A Critical β-Scale for Posterior Collapse in Dirichlet β-VAEs

Delphine Doutsas[1,2] **Bruno Figliuzzi**[1]
[1]Center for Statistics and Images (STIM), Mines Paris, PSL University, Fontainebleau, France
[2]Perceptual Systems Laboratory (LSP), École normale supérieure, PSL University, Paris, France
{delphine.doutsas, bruno.figliuzzi}@minesparis.psl.eu

Abstract

Posterior collapse is a common failure mode in variational autoencoders, but existing theory has mostly focused on Gaussian latent variables. Dirichlet VAEs are comparatively less understood, despite their use in settings where latent variables represent proportions, abundances, memberships, or convex mixture weights. We derive a closed-form local stability threshold β_c for collapse in symmetric Dirichlet β-VAEs with linear Gaussian decoders. This threshold shows that collapse in those models depends not only on the data spectrum, but also on the simplex dimension and prior concentration through a trigamma curvature factor induced by the Dirichlet KL. Although β_c is derived from a local analysis, we find that it accurately tracks the empirical onset of collapse beyond the local regime. We show its practical relevance in two real-world linear-mixture settings: hyperspectral unmixing and convex archetypal analysis. Across experiments, β/β_c consistently organizes collapse-related behavior, providing an interpretable coordinate and a practical way to narrow the search over the β hyperparameter before training. Taken together, these empirical findings suggest that β_c can serve, beyond its local derivation, as a practical reference point for anticipating collapse and guiding β selection in Dirichlet β-VAEs. Code at `https://github.com/dldou/dir-vae-collapse`.

1 Introduction

Variational autoencoders provide a general framework for learning latent-variable models with amortized inference (Kingma and Welling, 2014; Rezende et al., 2014). The β-VAE objective modifies the ELBO by multiplying the KL regularizer by a scalar β, initially as a way to encourage disentangled representations (Higgins et al., 2017; Burgess et al., 2018; Locatello et al., 2019). From an information-theoretic viewpoint, however, the ELBO does not by itself guarantee that the latent variable carries information about the input: it trades reconstruction distortion against a rate term, and increasing pressure on the rate can drive the mutual information toward zero (Alemi et al., 2018; Zhao et al., 2019). This observation is closely related to posterior collapse, the regime in which the approximate posterior becomes nearly equal to the prior.

Posterior collapse is now comparatively well studied for Gaussian-prior VAEs. In a general setting, optimality analyses derive fixed-point conditions for VAEs and show that β induces a tempered optimal posterior (Rezende and Viola, 2018; Camuto et al., 2021). In linear-Gaussian models, the VAE objective is closely connected to probabilistic PCA and the data covariance spectrum (Tipping and Bishop, 1999; Lucas et al., 2019; Rolinek et al., 2019). The linear-Gaussian analysis (Lucas et al., 2019) has been extended to a broader Gaussian linear latent-variable model that contains the linear-Gaussian-VAE as a special case (Wang and Ziyin, 2022). Characterizing its global minima yields exact conditions for complete and partial collapse. Explicit spectral thresholds have also been derived through learning-dynamics and high-dimensional analyses based on the data covariance

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

spectrum and decoder noise scale (Ichikawa and Hukushima, 2024, 2025). Related linear analyses cover conditional and hierarchical VAEs (Dang et al., 2024), while other work interprets collapse as a phase transition (Li et al., 2026). Training-dynamics studies provide another perspective, emphasizing inference lag and collapse-inducing local minima (He et al., 2019; Dai et al., 2020). These results give a quantitative language for Gaussian collapse.

In many applications, however, the relevant latent variables are not Euclidean or Gaussian but simplex-valued. Such variables arise when observations are modeled through proportions or convex weights, as in topic models (Blei et al., 2003), hyperspectral unmixing (Bioucas-Dias et al., 2012), archetypal analysis (Cutler and Breiman, 1994), and cell-type deconvolution (Miller et al., 2022; Nguyen et al., 2024). In these settings, the simplex coordinates are the object of interest: they represent abundances, memberships, or archetypal weights. Posterior collapse is therefore especially problematic, since the inferred proportions become input-independent while the model still appears to use a compositional latent space.

Dirichlet and simplex VAEs have been developed precisely for domains where such proportions are the object of interest. Existing work has addressed Dirichlet reparameterization and topic modeling (Joo et al., 2020; Burkhardt and Kramer, 2019), motif discovery (Kshirsagar et al., 2022), medical imaging (Harkness et al., 2023), and hyperspectral unmixing (Mantripragada and Qureshi, 2024; Doutsas and Figliuzzi, 2026). These works demonstrate that Dirichlet latents are useful in practice, but they mostly treat collapse as an empirical training issue or an architectural concern. What is missing is a local stability theory analogous to the Gaussian case: a formula that predicts when a Dirichlet posterior becomes locally stable at collapse and how that threshold depends on the simplex geometry, the prior concentration, the data spectrum, and the decoder noise.

This paper studies that question in the analytically tractable regime of a Dirichlet β-VAE with a linear-Gaussian decoder. We consider a symmetric prior $p(z) = \mathrm{Dir}(a\mathbf{1})$, a decoder $p_\theta(y \mid z) = \mathcal{N}(Wz + b, \sigma^2 I)$, and a Dirichlet variational posterior parameterized by its mean and concentration. Our main result is a closed-form local collapse threshold

$$\beta_c = \frac{\lambda_{\max}(\Sigma)}{\sigma^2} \frac{Ka + 1}{Ka^2 \psi_1(a)},$$

where K is the number of simplex coordinates, $\Sigma = \mathrm{Cov}(Y)$, and ψ_1 is the trigamma function. The numerator measures the largest reconstruction gain obtainable by an input-dependent tangent perturbation of the posterior mean, while the denominator is the Dirichlet-KL curvature opposing that perturbation. Empirically, β/β_c helps anticipate collapse in real-world applications, even beyond the assumptions of our local analysis.

2 A local collapse threshold for Dirichlet β-VAEs

Problem setup. Let $\mathcal{D} = \{y_i\}_{i=1}^N$ with $y_i \in \mathbb{R}^D$. Each observation is assumed to be the result of a low-dimensional latent composition $z_i \in \mathbb{R}^K$, with $K \ll D$, encoding proportions, abundances, or memberships over $K \geq 2$ unknown components. A common approximation consists in modeling the observed data as a linear mixture of these components. This yields the so-called linear mixing model (LMM) widely used in hyperspectral unmixing (Rasti et al., 2024; Bioucas-Dias et al., 2012), topic modeling (Blei et al., 2003), and cell-type deconvolution (Miller et al., 2022),

$$y_i = Wz_i + b + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_D), \qquad i = 1, \ldots, N,$$

or equivalently $y_i \mid z_i, W, b, \sigma^2 \sim \mathcal{N}(Wz_i + b, \sigma^2 I_D)$. Here, $W = [w_1, \ldots, w_K] \in \mathbb{R}^{D \times K}$ is an unknown dictionary, $b \in \mathbb{R}^D$ is an offset, and σ^2 is the noise variance. Since z_i represents mixture weights, it lies on the simplex

$$z_i \in \Delta^{K-1} := \{z \in \mathbb{R}^K : \sum_{k=1}^K z_k = 1,\ z_k \geq 0 \text{ for all } k\}.$$

The LMM gives a tractable setting in which to study posterior collapse locally and we restrict our analysis to this setting in the rest of the paper.

Dirichlet β-VAE. We consider a Dirichlet β-VAE with simplex-valued latent variable $z \in \Delta^{K-1}$ and linear-Gaussian decoder. The prior and amortized posterior are $p(z) = \mathrm{Dir}(\alpha)$ and $q_\phi(z \mid y) =$

$\mathrm{Dir}(\gamma_\phi(y))$. Recall that $\mathrm{Dir}(\eta)$ is a distribution on the simplex whose parameter $\eta \in \mathbb{R}^K_{>0}$ encodes both a mean η/η_0 and a total concentration $\eta_0 = \sum_k \eta_k$. Accordingly, we write

$$\gamma_\phi(y) = s_\phi(y) m_\phi(y), \qquad s_\phi(y) = \sum_{k=1}^{K} \gamma_{\phi,k}(y) > 0, \qquad m_\phi(y) = \frac{\gamma_\phi(y)}{s_\phi(y)} \in \Delta^{K-1},$$

where $m_\phi(y)$ is the posterior mean composition, while $s_\phi(y)$ is its concentration. The decoder is $p_\theta(y \mid z) = \mathcal{N}(Wz + b, \sigma^2 I_D)$, and the population negative β-ELBO is, up to additive constants,

$$\mathcal{L}_\beta(\phi, \theta) = \mathbb{E}_Y \left[\mathbb{E}_{q_\phi(z|Y)} \left[\frac{\|Y - Wz - b\|^2}{2\sigma^2} \right] + \beta \, \mathrm{KL}(q_\phi(z \mid Y) \| p(z)) \right]. \tag{1}$$

The scalar $\beta > 0$ is a hyperparameter fixed before training and varied in our experiments. When $\beta = 1$, (1) reduces to the standard VAE objective. Varying β changes the strength of the prior regularization and can induce different latent representations.

Collapse regime. The local analysis focuses on the degenerate regime in which the encoder is independent of the input. We say that the encoder has collapsed when $q_\phi(z|y) = p(z)$, or equivalently when $\gamma_\phi(y) = \alpha$. Throughout, we assume $s_0 = \sum_{k=1}^K \alpha_k > 0$, $m_0 := \alpha/s_0 \in \mathrm{int}(\Delta^{K-1})$, ensuring that the collapsed posterior mean lies in the simplex interior.

Empirically, we observe that past a certain β point the posterior collapses. This motivates the notion of a β-collapse threshold. We propose to identify a local threshold β_c through a local stability analysis of the collapsed regime. First, we characterize the stationary collapsed manifold $\mathcal{M}_{\mathrm{coll}}$.

Theorem 1 (Geometry of stationary collapse)**.** *Let $\mathcal{T} := \{v \in \mathbb{R}^K : \mathbf{1}^\top v = 0\}$ denote the tangent space of the simplex, and let $\bar{y} = \mathbb{E}[Y]$.*

(i) Condition for encoder stationarity at collapse. *Given (W, b), the collapsed posterior is stationary w.r.t. admissible encoder perturbations, i.e., perturbations that stay within the simplex, if and only if $Wv = 0 \; \forall v \in \mathcal{T}$ (or equivalently, $W = c\mathbf{1}^\top$ for some $c \in \mathbb{R}^D$). Thus no condition is imposed on b, and stationarity holds exactly when all decoder columns are identical.*

(ii) Decoder minimizers at collapsed stationarity. *Under collapsed encoder stationarity, the negative β-ELBO is minimized over (W, b) exactly on the collapsed stationary manifold*

$$\mathcal{M}_{\mathrm{coll}} = \left\{ q(z \mid y) \equiv p(z), \; W = c\mathbf{1}^\top, \; b = \bar{y} - c, \; c \in \mathbb{R}^D \right\}. \tag{2}$$

Proof sketch of (i). ($\Leftarrow$): Take any admissible encoder perturbation $m_t = m_0 + t\eta$, $s_t = s_0 + t\xi$, with $\eta \in \mathcal{T}$, $s_t > 0$. If $Wv = 0$ for all $v \in \mathcal{T}$, then inserting (m_t, s_t) in Eq 1 and differentiating at $t = 0$ gives zero first variation: the decoder is insensitive to tangent perturbations and the KL is stationary at $q = p$. ($\Rightarrow$): Isolating the variance term in Eq 1 with path $m_t = m_0$, $s_t = s_0 + t$ forces $\mathrm{tr}(WS_0W^\top) = 0$. Since $S_0 := (s_0 + 1)\mathrm{Cov}_p(z)$ has image $\mathcal{T}$, $W = c\mathbf{1}^\top$.
Proof sketch of (ii). At collapse, $q = p$, so the KL term of Eq 1 is constant w.r.t (W, b). For fixed W, minimizing Eq 1 over b gives $b^\star = \bar{y} - Wm_0$. Substituting $b^\star$ back into the objective, the remaining minimization over W is achieved exactly when the decoder has no tangent component, i.e. $Wv = 0$ for all $v \in \mathcal{T}$. By part (i), this is equivalent to $W = c\mathbf{1}^\top$ for some $c \in \mathbb{R}^D$. Hence $b^\star = \bar{y} - c$, since $\mathbf{1}^\top m_0 = 1$, which gives exactly $\mathcal{M}_{\mathrm{coll}}$. Full proofs are in Appendix A. □

To test escape from collapse, we perturb the collapsed manifold along small valid encoder-decoder paths that move away from it. Along such a path, the second-order approximation of the objective is $\mathcal{L}_\beta(t) \approx \mathcal{L}_\beta(0) - t^2 Q_\beta$ where $-Q_\beta$ is the quadratic form associated with the Hessian. Thus Q_β can be thought of as an escape direction, i.e $Q_\beta > 0$ means that the perturbation lowers the objective.

Theorem 2 (Local collapse threshold)**.** *Let $\Sigma = \mathrm{Cov}(Y)$, let $\lambda_{\max}(\Sigma) > 0$ be its largest eigenvalue, and assume $p(z) = \mathrm{Dir}(a, \ldots, a)$, with $a > 0$. Let ψ_1 denote the trigamma function. Define*

$$\beta_c = \frac{\lambda_{\max}(\Sigma)}{\sigma^2} \frac{Ka + 1}{Ka^2 \psi_1(a)}. \tag{3}$$

Then, near $\mathcal{M}_{\mathrm{coll}}$, after projecting decoder perturbations onto the simplex tangent space,

1. *if $\beta < \beta_c$, there exists a perturbation with $Q_\beta > 0$ and collapse is a saddle;*

2. *if $\beta > \beta_c$, then $-\mathcal{Q}_\beta$ is coercive, hence collapse is a strict local minimum;*

3. *if $\beta = \beta_c$ we cannot conclude.*

Proof sketch. We perturb a collapsed point $(W_0, b_0, m_0, s_0) \in \mathcal{M}_{\text{coll}}$ by considering paths $m_t = m_0 + t\eta$, $s_t = s_0 + t\xi$, $W_t = W_0 + tA$, and $b_t = b_0 + tu$, with t sufficiently small so that it is admissible, i.e. $m_t(Y) \in \Delta^{K-1}$ and $s_t(Y) > 0$. The second-order expansion of the loss leads to

$$\mathcal{L}_\beta(t) - \mathcal{L}_\beta(0) = -t^2 \mathcal{Q}_\beta(A, u, \eta, \xi) + o(t^2)$$

where $\mathcal{Q}_\beta(A, u, \eta, \xi)$ is the quadratic form associated to the Hessian (up to a factor -2) along the perturbation.

We want a condition on β such that there exists a direction (A, u, η, ξ) able to decrease $\mathcal{L}_\beta$ to second order, in particular, whether $\sup \mathcal{Q}_\beta(A, u, \eta, \xi) > 0$ over admissible perturbations. Fixing η, we remark that ξ only contributes a nonpositive term, maximized at $\xi = 0$, and that Q_β is concave in (A, u) and the supremum is attained up to a projection on the tangent space. We define the resulting optimal escape score by $\mathcal{Q}^\star_\beta(\eta) := \sup_{(A,u,\xi)} \mathcal{Q}_\beta(A, u, \eta, \xi)$, Thus an escape encoder direction $(\eta, 0)$ exists iff $\mathcal{Q}^\star_\beta(\eta) > 0$. To get β_{c}, we assume that the prior is symmetric yielding the Hessian $\mathcal{H}_\alpha = \psi_1(a) I_K - \psi_1(Ka)\mathbf{1}\mathbf{1}^\top$ and the closed-form optimal escape score

$$\mathcal{Q}^\star_\beta(\eta) = \frac{K(Ka+1)}{2\sigma^2} \left\| \mathbb{E}\left[(Y - \bar{y})\eta(Y)^\mathsf{T}\right] \right\|_{\text{F}}^2 - \frac{\beta}{2}\left(K^2 a^2 \psi_1(a) \mathbb{E}\left\|\eta(Y)\right\|^2\right). \tag{4}$$

Now, the question is how much reconstruction gain the posterior-mean perturbation $\eta(Y)$ can obtain. Cauchy-Schwarz gives $\left\| \mathbb{E}\left[(Y - \bar{y})\eta(Y)^\mathsf{T}\right] \right\|_{\text{F}}^2 \leq \lambda_{\max}(\Sigma) \mathbb{E}\left\|\eta(Y)\right\|^2$. Substituting into Eq. (4) gives

$$\mathcal{Q}^\star_\beta(\eta) \leq \frac{K^2 a^2 \psi_1(a)}{2}(\beta_{\text{c}} - \beta) \mathbb{E}\left\|\eta(Y)\right\|^2. \tag{5}$$

For $\beta < \beta_{\text{c}}$, let $e_{\max}$ be a top unit eigenvector of the data covariance Σ. The bound is sharp and can be reached by $\eta(Y) = [e_{\max}^\top (Y - \bar{y})]v$, with $v \in \mathcal{T}$ a fixed unit vector tangent to the simplex. Then, to ensure admissibility of perturbations, we use bounded truncations of $\eta(Y)$ to approach this bound arbitrarily closely. This gives a perturbation direction with $\mathcal{Q}_\beta > 0$, hence collapse is a saddle.

For $\beta > \beta_{\text{c}}$, the upper bound in Eq. (5) is strictly negative for $\eta \neq 0$. Together with the quadratic decoder and concentration costs, this gives coercivity of $-\mathcal{Q}_\beta$ after tangent projection. Uniform remainder control then yields a strict local minimum. Full details are in Appendix A. □

Eq. 4 makes the collapse escape mechanism explicit: it requires a reconstruction-active tangent variation of the posterior mean through $\mathbb{E}\left[(Y - \bar{y})\eta(Y)^\mathsf{T}\right]$. Perturbation on the concentration $\xi(Y)$ can only cost. Thus, acting only on ξ can be insufficient to escape collapse.

This equation also shows that β_{c} depends on the Dirichlet concentration a. This raises the calibration question: which collapse thresholds can be reached by tuning a, and is the required value of a unique? The following proposition answers this by characterizing the attainable range and proving one-to-one calibration on that range.

Proposition 1 (Prior-shift calibration law)**.** *Define, for $a > 0$, $K \geq 2$ and $\lambda_{\max}(\Sigma) > 0$, $g_K(a) := \frac{Ka+1}{Ka^2\psi_1(a)}$ and $\beta_c(a) = \frac{\lambda_{\max}(\Sigma)}{\sigma^2} g_K(a)$. Then $\beta_c(a) \in \left(\frac{\lambda_{\max}(\Sigma)}{K\sigma^2}, \frac{\lambda_{\max}(\Sigma)}{\sigma^2}\right)$, and every value in this interval is attained for a unique $a > 0$.*

Proof sketch. For fixed K, $\lambda_{\max}(\Sigma)$, and σ^2, the raw collapse threshold changes with the prior concentration through $g_K(a)$. The function g_K is differentiable and strictly increasing on $(0, \infty)$. Moreover, $\lim_{a\to 0} g_K(a) = 1/K$, $\lim_{a\to\infty} g_K(a) = 1$. Thus $\beta_c(a)$ is in the interval. □

Note that $\beta_c(a)$ increases with a, so a sparse prior ($a < 1$) makes collapse locally stable at lower β values, while a dense prior ($a > 1$) requires higher β to induce collapse. Conversely, $\beta_c(K)$ is strictly decreasing in K, thus increasing the latent dimension makes collapse locally stable at lower β values.

Table 1 compares our local result for Dirichlet latent variables with the global Gaussian characterization (Wang and Ziyin, 2022), where the encoder and decoder are linear and the latent space is

Table 1: Gaussian and Dirichlet collapse with the same decoder $p_\theta(y \mid z) = \mathcal{N}(Wz + b, \sigma^2 I_D)$ and fixed σ^2. C is a learned diagonal covariance. At collapse, both models satisfy $q_\phi = p$ and predict $\bar{y}$.

	Gaussian linear β-VAE	**Dirichlet β-VAE**
Prior	$p(z) = \mathcal{N}(0, I)$	$p(z) = \mathrm{Dir}(a\mathbf{1})$
Posterior	$q_\phi(z \mid y) = \mathcal{N}(V(y - \bar{y}), C)$	$q_\phi(z \mid y) = \mathrm{Dir}(\gamma_\phi(y))$ with $\gamma_\phi(y) = s_\phi(y) m_\phi(y)$
Posterior at collapse	$V^* = 0, \quad C^* = I$ $\Longrightarrow q^*(z \mid y) = \mathcal{N}(0, I) = p(z)$	$m^*(y) = \frac{1}{K}\mathbf{1}, \quad s^*(y) = Ka$ $\Longrightarrow q^*(z \mid y) = \mathrm{Dir}(a\mathbf{1}) = p(z)$
Decoder at collapse	$W^* = 0, \quad b^* = \bar{y}$ $\Longrightarrow \mathbb{E}_\theta[Y \mid z] = \bar{y}$	$W^* = c\mathbf{1}^\top, \quad b^* = \bar{y} - c$ $\Longrightarrow W^* z + b^* = \bar{y}, \quad \forall z \in \Delta^{K-1}$
Threshold	$\beta_c^{\mathrm{Gauss}} = \dfrac{\lambda_{\max}(\Sigma_Y)}{\sigma^2}$	$\beta_c^{\mathrm{Dir}} = \dfrac{\lambda_{\max}(\Sigma_Y)}{\sigma^2} \dfrac{Ka + 1}{Ka^2 \psi_1(a)}$
Theoretical scope	Exact global minima of a fully linear Gaussian β-VAE. β_c^{Gauss} marks when complete collapse is globally optimal.	Local stability analysis on collapse manifold $\mathcal{M}_{\mathrm{coll}}$. β_c^{Dir} marks when collapse is locally optimal. Only the decoder is linear Gaussian.

Gaussian. Interestingly, both thresholds share the spectral factor $\lambda_{\max}(\Sigma)/\sigma^2$, but the Dirichlet case adds a KL curvature term $g_K(a)$.

Finally, Eq. 3 motivates the following practical collapse-risk estimation algorithm.

Practical collapse-risk estimation algorithm. Given a dataset and a Dirichlet β-VAE with latent dimension K, symmetric prior concentration a, KL weight β, and decoder noise scale σ^2, proceed as follows:

1. Estimate the data covariance Σ and $\lambda_{\max}(\Sigma)$ from the centered training data.
2. Set σ^2 from the decoder likelihood, or from an application-specific residual-noise estimate.
3. Compute $\widehat{\beta}_c = \frac{\widehat{\lambda}_{\max}(\Sigma)}{\sigma^2} \frac{Ka+1}{Ka^2\psi_1(a)}$ and $\frac{\beta}{\widehat{\beta}_c}$.
4. Treat $\beta/\widehat{\beta}_c \gtrsim 1$ as high collapse risk.

3 Experiments

Despite the local nature of our analysis in Section 2 and its restrictive assumptions, we test whether its predictions can help practitioners anticipate collapse. In this work, we consider blind source separation in hyperspectral imaging (HSI) and convex archetypal analysis (CAA), where some of these assumptions may not hold. Our experiments include initialization far from collapse, nonnegativity constraints on the decoder or shallow non-linear decoders, and an ablation study on encoder capacity. We also test the prior-shift calibration of Proposition 1 on synthetic data. Numerical checks of the theoretical results are reported in Appendix C.

3.1 Common protocol and metrics

Model and training. Unless stated otherwise, for each fixed β, we train a Dirichlet β-VAE with an amortized encoder and a linear decoder by minimizing the negative β-ELBO. The default encoder is a two-hidden-layer Tanh MLP with width 128 followed by a softplus output plus a constant $\gamma_{\min} = 10^{-4}$ for numerical stability. The decoder is a linear map $Wz + b$, except for HSI experiments where $b = 0$ and nonnegativity is enforced on the decoder weights by softplus or by clamping. We use Adam (Kingma and Ba, 2014) with learning rate $3 \cdot 10^{-3}$, a batch size of 256 and no weight decay. The KL term is computed analytically, while the reconstruction term is evaluated either analytically

or by Monte Carlo sampling. Parameter σ^2 is known in synthetic experiments and estimated for HSI and image datasets. More details on the loss can be found in Appendix C.2. We compute $\widehat{\beta}_{\mathrm{c}}$ using the practical collapse-risk estimation algorithm of Section 2 and select fixed values of $\beta/\widehat{\beta}_{\mathrm{c}}$ for training.

Collapse metrics. Following Section 2, we propose to track three complementary metrics defined below. First, we monitor the average KL divergence between the posterior and the prior, $\overline{\mathrm{KL}}_\phi$. Second, the projected decoder norm $\|WP_T\|_F$ quantifies the deviation of the columns of W from their mean since all columns coincide at exact collapse. Third, we track U_m to probe the latent usage of the model, quantifying the variability of posterior mean compositions across samples. At collapse, the encoder assigns the same mean composition to every sample. These metrics vanish at exact collapse.

$$\overline{\mathrm{KL}}_\phi = \frac{1}{N}\sum_{i=1}^{N} \mathrm{KL}(\mathrm{Dir}(\gamma_i) \,\|\, \mathrm{Dir}(a\mathbf{1}_K))\,, \quad \|W_T\|_F = \|WP_T\|_F, \quad U_m = \frac{1}{N}\sum_{i=1}^{N} \|m_i - \bar{m}\|^2, \tag{6}$$

where $m_i = \gamma_i / \sum_k \gamma_{ik}$, $\bar{m} = \frac{1}{N}\sum_i m_i$, $P_T = I_K - K^{-1}\mathbf{1}_K\mathbf{1}_K^\top$ is the orthogonal projector onto the simplex tangent space and W is the decoder matrix. See Appendix B for more details on implementation. Finally, all results are reported against $\beta/\widehat{\beta}_{\mathrm{c}}$.

3.2 Practical simplex inverse problems

We first test whether $\beta/\widehat{\beta}_{\mathrm{c}}$ organizes collapse in settings where the simplex coordinates have a direct semantic interpretation. These experiments are stress tests of the threshold β_{c} on real applications, not benchmark comparisons.

Hyperspectral unmixing. In blind hyperspectral unmixing, each pixel spectrum is modeled as a convex combination of nonnegative endmember spectra (Bioucas-Dias et al., 2012; Rasti et al., 2024),

$$y_i \approx W z_i, \qquad z_i \in \Delta^{K-1}, \qquad W \in \mathbb{R}_{\geq 0}^{d\times K}.$$

Here z_i is the abundance vector and the columns of W are the learned endmember spectra. We use the Samson dataset which has $N = 95 \times 95 = 9025$ pixels, $d = 156$ spectral bands, and $K = 3$ ground-truth endmembers (Zhu, 2017). The scene is normalized by its global maximum reflectance. Additional HSI experiments on the JasperRidge dataset can be found in Appendix D.

Training settings. We train the model with symmetric prior concentration $a = 1$, Xavier-uniform initialization (Glorot and Bengio, 2010), and enforce the nonnegative constraint on the decoder either by adding a softplus activation or by clamping after each optimization step. We sweep $\beta/\widehat{\beta}_{\mathrm{c}} \in \{0.1, 0.5, 0.8, 0.9, 0.95, 1.0, 1.1, 1.2, 1.5, 2.0\}$, using five seeds and 600 epochs. The empirical collapse scale uses the largest eigenvalue of the empirical covariance and a residual noise estimate obtained after subtracting a centered rank-K truncated SVD approximation. For Samson, this gives $\lambda_{\max}(\widehat{\Sigma}) = 2.6895$, $\widehat{\sigma}^2 = 3.1554 \times 10^{-5}$, and $\widehat{\beta}_{\mathrm{c}} = 6.91 \times 10^4$. The rank-$K$ residual estimator is used as a practical fixed noise calibration.

Metrics. Given ground-truth endmembers $W^* = [w_1^*, \ldots, w_K^*]$ and learned endmembers $\widehat{W} = [\hat{w}_1, \ldots, \hat{w}_K]$, which are the decoder's weights, we report the RMSE on observations, the spectral angle distance (SAD) between ground-truth and learned endmembers, with the Hungarian algorithm to find the optimal matching permutation π, and the RMSE between the ground-truth abundances z_i and the learned abundances $\hat{z}_i$ (posterior mean), after aligning the learned abundances with the same permutation π. We define these metrics as follows:

$$\mathrm{RMSE}_{\mathrm{rec}}(Y, \widehat{Y}) = \left(\frac{1}{dN}\sum_{i=1}^{N} \|y_i - \hat{y}_i\|_2^2\right)^{1/2}, \qquad \mathcal{L}_{\mathrm{SAD}}(W^\star, \widehat{W}) = \frac{1}{K}\sum_{k=1}^{K} s(w_k^*, \hat{w}_{\pi(k)}),$$

$$\mathcal{L}_{\mathrm{ARMSE}}(Z^\star, \widehat{Z}\,; \pi) = \left(\frac{1}{KN}\sum_{k=1}^{K}\sum_{i=1}^{N} \left(z_{i,k}^\star - \hat{z}_{i,\pi(k)}\right)^2\right)^{1/2},$$

where $s(u, v) = \arccos(\langle u, v\rangle / (\|u\|_2 \|v\|_2))$. For more results on HSI, see Appendix D.

Table 2: **Samson HSI collapse diagnostics.** We use a fully connected decoder W with softplus activation to enforce nonnegativity, prior concentration $a = 1$, and Xavier-uniform initialization. The diagnostics correspond to the softplus curve in Figure 1. Values are mean $\pm$ standard deviation over five seeds.

$\frac{\beta}{\hat{\beta}_c}$	$\overline{\mathrm{KL}_\phi}$	U_m	$\|W_T\|_F$
0.1	$1.6 \pm 5.03 \times 10^{-3}$	$1.89 \times 10^{-1} \pm 1.37 \times 10^{-3}$	$4.05 \pm 1.26 \times 10^{-3}$
0.5	$4.07 \times 10^{-1} \pm 1.78 \times 10^{-3}$	$5.77 \times 10^{-2} \pm 2.76 \times 10^{-4}$	$3.33 \pm 3.25 \times 10^{-3}$
0.8	$1.75 \times 10^{-1} \pm 5.52 \times 10^{-3}$	$2.44 \times 10^{-2} \pm 6.34 \times 10^{-4}$	$2.73 \pm 6.87 \times 10^{-3}$
0.9	$1.01 \times 10^{-1} \pm 8.97 \times 10^{-4}$	$1.4 \times 10^{-2} \pm 1.14 \times 10^{-4}$	$2.17 \pm 5.88 \times 10^{-3}$
0.95	$6.01 \times 10^{-2} \pm 6.8 \times 10^{-4}$	$8.34 \times 10^{-3} \pm 1.02 \times 10^{-4}$	$1.73 \pm 5.22 \times 10^{-3}$
1.0	$(3.1 \pm 5.17) \times 10^{-4}$	$(3.78 \pm 6.11) \times 10^{-7}$	$(3.27 \pm 3.34) \times 10^{-3}$
1.1	$(2.57 \pm 4.22) \times 10^{-5}$	$7.48 \times 10^{-7} \pm 1.17 \times 10^{-6}$	$1.11 \times 10^{-3} \pm 4.79 \times 10^{-4}$
1.2	$(3.37 \pm 5.75) \times 10^{-4}$	$(1.37 \pm 2.27) \times 10^{-5}$	$3.53 \times 10^{-4} \pm 9.53 \times 10^{-5}$
1.5	$(2.08 \pm 1.91) \times 10^{-4}$	$(3.43 \pm 5.68) \times 10^{-6}$	$(2.56 \pm 1.23) \times 10^{-4}$
2.0	$(3.21 \pm 4.69) \times 10^{-5}$	$(1.16 \pm 1.96) \times 10^{-6}$	$(3.30 \pm 3.13) \times 10^{-4}$

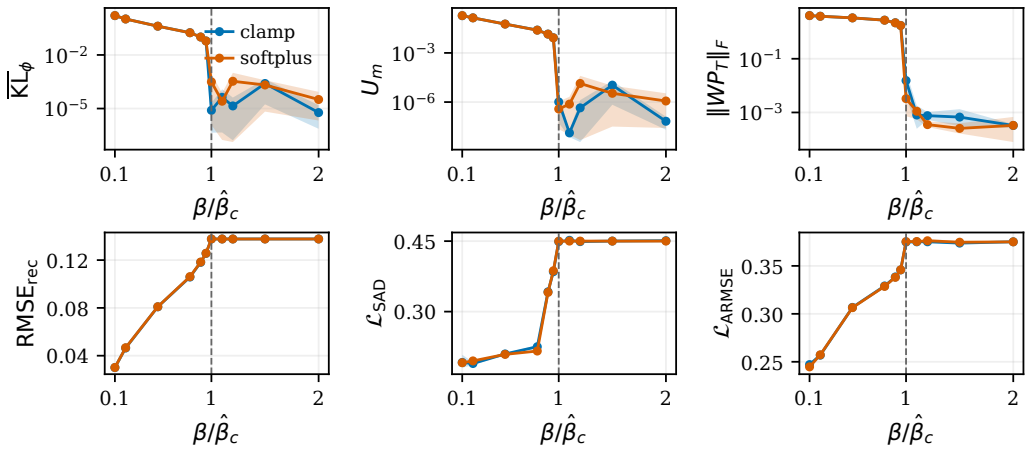


Figure 1: **Samson HSI sweep.** Curves show means over five seeds, with shaded bands spanning the minimum-maximum range. The dashed vertical line marks $\beta/\widehat{\beta_c} = 1$, the expected collapse threshold. Both nonnegative decoders show large latent activity below the predicted threshold and nearly zero activity at, or just above it. The unmixing metrics follow the same sharp transition, indicating the loss of latent information.

