UNDERSTANDING FEATURE LEARNING DYNAMICS OF SIGREG VIA BHEP STATISTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised training with LeJEPA relies on the SIGReg regularizer, which pulls the embedding distribution toward an isotropic Gaussian and prevents dimensional collapse. However, its random projections make the learning dynamics hard to analyze directly. MMD-Gauss, a closed-form approximation of SIGReg, is a bandwidth mixture of BHEP statistics, and therefore we adopt BHEP as a tractable proxy that shares the same target distribution. Leveraging this theoretical proxy, we derive closed-form ODEs governing the macroscopic feature learning dynamics, exact under the BHEP proxy and our stated assumptions. Based on this formulation, we identify two phases: an initial rapid-growth *Explosion Phase* in which leading covariance eigenvalues overshoot, followed by a self-stabilizing *Shrinking Phase* in which they return toward the isotropic target while later features emerge one after another after increasingly long delays. We show that the eigenvalue dynamics are governed by a global **engine** α, which drives every dormant feature and collapses as the top features saturate, and a **brake** β, which plateaus early, so the target eigenvalue $T = \alpha/\beta$ steps down with each saturated feature. For MMD-Gauss and SIGReg the bandwidth mixture leaves no single engine and brake, but the force on a dormant feature and the common target eigenvalue survive and decrease in the same way, so the two phases and the stepwise delay carry over.

1 INTRODUCTION

Self-supervised learning (SSL) has revolutionized representation learning by enabling models to extract features from unlabeled data. However, a fundamental challenge in SSL is avoiding dimensional collapse, where representations degenerate into a low-dimensional manifold or a single constant vector (Jing et al., 2022). To mitigate this issue, a range of regularization strategies have been introduced. For instance, Barlow Twins (Zbontar et al., 2021) uses implicit batch normalization and the off-diagonal entries of the cross-correlation matrix to decorrelate features, VICReg (Bardes et al., 2022) employs explicit variance and covariance regularization, SimCLR (Chen et al., 2020) leverages negative pairs, and JEPA (Assran et al., 2023) and DINO (Caron et al., 2021) utilize stop-gradient mechanisms with exponential moving averages, and He et al. (2024) mitigate collapse with an orthogonality regularization on the weight matrices.

LeJEPA (Balestriero & LeCun, 2025) introduced the SIGReg regularizer, which encourages the embedding distribution to match an isotropic Gaussian, preventing collapse and improving downstream task performance. LeJEPA demonstrates remarkable performance, particularly on in-domain tasks, without the need for exhaustive hyperparameter tuning, relying on only a single tunable parameter. However, while such approaches constrain the overall embedding distribution, it remains unclear how individual features are actually learned during training.

Although prior analyses have characterized such per-feature dynamics under Barlow Twins (Simon et al., 2023; Kim & Lee, 2026), these results do not transfer to LeJEPA, where feature dynamics are strongly coupled, making the independent per-feature analysis adopted in prior work inapplicable. Moreover, directly analyzing the learning dynamics induced by SIGReg is challenging, as it relies on random projection to utilize the computational efficiency while maintaining the isotropic Gaussian distribution. Recent work (Zimmermann et al., 2025) addressed this by providing a closed-form MMD-Gauss approximation. However, even after the inverse-square-root approximation of its

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

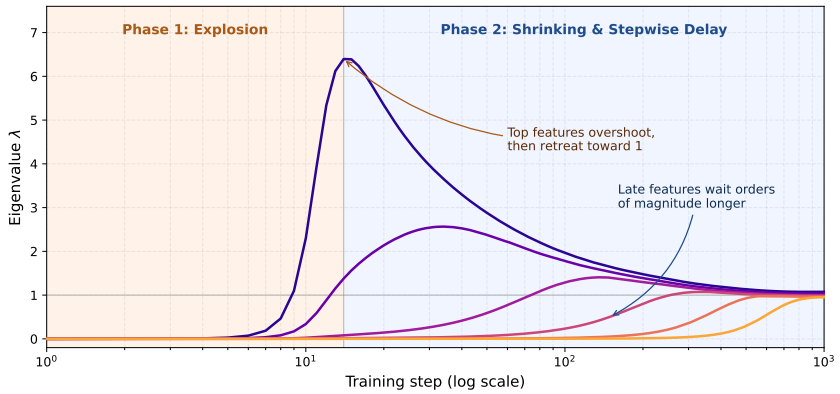


Figure 1: **Learning dynamics of the BHEP proxy.** Eigenvalue trajectories of the embedding covariance over training steps (log scale) show two phases. In *Phase 1 (Explosion)*, a few top principal features rapidly grow and overshoot the target. In *Phase 2 (Shrinking and Stepwise Delay)*, these top features slowly retreat, then lower-ranked features sequentially emerge after increasing delays.

Kummer-function kernel (Knop et al., 2020), the resulting formulation itself remains highly nonlinear and intractable for an explicit ODE derivation.

In this work, we bridge this gap with a Gamma-integral transform, under which MMD-Gauss becomes a Gamma-weighted mixture of BHEP-type statistics (Baringhaus & Henze, 1988). While the resulting integral term still poses analytical hurdles for a purely closed-form solution, this structural equivalence identifies the BHEP statistic as the theoretical bridge to the MMD-Gauss objective. Consequently, we adopt the BHEP statistic as a tractable proxy to derive exact ODEs and map the macroscopic learning trajectories.

Figure 1 shows the resulting dynamics: a few top features grow fast and overshoot, then shrink as later features emerge one by one after increasingly long delays. The same overshoot-shrink pattern has been empirically reported for SimCLR and VICReg (Simon et al., 2023), and discussed for Barlow Twins (Kim & Lee, 2026), so we set out to explain the mechanism behind it for Gaussian regularization.

Concretely, our contributions are:

- **Closed-form ODE for BHEP feature dynamics.** We derive a complete closed-form ODE for each embedding eigenvalue λ_k, exposing an interpretable engine-brake structure, $\dot{\lambda}_k \propto \lambda_k [\alpha - \beta\lambda_k]$ governed by two macroscopic global forces, the engine α and the brake β.
- **Two-phase characterization of training.** We show that the resulting trajectory cleanly separates into an initial *Explosion Phase*, where an unsuppressed α drives top features to overshoot, and a subsequent *Shrinking Phase*, where the asymmetric decay of α relative to a quickly-plateauing β pulls the dynamic target $T = \alpha/\beta$ down toward the isotropic equilibrium $\lambda^* = 1$, mechanistically explaining the stepwise delay of later features.
- **Transfer to SIGReg.** MMD-Gauss, which approximates SIGReg, writes the per-feature force as an integral of engine-brake terms. The integral has no single engine and brake, but the force on a dormant feature (the engine α for BHEP) and the target eigenvalue at which the force vanishes ($T = \alpha/\beta$ for BHEP) do survive. Both decrease for all three objectives (BHEP, SIGReg, and MMD-Gauss), so the two phases and the stepwise delay carry over.

2 Related Work

Learning dynamics in self-supervised learning. Prior work (Simon et al., 2023) provided a powerful framework for analyzing stepwise SSL dynamics by linearizing the network and aligning weights with data eigenvectors, which yields an explicit relationship between embedding eigenvalues and learnable weight factors of weight matrix. Littwin et al. (2024) analyze deep linear networks over

JEPA (Assran et al., 2023) and show that predicting in latent space biases learning toward high-influence features. Eigenvalue dynamics of non-contrastive objectives were analyzed by Tian et al. (2021). Furthermore, Betser et al. (2026) observed that under InfoNCE the alignment (invariance) term saturates early, in some regimes before uniformity, and showed that at this plateau the population objective drives the representation toward a Gaussian distribution. In this work, we build upon these frameworks to isolate and analyze the intricate regularization dynamics of SSL.

Isotropic Gaussian regularization. Regularizing representations towards isotropic Gaussian distributions has shown significant empirical success, notably in LeJEPA (Balestriero & LeCun, 2025) via the SIGReg loss. Recently, Zimmermann et al. (2025) analyzed the theoretical properties of such regularizers, proving that the expectation of the sliced Epps-Pulley (EP) test over infinite projection directions is equivalent to the Maximum Mean Discrepancy (MMD) with a specific kernel related to Kummer's function, which can be approximated by an inverse-square-root kernel in the high-dimensional regime ($d > 20$), an approximation that first appeared in Knop et al. (2020). The Gaussian-kernel MMD between a sample and the standard normal is a classical object whose biased estimator coincides with the BHEP statistic of the Henze-Zirkler test (Henze & Zirkler, 1990; Rustamov, 2021).

