df_{\theta_{\text{sp}}}(x)[h] = \delta v_{21}^\top \left(\Gamma U_{11}^\top x + m_{-1} + \Delta m \sigma(s_0 + s(\theta_{\text{sp}})^\top x) \right),$$

and

$$D^2 f_{\theta_{\text{sp}}}(x)[h, h] = 2\delta v_{21}^\top \Gamma \delta U_{11}^\top x.$$

By Lemma B.11 (and Lemma B.9), we can deduce that

$$\begin{aligned} D^2 \mathcal{R}(\theta_{\text{sp}})[h, h] &= \mathbb{E} \left[\sigma(f_\theta(X))(1 - \sigma(f_\theta(X))) (Df_\theta(X)[h])^2 + (\sigma(f_\theta(X)) - \eta(X)) D^2 f_\theta(X)[h, h] \right] \\ &= C + 2\delta v_{21}^\top \Gamma \delta U_{11}^\top \nabla_w \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) \end{aligned}$$

where C does not depend on δU_{11} .

Using Lemma B.16, we know that $\nabla_w \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) \neq 0$. We can thus choose a fixed δv_{21} such that $\Gamma \delta v_{21} \neq 0$ and set

$$\delta U_{11}^{(t)} = -t \nabla_w \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) (\Gamma \delta v_{21})^\top, \quad t > 0.$$

Then

$$\delta v_{21}^\top \Gamma (\delta U_{11}^{(t)})^\top \nabla_w \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) = -t \left\| \nabla_w \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) \right\|^2 \|\Gamma \delta v_{21}\|^2.$$

Thus

$$D^2 \mathcal{R}(\theta_{\text{sp}})[h_t, h_t] = C - 2t \left\| \nabla_w \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) \right\|^2 \|\Gamma \delta v_{21}\|^2.$$

For t large enough,

$$D^2 \mathcal{R}(\theta_{\text{sp}})[h_t, h_t] < 0.$$

Taking

$$\delta U_{11}^{(t)} = +t \nabla_w \tilde{\mathcal{R}}(\Phi(\theta_{\text{sp}})) (\Gamma \delta v_{21})^\top, \quad t > 0.$$

leads to

$$D^2 \mathcal{R}(\theta_{\text{sp}})[h_t, h_t] > 0.$$

for t large enough. Hence θ_{sp} is a strict saddle point. $\square$

Proof of Theorem 3.3. Corollary B.18 gives $\Theta^* = \Theta_1 \cup \Theta_2$. Under $m_1 \neq -m_{-1}$, Lemma B.21 shows that every critical point outside $\{v_{21} = 0\}$ is globally optimal, while Lemma B.22 supplies a negative-curvature direction at every nonoptimal critical point. Such a point cannot be a local minimum. Hence every critical point is either globally minimizing or a strict saddle, and all local minima are global. $\square$

Why a benign landscape is not a global convergence theorem. We have classified finite critical points, but this does not bound a gradient-flow trajectory. The next construction exhibits noncompact near-optimal directions: as the sigmoid amplitude grows and its direction shrinks, its constant and linear contributions are canceled by the affine part. The predictor approaches f^* even though the effective parameters diverge.

B.5.5 Minimizing sequences at infinity (proof of Proposition 3.4)

Proposition B.23 (Minimizers at infinity). *For $c \in \mathbb{R}^d$ and $R > 0$, let*

$$\zeta_R = (w^* - \sigma'(s_0)c, b^* - \sigma(s_0)R, R, c/R).$$

Then $g_{\zeta_R} \rightarrow f^*$ in $L^1(\mu_X)$ and $\tilde{\mathcal{R}}(\zeta_R) \rightarrow \mathcal{R}^*$ as $R \rightarrow \infty$.

Proof of Proposition 3.4. This is the statement restated in Proposition B.23. Let $c \in \mathbb{R}^d$, and for $r > 0$ define

$$\zeta_R = \left(w^* - \sigma'(s_0)c, b^* - \sigma(s_0)R, R, \frac{c}{R} \right).$$

For $x \in \mathbb{R}^d$, we have

$$g_{\zeta_R}(x) = (w^* - \sigma'(s_0)c)^\top x + b^* - \sigma(s_0)R + R\sigma\left(s_0 + \frac{c^\top x}{R}\right).$$

Therefore

$$g_{\zeta_R}(x) - f^*(x) = R\left[\sigma\left(s_0 + \frac{c^\top x}{R}\right) - \sigma(s_0) - \sigma'(s_0)\frac{c^\top x}{R}\right].$$

Using Taylor expansion at order 2 around s_0 for fixed x , as $R \rightarrow +\infty$, we get

$$\sigma\left(s_0 + \frac{c^\top x}{R}\right) = \sigma(s_0) + \sigma'(s_0)\frac{c^\top x}{R} + \frac{1}{2}\sigma''(s_0)\frac{(c^\top x)^2}{R^2} + o\left(\frac{1}{R^2}\right).$$

Hence

$$g_{\zeta_R}(x) - f^*(x) = \frac{1}{2}\sigma''(s_0)\frac{(c^\top x)^2}{R} + o\left(\frac{1}{R}\right).$$

In particular,

$$g_{\zeta_R}(x) \rightarrow f^*(x) \quad \text{for every } x \in \mathbb{R}^d.$$

We now justify convergence in $L^1(\mu_X)$. Since the sigmoid is smooth and σ'' is bounded, Taylor's formula with integral remainder gives a constant $C > 0$ such that, for every $R \geq 1$ and every x ,

$$|g_{\zeta_R}(x) - f^*(x)| \leq \frac{C}{R}(c^\top x)^2 \leq C(c^\top x)^2$$

Since X is a mixture of non-degenerate Gaussian distributions, it has finite moments of all orders. In particular, $\mathbb{E}[(c^\top X)^2] < +\infty$. By dominated convergence, we obtain

$$\|g_{\zeta_R} - f^*\|_{L^1(\mu_X)} \rightarrow 0.$$

It remains to pass to the risk. The log-exp function is 1-Lipschitz. Hence

$$\left| \tilde{\mathcal{R}}(\zeta_R) - \mathcal{R}^* \right| = \left| \mathbb{E} [\log(1 + \exp(-Y g_{\zeta_R}(X))) - \log(1 + \exp(-Y f^*(X)))] \right| \leq \mathbb{E} [|g_{\zeta_R}(X) - f^*(X)|].$$

Since the right-hand side converges to 0, we obtain

$$\tilde{\mathcal{R}}(\zeta_R) \rightarrow \mathcal{R}^*.$$

This proves the claim. $\square$

The preceding results establish Theorems 3.1 and 3.3, Lemma 3.2, and Proposition 3.4. Their dynamical implication is deliberately limited: the absence of nonoptimal local minima does not supply the boundedness needed for a global convergence argument. Conversely, the escaping sequence above is not itself a gradient-flow trajectory and does not prove that such trajectories diverge. We therefore turn to regimes where the dynamics can be controlled: a neighborhood of the regular component Θ_1 , and an invariant subspace in the balanced symmetric setting.

B.6 Normal curvature (preparation for Theorem 3.5)

Throughout this subsection, assume $m_1 \neq -m_{-1}$. We now move from the classification of critical points to the geometry needed for local convergence. Since optimal parameters are not isolated, positive definiteness of the full Hessian is neither available nor desirable: directions tangent to a manifold of equivalent Bayes realizations must be flat. The relevant question is whether these are the only flat directions.

We apply the Morse–Bott lemma at a point

$$\theta^* \in \Theta_1^{\text{reg}} := \{\theta \in \Theta_1 : s(\theta) \neq 0\},$$

in order to study the local convergence of the gradient flow near this critical manifold.

The key point is that Θ_1^{reg} is a Morse–Bott critical manifold for the risk $\mathcal{R}$. In other words, at every point $\theta^* \in \Theta_1^{\text{reg}}$, the tangent space to the manifold coincides with the kernel of the Hessian:

$$T_{\theta^*}\Theta_1^{\text{reg}} = \ker \nabla^2 \mathcal{R}(\theta^*).$$

Equivalently, the Hessian is positive definite in the directions normal to the critical manifold. This means that, although the risk is flat along the manifold of minimizers, it has a quadratic behavior transversally to it.

The Morse–Bott lemma, recalled in Lemma B.25, then formalizes that, up to a local diffeomorphism, the risk $\mathcal{R}$ is locally equivalent to a quadratic function in the normal directions to Θ_1^{reg} . This normal form is the main ingredient in the proof of local convergence of the gradient flow toward the manifold of global minimizers (Theorem 3.5).

We first state the geometric criterion and its local normal form, adapted from Rebjock and Boumal (2025) and Banyaga and Hurtubise (2004) to our local setting.

Definition B.24 (Morse–Bott property at a point from Rebjock and Boumal 2025). *Let $f : M \rightarrow \mathbb{R}$ be a $\mathcal{C}^2$ function on a smooth manifold M , and let $\bar{x} \in M$ be a local minimum of f . Let*

$$S := \{x \in M : \nabla f(x) = 0, f(x) = f(\bar{x})\}$$

denote the associated critical set. We say that f satisfies the Morse–Bott property at $\bar{x}$ if, in a neighborhood of $\bar{x}$, the set S is a $\mathcal{C}^1$ submanifold and

$$\ker \nabla^2 f(\bar{x}) = T_{\bar{x}}S.$$

In other words, the Hessian is non-degenerate in the normal directions to the local critical submanifold.

Lemma B.25 (Morse–Bott lemma, local version adapted from Banyaga and Hurtubise 2004). *Let M be a smooth manifold of dimension m , let $f : M \rightarrow \mathbb{R}$ be a smooth function, and let $p \in M$ be a critical point of f . Assume that f satisfies the Morse–Bott property at p . Let C be the local critical submanifold through p , and let $n := \dim C$. Then there exist an open neighborhood U of p in M and a smooth chart*

$$\psi : U \rightarrow \mathbb{R}^n \times \mathbb{R}^{m-n}$$

such that:

- (a) $\psi(p) = 0$;
- (b) $\psi(U \cap C) = \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^{m-n} : y = 0\}$;
- (c) *there exists an integer $k \leq m - n$ such that*

$$(f \circ \psi^{-1})(x, y) = f(p) - y_1^2 - \cdots - y_k^2 + y_{k+1}^2 + \cdots + y_{m-n}^2.$$

Here k is the index of the Hessian of f at p restricted to the normal directions to C .

Proof. The proof can be found in [Banyaga and Hurtubise \(2004\)](#). The version we enounced can be deduced by considering the function f restricted to the neighborhood where it satisfies the Morse-Bott property. $\square$

We recall the following definitions :

$$\Theta_1 := \Phi^{-1}(\mathcal{E}_1), \quad \Theta_2 := \Phi^{-1}(\mathcal{E}_2).$$

$$\Theta_1^{\text{reg}} := \{\theta \in \Theta_1 : s(\theta) \neq 0\}.$$

Here, the condition $s(\theta) \neq 0$ will play an important role in the sequel, as it ensures the non-degeneracy of the Hessian at the point under consideration.

Lemma B.26. *Let $\theta^* \in \Theta_1^{\text{reg}}$, and let $\mathcal{S}$ denote the connected component of Θ_1^{reg} containing θ^* . Then $\mathcal{R}$ satisfies the Morse-Bott property at θ^* . More precisely:*

(i) $\mathcal{S}$ is a $\mathcal{C}^1$ submanifold in a neighborhood of θ^* ;

(ii)

$$\ker \nabla^2 \mathcal{R}(\theta^*) = T_{\theta^*} \mathcal{S}.$$

The proof has two parts. The preimage theorem first identifies Θ_1^{reg} as a smooth constraint manifold and computes its tangent space. The Hessian identities from [Appendix B.4](#) then show that precisely these tangent directions have zero curvature. We recall the differential-topological statement needed for the first step.

Definition B.27. *Let X and Y be smooth manifolds, and let $f : X \rightarrow Y$ be a smooth map. We say that a point $y \in Y$ is a regular value of f if for every $x \in f^{-1}(y)$, the differential $df_x : T_x X \rightarrow T_y Y$ is surjective.*

Lemma B.28 (Preimage theorem). *Let X and Y be smooth manifolds, let $f : X \rightarrow Y$ be a smooth map, and let $y \in Y$ be a regular value of f . Then the preimage $f^{-1}(y)$ is a smooth submanifold of X . If moreover $y \in \text{Im}(f)$, then the codimension of $f^{-1}(y)$ in X is equal to $\dim Y$. Furthermore, for every $x \in f^{-1}(y)$, the tangent space satisfies*

$$T_x(f^{-1}(y)) = \ker(df_x).$$

Proof of Lemma B.26. We first show that the connected component $\mathcal{S}$ of Θ_1^{reg} containing θ^* is a smooth submanifold. Let $\mathcal{O} := \{\theta \in \Theta : s(\theta) \neq 0\}$. Since $s : \Theta \rightarrow \mathbb{R}^d$ is continuous and $\mathbb{R}^d \setminus \{0\}$ is open, we have that $\mathcal{O} = s^{-1}(\mathbb{R}^d \setminus \{0\})$ is an open subset of Θ (and thus a smooth manifold).