Figure 1 and Table 2 show a sharp transition near the predicted threshold $\beta/\widehat{\beta_c} = 1$. The model is active for $\beta/\widehat{\beta_c} < 1$ and the corresponding task metrics are in the informative regime. At $\beta/\widehat{\beta_c} = 0.95$, around the predicted threshold, the representation contracts. Above the threshold, the posterior is essentially inactive. These results demonstrate that, despite the limitations of the theoretical framework, β_c remains a practical scale for anticipating collapse in real HSI applications.

Convex archetypal analysis. As a complementary visual test, we use a convex archetypal reconstruction task on image triplets (Alcacer et al., 2026). Images are reconstructed as convex combinations of prototypes W,

$$\hat{y}_i = W z_i + b, \qquad z_i \in \Delta^{K-1}.$$

For MNIST (LeCun et al., 2010) and SVHN (Netzer et al., 2011), we select three classes, $0, 1, 7$ for illustration and set $K = 3$ so that the mean composition can be visualized in the 2-dimensional simplex. The class labels are not given to the model and are used only for coloring the simplex plots. In this representation, the decoder's columns plus the bias define the archetypes at the vertices of the simplex, and the posterior mean compositions are plotted as points in the simplex. Therefore, collapse has a geometric signature: all posterior means contract toward the prior's mean, the barycenter of the

simplex here, and the decoder archetypes become indistinguishable. For each dataset, we train the Dirichlet β-VAE described in Sec. 3.1 with prior concentration parameter $a = 1$ and for 3000 epochs. All quantities are reported against $\beta/\widehat{\beta}_c$.

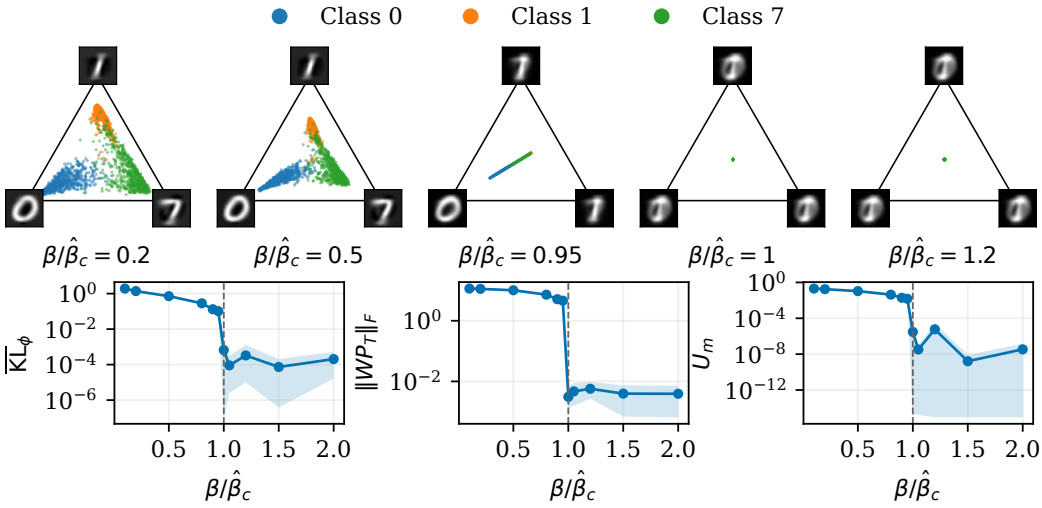


Figure 2: **CAA on MNIST classes** $0, 1, 7$**.** Top: posterior mean visualization in the $K = 3$ simplex, with learned decoder archetypes shown at the vertices. Above the threshold $\beta/\widehat{\beta}_c = 1$, all posterior means contract to the barycenter and vertices become indistinguishable. Bottom: collapse diagnostics. Curves show means over five seeds for the average encoder KL, the tangent decoder norm $\|WP_T\|_F$, and the posterior-mean usage U_m. Shaded bands span the minimum-maximum range across seeds. The dashed vertical line marks the predicted local collapse threshold. Below the threshold, posterior means occupy the simplex in a class-dependent way. Above the threshold $\beta/\widehat{\beta}_c = 1$, the occupancy contracts to the barycenter.

Table 3: **MNIST convex archetypal collapse summary.** MNIST classes $0, 1, 7$, $K = 3$, symmetric prior concentration $a = 1$, and 3000 epochs. Values are mean $\pm$ standard deviation over five seeds. All the collapse metrics almost vanish for $\beta/\widehat{\beta}_c \geq 1$.

$\beta/\widehat{\beta}_c$	average KL	$\|W_T\|_F$	U_m
0.2	$1.39 \pm 2.13 \times 10^{-3}$	$1.13 \times 10 \pm 3.69 \times 10^{-3}$	$1.77 \times 10^{-1} \pm 5.60 \times 10^{-4}$
0.5	$7.18 \times 10^{-1} \pm 5.50 \times 10^{-3}$	$1.01 \times 10 \pm 1.16 \times 10^{-2}$	$1.07 \times 10^{-1} \pm 1.35 \times 10^{-3}$
0.95	$1.03 \times 10^{-1} \pm 1.15 \times 10^{-3}$	$4.60 \pm 9.90 \times 10^{-3}$	$1.51 \times 10^{-2} \pm 4.45 \times 10^{-5}$
1.0	$6.58 \times 10^{-4} \pm 1.23 \times 10^{-3}$	$(3.09 \pm 2.43) \times 10^{-3}$	$(3.04 \pm 6.77) \times 10^{-6}$
1.2	$(3.28 \pm 5.37) \times 10^{-4}$	$(5.72 \pm 2.72) \times 10^{-3}$	$5.89 \times 10^{-6} \pm 1.30 \times 10^{-5}$

Figure 2 and Table 3 show the same transition in a semantic visualization and in the three collapse diagnostics. For $\beta/\widehat{\beta}_c = 0.2$, the three digit classes occupy distinct regions of the simplex and the vertices correspond to recognizable digit archetypes. At $\beta/\widehat{\beta}_c = 0.5$, the representation remains informative but is visibly contracted. By $\beta/\widehat{\beta}_c = 0.95$, the occupancy has nearly collapsed to a 1-dimensional space, suggesting loss of one effective tangent degree of freedom. At $\beta/\widehat{\beta}_c = 1$ and above, the posterior means concentrate near the barycenter and the representation becomes nearly input-independent which is characteristic of a collapsed posterior. Quantitatively, the collapse metrics follow the same pattern and drop sharply for $\beta/\widehat{\beta}_c \geq 1$.

We repeat the experiment on SVHN classes $0, 1, 7$, using 9000 images, $K = 3$ and $a = 1$. The encoder has two hidden layers of width 512 since the input dimension is larger ($32 \times 32 \times 3 = 3072$). We use Xavier-uniform initialization and train for 3000 epochs with Adam and learning rate 3×10^{-4} for learning stability. The full protocol and additional results are given in Appendix E.2.

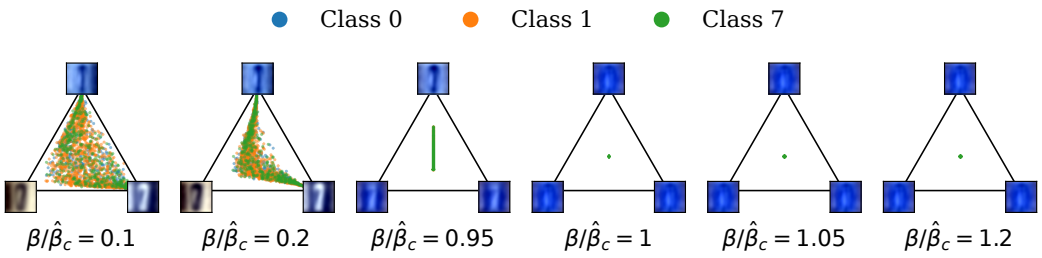


Figure 3: **Convex archetypal analysis on SVHN classes** $0, 1, 7$**.** Posterior-mean compositions and decoded archetypes with the linear decoder, $a = 1$, and seed 0. Each ratio $\beta/\hat{\beta}_c$ is shown below its panel. Class labels are used only to color the points.

Figure 3 shows the same contraction on color images. However, visual inspection suggests that the learned archetypes correspond to variations in lighting or contrast rather than digit identity. Similarly, posterior means contract to a 1-dimensional subspace near $\beta/\hat{\beta}_c = 0.95$ and concentrate at the barycenter at $\beta/\hat{\beta}_c = 1$. Table 17 shows that the collapse metrics follow a similar pattern across five seeds and almost vanish around the threshold. More details on the training protocol and diagnostics are given in Appendix E.2. We also report a replication on FashionMNIST classes $1, 3, 8$ in Appendix E.3, with the corresponding simplex figure, training protocol, and detailed diagnostics.

3.3 Decoder nonlinearity and encoder capacity

Further experiments test on Samson and MNIST whether β_c remains predictive of the collapse behavior in the presence of shallow nonlinear decoders using the linear β_c as an empirical reference. Figure 4 shows that the posterior KL and U_m drop sharply between $\beta/\beta_c = 1$ and 1.05.

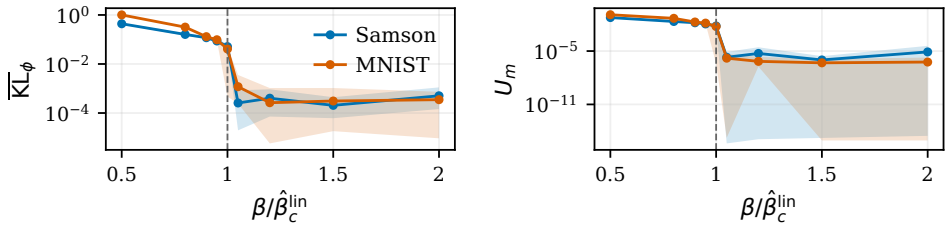


Figure 4: **Nonlinear decoders.** Means over five seeds, with minimum-maximum bands. The dashed vertical line marks the threshold predicted for the linear model.

To test the encoder capacity, we also perform experiments on the same datasets where we retain linear decoders and compare encoders with no hidden layer, the reference architecture, and deeper or wider networks. Posterior activity decreases near the same reference scale, while residual KL varies across architectures and training settings. Architectures, training protocols, and full results are given in Appendices F and G.

3.4 Prior-shift calibration

The final experiment tests the prior-calibration prediction of Proposition 1: changing the symmetric Dirichlet concentration a should shift the raw collapse scale according to $\beta_c(a)$, while the normalized coordinate $\beta/\beta_c(a)$ should remain comparable across priors.

Dataset. We use a synthetic dataset to control the data-generating prior, set to $a_{\mathrm{data}} = 1$. We use $N = 8000$ samples, $d = 150$, $K = 20$. The mixing matrix W_{data} is generated with $W_T = \tau GH$, where $G_{ij} \sim \mathcal{N}(0, 1)$ and $H = I_K - K^{-1}\mathbf{1}\mathbf{1}^\top$. The noise is normalized by $\sigma = \sigma_{\mathrm{factor}} s_{\mathrm{sig}}$, giving an SNR lower bound of approximately 9 dB along the leading eigenvector. More details can be found in Appendix C.1.

Training settings. Keeping the dataset fixed, we train the Dirichlet VAE described in Sec. 3.1. For each $a \in \{0.1, 0.2, 0.3, 0.5, 0.8, 1, 1.5, 2, 4, 6\}$, we compute $\beta_c(a)$ from the empirical covariance and train on a fixed grid of $\beta/\beta_c(a)$. Models use Xavier-uniform initialization, full-batch Adam for 2000 epochs, learning rate of 3×10^{-3} and no weight decay. We run three initialization seeds per $(a, \beta/\beta_c(a))$ configuration.

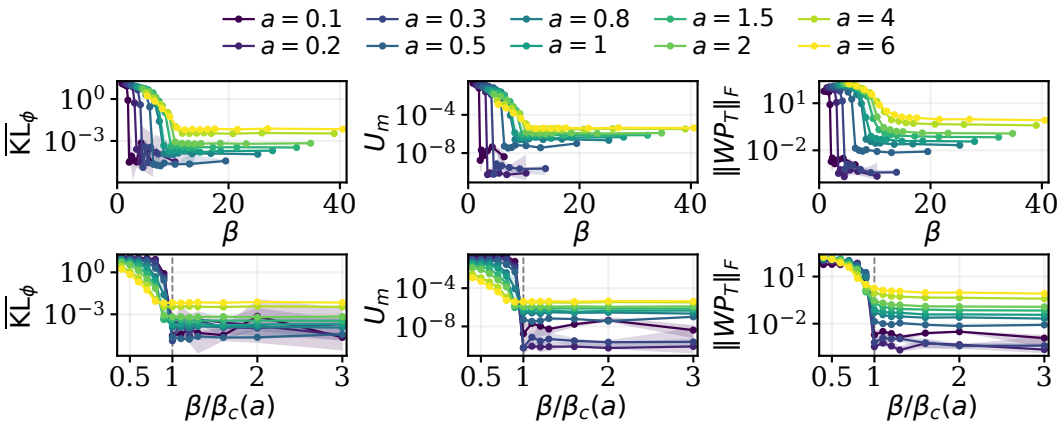


Figure 5: **Prior-shift calibration.** Top: collapse metrics against raw β. Bottom: same metrics against $\beta/\beta_c(a)$. Curves show medians over three seeds, with min-max bands. The threshold increases as a increases, and normalizing β by $\beta_c(a)$ approximately aligns the transitions across priors.

Results. Figure 5 shows that the prior shifts the collapse threshold on the raw β axis for all collapse metrics, but the curves realign when plotted against $\beta/\beta_c(a) \in \{0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.6, 2.0, 3.0\}$. This confirms Prop. 1. Full numerical results and the $K = 50$ counterpart are in Appendix C.6.

4 Limitations

Our theoretical result is a local stability statement around the collapsed manifold. It identifies when collapse becomes locally stable or unstable to second order, but it does not prove global stability of VAE training. The closed-form threshold also relies on restrictive assumptions: a symmetric Dirichlet prior, a linear-Gaussian decoder, and a fixed likelihood noise scale. These assumptions make the role of simplex geometry and Dirichlet KL curvature explicit, but they do not cover the full range of decoder architectures such as deep decoders. Another practical limitation is the calibration of β_c. The threshold depends on $\lambda_{\max}(\Sigma)$ and σ^2. While the covariance term is directly estimable from data, the likelihood noise is often unknown in real applications and must be specified or approximated. Poor noise calibration can therefore shift the numerical collapse boundary. For this reason, we interpret β/β_c as a normalized collapse-risk coordinate rather than as an exact universal decision rule. The experiments go beyond the assumptions of the theorems, including random initialization, shallow non-linear or non-negative decoders, and real-world datasets with estimated likelihood noise. The empirical results support the usefulness of β/β_c as an organizing scale, but they should be read as stress tests of the collapse threshold rather than as guarantees. Finally, β_c predicts the onset of collapse, not the task-optimal value of β.

5 Conclusion

We derived a closed-form local collapse threshold for symmetric Dirichlet β-VAEs with linear-Gaussian decoders. The threshold combines the data spectrum, decoder noise, simplex dimension, and Dirichlet KL curvature into a single scale. Although the derivation is local, β_c empirically organizes collapse behavior across our experiments, suggesting that β/β_c is a useful practical coordinate for diagnosing posterior-collapse regimes in Dirichlet VAEs. Thus, in this setting, collapse is better tracked by the dimensionless ratio β/β_c than by the raw KL weight β.

Acknowledgments and Disclosure of Funding

We gratefully thank Emile Dolicanin (Sorbonne University, Paris, France) for carefully reviewing the mathematical proofs. Any remaining errors are our own. We are also grateful to the anonymous reviewers for their constructive comments and suggestions, which have helped improve the quality of this work. Finally, we thank The Transition Institute 1.5 - Mines Paris - PSL University for supporting this work.

The authors declare no competing interests.

References

Aleix Alcacer, Irene Epifanio, Sebastian Mair, and Morten Mørup. A survey on archetypal analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a broken ELBO. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 159–168. PMLR, 2018.

José M. Bioucas-Dias, Antonio Plaza, Nicolas Dobigeon, Mario Parente, Qian Du, Paul Gader, and Jocelyn Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2):354–379, 2012. doi: 10.1109/JSTARS.2012.2194696.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE, 2018.

Sophie Burkhardt and Stefan Kramer. Decoupling sparsity and smoothness in the Dirichlet variational autoencoder topic model. *Journal of Machine Learning Research*, 20(131):1–27, 2019.

Alexander Camuto, Matthew Willetts, Stephen Roberts, Chris Holmes, and Tom Rainforth. Towards a theoretical understanding of the robustness of variational autoencoders. In *International Conference on Artificial Intelligence and Statistics*, pages 3565–3573. PMLR, 2021.

Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994. doi: 10.1080/00401706.1994.10485840.

Bin Dai, Ziyu Wang, and David Wipf. The usual suspects? Reassessing blame for VAE posterior collapse. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2313–2322. PMLR, 2020.

Hien Dang, Tho-Huu Tran, Tan Nguyen, and Nhat Ho. Beyond vanilla variational autoencoders: Detecting posterior collapse in conditional and hierarchical variational autoencoders. In *International Conference on Learning Representations*, volume 2024, pages 9160–9198, 2024.

Delphine Doutsas and Bruno Figliuzzi. Effective-dimension control of beta in dirichlet beta-VAE for blind hyperspectral unmixing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 7762–7771, 2026.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

Rachael Harkness, Alejandro F. Frangi, Kieran Zucker, and Nishant Ravikumar. Learning disentangled representations for explainable chest X-ray classification using Dirichlet VAEs. In *Medical Imaging 2023: Image Processing*, 2023. doi: 10.1117/12.2654345.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. In *International Conference on Learning Representations*, 2019.

Irina Higgins, Loic Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Yuma Ichikawa and Koji Hukushima. Learning dynamics in linear VAE: Posterior collapse threshold, superfluous latent space pitfalls, and speedup with KL annealing. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1936–1944. PMLR, 2024.

Yuma Ichikawa and Koji Hukushima. High-dimensional asymptotics of VAEs: Threshold of posterior collapse and dataset-size dependence of rate-distortion curve. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(7):073402, 2025. doi: 10.1088/1742-5468/adde3e.

Weonyoung Joo, Wonsung Lee, Sungrae Park, and Il-Chul Moon. Dirichlet variational autoencoder. *Pattern Recognition*, 107:107514, 2020. doi: 10.1016/j.patcog.2020.107514.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Meghana Kshirsagar, Han Yuan, Juan Lavista Ferres, and Christina Leslie. Bindvae: Dirichlet variational autoencoders for de novo motif discovery from accessible chromatin. *Genome biology*, 23(1):174, 2022.

Yann LeCun, Corinna Cortes, and C. J. Burges. MNIST handwritten digit database, 2010. URL `http://yann.lecun.com/exdb/mnist/`.

Zhen Li, Fan Zhang, Zheng Zhang, and Yu Chen. Posterior collapse as a phase transition in variational autoencoders. *Physica A: Statistical Mechanics and its Applications*, 683:131228, 2026. doi: 10.1016/j.physa.2025.131228.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schoelkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124. PMLR, 2019.

James Lucas, George Tucker, Roger Grosse, and Mohammad Norouzi. Don't blame the ELBO! A linear VAE perspective on posterior collapse. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Kiran Mantripragada and Faisal Z. Qureshi. Hyperspectral pixel unmixing with latent Dirichlet variational autoencoder. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024. doi: 10.1109/TGRS.2024.3357589.

Brendan F. Miller, Feiyang Huang, Lyla Atta, Arpan Sahoo, and Jean Fan. Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. *Nature Communications*, 13(1):2339, 2022. doi: 10.1038/s41467-022-30033-z.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Hung Nguyen, Ha Nguyen, Duc Tran, Sorin Draghici, and Tin Nguyen. Fourteen years of cellular deconvolution: methodology, applications, technical evaluation and outstanding challenges. *Nucleic Acids Research*, 52(9):4761–4783, 2024.

Behnood Rasti, Alexandre Zouaoui, Julien Mairal, and Jocelyn Chanussot. Image processing and machine learning for hyperspectral unmixing: An overview and the hysupp python package. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–31, 2024.

Danilo Jimenez Rezende and Fabio Viola. Taming VAEs. *arXiv preprint arXiv:1810.00597*, 2018.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286. PMLR, 2014.

Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue PCA directions (by accident). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12406–12415, 2019.

Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B*, 61(3):611–622, 1999. doi: 10.1111/1467-9868.00196.

Zihao Wang and Liu Ziyin. Posterior collapse of a linear latent variable model. *Advances in Neural Information Processing Systems*, 35:37537–37548, 2022.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. InfoVAE: Balancing learning and inference in variational autoencoders. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, pages 5885–5892, 2019. doi: 10.1609/aaai.v33i01.33015885.

Feiyun Zhu. Hyperspectral unmixing: ground truth labeling, datasets, benchmark performances and survey. *arXiv preprint arXiv:1708.05125*, 2017.

Appendix contents

A Additional details and proofs of the theoretical results 14
B Collapse metrics 30
C Theorem validation experiments 31
D Hyperspectral imaging (HSI): unmixing details and additional experiments 46
E Convex archetypal analysis experiments: supplementary details 49
F Nonlinear decoder experiments 53
G Encoder architecture ablation 56
H NeurIPS Paper Checklist 62

A Additional details and proofs of the theoretical results

A.1 Dirichlet preliminaries

For $a > 0$, we denote by S_a the covariance of the symmetric Dirichlet distribution:

$$S_a := \mathrm{Cov}_{z\sim\mathrm{Dir}(a\mathbf{1})}(z) = \frac{1}{K(Ka+1)}P_{\mathrm{T}}, \qquad P_{\mathrm{T}} := I_K - \frac{1}{K}\mathbf{1}\mathbf{1}^\top.$$

Lemma 1 (Second-order derivatives of the Dirichlet KL divergence)**.** *Let $K \geq 2$. Assume $p(z) = \mathrm{Dir}(\alpha_0)$ with $\alpha_0 \in \mathbb{R}^K_{>0}$ fixed, denote $\alpha_{0,+} := \sum_{k=1}^K \alpha_{0,k}$, and define*

$$\kappa(\gamma) := \mathrm{KL}(\mathrm{Dir}(\gamma) \,\|\, \mathrm{Dir}(\alpha_0)) \text{ for } \gamma \in \mathbb{R}^K_{>0}.$$

Then,

1. *κ is C^∞ on $\mathbb{R}^K_{>0}$, and*
$$\kappa(\alpha_0) = 0, \qquad \nabla_\gamma \kappa(\alpha_0) = 0.$$
Its Hessian at α_0 is
$$H_\gamma := \nabla^2_\gamma \kappa(\alpha_0) = \mathrm{Diag}\big(\psi_1(\alpha_{0,1}), \ldots, \psi_1(\alpha_{0,K})\big) - \psi_1(\alpha_{0,+})\mathbf{1}\mathbf{1}^\top, \tag{7}$$
where $\psi_0 = \Gamma'/\Gamma$ is the digamma function and $\psi_1 = \psi_0'$ is the trigamma function.
2. *If the prior is symmetric, i.e., $\alpha_0 = a\mathbf{1}$ for some $a > 0$, the Hessian of the Dirichlet-Dirichlet KL at $\gamma = a\mathbf{1}$ is*
$$H_a = \psi_1(a) I_K - \psi_1(Ka)\mathbf{1}\mathbf{1}^\top.$$

Proof of Lemma 1.

1. The closed-form Dirichlet KL is
$$\begin{aligned}\mathrm{KL}\left(\mathrm{Dir}(\gamma)\|\mathrm{Dir}(\alpha_0)\right) = \log\frac{\Gamma(\sum_{k=1}^K \gamma_k)}{\Gamma(\sum_{k=1}^K \alpha_{0,k})} + \sum_{k=1}^K \log\frac{\Gamma(\alpha_{0,k})}{\Gamma(\gamma_k)} \\ + \sum_{k=1}^K (\gamma_k - \alpha_{0,k})\left[\psi_0(\gamma_k) - \psi_0(\sum_{k=1}^K \gamma_k)\right].\end{aligned}$$
Since $\log\Gamma$ and ψ_0 are smooth on $\mathbb{R}_{>0}$, this expression shows that κ is smooth on $\mathbb{R}^K_{>0}$. Write $\psi_2 := \psi_1'$.

For all $i = 1, \ldots, K$, we get
$$\frac{\partial}{\partial\gamma_i}\mathrm{KL}\left(\mathrm{Dir}(\gamma) \,\|\, \mathrm{Dir}(\alpha_0)\right) = (\gamma_i - \alpha_{0,i})\psi_1(\gamma_i) - \left(\sum_{k=1}^K \gamma_k - \sum_{k=1}^K \alpha_{0,k}\right)\psi_1\left(\sum_{\ell=1}^K \gamma_\ell\right).$$

Differentiating once more, the Hessian entries are

$$\frac{\partial}{\partial\gamma_j}\frac{\partial}{\partial\gamma_i}\mathrm{KL}\left(\mathrm{Dir}(\gamma)\,\|\,\mathrm{Dir}(\alpha_0)\right) = \mathbf{1}_{\{i=j\}}\psi_1(\gamma_i) + \mathbf{1}_{\{i=j\}}(\gamma_i-\alpha_{0,i})\psi_2(\gamma_i) - \psi_1\left(\sum_{\ell=1}^K\gamma_\ell\right) - \left(\sum_{k=1}^K\gamma_k - \sum_{k=1}^K\alpha_{0,k}\right)\psi_2\left(\sum_{\ell=1}^K\gamma_\ell\right)$$

that can be rewritten in matrix form as

$$\begin{aligned}&\nabla^2_\gamma\mathrm{KL}\left(\mathrm{Dir}(\gamma)\,\|\,\mathrm{Dir}(\alpha_0)\right)\\ &= \mathrm{Diag}\left(\psi_1(\gamma_1),\ldots,\psi_1(\gamma_K)\right) + \mathrm{Diag}\left((\gamma_1-\alpha_{0,1})\psi_2(\gamma_1),\ldots,(\gamma_K-\alpha_{0,K})\psi_2(\gamma_K)\right)\\ &\quad - \left[\psi_1\left(\sum_{\ell=1}^K\gamma_\ell\right) + \left(\sum_{k=1}^K\gamma_k - \sum_{k=1}^K\alpha_{0,k}\right)\psi_2\left(\sum_{\ell=1}^K\gamma_\ell\right)\right]\mathbf{1}\mathbf{1}^\top.\end{aligned}$$

At $\gamma=\alpha_0$, the closed-form expression and the gradient give $\kappa(\alpha_0)=0$ and $\nabla_\gamma\kappa(\alpha_0)=0$. In the Hessian, the terms containing $\gamma_i-\alpha_{0,i}$ or $\sum_k\gamma_k-\sum_k\alpha_{0,k}$ vanish, yielding Equation (7).

2. The expression for H_a follows from part 1 by setting $\alpha_{0,i}=a$.

□

A.2 Proof of Equation 1

For completeness, we derive Eq. 1 from the population negative β-ELBO. For $Y\in\mathbb{R}^D$, assume the likelihood is Gaussian, $p_\theta(y\mid z)=\mathcal{N}(Wz+b,\sigma^2 I_D)$, with fixed $\sigma^2>0$, the prior is $p(z)=\mathrm{Dir}(\alpha)$, and the amortized posterior is $q_\phi(z\mid y)=\mathrm{Dir}(\gamma_\phi(y))$. Then,

$$\begin{aligned}&\mathbb{E}_Y\left[\mathbb{E}_{q_\phi(z|Y)}\left[-\log p_\theta(Y\mid z)\right]+\beta\,\mathrm{KL}(q_\phi(z\mid Y)\|p(z))\right]\\ &= \mathbb{E}_Y\left[\mathbb{E}_{q_\phi(z|Y)}\left[\frac{D}{2}\log(2\pi\sigma^2)+\frac{1}{2\sigma^2}\|Y-Wz-b\|^2\right]+\beta\,\mathrm{KL}(q_\phi(z\mid Y)\|p(z))\right]\\ &= \frac{D}{2}\log(2\pi\sigma^2)+\mathcal{L}_\beta(\phi,\theta).\end{aligned}$$

Since σ^2 is fixed, the first term in the last line is constant. Dropping this constant leaves the minimizers unchanged and yields the objective in Eq. 1.

A.3 Admissible perturbations

For the prior $p(z)=\mathrm{Dir}(\alpha)$, write $s_0:=\sum_{k=1}^K\alpha_k$ and $m_0:=\alpha/s_0$. Throughout the local analysis, assume $Y\in L^2(P_Y;\mathbb{R}^D)$ and $\sigma^2>0$. The simplex tangent space and its orthogonal projector are

$$\mathcal{T}=\{v\in\mathbb{R}^K:\mathbf{1}^\top v=0\},\qquad P_{\mathrm{T}}=I_K-\frac{1}{K}\mathbf{1}\mathbf{1}^\top.$$

Indeed, for every $v\in\mathbb{R}^K$, $\mathbf{1}^\top P_{\mathrm{T}}v=0$ and $v-P_{\mathrm{T}}v=K^{-1}(\mathbf{1}^\top v)\mathbf{1}\in\mathcal{T}^\perp$. Thus $P_{\mathrm{T}}v\in\mathcal{T}$ and the residual is orthogonal to $\mathcal{T}$, which proves that P_{T} is its orthogonal projector. An admissible posterior perturbation around collapse is a pair

$$\eta\in L^2(P_Y;\mathcal{T}),\qquad \xi\in L^2(P_Y;\mathbb{R}),$$

such that there exists $\varepsilon>0$ for which, for P_Y-almost every Y and all $|t|<\varepsilon$,

$$m_t(Y)=m_0+t\eta(Y)\in\mathrm{int}(\Delta^{K-1}),\qquad s_t(Y)=s_0+t\xi(Y)>0.$$

Admissibility on this common two-sided interval implies $\eta,\xi\in L^\infty(P_Y)$, by evaluating these constraints at $t=\pm\varepsilon/2$. A full admissible encoder-decoder direction is a quadruple (A,u,η,ξ), with $A\in\mathbb{R}^{D\times K}$, $u\in\mathbb{R}^D$, and (η,ξ) as above, along the path

$$W_t=W_0+tA,\qquad b_t=b_0+tu,\qquad q_t(z|Y)=\mathrm{Dir}(s_t(Y)m_t(Y)).$$

Local neighborhoods are defined in $\mathbb{R}^{D\times K}\times\mathbb{R}^D\times L^\infty(P_Y;\mathcal{T})\times L^\infty(P_Y;\mathbb{R})$ by the norm

$$\|(A,u,\eta,\xi)\|_{\mathrm{par}}:=\|A\|_F+\|u\|_2+\|\eta\|_{L^\infty(P_Y)}+\|\xi\|_{L^\infty(P_Y)},\tag{8}$$

where $\|\eta\|_{L^\infty(P_Y)}=\mathrm{ess\,sup}_{Y\sim P_Y}\|\eta(Y)\|_2$ and $\|\xi\|_{L^\infty(P_Y)}=\mathrm{ess\,sup}_{Y\sim P_Y}|\xi(Y)|$. The norms on A and u are the Frobenius and Euclidean norms, respectively.

A.4 Proof of Theorem 1

We observe that past a certain threshold β_{coll}, when $\beta > \beta_{\text{coll}}$, the model stays stuck in a collapsed state or around it. This suggests that collapse may correspond to a local minimum in the negative β-ELBO landscape. The main idea of the paper is to study the local geometry around those stationary points by performing a Taylor expansion around them. However, we need to exhibit the properties of such points. The purpose of Theorem 1 is to characterize the collapsed points that are stationary points too.

We start by recalling the definition of encoder collapse in VAEs.

Definition 1 (Collapsed encoder)**.** *Let y be an observation and z be a latent variable. Let $p(z) = \mathrm{Dir}(\alpha)$ be the prior distribution over the latent variable, with $\alpha \in \mathbb{R}^K_{>0}$, and let $q_\phi(z|y) = \mathrm{Dir}(\gamma_\phi(y))$ be the amortized posterior distribution over the latent variable, with $\gamma_\phi(y) \in \mathbb{R}^K_{>0}$. We say that the encoder has collapsed when $q_\phi(z|y) = p(z)$, or equivalently when $\gamma_\phi(y) = \alpha$ for all y.*

Empirically we observe a persistent collapse for $\beta > \beta_{\text{coll}}$. Therefore, we seek collapsed configurations at which the objective is jointly stationary with respect to encoder and decoder. The first part of Theorem 1 determines the condition on a fixed decoder under which a collapsed encoder is a stationary point. The second part gives the explicit form of the decoder minimizing the objective function when the encoder is held collapsed, yielding the collapsed stationary manifold.

Theorem 1 (Geometry of stationary collapse)**.** *Let $\mathcal{T} := \{v \in \mathbb{R}^K : \mathbf{1}^\top v = 0\}$ denote the tangent space of the simplex, and let $\bar{y} = \mathbb{E}[Y]$.*

(i) Condition for encoder stationarity at collapse. *Given (W, b), the collapsed posterior is stationary w.r.t. admissible encoder perturbations, i.e., perturbations that stay within the simplex, if and only if $Wv = 0\ \forall v \in \mathcal{T}$ (or equivalently, $W = c\mathbf{1}^\top$ for some $c \in \mathbb{R}^D$). Thus no condition is imposed on b, and stationarity holds exactly when all decoder columns are identical.*

(ii) Decoder minimizers at collapsed stationarity. *Under collapsed encoder stationarity, the negative β-ELBO is minimized over (W, b) exactly on the collapsed stationary manifold*

$$\mathcal{M}_{\text{coll}} = \left\{ q(z \mid y) \equiv p(z),\ W = c\mathbf{1}^\top,\ b = \bar{y} - c,\ c \in \mathbb{R}^D \right\}. \tag{2}$$

Proof. Assume $\mathbb{E}\,\|Y\|^2 < \infty$ and set $\Sigma = \mathrm{Cov}(Y)$.