3 SIGREG TO BHEP PROXY

3.1 PRELIMINARY: EPPS-PULLEY TEST

The SIGReg regularizer evaluates the Gaussianity of the embedding space using the empirical characteristic function based on the Epps-Pulley test (Epps & Pulley, 1983).

Definition 1 (Epps-Pulley Test)**.** *The loss function measures the difference between an empirical characteristic function $\hat{\phi}_X(t) = \frac{1}{n}\sum_{j=1}^{n} e^{itx_j}$ of data and Gaussian characteristic function $\phi(t) = e^{-\frac{1}{2}t^2}$, weighted by the weight window function $w(t)$.*

$$EP = n\int_{-\infty}^{\infty} |\hat{\phi}_X(t) - \phi(t)|^2 w(t)dt \tag{1}$$

In our runs we divide this term by n, and the coefficients of BHEP and MMD-Gauss are multiplied by $\sqrt{2\pi}$ to match its scale (Appendix C).

Although this formulation only checks the distribution of a scalar quantity, SIGReg applies random projections using random unit vectors, thereby mapping high-dimensional embeddings within a batch to scalar distributions. With sufficient projections, SIGReg effectively enforces an isotropic Gaussian structure, based on the theorem (Cramér & Wold, 1936) that a distribution is determined by its one-dimensional projections, so that if every projection is standard Gaussian, the distribution is the isotropic Gaussian.

3.2 BRIDGING SIGREG TO BHEP VIA THE GAMMA INTEGRATION TRICK

Tracking exact learning dynamics of the SIGReg objective is highly intractable due to the random projections involved in its optimization step. As noted in recent work (Zimmermann et al., 2025), bridging this to a Maximum Mean Discrepancy (MMD) formulation removes the randomness. For a sufficiently large embedding dimension ($d > 20$), the MMD form of SIGReg, whose kernel is Kummer's function, is approximated by an inverse multiquadric kernel through the inverse-square-root approximation that already appears in Knop et al. (2020):

Definition 2 (MMD-Gauss loss)**.**

$$\widehat{\textit{MMD-Gauss}}^2 \approx \frac{1}{n(n-1)}\sum_{j\neq i}\left(1 + \frac{2}{2d-3}\|z_i - z_j\|_2^2\right)^{-1/2} - \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}_y\left(1 + \frac{2}{2d-3}\|z_i - y\|_2^2\right)^{-1/2} + C \tag{2}$$

Here, $y \sim \mathcal{N}(0, I_d)$, d indicates the dimension of embedding, and z_i, z_j are the embeddings within a batch. The expectation over the target is evaluated in closed form (Appendix H), so no target samples

are drawn. Zimmermann et al. (2025) apply the same approximation one step earlier, after absorbing the target into the pull term, which replaces the second term by $\frac{\sqrt{2}}{n}\sum_i(1+\frac{1}{2d-3}\|z_i\|^2)^{-1/2}$. *We address the difference in Appendix I.*

However, due to its inverse polynomial structure, this formulation remains highly non-linear and prevents the direct derivation of closed-form ODEs. To bridge this mathematical gap and leverage a tractable objective, we use a structural link using the Gamma function integral identity:

Definition 3 (Gamma function integral identity)**.**

$$\frac{1}{\sqrt{Z}} = \frac{1}{\sqrt{\pi}}\int_0^\infty t^{-1/2}e^{-Zt}dt$$

By applying this integral transform before the expectation, which is the Gaussian scale-mixture form of a radial kernel (Schoenberg's theorem (Schoenberg, 1938; Tolstikhin et al., 2016), noted for this kernel by Zimmermann et al. (2025)), the intractable inverse polynomial structure can be represented as a continuous mixture of exponential functions (using $s = \frac{2}{2d-3}$):

$$\widehat{\text{MMD-Gauss}}^2 \approx \frac{1}{\sqrt{\pi}}\int_0^\infty t^{-1/2}e^{-t}\Big[P_1(t) - \frac{2}{(1+2st)^{d/2}}P_2(t)\Big]dt + C,$$
$$P_1(t) = \frac{1}{n(n-1)}\sum_{j\neq i}\exp\left(-st\|z_i - z_j\|^2\right), \quad P_2(t) = \frac{1}{n}\sum_{i=1}^n \exp\left(-\frac{st}{1+2st}\|z_i\|^2\right) \tag{3}$$

Analyzing the exact ODEs inside this integral needs t-dependent coefficients (Appendix H), but its structure is simple: each slice of Eq. 3 is a Gaussian-kernel statistic with bandwidth $2st$, exactly the Baringhaus-Henze-Epps-Pulley (BHEP) statistic (Baringhaus & Henze, 1988) defined next.

Definition 4 (BHEP Loss)**.** *The Baringhaus-Henze Epps-Pulley test for multivariate Gaussian:*

$$\textit{BHEP}_{n,\beta_{bw}} = \frac{1}{n}\sum_{j,k=1}^n \exp\left(-\frac{\beta_{bw}^2\|z_j - z_k\|^2}{2}\right)$$
$$- \frac{2}{(1+\beta_{bw}^2)^{d/2}}\sum_{j=1}^n \exp\left(-\frac{\beta_{bw}^2\|z_j\|^2}{2(1+\beta_{bw}^2)}\right) + \frac{n}{(1+2\beta_{bw}^2)^{d/2}} \tag{4}$$

For calculation, we divide this statistic by n *to match the scaling with other regularizers, and drop the diagonal terms* $j = k$ *of the first sum in the formulation (details in Appendix A).*

Term by term, Eq. 3 reads (in the normalized form of Definition 4):

$$\widehat{\text{MMD-Gauss}}^2 \approx \frac{1}{\sqrt{\pi}}\int_0^\infty t^{-1/2}e^{-t}\,\text{BHEP}_{n,\beta_{bw}}\Big|_{\beta_{bw}^2=2st}dt, \tag{5}$$

so MMD-Gauss is a $t^{-1/2}e^{-t}$-weighted mixture of BHEP statistics, and BHEP with $\beta_{bw}^2 = 1/d$, the reference scale of Rustamov (2021), is the slice at $t \approx \frac{1}{2}$. We therefore adopt it as a *tractable surrogate proxy* for the macroscopic learning dynamics.

Assumptions for dynamics analysis. For a stable analysis of the dynamics, we assume sufficient dimensionality d and sample size n, where $(n > d)$. Following the linear network framework (Simon et al., 2023), we align the weight matrix with top-d components of the input covariance matrix $C_X = \frac{1}{n}X^TX$. We also assume that the projections of the input data onto the top eigenvectors (the principal components) follow a multivariate Gaussian distribution. Finally, we do not consider the invariance term to affect the regularization dynamics, assuming it is already saturated, so that the regularizer alone drives the feature dynamics (like in (Betser et al., 2026)).

Feature and eigenvalue We use feature k to refer to the k-th eigenvector of C_Z, and λ_k (its eigenvalue) to refer to the scalar magnitude of that feature. Statements about features emerging, saturating, or being delayed refer to the behavior of λ_k.

4 LEARNING DYNAMICS OF BHEP

This section develops a closed-form characterization of the BHEP feature learning dynamics. We proceed in four steps. Section 4.1 introduces the weight-alignment ansatz that decouples each embedding eigenvalue into an input variance and a learnable weight factor, providing the parametrization for our analysis. Section 4.2 derives the closed-form ODE for individual eigenvalues, then categorizes the components of the resulting ODE. Building on this, Section 4.3 identifies two distinct phases of learning, an initial *Explosion* driven by an unsuppressed engine α followed by a *Shrinking* phase governed by the asymmetric decay of α relative to a quickly plateauing brake β, which together produce the stepwise delay observed in practice. Finally, Section 4.4 extends the dynamics to MMD-Gauss and SIGReg.

4.1 WEIGHT ALIGNMENT AND EIGENVALUE DECOUPLING

To analytically trace the macroscopic learning dynamics within our closed-form BHEP proxy, we build upon the weight alignment framework established by Simon et al. (2023). We assume that the weight matrix W aligns with the principal components of the input data X during training.

Under this alignment assumption, the embedding covariance matrix $C_Z = \frac{1}{n} Z^T Z \in \mathbb{R}^{d\times d}$ is perfectly decoupled into orthogonal feature directions. As detailed in Appendix G, this yields a direct relationship where each macroscopic embedding eigenvalue λ_k is strictly determined by the product of the constant input variance σ_k^2 and its corresponding learnable weight factor s_k^2:

$$\lambda_k = \sigma_k^2 s_k^2 \tag{6}$$

By mapping the gradient flow from the global regularization loss down to the scalar parameters s_k, we can directly formulate how individual features evolve over time.