Next, define

$$C : \Theta \rightarrow \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}, \quad C(\theta) = (w(\theta) - w^*, b(\theta) - b^*, r(\theta)).$$

By definition of Θ_1 , we have

$$\Theta_1 = \{\theta \in \Theta : w(\theta) = w^*, b(\theta) = b^*, r(\theta) = 0\} = C^{-1}(0).$$

Therefore,

$$\Theta_1^{\text{reg}} = \Theta_1 \cap \mathcal{O} = C^{-1}(0) \cap \mathcal{O} = (C|_{\mathcal{O}})^{-1}(0),$$

where $C|_{\mathcal{O}}$ denotes the restriction of C to the smooth open submanifold $\mathcal{O}$.

Let now $\theta \in \Theta_1^{\text{reg}} = (C|_{\mathcal{O}})^{-1}(0)$. Since $\mathcal{O}$ is open, we have $D(C|_{\mathcal{O}})(\theta) = DC(\theta)$. Moreover, $DC(\theta)$ is surjective. Indeed, C is obtained by keeping only the (w, b, r) -components of Φ , so $DC(\theta)$ is the composition of $D\Phi(\theta)$ with the canonical projection onto $\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}$. By [Lemma B.20](#), $D\Phi(\theta)$ is surjective (because $m_1 \neq -m_{-1}$ by assumption and $v_{21} \neq 0$ since $U_{11}\Gamma v_{21} = w = w^* \neq 0$). Therefore $DC(\theta)$ is surjective as well. Thus, 0 is a regular value of $C|_{\mathcal{O}}$.

By the preimage theorem [B.28](#) we can then deduce that $\Theta_1^{\text{reg}} = (C|_{\mathcal{O}})^{-1}(0)$ is a smooth submanifold of $\mathcal{O}$, hence of Θ . Its codimension is $\dim(\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}) = d + 2$. In particular, for every $\theta \in \Theta_1^{\text{reg}}$,

$$T_{\theta} \Theta_1^{\text{reg}} = \ker DC(\theta).$$

Finally, since $\mathcal{S}$ is a connected component of the smooth manifold Θ_1^{reg} , it is itself a smooth submanifold in a neighborhood of θ^* . In particular, $\text{codim } \mathcal{S} = d + 2$ and for every $\theta \in \mathcal{S}$,

$$T_\theta \mathcal{S} = \ker DC(\theta).$$

It remains to prove that

$$\ker \nabla^2 \mathcal{R}(\theta^*) = T_{\theta^*} \mathcal{S}.$$

We already know that $T_{\theta^*} \mathcal{S} = \ker DC(\theta^*)$, thus it is enough to show that

$$\ker \nabla^2 \mathcal{R}(\theta^*) = \ker DC(\theta^*).$$

Set $\zeta^* := \Phi(\theta^*) \in \mathcal{E}_1$. Since $\mathcal{R} = \tilde{\mathcal{R}} \circ \Phi$, the chain rule yields that for every $h, h' \in \Theta$,

$$D^2 \mathcal{R}(\theta^*)(h, h') = D^2 \tilde{\mathcal{R}}(\zeta^*)(D\Phi(\theta^*)h, D\Phi(\theta^*)h') + D\tilde{\mathcal{R}}(\zeta^*)(D^2 \Phi(\theta^*)(h, h')).$$

Now $\zeta^* = \Phi(\theta^*) \in \mathcal{E}^* = \text{argmin } \tilde{\mathcal{R}}$, we have $D\tilde{\mathcal{R}}(\zeta^*) = 0$ and therefore,

$$D^2 \mathcal{R}(\theta^*)(h, h') = D^2 \tilde{\mathcal{R}}(\zeta^*)(D\Phi(\theta^*)h, D\Phi(\theta^*)h').$$

We now prove the first inclusion

$$\ker DC(\theta^*) \subset \ker \nabla^2 \mathcal{R}(\theta^*).$$

Let $h \in \ker DC(\theta^*)$. The condition $DC(\theta^*)h = 0$ means that the first three components of $D\Phi(\theta^*)h$ vanish. Hence

$$D\Phi(\theta^*)h = (0, 0, 0, \delta s)$$

for some $\delta s \in \mathbb{R}^d$. By Lemma B.13, since $\theta^* \in \Theta_1^{\text{reg}}$ and thus $s(\theta^*) \neq 0$, we have $(0, 0, 0, \delta s) \in \ker \nabla^2 \tilde{\mathcal{R}}(\zeta^*)$. Therefore,

$$D^2 \mathcal{R}(\theta^*)(h, h) = D^2 \tilde{\mathcal{R}}(\zeta^*)(D\Phi(\theta^*)h, D\Phi(\theta^*)h) = 0,$$

so $h \in \ker \nabla^2 \mathcal{R}(\theta^*)$. This proves

$$\ker DC(\theta^*) \subset \ker \nabla^2 \mathcal{R}(\theta^*).$$

Conversely, let

$$h \in \ker \nabla^2 \mathcal{R}(\theta^*).$$

Then

$$0 = D^2 \mathcal{R}(\theta^*)(h, h) = D^2 \tilde{\mathcal{R}}(\zeta^*)(D\Phi(\theta^*)h, D\Phi(\theta^*)h).$$

Since $\zeta^* \in \mathcal{E}_1$ with $s(\theta^*) \neq 0$, Lemma B.13 implies that

$$D\Phi(\theta^*)h = (0, 0, 0, \delta s')$$

for some $\delta s' \in \mathbb{R}^d$. Hence the first three components vanish, that is,

$$DC(\theta^*)h = 0.$$

Thus $h \in \ker DC(\theta^*)$ and we obtain

$$\ker \nabla^2 \mathcal{R}(\theta^*) \subset \ker DC(\theta^*).$$

We conclude that

$$\ker \nabla^2 \mathcal{R}(\theta^*) = \ker DC(\theta^*) = T_{\theta^*} \mathcal{S}.$$

□

B.7 Local gradient-flow convergence (proof of Theorem 3.5)

The Morse–Bott structure provides local coordinates in which the excess risk is quadratic in the normal variables and independent of the tangential variables. To obtain convergence of the original Euclidean gradient flow, however, we must also account for the metric induced by this coordinate change. The argument below shows exponential decay in the normal directions, controls the total tangential displacement, and chooses an initial neighborhood from which the trajectory cannot leave the coordinate chart. This yields convergence to a single Bayes-optimal parameter, not merely approach to the set of minimizers.

The following statement makes the neighborhoods and the normal decay estimate in Theorem 3.5 explicit.

Notation. We recall the parameterization and the regular component used in the local argument.

$$\theta = (U_{11}, u_{12}, v_{21}, v_{22}) \in \mathbb{R}^{d \times d} \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}, \quad \Delta m := m_1 - m_{-1},$$

and define

$$\begin{aligned} w(\theta) &:= U_{11} \Gamma v_{21}, & b(\theta) &:= v_{21}^\top m_{-1} - v_{22}, \\ r(\theta) &:= v_{21}^\top \Delta m + 2v_{22}, & s(\theta) &:= 2u_{12} + U_{11} \Delta m. \end{aligned}$$

We consider the population risk

$$\mathcal{R}(\theta) = \tilde{\mathcal{R}}(w(\theta), b(\theta), r(\theta), s(\theta)).$$

We also consider

$$\Theta_1 := \left\{ \theta : w(\theta) = w^*, \quad b(\theta) = b^*, \quad r(\theta) = 0 \right\},$$

and its regular part

$$\Theta_1^{\text{reg}} := \left\{ \theta \in \Theta_1 : s(\theta) \neq 0 \right\}.$$

For a fixed point $\theta^* \in \Theta_1^{\text{reg}}$, we denote by $\mathcal{S}$ the connected component of Θ_1^{reg} containing θ^* .

Theorem B.29 (Local convergence of the gradient flow near a regular critical component). *Assume that $m_1 \neq -m_{-1}$, $\theta^* \in \Theta_1^{\text{reg}}$. Let $\mathcal{S}$ be the connected component of Θ_1^{reg} containing θ^* . Then there exist:*

- an open neighborhood U of θ^* in the parameter space,
- an open neighborhood $V \subset \mathbb{R}^n \times \mathbb{R}^{d+2}$ of 0, where $n = \dim \mathcal{S} = d^2 + d - 1$,
- a smooth diffeomorphism

$$\psi : U \longrightarrow V, \quad \psi(\theta^*) = 0,$$

such that

$$\psi(U \cap \mathcal{S}) = V \cap \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^{d+2} : y = 0\},$$

and

$$(\mathcal{R} \circ \psi^{-1})(x, y) = \mathcal{R}(\mathcal{S}) + |y|^2 \quad \text{for all } (x, y) \in V.$$

Moreover, there exists an open neighborhood $U_0 \subset U$ of θ^* such that, for every initial condition $\theta_0 \in U_0$, the Euclidean gradient flow

$$\dot{\theta}_t = -\nabla \mathcal{R}(\theta_t), \quad \theta_{t=0} = \theta_0,$$

is defined for all $t \geq 0$, remains in U , and converges to a point $\theta_\infty \in \mathcal{S}$.

More precisely, if we write

$$z_t := (x_t, y_t) := \psi(\theta_t),$$

then $z_t \in V$ for all $t \geq 0$, x_t converges in $\mathbb{R}^n$, and there exists a constant $c > 0$ such that

$$|y_t| \leq e^{-ct} |y_0| \quad \text{for all } t \geq 0.$$

In particular,

$$\text{dist}(\theta_t, \mathcal{S}) \longrightarrow 0 \quad \text{and} \quad \theta_t \longrightarrow \theta_\infty \in \mathcal{S} \quad \text{as } t \rightarrow +\infty.$$

where the convergence rates are exponentials.

Proof of Theorem 3.5. We prove the detailed version stated in Theorem B.29, which implies the main-text result. The argument is divided into the following steps.

Step 1: Rewriting the flow in Morse–Bott coordinates. By Lemma B.26, we can use the Morse–Bott theorem (Lemma B.25) Let

$$\psi : U \rightarrow V \subset \mathbb{R}^n \times \mathbb{R}^{d+2}$$

be the local diffeomorphism given by the Morse–Bott lemma at θ^* , and define

$$F := \mathcal{R} \circ \psi^{-1}.$$

By construction, F is the expression of the risk in local coordinates adapted to the critical manifold $\mathcal{S}$: the variable

$$z = (x, y) \in \mathbb{R}^n \times \mathbb{R}^{d+2}$$

splits into tangent coordinates x along $\mathcal{S}$ and normal coordinates y , along which the risk has a quadratic structure.

Now let $(\theta_t)_{t \geq 0}$ be a solution of the Euclidean gradient flow

$$\dot{\theta}_t = -\nabla \mathcal{R}(\theta_t), \quad \theta_{t=0} = \theta_0,$$

as long as $\theta_t \in U$, and define

$$z_t := \psi(\theta_t).$$

Differentiating the identity $z_t = \psi(\theta_t)$, we obtain

$$\dot{z}_t = D\psi(\theta_t)\dot{\theta}_t = -D\psi(\theta_t)\nabla \mathcal{R}(\theta_t).$$

Set

$$J(z) := D\psi(\psi^{-1}(z)), \quad K(z) := D\psi^{-1}(z) = J(z)^{-1}.$$

Next, by the chain rule, we have that

$$\nabla \mathcal{R}(\psi^{-1}(z)) = \nabla(F \circ \psi)(\psi^{-1}(z)) = D\psi(\psi^{-1}(z))^\top \nabla F(z).$$

Substituting this into the previous identity gives

$$\dot{z}_t = -J(z_t)J(z_t)^\top \nabla F(z_t).$$

Therefore, defining $G(z) := J(z)J(z)^\top$, the flow in the z -coordinates takes the form

$$\dot{z}_t = -G(z_t)\nabla F(z_t).$$

Finally, the map G is continuous on V , and for every $z \in V$, the matrix $G(z)$ is symmetric positive definite, with inverse

$$G(z)^{-1} = K(z)^\top K(z) = D\psi^{-1}(z)^\top D\psi^{-1}(z).$$

This shows that the original Euclidean gradient flow for $\mathcal{R}$ is equivalent, in the coordinates given by ψ , to a gradient-type flow for F in the tangent/normal splitting.