Proof of (i). ($\Leftarrow$): Assume that $Wv = 0$ for every $v \in \mathcal{T}$. Then there exists some $c \in \mathbb{R}^D$ such that $W = c\mathbf{1}^\top$. Hence, for every $z \in \Delta^{K-1}$,

$$Wz = c\mathbf{1}^\top z = c.$$

Therefore the decoder output $Wz + b$ is independent of z on the simplex: the reconstruction term is locally independent of the encoder distribution. The KL term is stationary at $q = p$, since $q = p$ is its minimizer. Thus the collapsed encoder is stationary with respect to all admissible encoder perturbations.

($\Rightarrow$): Assume that the collapsed encoder is stationary for fixed (W, b). Denote the prior's total concentration by $\alpha_0 = \mathbf{1}^\top \alpha$, its mean by $m_0 = \alpha/\alpha_0$, and its scaled covariance by $S_0 = \mathrm{diag}(m_0) - m_0 m_0^\top = (\alpha_0 + 1)\mathrm{Cov}_p(z)$. As stationarity holds for all admissible encoder perturbations, we consider the following particular admissible path for t sufficiently small,

$$m_t(Y) = m_0, \qquad s_t(Y) = \alpha_0 + t, \qquad q_t(z|Y) = \mathrm{Dir}(s_t m_t),$$

where the mean is fixed and the concentration varies. The Dirichlet covariance formula gives

$$\mathrm{Cov}_{q_t}(z) = \frac{S_0}{\alpha_0 + t + 1}.$$

Using the bias-variance decomposition, we can write the reconstruction term as

$$\begin{aligned} R(t) &= \mathbb{E}_Y \mathbb{E}_{q_t(z|Y)} \left[\frac{\|Y - Wz - b\|^2}{2\sigma^2} \right] \\ &= \frac{1}{2\sigma^2} \mathbb{E} \|Y - Wm_t - b\|^2 + \frac{1}{2\sigma^2} \mathrm{tr} \left(W \mathrm{Cov}_{q_t}(z) W^\top \right) \\ &= \frac{1}{2\sigma^2} \mathbb{E} \|Y - Wm_0 - b\|^2 + \frac{1}{2\sigma^2} \frac{\mathrm{tr}(W S_0 W^\top)}{\alpha_0 + t + 1}. \end{aligned}$$

Since the KL term is minimized at $q = p$, the first derivative of the KL term with respect to any admissible encoder perturbation is zero. Stationarity therefore implies

$$\left. \frac{d}{dt} \right|_{t=0} R(t) = 0 = -\frac{\mathrm{tr}(W S_0 W^\top)}{2\sigma^2(\alpha_0 + 1)^2} \implies \mathrm{tr}(W S_0 W^\top) = 0.$$

Since S_0 is positive semidefinite, we have

$$\mathrm{tr}(W S_0 W^\top) = \left\| W S_0^{1/2} \right\|_{\mathrm{F}}^2,$$

so $W S_0^{1/2} = 0$. Because m_0 is in the interior of the simplex, we have $\mathrm{Im}(S_0) = \mathcal{T}$, this gives $Wv = 0$ for every $v \in \mathcal{T}$. Equivalently, because every row of W belongs to the normal space $\mathcal{T}^\perp = \mathrm{span}\{\mathbf{1}^\top\}$, we get $W = c\mathbf{1}^\top$ for some $c \in \mathbb{R}^D$. This proves the fixed-decoder stationarity condition.

Proof of (ii). ($\Rightarrow$): Under the collapse constraint $q = p$, the KL term is zero. Therefore, the collapsed objective is simply the reconstruction term. Using its bias-variance decomposition together with the independence of z from Y, we have

$$\mathcal{L}_\beta^{\mathrm{coll}}(W, b) = \frac{1}{2\sigma^2} \mathbb{E} \left[\|Y - Wm_0 - b\|^2 \right] + \frac{1}{2\sigma^2(\alpha_0 + 1)} \mathrm{tr}(W S_0 W^\top).$$

For fixed W, the first term is minimized over b at $b^\star = \bar{y} - Wm_0$. Substituting this value gives

$$\mathcal{L}_\beta^{\mathrm{coll}}(W, b^\star) = \frac{1}{2\sigma^2} \mathrm{tr}(\Sigma) + \frac{1}{2\sigma^2(\alpha_0 + 1)} \mathrm{tr}(W S_0 W^\top),$$

The second term is nonnegative, and is minimized exactly when $\mathrm{tr}(W S_0 W^\top) = 0$. Since $S_0 \succeq 0$ and m_0 is in the interior of the simplex, we have $\mathrm{Im}(S_0) = \mathcal{T}$, and hence $Wv = 0$ for every $v \in \mathcal{T}$, or equivalently $W = c\mathbf{1}^\top$ for some $c \in \mathbb{R}^D$. Using $\mathbf{1}^\top m_0 = 1$, the optimal bias is then

$$b^\star = \bar{y} - Wm_0 = \bar{y} - c\mathbf{1}^\top m_0 = \bar{y} - c.$$

Therefore the collapsed minimizers are exactly

$$q = p, \qquad W^\star = c\mathbf{1}^\top, \qquad b^\star = \bar{y} - c, \qquad c \in \mathbb{R}^D.$$

This is precisely $\mathcal{M}_{\mathrm{coll}}$.

($\Leftarrow$): If $W = c\mathbf{1}^\top$ and $b = \bar{y} - c$, then for every $z \in \Delta^{K-1}$,

$$Wz + b = c\mathbf{1}^\top z + \bar{y} - c = \bar{y}.$$

Hence the reconstruction term is minimized under collapse. The KL term is minimized by $q = p$. This proves the characterization of $\mathcal{M}_{\mathrm{coll}}$. □

A.5 Proof of Theorem 2: Local collapse β-threshold

This subsection aims at proving Theorem 2 which intuitively exhibits a threshold β_c such that, for $\beta > \beta_c$, points in the collapsed manifold $\mathcal{M}_{\mathrm{coll}}$ are strict local minima of the negative β-ELBO after projecting decoder perturbations onto the simplex tangent space, while for $\beta < \beta_c$, they are saddle points. The proof is long and technical. We therefore propose to decompose it into lemmas as follows:

1. We perform a Taylor expansion of $\mathcal{L}_\beta$ around a point in $\mathcal{M}_{\text{coll}}$ (Lemma 2);
2. We use the KL Hessian at the prior and its symmetric expression (Lemma 1) to identify the KL contribution to the quadratic form. Under the symmetric prior, we separate the composition and concentration contributions and establish positive KL curvature in the concentration direction (Lemma 3);
3. *Saddle directions for $\beta < \beta_c$.*
   (a) We maximize $\mathcal{Q}_\beta$ over the decoder perturbations (A, u) and set $\xi = 0$, which maximizes the profiled score for fixed η, to obtain $\mathcal{Q}^\star_\beta(\eta)$ (Lemma 3);
   (b) We choose a mean perturbation using the projection of the data onto a top covariance eigenvector and use bounded truncations to make it admissible. This identifies the critical value β_c in Eq. 3 and gives an admissible joint perturbation with $\mathcal{Q}_\beta > 0$ for $\beta < \beta_c$;
   (c) We control the Taylor remainder along this fixed direction to obtain a decrease of $\mathcal{L}_\beta$ (Lemma 2).
4. We prove strict local minimality for $\beta > \beta_c$, with the $L^\infty(P_Y)$ topology on posterior perturbations.
   (a) We project the decoder perturbations onto the simplex tangent space and adjust the bias (Lemma 4);
   (b) We prove that $-\mathcal{Q}_\beta$ is coercive using a spectral bound and positive concentration curvature (Lemma 3);
   (c) We control the KL and reconstruction remainders uniformly. For sufficiently small perturbations, coercivity yields a strict local minimum after projection.
5. At $\beta = \beta_c$, the coercivity constant degenerates, so the second-order analysis does not determine local stability.

Theorem 2 (Local collapse threshold)**.** *Let $\Sigma = \mathrm{Cov}(Y)$, let $\lambda_{\max}(\Sigma) > 0$ be its largest eigenvalue, and assume $p(z) = \mathrm{Dir}(a, \dots, a)$, with $a > 0$. Let ψ_1 denote the trigamma function. Define*

$$\beta_c = \frac{\lambda_{\max}(\Sigma)}{\sigma^2} \frac{Ka + 1}{Ka^2 \psi_1(a)}. \tag{3}$$

Then, near $\mathcal{M}_{\text{coll}}$ in the topology specified in Appendix A.3, after projecting decoder perturbations onto the simplex tangent space,

1. *if $\beta < \beta_c$, there exists a perturbation with $Q_\beta > 0$ and collapse is a saddle;*
2. *if $\beta > \beta_c$, then $-\mathcal{Q}_\beta$ is coercive with respect to the decoder norms and the posterior $L^2(P_Y)$ norms, and collapse is a strict local minimum;*
3. *if $\beta = \beta_c$ we cannot conclude.*

Proof. Fix a collapse point $(W_0, b_0, p) \in \mathcal{M}_{\text{coll}}$ and consider an admissible perturbation (A, u, η, ξ), such that

$$W_t = W_0 + tA, \qquad b_t = b_0 + tu,$$
$$m_t(Y) = m_0 + t\eta(Y), \quad s_t(Y) = Ka + t\xi(Y).$$

Taylor expansion around (W_0, b_0, p). By Lemma 2, we have

$$\mathcal{L}_\beta(W_t, b_t, q_t) = \mathcal{L}_\beta(W_0, b_0, p) - t^2 \mathcal{Q}_\beta(A, u, \eta, \xi) + o(t^2),$$

where

$$\mathcal{Q}_\beta(A, u, \eta, \xi) = \frac{1}{\sigma^2} \left\langle A, \mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right] \right\rangle - \frac{1}{2\sigma^2} \|u + Am_0\|^2 - \frac{1}{2\sigma^2} \mathrm{tr}(AS_aA^\top)$$
$$- \frac{\beta}{2} \mathbb{E}\left[\delta\gamma(Y)^\top H_a \delta\gamma(Y)\right].$$

with $\delta\gamma(Y) = Ka\,\eta(Y) + \frac{\xi(Y)}{K}\mathbf{1}$.

Symmetric prior. Applying Lemma 3 to the KL term gives

$$\begin{aligned}\mathcal{Q}_\beta(A,u,\eta,\xi) = \quad & \frac{1}{\sigma^2}\left(\left\langle A, \mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]\right\rangle - \frac{1}{2}\left\|u+Am_0\right\|^2 - \frac{1}{2}\mathrm{tr}(AS_aA^\top)\right) \\ & - \frac{\beta}{2}\mathbb{E}\left[K^2a^2\psi_1(a)\left\|\eta(Y)\right\|^2 + \left(\frac{\psi_1(a)}{K} - \psi_1(Ka)\right)\xi(Y)^2\right].\end{aligned}$$

The quadratic form $\mathcal{Q}_\beta$ splits into a reconstruction part,

$$\frac{1}{\sigma^2}\left(\left\langle A, \mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]\right\rangle - \frac{1}{2}\left\|u+Am_0\right\|^2 - \frac{1}{2}\mathrm{tr}(AS_aA^\top)\right),$$

and a KL part,

$$-\frac{\beta}{2}\mathbb{E}\left[K^2a^2\psi_1(a)\left\|\eta(Y)\right\|^2 + \left(\frac{\psi_1(a)}{K} - \psi_1(Ka)\right)\xi(Y)^2\right].$$

Those 2 parts compete against each other. The KL part is always non-positive (Lemma 3), while the reconstruction part can be positive or negative depending on the encoder perturbation η.

The rest of the proof consists in finding the critical value β_c such that when $\beta < \beta_c$, the reconstruction dominates and collapse is a saddle, i.e. $\mathcal{Q}_\beta > 0$, while when $\beta > \beta_c$, the KL dominates and collapse is a strict local minimum, i.e. $\mathcal{Q}_\beta < 0$.

Saddle direction. We want to find the largest value of β for which there still exists a perturbation (A,u,η,ξ) that lets the reconstruction gain dominate the KL. By Lemma 3, the problem can be reduced to determining when there exists a mean encoder perturbation $\eta(Y) \in L_2(P_Y)$ such that

$$\mathcal{Q}^\star_\beta(\eta) := \frac{K(Ka+1)}{2\sigma^2}\left\|\mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]\right\|_F^2 - \frac{\beta}{2}K^2a^2\psi_1(a)\mathbb{E}\left[\left\|\eta(Y)\right\|^2\right] > 0.$$

Ideally, we would like to choose $h(Y) := \left\langle e, \tilde{Y}\right\rangle$ with $e \in \mathbb{R}^D$ the unit eigenvector associated to the largest eigenvalue $\lambda_{\max}(\Sigma)$ and define the following perturbation of the encoder's posterior mean:

$$\eta^\star(Y) := h(Y)v$$

where $v \in \mathcal{T}$ is a unit vector. Since Y has a finite second moment by hypothesis, $h(Y) \in L^2(P_Y)$ and $\eta^\star(Y) \in L^2(P_Y)$ but this doesn't guarantee that $\eta^\star(Y)$ is an admissible perturbation. To ensure it, take bounded truncations

$$h_n(Y) := h(Y)\mathbf{1}_{\{|h(Y)|\le n\}}, \quad \eta_n(Y) := h_n(Y)v$$

which are in $L^\infty(P_Y) \subset L^2(P_Y)$ and hence admissible. Then η_n is bounded and tangent-valued, so for sufficiently small $|t|$, the path $m_t(Y) = m_0 + t\eta_n(Y)$ remains in the simplex almost surely. Moreover, $h_n \to h$ in $L^2(P_Y)$ and Y has a finite second moment, so Cauchy-Schwarz gives

$$\begin{aligned}\left\|\mathbb{E}\left[\tilde{Y}\eta_n(Y)^\top\right] - \mathbb{E}\left[\tilde{Y}\eta^\star(Y)^\top\right]\right\|_F &= \left\|\mathbb{E}\left[\tilde{Y}(h_n(Y)-h(Y))\right]\right\| \\ &\le \mathbb{E}\left[\left\|\tilde{Y}\right\|\left|h_n(Y)-h(Y)\right|\right] \\ &\le \left(\mathbb{E}\left\|\tilde{Y}\right\|^2\right)^{1/2}\left|\left|h_n(Y)-h(Y)\right|\right|_{L^2} \\ &\to 0 \ \text{ as } n \text{ tends to } \infty.\end{aligned}$$

Following the same reasoning, we also get

$$\mathbb{E}\left[\left\|\eta_n(Y)\right\|^2\right] = \mathbb{E}\left[h_n(Y)^2\right] \to \mathbb{E}\left[\left\|h(Y)\right\|^2\right].$$

Using the definition of $h(Y)$, we have

$$\mathbb{E}\left[\left\|h(Y)\right\|^2\right] = \mathbb{E}\left[\left\langle e, \tilde{Y}\right\rangle^2\right] = \lambda_{\max}(\Sigma) \quad \text{and} \quad \mathbb{E}\left[\tilde{Y}\eta^\star(Y)^\top\right] = \Sigma ev^\top = \lambda_{\max}(\Sigma)ev^\top.$$

For $\beta < \beta_c$, substituting these expressions into $\mathcal{Q}^\star_\beta(\eta^\star)$ gives

$$\begin{aligned}\mathcal{Q}^\star_\beta(\eta^\star) &= \frac{K(Ka+1)}{2\sigma^2}\lambda_{\max}(\Sigma)^2 - \frac{\beta}{2}K^2a^2\psi_1(a)\lambda_{\max}(\Sigma) \\ &= \frac{\lambda_{\max}(\Sigma)}{2}K^2a^2\psi_1(a)\left(\underbrace{\frac{\lambda_{\max}(\Sigma)}{\sigma^2}\frac{Ka+1}{Ka^2\psi_1(a)}}_{=:\,\beta_c} - \beta\right) > 0.\end{aligned}$$

This gives the critical value β_c defined in Eq. 3. To conclude on the saddle point, since $\mathcal{Q}^\star_\beta(\eta_n) \to \mathcal{Q}^\star_\beta(\eta^\star) > 0$ for $\beta < \beta_c$, we can choose n sufficiently large so that $\mathcal{Q}^\star_\beta(\eta_n) > 0$.

Fix such an n. By Lemma 2, the remainder is $o(t^2)$ for this fixed perturbation. Since $\mathcal{Q}^\star_\beta(\eta_n)$ is strictly positive and independent of t, for all sufficiently small nonzero t, we have

$$\begin{aligned}\mathcal{L}_\beta(W_t, b_t, q_t) - \mathcal{L}_\beta(W_0, b_0, p) &= -t^2\mathcal{Q}^\star_\beta(\eta_n) + o(t^2) \\ &< -\frac{t^2}{2}\mathcal{Q}^\star_\beta(\eta_n) < 0.\end{aligned}$$

This gives a loss-decreasing direction whenever $\beta < \beta_c$, so collapse is a saddle.

Condition for collapse to be a local minimum. Our goal is now to show that for $\beta > \beta_c$, we can find a neighborhood of (W_0, b_0, p) in which $\mathcal{L}_\beta$ is strictly larger than at (W_0, b_0, p) except along the flat directions and at (W_0, b_0, p), i.e.

$$\mathcal{L}_\beta(W_t, b_t, q_t) - \mathcal{L}_\beta(W_0, b_0, p) = -t^2\mathcal{Q}_\beta(A, u, \eta, \xi) + o(t^2) > 0.$$

To remove the flat directions from the decoder perturbations (A, u), we project A onto the simplex tangent space and adjust the bias, as given in Lemma 4 yielding $(AP_{\mathrm{T}}, u + Am_0)$. To simplify the notations, we denote the projected pair by (A, u) for the rest of the proof.

Let $r > 0$ and consider the set of admissible perturbations (A, u, η, ξ) such that $\|(A, u, \eta, \xi)\|_{\mathrm{par}} < r$ where $\|(A, u, \eta, \xi)\|_{\mathrm{par}}$ is the norm defined in Eq. 8. First, we show that $-\mathcal{Q}_\beta$ is coercive for $\beta > \beta_c$ with respect to the decoder norms and the L^2 norms of η and ξ. Since $L^\infty(P_Y) \subset L^2(P_Y)$, the lower bound found in L^2 also applies to the set of admissible perturbations in the neighborhood defined previously. Second, we control the remainder uniformly over this L^∞ neighborhood. Together, these establish strict local minimality after projection.

Let's prove the coercivity of $-\mathcal{Q}_\beta$. By Lemma 4, injecting the projected pair (A, u) into $\mathcal{Q}_\beta$ gives

$$\begin{aligned}-\mathcal{Q}_\beta(A, u, \eta, \xi) =& \frac{1}{2\sigma^2}\|u\|^2 + \frac{1}{2\sigma^2 K(Ka+1)}\|A\|_F^2 - \frac{1}{\sigma^2}\left\langle A, \mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]\right\rangle \\ &+ \frac{\beta}{2}K^2a^2\psi_1(a)\mathbb{E}\left[\|\eta(Y)\|^2\right] + \frac{\beta}{2}\left(\frac{\psi_1(a)}{K} - \psi_1(Ka)\right)\mathbb{E}\left[\|\xi(Y)\|^2\right].\end{aligned}$$

The inner product can be bounded by the Cauchy-Schwarz inequality,

$$\left\langle A, \mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]\right\rangle \le \|A\|_F\left\|\mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]\right\|_F \le \lambda_{\max}(\Sigma)^{1/2}\|A\|_F\|\eta\|_{L^2}.$$

Denote

$$x := \frac{\|A\|_F}{\sqrt{2\sigma^2 K(Ka+1)}}, \qquad y := \sqrt{\frac{\beta K^2a^2\psi_1(a)}{2}}\|\eta\|_{L^2}.$$

Using the definition of β_c, the bound becomes

$$-\mathcal{Q}_\beta(A, u, \eta, \xi) \ge \frac{1}{2\sigma^2}\|u\|^2 + \frac{\beta}{2}\left(\frac{\psi_1(a)}{K} - \psi_1(Ka)\right)\|\xi\|_{L^2}^2 + x^2 + y^2 - 2\sqrt{\frac{\beta_c}{\beta}}xy.$$

Since $2xy \le x^2 + y^2$ and $\sqrt{\frac{\beta_c}{\beta}} < 1$, we have

$$x^2 + y^2 - 2\sqrt{\frac{\beta_c}{\beta}}xy \ge \left(1 - \sqrt{\frac{\beta_c}{\beta}}\right)(x^2 + y^2).$$

The negative cross term therefore cannot compensate for the two positive terms. Together with the positive coefficients of the bias and concentration terms, this gives

$$-\mathcal{Q}_\beta(A,u,\eta,\xi) \geq c_\beta \left(\|A\|_F^2 + \|u\|^2 + \|\eta\|_{L^2}^2 + \|\xi\|_{L^2}^2 \right),$$

where

$$c_\beta := \min\Bigg\{ \frac{1}{2\sigma^2}, \frac{\beta}{2}\left(\frac{\psi_1(a)}{K} - \psi_1(Ka) \right), \\ \frac{1-\sqrt{\beta_c/\beta}}{2\sigma^2 K(Ka+1)}, \frac{(1-\sqrt{\beta_c/\beta})\beta K^2 a^2 \psi_1(a)}{2} \Bigg\} > 0.$$

The positivity of the concentration coefficient follows from Lemma 3. This proves coercivity.

Let's now consider the remainder term. Proof of Lemma 2 already shows that the remainder term is $o(t^2)$ for each fixed perturbation. To establish the local minimum, we need to bound it uniformly.

Recall $m_0 = \mathbf{1}/K$ and $s_0 = Ka$. We consider perturbations satisfying

$$\|\eta\|_{L^\infty} + \|\xi\|_{L^\infty} < r_0 := \min\left\{ \frac{1}{2K}, \frac{s_0}{2} \right\},$$

which remain admissible for $t \in [0,1]$. For almost every Y and $t \in [0,1]$, the pair $(m_0 + t\eta(Y), s_0 + t\xi(Y))$ then belongs to the compact set

$$\mathcal{K} = \left\{ m \in \Delta^{K-1} : \frac{1}{2K} \leq m_k \leq \frac{3}{2K} \text{ for all } k \right\} \times \left[\frac{s_0}{2}, \frac{3s_0}{2} \right].$$

Thus each coordinate of the Dirichlet parameter $(s_0 + t\xi(Y))(m_0 + t\eta(Y))$ is in $[a/4, 9a/4]$. Write $\gamma = (m_0 + \eta)(s_0 + \xi)$.

For the KL remainder, define

$$F(m,s) := \mathrm{KL}\left(\mathrm{Dir}(sm) \,\|\, \mathrm{Dir}(a\mathbf{1})\right), \qquad h(Y) := (\eta(Y), \xi(Y)).$$

The function F is smooth on a neighborhood of $\mathcal{K}$. With $g_Y(t) = F(m_0 + t\eta(Y), s_0 + t\xi(Y))$, the integral remainder from the proof of Lemma 2, evaluated at $t = 1$, gives

$$R_{\mathrm{KL}}(Y) = \int_0^1 (1-t)\Big[D^2F(m_0 + t\eta(Y), s_0 + t\xi(Y)) \\ - D^2F(m_0, s_0) \Big][h(Y), h(Y)]\, dt.$$

Since D^2F is continuous on the compact set $\mathcal{K}$, Heine's theorem gives uniform continuity. Thus, for every $\varepsilon > 0$, there exists $r_\varepsilon > 0$ such that $r \leq r_\varepsilon$ gives

$$|R_{\mathrm{KL}}(Y)| \leq \frac{\varepsilon}{2}\left(\|\eta(Y)\|^2 + |\xi(Y)|^2 \right).$$

Hence

$$\mathbb{E}[|R_{\mathrm{KL}}(Y)|] \leq \frac{\varepsilon}{2}\left(\|\eta\|_{L^2}^2 + \|\xi\|_{L^2}^2 \right).$$

For reconstruction, the expansion in the proof of Lemma 2 gives the remainder

$$R_{\mathrm{rec}} = \frac{1}{2\sigma^2}\mathbb{E}\left[2u^\top A\eta + \|A\eta\|^2 \right] + \frac{1}{2\sigma^2}\mathbb{E}\mathrm{tr}\left(A[\mathrm{Cov}_\gamma - S_a]A^\top \right).$$

First, since $\|\eta\|_{L^\infty} < r$, Cauchy-Schwarz gives

$$2\left|\mathbb{E}[u^\top A\eta]\right| \leq 2r\,\|u\|\,\|A\|_F \leq r\left(\|u\|^2 + \|A\|_F^2 \right), \\ \mathbb{E}\,\|A\eta\|^2 \leq r^2\,\|A\|_F^2.$$

For the covariance term, using $m_0 = \mathbf{1}/K$ and $Am_0 = 0$ in the Dirichlet covariance formula gives

$$\mathrm{tr}\left(A[\mathrm{Cov}_\gamma - S_a]A^\top \right) = \frac{\mathrm{tr}\left(A\,\mathrm{Diag}(\eta)A^\top \right) - \|A\eta\|^2}{s_0 + \xi + 1} - \frac{\xi\,\|A\|_F^2}{K(s_0+1)(s_0+\xi+1)}.$$

Since $s_0 + \xi > 0$, the denominators are at least above one. Using $|\eta_k| \le \|\eta\|$ and $\|\eta\| + |\xi| < r$, we obtain, almost surely,

$$\left|\mathrm{tr}\left(A[\mathrm{Cov}_\gamma - S_a]A^\top\right)\right| \le \left(\|\eta\| + \|\eta\|^2 + |\xi|\right)\|A\|_F^2 \le (r + r^2)\|A\|_F^2 .$$

Taking expectations and combining these bounds, we obtain, for $r \le r_0 \le \frac{1}{2K} \le 1$,

$$|R_{\mathrm{rec}}| \le \frac{r\|u\|^2 + 2r(1+r)\|A\|_F^2}{2\sigma^2} \le \frac{2r}{\sigma^2}\left(\|A\|_F^2 + \|u\|^2\right).$$

Choose $\varepsilon = c_\beta/\beta$ and

$$0 < r \le \min\left(r_0, r_\varepsilon, \frac{c_\beta \sigma^2}{4}\right).$$

Since the loss increment equals $-\mathcal{Q}_\beta + R_{\mathrm{rec}} + \beta\mathbb{E}[R_{\mathrm{KL}}]$, the remainder bounds and coercivity yield

$$\begin{aligned}\mathcal{L}_\beta(W_0 + A, b_0 + u, m_0 + \eta, s_0 + \xi) - \mathcal{L}_\beta(W_0, b_0, m_0, s_0)\\ \ge \frac{c_\beta}{2}\left(\|A\|_F^2 + \|u\|^2 + \|\eta\|_{L^2}^2 + \|\xi\|_{L^2}^2\right) > 0.\end{aligned}$$

Finally, since $L^\infty(P_Y) \subset L^2(P_Y)$ the lower bounds apply to all posterior perturbations in this uniform neighborhood. The loss therefore increases strictly, proving that collapse is a strict local minimum after projection. □

Lemma 2 (Taylor expansion of $\mathcal{L}_\beta$ around collapse and quadratic form $\mathcal{Q}_\beta$ associated to the Hessian)**.** *Assume the symmetric prior $p(z) = \mathrm{Dir}(a\mathbf{1})$ with $a > 0$. Let (W_0, b_0, q_0) be a collapsed point with*

$$W_0 = c\mathbf{1}^\top, \qquad b_0 = \bar{y} - c, \qquad m_0 = \frac{1}{K}\mathbf{1}, \qquad s_0 = Ka.$$

Let (A, u, η, ξ) be an admissible direction and consider the path

$$W_t = W_0 + tA, \qquad b_t = b_0 + tu, \qquad m_t(Y) = m_0 + t\eta(Y), \qquad s_t(Y) = Ka + t\xi(Y).$$

Define

$$\delta\gamma(Y) = Ka\,\eta(Y) + \frac{\xi(Y)}{K}\mathbf{1}.$$

Then

$$\mathcal{L}_\beta(W_t, b_t, q_t) = \mathcal{L}_\beta(W_0, b_0, p) - t^2\mathcal{Q}_\beta(A, u, \eta, \xi) + o(t^2),$$

where

$$\begin{aligned}\mathcal{Q}_\beta(A, u, \eta, \xi) = \frac{1}{\sigma^2}\left\langle A, \mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]\right\rangle - \frac{1}{2\sigma^2}\|u + Am_0\|^2 - \frac{1}{2\sigma^2}\mathrm{tr}(AS_aA^\top)\\ - \frac{\beta}{2}\mathbb{E}\left[\delta\gamma(Y)^\top H_a \delta\gamma(Y)\right].\end{aligned} \tag{9}$$

Equivalently, $-\mathcal{Q}_\beta$ is the second-order curvature of the minimized loss along this path.

Proof of Lemma 2. Let (W_0, b_0, m_0, s_0) be a collapsed point where the prior parameter vector α_0 is decomposed into mean and concentration such that $m_0 = \mathbf{1}/K$ and $s_0 = Ka$. Let us consider the following admissible perturbations:

$$\begin{aligned} W_t = W_0 + tA, \qquad & b_t = b_0 + tu,\\ m_t(Y) = m_0 + t\eta(Y), \quad & s_t(Y) = Ka + t\xi(Y).\end{aligned}$$

Note that $\gamma_t(Y) = s_t(Y)m_t(Y)$. Then

$$\gamma_t(Y) = a\mathbf{1} + t\delta\gamma(Y) + t^2\xi(Y)\eta(Y),$$

where

$$\delta\gamma(Y) = Ka\,\eta(Y) + \frac{\xi(Y)}{K}\mathbf{1}.$$

Define $g_Y(t) := \mathrm{KL}\left(\mathrm{Dir}(\gamma_t(Y))\|\mathrm{Dir}(a\mathbf{1})\right)$. By the chain rule, $g_Y''(0) = \delta\gamma(Y)^\top H_a \delta\gamma(Y)$ and $g_Y'(0) = g_Y(0) = 0$. To obtain the second-order expansion of the KL term for almost every Y, we use Taylor's formula at order one with integral remainder,

$$g_Y(t) = \int_0^t (t-\tau) g_Y''(\tau)\, d\tau.$$

By adding and subtracting $g_Y''(0)$ in the integral, and using $\int_0^t (t-\tau)\, d\tau = t^2/2$, we get

$$g_Y(t) = \frac{t^2}{2}\delta\gamma(Y)^\top H_a \delta\gamma(Y) + r_Y(t),$$

where

$$r_Y(t) := \int_0^t (t-\tau)\left[g_Y''(\tau) - g_Y''(0)\right]\, d\tau.$$

We now justify that $\mathbb{E}_Y[r_Y(t)] = o(t^2)$. Since the perturbations are admissible, $\eta, \xi \in L^\infty(P_Y)$. Let us choose

$$t_0 := \min\left\{\frac{1}{2K(1+\|\eta\|_{L^\infty})}, \frac{Ka}{2(1+\|\xi\|_{L^\infty})}\right\} > 0.$$

For all $|t| \le t_0$ and almost every Y, we have

$$\frac{1}{2K} \le m_{t,k}(Y) \le \frac{3}{2K} \quad \text{for all } k, \qquad \frac{Ka}{2} \le s_t(Y) \le \frac{3Ka}{2}.$$

Thus $\gamma_t(Y)$ belongs to the compact set $[a/4, 9a/4]^K \subset \mathbb{R}^K_{>0}$. Moreover, differentiating the path gives

$$\gamma_t'(Y) = \delta\gamma(Y) + 2t\xi(Y)\eta(Y), \qquad \gamma_t''(Y) = 2\xi(Y)\eta(Y).$$

These derivatives are bounded independently of Y and t on this interval. Since the KL is smooth, its first and second derivatives are bounded on this compact set. By the chain rule, there exists a finite constant C, independent of Y and t, such that $|g_Y''(t)| \le C$.

For $0 < |t| \le t_0$ and $u \in [0,1]$, we have $|ut| \le t_0$. By the triangle inequality, $|g_Y''(ut) - g_Y''(0)| \le |g_Y''(ut)| + |g_Y''(0)| \le 2C$. Since $1-u \ge 0$, the integral expression of the Taylor remainder gives

$$\begin{aligned}
\frac{|r_Y(t)|}{t^2} &= \frac{1}{t^2}\left|\int_0^t (t-\tau)\left[g_Y''(\tau) - g_Y''(0)\right]\, d\tau\right| \\
&\overset{\substack{\text{change of var.}\\ u=\tau/t}}{=} \left|\int_0^1 (1-u)\left[g_Y''(ut) - g_Y''(0)\right]\, du\right| \\
&\le \int_0^1 (1-u)\left|g_Y''(ut) - g_Y''(0)\right|\, du \\
&\le 2C\int_0^1 (1-u)\, du = C.
\end{aligned}$$

Since g_Y'' is continuous, $r_Y(t)/t^2 \to 0$ for almost every Y and dominated convergence gives

$$\frac{\mathbb{E}_Y[r_Y(t)]}{t^2} = \mathbb{E}_Y\left[\frac{r_Y(t)}{t^2}\right] \longrightarrow 0.$$

Therefore,

$$\mathbb{E}_Y\left[\mathrm{KL}\left(\mathrm{Dir}(\gamma_t(Y))\,\|\,\mathrm{Dir}(a\mathbf{1})\right)\right] = \frac{t^2}{2}\mathbb{E}_Y\left[\delta\gamma(Y)^\top H_a \delta\gamma(Y)\right] + o(t^2). \tag{10}$$

Reconstruction term. We now expand the reconstruction term. For a Dirichlet posterior with mean $m_t(Y)$ and covariance $\Sigma_t(Y) := \mathrm{Cov}_{q_t(z|Y)}(z)$, we can split it into its mean and covariance parts,

$$\mathbb{E}_{q_t(z|Y)}\left[\|Y - W_t z - b_t\|^2\right] = \|Y - W_t m_t(Y) - b_t\|^2 + \mathrm{tr}\left(W_t \Sigma_t(Y) W_t^\top\right). \tag{11}$$

First consider the mean part. Since $W_0 = c\mathbf{1}^\top$, $b_0 = \bar{y} - c$, and $\mathbf{1}^\top \eta(Y) = 0$ for some $c \in \mathbb{R}^D$, we have

$$W_0 m_0 + b_0 = \bar{y}, \qquad W_0 \eta(Y) = c\mathbf{1}^\top \eta(Y) = 0.$$

Define also $v := u + Am_0$ and $\tilde{Y} = Y - \bar{y}$. Thus, by using $\mathbb{E}_Y[\tilde{Y}] = 0$, we obtain

$$\mathbb{E}_Y\left[\|Y - W_t m_t(Y) - b_t\|^2\right] = \mathbb{E}_Y\left[\left\|\widetilde{Y}\right\|^2 + t^2\|v\|^2 - 2t^2\widetilde{Y}^\top A\eta(Y) + R_t(Y)\right],$$

where

$$R_t(Y) := 2t^3 v^\top A\eta(Y) + t^4\|A\eta(Y)\|^2.$$

Since $\eta \in L^\infty(P_Y) \subset L^2(P_Y)$ and $Y \in L^2(P_Y)$, all terms obtained by expanding the square are integrable. For the fixed perturbation, the cubic and quartic terms have finite coefficients independent of t, so their contribution is $o(t^2)$. Therefore,

$$\begin{aligned}\mathbb{E}_Y\left[\|Y - W_t m_t(Y) - b_t\|^2\right] = \mathbb{E}_Y\left\|\widetilde{Y}\right\|^2 + t^2\|u + Am_0\|^2 \\ - 2t^2\left\langle A, \mathbb{E}_Y\left[\widetilde{Y}\eta(Y)^\top\right]\right\rangle + o(t^2),\end{aligned} \tag{12}$$

where $\langle A, B\rangle = \mathrm{tr}(A^\top B)$ denotes the Frobenius inner product.