4.2 CLOSED-FORM DERIVATION

Using the BHEP proxy and decoupled framework, we characterize the exact learning trajectories. By continuous-time gradient descent $\dot{s}_k = -\frac{\partial L}{\partial s_k}$, and the chain rule, the eigenvalue evolves as $\dot{\lambda}_k = 2\sigma_k^2 s_k \dot{s}_k$, which yields the following ODE for the BHEP objective (under assumptions provided in Section 3.2):

Theorem 1 (Eigenvalue ODE)**.** *The time derivative of the k-th embedding eigenvalue λ_k is governed by the interplay of Local and Global forces (Full derivation in Appendix A and Appendix B):*

$$\dot{\lambda}_k = \underbrace{2\sigma_k^2}_{\textit{Local force}} \cdot \lambda_k \cdot \underbrace{u_k \cdot [\alpha - \beta\lambda_k]}_{\textit{Global force}}, \tag{7}$$

As written above, the overall learning dynamics depend on two forces. The *Local force*, which is invariant along learning, relies strictly on the input covariance eigenvalues. This implies that features with larger input variance learn faster, aligning with the stepwise nature proposed in Simon et al. (2023). The *Global force*, however, evolves dynamically. Dormant features ($\lambda_k \approx 0$) are solely powered by the ***driving engine term*** α. Conversely, already active features are bounded by a dynamic target, experiencing a ***regularizing braking term*** β.

Definition 5. *The engine, the brake, and the target eigenvalue T are given by:*

$$\alpha = c_1 P_1 - \gamma c_2 P_2, \qquad \beta = c_1 c_2 (\gamma P_2 - P_1), \qquad T = \frac{\alpha}{\beta} = \frac{c_1 P_1 - \gamma c_2 P_2}{c_1 c_2 (\gamma P_2 - P_1)}\,. \tag{8}$$

To provide a clear understanding of the derived ODE, we categorize the mathematical components based on their origins and roles within the learning trajectory:

- **Global interaction potentials (P_1, P_2):** These terms encapsulate the aggregate state of the entire embedding space. Under the Gaussian and alignment assumptions the two sums of Definition 4 become products over all eigenvalues, $P_1 = \prod_{i=1}^{d} (1 + c_1\lambda_i)^{-1/2}$ (Push) and $P_2 = \prod_{i=1}^{d} (1 + c_2\lambda_i)^{-1/2}$ (Pull) (Appendix A). The names reflect that the first term pushes embeddings apart while the second pulls them toward the origin.

- **Coefficients (c_1, c_2, γ):** These are structural constants introduced purely for notational simplicity during the algebraic derivation. They correspond to $c_1 = \frac{2}{d}$ and $c_2 = \frac{1}{d+1}$, alongside the coefficient $\gamma = \frac{2}{(1+\frac{1}{d})^{d/2}} \approx \frac{2}{\sqrt{e}}$.
- **Denominator (u_k):** The term $u_k = [(1 + c_1\lambda_k)(1 + c_2\lambda_k)]^{-1}$ emerges algebraically to structure the gradients into an explicit ODE format. It equals 1 for a dormant feature and only rescales the speed of a large eigenvalue, so it changes neither the engine α nor the target $T = \alpha/\beta$.

4.3 Observable phenomena: two phases of learning

Every feature is coupled through the shared target $T = \alpha/\beta$: because α and β depend on the whole spectrum, the early features reshape the landscape of the later ones. We formalize this in two phases, an initial *Explosion Phase* and a subsequent *Shrinking and Stepwise Delay Phase* (the Shrinking Phase for short), which explain the overshoot and the delayed emergence of Figure 1.

4.3.1 Phase 1: eigenvalue explosion and engine collapse

Initial explosion. At initialization ($\lambda_k \approx 0$), the global force is overwhelmingly dominated by α. α starts at $\mathcal{O}(1/d)$ ($\approx 0.794/d$ at $d = 128$), while β remains at a much weaker $\mathcal{O}(1/d^2)$ scale ($\approx 0.428/d^2$). This applies a profound driving force across all features.

Instantaneous engine collapse. The mere emergence of a single dominant feature severely suppresses the system's driving force (Figure 2). Consider the extreme case where only the first feature saturates ($m = 1$, $\lambda_1^* \approx 0.658d$). The expansion engine α is instantly halved (from $0.794/d$ to $0.376/d$), while the brake β slightly stiffens (from $0.428/d^2$ to $0.572/d^2$). Figure 2 illustrates these overall dynamics with respect to the number of saturated features for $d = 128$.

4.3.2 Phase 2: shrinking and stepwise delay

Asymmetric decay: the engine vs. the brake. As training continues, more features gradually emerge. Suppose m features have saturated to their local peaks, while the other $(d - m)$ features remain close to 0. In particular the eigenvalue mean $\bar{\lambda} = \frac{1}{d}\sum_{i=1}^{d} \lambda_i$ follows:

$$\bar{\lambda} = \frac{(d+1)(K[m,d]-1)}{2(d+1) - dK[m,d]} \cdot m \approx \frac{e^{\frac{1}{m+2}} - 1}{2 - e^{\frac{1}{m+2}}} \cdot m \approx 1 - \frac{1}{2(m+2)} \quad \text{(for large } m\text{)}, \tag{9}$$

where $K[m,d] = (1 + \frac{1}{d})^{\frac{d+2}{m+2}}$. So $\bar{\lambda}$ steadily rises from ≈ 0.658 toward the ideal isotropic state of 1 as features saturate, shown in Figure 2 (Left).

As features saturate one after another, a profound asymmetry emerges in the global forces (Figure 2 (Right)). The brake β relies on the slower-decaying P_2 and rapidly plateaus, reaching over 95% of its maximum within just roughly 10 active features. In contrast, the engine α suffers drastic term cancellation, losing over 90% of its initial power and collapsing to the $\mathcal{O}(1/d^2)$ scale to match β. (See Appendix D for a small-λ analysis.)

The mechanism of stepwise delay. With β effectively acting as a constant wall, the target eigenvalue $T = \alpha/\beta$ and the learning speed of inactive features ($\dot{\lambda}_k \propto \alpha, \lambda_k \approx 0$) become strictly proportional to the collapsing engine α. The explosion of the first few features instantly cripples the engine's power, so that subsequent features are subjected to a severely depleted driving force and learn in a sequentially delayed manner.

Asymptotic convergence. The decreasing target T forces the early-exploded features to retreat, initiating the *Shrinking Phase*. The process concludes at a stable equilibrium where the engine and brake perfectly match ($\alpha = \beta$), $\lambda_k \to 1$ for all features: the isotropic Gaussian.

4.4 Universality of the dynamics and asymptotic convergence

Per-bandwidth structure of MMD-Gauss. While our macroscopic phase analysis primarily leverages the strictly closed-form BHEP proxy, our framework extends to the MMD-Gauss objective of

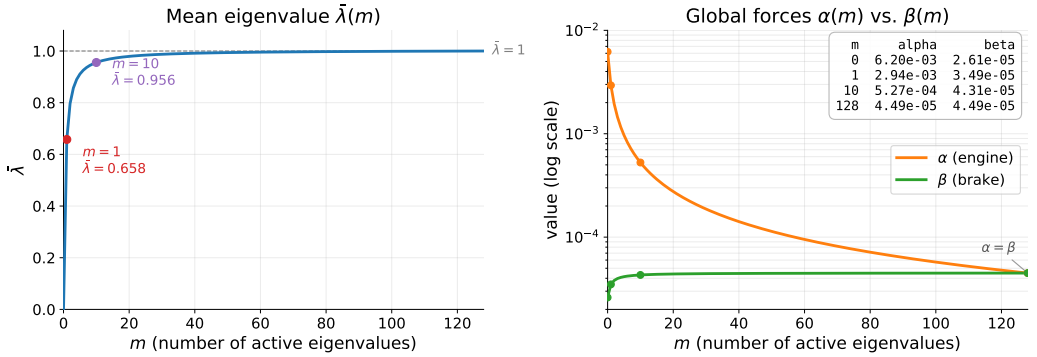


Figure 2: **Learning dynamics of the $\bar{\lambda}, \alpha, \beta$ on $d = 128$ setting:** With only a handful of saturated features, engine α is suppressed immediately, while the brake β is stiffened into plateau. $\bar{\lambda}$ rapidly approaches 1.