Step 2: Writing the flow in the (x, y) -coordinates and preliminary estimates.

By the Morse–Bott normal form, for every $(x, y) \in V \subset \mathbb{R}^n \times \mathbb{R}^{d+2}$, we have

$$F(x, y) = \mathcal{R}(\mathcal{S}) + |y|^2,$$

hence

$$\nabla F(x, y) = \begin{pmatrix} 0 \\ 2y \end{pmatrix}.$$

From Step 1, the flow in the z -coordinates satisfies

$$\dot{z}_t = -G(z_t)\nabla F(z_t), \quad z_t = (x_t, y_t).$$

We write the symmetric positive definite matrix $G(x, y)$ in block form as

$$G(x, y) = \begin{pmatrix} A(x, y) & B(x, y) \\ B(x, y)^\top & C(x, y) \end{pmatrix},$$

where

$$A(x, y) \in \mathbb{R}^{n \times n}, \quad B(x, y) \in \mathbb{R}^{n \times (d+2)}, \quad C(x, y) \in \mathbb{R}^{(d+2) \times (d+2)}.$$

Therefore,

$$\begin{pmatrix} \dot{x}_t \\ \dot{y}_t \end{pmatrix} = - \begin{pmatrix} A(x_t, y_t) & B(x_t, y_t) \\ B(x_t, y_t)^\top & C(x_t, y_t) \end{pmatrix} \begin{pmatrix} 0 \\ 2y_t \end{pmatrix},$$

that is,

$$\dot{x}_t = -2B(x_t, y_t)y_t, \quad \dot{y}_t = -2C(x_t, y_t)y_t.$$

We now fix a compact set $K := \overline{B}_{\mathbb{R}^n}(0, \rho_x) \times \overline{B}_{\mathbb{R}^{d+2}}(0, \rho_y) \subset V$. Our goal will be to show that, on the compact set K , the eigenvalues of B and C are bounded below, which will allow us to use Grönwall's lemma to obtain exponential convergence on this compact set, and then to adjust the size of the permitted initialization perimeter to ensure that convergence occurs before leaving the compact set.

Since C is continuous on V and takes values in the set of symmetric positive definite matrices, and since the map $M \mapsto \lambda_{\min}(M)$ is continuous on the space of symmetric matrices, the function $z \mapsto \lambda_{\min}(C(z))$ is continuous on K . Moreover, it is strictly positive on K . Hence we may define

$$\lambda := \min_{z \in K} \lambda_{\min}(C(z)) > 0.$$

We also set

$$M := \sup_{z \in K} \|B(z)\|_{\text{op}} < +\infty.$$

For the remainder of the proof, we choose $\varepsilon_x, \varepsilon_y > 0$ such that

$$\varepsilon_x + \frac{M}{\lambda} \varepsilon_y < \rho_x/2, \quad \varepsilon_y < \rho_y/2,$$

and we consider an initial condition $\theta_{t=0} = \theta_0 = \psi^{-1}(x_0, y_0)$ such that

$$|x_0| < \varepsilon_x, \quad |y_0| \leq \varepsilon_y.$$

Finally, we define

$$T^* := \sup \left\{ T > 0 : z_t \in \text{int}(K) \text{ for all } t \in [0, T] \right\}.$$

We will show below that $T^* = +\infty$.

Step 3: Control in the normal directions.

For $t \in [0, T^*)$, we have

$$\dot{y}_t = -2C(x_t, y_t)y_t.$$

Define $V(y) := \|y\|^2$ so that

$$\frac{d}{dt}V(y_t) = 2\langle y_t, \dot{y}_t \rangle = -4\langle y_t, C(x_t, y_t)y_t \rangle.$$

Since $z_t = (x_t, y_t) \in \text{int}(K) \subset K$ for every $t < T^*$, and by definition of λ ,

$$\langle y, C(z)y \rangle \geq \lambda \|y\|^2 \quad \text{for all } z \in K, \ y \in \mathbb{R}^{d+2},$$

we obtain

$$\frac{d}{dt}V(y_t) \leq -4\lambda V(y_t).$$

By Grönwall's lemma,

$$V(y_t) \leq e^{-4\lambda t}V(y_0) \quad \text{for all } t < T^*.$$

Equivalently,

$$\|y_t\| \leq e^{-2\lambda t} \|y_0\| \quad \text{for all } t < T^*.$$

Step 4: Control in the tangent directions.

For $t \in [0, T^*)$, we also have

$$\dot{x}_t = -2B(x_t, y_t)y_t.$$

Hence, using the definition of M ,

$$\|\dot{x}_t\| \leq 2\|B(x_t, y_t)\|_{\text{op}} \|y_t\| \leq 2M\|y_t\|.$$

Combining this with the estimate from Step 3 gives

$$\|\dot{x}_t\| \leq 2Me^{-2\lambda t} \|y_0\|.$$

Integrating on $[0, t] \subset [0, T^*)$, we obtain

$$\|x_t\| \leq \|x_0\| + \int_0^t \|\dot{x}_s\| ds \leq \|x_0\| + 2M\|y_0\| \int_0^t e^{-2\lambda s} ds.$$

Therefore,

$$\|x_t\| \leq \|x_0\| + \frac{M}{\lambda} (1 - e^{-2\lambda t}) \|y_0\| \leq \|x_0\| + \frac{M}{\lambda} \|y_0\| \quad \text{for all } t < T^*.$$

In particular,

$$\|x_t\| < \varepsilon_x + \frac{M}{\lambda} \varepsilon_y \leq \rho_x/2,$$

and from Step 3,

$$\|y_t\| \leq \|y_0\| \leq \varepsilon_y \leq \rho_y/2.$$

Thus $z_t = (x_t, y_t)$ cannot reach the boundary of K on $[0, T^*)$. By continuity and the definition of T^* , this implies

$$T^* = +\infty.$$

Conclusion. By Step 3, we have

$$\|y_t\| \leq e^{-2\lambda t} \|y_0\| \quad \text{for all } t \geq 0,$$

hence

$$y_t \longrightarrow 0 \quad \text{as } t \rightarrow +\infty.$$

Moreover, by Step 4,

$$\|x_t\| \leq \|x_0\| + \frac{M}{\lambda} \|y_0\| < \rho_x \quad \text{for all } t \geq 0,$$

so (x_t) remains in the compact ball $\overline{B}(0, \rho_x)$. Since

$$\int_0^{+\infty} \|\dot{x}_t\| dt \leq 2M\|y_0\| \int_0^{+\infty} e^{-2\lambda t} dt < +\infty,$$

the curve (x_t) has finite length and therefore converges to some point

$$x_\infty \in \overline{B}(0, \rho_x).$$

Consequently,

$$z_t = (x_t, y_t) \longrightarrow (x_\infty, 0) \quad \text{as } t \rightarrow +\infty.$$

Since ψ is a diffeomorphism, it follows that

$$\theta_t = \psi^{-1}(z_t) \longrightarrow \theta_\infty := \psi^{-1}(x_\infty, 0).$$

Finally, because

$$(x_\infty, 0) \in V \cap \{(x, y) : y = 0\} = \psi(U \cap \mathcal{S}),$$

we obtain

$$\theta_\infty \in \mathcal{S} \cap U.$$

The exponential convergence in the original coordinate θ_t can be deduced from the fact that ψ is a locally lipschitz, hence $\|\theta_t - \theta_\infty\| \leq L\|z_t - z_\infty\|$. This concludes the proof. $\square$

B.8 Constant-gate geometry (complement to Theorem 3.5)

The two Bayes representation mechanisms need not yield the same local geometry. On the regular part of Θ_1 , a nonconstant gate has zero amplitude, and the preceding argument identifies all flat directions with motion along the minimizer manifold. On $\mathcal{E}_2$, the gate is constant: a perturbation of its direction can be compensated to first order by a change in the affine coefficient. The following calculation identifies the resulting additional Hessian kernel.

Proposition B.30 (Failure of the Morse–Bott property). *Let $\zeta^* = (w^*, b^* - r^* \sigma(s_0), r^*, 0) \in \mathcal{E}_2$ such that $r^* \neq 0$. Then*

$$\ker \nabla^2 \tilde{\mathcal{R}}(\zeta^*) = T_{\zeta^*} \mathcal{E}_2 \oplus E_{\text{flat}},$$

where

$$E_{\text{flat}} = \{(-r^* \sigma'(s_0) \delta s, 0, 0, \delta s) : \delta s \in \mathbb{R}^d\}.$$

In particular, the Hessian kernel has dimension $d + 1$, whereas $T_{\zeta^*} \mathcal{E}_2$ has dimension one.

Corollary B.31 (Failure of the Morse–Bott property in attention coordinates). *Assume $m_1 \neq -m_{-1}$ and let $\theta^* \in \Theta_2$ such that $r(\theta^*) \neq 0$. Then*

$$\dim \ker \nabla^2 \mathcal{R}(\theta^*) - \dim T_{\theta^*} \Theta_2 = d > 0.$$

Thus the constant-gate component Θ_2 is not Morse–Bott.

Proof of Proposition B.30 and Corollary B.31. Let $r^* \in \mathbb{R} \setminus \{0\}$, and

$$\zeta^* = (w^*, b^* - r^* \sigma(s_0), r^*, 0) \in \mathcal{E}_2$$

A direct computation gives:

$$T_{\zeta^*} \mathcal{E}_2 = \{(0, -\sigma(s_0) \delta r, \delta r, 0) : \delta r \in \mathbb{R}\}.$$

. Moreover, the same computation as Proposition B.12 gives :

$$\nabla^2 \tilde{\mathcal{R}}(\zeta^*) = \mathbb{E} \left[\eta(X) (1 - \eta(X)) \nabla g_{\zeta^*}(X) \nabla g_{\zeta^*}(X)^\top \right],$$

Since $g_\zeta(x) = w^\top x + b + r \sigma(s_0 + s^\top x)$, we obtain:

$$\nabla g_{\zeta^*}(X) = \begin{pmatrix} X \\ 1 \\ \sigma(s_0) \\ r^* \sigma'(s_0) X \end{pmatrix}.$$

One can deduce :

$$D^2 \tilde{\mathcal{R}}(\zeta^*)[\delta \zeta, \delta \zeta] = E \left[\eta(X) (1 - \eta(X)) ((\delta w + r^* \sigma'(s_0) \delta s)^\top X + \delta b + \sigma(s_0) \delta r)^2 \right]$$

Hence,

$$\ker \nabla^2 \tilde{\mathcal{R}}(\zeta^*) = \{(-r^* \sigma'(s_0) \delta s, -\sigma(s_0) \delta r, \delta r, \delta s) : \delta r \in \mathbb{R}, \delta s \in \mathbb{R}^d\}$$

This proves the proposition.

For the corollary, Lemma B.20 combined with the fact that $v_{21}(\theta^*) \neq 0$ shows that $D\Phi(\theta^*)$ is surjective. We have $\nabla^2 \mathcal{R}(\theta^*) = D\Phi(\theta^*)^\top \nabla^2 \tilde{\mathcal{R}}(\zeta^*) D\Phi(\theta^*)$, so $\ker \nabla^2 \mathcal{R}(\theta^*) = D\Phi(\theta^*)^{-1} \ker \nabla^2 \tilde{\mathcal{R}}(\zeta^*)$. Moreover, because $D\Phi(\theta^*)$ is surjective, we have $T_{\theta^*} \Theta_2 = D\Phi(\theta^*)^{-1} T_{\zeta^*} \mathcal{E}_2$ (see [Guillemin and Pollack \(1974\)](#) Chap. 1 §5 exercise 5). We can conclude using Rank-nullity theorem. Indeed, for any subspace $V \subset \mathbb{R}^{2d+2} = \text{Im } D\Phi(\theta^*)$:

$$\dim D\Phi(\theta^*)^{-1}(V) = \dim \ker D\Phi(\theta^*) + \dim V$$

Thus

$$\dim \ker \nabla^2 \mathcal{R}(\theta^*) - \dim T_{\theta^*} \Theta_2 = \dim \ker \nabla^2 \tilde{\mathcal{R}}(\zeta^*) - \dim T_{\zeta^*} \mathcal{E}_2 = d$$

□

These extra directions distinguish the constant-gate branch from the regular zero-amplitude branch. At points of $\mathcal{E}_2$ with $r^* \neq 0$, the local minimizer set is the constant-gate branch, yet the Hessian kernel is strictly larger than its tangent space. At $r^* = 0$, the two effective branches meet, so the regular-manifold picture also requires separate care. Under the asymmetry assumption of the corollary, the same excess of flat directions transfers to attention coordinates.