Next consider the posterior-covariance part. For every Dirichlet distribution on the simplex, the covariance matrix annihilates $\mathbf{1}$:

$$\Sigma_t(Y)\mathbf{1} = 0, \qquad \mathbf{1}^\top\Sigma_t(Y) = 0.$$

Since $W_0 = c\mathbf{1}^\top$, this implies

$$W_0\Sigma_t(Y) = 0, \qquad \Sigma_t(Y)W_0^\top = 0$$

for every t. Therefore,

$$\begin{aligned}\mathrm{tr}\left(W_t\Sigma_t(Y)W_t^\top\right) &= \mathrm{tr}\left((W_0 + tA)\Sigma_t(Y)(W_0 + tA)^\top\right) \\ &= t^2\mathrm{tr}\left(A\Sigma_t(Y)A^\top\right).\end{aligned}$$

Moreover, $\Sigma_t(Y) = S_a + o(1)$, where

$$S_a = \mathrm{Cov}_{z\sim\mathrm{Dir}(a\mathbf{1})}(z) = \frac{1}{K(Ka+1)}P_{\mathrm{T}}.$$

Adding and subtracting S_a gives

$$\begin{aligned}t^2\mathrm{tr}\left(A\Sigma_t(Y)A^\top\right) &= t^2\mathrm{tr}(AS_aA^\top) \\ &\quad + t^2\mathrm{tr}\left(A[\Sigma_t(Y) - S_a]A^\top\right).\end{aligned}$$

The first term is the quadratic contribution, while the second contributes to the remainder. Hence

$$\mathbb{E}\left[\mathrm{tr}\left(W_t\Sigma_t(Y)W_t^\top\right)\right] = t^2\mathrm{tr}(AS_aA^\top) + o(t^2). \tag{13}$$

Combining Eqs. (11), (12), and (13), the reconstruction contribution to the negative ELBO is

$$\begin{aligned}&\frac{1}{2\sigma^2}\mathbb{E}\left[\mathbb{E}_{q_t(z|Y)}\left[\|Y - W_t z - b_t\|^2\right]\right] \\ &= \frac{1}{2\sigma^2}\mathbb{E}\left\|\widetilde{Y}\right\|^2 \\ &\quad + t^2\left[\frac{1}{2\sigma^2}\|u + Am_0\|^2 - \frac{1}{\sigma^2}\left\langle A, \mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]\right\rangle + \frac{1}{2\sigma^2}\mathrm{tr}(AS_aA^\top)\right] + o(t^2).\end{aligned}$$

Adding the KL expansion in Eq. (10), we obtain

$$\begin{aligned}\mathcal{L}_\beta(W_t, b_t, q_t) = \mathcal{L}_\beta(W_0, b_0, p) \\ + t^2\Bigg[\frac{1}{2\sigma^2}\|u + Am_0\|^2 - \frac{1}{\sigma^2}\left\langle A, \mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]\right\rangle + \frac{1}{2\sigma^2}\mathrm{tr}(AS_aA^\top) \\ + \frac{\beta}{2}\mathbb{E}\left[\delta\gamma(Y)^\top H_a\delta\gamma(Y)\right]\Bigg] + o(t^2).\end{aligned}$$

By definition, the escape score is the negative of the loss curvature. Thus

$$\mathcal{L}_\beta(W_t, b_t, q_t) = \mathcal{L}_\beta(W_0, b_0, p) - t^2 \mathcal{Q}_\beta(A, u, \eta, \xi) + o(t^2),$$

with

$$\mathcal{Q}_\beta(A, u, \eta, \xi) = \frac{1}{\sigma^2} \left\langle A, \mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]\right\rangle - \frac{1}{2\sigma^2}\left\|u + Am_0\right\|^2 - \frac{1}{2\sigma^2}\mathrm{tr}(AS_aA^\top) - \frac{\beta}{2}\mathbb{E}\left[\delta\gamma(Y)^\top H_a \delta\gamma(Y)\right].$$

This is Eq. (9). □

Lemma 3 (Decoder-profiled escape score)**.** *Under the assumptions of Lemma 2, with the symmetric prior $p(z) = \mathrm{Dir}(a\mathbf{1})$, $a > 0$, let $\beta \geq 0$ and fix an admissible mean perturbation η. Then, maximizing over admissible perturbations,*

$$\begin{aligned}\max_{A,u,\xi} \mathcal{Q}_\beta(A, u, \eta, \xi) &= \max_{A,u} \mathcal{Q}_\beta(A, u, \eta, 0) \\ &= \frac{K(Ka+1)}{2\sigma^2}\left\|\mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]\right\|_{\mathrm{F}}^2 - \frac{\beta K^2 a^2 \psi_1(a)}{2}\mathbb{E}\left[\|\eta(Y)\|^2\right] =: \mathcal{Q}^\star_\beta(\eta).\end{aligned} \tag{14}$$

For a fixed η, the maximum is attained by any admissible perturbation $(A^*, u^*, \xi^*(Y))$ satisfying

$$A^\star P_{\mathrm{T}} = K(Ka+1)\mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right], \qquad u^\star = -A^\star m_0, \qquad \xi^*(Y) = 0.$$

For $\beta > 0$, these conditions characterize all maximizers, with $\xi^*(Y) = 0$ almost surely.

Proof of Lemma 3. Under the symmetric prior $p(z) = \mathrm{Dir}(a\mathbf{1})$, we have $m_0 = \mathbf{1}/K$ and $S_a = P_{\mathrm{T}}/[K(Ka+1)]$. Consider a perturbation $(A, u, \eta(Y), \xi(Y))$, where the posterior perturbation $(\eta(Y), \xi(Y))$ is admissible. Substituting S_a into Eq. (9) gives

$$\mathcal{Q}_\beta(A, u, \eta, \xi) = \frac{1}{\sigma^2}\left\langle A, \mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]\right\rangle - \frac{1}{2\sigma^2}\left\|u + Am_0\right\|^2 - \frac{1}{2\sigma^2 K(Ka+1)}\left\|AP_{\mathrm{T}}\right\|_F^2 - \frac{\beta}{2}\mathbb{E}\left[\delta\gamma(Y)^\top H_a \delta\gamma(Y)\right],$$

where $\delta\gamma(Y) = Ka\,\eta(Y) + \xi(Y)\mathbf{1}/K$.

Since $\mathbf{1}^\top\eta(Y) = 0$ almost surely, we have $\mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]\mathbf{1} = 0$ and $\mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]P_{\mathrm{T}} = \mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]$. The KL term in Eq. (9) is independent of (A, u), so we first maximize the decoder contribution,

$$\mathcal{Q}_{\beta,\mathrm{dec}}(A, u) = \frac{1}{\sigma^2}\left\langle AP_{\mathrm{T}}, \mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]\right\rangle - \frac{1}{2\sigma^2}\left\|u + Am_0\right\|^2 - \frac{1}{2\sigma^2 K(Ka+1)}\left\|AP_{\mathrm{T}}\right\|_F^2.$$

Completing the square, we get

$$\mathcal{Q}_{\beta,\mathrm{dec}}(A, u) = \frac{K(Ka+1)}{2\sigma^2}\left\|\mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]\right\|_F^2 - \frac{\left\|AP_{\mathrm{T}} - K(Ka+1)\mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]\right\|_F^2}{2\sigma^2 K(Ka+1)} - \frac{1}{2\sigma^2}\left\|u + Am_0\right\|^2.$$

The first term does not depend on (A, u) and the two other terms are non-positive. The maximum is therefore attained when both squares vanish, which gives

$$u^\star = -A^\star m_0, \qquad A^\star P_{\mathrm{T}} = K(Ka+1)\mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]. \tag{15}$$

Since $\mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]P_{\mathrm{T}} = \mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]$, all solutions can be written as

$$A^\star = K(Ka+1)\mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right] + G(I_K - P_{\mathrm{T}}), \qquad G \in \mathbb{R}^{D\times K}. \tag{16}$$

where G can be any matrix in $\mathbb{R}^{D\times K}$, with its effect compensated by $u^\star$ (see Lemma 4).

At the maximum $(A^\star, u^\star)$, only the first term remains. Adding back the KL contribution gives

$$\mathcal{Q}^\star_\beta(\eta,\xi) = \frac{K(Ka+1)}{2\sigma^2}\left\|\mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]\right\|_{\mathrm{F}}^2 - \frac{\beta}{2}\mathbb{E}\left[\delta\gamma(Y)^\top H_a \delta\gamma(Y)\right]. \tag{17}$$

For the KL term, Lemma 1(2) gives the symmetric Hessian. Together with the definition of $\delta\gamma(Y)$, we have

$$\delta\gamma(Y) = Ka\,\eta(Y) + \frac{\xi(Y)}{K}\mathbf{1}, \quad \text{and} \quad H_a = \psi_1(a)I_K - \psi_1(Ka)\mathbf{1}\mathbf{1}^\top. \tag{18}$$

The mixed term vanishes because $\mathbf{1}^\top\eta(Y) = 0$, and $\mathbf{1}^\top\mathbf{1} = K$, i.e.

$$\eta(Y)^\top H_a \mathbf{1} = \left(\psi_1(a) - K\psi_1(Ka)\right)\eta(Y)^\top\mathbf{1} = 0. \tag{19}$$

Moreover,

$$\eta(Y)^\top H_a \eta(Y) = \psi_1(a)\left\|\eta(Y)\right\|^2, \tag{20}$$

and

$$\frac{1}{K^2}\mathbf{1}^\top H_a\mathbf{1} = \frac{1}{K^2}\left(K\psi_1(a) - K^2\psi_1(Ka)\right) = \frac{\psi_1(a)}{K} - \psi_1(Ka). \tag{21}$$

Combining Eqs. (18), (19), (20), and (21) gives

$$\delta\gamma(Y)^\top H_a\delta\gamma(Y) = K^2a^2\psi_1(a)\left\|\eta(Y)\right\|^2 + \left(\frac{\psi_1(a)}{K} - \psi_1(Ka)\right)\xi(Y)^2. \tag{22}$$

By Lemma 5, for every $x > 0$,

$$-\frac{\psi_2(x)}{\psi_1(x)} > \frac{2x+2}{x(2x+1)} > \frac{1}{x}.$$

Multiplying by $x\psi_1(x) > 0$ gives $-x\psi_2(x) > \psi_1(x)$, so by the product rule,

$$\frac{d}{dx}\left[x\psi_1(x)\right] = \psi_1(x) + x\psi_2(x) < 0.$$

Thus $x\psi_1(x)$ is strictly decreasing. Since $K \geq 2$, we have $a < Ka$, which gives

$$a\psi_1(a) > Ka\psi_1(Ka), \qquad \text{hence} \qquad \frac{\psi_1(a)}{K} - \psi_1(Ka) > 0.$$

Substituting Eq. (22) into the profiled score in Eq. (17) yields

$$\mathcal{Q}^\star_\beta(\eta,\xi) = \frac{K(Ka+1)}{2\sigma^2}\left\|\mathbb{E}\left[\tilde{Y}\eta(Y)^\top\right]\right\|_{\mathrm{F}}^2 \tag{23}$$

$$- \frac{\beta}{2}\mathbb{E}\left[K^2a^2\psi_1(a)\left\|\eta(Y)\right\|^2 + \left(\frac{\psi_1(a)}{K} - \psi_1(Ka)\right)\xi(Y)^2\right]. \tag{24}$$

The concentration coefficient is positive, so for $\beta \geq 0$ the admissible choice $\xi = 0$ maximizes the profiled score for fixed η. Substituting $\xi = 0$ in Eq. (23) gives the stated expression for $\mathcal{Q}^\star_\beta(\eta)$ in Eq. (14). □

Lemma 4 (Decoder invariance under compensating changes in W and b)**.** *For every $d \in \mathbb{R}^D$, the transformation*

$$(W, b) \longmapsto (W + d\mathbf{1}^\top, b - d)$$

leaves the decoder map $z \mapsto Wz + b$ unchanged on the simplex, and hence leaves $\mathcal{L}_\beta(W, b, q)$ unchanged for every admissible posterior q. In particular, changes in c within the collapsed manifold are compensated by changes in b.

Proof of Lemma 4. For every $z \in \Delta^{K-1}$, one has $\mathbf{1}^\top z = 1$. Therefore

$$(W + d\mathbf{1}^\top)z + b - d = Wz + b + d(\mathbf{1}^\top z) - d = Wz + b.$$

Thus the decoder map is unchanged on the simplex. The reconstruction term is therefore unchanged, and the KL term does not depend on (W, b). Moreover,

$$(W + d\mathbf{1}^\top)P_{\mathrm{T}} = WP_{\mathrm{T}} + d\mathbf{1}^\top P_{\mathrm{T}} = WP_{\mathrm{T}},$$

since $\mathbf{1}^\top P_{\mathrm{T}} = 0$. Hence the tangent decoder component WP_{T} is unchanged by this transformation. □

A.6 Proof of Proposition 1

Lemma 5 (Trigamma ratio inequality)**.** *Let ψ_1 denote the trigamma function and let $\psi_2 = \psi_1'$ be the polygamma function of order two. Then, for every $a > 0$,*

$$-\frac{\psi_2(a)}{\psi_1(a)} > \frac{2a+2}{a(2a+1)}.$$

Proof. For $a > 0$, we use the integral representations

$$\psi_1(a) = \int_0^\infty \frac{te^{-at}}{1-e^{-t}}\,dt, \qquad -\psi_2(a) = \int_0^\infty \frac{t^2e^{-at}}{1-e^{-t}}\,dt.$$

Set $h(t) := \frac{t}{1-e^{-t}}$. Then h extends smoothly to $t = 0$, with $h(0) = 1$, and

$$\psi_1(a) = \int_0^\infty h(t)e^{-at}\,dt, \qquad -\psi_2(a) = \int_0^\infty th(t)e^{-at}\,dt.$$

For brevity, write

$$H(a) := \int_0^\infty h(t)e^{-at}\,dt = \psi_1(a), \qquad T(a) := \int_0^\infty th(t)e^{-at}\,dt = -\psi_2(a).$$

All integrations by parts below are justified because the exponential factor e^{-at} dominates the polynomial growth of h, h', and h'' at infinity, while the boundary terms at 0 are finite.

First,

$$\int_0^\infty h'(t)e^{-at}\,dt = aH(a) - 1.$$

Second,

$$\int_0^\infty th'(t)e^{-at}\,dt = aT(a) - H(a).$$

Third,

$$\int_0^\infty th''(t)e^{-at}\,dt = a^2T(a) - 2aH(a) + 1.$$

Therefore

$$\begin{aligned}
&\int_0^\infty \Big(2th''(t) + (t+2)h'(t) - h(t)\Big)e^{-at}\,dt \\
&= 2\big(a^2T(a) - 2aH(a) + 1\big) + \big(aT(a) - H(a)\big) + 2\big(aH(a) - 1\big) - H(a) \\
&= a(2a+1)T(a) - 2(a+1)H(a).
\end{aligned}$$

Thus it remains to prove that

$$K(t) := 2th''(t) + (t+2)h'(t) - h(t)$$

is strictly positive for every $t > 0$. A direct calculation gives

$$K(t) = \frac{e^t}{(e^t-1)^3}\Big(t^2e^t + 3t^2 - 6te^t + 6t + 2e^{2t} - 4e^t + 2\Big).$$

The prefactor is strictly positive for $t > 0$. It remains to check the sign of the term in parentheses. Define

$$F(t) := t^2e^t + 3t^2 - 6te^t + 6t + 2e^{2t} - 4e^t + 2.$$

We have $F(0) = 0$, $F'(0) = 0$, and $F''(0) = 0$. Moreover, $F'''(t) = e^t\big(t^2 + 16(e^t - 1)\big) > 0$ for $t > 0$. Therefore, $F'''(t) > 0$ for $t > 0$, and hence $F''(t), F'(t), F(t) > 0$ for $t > 0$. Consequently,

$$K(t) > 0 \qquad \text{for every } t > 0.$$

It follows that

$$a(2a+1)T(a) - 2(a+1)H(a) = \int_0^\infty K(t)e^{-at}\,dt > 0.$$

Substituting $T(a) = -\psi_2(a)$ and $H(a) = \psi_1(a)$, we obtain

$$a(2a+1)(-\psi_2(a)) - 2(a+1)\psi_1(a) > 0.$$

Since $a(2a+1)\psi_1(a) > 0$, division gives

$$-\frac{\psi_2(a)}{\psi_1(a)} > \frac{2a+2}{a(2a+1)}.$$

This proves the claim. □

Proposition 1 (Prior-shift calibration law)**.** *Let $K \geq 2$, assume the symmetric prior $p(z) = \mathrm{Dir}(a\mathbf{1})$ and $\lambda_{\max}(\Sigma) > 0$. For fixed K, $\lambda_{\max}(\Sigma)$, and σ^2, define*

$$g_K(a) = \frac{Ka+1}{Ka^2\psi_1(a)}$$

such that we can rewrite the collapse threshold as $\beta_{\mathrm{c}}(a) = \frac{\lambda_{\max}(\Sigma)}{\sigma^2} g_K(a)$. Then g_K is smooth and strictly increasing on $(0, \infty)$, with

$$\lim_{a\to 0} g_K(a) = \frac{1}{K}, \qquad \lim_{a\to\infty} g_K(a) = 1.$$

Consequently,

$$\beta_{\mathrm{c}}(a) \in \left(\frac{\lambda_{\max}(\Sigma)}{K\sigma^2}, \frac{\lambda_{\max}(\Sigma)}{\sigma^2} \right),$$

and every threshold value in this interval is attained by a unique $a > 0$. Equivalently, for a fixed β and target collapse ratio $(\beta/\beta_{\mathrm{c}})^\star = \frac{\beta}{\beta_{\mathrm{c}}(a)}$, the required concentration is determined by the scalar equation

$$g_K(a) = \frac{\sigma^2\beta}{(\beta/\beta_{\mathrm{c}})^\star \, \lambda_{\max}(\Sigma)},$$

whenever the right-hand side lies in $(1/K, 1)$.

Proof of Proposition 1. g_K is smooth. The function $g_K(a)$ is a ratio of smooth functions, and the denominator $Ka^2\psi_1(a)$ is strictly positive for all $a > 0$. Therefore, $g_K(a)$ is smooth on $(0, \infty)$.

Limiting behavior of g_K. We next compute the limiting behavior of g_K. In both limits, we will use the series expansions, i.e.

$$\psi_1(a) = \sum_{n=0}^{\infty} \frac{1}{(a+n)^2}.$$

As $a \to 0$, the leading term of the series is $1/a^2$, and the sum of the remaining terms is bounded by $\pi^2/6$. Hence,

$$\psi_1(a) = \frac{1}{a^2} + O(1).$$

Therefore, $a^2\psi_1(a) \to 1$ and $Ka + 1 \to 1$, and hence $\lim_{a\to 0} g_K(a) = \frac{1}{K}$.

To study the case $a \to \infty$, we use the equality

$$\frac{1}{x^2} = \frac{1}{x} - \frac{1}{x+1} + \frac{1}{2x^2} - \frac{1}{2(x+1)^2} + \frac{1}{2x^2(x+1)^2}.$$

Hence, we can write

$$\psi_1(a) = \frac{1}{a} + \frac{1}{2a^2} + \frac{1}{2}\sum_{n=0}^{\infty} \frac{1}{(a+n)^2(a+n+1)^2}$$

where the last term is $O(a^{-3})$ as $a \to \infty$, by comparison with $\sum_{n=0}^{\infty}(a+n)^{-4} \leq a^{-4} + 1/(3a^3)$. Thus

$$Ka^2\psi_1(a) = Ka + \frac{K}{2} + O(a^{-1}),$$

and consequently

$$\lim_{a\to\infty} g_K(a) = \lim_{a\to\infty} \frac{Ka+1}{Ka+\frac{K}{2}+O(a^{-1})} = 1.$$

Strict monotonicity. It remains to prove strict monotonicity. Differentiating the logarithm gives

$$\frac{d}{da}\log g_K(a) = \frac{K}{Ka+1} - \frac{2}{a} - \frac{\psi_2(a)}{\psi_1(a)},$$

where $\psi_2 = \psi_1'$ is the polygamma function of order two. Thus $g_K'(a) > 0$ is equivalent to

$$-\frac{\psi_2(a)}{\psi_1(a)} > \frac{2}{a} - \frac{K}{Ka+1} = \frac{Ka+2}{a(Ka+1)}.$$

For $K \geq 2$, the map

$$K \mapsto \frac{Ka+2}{a(Ka+1)}$$

is decreasing, because

$$\frac{\partial}{\partial K}\left(\frac{Ka+2}{a(Ka+1)}\right) = -\frac{1}{(Ka+1)^2} < 0.$$

Hence

$$\frac{Ka+2}{a(Ka+1)} \leq \frac{2a+2}{a(2a+1)}.$$

Using the trigamma ratio inequality from Lemma 5,

$$-\frac{\psi_2(a)}{\psi_1(a)} > \frac{2a+2}{a(2a+1)}, \qquad a > 0,$$

we obtain

$$-\frac{\psi_2(a)}{\psi_1(a)} > \frac{Ka+2}{a(Ka+1)} = \frac{2}{a} - \frac{K}{Ka+1}.$$

Therefore $\frac{d}{da}\log g_K(a) > 0$, and since $g_K(a) > 0$, this implies $g_K'(a) > 0$ for all $a > 0$.

Together with the endpoint limits, strict monotonicity implies that

$$g_K : (0,\infty) \to \left(\frac{1}{K}, 1\right)$$

is a bijection. Multiplication by $\lambda_{\max}(\Sigma)/\sigma^2$ gives $\beta_{\mathrm{c}}(a) \in \left(\frac{\lambda_{\max}(\Sigma)}{K\sigma^2}, \frac{\lambda_{\max}(\Sigma)}{\sigma^2}\right)$, and every value in this interval is attained by a unique $a > 0$.

Fixed ratio and prior calibration. Finally, imposing a target normalized collapse ratio $(\beta/\beta_{\mathrm{c}})^\star = \frac{\beta}{\beta_{\mathrm{c}}(a)}$ and substituting the expression of $\beta_{\mathrm{c}}(a)$ gives

$$(\beta/\beta_{\mathrm{c}})^\star = \frac{\beta\sigma^2}{\lambda_{\max}(\Sigma) g_K(a)}.$$

Rearranging yields the inverse calibration equation

$$g_K(a) = \frac{\sigma^2\beta}{(\beta/\beta_{\mathrm{c}})^\star \lambda_{\max}(\Sigma)}.$$

When the right-hand side lies in the attainable range $(1/K, 1)$, strict monotonicity gives a unique prior concentration a corresponding to the target collapse ratio. □

B Collapse metrics

We evaluate the following diagnostics on the training samples. For $q_\phi(z \mid y_i) = \mathrm{Dir}(\gamma_i)$ and $W \in \mathbb{R}^{D\times K}$, let

$$m_i = \frac{\gamma_i}{\sum_{k=1}^{K} \gamma_{ik}}, \qquad \bar{m} = \frac{1}{N}\sum_{i=1}^{N} m_i.$$

Average posterior KL.

$$\overline{\mathrm{KL}}_\phi = \frac{1}{N}\sum_{i=1}^{N} \mathrm{KL}(\mathrm{Dir}(\gamma_i) \,\|\, \mathrm{Dir}(a\mathbf{1}_K)) \,.$$

It measures deviation from the prior and is zero exactly when every evaluated posterior equals the prior.

Tangent norm of the mixing matrix.

$$\|W P_T\|_F.$$

It measures the dependence of the reconstruction mean on latent composition and vanishes when all columns of W coincide.

Variation of posterior means. The latent-usage diagnostic U_m measures variation of posterior mean compositions across samples, with zero indicating identical means but not necessarily identical posteriors. For numerical implementation, we compute the same quantity in explicit tangent coordinates. Let $B \in \mathbb{R}^{K\times(K-1)}$ be a fixed orthonormal basis of the simplex tangent space, so that $B^\top B = I_{K-1}$ and $BB^\top = P_T$. Since $\mathbf{1}_K^\top(m_i - \bar{m}) = 0$, we have

$$U_m = \frac{1}{N}\sum_{i=1}^{N} \|m_i - \bar{m}\|^2 = \frac{1}{N}\sum_{i=1}^{N} \|B^\top(m_i - \bar{m})\|^2.$$

The implementation uses the second expression to compute U_m from the coordinates of $m_i - \bar{m}$ in this orthonormal basis.

C Theorem validation experiments

This appendix reports controlled numerical checks of the theoretical results of Section A. We test:

- The local quadratic expansion of the β-ELBO around the collapsed manifold (Lemma 2), and the sign change of the curvature at $\beta = \beta_c$ predicted by Theorem 2;
- The composition versus concentration mechanism of escape from collapse when the prior is symmetric predicted by Lemma 3;
- The invariance to any variation in the normal space of the simplex tangent space proven by Lemma 4;
- Proposition 1 on the effect of prior shift on the collapse threshold.

C.1 Synthetic dataset

Unless stated otherwise, for all the experiments in this section, we use a synthetic dataset with the parameters in Table 4.

Table 4: Parameters used to generate the synthetic dataset.

Parameter	Symbol	Value
Number of observations	N	8000
Observation dimension	D	150
Latent dimension	K	20
Dirichlet parameter per component	a	1
Mixing matrix scale	τ	4
Noise multiplier	σ_{factor}	1
Random seed for dataset generation	seed	0

We generate N observations from the Dirichlet linear Gaussian model

$$y_i = Wz_i + b + \varepsilon_i, \qquad i = 1, \dots, N,$$
$$z_i \overset{\text{i.i.d.}}{\sim} \text{Dir}(a\mathbf{1}_K), \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I_D).$$

Here $W \in \mathbb{R}^{D\times K}$ is a fixed mixing matrix, $b \in \mathbb{R}^D$ is the offset, and the latent compositions and noise vectors are independent. To simplify the analysis, we center the distribution of the observations so that the theoretical mean is zero: $\mathbb{E}[Y] = 0$. Writing $c := \frac{1}{K} W\mathbf{1}_K \in \mathbb{R}^D$ for the mean of the columns of W, we choose $b = 0$ and impose $c = 0$, equivalently projecting the mixing matrix $W = WP_T$ with the projection operator $P_T = I_K - \frac{1}{K}\mathbf{1}\mathbf{1}^\top$ onto the simplex tangent space. Indeed, since it can always be decomposed into its tangent and normal components, we can write

$$Y = WZ + b + \varepsilon = (WP_T + c\mathbf{1}^\top)Z + b + \varepsilon = W_T Z + (b + c) + \varepsilon.$$

Hence, because $\mathbb{E}[Z] = \frac{1}{K}\mathbf{1}$ and $b = 0$, $c = 0$, we have $\mathbb{E}[Y] = 0$.

To build W, we draw a random Gaussian matrix $G \in \mathbb{R}^{D\times K}$ with independent entries $G_{dk} \sim \mathcal{N}(0, 1)$ and form

$$W = W_T := \tau G P_T.$$

Then, conditional on this fixed matrix, we generate $y_i = Wz_i + \varepsilon_i$. Note that in practice, finite-sample fluctuations might induce a non zero empirical mean $\bar{y} = \frac{1}{N}\sum_i y_i$. Therefore, at collapse, the decoder offset will be set to $b_0 = \bar{y}$ to match the empirical mean of the data.

To control the Signal-to-Noise Ratio (SNR), define $\sigma^2 := \sigma^2_{\text{factor}} s^2_{\text{sig}}$, where

$$s^2_{\text{sig}} = \frac{1}{D}\,\text{tr}(W_T S_a W_T^\top) = \frac{1}{D}\,\text{tr}(\Sigma_{sig})$$

is the average signal power per dimension, and σ_{factor} is a multiplicative factor that controls the overall noise level. Using $\text{tr}(\Sigma_{sig}) \le (K-1)\lambda_{\max}(\Sigma_{sig})$ we get the following lower bound on the

SNR in the maximal signal direction:

$$\frac{\lambda_{\max}(\Sigma_{sig})}{\sigma^2} \geq \frac{D}{K-1}\frac{1}{\sigma_{\text{factor}}^2}.$$

Setting $\sigma_{\text{factor}} = 1$ makes the expected total signal and noise powers equal (an overall SNR of 0 dB) while guaranteeing an SNR of at least $D/(K-1)$ in the maximal signal direction (≈ 9 dB for $D = 150$ and $K = 20$).

Finally, the dataset is generated with random seed 0 in double precision to limit numerical errors.

C.2 Losses

For the three numerical checks below, we use the empirical negative β-ELBO with either an analytic or a Monte Carlo reconstruction term. The Gaussian additive constant is omitted and σ^2 is fixed. In both cases, the KL is evaluated in closed form and no additional regularization of W is used.

Write $p(z) = \text{Dir}(a\mathbf{1}_K)$ and $q_i(z) = \text{Dir}(\gamma_i)$, where γ_i is either a free parameter for sample i or the encoder output $\gamma_\phi(y_i)$. Let $s_i = \mathbf{1}_K^\top \gamma_i$, $m_i = \gamma_i / s_i$, and $\Sigma_i = \text{Cov}_{q_i}(z) = (s_i+1)^{-1}(\text{Diag}(m_i) - m_i m_i^\top)$.

Exact loss. Evaluating the reconstruction expectation analytically gives

$$\mathcal{L}_\beta^{\text{exact}} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\|y_i - Wm_i - b\|_2^2 + \text{tr}(W\Sigma_i W^\top)}{2\sigma^2} + \beta\,\text{KL}(q_i\|p)\right]. \tag{25}$$

This loss is used for the local curvature check (Sec. C.3) and the composition versus concentration check (Sec. C.4). In the latter, it is minimized over (W, b) for each fixed set of Dirichlet parameters. In the invariance check (Sec. C.5), it is the training loss for the *oracle exact* and *amortized exact* regimes and the evaluation loss for all three regimes.

Monte Carlo loss. Replacing the reconstruction expectation by a sample average gives

$$\widehat{\mathcal{L}}_\beta^{\text{MC}} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2\sigma^2 S}\sum_{r=1}^{S}\|y_i - Wz_i^{(r)} - b\|_2^2 + \beta\,\text{KL}(q_i\|p)\right], \tag{26}$$

where $z_i^{(r)} \overset{\text{i.i.d.}}{\sim} q_i$ and $S = 4$. This estimator is used to train the *amortized practical* regime in Sec. C.5. We evaluate the model with the exact loss in Eq. (25) to compare regimes without the Monte Carlo estimation noise.

Table 5: Losses and variational parameterizations in the theorem validation experiments. Exact refers to Eq. (25), with analytic reconstruction expectations, and MC to Eq. (26).

Experiment	Loss	Inference setup
Local curvature (Sec. C.3)	Exact	Built posterior perturbation path
Composition vs. concentration (Sec. C.4)	Exact, minimized over (W, b)	Built posterior perturbation path
Invariance (Sec. C.5)	Exact for training and evaluation	*Oracle exact*, 1200 epochs
	Exact for training and evaluation	*Amortized exact*, 3000 epochs
	MC for training, exact for evaluation	*Amortized practical*, 3000 epochs
Prior shift (Sec. C.6)	Exact for training and evaluation	*Amortized exact*, 2000 epochs

C.3 Numerical check of Lemma 2 and Theorem 2: local curvature validation

Lemma 2 predicts that the β-ELBO is locally quadratic around the collapsed manifold,

$$\mathcal{L}_\beta(W_t, b_t, q_t) - \mathcal{L}_\beta(W_0, b_0, q_0) = \mathcal{L}(t) - \mathcal{L}(0) = -t^2 Q_\beta(A, u, \eta, \xi) + o(t^2),$$

and Theorem 2 that the curvature of the β-ELBO changes sign at $\beta = \beta_c$. In this section, we check the curvature and its sign change numerically along a perturbation path from a collapsed point.

Experimental protocol. To fix the sign of the second-order Taylor expansion term, we propose to report the loss curvature (up to a factor of 2) $\kappa_\beta(A, u, \eta, \xi) = -Q_\beta(A, u, \eta, \xi)$. Thus, along a perturbation path,

$$\mathcal{L}_\beta(t) - \mathcal{L}_\beta(0) = t^2 \kappa_\beta(A, u, \eta, \xi) + o(t^2),$$

negative loss curvature $\kappa_\beta < 0$, equivalently $\mathcal{Q}_\beta > 0$, corresponds to a local escape direction from collapse.