Definition 2. Deriving the continuous-time gradient descent under the Gamma integral representation (detailed in Appendix H.2), the learning dynamics for MMD-Gauss yield:

$$\dot{\lambda}_k = 2\sigma_k^2 \lambda_k F(\lambda_k), \qquad F(\lambda_k) = \frac{1}{\sqrt{\pi}} \int_0^\infty u_{t,k}\, t^{1/2} e^{-t} \big[\alpha_t - \beta_t \lambda_k\big]\, dt. \tag{10}$$

F is the global force of Theorem 1: for BHEP it is the single term $u_k[\alpha - \beta\lambda_k]$, for MMD-Gauss it is the same term averaged over the bandwidth t, with α_t, β_t the BHEP coefficients at bandwidth $2st$ divided by t (the factor of t is moved into the weight $t^{1/2}e^{-t}$, Appendix H.3, Eq. 41) and $u_{t,k}$ a positive weight that depends on λ_k and equals 1 at $\lambda_k = 0$.

What does the BHEP analysis say about SIGReg? Its conclusions reach SIGReg through MMD-Gauss. In the integral above the weight $u_{t,k}$ varies with t, so after the integration the force is no longer affine in λ_k and there is no single engine and brake. Two quantities survive the integration: the force on a dormant feature, $F(0) = \frac{1}{\sqrt{\pi}} \int_0^\infty u_{t,0}\, t^{1/2} e^{-t} \big[\alpha_t - \beta_t \cdot 0\big]\, dt = \frac{1}{\sqrt{\pi}} \int_0^\infty t^{1/2} e^{-t} \alpha_t\, dt$, which plays the role of α for BHEP, and the target T at which F vanishes ($F(T) = 0$), which is common to all features and for BHEP equals $T = \alpha/\beta$. Appendix L evaluates both for all objectives on the same ladder of saturated features: both $F(0)$ and target T decrease as features saturate, so the stepwise delay and the overshoot followed by shrinking carry over. All three objectives share the fixed point $\lambda^* = 1$ exactly (Appendix H.5). On the equilibrium ladder, the one visible difference is the path of the mean: BHEP climbs to 1 from below, while MMD-Gauss and SIGReg overshoot and relax back to 1.

5 Experiments

5.1 Two phases on the toy model and on CIFAR-10

To empirically validate our theoretical framework, we evaluate the eigenvalue trajectories on a 32-dimensional toy setting. The first row of Figure 3 illustrates the complete lifecycle of the feature learning dynamics. Each graph corresponds to SIGReg, MMD-Gauss (approximated using Gauss-Laguerre quadrature), and BHEP.

Phase transitions and convergence. As shown in the figure, we clearly observe the macroscopic phases predicted by our theory. Several features rapidly peak during the initial *Explosion Phase*. Each feature that saturates lowers the engine, so the target eigenvalue (black line: $T = \alpha/\beta$ for BHEP, the root of $F(T) = 0$ for MMD-Gauss) steps down as features saturate one after another. This decrease triggers the *Shrinking Phase*, forcing the exploded features to slowly retreat and ultimately converge toward the isotropic fixed point ($T \to \lambda^*$). For MMD-Gauss the force is not affine in λ_k, but its root is common to all features (Appendix H.4), so a single target line is drawn.

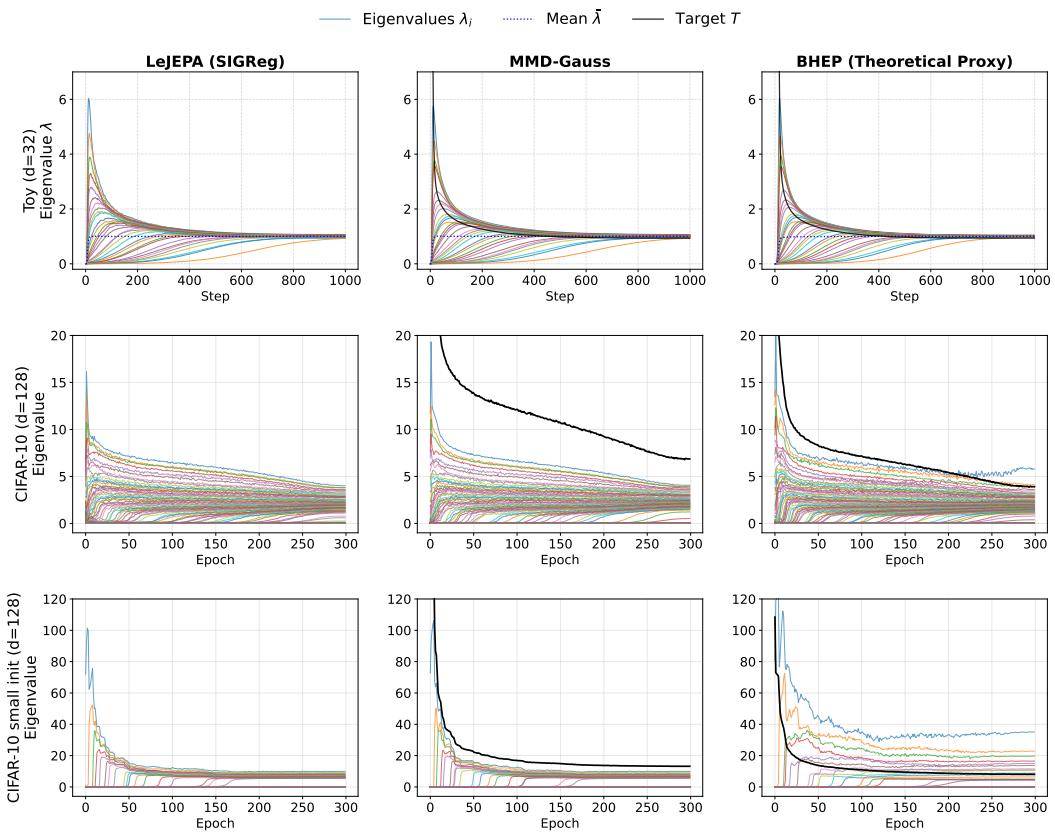


Figure 3: **Eigenvalue learning dynamics on different regularizers and different settings.** Rows: toy model ($d = 32$), CIFAR-10 ResNet-18 with default and with small initialization ($d = 128$). The black line is the target T: $T = \alpha/\beta$ for BHEP and the root of the per-feature force ($F(T) = 0$) for MMD-Gauss (Appendix H). Each panel shows a representative single run, intended as a qualitative illustration of the predicted dynamics rather than as statistical validation.

The origin of stepwise delay. Although the input variances σ_k^2 differ by only about $4\times$, the convergence times across features span orders of magnitude. The reason is the collapse of the engine: each saturating feature depletes α (for BHEP) or $F(0)$ (for MMD-Gauss) by far more than the initial variance gap, so the remaining features are left with a profoundly weakened driving force and emerge one after another.

Practical experiment with CIFAR-10. To validate our theory in a practical setting, Figure 3 shows real training runs on the CIFAR-10 dataset (Krizhevsky, 2009) with a ResNet-18 backbone (He et al., 2016), obtained by replacing the regularizer of LeJEPA in the code of Terver et al. (2026). In the MMD-Gauss panels the target is the root of F computed from the logged slice potentials of the run. Under small initialization and slow learning rate, the stepwise emergence that happens within the first epoch in the standard setting is stretched out and becomes visible.

The forces behind the trajectories. Figure 4 shows the mechanism for both objectives. For BHEP the brake β reaches its fixed-point value β^* within a few epochs and stays there while the engine α drops sharply, so the target $T = \alpha/\beta$ falls exactly as the engine falls. MMD-Gauss has no single brake, so we ask whether the dormant-feature force $F(0)$ carries the same information as the target T. It does: $F(0)$ and T decrease together and their ratio stays within about 10% of the value $\hat{\beta}$ it takes at the isotropic fixed point (Appendix M.1), so MMD-Gauss behaves as if it had a constant brake, and with the right axis scaled by $\hat{\beta}$ the two curves fall on top of each other. The constant is derived at the fixed point only, so along the path this is an empirical regularity. In both cases the decrease of the target is the decrease of the engine.