This calculation explains why the preceding Morse–Bott argument cannot be reused near Θ_2 . It does not, by itself, rule out convergence there or determine a slower rate: those questions require control beyond the Hessian. The diagnostics in Figure 1 illustrate the contrast between the two branches without supplying such a convergence theorem.

B.9 Reduced-risk estimates and conserved balance (preparation for Theorem 3.6)

We now assume $m_1 = -m_{-1} =: m$ and $\pi_+ = 1/2$. The regularity argument based on $D\Phi$ no longer applies, but symmetry offers another route: on $\mathcal{I}_K$, the bias and the sigmoid contribution vanish, and the risk reduces to the linear logistic objective $L(w)$ with $w = U_{11}\Gamma v_{21}$.

Convexity in w alone is not enough, because training still evolves two factors. The lemmas below supply the additional controls: coercivity bounds the effective predictor, positive curvature on the resulting compact region gives a uniform convexity constant, and a conserved balance between the factors prevents the effective dynamics from stalling under the prescribed initialization. These ingredients are assembled in Appendix B.12.

Lemma B.32 (Coercivity of the reduced linear risk). *Let*

$$L(w) := \tilde{\mathcal{R}}(w, 0, 0, 0) = \mathbb{E}[\ell_{\log}(Y, w^\top X)].$$

Then L is coercive, with a linear lower bound. More precisely, there exists a constant $C_1 > 0$ such that, for every $w \in \mathbb{R}^d$,

$$L(w) \geq C_1 \|w\|.$$

In particular,

$$L(w) \longrightarrow +\infty \quad \text{as } \|w\| \rightarrow +\infty.$$

Proof. Since $m_{-1} = -m_1$, we first observe that $YX \sim \mathcal{N}(m_1, \Gamma)$. For every $(y, t) \in \{-1, 1\} \times \mathbb{R}$, the logistic loss satisfies

$$\ell_{\log}(y, t) = \log(1 + \exp(-yt)) \geq (-yt)_+.$$

Therefore, for every $w \in \mathbb{R}^d$,

$$L(w) = \mathbb{E}[\ell_{\log}(Y, w^\top X)] \geq \mathbb{E}[(-w^\top YX)_+].$$

Set

$$Z_w \sim \mathcal{N}(\mu_w, \sigma_w^2), \quad \mu_w := -w^\top m_1, \quad \sigma_w^2 := w^\top \Gamma w.$$

Thus, for $w \neq 0$, using Lemma B.6,

$$\mathbb{E}[(-w^\top YX)_+] = \mathbb{E}[(Z_w)_+] = \mu_w \Phi\left(\frac{\mu_w}{\sigma_w}\right) + \sigma_w \varphi\left(\frac{\mu_w}{\sigma_w}\right).$$

So

$$\mathbb{E}[(Z_w)_+] = \sigma_w [q_w \Phi(q_w) + \varphi(q_w)], \quad q_w := \frac{\mu_w}{\sigma_w}.$$

We show that the term in brackets is bounded from below by a positive constant. Since Γ is symmetric positive definite, by Cauchy Schwarz inequality we get

$$|w^\top m_1| = \left\langle \Gamma^{1/2} w, \Gamma^{-1/2} m_1 \right\rangle \leq (w^\top \Gamma w)^{1/2} (m_1^\top \Gamma^{-1} m_1)^{1/2}.$$

Hence

$$|q_w| \leq M, \quad M := (m_1^\top \Gamma^{-1} m_1)^{1/2}.$$

Let $g(t) := t\Phi(t) + \varphi(t)$, the function g is continuous and strictly positive on $\mathbb{R}$, since (still by Lemma B.6)

$$g(t) = \mathbb{E}[G_+], \quad G \sim \mathcal{N}(t, 1).$$

Therefore, we can define

$$c_0 := \inf_{w \neq 0} q_w \Phi(q_w) + \varphi(q_w) = \min_{t \in [-M, M]} g(t) > 0.$$

It follows that, for every $w \neq 0$,

$$\mathbb{E}[(Z_w)_+] \geq c_0 \sigma_w.$$

Moreover, since Γ is symmetric positive definite,

$$\sigma_w = (w^\top \Gamma w)^{1/2} \geq \sqrt{\lambda_{\min}(\Gamma)} \|w\|.$$

Consequently,

$$L(w) \geq \mathbb{E}[(Z_w)_+] \geq c_0 \sqrt{\lambda_{\min}(\Gamma)} \|w\|.$$

Thus the desired inequality holds with

$$C_1 := c_0 \sqrt{\lambda_{\min}(\Gamma)} > 0$$

The case $w = 0$ is immediate since $L(0) = \log 2 \geq 0$. □

Lemma B.33 (Local strong convexity along the effective trajectory). *Let $(\theta_t)_{t \geq 0}$ be a gradient-flow trajectory initialized at $\theta_0 \in \mathcal{I}_K$, and let*

$$w_t = U_{11}(t) \Gamma v_{21}(t).$$

Define

$$\mathcal{C}_{\theta_0} := \overline{\text{Conv}(\{w_t : t \geq 0\} \cup \{w^*\})}.$$

Then there exists a constant $\mu_0 = \mu_0(\theta_0) > 0$ such that L is μ_0 -strongly convex on $\mathcal{C}_{\theta_0}$. In other words, for every $w, w' \in \mathcal{C}_{\theta_0}$,

$$L(w') \geq L(w) + \langle \nabla L(w), w' - w \rangle + \frac{\mu_0}{2} \|w' - w\|^2.$$

Proof. Since the trajectory remains in $\mathcal{I}_K$, one has

$$\mathcal{R}(\theta_t) = L(w_t) \quad \text{for every } t \geq 0.$$

Moreover, $\mathcal{R}$ is non-increasing along the gradient flow. Hence

$$L(w_t) = \mathcal{R}(\theta_t) \leq \mathcal{R}(\theta_0) = L(w_0) \quad \text{for every } t \geq 0.$$

This implies that $(w_t)_{t \geq 0}$ is bounded. Indeed, if

$$\limsup_{t \rightarrow +\infty} \|w_t\| = +\infty,$$

then, by Lemma B.32, one would have

$$\limsup_{t \rightarrow +\infty} L(w_t) = +\infty,$$

which contradicts the uniform bound

$$L(w_t) \leq L(w_0).$$

Thus $(w_t)_{t \geq 0}$ is bounded. It follows that $\{w_t : t \geq 0\} \cup \{w^*\}$ is bounded, and therefore, because we work in finite dimension, $\mathcal{C}_{\theta_0} = \overline{\text{Conv}(\{w_t : t \geq 0\} \cup \{w^*\})}$ is compact.

Now L is the population logistic risk restricted to linear scores,

$$L(w) = \mathbb{E}[\ell_{\log}(Y, w^\top X)].$$

Its Hessian is

$$\nabla^2 L(w) = \mathbb{E}[\sigma'(w^\top X) X X^\top].$$

Under the non-degeneracy of the LDA distribution, this Hessian is positive definite for every $w \in \mathbb{R}^d$. Since $\nabla^2 L$ is continuous and $\mathcal{C}_{\theta_0}$ is compact, the map

$$w \mapsto \lambda_{\min}(\nabla^2 L(w))$$

attains a positive minimum on $\mathcal{C}_{\theta_0}$. We define

$$\mu_0 := \min_{w \in \mathcal{C}_{\theta_0}} \lambda_{\min}(\nabla^2 L(w)) > 0.$$

Therefore,

$$\nabla^2 L(w) \succeq \mu_0 I \quad \text{for every } w \in \mathcal{C}_{\theta_0}.$$

This is exactly the μ_0 -strong convexity of L on $\mathcal{C}_{\theta_0}$. □

Lemma B.34 (Conservation of the balance on $\mathcal{I}_K$). *Let*

$$\theta_t = (U_{11}(t), u_{12}(t), v_{21}(t), v_{22}(t))$$

be a gradient-flow trajectory such that

$$\theta_t \in \mathcal{I}_K \quad \text{for every } t \geq 0.$$

Then

$$\|U_{11}(t)\|_F^2 - \|v_{21}(t)\|^2 = \|U_{11}(0)\|_F^2 - \|v_{21}(0)\|^2 \quad \text{for every } t \geq 0.$$

Proof. On $\mathcal{I}_K$, the flow reduces to

$$\dot{U}_{11}(t) = -G_{w_t}(\Gamma v_{21}(t))^\top, \quad \dot{v}_{21}(t) = -\Gamma U_{11}(t)^\top G_{w_t},$$

where

$$G_{w_t} := \mathbb{E}[(\sigma(w_t^\top X) - \eta(X))X] = \nabla L(w_t), \quad w_t = U_{11}(t)\Gamma v_{21}(t).$$

(see proof of Lemma B.37 for more details)

Set

$$F(t) := \|U_{11}(t)\|_F^2 - \|v_{21}(t)\|^2.$$

Then, using $\Gamma = \Gamma^\top$,

$$\begin{aligned} F'(t) &= 2\langle U_{11}(t), \dot{U}_{11}(t) \rangle_F - 2\langle v_{21}(t), \dot{v}_{21}(t) \rangle \\ &= -2G_{w_t}^\top U_{11}(t)\Gamma v_{21}(t) + 2v_{21}(t)^\top \Gamma U_{11}(t)^\top G_{w_t} \\ &= -2G_{w_t}^\top w_t + 2G_{w_t}^\top w_t \\ &= 0. \end{aligned}$$

Thus $F(t)$ is constant along the flow, which proves the claim. □

B.10 Feasibility of the Krylov initialization (complement to Theorem 3.6)

The convergence theorem requires a nonzero initialization in $K^\perp$, so it is useful to determine when this space is available. The answer is spectral: successive applications of Γ to m separate the distinct eigenvalues represented in m , but cannot generate every direction inside a repeated eigenspace. The following proposition states the resulting criterion.

Proposition B.35. *For symmetric positive-definite Γ , the Krylov space $K = \text{span}\{m, \Gamma m, \dots, \Gamma^{d-1}m\}$ equals $\mathbb{R}^d$ if and only if Γ has d distinct eigenvalues and the projection of m onto every eigenspace of Γ is nonzero.*

Proof. Since Γ is symmetric, there exists an orthonormal basis $(e_1, \dots, e_d)$ of $\mathbb{R}^d$ made of eigenvectors of Γ . We write

$$\Gamma e_i = \lambda_i e_i, \quad i = 1, \dots, d,$$

and

$$m = \sum_{i=1}^d m_i e_i.$$

We first prove the direct implication by contraposition. Assume first that Γ does not have d distinct eigenvalues. Without loss of generality, suppose that $\lambda_1 = \lambda_2$. Let P be the orthogonal projection onto $E := \text{span}(e_1, e_2)$. Then, for every $k \geq 0$,

$$P\Gamma^k m = \lambda_1^k (m_1 e_1 + m_2 e_2).$$

Therefore,

$$PK = \text{span}\{P\Gamma^k m : k = 0, \dots, d-1\} \subset \text{span}\{m_1 e_1 + m_2 e_2\}.$$

Hence $\dim(PK) \leq 1$, whereas $\dim(E) = 2$. Thus K cannot be equal to $\mathbb{R}^d$.

Assume now that the component of m in one eigenspace is zero. For instance, suppose that the projection of Δm onto the eigenspace associated with λ_1 is zero, which means $m_1 = 0$. Let P_1 be the orthogonal projection onto that eigenspace. We have for every $k \geq 0$,

$$P_1 \Gamma^k m = \Gamma^k P_1 m = 0.$$

Therefore,

$$P_1 K = \{0\},$$

so K cannot be equal to $\mathbb{R}^d$. This proves that if $K = \mathbb{R}^d$, then Γ has d distinct eigenvalues and m has a nonzero component in each eigenspace of Γ .

Conversely, assume that Γ has d distinct eigenvalues $\lambda_1, \dots, \lambda_d$ and that the components of m in the corresponding eigendirections are all nonzero (which means $m_i \neq 0$ for $i = 1, \dots, d$).