Perturbation path. We initialize the model at the collapsed manifold point,

$$q_i(z) = \mathrm{Dir}(a\mathbf{1}_K), \qquad W_0 = 0, \qquad b_0 = \bar{y}.$$

Note that we can directly set concentration direction $\xi = 0$ since it contributes only to the KL curvature, which is always positive. Then considering the optimal decoder response (A^*, u^*) given by Eqs. (16) and (15), it remains to construct a perturbation of the posterior mean in the simplex tangent space, $\eta(Y) \in T$, to reach the upper bound of the LHS of the following inequality,

$$\left\| \mathbb{E}\tilde{Y}\eta(Y)^T \right\|_F^2 \leq \lambda_{\max}(\mathrm{Cov}(Y)) \, \|\eta\|_2^2 \,. \tag{27}$$

In the symmetric case, changing only the unit tangent direction q_j while keeping the same scalar products $\langle u_{\max}, y_i - \bar{y} \rangle$ gives the same profiled curvature. Therefore, we consider perturbations on the simplex along each vector q_j of an orthonormal basis $q_1, \ldots, q_{K-1}$ of the simplex tangent space, obtained by applying an SVD on the projector operator $P_T = \mathbb{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$. For each direction $j = 1, \ldots, K-1$, we construct a perturbation

$$m_t(y) = m_0 + t\eta_j(y), \qquad s_t(y) = s_0, \qquad W_t = W_0 + tA_j^*, \qquad b_t = b_0 + tu_j^*.$$

We build η_j to reach the upper bound of the inequality (27). Let $u_{\max}$ be the unit eigenvector associated with the largest eigenvalue $\lambda_{\max}(\hat{\Sigma}(Y))$ of the empirical covariance

$$\widehat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y})(y_i - \bar{y})^\top.$$

For each observation y_i, define

$$\eta_j(y_i) = c_\eta \, \langle u_{\max}, y_i - \bar{y} \rangle \, q_j, \qquad j = 1, \ldots, K-1,$$

where

$$c_\eta = \frac{1}{K \max_{i,j} |\langle u_{\max}, y_i - \bar{y} \rangle|}.$$

This last choice ensures the admissibility of the perturbations, i.e., that $\|\eta_j(y_i)\|_\infty \leq 1/K$ for all i, j, so that the perturbed mean $m_t(y_i)$ remains in the interior of the simplex for all amplitudes $|t| < 1$, which includes all amplitudes used in the checks below.

Quadratic coefficient and sign change. Define the closed-form curvature coefficient along the path j as $\kappa_j(\beta) := \kappa_\beta(A_j, u_j, \eta_j, 0)$. For each path j and each ratio $\beta/\beta_c \in \{0.5, 0.8, 1, 1.2, 1.5\}$, we evaluate the loss at $t = -h, 0, h$ with $h = 10^{-3}$, and compare the closed-form coefficient κ_β to the centered finite-difference estimate

$$\widehat{\kappa}_j^{\mathrm{fd}}(\beta) = \frac{\mathcal{L}_\beta(h) + \mathcal{L}_\beta(-h) - 2\mathcal{L}_\beta(0)}{2h^2}.$$

The measured coefficient $\widehat{\kappa}_j^{\mathrm{fd}}(\beta)$ should be negative below β_c, positive above it and close to the theoretical curvature value. Results are reported in Table 6 and can be seen in Fig. 6a. For each fixed $\beta \neq \beta_c$, the median relative error column of Table 6 reports

$$\mathrm{median}_{j=1,\ldots,K-1} \frac{|\widehat{\kappa}_j^{\mathrm{fd}}(\beta) - \kappa_j(\beta)|}{|\kappa_j(\beta)|}.$$

Since at $\beta = \beta_c$ the predicted coefficient vanishes, a relative error is not meaningful.

Curvature around the collapse point. We evaluate the full curve $\Delta\mathcal{L}_\beta(t) := \mathcal{L}_\beta(t) - \mathcal{L}_\beta(0)$ on the grid

$$t \in \{-0.1, -0.075, -0.05, -0.025, -0.01, 0, 0.01, 0.025, 0.05, 0.075, 0.1\}.$$

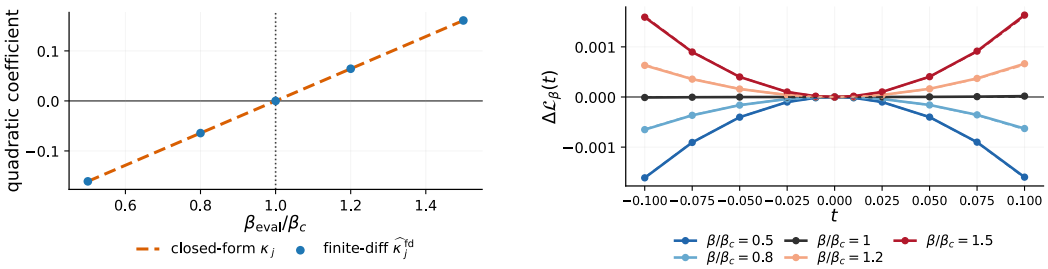
and compare it to the predicted parabola $t^2\kappa_j(\beta)$ with the same β grid as previously.

Results are shown in Fig. 6b. The horizontal coordinate is therefore t, and each curve corresponds to a fixed β/β_c. Table 7 reports the quantities mentioned above and the median absolute relative error,

$$\text{median} \left| \frac{\Delta\mathcal{L}_\beta(t)}{t^2\kappa_j(\beta)} - 1 \right|,$$

computed for each fixed $|t|$ over the $K-1$ tangent directions, $\beta/\beta_c \in \{0.5, 0.8, 1.2, 1.5\}$, and both signs $t = \pm|t|$. This error is expressed as a percentage in the median abs. deviation column.

Results. Globally, both Figures show that the empirical local parabola agrees with the theoretical one $\kappa_j(\beta)$ (Fig. 6b) and that the sign change occurs at $\beta_{\text{eval}}/\beta_c = 1$ (Fig. 6a). This is also confirmed either by the finite-difference check in Table 6, where the closed-form and finite-difference signs agree with a very low median error (median below 5×10^{-7}) for all tangent directions, or by Table 7 where the ratio $\Delta\mathcal{L}_\beta(t)/(t^2\kappa_j)$ is close to one for all tangent directions, evaluation values and perturbation amplitudes. As $|t|$ increases, the perturbation moves away from the collapsed point and the quadratic approximation becomes less accurate. This is consistent with the higher-order remainder in the Taylor expansion.



(a) Local loss coefficient and sign

(b) Local parabolas

Figure 6: **Local curvature and sign-change validation.** (a) Finite-difference loss curvatures κ_j^{fd} around the collapsed point overlap the theoretical coefficient κ_j and change sign as expected at $\beta_{\text{eval}}/\beta_c = 1$. Faint marks show tangent directions. (b) Along a representative perturbation path around the collapse point, the measured objective variations follow the predicted parabolas $t^2\kappa_j(\beta)$ across values of β/β_c.

Table 6: **Finite-difference validation of local curvature.** For each value $\beta_{\text{eval}}/\beta_c$, we report the median closed-form curvature κ_j, the median finite-difference estimate $\widehat{\kappa}_j^{\text{fd}}$, and median relative error over the $K-1=19$ tangent directions. $\beta_{\text{eval}}/\beta_c = 1$ is degenerate, so the relative error is undefined.

$\beta_{\text{eval}}/\beta_c$	median κ_j	median $\widehat{\kappa}_j^{\text{fd}}$	sign agreement	median rel. error
0.5	-1.86×10^{-1}	-1.86×10^{-1}	19/19	1.86×10^{-8}
0.8	-7.45×10^{-2}	-7.45×10^{-2}	19/19	4.76×10^{-7}
1.0	0.00	5.68×10^{-8}	degenerate	n/a
1.2	7.45×10^{-2}	7.45×10^{-2}	19/19	4.76×10^{-7}
1.5	1.86×10^{-1}	1.86×10^{-1}	19/19	3.24×10^{-7}

Table 7: **Agreement between measured objective variations and predicted decoder profiled quadratic curves.** For each perturbation magnitude $|t|$, we report statistics of the ratio $\Delta\mathcal{L}_\beta(t)/(t^2\kappa_j)$ over all $K-1=19$ tangent directions, the four nondegenerate evaluation values $\beta_{\text{eval}}/\beta_{\text{c}} \in \{0.5, 0.8, 1.2, 1.5\}$, and the two signs $t=\pm|t|$. Thus each row contains $19\times 4\times 2 = 152$ scalar ratios. Values close to one indicate agreement with the second-order prediction.

$\lvert t\rvert$	number of ratios	median ratio	median abs. deviation	max abs. deviation
0.010	152	1.00	$1.45\times10^{-1}\%$	$3.13\times10^{-1}\%$
0.025	152	1.00	$3.71\times10^{-1}\%$	$8.24\times10^{-1}\%$
0.050	152	1.00	$7.38\times10^{-1}\%$	1.79%
0.075	152	1.00	$9.89\times10^{-1}\%$	2.89%
0.100	152	9.99×10^{-1}	1.16%	4.14%

C.4 Numerical check of Lemma 3: composition vs. concentration mechanism

We next check the composition versus concentration mechanism of escape from collapse when the prior is symmetric predicted by Lemma 3. This lemma gives, for the best decoder response to perturbations, with $\kappa_\beta(\eta,\xi) = -\mathcal{Q}^\star_\beta(\eta,\xi)$,

$$\begin{aligned}\kappa_\beta(\eta(Y),\xi(Y)) = \quad & -\frac{K(Ka+1)}{2\sigma^2}\left\|\mathbb{E}\left[(Y-\bar{y})\eta(Y)^\top\right]\right\|_{\mathrm{F}}^2 \\ & +\frac{\beta}{2}\left(K^2a^2\psi_1(a)\mathbb{E}\left[\|\eta(Y)\|^2\right]+\left(\frac{\psi_1(a)}{K}-\psi_1(Ka)\right)\mathbb{E}\left[\xi(Y)^2\right]\right).\end{aligned}$$

This can be interpreted as a trade-off between the reconstruction (the first term) and the KL curvatures (the second term), with $\eta(Y)$ and $\xi(Y)$ perturbing the posterior mean and concentration respectively. If reconstruction dominates then $\kappa_\beta(\eta,0) < 0$ (only possible when $\beta < \beta_{\text{c}}$) and the model can escape from collapse. Otherwise, if the KL dominates, the model remains collapsed. Since pure concentration perturbations have $\eta = 0$ and only contribute a nonnegative KL cost, in theory they should not create an escape direction.

Experimental protocol. *Perturbation path.* We reuse the same collapsed point as in Sec. C.3. In particular, the Dirichlet parameters are initialized at

$$m_0 = \frac{1}{K}\mathbf{1}_K, \qquad s_0 = Ka, \qquad \text{hence,} \qquad \gamma_0 = s_0 m_0 = a\mathbf{1}_K.$$

To perturb the posterior parameter, we consider a small perturbation td_i such that

$$\gamma_i(t) = \gamma_0 + td_i, \qquad \text{with } d_i = s_0\eta_i + \xi_i m_0 =: d_i^{\text{comp}} + d_i^{\text{conc}}.$$

We want to construct two types of perturbations: one is purely compositional, i.e. $\eta_i \neq 0$ and $\xi_i = 0$, while the other changes only the concentration, i.e. $\eta_i = 0$ and $\xi_i \neq 0$.

Let $h_i =< u_{\max}, y_i - \bar{y} >$ be the projection of the observation y_i onto the leading empirical covariance direction $u_{\max}$ and let q be the first vector of the orthonormal basis built in Sec. C.3. We then define the two perturbations as

$$\begin{aligned} d_i^{\text{comp}} &= \frac{s_0 h_i q}{\sqrt{\frac{1}{N}\sum_{j=1}^N \|s_0 h_j q\|^2}} = \frac{h_i}{\sqrt{\lambda_{\max}(\hat{\Sigma}(Y))}}q, \\ d_i^{\text{conc}} &= \frac{h_i m_0}{\sqrt{\frac{1}{N}\sum_{j=1}^N \|h_j m_0\|^2}} = \frac{\sqrt{K}h_i}{\sqrt{\lambda_{\max}(\hat{\Sigma}(Y))}}m_0. \end{aligned}$$

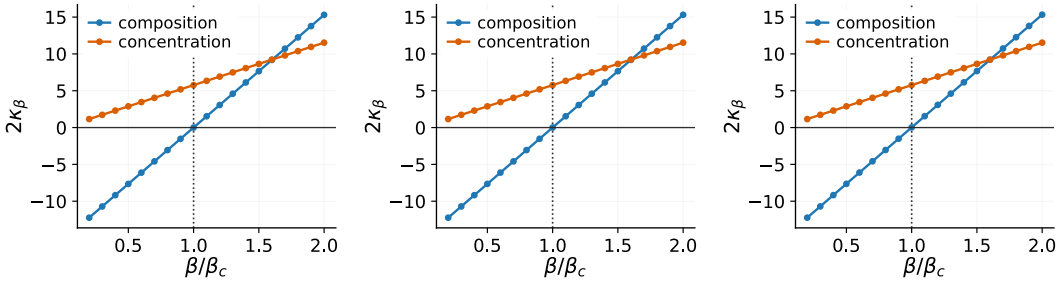
where we normalized each perturbation to ensure that amplitudes are comparable.

For each perturbation path $\gamma_i(t)$, we then recompute the optimal decoder parameters (W, b) by solving the linear least-squares problem associated with the exact reconstruction term in the loss of Eq. (25).

We then compute the centered loss curvature κ_β along the perturbation path using the finite-difference scheme as in Sec. C.3:

$$\widehat{\kappa}_\beta(d) = \frac{\mathcal{L}_\beta(\gamma_0 + td) + \mathcal{L}_\beta(\gamma_0 - td) - 2\mathcal{L}_\beta(\gamma_0)}{2t^2}.$$

Results. The results are shown in Fig. 7 and in Table 8. Clearly, composition perturbations become negative-curvature directions below β_c, while concentration perturbations do not allow the model to escape from collapse for any value of β. This confirms that the escape mechanism from collapse is driven by the tangent motion of the posterior mean rather than by concentration changes alone.



(a) $t = 0.005$ (b) $t = 0.01$ (c) $t = 0.02$

Figure 7: **Composition vs. concentration escape from collapse mechanism.** Loss curvatures $2\kappa_\beta$ are reported for composition and concentration perturbations in Dirichlet-parameter space. Pure composition perturbations of the posterior mean have negative curvature below $\beta_{\text{eval}}/\beta_c = 1$, while the concentration-only direction remains positive for any $\beta > 0$. Thus, for a symmetric prior and at collapse, only perturbation of the posterior mean can create an escape direction.

Table 8: **Composition and concentration curvatures for three perturbation steps.** Finite-difference estimates $2\widehat{\kappa}_\beta$ for $K = 20$ and $a = 1$, rounded to two decimal places in the mantissa.

Perturbation	β/β_c	$t = 0.005$	$t = 0.01$	$t = 0.02$
Composition	0.5	-7.65	-7.65	-7.64
Composition	1.0	1.00×10^{-3}	4.01×10^{-3}	1.60×10^{-2}
Composition	2.0	1.53×10	1.53×10	1.53×10
Concentration	0.5	2.88	2.88	2.88
Concentration	1.0	5.76	5.76	5.76
Concentration	2.0	1.15×10	1.15×10	1.15×10

C.5 Numerical check of Lemma 4: β-ELBO invariance to normal space directions

Lemma 4 predicts that a change in the component of W normal to the tangent space of the simplex that is compensated by the bias b preserves both the reconstruction and the β-ELBO:

$$\mathcal{L}_\beta(\tilde{W} := W + c\mathbf{1}^\top, \tilde{b} := b - c) = \mathcal{L}_\beta(W, b), \qquad c \in \mathbb{R}^D.$$

This has several implications for the practitioners. The parameter norm $\|W\|_F$ can drift and penalizing it can be misleading. Moreover, regularization by penalization of $\|W\|_F$ can act differently for two decoder parameterizations with the same tangent component WP_T hence producing the same reconstruction.

In this experiment, we demonstrate that varying the component of W normal to the simplex tangent space, with a compensating change in b, leaves the β-ELBO unchanged. We compare unconstrained training with training in which W is projected onto the simplex tangent space and its column mean $c = \frac{1}{K}W\mathbf{1}$ is absorbed into b.

Experimental protocol. *Training regimes.* We consider three training regimes to verify how amortized inference and Monte Carlo estimation affect the comparison between unconstrained decoder and projected decoder training. The first one is the *oracle exact* regime, where we have one free variational parameter γ_i per sample and no amortized encoder. The second one is the *amortized exact* regime, where we use a shared encoder $\gamma_\phi(y_i)$ but keep the closed-form reconstruction term. The third one is the *amortized practical* regime, where we use the same shared encoder but replace the closed-form reconstruction expectation by a Monte Carlo estimator. The losses used by each regime are defined in Sec. C.2 and summarized in Table 5.

Model architecture. For the two amortized regimes, the shared encoder is

$$\gamma_\phi(y_i) = \mathrm{softplus}(f_\phi(y_i)) + \gamma_{\min}.$$

Here f_ϕ is a two-hidden-layer Tanh MLP with hidden width 128, and $\gamma_{\min} = 10^{-4}$. The function $\mathrm{softplus}(u) = \log(1 + e^u)$ is applied componentwise to ensure positive Dirichlet parameters. The oracle exact regime applies the same transformation to per-sample vectors γ_i.

Training settings. We compute the threshold at $\beta_c = 9.3034$ following the practical collapse-risk estimation algorithm using the true σ^2 of the dataset. We sweep

$$\beta/\beta_c \in \{0.1, 0.3, 0.5, 0.7, 0.9, 1.0, 1.1, 1.3, 1.5\}.$$

For each value of β, we train all three model regimes (oracle exact, amortized exact, and amortized practical) under both decoder parameterizations (unconstrained and projected), using three initialization seeds $0, 1, 2$.

The encoder weights use Xavier uniform initialization, with gain $5/3$ computed using PyTorch's `calculate_gain("tanh")` for the two hidden layers and the default gain 1 for the output layer. All encoder biases are initialized to zero. The mixing matrix uses Xavier uniform initialization with default parameters, and b is initialized to zero.

All runs use full-batch Adam with no weight decay to avoid mini-batch noise. We use Adam to help optimization move away from collapse when small escape directions exist. The oracle exact regime is trained for 1200 epochs with learning rate $5 \cdot 10^{-2}$, since the variational parameters converge faster. The two amortized regimes are trained for 3000 epochs with learning rate $3 \cdot 10^{-3}$. Diagnostics are recorded every 100 epochs.

In the unconstrained parameterization, W and b are optimized directly. In the projected parameterization, we project the mixing matrix onto the simplex tangent space and adjust b after initialization and after every optimizer step by applying

$$c(W) = \frac{1}{K} W\mathbf{1}, \qquad W \leftarrow W - c(W)\mathbf{1}^\top, \qquad b \leftarrow b + c(W)$$

to enforce $W\mathbf{1} = 0$.

At every recorded epoch t, we measure the column mean $c_t = \frac{1}{K} W_t\mathbf{1}$, the norm of the column mean $\|c_t\|$, the decoder norm $\|W_t\|_F$, the projected decoder norm $\|W_t P_T\|_F$, the training objective depending on the regime, and the closed-form objective $R_{\mathrm{exact}} + \beta\,\mathrm{KL}$. For the amortized-practical regime, this exact objective is an evaluation diagnostic, not the stochastic objective used for training. Curves report the median over the three seeds, and shaded bands span the minimum-maximum range. The numerical tables report median $[\min, \max]$ over the same seeds.

Results. The diagnostic confirms the role of the tangent component of the mixing matrix. In the unconstrained parameterization, optimization can change the normal component of W, compensated by the offset b, producing large changes in $\|c_t\|$ and in the raw norm $\|W_t\|_F$. In the projected parameterization, the same decoder function is represented with $W\mathbf{1} = 0$, so the column mean is set to zero. The quantities relevant to collapse are therefore the tangent component of the mixing matrix WP_T, the constant-output condition $Wz + b = \bar{y}$ at collapse, and the transverse quadratic form used in the main stability theorem.

Because the vertical scale of Fig. 8 can make the seed-to-seed bands visually negligible, Tables 9–12 report the numerical values behind the summary curves. Each entry is the median over the three initialization seeds, followed by the minimum and maximum across seeds in brackets. These brackets are descriptive seed-to-seed ranges, not confidence intervals.

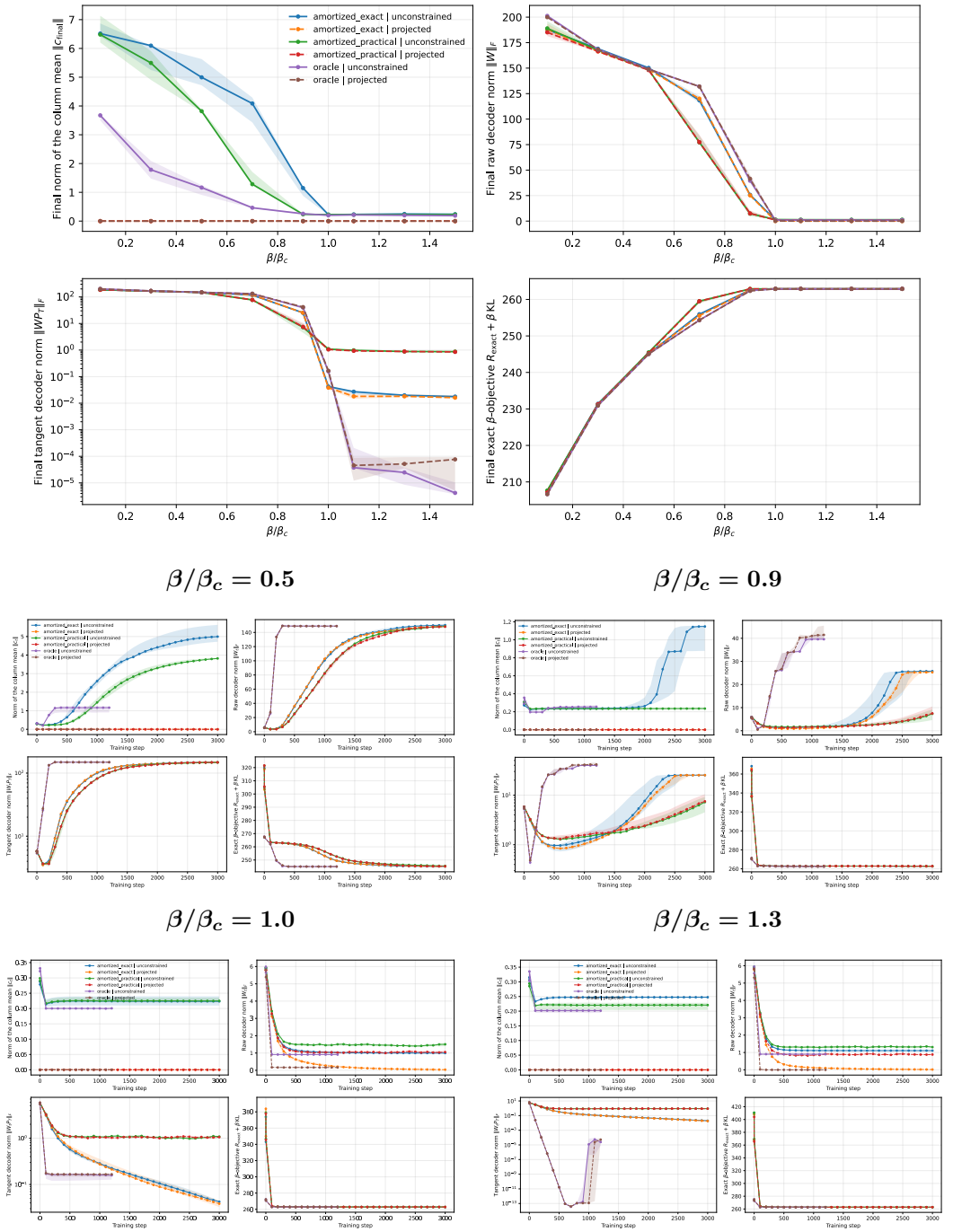


Figure 8: Diagnostics of projection onto the simplex tangent space. Top: final checkpoint summary across β/β_c. Bottom: representative training trajectories below, near, and above the collapse threshold. Each trajectory panel uses the same four metrics: norm of the column mean $\|c_t\|$, raw decoder norm $\|W_t\|_F$, tangent decoder norm $\|W_t P_T\|_F$, and deterministic exact β-objective $R_{\text{exact}} + \beta\,\text{KL}$. Solid curves correspond to the unconstrained decoder parameterization and dashed curves to the projected parameterization. Curves show medians over three initialization seeds. Shaded bands span the minimum-maximum range. Numerical final-checkpoint values are reported in Tables 9–12.

Table 9: Projection diagnostic for the final norm of the column mean $\|c_{\mathrm{final}}\|$. Values are median $[\min, \max]$ over $n = 3$ initialization seeds. U denotes unconstrained training and P denotes projected training.

β/β_c	Mode	Oracle	Amortized exact	Amortized practical
0.1	U	$3.67\,[3.50, 3.70]$	$6.52\,[6.42, 6.87]$	$6.48\,[6.21, 7.14]$
	P	$1.38\,[1.36, 1.51] \times 10^{-15}$	$1.36\,[1.26, 1.60] \times 10^{-15}$	$1.28\,[1.25, 1.41] \times 10^{-15}$
0.3	U	$1.78\,[1.48, 2.09]$	$6.10\,[5.25, 6.10]$	$5.50\,[4.91, 5.91]$
	P	$1.22\,[1.11, 1.24] \times 10^{-15}$	$1.23\,[1.11, 1.24] \times 10^{-15}$	$1.30\,[1.08, 1.39] \times 10^{-15}$
0.5	U	$1.17\,[0.91, 1.25]$	$5.00\,[4.72, 5.63]$	$3.82\,[3.79, 3.87]$
	P	$1.09\,[1.02, 1.14] \times 10^{-15}$	$1.08\,[1.04, 1.17] \times 10^{-15}$	$10.20\,[9.34, 10.30] \times 10^{-16}$
0.7	U	$4.63\,[4.48, 4.96] \times 10^{-1}$	$4.09\,[3.43, 4.29]$	$1.28\,[1.26, 1.71]$
	P	$9.03\,[8.57, 9.89] \times 10^{-16}$	$8.62\,[8.14, 8.74] \times 10^{-16}$	$5.17\,[4.91, 5.51] \times 10^{-16}$
0.9	U	$2.56\,[2.26, 2.72] \times 10^{-1}$	$1.15\,[0.88, 1.16]$	$2.35\,[2.35, 2.49] \times 10^{-1}$
	P	$2.67\,[2.44, 2.83] \times 10^{-16}$	$1.81\,[1.30, 2.33] \times 10^{-16}$	$4.91\,[4.51, 7.37] \times 10^{-17}$
1.0	U	$2.01\,[1.98, 2.03] \times 10^{-1}$	$2.23\,[2.09, 2.40] \times 10^{-1}$	$2.25\,[2.20, 2.36] \times 10^{-1}$
	P	$1.17\,[1.13, 1.85] \times 10^{-18}$	$2.42\,[2.22, 3.36] \times 10^{-19}$	$7.68\,[7.42, 7.72] \times 10^{-18}$
1.1	U	$2.15\,[2.10, 2.18] \times 10^{-1}$	$2.32\,[2.31, 2.44] \times 10^{-1}$	$2.22\,[2.09, 2.27] \times 10^{-1}$
	P	$7.85\,[3.81, 17.10] \times 10^{-19}$	$13.60\,[7.31, 20.30] \times 10^{-20}$	$6.99\,[5.92, 8.71] \times 10^{-18}$
1.3	U	$2.02\,[1.94, 2.08] \times 10^{-1}$	$2.48\,[2.45, 2.48] \times 10^{-1}$	$2.21\,[2.04, 2.28] \times 10^{-1}$
	P	$4.74\,[2.02, 10.70] \times 10^{-19}$	$1.34\,[1.11, 1.96] \times 10^{-19}$	$6.20\,[5.26, 6.22] \times 10^{-18}$
1.5	U	$1.89\,[1.87, 2.01] \times 10^{-1}$	$2.38\,[2.38, 2.56] \times 10^{-1}$	$2.26\,[2.11, 2.49] \times 10^{-1}$
	P	$119.00\,[6.59, 304.00] \times 10^{-20}$	$1.23\,[1.18, 1.27] \times 10^{-19}$	$5.87\,[5.86, 6.37] \times 10^{-18}$

Table 10: Projection diagnostic for the final raw decoder norm $\|W\|_F$. Values are median $[\min, \max]$ over $n = 3$ initialization seeds. U denotes unconstrained training and P denotes projected training.

β/β_c	Mode	Oracle	Amortized exact	Amortized practical
0.1	U	$2.01\,[1.99, 2.01] \times 10^2$	$1.88\,[1.87, 1.89] \times 10^2$	$1.89\,[1.88, 1.96] \times 10^2$
	P	$2.00\,[2.00, 2.01] \times 10^2$	$1.88\,[1.84, 1.94] \times 10^2$	$1.85\,[1.81, 1.88] \times 10^2$
0.3	U	$1.68\,[1.68, 1.68] \times 10^2$	$1.69\,[1.68, 1.69] \times 10^2$	$1.68\,[1.67, 1.69] \times 10^2$
	P	$1.68\,[1.68, 1.68] \times 10^2$	$1.67\,[1.66, 1.67] \times 10^2$	$1.66\,[1.66, 1.67] \times 10^2$
0.5	U	$1.49\,[1.49, 1.49] \times 10^2$	$1.50\,[1.50, 1.51] \times 10^2$	$1.48\,[1.48, 1.49] \times 10^2$
	P	$1.49\,[1.49, 1.49] \times 10^2$	$1.48\,[1.48, 1.49] \times 10^2$	$1.48\,[1.48, 1.48] \times 10^2$
0.7	U	$1.32\,[1.32, 1.32] \times 10^2$	$1.19\,[1.16, 1.23] \times 10^2$	$7.72\,[7.69, 8.39] \times 10$
	P	$1.32\,[1.32, 1.32] \times 10^2$	$1.20\,[1.20, 1.21] \times 10^2$	$7.74\,[7.61, 8.47] \times 10$
0.9	U	$3.96\,[3.96, 4.14] \times 10$	$2.58\,[2.56, 2.58] \times 10$	$7.30\,[4.60, 9.30]$
	P	$4.15\,[3.96, 4.55] \times 10$	$2.53\,[2.53, 2.54] \times 10$	$7.50\,[7.00, 10.50]$
1.0	U	$9.00\,[9.00, 9.00] \times 10^{-1}$	$1.00\,[0.90, 1.10]$	$1.50\,[1.40, 1.50]$
	P	$2.00\,[2.00, 2.00] \times 10^{-1}$	$4.00\,[3.00, 4.00] \times 10^{-2}$	$1.10\,[1.10, 1.20]$
1.1	U	$1.00\,[0.90, 1.00]$	$1.00\,[1.00, 1.10]$	$1.40\,[1.40, 1.40]$
	P	$4.60\,[1.20, 8.90] \times 10^{-5}$	$2.00\,[1.00, 2.00] \times 10^{-2}$	$9.00\,[9.00, 10.00] \times 10^{-1}$
1.3	U	$9.00\,[9.00, 9.00] \times 10^{-1}$	$1.10\,[1.10, 1.10]$	$1.30\,[1.30, 1.40]$
	P	$5.20\,[3.60, 9.30] \times 10^{-5}$	$2.00\,[2.00, 2.00] \times 10^{-2}$	$9.00\,[9.00, 9.00] \times 10^{-1}$
1.5	U	$9.00\,[8.00, 9.00] \times 10^{-1}$	$1.10\,[1.10, 1.10]$	$1.30\,[1.30, 1.40]$
	P	$7.70\,[0.51, 9.30] \times 10^{-5}$	$2.00\,[2.00, 2.00] \times 10^{-2}$	$9.00\,[9.00, 9.00] \times 10^{-1}$

Table 11: Projection diagnostic for the final tangent decoder norm $\|W_T\|_F$. Values are median $[\min, \max]$ over $n = 3$ initialization seeds. U denotes unconstrained training and P denotes projected training.