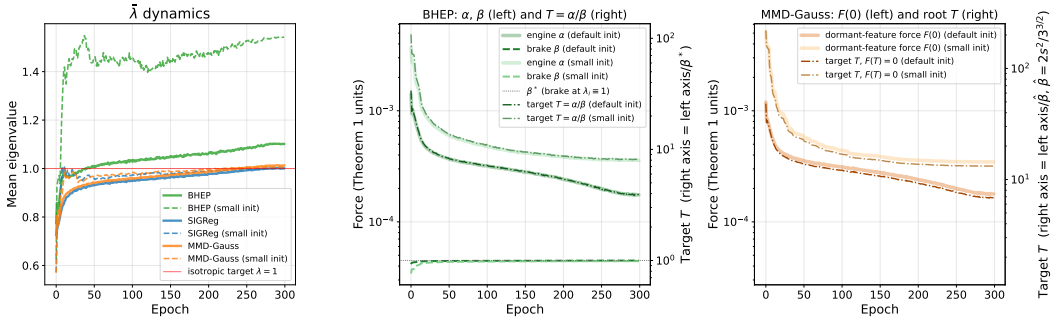


Figure 4: **Mean eigenvalue, global forces, and target on CIFAR-10.** Left: $\bar{\lambda}$ of six runs. Except for BHEP, $\bar{\lambda}$ converges to 1. Middle: BHEP engine α and brake β (left axis) and target $T = \alpha/\beta$ (right axis). Right: MMD-Gauss dormant-feature force $F(0)$ (left axis) and target T, where $F(T) = 0$ (right axis). In both panels the right axis is the left axis divided by the value of $F(0)/T$ at the isotropic fixed point, $\beta^* = c_1 c_2 (1 + \frac{2}{d})^{-d/2}$ for BHEP and $\hat{\beta} = 2s^2/3^{3/2}$ for MMD-Gauss (Appendix M.1), so a force curve lands on its target curve wherever their ratio has reached the fixed-point value: the brake β settles at β^* after a few epochs, and $T \cdot \hat{\beta}$ of MMD-Gauss stays around $F(0)$ along the whole run. Curves show representative single runs.

5.2 Deviations from the idealized predictions

The end points of SIGReg and MMD-Gauss agree with the fixed point $\bar{\lambda} = 1$ of Section 4.4, but neither path is the one the equilibrium ladder predicts: on the ladder BHEP approaches 1 from below and SIGReg and MMD-Gauss overshoot it slightly, whereas here BHEP overshoots ($\bar{\lambda} > 1.4$ under the small initialization) and the other two rise toward 1 from below without a visible overshoot, and the MMD-Gauss eigenvalues stay below the root target for the whole run while the BHEP eigenvalues cross it.

We attribute the difference to two effects outside the idealized assumptions, the **unsaturated invariance term** and the **strongly non-Gaussian embeddings** of BHEP. An unsaturated invariance term subtracts from each feature a force proportional to its augmentation ratio, so the emerged features settle below the target T at different levels and form a band rather than a single line (Appendix K.1). A non-Gaussian embedding moves the push and pull terms away from their Gaussian values, which shifts α and hence T. In our runs the radial fourth cumulant is large, most of all for BHEP under small initialization, and the first-order correction moves the fixed point in the direction of the overshoot of $\bar{\lambda}$ observed there (Appendix K.2). Appendix K describes how each of these assumption violations, including the two effects above, affects the dynamics, and Appendix M provides additional experiments with different datasets and backbone architectures.

6 Conclusions, Limitations, and Future Work

We presented a closed-form analysis of LeJEPA's feature learning dynamics through the BHEP proxy. With the two global forces α and β we have shown why the dynamics unfold in two phases, and that the isotropic state is the fixed point of all three objectives, with the dormant-feature force and the target eigenvalue carrying the same two-phase picture from BHEP to SIGReg.

Our framework has several limitations. The two assumptions most visibly violated in our runs, the unsaturated invariance term and the non-Gaussian embeddings that accompany the overshoot, are analyzed in Appendix K together with the corrections we derived for them, and the remaining ones, the alignment and large-n, d assumptions, the structural gap between BHEP and SIGReg beyond the equilibrium ladder, and the $O(n^2 d)$ cost of the pairwise proxies, are discussed in Appendix N.

Going forward, we hope to analyze the learning dynamics of VICReg (Bardes et al., 2022) and SimCLR (Chen et al., 2020), as they share similar learning dynamics (Simon et al., 2023), or even similar embedding distribution (Wang & Isola, 2020; Betser et al., 2026).

AI USE STATEMENT

Required tasks used In this work, we used generative AI tools for the following tasks with required disclosure: developing the computer-assisted proof of Appendix Q jointly with the authors, who re-derived and checked every step; formulating and checking several secondary mathematical claims, namely the fixed-point value of $F(0)/T$ (Appendix M.1), the root-existence argument of Appendix H.4, and the first-order non-Gaussian correction of Appendix K.2; proposing and refining one hypothesis, that a non-Gaussian embedding shifts the fixed point upward and that a single bandwidth lets BHEP trade it for a higher mean (Appendix K.2); assisting the implementation of the methods (the modifications of the training code, the logging, and the analysis code, including the released version); providing feedback on implementation details of the experiments, while the experimental design is the authors'; assisting in the interpretation of some results, in particular the comparison of the three objectives on the equilibrium ladder and the consistency checks between the reported numbers and the theory; and translating drafts.

Required tasks not used We did not use generative AI tools to generate the synthetic data sets (the toy data are Gaussian samples fully specified by the authors in Appendix O, the tool only wrote the code that samples them), to develop the theoretical framework, or to formulate the main claims of the paper (the BHEP reformulation, Theorem 1, and the two-phase characterization). Cleaning and reformatting of data sets and qualitative or thematic data analysis are not applicable to this work.

Recommended tasks used Additionally, we used generative AI tools to edit the paper for readability, to draft parts of the appendix and the figure captions, to identify related literature and to check citations and possible overlap with prior work, to format and check references, to suggest parameters for the toy settings and the structure of Section 5, and to write and edit software code, including the scripts that produce the figures.

Review over AI-assisted work We have reviewed all AI-assisted work: the generated code was checked against independent re-computations of the printed numbers and against the training logs, the proof steps were re-derived by the authors and the constants of Table 6 are certified by interval arithmetic, and every retrieved reference was checked against the original publication. The core framework of the paper, the BHEP reformulation, Theorem 1, the two-phase characterization, and all real-data experiments were conceived and carried out by the authors.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is a theoretical and empirical study of the training dynamics of self-supervised regularizers. It involves no human subjects and no data collection. Experiments use the public CIFAR-10, STL-10, and CelebA datasets under their respective licenses. CelebA contains face images, and we use it only to measure covariance spectra of learned representations and do not perform any identity-related task. We do not foresee direct harmful applications of the results.

REPRODUCIBILITY STATEMENT

Our code is provided as supplementary material. Every figure and table of the paper has a script named after it (for example `[Fig 4]`), so that the code behind each result can be found directly, and the settings needed to read the results are listed in Appendix O.

The code is grouped in three parts. `toy` contains the linear toy models on synthetic Gaussian data (Figures 1, 3, 5, 6, 7), which run in minutes on a CPU and reproduce the figures in the paper. `theory` contains the closed-form and certified computations (Figure 2, Tables 3, 4, and 6, and the band of Figure 10); the interval-arithmetic script re-derives every enclosure of Table 6 and reports whether all checks pass. `cifar_delta` contains the files we modified or added on top of the EB-JEPA library of Terver et al. (2026) (the three regularizers, the per-epoch logging of the eigenvalues and forces, and one configuration template covering all real-data runs), together with the analysis scripts that produce Figures 4, 8, 9, and 10, Tables 1 and 2, and the κ_{rad} measurement from the

logged runs and checkpoints. The analysis scripts read training logs through placeholders for the run identifiers, so reproducing the real-data results requires re-training the runs with the released configuration.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15619–15629, 2023.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Ludwig Baringhaus and Norbert Henze. A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, 35(1):339–348, 1988.

Roy Betser, Eyal Gofer, Meir Yossef Levi, and Guy Gilboa. InfoNCE induces gaussian distribution. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=BlSH7gNQSq.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Harald Cramér and Herman O. A. Wold. Some theorems on distribution functions. *Journal of The London Mathematical Society-second Series*, pp. 290–294, 1936. URL https://api.semanticscholar.org/CorpusID:122761325.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

T. W. Epps and Lawrence B. Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726, 12 1983. ISSN 0006-3444. doi: 10.1093/biomet/70.3.723. URL https://doi.org/10.1093/biomet/70.3.723.