We prove that the family

$$m, \Gamma m, \dots, \Gamma^{d-1} m$$

is linearly independent. Let $a_1, \dots, a_d \in \mathbb{R}$ be such that

$$\sum_{\ell=1}^d a_\ell \Gamma^{\ell-1} m = 0.$$

We have

$$\Gamma^{\ell-1} m = \sum_{i=1}^d m_i \lambda_i^{\ell-1} e_i.$$

Therefore,

$$0 = \sum_{\ell=1}^d a_\ell \Gamma^{\ell-1} m = \sum_{i=1}^d m_i \left(\sum_{\ell=1}^d a_\ell \lambda_i^{\ell-1} \right) e_i.$$

Since $(e_1, \dots, e_d)$ is a basis, it follows that for every $i = 1, \dots, d$,

$$m_i \left(\sum_{\ell=1}^d a_\ell \lambda_i^{\ell-1} \right) = 0.$$

Since $m_i \neq 0$, we obtain

$$\sum_{\ell=1}^d a_\ell \lambda_i^{\ell-1} = 0, \quad i = 1, \dots, d.$$

Now define the polynomial

$$P(X) := \sum_{\ell=1}^d a_\ell X^{\ell-1}.$$

Then

$$P(\lambda_i) = 0, \quad i = 1, \dots, d.$$

Hence P is a polynomial of degree at most $d - 1$ having d distinct roots. Therefore P is identically zero. Consequently,

$$a_1 = \dots = a_d = 0.$$

Thus the family

$$m, \Gamma m, \dots, \Gamma^{d-1} m$$

is linearly independent. Since it contains d vectors in $\mathbb{R}^d$, it is a basis of $\mathbb{R}^d$. This proves $K = \mathbb{R}^d$. $\square$

Remark B.36 (Initialization and the codimension of the Krylov space). *The only nontrivial initialization constraint is $v_{21}(0) \in K^\perp \setminus \{0\}$. When $\dim K \ll d$, an isotropic random initialization has only a fraction of order $\dim(K)/d$ of its energy in K . In particular, if $\Gamma = \gamma^2 I_d$, then $K = \text{span}(m)$, so a normalized Gaussian initialization lies at distance $O_{\mathbb{P}}(d^{-1/2})$ from $K^\perp$.*

The proximity of a random initialization to $K^\perp$ should not be confused with exact membership: the convergence theorem below assumes the latter. When $K \neq \mathbb{R}^d$, this constraint can be enforced by choosing a nonzero vector in $K^\perp$.

B.11 Invariance of the Krylov subspace (preparation for Theorem 3.6)

The Krylov constraint is designed to be preserved by the flow. Orthogonality to m alone would not suffice, since the factor dynamics also involve multiplication by Γ . By construction, K and $K^\perp$ are Γ -invariant. Together with $U_{11}P_K = 0$, this closes the dynamics on $\mathcal{I}_K$ and keeps the predictor linear throughout training.

Lemma B.37. *In the symmetric balanced setting, if $\theta_0 \in \mathcal{I}_K$, then the gradient-flow trajectory remains in $\mathcal{I}_K$ for every $t \geq 0$.*

Proof. Let $\theta^K = (U_{11}, u_{12}, v_{21}, v_{22}) \in \mathcal{I}_K$. By definition of $\mathcal{I}_K$, one has

$$u_{12} = 0, \quad v_{22} = 0, \quad v_{21} \in K^\perp, \quad U_{11}P_K = 0.$$

Since $\Delta m \in K$, this implies $v_{21}^\top \Delta m = 0$ and $U_{11} \Delta m = 0$. Therefore, on $\mathcal{I}_K$,

$$b(\theta^K) = 0, \quad r(\theta^K) = 0, \quad s(\theta^K) = 0.$$

We now set $G_w := \mathbb{E}[(\sigma(w^\top X) - \eta(X))X]$. Using Lemma B.10 (and the fact that $s_0 = 0$ in the symmetric setting), we immediately obtain

$$\nabla_{u_{12}} \mathcal{R}(\theta^K) = 0, \quad \nabla_{v_{22}} \mathcal{R}(\theta^K) = 0, \quad \nabla_{U_{11}} \mathcal{R}(\theta^K) = G_w(\Gamma v_{21})^\top, \quad \nabla_{v_{21}} \mathcal{R}(\theta^K) = \Gamma U_{11}^\top G_w.$$

We now show that the vector field of the gradient flow stays in $\mathcal{I}_K$. We already know that at $\theta^K \in \mathcal{I}_K$, $\dot{u}_{12} = 0$ and $\dot{v}_{22} = 0$. Let $x_K \in K$. Since $v_{21} \in K^\perp$ and $K^\perp$ is stable under Γ , one has $\Gamma v_{21} \in K^\perp$. Therefore

$$\dot{U}_{11} x_K = -\nabla_{U_{11}} \mathcal{R}(\theta^K) x_K = -G_w(\Gamma v_{21})^\top x_K = 0.$$

Since this holds for every $x_K \in K$, we obtain $\dot{U}_{11} P_K = 0$.

It remains to check the constraint on v_{21} . Since $U_{11} P_K = 0$, one has $\text{Im}(U_{11}^\top) \subset K^\perp$ and thus $U_{11}^\top G_w \in K^\perp$. Because $K^\perp$ is stable under Γ , it follows that

$$\dot{v}_{21} = -\nabla_{v_{21}} \mathcal{R}(\theta^K) = -\Gamma U_{11}^\top G_w \in K^\perp.$$

We have therefore shown that, whenever $\theta_t \in \mathcal{I}_K$,

$$\dot{u}_{12}(t) = 0, \quad \dot{v}_{22}(t) = 0, \quad \dot{U}_{11}(t) P_K = 0, \quad \dot{v}_{21}(t) \in K^\perp.$$

Equivalently,

$$\dot{\theta}_t \in \mathcal{I}_K$$

It follows that if $\theta_0 \in \mathcal{I}_K$, then the whole trajectory remains in $\mathcal{I}_K$. $\square$

B.12 Global convergence on the invariant subspace (proof of Theorem 3.6)

We now combine invariance with the controls established above. The proof first derives a preconditioned gradient flow for $w_t = U_{11}(t)\Gamma v_{21}(t)$. The conserved balance and nonzero initialization provide a uniform lower bound on its preconditioner, while coercivity keeps w_t in a region of uniform positive curvature. These facts yield exponential convergence of the risk and effective predictor. A final estimate then bounds the factors and makes their velocities integrable, establishing convergence in the original attention parameters as well.

Proof of Theorem 3.6. Let $\theta_0 = (U_{11}(0), u_{12}(0), v_{21}(0), v_{22}(0))$ satisfy the initialization assumptions of the theorem, namely

$$U_{11}(0) = 0, \quad u_{12}(0) = 0, \quad v_{22}(0) = 0, \quad v_{21}(0) \in K^\perp \setminus \{0\}.$$

Then $\theta_0 \in \mathcal{I}_K$. By Lemma B.37, the whole trajectory remains in $\mathcal{I}_K$. In particular,

$$u_{12}(t) = 0, \quad v_{22}(t) = 0, \quad b_t = r_t = s_t = 0.$$

We first prove the convergence of the effective variable $w_t := U_{11}(t)\Gamma v_{21}(t)$. Set

$$L(w) := \tilde{\mathcal{R}}(w, 0, 0, 0).$$

On $\mathcal{I}_K$, one has $\mathcal{R}(\theta_t) = L(w_t)$. Moreover, the gradient flow reduces to

$$\dot{U}_{11}(t) = -\nabla L(w_t)(\Gamma v_{21}(t))^\top, \quad \dot{v}_{21}(t) = -\Gamma U_{11}(t)^\top \nabla L(w_t).$$

Therefore,

$$\begin{aligned} \dot{w}_t &= \dot{U}_{11}(t)\Gamma v_{21}(t) + U_{11}(t)\Gamma \dot{v}_{21}(t) \\ &= -\|\Gamma v_{21}(t)\|^2 \nabla L(w_t) - U_{11}(t)\Gamma^2 U_{11}(t)^\top \nabla L(w_t). \end{aligned}$$

Thus the effective dynamics can be written as

$$\dot{w}_t = -G_t \nabla L(w_t),$$

where

$$G_t := \|\Gamma v_{21}(t)\|^2 I_d + U_{11}(t)\Gamma^2 U_{11}(t)^\top.$$

By Lemma B.34,

$$\|U_{11}(t)\|_F^2 - \|v_{21}(t)\|^2 = \|U_{11}(0)\|_F^2 - \|v_{21}(0)\|^2 = -\|v_{21}(0)\|^2 =: C_0 < 0$$

Hence, for every $t \geq 0$,

$$\|v_{21}(t)\|^2 = \|U_{11}(t)\|_F^2 - C_0 \geq -C_0.$$

We set $\nu := \sqrt{-C_0} > 0$ so that

$$\|v_{21}(t)\| \geq \nu \quad \text{for every } t \geq 0.$$

Since $\Gamma = \Gamma^\top \succ 0$,

$$\|\Gamma v_{21}(t)\| \geq \lambda_{\min}(\Gamma)\|v_{21}(t)\| \geq \lambda_{\min}(\Gamma)\nu.$$

Moreover, $U_{11}(t)\Gamma^2 U_{11}(t)^\top$ is positive semidefinite. Hence

$$G_t \succeq \|\Gamma v_{21}(t)\|^2 I_d \succeq \alpha I_d,$$

where $\alpha := \lambda_{\min}(\Gamma)^2 \nu^2 > 0$.

By Lemma B.33, L is μ_0 -strongly convex on

$$\mathcal{C}_{\theta_0} = \overline{\text{Conv}(\{w_t : t \geq 0\} \cup \{w^*\})}.$$

Thus, for every $t \geq 0$,

$$\|\nabla L(w_t)\|^2 \geq 2\mu_0(L(w_t) - L(w^*)).$$

Using the effective dynamics, we get

$$\begin{aligned}
\frac{d}{dt}(L(w_t) - L(w^*)) &= \langle \nabla L(w_t), \dot{w}_t \rangle \\
&= -\langle \nabla L(w_t), G_t \nabla L(w_t) \rangle \\
&\leq -\alpha \|\nabla L(w_t)\|^2 \\
&\leq -2\alpha\mu_0(L(w_t) - L(w^*)).
\end{aligned}$$

By Grönwall's inequality,

$$L(w_t) - L(w^*) \leq e^{-2\alpha\mu_0 t}(L(w_0) - L(w^*)).$$

In particular, the risk converges exponentially:

$$\mathcal{R}(\theta_t) - \mathcal{R}^* = L(w_t) - L(w^*) \lesssim e^{-2\alpha\mu_0 t}.$$

Since L is μ_0 -strongly convex on $\mathcal{C}_{\theta_0}$, we also have

$$L(w_t) - L(w^*) \geq \frac{\mu_0}{2} \|w_t - w^*\|^2.$$

Therefore

$$\|w_t - w^*\| \lesssim e^{-\alpha\mu_0 t}.$$

We finally prove the convergence of the factors $U_{11}(t)$ and $v_{21}(t)$. We first show that they are bounded. The function L is smooth. Indeed,

$$\nabla^2 L(w) = \mathbb{E} [\sigma'(w^\top X) X X^\top],$$

and since $0 \leq \sigma' \leq 1/4$, one has

$$\nabla^2 L(w) \preceq \frac{1}{4} \mathbb{E}[X X^\top].$$

Thus L is M -smooth with $M := \frac{1}{4} \|\mathbb{E}[X X^\top]\|_{\text{op}}$. In particular,

$$\|\nabla L(w_t)\| = \|\nabla L(w_t) - \nabla L(w^*)\| \leq M \|w_t - w^*\| \lesssim e^{-\alpha\mu_0 t}.$$

Set

$$S(t) := \|U_{11}(t)\|_F^2 + \|v_{21}(t)\|^2.$$

Using the reduced dynamics on $\mathcal{I}_K$, we get

$$\begin{aligned}
S'(t) &= 2\langle U_{11}(t), \dot{U}_{11}(t) \rangle_F + 2\langle v_{21}(t), \dot{v}_{21}(t) \rangle \\
&= -2\langle U_{11}(t), \nabla L(w_t) (\Gamma v_{21}(t))^\top \rangle_F - 2\langle v_{21}(t), \Gamma U_{11}(t)^\top \nabla L(w_t) \rangle \\
&= -4\langle w_t, \nabla L(w_t) \rangle.
\end{aligned}$$

Hence

$$|S'(t)| \leq 4\|w_t\| \|\nabla L(w_t)\| \lesssim e^{-\alpha\mu_0 t},$$

because $(w_t)_{t \geq 0}$ is bounded. Therefore $S' \in L^1(\mathbb{R}_+)$, and

$$S(t) = S(0) + \int_0^t S'(\tau) d\tau$$

converges to a finite limit as $t \rightarrow +\infty$. In particular,

$$\sup_{t \geq 0} \|U_{11}(t)\|_F < +\infty, \quad \sup_{t \geq 0} \|v_{21}(t)\| < +\infty.$$