β/β_c	Mode	Oracle	Amortized exact	Amortized practical
0.1	U	$2.01\,[1.99, 2.01] \times 10^2$	$1.86\,[1.85, 1.87] \times 10^2$	$1.87\,[1.86, 1.93] \times 10^2$
	P	$2.00\,[2.00, 2.01] \times 10^2$	$1.88\,[1.84, 1.94] \times 10^2$	$1.85\,[1.81, 1.88] \times 10^2$
0.3	U	$1.68\,[1.68, 1.68] \times 10^2$	$1.67\,[1.66, 1.68] \times 10^2$	$1.66\,[1.66, 1.67] \times 10^2$
	P	$1.68\,[1.68, 1.68] \times 10^2$	$1.67\,[1.66, 1.67] \times 10^2$	$1.66\,[1.66, 1.67] \times 10^2$
0.5	U	$1.49\,[1.49, 1.49] \times 10^2$	$1.49\,[1.48, 1.49] \times 10^2$	$1.48\,[1.47, 1.48] \times 10^2$
	P	$1.49\,[1.49, 1.49] \times 10^2$	$1.48\,[1.48, 1.49] \times 10^2$	$1.48\,[1.48, 1.48] \times 10^2$
0.7	U	$1.32\,[1.32, 1.32] \times 10^2$	$1.17\,[1.15, 1.21] \times 10^2$	$7.70\,[7.67, 8.36] \times 10$
	P	$1.32\,[1.32, 1.32] \times 10^2$	$1.20\,[1.20, 1.21] \times 10^2$	$7.74\,[7.61, 8.47] \times 10$
0.9	U	$3.96\,[3.96, 4.14] \times 10$	$2.53\,[2.53, 2.53] \times 10$	$7.20\,[4.50, 9.20]$
	P	$4.15\,[3.96, 4.55] \times 10$	$2.53\,[2.53, 2.54] \times 10$	$7.50\,[7.00, 10.50]$
1.0	U	$2.00\,[1.00, 2.00] \times 10^{-1}$	$4.00\,[4.00, 4.00] \times 10^{-2}$	$1.10\,[1.00, 1.10]$
	P	$2.00\,[2.00, 2.00] \times 10^{-1}$	$4.00\,[3.00, 4.00] \times 10^{-2}$	$1.10\,[1.10, 1.20]$
1.1	U	$3.80\,[3.60, 3.80] \times 10^{-5}$	$3.00\,[2.00, 3.00] \times 10^{-2}$	$1.00\,[0.90, 1.00]$
	P	$4.60\,[1.20, 8.90] \times 10^{-5}$	$2.00\,[1.00, 2.00] \times 10^{-2}$	$9.00\,[9.00, 10.00] \times 10^{-1}$
1.3	U	$2.50\,[0.85, 3.70] \times 10^{-5}$	$2.00\,[2.00, 2.00] \times 10^{-2}$	$9.00\,[9.00, 9.00] \times 10^{-1}$
	P	$5.20\,[3.60, 9.30] \times 10^{-5}$	$2.00\,[2.00, 2.00] \times 10^{-2}$	$9.00\,[9.00, 9.00] \times 10^{-1}$
1.5	U	$4.20\,[3.70, 10.00] \times 10^{-6}$	$2.00\,[2.00, 2.00] \times 10^{-2}$	$9.00\,[9.00, 9.00] \times 10^{-1}$
	P	$7.70\,[0.51, 9.30] \times 10^{-5}$	$2.00\,[2.00, 2.00] \times 10^{-2}$	$9.00\,[9.00, 9.00] \times 10^{-1}$

Table 12: Projection diagnostic for the final exact β-objective $R_{\mathrm{exact}} + \beta\,\mathrm{KL}$. Values are median $[\min, \max]$ over $n = 3$ initialization seeds. U denotes unconstrained training and P denotes projected training.

β/β_c	Mode	Oracle	Amortized exact	Amortized practical
0.1	U	$2.07\,[2.07, 2.07] \times 10^2$	$2.08\,[2.07, 2.08] \times 10^2$	$2.08\,[2.07, 2.08] \times 10^2$
	P	$2.07\,[2.07, 2.07] \times 10^2$	$2.07\,[2.07, 2.08] \times 10^2$	$2.07\,[2.07, 2.08] \times 10^2$
0.3	U	$2.31\,[2.31, 2.31] \times 10^2$	$2.31\,[2.31, 2.32] \times 10^2$	$2.31\,[2.31, 2.31] \times 10^2$
	P	$2.31\,[2.31, 2.31] \times 10^2$	$2.31\,[2.31, 2.31] \times 10^2$	$2.31\,[2.31, 2.32] \times 10^2$
0.5	U	$2.45\,[2.45, 2.45] \times 10^2$	$2.45\,[2.45, 2.45] \times 10^2$	$2.46\,[2.45, 2.46] \times 10^2$
	P	$2.45\,[2.45, 2.45] \times 10^2$	$2.45\,[2.45, 2.46] \times 10^2$	$2.45\,[2.45, 2.45] \times 10^2$
0.7	U	$2.54\,[2.54, 2.54] \times 10^2$	$2.56\,[2.55, 2.56] \times 10^2$	$2.60\,[2.59, 2.60] \times 10^2$
	P	$2.54\,[2.54, 2.54] \times 10^2$	$2.56\,[2.55, 2.56] \times 10^2$	$2.60\,[2.59, 2.60] \times 10^2$
0.9	U	$2.62\,[2.62, 2.62] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$
	P	$2.62\,[2.62, 2.62] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$
1.0	U	$2.63\,[2.63, 2.63] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$
	P	$2.63\,[2.63, 2.63] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$
1.1	U	$2.63\,[2.63, 2.63] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$
	P	$2.63\,[2.63, 2.63] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$
1.3	U	$2.63\,[2.63, 2.63] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$
	P	$2.63\,[2.63, 2.63] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$
1.5	U	$2.63\,[2.63, 2.63] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$
	P	$2.63\,[2.63, 2.63] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$	$2.63\,[2.63, 2.63] \times 10^2$

C.6 Prior Shift Experiments

This section provides additional results for the prior-shift experiments of Section 3.4. It demonstrates on 2 synthetic datasets, generated with $K = 20$ and $K = 50$ but keeping the other parameters as in Table 4 (in particular $a_{\text{data}} = 1$), the monotonic relation between the prior concentration a and the predicted scale $\beta_c(a)$.

Experimental protocol. For each training run, we vary the prior concentration a appearing in the KL term of the loss in Eq. (25). It is swept over

$$a \in \{0.1, 0.2, 0.3, 0.5, 0.8, 1, 1.5, 2, 4, 6\}.$$

Then, for each a, we compute the predicted scale $\beta_c(a)$ from the empirical covariance of the dataset and the true noise variance σ^2, following the practical collapse-risk estimation algorithm. Varying

$$\beta/\beta_c(a) \in \{0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.6, 2.0, 3.0\},$$

for each pair $(a, \beta/\beta_c(a))$, we train a new model in the *amortized exact* regime with the loss in Eq. (25). The encoder architecture and Xavier-uniform initialization follow Sec. C.5. Each run uses 2000 epochs of full-batch Adam with learning rate 3×10^{-3}, no weight decay, and float64 precision. We repeat each setting with three initialization seeds $0, 1, 2$. We record KL_ϕ, WP_T, and U_m every 500 epochs and summarize their final values in the tables by the median and the minimum and maximum across seeds.

Results. Figure 5 in the main text shows all three diagnostics for the $K = 20$ dataset, with medians over three initialization seeds. On the raw β axis, changing a shifts the transition over a wide range. After plotting the same runs against $\beta/\beta_c(a)$, the encoder-side diagnostics realign around the predicted scale. The tangent decoder norm $\|WP_T\|_F$ follows the same broad trend, but should be read as a geometric decoder diagnostic: in finite-time training, small residual tangent weights may remain even when the posterior means have essentially collapsed.

Figure 9 repeats the same experiment for $K = 50$. The larger simplex dimension changes the raw scale of $\beta_c(a)$, but the qualitative behavior is the same: raw β is not comparable across priors, whereas $\beta/\beta_c(a)$ organizes the transition.

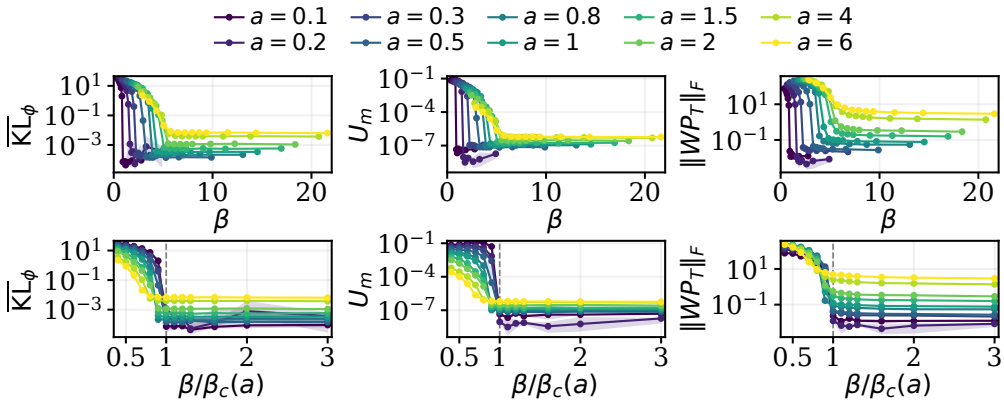


Figure 9: **Prior-shift diagnostics for $K = 50$.** Same layout as Figure 5: the three columns show KL, U_m, and $\|WP_T\|_F$, with raw β above and $\beta/\beta_c(a)$ below. The normalized coordinate $\beta/\beta_c(a)$ again aligns the collapse transition across Dirichlet prior concentrations. Curves show medians over three initialization seeds. Shaded bands span the minimum-maximum range.

Table 13: Prior-shift trained diagnostics for $K = 20$ with selected prior concentrations. Values are median $[\min, \max]$ over $n = 3$ seeds. Powers of ten apply to the median and both bounds. Rows are grouped by prior a and observed ratio $\beta/\beta_{\mathrm{c}}(a)$.

a	$\beta/\beta_{\mathrm{c}}(a)$	KL_ϕ	$\|W_T\|_F$	U_m
0.1	0.5	$1.2\,[1.1, 1.2] \times 10$	$6.0\,[5.6, 6.0] \times 10$	$2.1\,[2.1, 2.1] \times 10^{-1}$
0.1	0.9	$8.1\,[8.1, 8.2] \times 10^{-1}$	$8.2\,[8.2, 8.2]$	$6.2\,[6.2, 6.2] \times 10^{-2}$
0.1	1	$3.0\,[1.5, 21.8] \times 10^{-5}$	$19.3\,[7.3, 43.9] \times 10^{-4}$	$17.3\,[1.1, 286.0] \times 10^{-10}$
0.1	1.2	$83.4\,[8.3, 1220.0] \times 10^{-6}$	$23.4\,[9.1, 27.8] \times 10^{-4}$	$105.0\,[8.9, 181.0] \times 10^{-10}$
0.5	0.5	$8.1\,[8.0, 8.1]$	$1.2\,[1.2, 1.2] \times 10^{2}$	$7.5\,[7.4, 7.5] \times 10^{-2}$
0.5	0.9	$2.7\,[2.7, 2.7] \times 10^{-1}$	$1.8\,[1.8, 1.8] \times 10$	$5.0\,[5.0, 5.0] \times 10^{-3}$
0.5	1	$2.3\,[1.2, 3.0] \times 10^{-5}$	$1.4\,[1.3, 2.1] \times 10^{-2}$	$5.2\,[3.0, 6.8] \times 10^{-8}$
0.5	1.2	$1.7\,[1.6, 2.7] \times 10^{-5}$	$8.8\,[8.7, 11.4] \times 10^{-3}$	$3.8\,[3.0, 7.1] \times 10^{-8}$
0.8	0.5	$6.5\,[6.5, 6.5]$	$1.3\,[1.3, 1.4] \times 10^{2}$	$4.5\,[4.4, 4.5] \times 10^{-2}$
0.8	0.9	$2.5\,[2.5, 18.5] \times 10^{-2}$	$9.5\,[9.5, 22.2]$	$8.8\,[8.5, 163.0] \times 10^{-5}$
0.8	1	$12.2\,[9.8, 13.0] \times 10^{-5}$	$5.1\,[4.2, 6.0] \times 10^{-2}$	$2.6\,[2.1, 3.1] \times 10^{-7}$
0.8	1.2	$1.1\,[0.9, 1.4] \times 10^{-4}$	$2.8\,[2.6, 3.3] \times 10^{-2}$	$2.2\,[1.7, 3.1] \times 10^{-7}$
1	0.5	$5.7\,[5.7, 5.8]$	$1.4\,[1.4, 1.4] \times 10^{2}$	$3.2\,[3.2, 3.3] \times 10^{-2}$
1	0.9	$4.7\,[3.0, 74.8] \times 10^{-4}$	$12.6\,[7.8, 59.9] \times 10^{-1}$	$13.1\,[7.7, 226.0] \times 10^{-7}$
1	1	$1.8\,[1.8, 2.1] \times 10^{-4}$	$7.9\,[7.5, 10.5] \times 10^{-2}$	$4.0\,[3.9, 5.0] \times 10^{-7}$
1	1.2	$2.0\,[2.0, 2.1] \times 10^{-4}$	$4.9\,[4.7, 5.1] \times 10^{-2}$	$4.7\,[4.5, 4.7] \times 10^{-7}$
2	0.5	$3.3\,[3.1, 3.4]$	$1.5\,[1.5, 1.6] \times 10^{2}$	$8.9\,[8.2, 9.2] \times 10^{-3}$
2	0.9	$8.4\,[5.9, 10.6] \times 10^{-4}$	$13.4\,[8.0, 19.7] \times 10^{-1}$	$1.6\,[1.1, 2.0] \times 10^{-6}$
2	1	$6.8\,[6.4, 7.1] \times 10^{-4}$	$2.8\,[2.5, 3.1] \times 10^{-1}$	$1.2\,[1.2, 1.3] \times 10^{-6}$
2	1.2	$6.2\,[6.1, 6.7] \times 10^{-4}$	$1.6\,[1.6, 1.6] \times 10^{-1}$	$1.1\,[1.1, 1.2] \times 10^{-6}$
4	0.5	$1.1\,[1.1, 1.2]$	$1.3\,[1.3, 1.3] \times 10^{2}$	$1.2\,[1.2, 1.3] \times 10^{-3}$
4	0.9	$3.5\,[3.5, 3.8] \times 10^{-3}$	$2.9\,[1.9, 2.9]$	$3.5\,[3.3, 3.6] \times 10^{-6}$
4	1	$3.5\,[3.3, 3.6] \times 10^{-3}$	$8.7\,[8.2, 10.4] \times 10^{-1}$	$3.3\,[3.2, 3.5] \times 10^{-6}$
4	1.2	$3.7\,[3.2, 3.8] \times 10^{-3}$	$5.6\,[5.4, 6.3] \times 10^{-1}$	$3.6\,[3.0, 3.7] \times 10^{-6}$
6	0.5	$6.9\,[6.9, 7.4] \times 10^{-1}$	$1.2\,[1.2, 1.3] \times 10^{2}$	$5.2\,[5.1, 5.5] \times 10^{-4}$
6	0.9	$6.3\,[6.2, 6.8] \times 10^{-3}$	$3.1\,[2.9, 3.3]$	$3.7\,[3.7, 4.1] \times 10^{-6}$
6	1	$6.9\,[6.7, 6.9] \times 10^{-3}$	$1.8\,[1.8, 1.9]$	$4.1\,[4.0, 4.1] \times 10^{-6}$
6	1.2	$7.0\,[6.9, 7.1] \times 10^{-3}$	$1.2\,[1.2, 1.3]$	$4.1\,[4.1, 4.4] \times 10^{-6}$

Table 14: Prior-shift trained diagnostics for $K = 50$ with all prior concentrations. Values are median $[\min, \max]$ over $n = 3$ seeds. Powers of ten apply to the median and both bounds. Rows are grouped by prior a and observed ratio $\beta/\beta_c(a)$.

a	$\beta/\beta_c(a)$	KL_ϕ	$\|W_T\|_F$	U_m
0.1	0.5	$2.7\,[2.7, 2.8] \times 10$	$7.4\,[7.3, 7.6] \times 10$	$8.1\,[8.0, 8.1] \times 10^{-2}$
0.1	0.9	$2.0\,[1.1, 2.1]$	$8.7\,[5.7, 8.7]$	$5.1\,[3.0, 5.2] \times 10^{-2}$
0.1	1	$7.2\,[5.4, 7.5] \times 10^{-5}$	$2.3\,[2.2, 2.6] \times 10^{-2}$	$4.5\,[4.0, 5.5] \times 10^{-8}$
0.1	1.2	$7.8\,[6.8, 8.6] \times 10^{-5}$	$1.7\,[1.4, 1.9] \times 10^{-2}$	$5.2\,[4.1, 5.2] \times 10^{-8}$
0.5	0.5	$1.7\,[1.6, 1.7] \times 10$	$1.8\,[1.8, 1.8] \times 10^{2}$	$2.7\,[2.7, 2.7] \times 10^{-2}$
0.5	0.9	$4.5\,[3.0, 24.5] \times 10^{-2}$	$10.2\,[8.3, 19.0]$	$2.9\,[2.0, 75.7] \times 10^{-5}$
0.5	1	$1.5\,[1.5, 1.6] \times 10^{-4}$	$4.4\,[4.4, 4.5] \times 10^{-2}$	$7.3\,[7.2, 7.4] \times 10^{-8}$
0.5	1.2	$1.4\,[1.4, 1.7] \times 10^{-4}$	$3.5\,[3.3, 3.5] \times 10^{-2}$	$7.1\,[6.9, 7.2] \times 10^{-8}$
0.8	0.5	$1.3\,[1.3, 1.3] \times 10$	$2.1\,[2.1, 2.1] \times 10^{2}$	$1.4\,[1.4, 1.5] \times 10^{-2}$
0.8	0.9	$2.1\,[2.1, 3.9] \times 10^{-4}$	$1.7\,[1.2, 9.2] \times 10^{-1}$	$1.1\,[1.0, 2.0] \times 10^{-7}$
0.8	1	$2.2\,[2.1, 2.2] \times 10^{-4}$	$8.5\,[8.2, 8.8] \times 10^{-2}$	$10.7\,[10.0, 10.9] \times 10^{-8}$
0.8	1.2	$2.2\,[2.1, 2.3] \times 10^{-4}$	$6.8\,[6.7, 7.1] \times 10^{-2}$	$1.1\,[1.1, 1.1] \times 10^{-7}$
1	0.5	$1.1\,[1.1, 1.1] \times 10$	$2.2\,[2.2, 2.2] \times 10^{2}$	$9.9\,[9.7, 9.9] \times 10^{-3}$
1	0.9	$3.8\,[3.6, 25.7] \times 10^{-4}$	$5.4\,[3.8, 38.2] \times 10^{-1}$	$1.5\,[1.4, 12.2] \times 10^{-7}$
1	1	$3.2\,[3.2, 3.6] \times 10^{-4}$	$1.2\,[1.1, 1.3] \times 10^{-1}$	$1.3\,[1.2, 1.4] \times 10^{-7}$
1	1.2	$3.5\,[3.4, 3.5] \times 10^{-4}$	$9.8\,[9.7, 10.2] \times 10^{-2}$	$1.3\,[1.3, 1.4] \times 10^{-7}$
2	0.5	$4.4\,[4.3, 4.5]$	$2.1\,[2.0, 2.1] \times 10^{2}$	$1.6\,[1.6, 1.7] \times 10^{-3}$
2	0.9	$1.2\,[1.2, 1.3] \times 10^{-3}$	$1.2\,[1.0, 1.3]$	$3.1\,[2.9, 3.3] \times 10^{-7}$
2	1	$1.2\,[1.2, 1.2] \times 10^{-3}$	$5.6\,[5.3, 6.3] \times 10^{-1}$	$2.8\,[2.8, 3.1] \times 10^{-7}$
2	1.2	$1.1\,[1.1, 1.1] \times 10^{-3}$	$4.0\,[3.9, 4.3] \times 10^{-1}$	$2.8\,[2.6, 2.8] \times 10^{-7}$
4	0.5	$1.6\,[1.6, 1.7]$	$1.9\,[1.9, 1.9] \times 10^{2}$	$2.8\,[2.8, 2.9] \times 10^{-4}$
4	0.9	$4.2\,[4.2, 4.3] \times 10^{-3}$	$3.7\,[3.6, 3.8]$	$5.9\,[5.9, 6.0] \times 10^{-7}$
4	1	$3.9\,[3.7, 4.0] \times 10^{-3}$	$2.5\,[2.4, 2.5]$	$5.4\,[5.0, 5.6] \times 10^{-7}$
4	1.2	$3.9\,[3.8, 3.9] \times 10^{-3}$	$2.0\,[2.0, 2.0]$	$5.3\,[5.1, 5.4] \times 10^{-7}$
6	0.5	$8.7\,[8.6, 9.2] \times 10^{-1}$	$1.6\,[1.6, 1.7] \times 10^{2}$	$1.0\,[1.0, 1.1] \times 10^{-4}$
6	0.9	$7.1\,[7.0, 7.7] \times 10^{-3}$	$6.3\,[6.2, 6.6]$	$6.6\,[6.4, 7.2] \times 10^{-7}$
6	1	$6.7\,[6.7, 6.8] \times 10^{-3}$	$4.9\,[4.7, 4.9]$	$6.2\,[6.1, 6.2] \times 10^{-7}$
6	1.2	$6.9\,[6.5, 7.0] \times 10^{-3}$	$4.1\,[4.0, 4.1]$	$6.3\,[5.8, 6.3] \times 10^{-7}$

D Hyperspectral imaging (HSI): unmixing details and additional experiments

Hyperspectral unmixing represents each pixel spectrum as a mixture of endmember spectra, with nonnegative abundances summing to one. We use $y_i \mid z_i \sim \mathcal{N}(Wz_i, \sigma^2 I_D)$ with $W \geq 0$, $z_i \in \Delta^{K-1}$, and no decoder bias. The columns of W are the endmembers and the posterior mean m_i gives the estimated abundances. We test the relevance of the local collapse scale on Samson and JasperRidge (Zhu, 2017).

Samson contains 9025 pixels on a 95×95 grid, with $D = 156$ bands and $K = 3$ endmembers. JasperRidge contains 10000 pixels on a 100×100 grid, with $D = 198$ and $K = 4$. Both datasets are loaded from MATLAB files, flattened in MATLAB column-major order, and normalized by dividing all observations y_i by the global maximum of the dataset, i.e. $\max_{i,d} y_{id}$. The ground-truth endmembers and abundance maps are used for evaluation only.

For each dataset, we center the data matrix Y and retain its leading K singular components to form $\hat{Y}_c$ using singular value decomposition (SVD). We set

$$\hat{\sigma}^2 = \frac{\|Y_c - \hat{Y}_c\|_F^2}{ND}$$

and compute $\hat{\beta}_c$ from Equation (3). The same noise variance is kept fixed throughout each sweep. The rank-K residual allows one additional direction beyond the $K - 1$ dimensional tangent space to account for structured spectral variation.

D.1 Hyperspectral unmixing on Samson

Model architecture. The encoder is an MLP with two hidden layers of width 128 and `tanh` activations. Its concentrations are $\gamma_\phi(y) = \mathrm{softplus}(f_\phi(y)) + 10^{-4}$. The prior is $\mathrm{Dir}(\mathbf{1}_3)$. For the quantitative sweep, the nonnegative mixing matrix is parameterized either as $W = 10^{-8} + \mathrm{softplus}(V)$ where V is the decoder's fully-connected layer, or by clamping its entries to nonnegative values after each optimizer update. In both cases the bias is fixed to zero.

Training protocol. For the loss, we use the negative β-ELBO with a closed-form reconstruction and train for 600 epochs with Adam, learning rate 3×10^{-3}, no weight decay, batch size of 256, and Xavier-uniform initialization. The runs use seeds $\{0, 1, 2, 3, 4\}$ and the grid

$$\frac{\beta}{\hat{\beta}_c} \in \{0.1, 0.2, 0.5, 0.8, 0.9, 0.95, 1.0, 1.1, 1.2, 1.5, 2.0\}.$$

On Samson, we estimate the noise at $\hat{\sigma}^2 = 3.1554 \times 10^{-5}$, $\lambda_{\max}(\widehat{\Sigma}) = 2.6895$ and the threshold at $\hat{\beta}_c \simeq 6.909 \times 10^4$. The simplex and abundance-map visualizations in Figs. 10 and 11 are obtained with seed 0.

Results. Figure 1 and Table 2 report the Samson sweep and the corresponding softplus diagnostics. The simplex and abundance panels illustrate the progressive loss of input-dependent latent structure. Figures 10 and 11 show the qualitative sharp transition visually: the learned posterior-mean compositions collapse to the simplex center, hence all abundance maps become uniform. The quantitative diagnostics show this transition in both the collapse and reconstruction metrics. This is consistent with the interpretation that $\hat{\beta}_c$ predicts representation collapse rather than task-optimal regularization.

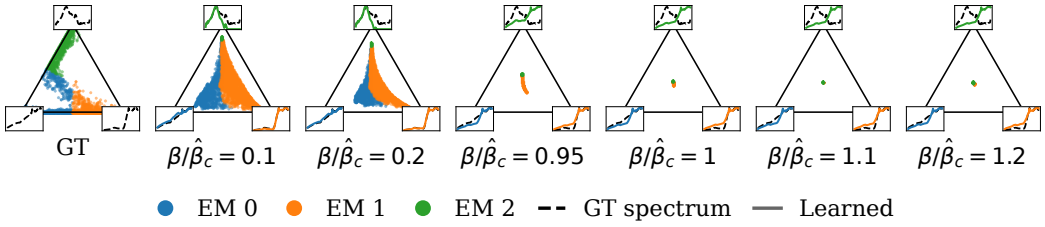


Figure 10: Samson posterior-mean compositions with $a = 1$, the `clamp` nonnegative mixing matrix, Xavier-uniform initialization, and seed 0. Each ratio $\beta/\hat{\beta}_c$ is shown below its panel.

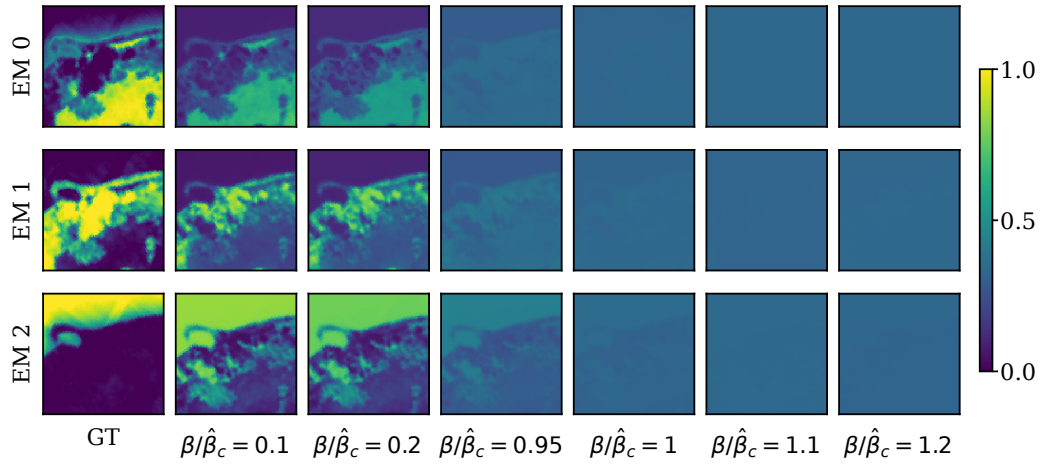


Figure 11: Samson abundance maps for $a = 1$, with the `clamp` nonnegative mixing matrix, Xavier-uniform initialization, and seed 0. Rows correspond to endmembers and columns compare the ground truth with the learned abundance maps at the ratios $\beta/\hat{\beta}_c$ shown below. All maps use the same color scale.

D.2 Hyperspectral unmixing on JasperRidge

Model architecture. We use the same encoder and nonnegative mixing-matrix parameterizations as for Samson, with input dimension 198, $K = 4$, and prior $\mathrm{Dir}(\mathbf{1}_4)$.

Training protocol. The optimizer, initialization, 600 epochs, five seeds, and normalized $\beta/\hat{\beta}_c$ grid are the same as for Samson. The batch size is 512, giving 12000 optimizer updates per run. The calibration gives $\hat{\sigma}^2 = 7.4760 \times 10^{-5}$, $\lambda_{\max}(\widehat{\Sigma}) = 4.8295$, and $\hat{\beta}_c \simeq 49{,}090$. The abundance maps in Fig. 12 are obtained from the earlier Monte Carlo sweep with 4 reconstruction samples and 3000 minibatch updates.

Results. For both nonnegative mixing-matrix parameterizations, low values of $\beta/\hat{\beta}_c$ produce input-dependent posteriors and nonzero tangent activity. Around $\beta/\hat{\beta}_c = 1$, the continuous collapse diagnostics sharply decrease. The abundance maps lose spatial structure, and RMSE and SAD plateau near the constant-mixture regime. This is consistent with interpreting $\hat{\beta}_c$ as a collapse scale rather than task-optimal regularization.

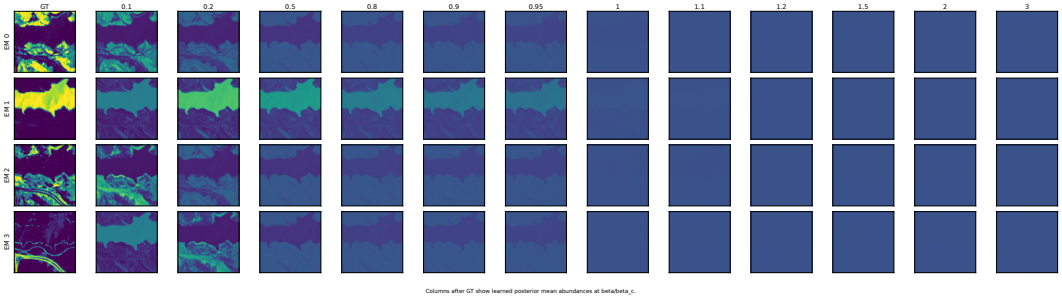


Figure 12: Representative JasperRidge posterior-mean abundance maps with $a = 1$, the `clamp` nonnegative mixing matrix, Xavier-uniform initialization, and seed 0. These abundance maps come from the Monte Carlo sweep and illustrate the loss of information, hence spatial structure.

Table 15: JasperRidge sweep with the `softplus` nonnegative mixing matrix and $a = 1$. Metrics give median $\pm$ half the interquartile range. The lower panel also gives the corresponding KL weights. The sweep contains seeds $\{0, 1, 2, 3, 4\}$.

$\beta/\hat{\beta}_c$	$\overline{\mathrm{KL}}_\phi$	U_m	$\|W_T\|_F$
0.100	1.8 ± 0.0015	$(1.7 \pm 0.005) \times 10^{-1}$	$5.5 \pm 1.1 \times 10^{-3}$
0.200	1.1 ± 0.001	$(1.1 \pm 0.0) \times 10^{-1}$	$5.2 \pm 2.9 \times 10^{-4}$
0.500	$(3.8 \pm 0.035) \times 10^{-1}$	$(3.6 \pm 0.2) \times 10^{-2}$	$4.2 \pm 3.3 \times 10^{-1}$
0.800	$(1.6 \pm 0.005) \times 10^{-1}$	$(1.4 \pm 0.0) \times 10^{-2}$	$3.9 \pm 5.4 \times 10^{-3}$
0.900	$(1.0 \pm 0.005) \times 10^{-1}$	$(8.6 \pm 0.07) \times 10^{-3}$	$3.5 \pm 2.4 \times 10^{-2}$
0.950	$(2.1 \pm 0.6) \times 10^{-2}$	$(1.6 \pm 0.52) \times 10^{-3}$	$1.6 \pm 2.3 \times 10^{-1}$
1.000	$(9.7 \pm 5.9) \times 10^{-5}$	$(2.9 \pm 1.9) \times 10^{-6}$	$2.0 \times 10^{-2} \pm 4.1 \times 10^{-3}$
1.100	$(2.9 \pm 0.75) \times 10^{-5}$	$(7.4 \pm 0.85) \times 10^{-7}$	$1.7 \times 10^{-3} \pm 1.8 \times 10^{-4}$
1.200	$(1.7 \pm 0.75) \times 10^{-5}$	$(4.1 \pm 2.9) \times 10^{-7}$	$1.7 \times 10^{-3} \pm 2.8 \times 10^{-4}$
1.500	$(2.4 \pm 4.8) \times 10^{-6}$	$(8.5 \pm 12.0) \times 10^{-8}$	$1.4 \times 10^{-3} \pm 1.8 \times 10^{-4}$
2.000	$(2.0 \pm 1.6) \times 10^{-5}$	$(4.7 \pm 3.4) \times 10^{-7}$	$1.3 \times 10^{-3} \pm 1.9 \times 10^{-4}$

$\beta/\hat{\beta}_c$	$\beta/10^3$	RMSE	SAD
0.100	4.9	$4.8 \times 10^{-2} \pm 1.9 \times 10^{-5}$	$2.3 \times 10^{-1} \pm 3.6 \times 10^{-4}$
0.200	9.8	$7.1 \times 10^{-2} \pm 7.1 \times 10^{-6}$	$2.4 \times 10^{-1} \pm 1.7 \times 10^{-4}$
0.500	2.5×10	$1.2 \times 10^{-1} \pm 2.1 \times 10^{-3}$	$2.9 \times 10^{-1} \pm 1.1 \times 10^{-3}$
0.800	3.9×10	$1.4 \times 10^{-1} \pm 9.2 \times 10^{-5}$	$2.6 \times 10^{-1} \pm 3.2 \times 10^{-3}$
0.900	4.4×10	$1.5 \times 10^{-1} \pm 1.5 \times 10^{-4}$	$2.3 \times 10^{-1} \pm 6.6 \times 10^{-3}$
0.950	4.7×10	$1.6 \times 10^{-1} \pm 1.3 \times 10^{-3}$	$4.2 \times 10^{-1} \pm 7.5 \times 10^{-3}$
1.000	4.9×10	$1.7 \times 10^{-1} \pm 6.7 \times 10^{-7}$	$4.5 \times 10^{-1} \pm 1.3 \times 10^{-4}$
1.100	5.4×10	$1.7 \times 10^{-1} \pm 4.8 \times 10^{-7}$	$4.5 \times 10^{-1} \pm 6.6 \times 10^{-5}$
1.200	5.9×10	$1.7 \times 10^{-1} \pm 2.7 \times 10^{-7}$	$4.5 \times 10^{-1} \pm 2.0 \times 10^{-4}$
1.500	7.4×10	$1.7 \times 10^{-1} \pm 4.0 \times 10^{-7}$	$4.5 \times 10^{-1} \pm 5.1 \times 10^{-5}$
2.000	9.8×10	$1.7 \times 10^{-1} \pm 1.1 \times 10^{-7}$	$4.5 \times 10^{-1} \pm 8.9 \times 10^{-5}$

E Convex archetypal analysis experiments: supplementary details

Convex archetypal analysis approximates each observation as a convex combination of a small set of archetypes. Here, we learn the archetypes freely, without requiring them to be convex combinations of the observations. The combination coefficients are nonnegative and sum to one:

$$Wz + b = \sum_{k=1}^{K} z_k W_{:k} + b, \qquad z \in \Delta^{K-1},$$

where $W_{:k}$ denotes the k-th column of W.