Junlin He, Jinxiao Du, and Wei Ma. Preventing dimensional collapse in self-supervised learning via orthogonality regularization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 95579–95606. Curran Associates, Inc., 2024. doi: 10.52202/079017-3028. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad7922fd4650f8aba5d8b067e622ca84-Paper-Conference.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

N. Henze and B. Zirkler. A class of invariant consistent tests for multivariate normality. *Communications in Statistics - Theory and Methods*, 19(10):3595–3617, 1990. doi: 10.1080/03610929008830400. URL https://doi.org/10.1080/03610929008830400.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=YevsQ05DEN7.

Juhwan Kim and Sungyoon Lee. Stepwise feature learning in self-supervised learning, 2026. URL https://openreview.net/forum?id=P23YpnH3kq.

Szymon Knop, Przemysław Spurek, Jacek Tabor, Igor Podolak, Marcin Mazur, and Stanisław Jastrzebski. Cramer-wold auto-encoder. *Journal of Machine Learning Research*, 21(164):1–28, 2020. URL http://jmlr.org/papers/v21/19-560.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

Etai Littwin, Omid Saremi, Madhu Advani, Vimal Thilak, Preetum Nakkiran, Chen Huang, and Joshua Susskind. How jepa avoids noisy features: The implicit bias of deep linear self distillation networks. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 91300–91336. Curran Associates, Inc., 2024. doi: 10.52202/079017-2897. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a600f0f740605205133553cb74a1c107-Paper-Conference.pdf.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Raif M Rustamov. Closed-form expressions for maximum mean discrepancy with applications to wasserstein auto-encoders. *Stat*, 10(1):e329, 2021.

I. J. Schoenberg. Metric spaces and completely monotone functions. *Annals of Mathematics*, 39 (4):811–841, 1938. ISSN 0003486X, 19398980. URL http://www.jstor.org/stable/1968466.

James B Simon, Maksis Knutins, Liu Ziyin, Daniel Geisz, Abraham J Fetterman, and Joshua Albrecht. On the stepwise nature of self-supervised learning. In *International Conference on Machine Learning*, pp. 31852–31876. PMLR, 2023.

Basile Terver, Randall Balestriero, Megi Dervishi, David Fan, Quentin Garrido, Tushar Nagarajan, Koustuv Sinha, Wancong Zhang, Mike Rabbat, Yann LeCun, et al. A lightweight library for energy-based joint-embedding predictive architectures. *arXiv preprint arXiv:2602.03604*, 2026.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10268–10278. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/tian21a.html.

Ilya O Tolstikhin, Bharath K. Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximum mean discrepancy with radial kernels. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/5055cbf43fac3f7e2336b27310f0b9ef-Paper.pdf.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

Eric Zimmermann, Harley Wiltzer, Justin Szeto, David Alvarez-Melis, and Lester Mackey. Kerjepa: Kernel discrepancies for euclidean self-supervised learning. *arXiv preprint arXiv:2512.19605*, 2025.

A TRANSFORMATION OF THE BHEP OBJECTIVE INTO A PRODUCT OF EIGENVALUES

Using $\beta_{bw}^2 = \frac{1}{d}$, then the overall formula becomes:

$$L = \frac{1}{n^2}\sum_{j,k=1}^{n} \exp\left(-\frac{\|z_j - z_k\|^2}{2d}\right) - \frac{2}{n(1+\frac{1}{d})^{d/2}}\sum_{j=1}^{n}\exp\left(-\frac{\|z_j\|^2}{2(d+1)}\right) + \frac{1}{(1+\frac{2}{d})^{d/2}} \tag{11}$$

As we assume the projections of the input data onto the top eigenvectors (the principal components) follow a multivariate Gaussian distribution ($x \sim \mathcal{N}(0, \Sigma)$), we can use the formula:[1]

$$\mathbb{E}_x\left[\exp\left(-\frac{1}{2}x^T M x\right)\right] = \det(I + \Sigma M)^{-1/2} \tag{12}$$

Applying this identity with $\Sigma = 2C_z, M = \frac{1}{d}I_d$ for the first term (since $z_i - z_j \sim \mathcal{N}(0, 2C_z)$) and $\Sigma = C_z, M = \frac{1}{d+1}I_d$ for the second term, and the overall loss formula simplifies to a function of determinants:

Collecting the two prefactors into constants,

$$\begin{aligned} L &= \det\left(I + \frac{2}{d}C_Z\right)^{-1/2} - \frac{2}{(1+\frac{1}{d})^{d/2}}\det\left(I + \frac{1}{d+1}C_Z\right)^{-1/2} + \frac{1}{(1+\frac{2}{d})^{d/2}} \\ &= \det\left(I + \frac{2}{d}C_Z\right)^{-1/2} - \gamma\det\left(I + \frac{1}{d+1}C_Z\right)^{-1/2} + C \end{aligned} \tag{13}$$

where $\gamma = \frac{2}{(1+\frac{1}{d})^{d/2}} \approx \frac{2}{\sqrt{e}}, \quad C \approx \frac{1}{e}$ if d is sufficiently large.

Using $\lambda_k = \sigma_k^2 s_k^2$, we can derive loss as:

$$\begin{aligned} L &= \prod_{i=1}^{d}(1 + c_1\lambda_i)^{-1/2} - \gamma\prod_{i=1}^{d}(1 + c_2\lambda_i)^{-1/2} + C \\ &= \underbrace{\prod_{i=1}^{d}(1 + c_1\sigma_i^2 s_i^2)^{-1/2}}_{\text{Push: } P_1} - \gamma\underbrace{\prod_{i=1}^{d}(1 + c_2\sigma_i^2 s_i^2)^{-1/2}}_{\text{Pull: } P_2} + C \end{aligned} \tag{14}$$

where $c_1 = \frac{2}{d}, c_2 = \frac{1}{d+1}$.

B DERIVING THE CLOSED-FORM LEARNING DYNAMICS

Using the chain rule, $\frac{dL}{ds_k} = \frac{dL}{d\lambda_k} \cdot \frac{d\lambda_k}{ds_k}$,

[1] Definition 4 sums over all pairs, including $j = k$, whose terms equal $e^0 = 1$ regardless of the representation: this is the V-statistic, which Rustamov (2021) identify with the BHEP statistic of the Henze-Zirkler test. Because the population P_1 is an expectation over two *independent* representations, we drop the diagonal terms and weight the remaining pairs by $\frac{1}{n(n-1)}$, for both the loss and the logged forces, as in Definition 2. The resulting statistic is, term by term, the unbiased closed-form Gaussian-kernel MMD of Rustamov (2021, Proposition) with kernel width $\gamma^2 = d$, so we keep the name BHEP for the family while using its unbiased estimator. For large n the two forms differ only by a constant and a factor $\frac{n}{n-1}$ (0.4% at $n = 256$), and keeping the diagonal would raise P_1 by $\frac{1}{n}$ and, since α is a small difference of two $O(1)$ terms, α and T by $O(d/n)$, moving the stationary point of the isotropic slice to $1 - \frac{2}{n}$.

$$\frac{dL}{d\lambda_k} = P_1 \cdot \left(-\frac{1}{2}\right) \cdot \left(\frac{c_1}{1+c_1\lambda_k}\right) - \gamma \cdot P_2 \cdot \left(-\frac{1}{2}\right) \cdot \left(\frac{c_2}{1+c_2\lambda_k}\right) \tag{15}$$

Furthermore, since $\frac{d\lambda_k}{ds_k} = 2\sigma_k^2 s_k$ and $\dot{s}_k = -\frac{dL}{ds_k}$,

$$\begin{aligned}
\dot{s}_k &= \sigma_k^2 s_k \cdot \left[P_1 \cdot \frac{c_1}{1+c_1\lambda_k} - \gamma \cdot P_2 \cdot \frac{c_2}{1+c_2\lambda_k}\right] \\
&= \frac{\sigma_k^2 s_k}{(1+c_1\lambda_k)(1+c_2\lambda_k)}[(c_1P_1 - \gamma c_2 P_2) + c_1c_2(P_1 - \gamma P_2)\lambda_k] \\
&= \sigma_k^2 s_k \cdot u_k \cdot [(c_1P_1 - \gamma c_2P_2) - c_1c_2(\gamma P_2 - P_1)\sigma_k^2 s_k^2]
\end{aligned} \tag{16}$$

Where $u_k = \frac{1}{(1+c_1\lambda_k)(1+c_2\lambda_k)}$. This yields a formulation quite similar to that of Simon et al. (2023).