We now prove convergence of the factors. Since

$$\dot{v}_{21}(t) = -\Gamma U_{11}(t)^\top \nabla L(w_t),$$

we have

$$\|\dot{v}_{21}(t)\| \leq \|\Gamma\|_{\text{op}} \|U_{11}(t)\|_{\text{op}} \|\nabla L(w_t)\| \lesssim e^{-\alpha\mu_0 t}.$$

Thus $\dot{v}_{21} \in L^1(\mathbb{R}_+)$, and $v_{21}(t)$ converges. More precisely, there exists $v_\infty \in \mathbb{R}^d$ such that

$$v_{21}(t) \rightarrow v_\infty,$$

and

$$\|v_{21}(t) - v_\infty\| \leq \int_t^{+\infty} \|\dot{v}_{21}(s)\| ds \lesssim e^{-\alpha\mu_0 t}.$$

Similarly,

$$\dot{U}_{11}(t) = -\nabla L(w_t)(\Gamma v_{21}(t))^\top,$$

so

$$\|\dot{U}_{11}(t)\|_F \leq \|\nabla L(w_t)\| \|\Gamma\|_{\text{op}} \|v_{21}(t)\| \lesssim e^{-\alpha\mu_0 t}.$$

Hence $\dot{U}_{11} \in L^1(\mathbb{R}_+)$, and there exists $U_\infty \in \mathbb{R}^{d \times d}$ such that

$$U_{11}(t) \rightarrow U_\infty,$$

with

$$\|U_{11}(t) - U_\infty\|_F \leq \int_t^{+\infty} \|\dot{U}_{11}(s)\|_F ds \lesssim e^{-\alpha\mu_0 t}.$$

Finally, we obtain

$$w^\star = \lim_{t \rightarrow +\infty} w_t = \lim_{t \rightarrow +\infty} U_{11}(t) \Gamma v_{21}(t) = U_\infty \Gamma v_\infty.$$

□

B.13 LDA experiments (illustrations of Theorems 3.5 and 3.6)

Configuration	dimension	m_1	m_{-1}	Γ	π_+
C1	$d = 2$	$(1, 0)^\top$	$(0, 1)^\top$	I_2	0.6
C2	$d = 20$	e_1	$-e_1$	I_{20}	0.5

Table 1: LDA configurations used in the numerical experiments.

Figure	Problem	Dimension	Numerical method
Fig. 1	Config. C1	$d = 2$	GH, step_size = 10^{-4}
Fig. 4	Config. C2	$d = 20$	SGD, step_size = 10^{-2}
Fig. 5	Config. C2	$d = 20$	SGD, step_size = 10^{-2}

Table 2: Technical details of the numerical experiments associated with each figure.

Initialization in the balanced symmetric setting. We compare two initialization strategies for stochastic gradient descent in dimension $d = 20$, with balanced classes and opposite means. With Gaussian initialization of all parameters, Figure 4 illustrates the difficulty of approaching the Bayes risk in the displayed runs. By contrast, setting $U_{11}(0) = 0$, $u_{12}(0) = 0$, and $v_{22}(0) = 0$, while sampling $v_{21}(0) \sim \mathcal{N}(0, I_d)$, leads to a rapid, approximately exponential decrease toward a small excess risk (Figure 5). This initialization requires no knowledge of m or Γ . Although it does not enforce the exact constraint $v_{21}(0) \in K^\perp$ required by Theorem 3.6, these experiments suggest that the favorable behavior can persist beyond the invariant subspace.

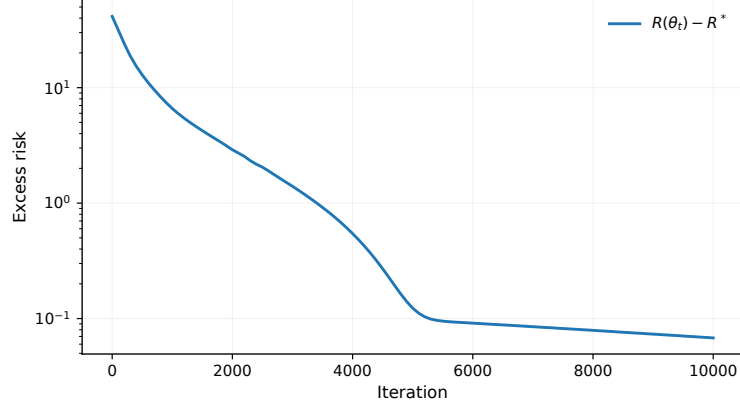


Figure 4: Risk evolution under stochastic gradient descent in the balanced symmetric setting with $d = 20$. All parameter entries are initialized from a standard Gaussian distribution.

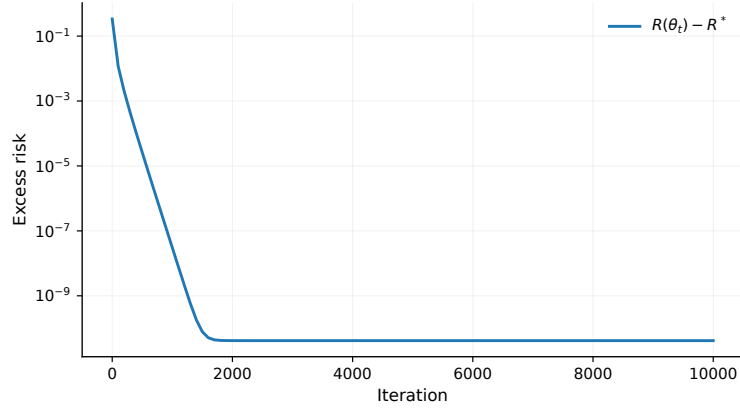


Figure 5: Risk evolution under stochastic gradient descent in the balanced symmetric setting with $d = 20$. Initialization: $U_{11}(0) = 0$, $u_{12}(0) = 0$, $v_{22}(0) = 0$, and $v_{21}(0) \sim \mathcal{N}(0, I_d)$.

Figure	$\mathbf{U}_{11}$	$\mathbf{u}_{12}$	$\mathbf{v}_{21}$	v_{22}	Noise
Fig. 1 ($\theta_t^{(1)}$)	$\begin{pmatrix} \frac{1}{2b_\star} & 0 \\ -\frac{1}{2b_\star} & 1 \end{pmatrix}$	$\begin{pmatrix} 10 \\ 10 \end{pmatrix}$	$\begin{pmatrix} 2b_\star \\ 0 \end{pmatrix}$	$-b_\star$	$\varepsilon = 10^{-6}$
Fig. 1 ($\theta_t^{(2)}$)	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} -\frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$	$\begin{pmatrix} 1 \\ -1 \end{pmatrix}$	$\frac{v_{21}^\top m_{-1} - b_\star + \pi_+ v_{21}^\top (m_1 - m_{-1})}{1 - 2\pi_+}$	$\varepsilon = 10^{-6}$

Table 3: Specific initializations used in numerical experiments.

C Proofs and Complements for Section 4

C.1 Bayes denoiser (derivation of Equation (10))

Proposition C.1 (Bayes denoiser for a Gaussian mixture). *Let $X \sim \sum_{k=1}^M \pi_k \mathcal{N}(m_k, \Gamma)$, $Z = X + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \xi^2 I_d)$ independent of X , and $S = \Gamma + \xi^2 I_d$. Then*

$$\mathbb{E}[X \mid Z = z] = \Gamma S^{-1} z + \xi^2 S^{-1} \sum_{k=1}^M \beta_k(z) m_k,$$

where

$$\beta_k(z) = \frac{\pi_k \mathcal{N}(z; m_k, S)}{\sum_{\ell=1}^M \pi_\ell \mathcal{N}(z; m_\ell, S)}.$$

Proof. Recalling that the conditional expectation is the minimizer of the mean squared error, the function $f^\star : z \mapsto \mathbb{E}[X \mid Z = z]$ is an optimal denoiser. It remains to compute this conditional expectation explicitly.

Let C denote the latent cluster variable, so that $\mathbb{P}(C = k) = \pi_k$ and $X \mid C = k \sim \mathcal{N}(m_k, \Gamma)$. By Bayes' formula,

$$\mathbb{P}(C = k \mid Z = z) = \frac{\pi_k \mathcal{N}(z; m_k, S)}{\sum_{\ell=1}^M \pi_\ell \mathcal{N}(z; m_\ell, S)} =: \beta_k(z).$$

Using the law of total expectation with respect to the latent variable C , we obtain

$$\mathbb{E}[X \mid Z = z] = \sum_{k=1}^M \mathbb{E}[X \mid Z = z, C = k] \mathbb{P}(C = k \mid Z = z).$$

It remains to compute $\mathbb{E}[X \mid Z = z, C = k]$. Conditionally on $C = k$, the pair (X, Z) is jointly Gaussian. Moreover, $\text{Cov}(X, Z \mid C = k) = \text{Cov}(X, X + \varepsilon \mid C = k) = \Gamma$. The Gaussian conditioning formula (e.g. see [Rasmussen and Williams \(2006\)](#)) therefore gives

$$\mathbb{E}[X \mid Z = z, C = k] = m_k + \Gamma S^{-1}(z - m_k).$$

Equivalently,

$$\mathbb{E}[X \mid Z = z, C = k] = \Gamma S^{-1} z + \xi^2 S^{-1} m_k.$$

Indeed,

$$I_d - \Gamma S^{-1} = (S - \Gamma) S^{-1} = \xi^2 S^{-1},$$

which explains the noise-variance coefficient ξ^2 . Plugging this expression into the previous decomposition and using that $\sum_{k=1}^M \beta_k(z) = 1$, we finally obtain

$$\mathbb{E}[X \mid Z = z] = \Gamma S^{-1} z + \sum_{k=1}^M \beta_k(z) \xi^2 S^{-1} m_k.$$

This proves the result.

It is worth noting that this result could also be derived from a multivariate version of Tweedie's formula, which states that

$$\mathbb{E}[X \mid Z = z] = z + \xi^2 \nabla_z \log p_Z(z).$$

□

Remark C.2 (High- and low-noise limits). *For fixed z , means, and covariance, as $\xi \rightarrow \infty$ the posterior weights $\beta_k(z)$ converge to the constant priors π_k , $\Gamma S^{-1} \rightarrow 0$, and $\xi^2 S^{-1} \rightarrow I_d$. Hence $f^*(z) \rightarrow \sum_k \pi_k m_k$. If Γ is positive definite, then as $\xi \downarrow 0$, $\Gamma S^{-1} \rightarrow I_d$ and $\xi^2 S^{-1} \rightarrow 0$, so $f^*(z) \rightarrow z$. The latter assertion does not apply to the degenerate clean mixture $\Gamma = 0$.*

C.2 Expressivity and risk separation (complete proof of Theorem 4.1)

Proof of Theorem 4.1. Proposition C.5 below constructs the two heads, including the case of opposite means. Proposition C.4 proves that a single head requires $\gamma^2 = \xi^2$. Conversely, at this balance $S = 2\xi^2 I_d$, and choosing $U = S^{-1}$, $V = I_d/2$ gives

$$T_{U,V}[\mu_Z](z) = \frac{1}{2}z + \frac{1}{2} \sum_k \beta_k(z) m_k = f^*(z).$$

Finally, Proposition C.6 proves the strict excess-risk gap, not merely failure of exact linear representation. Together these results establish all claims of the theorem. The degenerate complement stated after the theorem follows from Proposition C.3 and the same gap argument. □

C.3 Degenerate mixtures (complement to Theorem 4.1)

Proposition C.3 (Degenerate mixture: two heads are sufficient). *Assume $d \geq 2$, $\xi > 0$, $\Gamma = 0$, $M = 2$, positive mixture weights, and distinct means with $\|m_1\| = \|m_2\|$. Then a two-head softmax-attention predictor realizes the Bayes denoiser exactly. If, in addition, $d = 2$ and m_1, m_2 are linearly independent, no single head realizes it.*

Proof. We first prove the single-head obstruction. Since $\Gamma = 0$, the Bayes denoiser is the bounded posterior barycenter $f^*(z) = \beta_1(z)m_1 + \beta_2(z)m_2$. If one head represented it, comparison with

$$T_{U,V}[\mu_Z](z) = \xi^2 VUz + V(\alpha_1(z)m_1 + \alpha_2(z)m_2)$$

would force $VU = 0$, since the first term is linear and the target is bounded. Matching the two barycenters gives, up to exchanging the components, $Vm_i = m_i$ for $i = 1, 2$, while matching the posterior weights gives $U^\top(m_1 - m_2) = \xi^{-2}(m_1 - m_2)$. In dimension two, linear independence of m_1, m_2 implies $V = I_2$ in the first orientation. In the exchanged orientation, V swaps the two basis vectors and is also invertible. Thus, in either case $VU = 0$ forces $U = 0$, making the attention weights constant, a contradiction.