Datasets. We report complementary results on MNIST (LeCun et al., 2010), SVHN (Netzer et al., 2011), and FashionMNIST (Xiao et al., 2017). SVHN contains more natural images of street-view digits that have more complicated generative factors than MNIST or FashionMNIST. Class labels are used only for visualization. Images are converted to floating point tensors in $[0, 1]$ and flattened, without augmentation. MNIST and FashionMNIST have $D = 784$ pixels, and SVHN has $D = 3072$ values from 32×32 RGB images. Each dataset uses a fixed evaluation subset of 2000 training images.

Model. We use a Dirichlet β-VAE with $K = 3$ for visualization and prior $\mathrm{Dir}(\mathbf{1}_K)$. For each fixed β, we jointly optimize all encoder parameters ϕ and the decoder parameters W and b by minimizing the negative β-ELBO with closed-form reconstruction. After training, we extract the model's archetypes a_k by evaluating the trained decoder at each vertex e_k of the latent simplex,

$$a_k = We_k + b, \qquad Wz + b = \sum_{k=1}^{K} z_k a_k, \qquad z \in \Delta^{K-1}.$$

Here $e_k \in \mathbb{R}^K$ is the k-th standard basis vector, with a 1 in position k and zeros elsewhere.

β_c estimation. For each dataset, the noise variance is calibrated from the empirical covariance by removing its leading $K - 1$ eigenvalues,

$$\hat{\sigma}^2 = \frac{1}{D}\left(\mathrm{tr}(\widehat{\Sigma}) - \sum_{j=1}^{K-1} \hat{\lambda}_j\right).$$

This variance is used in the loss and kept fixed during training. The reference scale $\hat{\beta}_c$ is then computed using Equation (3). We report the ratios

$$\frac{\beta}{\hat{\beta}_c} \in \{0.1, 0.2, 0.5, 0.8, 0.9, 0.95, 1.0, 1.05, 1.2, 1.5, 2.0\}.$$

Collapse metrics are defined in Equation (6).

E.1 CAA on MNIST

Model architecture. We use a Dirichlet β-VAE with $K = 3$, $a = 1$, and a linear decoder (W, b). The encoder has two hidden layers of width 256 with `tanh` activations, followed by $\gamma_\phi(x) = \mathrm{softplus}(f_\phi(x)) + \epsilon$ with $\epsilon = 10^{-4}$ for numerical stability. The decoded archetypes are $We_k + b$.

Training protocol. We retain the classes $\{0, 1, 7\}$, giving 18,930 training images. The estimation gives $\hat{\sigma} = 0.20529$, $\lambda_{\max}(\widehat{\Sigma}) = 11.3121$, and $\hat{\beta}_c = 217.57$. We minimize the negative β-ELBO with closed-form reconstruction for 3000 epochs with Adam, a learning rate of 3×10^{-3}, batch size 256, no weight decay, and seeds $\{0, 1, 2, 3, 4\}$.

Table 16 and Fig. 13 report the MNIST results for classes $\{0, 1, 7\}$ with the full sweep. A non-collapsed model produces class-dependent occupancy of the simplex and distinct archetypes.

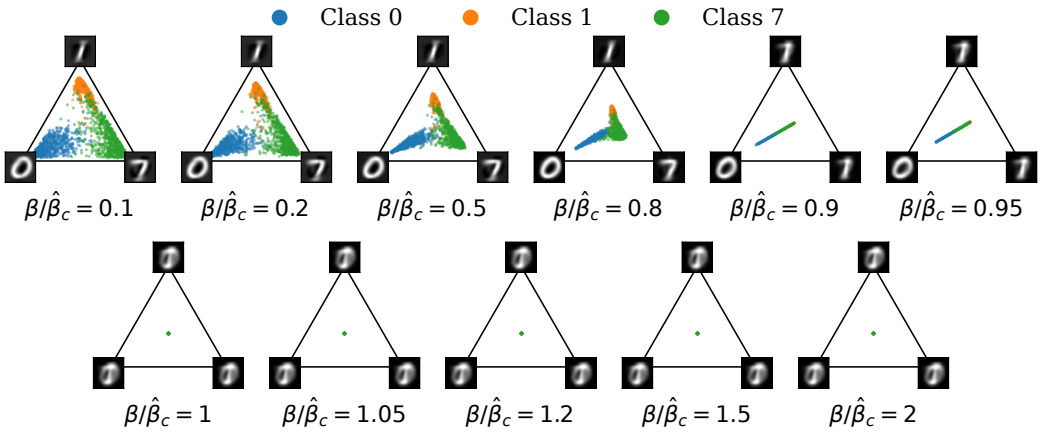


Figure 13: Full MNIST simplex sweep, extending Figure 2 to all ratios in Table 16. Panels show the posterior mean compositions of the same 2000 evaluation images from classes $\{0, 1, 7\}$, with $K = 3$, $a = 1$, and seed 0. Colors indicate class labels, and the images at the vertices are the learned decoder archetypes.

Table 16: CAA on MNIST classes $\{0, 1, 7\}$ with $K = 3$, $a = 1$, and closed-form reconstruction. Entries give mean $\pm$ sample standard deviation over the five runs at each ratio.

$\beta/\hat{\beta}_c$	$\overline{\mathrm{KL}}_\phi$	$\|W_T\|_F$	U_m
0.1	$1.9 \pm 2.6 \times 10^{-3}$	$1.2 \times 10 \pm 4.2 \times 10^{-3}$	$2.1 \times 10^{-1} \pm 8.3 \times 10^{-4}$
0.2	$1.4 \pm 2.1 \times 10^{-3}$	$1.1 \times 10 \pm 3.7 \times 10^{-3}$	$1.8 \times 10^{-1} \pm 5.6 \times 10^{-4}$
0.5	$7.2 \times 10^{-1} \pm 5.5 \times 10^{-3}$	$1.0 \times 10 \pm 1.2 \times 10^{-2}$	$1.1 \times 10^{-1} \pm 1.3 \times 10^{-3}$
0.8	$2.9 \times 10^{-1} \pm 2.8 \times 10^{-3}$	$7.3 \pm 1.2 \times 10^{-2}$	$4.3 \times 10^{-2} \pm 3.3 \times 10^{-4}$
0.9	$1.3 \times 10^{-1} \pm 3.0 \times 10^{-3}$	$5.1 \pm 1.2 \times 10^{-2}$	$1.9 \times 10^{-2} \pm 3.8 \times 10^{-4}$
0.95	$1.0 \times 10^{-1} \pm 1.2 \times 10^{-3}$	$4.6 \pm 9.9 \times 10^{-3}$	$1.5 \times 10^{-2} \pm 4.5 \times 10^{-5}$
1	$6.6 \times 10^{-4} \pm 1.2 \times 10^{-3}$	$(3.1 \pm 2.4) \times 10^{-3}$	$(3.0 \pm 6.8) \times 10^{-6}$
1.05	$9.0 \times 10^{-5} \pm 1.2 \times 10^{-4}$	$(4.7 \pm 2.6) \times 10^{-3}$	$(3.3 \pm 4.6) \times 10^{-8}$
1.2	$(3.3 \pm 5.4) \times 10^{-4}$	$(5.7 \pm 2.7) \times 10^{-3}$	$5.9 \times 10^{-6} \pm 1.3 \times 10^{-5}$
1.5	$(7.3 \pm 8.5) \times 10^{-5}$	$(3.9 \pm 3.0) \times 10^{-3}$	$(1.6 \pm 3.6) \times 10^{-9}$
2	$(2.0 \pm 2.0) \times 10^{-4}$	$(3.9 \pm 2.9) \times 10^{-3}$	$(3.4 \pm 6.5) \times 10^{-8}$

E.2 CAA on SVHN

Model architecture. The encoder has two hidden layers of width 512 with `tanh` activations followed by a softplus to ensure positivity, with $\gamma_\phi(x) = \mathrm{softplus}(f_\phi(x)) + \epsilon$ with $\epsilon = 10^{-4}$. The decoder is linear and we set $K = 3$ and $a = 1$.

Training protocol. We use 3000 training images from each of the digit classes $\{0, 1, 7\}$, for 9000 images in total, with data seed 0. The estimation gives $\hat{\sigma} \simeq 0.11953$ and $\hat{\beta}_c \simeq 3917.88$. We train for 3000 epochs using Adam, learning rate 3×10^{-4}, batch size 256, no weight decay, and seeds $\{0, 1, 2, 3, 4\}$. These runs use Xavier-uniform initialization. The normalized grid and projection of the mixing matrix are the same as for MNIST.

Results. Figure 14 extends the simplex plots of Figure 3 to the full sweep. Table 17 reports the full sweep. The mean diagnostics decrease sharply between $\beta/\hat{\beta}_c = 0.95$ and 1, consistent with a loss of input-dependent posterior means near the estimated collapse threshold.

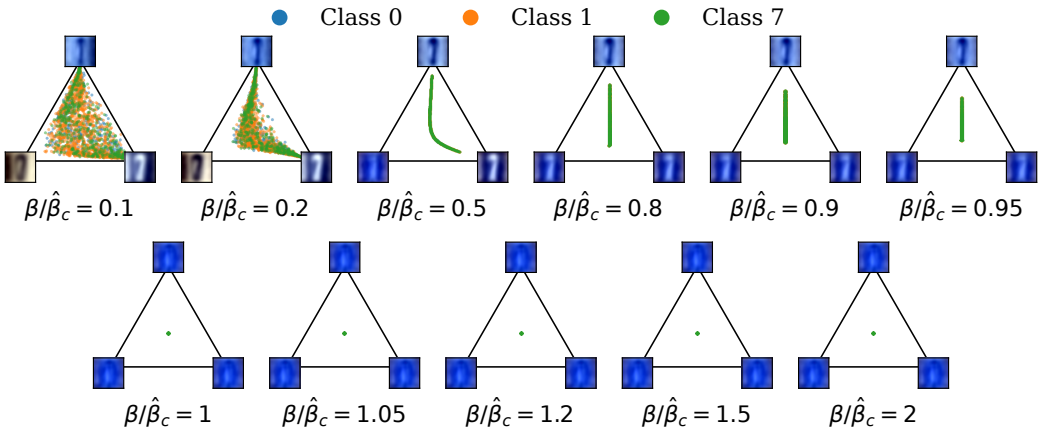


Figure 14: Full SVHN simplex sweep, extending Figure 3 to all ratios in Table 17. Panels show posterior mean compositions from classes $\{0, 1, 7\}$, with $K = 3$, $a = 1$, and seed 0. Colors indicate class labels, and the images at the vertices are the learned decoder archetypes.

Table 17: CAA on SVHN classes $\{0, 1, 7\}$ with $K = 3$, $a = 1$, and the exact loss. We report the mean $\pm$ standard deviation computed from the five runs at each ratio.

$\beta/\hat{\beta}_c$	$\overline{\mathrm{KL}}_\phi$	$\|W_T\|_F$	U_m
0.1	1.8 ± 0.039	$(2.5 \pm 0.0013) \times 10$	$(1.6 \pm 0.0043) \times 10^{-1}$
0.2	1.1 ± 0.0089	$(2.1 \pm 0.0015) \times 10$	$(1.4 \pm 0.015) \times 10^{-1}$
0.5	$(5.2 \pm 0.13) \times 10^{-1}$	$(1.8 \pm 0.003) \times 10$	$(7.2 \pm 0.18) \times 10^{-2}$
0.8	$(2.3 \pm 0.06) \times 10^{-1}$	$(1.4 \pm 0.0018) \times 10$	$(3.3 \pm 0.09) \times 10^{-2}$
0.9	$(1.5 \pm 0.087) \times 10^{-1}$	$(1.1 \pm 0.0022) \times 10$	$(2.1 \pm 0.14) \times 10^{-2}$
0.95	$(1.1 \pm 0.096) \times 10^{-1}$	9.9 ± 0.02	$(1.6 \pm 0.16) \times 10^{-2}$
1	$(2.8 \pm 6.2) \times 10^{-5}$	$(5.2 \pm 1.0) \times 10^{-4}$	$(6.7 \pm 8.4) \times 10^{-15}$
1.05	$(8.8 \pm 10.0) \times 10^{-5}$	$(7.0 \pm 1.6) \times 10^{-4}$	$(2.7 \pm 2.3) \times 10^{-15}$
1.2	$(2.3 \pm 4.7) \times 10^{-6}$	$(5.0 \pm 1.2) \times 10^{-4}$	$(2.4 \pm 2.5) \times 10^{-15}$
1.5	$(4.4 \pm 9.2) \times 10^{-5}$	$(7.9 \pm 4.1) \times 10^{-4}$	$(2.8 \pm 1.5) \times 10^{-15}$
2	$(8.9 \pm 12.0) \times 10^{-5}$	$(5.0 \pm 2.0) \times 10^{-4}$	$(2.9 \pm 1.6) \times 10^{-15}$

E.3 CAA on FashionMNIST

Model architecture. We use the same encoder and linear decoder as for MNIST in App. E.1, with $K = 3$, $a = 1$, and two hidden layers of width 256 with `tanh` activations.

Training protocol. We retain classes $\{1, 3, 8\}$, corresponding to trousers, dresses, and bags, giving 18,000 training images. The data seed is 0. The estimation gives $\hat{\sigma} = 0.20377$, $\lambda_{\max}(\widehat{\Sigma}) = 20.8967$, and $\hat{\beta}_c = 407.92$. We minimize the exact negative β-ELBO for 3000 epochs with Adam, learning rate 3×10^{-3}, batch size 256, no weight decay, and seeds $\{0, 1, 2, 3, 4\}$. We use Xavier-uniform initialization, as for MNIST. The preprocessing, evaluation subset size, and normalized sweep grid are given at the start of this appendix.

Results. Figure 15 gives a representative grid sweep on FashionMNIST and Table 18 reports the eleven ratios up to $\beta/\hat{\beta}_c = 2$. The simplex panels mirror those obtained on MNIST, and the collapse metrics also almost vanish at the predicted threshold. The archetypes are interpretable as the three classes, and the simplex occupancy contracts toward the barycenter as $\beta/\hat{\beta}_c$ approaches 1.

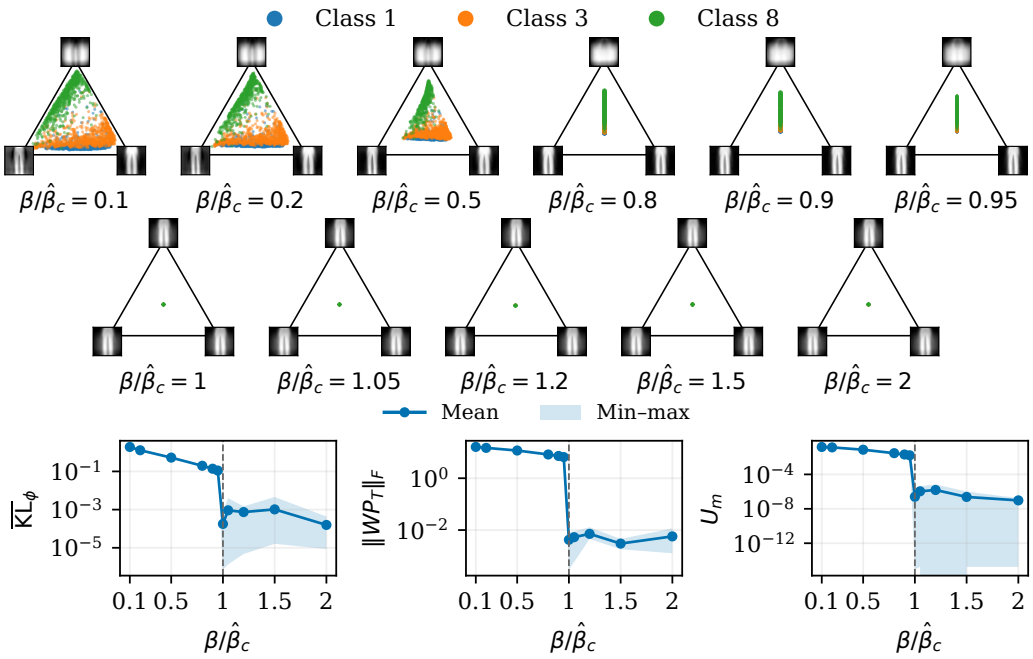


Figure 15: **Convex archetypal analysis on FashionMNIST classes** $1, 3, 8$**.** Top two rows: posterior mean compositions for all eleven ratios in Table 18, with $K = 3$ and seed 0, and learned archetypes at the vertices. Bottom: collapse diagnostics. Curves show means over five seeds for the average encoder KL, tangent decoder norm $\|W_T\|_F$, and posterior-mean usage U_m. Shaded bands show the minimum-maximum range across seeds. The dashed vertical line marks the predicted local collapse threshold.

Table 18: CAA on FashionMNIST classes $\{1, 3, 8\}$ with $K = 3$, $a = 1$, and the exact loss. Entries give mean $\pm$ sample standard deviation over the five runs at each ratio, rounded for display.

$\beta/\hat{\beta}_c$	$\overline{\mathrm{KL}}_\phi$	$\|W_T\|_F$	U_m
0.1	$1.9 \pm 8.7 \times 10^{-3}$	$1.6 \times 10 \pm 1.4 \times 10^{-2}$	$1.5 \times 10^{-1} \pm 1.7 \times 10^{-3}$
0.2	$1.3 \pm 4.7 \times 10^{-3}$	$1.5 \times 10 \pm 2.8 \times 10^{-2}$	$1.4 \times 10^{-1} \pm 9.2 \times 10^{-4}$
0.5	$5.3 \times 10^{-1} \pm 5.8 \times 10^{-3}$	$1.2 \times 10 \pm 4.4 \times 10^{-2}$	$7.4 \times 10^{-2} \pm 1.2 \times 10^{-3}$
0.8	$2.0 \times 10^{-1} \pm 2.0 \times 10^{-3}$	$8.2 \pm 3.5 \times 10^{-2}$	$2.9 \times 10^{-2} \pm 3.4 \times 10^{-4}$
0.9	$1.4 \times 10^{-1} \pm 1.3 \times 10^{-3}$	$7.2 \pm 2.7 \times 10^{-2}$	$2.1 \times 10^{-2} \pm 2.0 \times 10^{-4}$
0.95	$1.1 \times 10^{-1} \pm 4.1 \times 10^{-3}$	$6.5 \pm 3.9 \times 10^{-2}$	$1.6 \times 10^{-2} \pm 6.4 \times 10^{-4}$
1	$(1.8 \pm 3.7) \times 10^{-4}$	$(4.2 \pm 3.1) \times 10^{-3}$	$(2.5 \pm 5.6) \times 10^{-7}$
1.05	$9.2 \times 10^{-4} \pm 1.73 \times 10^{-3}$	$(5.3 \pm 3.4) \times 10^{-3}$	$(1.1 \pm 1.8) \times 10^{-6}$
1.2	$(7.4 \pm 6.5) \times 10^{-4}$	$(7.1 \pm 3.3) \times 10^{-3}$	$(1.5 \pm 3.0) \times 10^{-6}$
1.5	$(1.0 \pm 2.0) \times 10^{-3}$	$(3.0 \pm 1.4) \times 10^{-3}$	$(2.4 \pm 4.4) \times 10^{-7}$
2	$(1.6 \pm 1.7) \times 10^{-4}$	$(5.7 \pm 3.8) \times 10^{-3}$	$(9.5 \pm 6.9) \times 10^{-8}$

F Nonlinear decoder experiments

We further test the robustness of the collapse threshold on nonlinear decoders. We report results on the Samson (Zhu, 2017) hyperspectral dataset and MNIST (LeCun et al., 2010).

Model architecture. All models use $K = 3$ and the symmetric Dirichlet prior with $a = 1$. The encoder has two hidden layers of width 256 with `tanh` activations, followed by a linear output layer and $\gamma_\phi(y) = \mathrm{softplus}(f_\phi(y)) + \epsilon$ with $\epsilon = 10^{-4}$. The decoder has one hidden layer,

$$h(z) = \tanh(W_1 z + b_1), \qquad f_\theta(z) = \begin{cases} 10^{-8} + \mathrm{softplus}(W_2 h(z)), & \text{Samson,} \\ W_2 h(z) + b_2, & \text{MNIST.} \end{cases}$$

Its hidden width is 128 for Samson and 256 for MNIST.

Experimental protocol. Samson uses all 9025 normalized spectra, with the preprocessing described in Section D. For MNIST, we reuse the dataset in Section E. We keep the same estimated σ^2 with the same method to estimate $\hat{\beta}_{\mathrm{c}}$. The KL is computed analytically, while the reconstruction expectation uses 4 posterior samples per observation during training and 32 during evaluation. We train for 3000 epochs with Adam, learning rate 3×10^{-3}, batch size 256, no weight decay, and seeds $\{0, 1, 2, 3, 4\}$. All reported runs use Xavier-uniform initialization with zero biases. The `tanh` layers use the gain returned by `calculate_gain`, and the output layers use the default gain 1. The sweep is

$$\frac{\beta}{\hat{\beta}_c} \in \{0.5, 0.8, 0.9, 0.95, 1.0, 1.05, 1.2, 1.5, 2.0\}.$$

In addition to the average posterior KL and U_m, we use the decoder's local linear approximation, given by its Jacobian J_{f_θ} where f_θ is the decoder function, to quantify how its output changes along directions tangent to the simplex, thereby quantifying the decoder's sensitivity to latent variables. We then use $M = 512$ fixed samples $z_j \sim \mathrm{Dir}(\mathbf{1}_K)$ and compute

$$\mathcal{J}_T = \left(\frac{1}{M} \sum_{j=1}^{M} \| J_{f_\theta}(z_j) P_T \|_F^2 \right)^{1/2}.$$

This reduces to $\|W P_T\|_F$ for an affine decoder. We also track how much the learned decoder function f_θ differs from an affine function through

$$R_{\mathrm{aff}} = \frac{\frac{1}{M} \sum_{j=1}^{M} \left\| f_\theta(z_j) - \sum_{k=1}^{K} z_{jk} f_\theta(e_k) \right\|^2}{\frac{1}{M} \sum_{j=1}^{M} \| f_\theta(z_j) - \bar{f} \|^2 + 10^{-12}}, \qquad \bar{f} = \frac{1}{M} \sum_{j=1}^{M} f_\theta(z_j),$$

We normalize by variance to compare the affine interpolation error relative to the variations produced by the decoder across latent compositions. This ratio is reported only when the output variance exceeds 10^{-10}. It is not constrained to $[0, 1]$ and a small value indicates approximately affine behavior on the sampled points.

Results. Figures 16 and 18 report means and minimum-maximum bands across five seeds. The corresponding tables give mean $\pm$ standard deviation.

Samson. Overall, all the metrics display a sharp drop around the threshold, placing the collapse transition close to the one predicted by the linear analysis. Their joint decrease shows that the trained model approaches collapse. Although the decoder seems to retain a small sensitivity to latent compositions, $\mathcal{J}_T \simeq 10^{-2}$, the small R_{aff} above the transition indicates that its remaining output variations are well captured by its affine approximation. This suggests that the model approaches the configuration considered in our local analysis, where we study small joint perturbations of the posterior and an affine decoder around collapse.

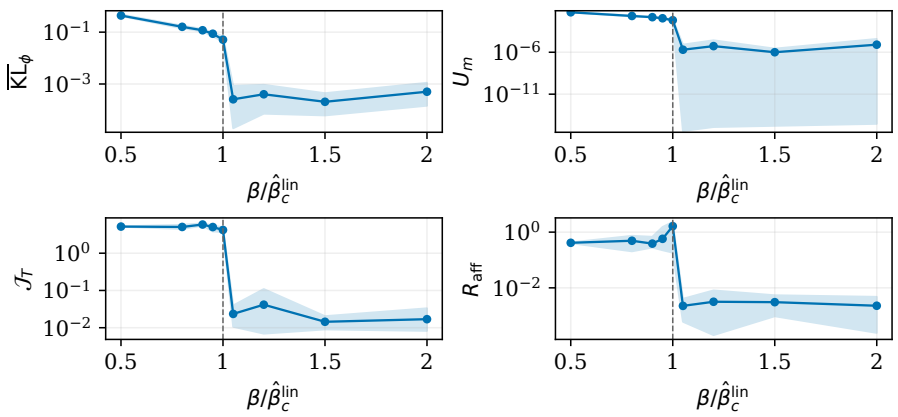


Figure 16: Nonlinear decoder on Samson with Xavier-uniform initialization and $a = 1$. Lines show means across five seeds and bands show the minimum-maximum range. The dashed line marks the linear reference scale.

Table 19: Nonlinear decoder on Samson, Xavier-uniform initialization only. Values give mean $\pm$ sample standard deviation across five seeds, computed from the runs and rounded for display.

$\beta/\hat{\beta}_c^{\mathrm{lin}}$	$\overline{\mathrm{KL}_\phi}$	U_m	$\mathcal{J}_T$	R_{aff}
0.5	$(4.4 \pm 0.24) \times 10^{-1}$	$(5.8 \pm 0.3) \times 10^{-2}$	5.2 ± 0.047	$(4.1 \pm 0.088) \times 10^{-1}$
0.8	$(1.6 \pm 0.14) \times 10^{-1}$	$(2.1 \pm 0.21) \times 10^{-2}$	5.1 ± 0.55	$(4.9 \pm 2.5) \times 10^{-1}$
0.9	$(1.2 \pm 0.09) \times 10^{-1}$	$(1.5 \pm 0.094) \times 10^{-2}$	5.9 ± 0.38	$(3.9 \pm 1.7) \times 10^{-1}$
0.95	$(8.7 \pm 1.2) \times 10^{-2}$	$(1.1 \pm 0.16) \times 10^{-2}$	5.0 ± 0.7	$(5.8 \pm 5.4) \times 10^{-1}$
1	$(5.1 \pm 1.0) \times 10^{-2}$	$(6.6 \pm 1.4) \times 10^{-3}$	4.2 ± 0.45	1.6 ± 0.87
1.05	$(2.6 \pm 3.4) \times 10^{-4}$	$(2.0 \pm 3.4) \times 10^{-6}$	$(2.3 \pm 1.1) \times 10^{-2}$	$(2.3 \pm 1.7) \times 10^{-3}$
1.2	$(4.0 \pm 3.4) \times 10^{-4}$	$(5.4 \pm 12.0) \times 10^{-6}$	$(4.2 \pm 4.5) \times 10^{-2}$	$(3.2 \pm 3.4) \times 10^{-3}$
1.5	$(2.1 \pm 1.6) \times 10^{-4}$	$(1.0 \pm 1.4) \times 10^{-6}$	$(1.4 \pm 0.43) \times 10^{-2}$	$(3.1 \pm 2.0) \times 10^{-3}$
2	$(5.0 \pm 3.8) \times 10^{-4}$	$(8.0 \pm 17.0) \times 10^{-6}$	$(1.7 \pm 1.1) \times 10^{-2}$	$(2.3 \pm 1.8) \times 10^{-3}$

MNIST. The same interval, between 1 and 1.05, contains a sharp decrease in the metrics. At larger ratios, the interpolation residual becomes very small, while the tangent sensitivity remains measurable which seems consistent with the results on Samson. Visually, the decoded images at the simplex vertices seem to be more recognizable than those produced by the linear decoder, which is consistent with the fact that the nonlinear decoder has more capacity to fit the data. Overall, these results suggest that the collapse threshold is robust to the choice of decoder architecture.

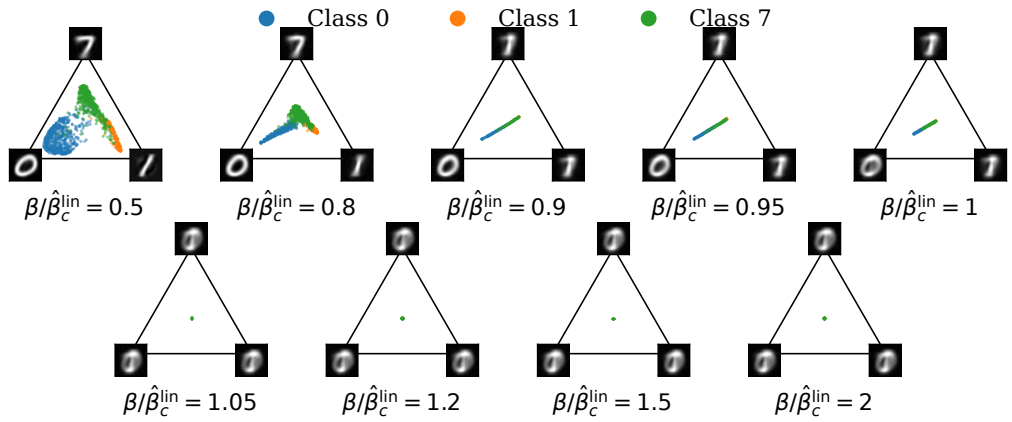


Figure 17: Posterior mean compositions on MNIST across the full sweep, with a nonlinear decoder, Xavier-uniform initialization, $a = 1$, and seed 0. Each point represents one evaluation image and is colored by its digit class. Small panels show the decoded images $f_\theta(e_k)$ at the simplex vertices, with contrast adjusted separately for display.

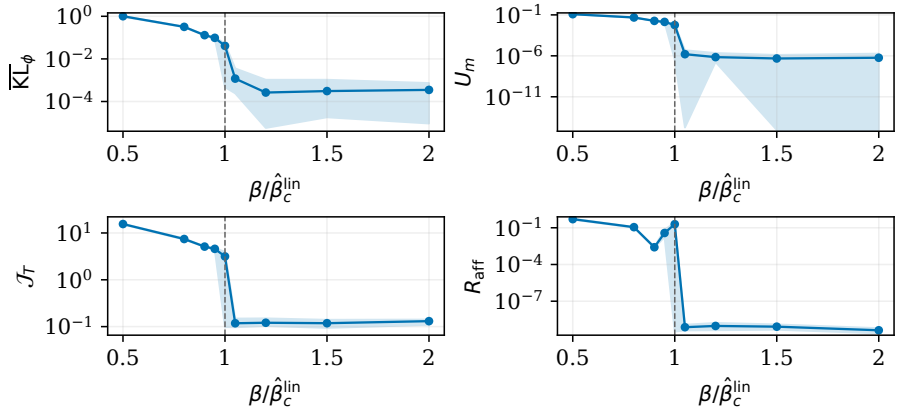


Figure 18: Nonlinear decoder on MNIST with Xavier-uniform initialization and $a = 1$. Lines show means across five seeds and bands show the minimum-maximum range. The dashed line marks the linear reference scale.

Table 20: Nonlinear decoder on MNIST, Xavier-uniform initialization only. Values give mean $\pm$ sample standard deviation across five seeds, computed from the runs and rounded for display.

$\beta/\hat{\beta}_c^{\mathrm{lin}}$	$\overline{\mathrm{KL}}_\phi$	U_m	$\mathcal{J}_T$	R_{aff}
0.5	1.0 ± 0.032	$(1.2 \pm 0.021) \times 10^{-1}$	$(1.6 \pm 0.025) \times 10$	$(5.1 \pm 0.36) \times 10^{-1}$
0.8	$(3.2 \pm 0.094) \times 10^{-1}$	$(4.8 \pm 0.11) \times 10^{-2}$	7.4 ± 0.17	$(1.1 \pm 0.071) \times 10^{-1}$
0.9	$(1.3 \pm 0.057) \times 10^{-1}$	$(1.8 \pm 0.026) \times 10^{-2}$	5.1 ± 0.05	$(2.6 \pm 0.27) \times 10^{-3}$
0.95	$(9.7 \pm 0.53) \times 10^{-2}$	$(1.4 \pm 0.081) \times 10^{-2}$	4.6 ± 0.13	$(3.7 \pm 2.6) \times 10^{-2}$
1	$(4.1 \pm 2.3) \times 10^{-2}$	$(5.7 \pm 3.2) \times 10^{-3}$	3.2 ± 1.7	$(2.0 \pm 1.1) \times 10^{-1}$
1.05	$(1.2 \pm 1.4) \times 10^{-3}$	$(1.6 \pm 2.2) \times 10^{-6}$	$(1.2 \pm 0.21) \times 10^{-1}$	$(7.5 \pm 2.8) \times 10^{-10}$
1.2	$(2.6 \pm 4.5) \times 10^{-4}$	$(6.9 \pm 8.4) \times 10^{-7}$	$(1.2 \pm 0.21) \times 10^{-1}$	$(9.5 \pm 5.0) \times 10^{-10}$
1.5	$(3.1 \pm 4.3) \times 10^{-4}$	$(4.7 \pm 4.9) \times 10^{-7}$	$(1.2 \pm 0.18) \times 10^{-1}$	$(8.3 \pm 3.4) \times 10^{-10}$
2	$(3.5 \pm 2.9) \times 10^{-4}$	$(5.8 \pm 7.9) \times 10^{-7}$	$(1.3 \pm 0.13) \times 10^{-1}$	$(4.2 \pm 1.7) \times 10^{-10}$

G Encoder architecture ablation

We perform an ablation study on the encoder architectures to examine how the collapse threshold depends on encoder capacity while retaining a linear decoder. The complementary experiments below use Samson (Zhu, 2017) and MNIST (LeCun et al., 2010). All results in this section use Xavier-uniform initialization.

Model architecture. Table 21 gives the four encoder architectures. Each hidden layer uses a `tanh` activation. In every case, the final encoder outputs are transformed into positive concentrations by $\gamma_\phi(y) = \mathrm{softplus}(f_\phi(y)) + \epsilon$ with $\epsilon = 10^{-4}$. Note that the architecture without a hidden layer therefore still has this nonlinear output transformation. All models use $K = 3$ and $a = 1$.