If we use $\dot{\lambda}_k = 2\sigma_k^2 s_k \cdot \dot{s}_k$ to track eigenvalue dynamics, then

$$\begin{aligned}
\dot{\lambda}_k &= 2\sigma_k^2 s_k \cdot \sigma_k^2 s_k u_k [(c_1P_1 - \gamma c_2P_2) + c_1c_2(P_1 - \gamma P_2)\lambda_k] \\
&= \underbrace{2\sigma_k^2}_{\text{Local force}} \cdot \lambda_k \cdot \underbrace{u_k[\alpha - \beta\lambda_k]}_{\text{Global force}}
\end{aligned} \tag{17}$$

C SIGREG - LOSS COEFFICIENT

In the pseudocode and official repository of LeJEPA (Balestriero & LeCun, 2025), $e^{-\frac{1}{2}t^2}$ is used for both $w(t)$ and target characteristic function $\phi(t)$, instead of $\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}$ for the window $w(t)$. So it is recommended to multiply $\sqrt{2\pi} \approx 2.5066$ to regularizer coefficient, if you want to match magnitude of other regularizers with SIGReg.

$$\begin{aligned}
EP &= \int_{-\infty}^{\infty} |\hat{\phi}_X(t) - \phi(t)|^2 w(t) dt \\
&= \int_{-\infty}^{\infty} (\hat{\phi}_X(t) - \phi(t)) \cdot (\overline{\hat{\phi}_X(t)} - \phi(t)) \cdot e^{-\frac{1}{2}t^2} dt \\
&= \int_{-\infty}^{\infty} (\hat{\phi}_X(t) \cdot \overline{\hat{\phi}_X(t)} - \phi(t) \cdot (\overline{\hat{\phi}_X(t)} + \hat{\phi}_X(t)) + \phi(t)^2) \cdot e^{-\frac{1}{2}t^2} dt \\
&= \int_{-\infty}^{\infty} \frac{1}{n^2} \sum_{j,k=1}^{n} e^{-\frac{1}{2}t^2 + it(z_j - z_k)} - \frac{1}{n}\sum_{j=1}^{n}\left(e^{-t^2 - itz_j} + e^{-t^2 + itz_j}\right) + e^{-\frac{3}{2}t^2} dt \\
&= \frac{\sqrt{2\pi}}{n^2} \sum_{j,k=1}^{n} e^{-\frac{1}{2}(z_j - z_k)^2} - \frac{\sqrt{2\pi}}{n}\sum_{j=1}^{n}\left(\frac{2}{\sqrt{2}} \cdot e^{-\frac{1}{4}(z_j)^2}\right) + \sqrt{\frac{2\pi}{3}} \\
&= \frac{\sqrt{2\pi}}{n^2} \sum_{j,k=1}^{n} e^{-\frac{1}{2}(z_j - z_k)^2} - \frac{\sqrt{2\pi}\cdot\sqrt{2}}{n}\sum_{j=1}^{n}\left(e^{-\frac{1}{4}(z_j)^2}\right) + \sqrt{\frac{2\pi}{3}} \\
&= \sqrt{2\pi} \cdot \left(\frac{1}{n^2} \sum_{j,k=1}^{n} e^{-\frac{1}{2}(z_j - z_k)^2} - \frac{\sqrt{2}}{n}\sum_{j=1}^{n}\left(e^{-\frac{1}{4}(z_j)^2}\right) + \frac{1}{\sqrt{3}}\right)
\end{aligned} \tag{18}$$

We consider the ratio with the first term, that becomes 1 at initialization. Averaged over directions, this loss is the Kummer-kernel MMD of Zimmermann et al. (2025), of which MMD-Gauss is the inverse-root approximation.

Population form under the Gaussian assumption. The bracket above is, up to the constant, the biased Gaussian-kernel MMD^2 between the projected sample $\{z_j\}$ and $\mathcal{N}(0,1)$. Under the Gaussian assumption of Appendix A the projection onto a unit direction θ is Gaussian, $z_j \sim \mathcal{N}(0, v)$ with

$v = \theta^T C_Z \theta$, so the two sample averages have closed-form expectations: $z_j - z_k \sim \mathcal{N}(0, 2v)$ gives $\mathbb{E}\left[e^{-\frac{1}{2}(z_j - z_k)^2}\right] = (1+2v)^{-1/2}$, and $\mathbb{E}\left[e^{-\frac{1}{4}z_j^2}\right] = (1+\frac{v}{2})^{-1/2}$, so that $\sqrt{2}\,(1+\frac{v}{2})^{-1/2} = 2\,(2+v)^{-1/2}$. Hence, after the $1/\sqrt{2\pi}$ correction, the population SIGReg loss of one projection is

$$h(v) = (1+2v)^{-1/2} - 2\,(2+v)^{-1/2} + \tfrac{1}{\sqrt{3}}, \tag{19}$$

and the full objective is its average over the uniformly random direction, $L_{\mathrm{SIGReg}} = \mathbb{E}_\theta[h(\theta^T C_Z \theta)]$, which vanishes exactly at $C_Z = I_d$. This is the form used in Appendix L, where $v = \sum_j \xi_j^2 \lambda_j$ in the eigenbasis of C_Z and ξ_j^2 are the squared components of θ.

D Small λ assumption

If all $\lambda_i \ll d$, and $u_k \approx 1$, and we can formulate the derivation as:

$$\begin{aligned} P_1 &= \prod_{i=1}^{d} \left(1 + \frac{2}{d}\lambda_i\right)^{-1/2} \\ &\approx \exp\left(-\frac{1}{2}\sum_{i=1}^{d} \ln\left(1 + \frac{2}{d}\lambda_i\right)\right) \\ &\approx \exp\left(-\frac{1}{2}\sum_{i=1}^{d} \frac{2}{d}\lambda_i\right) \\ &\approx \exp(-\bar{\lambda}) \end{aligned} \tag{20}$$

$$\begin{aligned} P_2 &= \prod_{i=1}^{d} \left(1 + \frac{1}{d+1}\lambda_i\right)^{-1/2} \\ &\approx \exp\left(-\frac{1}{2}\sum_{i=1}^{d} \ln\left(1 + \frac{1}{d+1}\lambda_i\right)\right) \\ &\approx \exp\left(-\frac{1}{2}\sum_{i=1}^{d} \frac{1}{d+1}\lambda_i\right) \\ &\approx \exp\left(-0.5\bar{\lambda}\right) \end{aligned} \tag{21}$$

Where $\bar{\lambda} = \frac{1}{d}\sum_{i=1}^{d} \lambda_i$, mean of eigenvalues.

This formulation can be applied to the early starting phase ($\lambda \approx 0$), or later shrinking phase ($\lambda \approx k \ll d$), or convergence ($\lambda = 1$).

So, the overall formulation would be (using $\gamma \approx \frac{2}{\sqrt{e}}$):

$$\begin{aligned} \dot{\lambda}_k &= 2\sigma_k^2 \cdot \lambda_k[(c_1 P_1 - \gamma c_2 P_2) - c_1 c_2(\gamma P_2 - P_1)\lambda_k] \\ &\approx 2\sigma_k^2 \cdot \lambda_k \left[\left(\frac{2}{d}e^{-\bar{\lambda}} - \frac{2}{\sqrt{e}(d+1)}e^{-0.5\bar{\lambda}}\right) - \frac{2}{d(d+1)}\left(\frac{2}{\sqrt{e}}e^{-0.5\bar{\lambda}} - e^{-\bar{\lambda}}\right)\lambda_k\right] \end{aligned} \tag{22}$$

This also shows, to the order of the approximation, that when all $\lambda_i = 1$, the overall global force becomes 0, corresponding to the global minimum of isotropic Gaussian.