We now construct two heads. We begin with the case where m_1 and m_2 are linearly independent. The idea is to let each head be responsible for one cluster. More precisely, we want the first head to recover the contribution of m_1 and the second head to recover the contribution of m_2 .

Let $\Delta m := m_1 - m_2$. We choose two value matrices V_1, V_2 such that

$$V_1 m_1 = m_1, \quad V_1 m_2 = 0,$$

and

$$V_2 m_1 = 0, \quad V_2 m_2 = m_2.$$

In other words, V_1 extracts the contribution of the first cluster, while V_2 extracts the contribution of the second cluster.

We then choose U_1 and U_2 so that their images are killed by the corresponding value matrices, while their attention scores still reproduce the optimal posterior weights. For instance, one may take

$$U_1 = \frac{1}{\xi^2} \frac{m_2 \Delta m^\top}{\langle m_2, \Delta m \rangle}, \quad U_2 = \frac{1}{\xi^2} \frac{m_1 \Delta m^\top}{\langle m_1, \Delta m \rangle}.$$

Since $\|m_1\| = \|m_2\|$ and $m_1 \neq m_2$, the denominators are nonzero.

With this choice, we have

$$\text{Im}(U_1) \subset \text{span}(m_2), \quad \text{Im}(U_2) \subset \text{span}(m_1),$$

and therefore

$$V_1 U_1 = 0, \quad V_2 U_2 = 0.$$

Thus, the linear parts of both attention heads vanish. Moreover,

$$U_1^\top \Delta m = \frac{1}{\xi^2} \Delta m, \quad U_2^\top \Delta m = \frac{1}{\xi^2} \Delta m.$$

Hence both heads produce the same posterior weight $\beta_1(z)$ as the optimal denoiser.

We now address the case $m_1 = -m_2 =: m$. Here, a single head is sufficient. Let $n \neq 0$ satisfy $\langle n, m \rangle = 0$, which is possible because $d \geq 2$. One can choose $U = \frac{(m+n)m^\top}{\xi^2 \|m\|^2}$ and $V = m \left(\frac{m^\top}{\|m\|^2} - \frac{n^\top}{\|n\|^2} \right)$. Indeed, $Vm = m$, $Vn = -m$, and hence $VU = 0$, while $U^\top m = \xi^{-2} m$. This gives the optimal denoiser. Setting a second head to zero also gives the asserted two-head representation. $\square$

C.4 Denoising-to-clustering transition (proof of Proposition 4.3)

Proof of Proposition 4.3. Let $c = (m_1 + m_2)/2$, $\delta = (m_1 - m_2)/2$, and $\lambda = \|\delta\|/\xi$. Equal norms between m_1 and m_2 imply $\langle c, \delta \rangle = 0$. Since the mixture is balanced and $\gamma = 0$, $f^*(z) = m_2 + \Delta m \cdot \sigma(\langle z, \Delta m \rangle / \xi^2)$. Using $2\sigma(2t) - 1 = \tanh(t)$ gives

$$f^*(z) = c + \delta \tanh\left(\frac{\langle z - c, \delta \rangle}{\xi^2}\right).$$

In particular, one step maps every query into the open segment joining the means. More precisely, for every $n \geq 1$, there exists $u_n \in (-1, 1)$ such that $z_n = c + u_n \delta$. Set $u_0 = \langle z_0 - c, \delta \rangle / \|\delta\|^2$ and $F(u) = \tanh(\lambda^2 u)$. We can rewrite the dynamic $z_{n+1} = f^*(z_n)$ in term of u_n as follows :

$$u_{n+1} = F(u_n), \quad z_n = c + u_n \delta \quad (n \geq 1),$$

and $|u_n| < 1$ for $n \geq 1$. The map F is odd, strictly increasing, and strictly concave on $(0, \infty)$.

Weak separation, including the threshold. If $\lambda \leq 1$, then for every $u > 0$, $0 < F(u) = \tanh(\lambda^2 u) < \lambda^2 u \leq u$. Thus a positive trajectory decreases and remains bounded below by zero. Its limit is a fixed point by continuity, and the same strict inequality shows that zero is the only nonnegative fixed point. Oddness gives the same conclusion for negative initial values. Therefore $u_n \rightarrow 0$ and $z_n \rightarrow c$, including when $\lambda = 1$. Boundedness alone would not suffice here; monotonicity supplies the convergence argument.

Strong separation and the basins of attraction. If $\lambda > 1$, let $H(u) = F(u) - u$. We have $H(0) = 0$, $H'(0) = \lambda^2 - 1 > 0$, and $H(1) < 0$. Strict concavity on $(0, \infty)$ implies that H has exactly one positive zero $u_\lambda \in (0, 1)$, with $H > 0$ on $(0, u_\lambda)$ and $H < 0$ on (u_λ, ∞) . For $0 < u < u_\lambda$, monotonicity of F gives $u < F(u) < u_\lambda$; for $u > u_\lambda$, it gives $u_\lambda < F(u) < u$. Every positive trajectory is therefore monotone, bounded, and converges to u_λ . By oddness, every negative trajectory converges to $-u_\lambda$. The initial value zero remains zero. Consequently the two limiting points are

$$m_{\lambda,1} = c + u_\lambda \delta, \quad m_{\lambda,2} = c - u_\lambda \delta.$$

Their basins are respectively $\{z : \langle z - c, \delta \rangle > 0\}$ and $\{z : \langle z - c, \delta \rangle < 0\}$. The separating hyperplane maps directly to c and has Lebesgue measure zero. Moreover, strict concavity yields $F'(u_\lambda) < F(u_\lambda)/u_\lambda = 1$, while $F' > 0$. The two outer fixed points are stable; the midpoint is unstable along the line of the means because $F'(0) = \lambda^2 > 1$.

Limit of large separation. For any $r \in (0, 1)$, $F(r) = \tanh(\lambda^2 r) \rightarrow 1$ as $\lambda \rightarrow \infty$. Hence eventually $H(r) > 0$, so $u_\lambda > r$. Since $u_\lambda < 1$, it follows that $u_\lambda \rightarrow 1$. For fixed means and $\xi \downarrow 0$, this proves $m_{\lambda,1} \rightarrow m_1$ and $m_{\lambda,2} \rightarrow m_2$. $\square$

C.5 Single-head necessity (Theorem 4.1, single-head claim)

Proposition C.4 (A single head is generically insufficient). *Assume $\Gamma = \gamma^2 I_d$ with $\gamma > 0$, and retain the two-component equal-norm setting of Section 4. If $\gamma^2 \neq \xi^2$, there is no pair (U, V) such that $T_{U,V}[\mu_Z] = f^*$.*

Proof. A single attention head takes the form

$$T_{U,V}[\mu_Z](z) = (\gamma^2 + \xi^2)VUz + V(\alpha_1(z)m_1 + (1 - \alpha_1(z))m_2) \quad \text{with} \quad \alpha_1(z) = \sigma\left(\log \frac{\pi_1}{\pi_2} + \langle Uz, m_1 - m_2 \rangle\right)$$

If we try to impose

$$T_{U,V}[\mu_Z](z) = f^*(z) \quad \text{for all } z \in \mathbb{R}^d,$$

then boundedness of the gated terms first forces equality of the linear parts (take $z = tx$ and divide by t). Put $\Delta m = m_1 - m_2$ and $a = \xi^2/(\gamma^2 + \xi^2)$. The remaining identity is between two nonconstant sigmoid maps. Their level hyperplanes coincide, so their score vectors are parallel. Taking the two limits along Δm identifies the endpoint values as am_1 and am_2 , possibly exchanged. In the same orientation, strict monotonicity of the sigmoid then identifies the linear argument within the sigmoid exactly. Thus the parameters must satisfy the following three conditions².

- (1) Matching of the linear part:

$$(\gamma^2 + \xi^2)VU = \frac{\gamma^2}{\gamma^2 + \xi^2}I_d,$$

- (2) Matching of the attention scores with the posterior weights:

$$U^\top(m_1 - m_2) = \frac{1}{\gamma^2 + \xi^2}(m_1 - m_2).$$

- (3) Preservation of the barycentric part with the correct scaling:

$$Vm_1 = \frac{\xi^2}{\gamma^2 + \xi^2}m_1, \quad Vm_2 = \frac{\xi^2}{\gamma^2 + \xi^2}m_2.$$

Since $\gamma^2 > 0$, the first condition now implies that both U and V are invertible.

From condition (1), we have

$$V = \frac{\gamma^2}{(\gamma^2 + \xi^2)^2}U^{-1}.$$

Using condition (3), we get, for $i = 1, 2$,

$$Um_i = \frac{\gamma^2}{\xi^2(\gamma^2 + \xi^2)}m_i, \quad i = 1, 2.$$

Thus m_1 and m_2 are eigenvectors of U associated with the same eigenvalue. Hence, we also have $U\Delta m = \frac{\gamma^2}{\xi^2(\gamma^2 + \xi^2)}\Delta m$. Therefore,

$$\Delta m^\top U^\top \Delta m = (U\Delta m)^\top \Delta m = \frac{\gamma^2}{\xi^2(\gamma^2 + \xi^2)}\|\Delta m\|^2.$$

On the other hand, condition (2) gives $U^\top \Delta m = \frac{1}{\gamma^2 + \xi^2}\Delta m$. Taking the scalar product with Δm , we obtain

$$\Delta m^\top U^\top \Delta m = \frac{1}{\gamma^2 + \xi^2}\|\Delta m\|^2.$$

²When $\pi_1 = \pi_2$, conditions (2) and (3) could alternatively be replaced by the sign-reversed score matching condition together with a permutation of the cluster means, but this does not affect the argument.

Since $m_1 \neq m_2$, comparing the two expressions yields

$$\frac{\gamma^2}{\xi^2(\gamma^2 + \xi^2)} = \frac{1}{\gamma^2 + \xi^2},$$

which is equivalent to $\gamma^2 = \xi^2$. For completeness, exchanging the endpoints instead gives $Vm_1 = am_2$, $Vm_2 = am_1$ and $U^\top \Delta m = -(\gamma^2 + \xi^2)^{-1} \Delta m$. The intercepts then require $\pi_1 = \pi_2$. Writing $q = \gamma^2/(\gamma^2 + \xi^2)^2$, the relation $VU = qI_d$ implies $U\Delta m = -(q/a)\Delta m$. Taking the scalar product with Δm again gives $q/a = (\gamma^2 + \xi^2)^{-1}$, hence the same necessary equality $\gamma^2 = \xi^2$. Consequently, if $\gamma^2 \neq \xi^2$, no single attention head can satisfy the three constraints simultaneously. $\square$

C.6 Two-head construction (Theorem 4.1, exact representation)

Proposition C.5 (Exact two-head representation in the nondegenerate case). *Assume $d \geq 2$, $\xi > 0$, $\Gamma = \gamma^2 I_d$ with $\gamma > 0$, $M = 2$, positive mixture weights, and distinct means with $\|m_1\| = \|m_2\|$. Then there exist two heads $(U_h, V_h)_{h=1}^2$ such that*

$$T_{U_1, V_1}[\mu_Z](z) + T_{U_2, V_2}[\mu_Z](z) = f^*(z) \quad \text{for every } z \in \mathbb{R}^d.$$

Proof. Let $\Delta m = m_1 - m_2$. We begin with the case $m_1 \neq -m_2$. We look for parameters (U_1, V_1) and (U_2, V_2) satisfying the following three conditions. First, both heads must reproduce the posterior weights:

$$U_i^\top \Delta m = \frac{1}{\gamma^2 + \xi^2} \Delta m, \quad i = 1, 2. \quad (\text{C1})$$

Second, each value matrix must select one cluster contribution, with the correct scaling:

$$V_i m_i = \frac{\xi^2}{\gamma^2 + \xi^2} m_i, \quad V_i m_j = 0, \quad (i, j) = (1, 2), (2, 1) \quad (\text{C2})$$

Third, the sum of the linear parts must reproduce the linear term of the optimal denoiser:

$$(\gamma^2 + \xi^2)(V_1 U_1 + V_2 U_2) = \frac{\gamma^2}{\gamma^2 + \xi^2} I_d. \quad (\text{C3})$$

These conditions are sufficient to exactly represent the optimal denoiser.