Table 21: Encoder architectures in the ablation. The listed widths describe the affine layers, including the final concentration-logit layer.

Architecture	Hidden layers	Layer dimensions
No hidden layer	0	$D \to 3$
Reference	2	$D \to 256 \to 256 \to 3$
Deep	3	$D \to 256 \to 256 \to 256 \to 3$
Wide	2	$D \to 1024 \to 1024 \to 3$

For Samson, the learned mixing matrix W is forced to be nonnegative, with either $W = 10^{-6} + \mathrm{softplus}(V)$ or clamping after each update. The decoder has no bias. For MNIST, the decoder is affine, $Wz + b$. Both the encoder and decoder are trained.

Experimental protocol. We reuse the same Samson dataset and MNIST dataset as in Appendix D and Appendix E, respectively. The same estimated σ^2 is used to compute $\hat{\beta}_c$. We minimize the negative β-ELBO with closed-form reconstruction for 3000 epochs using Adam, batch size 256, and no weight decay. Samson uses seeds $\{0, 1, 2, 3, 4\}$, and all reported MNIST ratios use seeds $\{0, 1, 2, 3, 4\}$. The learning rate is 3×10^{-3} except for the wide MNIST model, which uses 3×10^{-4} to ensure training stability. On Samson and MNIST, we sweep

$$\frac{\beta}{\hat{\beta}_c} \in \{0.1, 0.2, 0.5, 0.8, 0.9, 0.95, 1.0, 1.1, 1.2, 1.5, 2.0\}.$$

Results. The figures report means and minimum-maximum bands for the metrics (KL, WP_T and U_m). The tables give mean $\pm$ standard deviation. Overall, the results show that the collapse threshold is robust to the encoder architecture, with a sharp transition in the metrics occurring near $\hat{\beta}_c$ for all architectures.

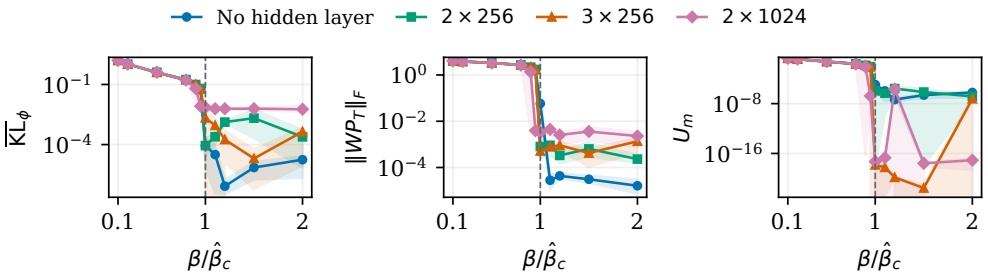


Figure 19: Encoder architecture ablation on Samson with the `softplus` nonnegative mixing matrix. All runs use Xavier-uniform initialization and $a = 1$. Lines show means and bands show the minimum-maximum range. Horizontal axes use a linear scale and vertical axes use a logarithmic scale.

Table 22: Encoder architecture ablation on Samson with the `softplus` nonnegative mixing matrix. Entries give mean $\pm$ sample standard deviation of $\overline{\mathrm{KL}}_\phi$, $\|W_T\|_F$, and U_m, rounded for display. Five seeds are used.

$\beta/\hat{\beta}_c$	$\overline{\mathrm{KL}}_\phi$	$\|W_T\|_F$	U_m
No hidden layer			
0.1	1.6 ± 0.0055	4.1 ± 0.0031	$(1.8 \pm 0.017) \times 10^{-1}$
0.2	1.0 ± 0.0032	3.9 ± 0.019	$(1.3 \pm 0.0057) \times 10^{-1}$
0.5	$(3.9 \pm 0.054) \times 10^{-1}$	3.3 ± 0.0046	$(5.3 \pm 0.069) \times 10^{-2}$
0.8	$(1.6 \pm 0.035) \times 10^{-1}$	2.7 ± 0.0008	$(2.2 \pm 0.046) \times 10^{-2}$
0.9	$(9.2 \pm 0.22) \times 10^{-2}$	2.1 ± 0.0095	$(1.3 \pm 0.029) \times 10^{-2}$
0.95	$(5.5 \pm 0.12) \times 10^{-2}$	1.6 ± 0.0046	$(7.5 \pm 0.16) \times 10^{-3}$
1	$(8.8 \pm 11.0) \times 10^{-5}$	$(5.8 \pm 4.9) \times 10^{-2}$	$(1.1 \pm 1.4) \times 10^{-5}$
1.1	$(3.2 \pm 5.3) \times 10^{-5}$	$(2.8 \pm 2.0) \times 10^{-5}$	$(1.1 \pm 1.8) \times 10^{-6}$
1.2	$(8.4 \pm 7.8) \times 10^{-7}$	$(4.4 \pm 1.2) \times 10^{-5}$	$(4.4 \pm 3.2) \times 10^{-8}$
1.5	$(7.1 \pm 6.0) \times 10^{-6}$	$(3.1 \pm 1.2) \times 10^{-5}$	$(2.4 \pm 1.5) \times 10^{-7}$
2	$(1.8 \pm 1.3) \times 10^{-5}$	$(1.6 \pm 1.2) \times 10^{-5}$	$(6.1 \pm 4.6) \times 10^{-7}$
Reference (2×256)			
0.1	1.6 ± 0.0038	4.0 ± 0.0037	$(1.9 \pm 0.0055) \times 10^{-1}$
0.2	1.0 ± 0.0011	3.8 ± 0.012	$(1.4 \pm 0.0015) \times 10^{-1}$
0.5	$(4.0 \pm 0.023) \times 10^{-1}$	3.3 ± 0.0063	$(5.7 \pm 0.029) \times 10^{-2}$
0.8	$(1.8 \pm 0.053) \times 10^{-1}$	2.7 ± 0.0034	$(2.5 \pm 0.084) \times 10^{-2}$
0.9	$(1.0 \pm 0.019) \times 10^{-1}$	2.2 ± 0.0096	$(1.4 \pm 0.029) \times 10^{-2}$
0.95	$(6.2 \pm 0.23) \times 10^{-2}$	1.7 ± 0.0062	$(8.6 \pm 0.26) \times 10^{-3}$
1	$(8.9 \pm 9.9) \times 10^{-5}$	$(8.0 \pm 7.5) \times 10^{-4}$	$(1.1 \pm 1.9) \times 10^{-6}$
1.1	$(2.5 \pm 1.5) \times 10^{-4}$	$(9.1 \pm 5.0) \times 10^{-4}$	$(4.7 \pm 4.6) \times 10^{-7}$
1.2	$(1.3 \pm 0.14) \times 10^{-3}$	$(3.4 \pm 2.2) \times 10^{-4}$	$(2.6 \pm 2.9) \times 10^{-6}$
1.5	$(2.1 \pm 3.1) \times 10^{-3}$	$(6.4 \pm 4.8) \times 10^{-4}$	$(7.2 \pm 12.4) \times 10^{-7}$
2	$(2.4 \pm 2.8) \times 10^{-4}$	$(2.3 \pm 1.1) \times 10^{-4}$	$(1.5 \pm 1.2) \times 10^{-7}$
Deep (3×256)			
0.1	1.6 ± 0.0059	4.0 ± 0.0037	$(1.9 \pm 0.0069) \times 10^{-1}$
0.2	1.0 ± 0.0053	3.8 ± 0.012	$(1.4 \pm 0.011) \times 10^{-1}$
0.5	$(4.0 \pm 0.06) \times 10^{-1}$	3.3 ± 0.0062	$(5.7 \pm 0.11) \times 10^{-2}$
0.8	$(1.7 \pm 0.017) \times 10^{-1}$	2.7 ± 0.0034	$(2.4 \pm 0.028) \times 10^{-2}$
0.9	$(1.0 \pm 0.0067) \times 10^{-1}$	2.2 ± 0.01	$(1.4 \pm 0.016) \times 10^{-2}$
0.95	$(6.1 \pm 0.17) \times 10^{-2}$	1.7 ± 0.0046	$(8.4 \pm 0.094) \times 10^{-3}$
1	$(2.1 \pm 3.6) \times 10^{-3}$	$(5.2 \pm 0.77) \times 10^{-4}$	$(1.5 \pm 2.5) \times 10^{-18}$
1.1	$(9.0 \pm 9.6) \times 10^{-4}$	$(7.8 \pm 1.1) \times 10^{-4}$	$(5.8 \pm 10.0) \times 10^{-19}$
1.2	$(1.8 \pm 2.8) \times 10^{-4}$	$(9.1 \pm 3.7) \times 10^{-4}$	$(1.5 \pm 0.2) \times 10^{-20}$
1.5	$(2.0 \pm 3.2) \times 10^{-5}$	$(4.2 \pm 2.8) \times 10^{-4}$	$(2.8 \pm 4.9) \times 10^{-22}$
2	$(4.4 \pm 3.8) \times 10^{-4}$	$(1.4 \pm 0.82) \times 10^{-3}$	$(6.0 \pm 10.3) \times 10^{-8}$
Wide (2×1024)			
0.1	1.6 ± 0.0091	4.0 ± 0.0039	$(1.9 \pm 0.018) \times 10^{-1}$
0.2	1.0 ± 0.0026	3.8 ± 0.0056	$(1.4 \pm 0.0063) \times 10^{-1}$
0.5	$(4.1 \pm 0.067) \times 10^{-1}$	3.3 ± 0.006	$(5.8 \pm 0.038) \times 10^{-2}$
0.8	$(1.7 \pm 0.1) \times 10^{-1}$	2.7 ± 0.064	$(2.3 \pm 0.15) \times 10^{-2}$
0.9	$(6.3 \pm 4.5) \times 10^{-2}$	1.4 ± 1.2	$(8.2 \pm 7.1) \times 10^{-3}$
0.95	$(8.3 \pm 2.8) \times 10^{-3}$	$(3.9 \pm 1.7) \times 10^{-3}$	$(1.7 \pm 2.2) \times 10^{-7}$
1	$(7.2 \pm 2.2) \times 10^{-3}$	$(3.3 \pm 1.8) \times 10^{-3}$	$(4.9 \pm 1.8) \times 10^{-18}$
1.1	$(6.2 \pm 1.4) \times 10^{-3}$	$(4.4 \pm 1.7) \times 10^{-3}$	$(2.1 \pm 1.8) \times 10^{-17}$
1.2	$(6.1 \pm 0.49) \times 10^{-3}$	$(2.6 \pm 1.3) \times 10^{-3}$	$(2.2 \pm 3.8) \times 10^{-6}$
1.5	$(6.2 \pm 1.1) \times 10^{-3}$	$(3.6 \pm 2.0) \times 10^{-3}$	$(2.6 \pm 1.5) \times 10^{-18}$
2	$(5.8 \pm 2.3) \times 10^{-3}$	$(2.3 \pm 1.4) \times 10^{-3}$	$(8.2 \pm 8.6) \times 10^{-18}$

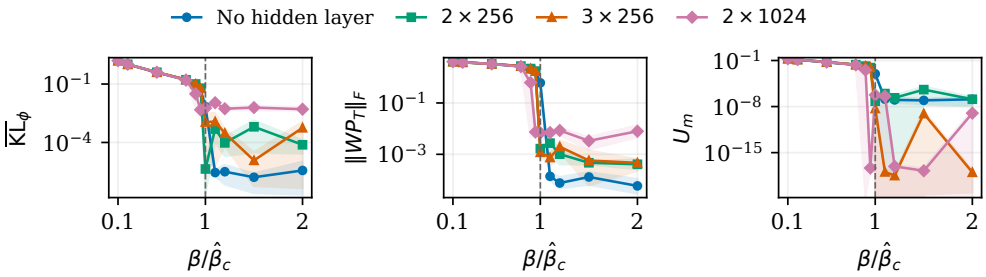


Figure 20: Encoder architecture ablation on Samson with the `clamp` nonnegative mixing matrix. All runs use Xavier-uniform initialization and $a = 1$. Lines show means and bands show the minimum-maximum range. Horizontal axes use a linear scale and vertical axes use a logarithmic scale.

Table 23: Encoder architecture ablation on Samson with the `clamp` nonnegative mixing matrix. Entries give mean ± standard deviation of $\overline{\mathrm{KL}}_\phi$, $\|W_T\|_F$, and U_m. Five seeds are used.

$\beta/\hat{\beta}_c$	$\overline{\mathrm{KL}}_\phi$	$\|W_T\|_F$	U_m
No hidden layer			
0.1	1.6 ± 0.0056	4.1 ± 0.019	$(1.8 \pm 0.015) \times 10^{-1}$
0.2	1.0 ± 0.0044	3.9 ± 0.0036	$(1.3 \pm 0.012) \times 10^{-1}$
0.5	$(3.9 \pm 0.031) \times 10^{-1}$	3.3 ± 0.0096	$(5.3 \pm 0.044) \times 10^{-2}$
0.8	$(1.7 \pm 0.038) \times 10^{-1}$	2.7 ± 0.0065	$(2.3 \pm 0.046) \times 10^{-2}$
0.9	$(9.2 \pm 0.11) \times 10^{-2}$	2.1 ± 0.0077	$(1.3 \pm 0.01) \times 10^{-2}$
0.95	$(5.3 \pm 0.05) \times 10^{-2}$	1.7 ± 0.0069	$(7.3 \pm 0.078) \times 10^{-3}$
1	$(6.9 \pm 0.19) \times 10^{-3}$	$(6.1 \pm 0.076) \times 10^{-1}$	$(9.4 \pm 0.26) \times 10^{-4}$
1.1	$(2.8 \pm 1.1) \times 10^{-6}$	$(1.4 \pm 0.49) \times 10^{-4}$	$(1.3 \pm 0.52) \times 10^{-7}$
1.2	$(3.1 \pm 3.4) \times 10^{-6}$	$(7.6 \pm 4.8) \times 10^{-5}$	$(1.0 \pm 0.85) \times 10^{-7}$
1.5	$(1.6 \pm 1.2) \times 10^{-6}$	$(1.3 \pm 0.68) \times 10^{-4}$	$(8.8 \pm 6.8) \times 10^{-8}$
2	$(3.6 \pm 5.8) \times 10^{-6}$	$(5.9 \pm 4.5) \times 10^{-5}$	$(1.3 \pm 1.8) \times 10^{-7}$
Reference (2×256)			
0.1	1.6 ± 0.0064	4.0 ± 0.013	$(1.9 \pm 0.0099) \times 10^{-1}$
0.2	1.0 ± 0.0057	3.8 ± 0.0037	$(1.4 \pm 0.0067) \times 10^{-1}$
0.5	$(4.0 \pm 0.023) \times 10^{-1}$	3.3 ± 0.012	$(5.7 \pm 0.03) \times 10^{-2}$
0.8	$(1.7 \pm 0.067) \times 10^{-1}$	2.7 ± 0.0096	$(2.3 \pm 0.11) \times 10^{-2}$
0.9	$(1.0 \pm 0.012) \times 10^{-1}$	2.2 ± 0.005	$(1.4 \pm 0.019) \times 10^{-2}$
0.95	$(6.1 \pm 0.16) \times 10^{-2}$	1.7 ± 0.0068	$(8.5 \pm 0.23) \times 10^{-3}$
1	$(4.3 \pm 4.7) \times 10^{-6}$	$(1.7 \pm 0.14) \times 10^{-3}$	$(6.5 \pm 9.0) \times 10^{-8}$
1.1	$(4.9 \pm 5.0) \times 10^{-4}$	$(2.7 \pm 1.5) \times 10^{-3}$	$(9.5 \pm 15.4) \times 10^{-7}$
1.2	$(9.5 \pm 9.6) \times 10^{-5}$	$(1.0 \pm 0.61) \times 10^{-3}$	$(2.1 \pm 2.7) \times 10^{-7}$
1.5	$(6.4 \pm 6.7) \times 10^{-4}$	$(4.7 \pm 0.59) \times 10^{-4}$	$(3.8 \pm 6.5) \times 10^{-6}$
2	$(7.6 \pm 5.2) \times 10^{-5}$	$(4.1 \pm 2.0) \times 10^{-4}$	$(1.3 \pm 1.3) \times 10^{-7}$
Deep (3×256)			
0.1	1.6 ± 0.0016	4.0 ± 0.014	$(1.9 \pm 0.0048) \times 10^{-1}$
0.2	1.0 ± 0.02	3.8 ± 0.0036	$(1.4 \pm 0.029) \times 10^{-1}$
0.5	$(4.1 \pm 0.017) \times 10^{-1}$	3.3 ± 0.01	$(5.7 \pm 0.024) \times 10^{-2}$
0.8	$(1.7 \pm 0.02) \times 10^{-1}$	2.7 ± 0.0082	$(2.4 \pm 0.029) \times 10^{-2}$
0.9	$(1.0 \pm 0.0049) \times 10^{-1}$	2.2 ± 0.0047	$(1.4 \pm 0.0079) \times 10^{-2}$
0.95	$(6.1 \pm 0.033) \times 10^{-2}$	1.7 ± 0.0036	$(8.5 \pm 0.088) \times 10^{-3}$
1	$(1.1 \pm 1.9) \times 10^{-3}$	$(1.2 \pm 0.46) \times 10^{-3}$	$(5.1 \pm 8.8) \times 10^{-9}$
1.1	$(1.2 \pm 1.8) \times 10^{-3}$	$(7.5 \pm 3.5) \times 10^{-4}$	$(1.2 \pm 2.0) \times 10^{-18}$
1.2	$(3.0 \pm 2.2) \times 10^{-4}$	$(2.0 \pm 1.2) \times 10^{-3}$	$(3.3 \pm 5.2) \times 10^{-19}$
1.5	$(1.2 \pm 1.7) \times 10^{-5}$	$(5.7 \pm 7.2) \times 10^{-4}$	$(8.5 \pm 14.7) \times 10^{-10}$
2	$(5.7 \pm 5.8) \times 10^{-4}$	$(4.6 \pm 2.2) \times 10^{-4}$	$(1.0 \pm 1.1) \times 10^{-18}$
Wide (2×1024)			
0.1	1.6 ± 0.02	4.0 ± 0.01	$(1.9 \pm 0.042) \times 10^{-1}$
0.2	1.0 ± 0.016	3.8 ± 0.0083	$(1.4 \pm 0.028) \times 10^{-1}$
0.5	$(4.1 \pm 0.045) \times 10^{-1}$	3.3 ± 0.015	$(5.6 \pm 0.12) \times 10^{-2}$
0.8	$(1.6 \pm 0.15) \times 10^{-1}$	2.6 ± 0.11	$(2.2 \pm 0.17) \times 10^{-2}$
0.9	$(3.1 \pm 4.1) \times 10^{-2}$	$(6.2 \pm 10.7) \times 10^{-1}$	$(3.4 \pm 6.0) \times 10^{-3}$
0.95	$(4.7 \pm 0.5) \times 10^{-3}$	$(7.4 \pm 2.9) \times 10^{-3}$	$(4.6 \pm 2.0) \times 10^{-18}$
1	$(5.7 \pm 0.73) \times 10^{-3}$	$(6.8 \pm 3.5) \times 10^{-3}$	$(5.8 \pm 6.6) \times 10^{-7}$
1.1	$(1.1 \pm 0.36) \times 10^{-2}$	$(7.0 \pm 1.9) \times 10^{-3}$	$(3.7 \pm 6.5) \times 10^{-7}$
1.2	$(5.4 \pm 1.2) \times 10^{-3}$	$(8.5 \pm 2.6) \times 10^{-3}$	$(8.1 \pm 5.3) \times 10^{-18}$
1.5	$(6.2 \pm 1.6) \times 10^{-3}$	$(3.4 \pm 1.6) \times 10^{-3}$	$(1.7 \pm 2.3) \times 10^{-18}$
2	$(5.1 \pm 1.2) \times 10^{-3}$	$(7.9 \pm 4.8) \times 10^{-3}$	$(1.0 \pm 1.7) \times 10^{-9}$

MNIST. The reference and deep encoders show a sharp KL decrease between $\beta/\hat{\beta}_c = 0.95$ and 1, while the architecture without a hidden layer retains a larger KL at 1. The wide model also has a small KL at 1, under its different learning rate. The results indicate that encoder architecture and optimization affect the empirical transition around the common reference scale.

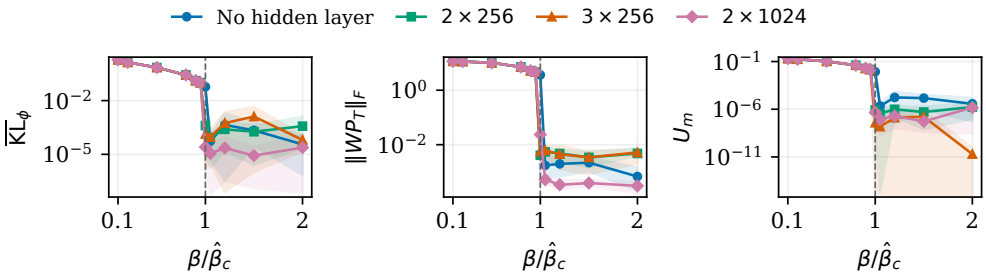


Figure 21: Encoder architecture ablation on MNIST. All runs use Xavier-uniform initialization and $a = 1$. Lines show means and bands show the minimum-maximum range. Horizontal axes use a linear scale and vertical axes use a logarithmic scale.

Table 24: Encoder architecture ablation on MNIST. Entries give mean $\pm$ standard deviation of $\overline{\mathrm{KL}}_\phi$, $\|W_T\|_F$, and U_m. Five seeds are used. The wide model has learning rate 3×10^{-4}, compared with 3×10^{-3} for the other architectures.

$\beta/\hat{\beta}_c$	$\overline{\mathrm{KL}}_\phi$	$\|W_T\|_F$	U_m
No hidden layer			
0.1	1.9 ± 0.007	$(1.2 \pm 0.00049) \times 10$	$(2.0 \pm 0.013) \times 10^{-1}$
0.2	1.4 ± 0.0024	$(1.1 \pm 0.00056) \times 10$	$(1.7 \pm 0.007) \times 10^{-1}$
0.5	$(7.1 \pm 0.022) \times 10^{-1}$	$(1.0 \pm 0.00069) \times 10$	$(1.0 \pm 0.0065) \times 10^{-1}$
0.8	$(2.6 \pm 0.042) \times 10^{-1}$	7.0 ± 0.02	$(3.9 \pm 0.074) \times 10^{-2}$
0.9	$(1.3 \pm 0.05) \times 10^{-1}$	5.1 ± 0.022	$(1.9 \pm 0.086) \times 10^{-2}$
0.95	$(9.5 \pm 0.17) \times 10^{-2}$	4.5 ± 0.018	$(1.4 \pm 0.028) \times 10^{-2}$
1	$(5.9 \pm 0.089) \times 10^{-2}$	3.6 ± 0.018	$(8.5 \pm 0.16) \times 10^{-3}$
1.05	$(5.4 \pm 6.0) \times 10^{-5}$	$(1.7 \pm 1.1) \times 10^{-3}$	$(2.2 \pm 1.8) \times 10^{-6}$
1.2	$(4.4 \pm 8.3) \times 10^{-4}$	$(2.0 \pm 1.1) \times 10^{-3}$	$(1.7 \pm 2.9) \times 10^{-5}$
1.5	$(2.2 \pm 2.6) \times 10^{-4}$	$(2.2 \pm 0.89) \times 10^{-3}$	$(1.5 \pm 1.3) \times 10^{-5}$
2	$(3.6 \pm 4.9) \times 10^{-5}$	$(6.8 \pm 4.8) \times 10^{-4}$	$(3.7 \pm 6.5) \times 10^{-6}$
Reference (2×256)			
0.1	1.9 ± 0.0052	$(1.2 \pm 0.00046) \times 10$	$(2.0 \pm 0.015) \times 10^{-1}$
0.2	1.4 ± 0.0025	$(1.1 \pm 0.00073) \times 10$	$(1.8 \pm 0.0085) \times 10^{-1}$
0.5	$(7.0 \pm 0.017) \times 10^{-1}$	$(1.0 \pm 0.00079) \times 10$	$(1.0 \pm 0.0065) \times 10^{-1}$
0.8	$(2.7 \pm 0.0097) \times 10^{-1}$	7.2 ± 0.021	$(4.1 \pm 0.029) \times 10^{-2}$
0.9	$(1.3 \pm 0.019) \times 10^{-1}$	5.2 ± 0.021	$(1.9 \pm 0.038) \times 10^{-2}$
0.95	$(1.0 \pm 0.02) \times 10^{-1}$	4.6 ± 0.019	$(1.5 \pm 0.017) \times 10^{-2}$
1	$(4.1 \pm 6.0) \times 10^{-4}$	$(4.1 \pm 1.4) \times 10^{-3}$	$(8.3 \pm 8.8) \times 10^{-7}$
1.05	$(8.7 \pm 5.4) \times 10^{-5}$	$(5.5 \pm 2.6) \times 10^{-3}$	$(3.4 \pm 5.7) \times 10^{-7}$
1.2	$(2.5 \pm 2.6) \times 10^{-4}$	$(4.8 \pm 1.5) \times 10^{-3}$	$(9.8 \pm 10.0) \times 10^{-7}$
1.5	$(1.8 \pm 2.0) \times 10^{-4}$	$(3.4 \pm 1.8) \times 10^{-3}$	$(5.1 \pm 7.7) \times 10^{-7}$
2	$(3.7 \pm 5.0) \times 10^{-4}$	$(4.8 \pm 1.4) \times 10^{-3}$	$(1.7 \pm 2.0) \times 10^{-6}$
Deep (3×256)			
0.1	1.9 ± 0.003	$(1.2 \pm 0.00048) \times 10$	$(2.1 \pm 0.0091) \times 10^{-1}$
0.2	1.4 ± 0.0058	$(1.1 \pm 0.00075) \times 10$	$(1.7 \pm 0.013) \times 10^{-1}$
0.5	$(7.1 \pm 0.11) \times 10^{-1}$	$(1.0 \pm 0.00076) \times 10$	$(1.1 \pm 0.023) \times 10^{-1}$
0.8	$(2.7 \pm 0.01) \times 10^{-1}$	7.2 ± 0.02	$(4.1 \pm 0.021) \times 10^{-2}$
0.9	$(1.3 \pm 0.018) \times 10^{-1}$	5.2 ± 0.022	$(1.9 \pm 0.033) \times 10^{-2}$
0.95	$(9.9 \pm 0.15) \times 10^{-2}$	4.6 ± 0.018	$(1.5 \pm 0.021) \times 10^{-2}$
1	$(1.4 \pm 2.8) \times 10^{-4}$	$(5.6 \pm 2.1) \times 10^{-3}$	$(3.8 \pm 5.2) \times 10^{-8}$
1.05	$(9.1 \pm 10.0) \times 10^{-5}$	$(6.2 \pm 2.0) \times 10^{-3}$	$(1.5 \pm 3.1) \times 10^{-8}$
1.2	$(5.6 \pm 9.6) \times 10^{-4}$	$(4.5 \pm 3.0) \times 10^{-3}$	$(1.3 \pm 2.3) \times 10^{-7}$
1.5	$(1.3 \pm 1.8) \times 10^{-3}$	$(3.4 \pm 1.8) \times 10^{-3}$	$(1.6 \pm 3.2) \times 10^{-7}$
2	$(6.4 \pm 4.2) \times 10^{-5}$	$(5.2 \pm 2.0) \times 10^{-3}$	$(2.0 \pm 4.4) \times 10^{-11}$
Wide (2×1024)			
0.1	1.9 ± 0.003	$(1.2 \pm 0.000095) \times 10$	$(2.0 \pm 0.0072) \times 10^{-1}$
0.2	1.4 ± 0.0018	$(1.1 \pm 0.00039) \times 10$	$(1.8 \pm 0.0062) \times 10^{-1}$
0.5	$(7.0 \pm 0.0069) \times 10^{-1}$	$(1.0 \pm 0.00058) \times 10$	$(1.1 \pm 0.00093) \times 10^{-1}$
0.8	$(2.7 \pm 0.014) \times 10^{-1}$	7.2 ± 0.0068	$(4.1 \pm 0.031) \times 10^{-2}$
0.9	$(1.3 \pm 0.0099) \times 10^{-1}$	5.2 ± 0.0094	$(1.9 \pm 0.017) \times 10^{-2}$
0.95	$(1.0 \pm 0.0066) \times 10^{-1}$	4.6 ± 0.0063	$(1.5 \pm 0.013) \times 10^{-2}$
1	$(2.6 \pm 3.9) \times 10^{-5}$	$(2.3 \pm 1.9) \times 10^{-2}$	$(4.7 \pm 4.4) \times 10^{-7}$
1.05	$(1.1 \pm 2.0) \times 10^{-5}$	$(5.2 \pm 2.0) \times 10^{-4}$	$(8.5 \pm 17.5) \times 10^{-8}$
1.2	$(2.3 \pm 3.9) \times 10^{-5}$	$(3.4 \pm 1.0) \times 10^{-4}$	$(1.9 \pm 1.4) \times 10^{-7}$
1.5	$(8.5 \pm 11.0) \times 10^{-6}$	$(3.9 \pm 0.94) \times 10^{-4}$	$(5.3 \pm 6.0) \times 10^{-8}$
2	$(2.4 \pm 5.2) \times 10^{-5}$	$(3.1 \pm 1.1) \times 10^{-4}$	$(1.5 \pm 3.3) \times 10^{-6}$

H NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's main theoretical contribution, namely the derivation of a closed-form local stability threshold β_c for posterior collapse in symmetric Dirichlet β-VAEs with linear Gaussian decoders, together with empirical validation beyond the local regime. They also describe the scope and limitations of the analysis, including the focus on symmetric Dirichlet priors and the local nature of the stability analysis.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discusses the limitations of the analysis in a dedicated section (section 4), including the focus on symmetric Dirichlet priors and the local nature of the stability analysis.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: The paper provides a full set of assumptions and a complete proof for each theoretical result. A short proof sketch is provided in the main paper, complemented by a formal proof in the supplemental material (see Section 2 and Appendix A).

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: the paper fully discloses all information necessary to reproduce the main experimental results, including data generation procedures, model architectures, optimization settings and metric definitions. These details are provided in Appendix B and Appendices in general.

   Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: We provide a minimal reproduction script for the synthetic, MNIST, and Fashion-MNIST experiments, with execution instructions. The complete experimental code will be made available in the GitHub repository: `https://github.com/dldou/dir-vae-collapse`.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The paper specifies all necessary experimental details to understand and interpret the results, including dataset usage (standard benchmarks such as MNIST and Fashion-MNIST), data handling procedures, model hyperparameters, optimization method, and training configurations. These details are provided in the main paper (section 3), appendix B, appendix C, and accompanying code.

   Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: The experimental results report variability across multiple random initializations (three seeds to five seeds depending on the experiment). For experiments repeated across seeds, tables report mean $\pm$ standard deviation, median [minimum, maximum], or median $\pm$ half the interquartile range for JasperRidge (Table 15). Shaded bands span the minimum-maximum range across seeds. These summaries describe variability across initializations and are not confidence intervals. This is clearly stated in the captions and consistently applied to all experimental evaluations supporting the main claims (see Appendix D.1 and related tables).

Guidelines:

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

8. **Experiments compute resources**

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

Answer: [No]

Justification: The experiments in this paper are small-scale and primarily serve to empirically validate theoretical predictions rather than to perform large-scale benchmarking or model training. Experiments use NVIDIA RTX A4500 Laptop, Tesla T4, and Tesla P100-SXM2-16GB GPUs, with one GPU per training run and no distributed training, with experiment-specific architectures and learning rates, as detailed in Section 3.1 and the corresponding appendices. Training uses Adam without weight decay. The datasets used (synthetic Dirichlet-linear Gaussian data, hyperspectral datasets such as Samson, and MNIST subsets) are small to moderate scale, and all models are trained independently at fixed β values without large hyperparameter sweeps. As a result, compute requirements are minimal and reproducible on a single modern GPU. We do not perform large-scale pretraining, distributed optimization, or extensive compute-intensive search procedures. We report the training protocols, but do not provide per-run execution times, memory requirements, or an estimate of the total computational cost.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research was conducted and reported in accordance with the NeurIPS Code of Ethics. The paper is primarily theoretical, with standard computational experiments and no human subjects, personal data collection, or other procedures raising specific ethical concerns.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [No]

    Justification: This work is primarily theoretical and focuses on the analysis of posterior collapse in Dirichlet β-VAEs, without introducing a deployable system or application-specific methodology. As a result, the paper does not include a dedicated discussion of broader societal impacts, although the presented analysis may indirectly inform the design and tuning of latent variable models in future applications.

    Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: This work is primarily theoretical and does not release high-risk models, generative systems, scraped datasets, or other artifacts with a clear risk of misuse. Therefore, no special safeguards for responsible release are required.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: The paper uses publicly available benchmark datasets, including MNIST, FashionMNIST, SVHN, and the hyperspectral datasets Samson and JasperRidge. These datasets are properly cited in the paper with references to their original sources, and their use follows the corresponding dataset licenses and terms of use as described in the respective documentation.

Guidelines:

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

13. **New assets**

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

Answer: [Yes]

Justification: We provide a minimal reproduction script with dependency information, argument descriptions, and example commands. The complete experimental code will be made available in the GitHub repository: `https://github.com/dldou/dir-vae-collapse`.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. All experiments are conducted on publicly available benchmark datasets and do not require participant recruitment, instructions, or compensation.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: An LLM was used in a limited capacity to assist in reviewing and sanity-checking some mathematical proofs for clarity and potential errors. The LLM was not part of the proposed method, training pipeline, or experimental evaluation, and all theoretical results were derived and verified by the authors.

    Guidelines:

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