Since m_1 and m_2 are non-collinear, they span a two-dimensional subspace $E := \text{span}(m_1, m_2)$. We fix a complementary subspace $E^\perp$ and define V_1, V_2 as follows. On E , we impose

$$V_1 m_1 = \frac{\xi^2}{\gamma^2 + \xi^2} m_1, \quad V_1 m_2 = 0,$$

and

$$V_2 m_1 = 0, \quad V_2 m_2 = \frac{\xi^2}{\gamma^2 + \xi^2} m_2.$$

On the complementary subspace $E^\perp$, we set

$$V_1 x = V_2 x = \frac{\xi^2}{2(\gamma^2 + \xi^2)} x, \quad x \in E^\perp.$$

With this definition, condition (C2) is satisfied. Moreover,

$$V_1 + V_2 = \frac{\xi^2}{\gamma^2 + \xi^2} I_d.$$

We now set

$$U_1 = \frac{\gamma^2}{\xi^2(\gamma^2 + \xi^2)} I_d + \frac{\xi^2 - \gamma^2}{\xi^2(\gamma^2 + \xi^2)} \frac{m_2 \Delta m^\top}{\langle m_2, \Delta m \rangle},$$

and

$$U_2 = \frac{\gamma^2}{\xi^2(\gamma^2 + \xi^2)} I_d + \frac{\xi^2 - \gamma^2}{\xi^2(\gamma^2 + \xi^2)} \frac{m_1 \Delta m^\top}{\langle m_1, \Delta m \rangle}.$$

We first check condition (C1). For U_1 , we have

$$U_1^\top \Delta m = \frac{\gamma^2}{\xi^2(\gamma^2 + \xi^2)} \Delta m + \frac{\xi^2 - \gamma^2}{\xi^2(\gamma^2 + \xi^2)} \frac{\Delta m \langle m_2, \Delta m \rangle}{\langle m_2, \Delta m \rangle}.$$

Hence $U_1^\top \Delta m = \frac{1}{\gamma^2 + \xi^2} \Delta m$. The same computation gives $U_2^\top \Delta m = \frac{1}{\gamma^2 + \xi^2} \Delta m$. Thus condition (C1) is verified.

It remains to verify condition (C3). By construction,

$$V_1 m_2 = 0, \quad V_2 m_1 = 0.$$

Therefore, using the definitions of U_1 and U_2 ,

$$V_1 U_1 = \frac{\gamma^2}{\xi^2(\gamma^2 + \xi^2)} V_1, \quad V_2 U_2 = \frac{\gamma^2}{\xi^2(\gamma^2 + \xi^2)} V_2.$$

Consequently,

$$V_1 U_1 + V_2 U_2 = \frac{\gamma^2}{\xi^2(\gamma^2 + \xi^2)} (V_1 + V_2) = \frac{\gamma^2}{(\gamma^2 + \xi^2)^2} I_d.$$

We obtain

$$(\gamma^2 + \xi^2)(V_1 U_1 + V_2 U_2) = \frac{\gamma^2}{\gamma^2 + \xi^2} I_d,$$

which is precisely condition (C3).

Thus conditions (C1), (C2) and (C3) all hold, so the associated two-head predictor satisfies

$$\hat{f}(z) = f^*(z), \quad \text{for all } z \in \mathbb{R}^d.$$

This concludes the proof.

For the case $m_1 = -m_2 =: m$, the construction is similar. We replace condition (C2) by the following condition, which splits the contribution between the two heads:

$$V_i m = \frac{\xi^2}{2(\gamma^2 + \xi^2)} m, \quad i \in \{1, 2\} \quad (\text{C2}')$$

Let $n \in \mathbb{R}^d$ be a unit vector such that $\langle m, n \rangle = 0$. We define

$$U_1 = \frac{1}{\gamma^2 + \xi^2} \left(I_d + \frac{nm^\top}{\|m\|} \right), \quad U_2 = \frac{1}{\gamma^2 + \xi^2} \left(I_d - \frac{nm^\top}{\|m\|} \right),$$

and

$$\begin{aligned} V_1 &= \frac{\xi^2}{2(\gamma^2 + \xi^2)} \frac{mm^\top}{\|m\|^2} + \frac{\gamma^2}{2(\gamma^2 + \xi^2)} \left(I_d - \frac{mm^\top}{\|m\|^2} \right) + \frac{\gamma^2 - \xi^2}{2(\gamma^2 + \xi^2)} \frac{mn^\top}{\|m\|}, \\ V_2 &= \frac{\xi^2}{2(\gamma^2 + \xi^2)} \frac{mm^\top}{\|m\|^2} + \frac{\gamma^2}{2(\gamma^2 + \xi^2)} \left(I_d - \frac{mm^\top}{\|m\|^2} \right) - \frac{\gamma^2 - \xi^2}{2(\gamma^2 + \xi^2)} \frac{mn^\top}{\|m\|}. \end{aligned}$$

Here is the verification. Write $s = \gamma^2 + \xi^2$, $e = m/\|m\|$, $P = ee^\top$, $Q = I_d - P$, and

$$A = \frac{\xi^2}{2s} P + \frac{\gamma^2}{2s} Q, \quad B = \frac{\gamma^2 - \xi^2}{2s} en^\top.$$

Then $U_{1,2} = s^{-1}(I_d \pm ne^\top)$ and $V_{1,2} = A \pm B$. Since $n \perp e$, $U_i^\top m = s^{-1}m$ and $V_i m = \xi^2 m/(2s)$, proving (C1) and (C2'). Finally,

$$\begin{aligned} V_1 U_1 + V_2 U_2 &= \frac{2}{s} (A + Bne^\top) \\ &= \frac{1}{s^2} (\xi^2 P + \gamma^2 Q + (\gamma^2 - \xi^2) P) = \frac{\gamma^2}{s^2} I_d. \end{aligned}$$

This is (C3), completing the opposite-means construction. □

C.7 Strict risk gap (Theorem 4.1, separation claim)

Proposition C.6 (Softmax attention strictly outperforms linear attention). *Let $\mathcal{F}_{\text{sm}}$ and $\mathcal{F}_{\text{lin}}$ denote the classes represented by finite multi-head softmax- and linear-attention networks, respectively. In the nontrivial two-component setting considered above,*

$$\inf_{f \in \mathcal{F}_{\text{sm}}} (R(f) - R(f^*)) = 0, \quad \inf_{f \in \mathcal{F}_{\text{lin}}} (R(f) - R(f^*)) > 0.$$

Lemma C.7. *Let f_{lin} be a network composed of linear attention layers, with an arbitrary number of heads. Then f_{lin} is a linear function of z .*

Proof of Lemma C.7. As with softmax attention, linear attention can be represented as an operator acting on probability measures. Let $\hat{\mu}_L$ be the empirical measure associated with the i.i.d. samples $Z_1, \dots, Z_L$ drawn from μ . Then

$$T_{\text{lin}}[\hat{\mu}_L](z) := \frac{1}{L} \sum_{k=1}^L (Z_k^\top U z) V Z_k \xrightarrow{L \rightarrow \infty} V \left(\int z' z'^\top d\mu(z') \right) U z = T_{\text{lin}}[\mu](z)$$

Suppose now that $\mu = \mu_Z$ is the distribution of a Gaussian mixture, $\mu_Z = \pi_1 \mathcal{N}(m_1, S) + \pi_2 \mathcal{N}(m_2, S)$. Then,

$$\begin{aligned} T_{\text{lin}}[\mu_Z](z) &= V \left(\pi_1 \mathbb{E}_{Z_1 \sim \mathcal{N}(m_1, S)} [Z_1 Z_1^\top] + \pi_2 \mathbb{E}_{Z_2 \sim \mathcal{N}(m_2, S)} [Z_2 Z_2^\top] \right) U z \\ &= V (\pi_1 m_1 m_1^\top + \pi_2 m_2 m_2^\top + S) U z, \end{aligned}$$

Therefore, $T_{\text{lin}}[\mu_Z](z)$ is linear in z . Since linear combinations and compositions of linear maps remain linear, any network composed of multi-head linear attention layers is also linear in z . This concludes the proof. $\square$

Proof of Proposition C.6. We have already shown that a single-layer softmax attention network with two heads can attain the Bayes risk. It remains to prove that linear attention cannot achieve the same risk.

Let $f_{\text{lin}} \in \mathcal{F}_{\text{lin}}$. By Lemma C.7, there exists a matrix A such that $f_{\text{lin}}(z) = Az$. Moreover, there exists a matrix A^* minimizing the risk over all linear functions of z , namely

$$R(A^*) = \inf_{A \in \mathbb{R}^{d \times d}} R(A).$$

Indeed, we have

$$\begin{aligned} R(A) &= \mathbb{E} [\|X - AZ\|^2] \\ &= \mathbb{E} [\|(I_d - A)X - A\varepsilon\|^2] \\ &= \mathbb{E} [\|(I_d - A)X\|^2] + \mathbb{E} [\|A\varepsilon\|^2] \\ &= \mathbb{E} [\|(I_d - A)X\|^2] + \xi^2 |A|_{\mathbb{F}}^2. \end{aligned}$$

Thus, the risk is coercive and continuous as a function of A , and therefore admits a minimizer A^* . Since $f^*(z) = \mathbb{E}[X | Z = z]$ is the Bayes estimator for the squared loss, the excess-risk can be written for every measurable predictor g as

$$R(g) - R(f^*) = \mathbb{E} [\|g(Z) - f^*(Z)\|^2]$$

Consequently,

$$\begin{aligned} R(f_{\text{lin}}) - R(f^*) &\geq R(A^*) - R(f^*) \\ &= \mathbb{E} [\|A^*Z - f^*(Z)\|^2]. \end{aligned}$$

Using the expression of the optimal denoiser we obtain

$$R(A^*) - R(f^*) = \mathbb{E} [\|(A^* - \Gamma S^{-1})Z - h(Z)\|^2].$$

where $h(z) = \xi^2 S^{-1} \sum_k \beta_k(z) m_k$. Assume, for contradiction, that this excess risk is equal to zero. It follows that

$$(A^* - \Gamma S^{-1})z = h(z)$$

first μ_Z -almost everywhere, and then for every $z \in \mathbb{R}^d$: both sides are continuous and μ_Z has an everywhere positive density because $\xi > 0$. The function h is bounded so it forces

$$A^* - \Gamma S^{-1} = 0 \quad \text{and} \quad h = 0.$$

This is impossible : along the two directions $z = \pm t(m_1 - m_2)$, the $h(z)$ converge to the two different endpoints, so h is nonconstant. Therefore,

$$R(A^*) - R(f^*) > 0,$$

and thus

$$\inf_{f \in \mathcal{F}_{\text{lin}}} (R(f) - R(f^*)) > 0.$$

This concludes the proof. $\square$

C.8 Residual in-context denoising (proof of Proposition 4.4)

Proof of Proposition 4.4. Let $S = (\gamma^2 + \xi^2)I_d$ and set

$$a = \frac{\xi^2}{\gamma^2 + \xi^2}, \quad U_1 = S^{-1}, \quad V_1 = aI_d, \quad U_2 = V_2 = 0.$$

For any admissible noisy mixture μ_Z , Proposition 2.1 and the equal-norm assumption give

$$T_{U_1, V_1}[\mu_Z](z) = az + a \sum_{k=1}^2 \beta_k(z) m_k, \quad T_{U_2, V_2}[\mu_Z](z) = 0,$$

where the $\beta_k(z)$ are the posterior component probabilities. Hence, with $W_{\text{out}} = (I_d, -2aI_d)$,

$$\begin{aligned} W_{\text{out}} \begin{pmatrix} T_{U_1, V_1}^{\text{res}}[\mu_Z](z) \\ T_{U_2, V_2}^{\text{res}}[\mu_Z](z) \end{pmatrix} &= (1+a)z + a \sum_k \beta_k(z) m_k - 2az \\ &= (1-a)z + a \sum_k \beta_k(z) m_k \\ &= \mathbb{E}[X \mid Z = z]. \end{aligned}$$

The matrices depend only on the shared variances γ^2 and ξ^2 ; the prompt-specific means and weights enter solely through the attention operator. In addition, conditional expectation gives, for every task,

$$R_{\mu_Z}(f) - R_{\mu_Z}(f^*) = \mathbb{E}[\|f[\mu_Z](Z) - f^*[\mu_Z](Z)\|^2 \mid \mu_Z].$$

Integrating this nonnegative identity over any task distribution with finite risk proves optimality for the in-context objective, since the constructed predictor makes the right-hand side zero for every task. This proves Proposition 4.4. $\square$