E Learning dynamics at start

We start from Eq. 16, without replacing denominator into u_k, and use $\dot{\lambda}_k = 2\sigma_k^2 s_k \cdot \dot{s}_k$, then the form would be (Using $\gamma \approx 2/\sqrt{e}$):

$$\dot{\lambda}_k = 2\sigma_k^2 \lambda_k \cdot \left[\frac{2}{d} \cdot \frac{\prod_{i=1}^{d}(1 + c_1\lambda_i)^{-1/2}}{1 + c_1\lambda_k} - \frac{2}{\sqrt{e}(d+1)} \cdot \frac{\prod_{i=1}^{d}(1 + c_2\lambda_i)^{-1/2}}{1 + c_2\lambda_k}\right]. \tag{23}$$

Assuming that the biggest eigenvalue λ_1 is the only eigenvalue being learned, then $\lambda_{i\neq 1} \approx 0$. Then, we can rewrite formula for λ_1 as:

$$\begin{aligned}\dot{\lambda}_1 &= 2\sigma_1^2\lambda_1 \cdot \left[\frac{2}{d}\cdot(1+c_1\lambda_1)^{-3/2} - \frac{2}{\sqrt{e}(d+1)}\cdot(1+c_2\lambda_1)^{-3/2}\right] \\ &\approx 2\sigma_1^2\lambda_1 \cdot \left[\frac{2}{d}\cdot(1+\frac{2}{d}\lambda_1)^{-3/2} - \frac{2}{\sqrt{e}d}\cdot(1+\frac{1}{d}\lambda_1)^{-3/2}\right] \\ &\approx \frac{2\sigma_1^2}{d}\lambda_1 \cdot \left[2(1+\frac{2}{d}\lambda_1)^{-3/2} - \frac{2}{\sqrt{e}}\cdot(1+\frac{1}{d}\lambda_1)^{-3/2}\right] \end{aligned} \tag{24}$$

where the dimension d is big enough (to approximate $\frac{1}{d} \approx \frac{1}{d+1}$).

Using first-order Taylor series to approximate, and retrieve $\dot{s}_1$ by using $\dot{\lambda}_k = 2\sigma_k^2 s_k \cdot \dot{s}_k$, then the formula becomes:

$$\begin{aligned}\dot{s}_1 &\approx \frac{1}{d}\sigma_1^2 \cdot s_1 \cdot \left[\left(2-\frac{2}{\sqrt{e}}\right) - \left(6-\frac{3}{\sqrt{e}}\right)\cdot\frac{1}{d}\sigma_1^2 s_1^2\right] \\ &\approx \frac{1}{d}\sigma_1^2 \cdot s_1 \cdot \left[A - B\cdot\frac{1}{d}\sigma_1^2 s_1^2\right] \end{aligned} \tag{25}$$

Also, we can obtain the critical time τ_1 for this approximation:

$$\tau_1 = \frac{-\log\left(s_1^2(0)\frac{\frac{1}{d^2}\sigma_1^4 B}{\frac{1}{d}\sigma_1^2 A}\right)}{2\cdot\frac{1}{d}\sigma_1^2 A} = d\cdot\frac{-\log\left(\frac{\sigma_1^2}{d}\cdot s_1^2(0)\cdot\frac{(6-\frac{3}{\sqrt{e}})}{(2-\frac{2}{\sqrt{e}})}\right)}{2\sigma_1^2\cdot(2-\frac{2}{\sqrt{e}})}. \tag{26}$$

This approximated formulation mirrors the stepwise ODE derived in prior work on Barlow Twins (Simon et al., 2023). Powered by this initial force, the top features enter the rapid explosion phase.

The Taylor approximation is not accurate at the peak itself, but this discrepancy does not invalidate its analytical utility. Because the initial eigenvalue growth is exponential, the vast majority of the training time (critical time τ_1) is spent in the near-zero regime where the expansion holds, so the estimated timing of emergence τ_1 remains robust. Moreover, both the true peak time and the approximated critical time scale linearly with d.

F EIGENVALUE MEAN MOVEMENT

Let's generalize from Appendix E. Let's assume there are m eigenvalues that are already trained $\lambda_i = \lambda_k$, and the other $(d-m)$ untrained $\lambda_i = 0$. Then we can expand formula as (using $\gamma \approx \frac{2}{\sqrt{e}}, c_2 \approx \frac{1}{d}$):

$$\begin{aligned}\dot{\lambda}_k &= 2\sigma_k^2\lambda_k \cdot \left[\frac{2}{d}\cdot\frac{\prod_{i=1}^{d}(1+c_1\lambda_i)^{-1/2}}{1+c_1\lambda_k} - \frac{2}{\sqrt{e}(d+1)}\cdot\frac{\prod_{i=1}^{d}(1+c_2\lambda_i)^{-1/2}}{1+c_2\lambda_k}\right] \\ &\approx \frac{4\sigma_k^2}{d}\lambda_k \cdot \left[(1+\frac{2}{d}\lambda_k)^{-\frac{m+2}{2}} - \frac{1}{\sqrt{e}}\cdot(1+\frac{1}{d}\lambda_k)^{-\frac{m+2}{2}}\right] \end{aligned} \tag{27}$$

As all those eigenvalues are assumed to have converged, their corresponding $\dot{\lambda}_k = 0$. Then,

$$\begin{aligned}(1+\frac{2}{d}\lambda_k)^{-\frac{m+2}{2}} &= e^{-\frac{1}{2}}\cdot(1+\frac{1}{d}\lambda_k)^{-\frac{m+2}{2}} \\ (1+\frac{2}{d}\lambda_k) &= e^{\frac{1}{m+2}}\cdot(1+\frac{1}{d}\lambda_k) \\ \therefore \lambda_k^* &= \frac{e^{\frac{1}{m+2}}-1}{2-e^{\frac{1}{m+2}}}d \end{aligned} \tag{28}$$

The overall eigenvalue mean, $\bar{\lambda} = \frac{1}{d}\sum_{i=1}^{d}\lambda_i = \frac{e^{\frac{1}{m+2}}-1}{2-e^{\frac{1}{m+2}}}m$. This approximated eigenvalue mean starts from $\bar{\lambda} \approx 0.655$, and gradually increases with number of learned eigenvalues, and get close to 1 at convergence. Since this uses the approximation, the mean does not converge exactly to 1 but slightly below it, by $\mathcal{O}(\frac{1}{d})$.

F.1 Exact mean eigenvalue without approximation

Without any approximation (γ, c_2) used above, the fixed point at $m = d$ is exactly $\bar{\lambda}^* = 1$:

$$\begin{aligned}\dot{\lambda}_k &= 2\sigma_k^2\lambda_k \cdot \left[\frac{2}{d}\cdot\frac{\prod_{i=1}^{d}(1+c_1\lambda_i)^{-1/2}}{1+c_1\lambda_k} - \frac{2}{(1+1/d)^{d/2}(d+1)}\cdot\frac{\prod_{i=1}^{d}(1+c_2\lambda_i)^{-1/2}}{1+c_2\lambda_k}\right] \\ &= 4\sigma_k^2\lambda_k \cdot \left[\frac{1}{d}\cdot(1+\frac{2}{d}\lambda_k)^{-\frac{m+2}{2}} - \frac{1}{(1+1/d)^{d/2}(d+1)}\cdot(1+\frac{1}{d+1}\lambda_k)^{-\frac{m+2}{2}}\right] \end{aligned} \tag{29}$$

This should be 0, then

$$\frac{1}{d}\cdot(1+\frac{2}{d}\lambda_k)^{-\frac{m+2}{2}} = \frac{1}{(1+1/d)^{d/2}(d+1)}\cdot(1+\frac{1}{d+1}\lambda_k)^{-\frac{m+2}{2}} \tag{30}$$

Let's assume $K[m,d] = \left(\frac{(1+1/d)^{d/2}(d+1)}{d}\right)^{2/(m+2)} = (1+\frac{1}{d})^{\frac{d+2}{m+2}}$, then

$$(1+\frac{2}{d}\lambda_k) = K[m,d]\cdot(1+\frac{1}{d+1}\lambda_k) \tag{31}$$

$$\therefore \lambda^* = \frac{d(d+1)(K[m,d]-1)}{2(d+1)-dK[m,d]} \tag{32}$$

And the corresponding eigenvalue mean would be: $\bar{\lambda}^* = \frac{m(d+1)(K[m,d]-1)}{2(d+1)-dK[m,d]}$.

At $m = d$ we have $K[d,d] = 1+\frac{1}{d}$, so $\lambda^* = \frac{(d+1)}{2(d+1)-(d+1)} = 1$ and $\bar{\lambda}^* = 1$ exactly, as claimed.

G Derivation of weight alignment and eigenvalue decoupling

In this section, we provide the detailed derivation for the eigenvalue decoupling (Eq. 6 in the main text), adopting the framework utilized to analyze stepwise learning in self-supervised learning architectures (Simon et al., 2023).

Let $C_X = \frac{1}{n}X^TX = V\Sigma^2V^T$ be the eigendecomposition of the input covariance, where $\Sigma^2 = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_k^2)$ contains the input variances along the principal directions. We assume the weight matrix aligns with the top d eigenvectors of C_X, that is, the singular value decomposition of W has them as its left singular vectors:

$$W = V^{(\le d)}SQ^T \tag{33}$$

where $V^{(\le d)} \in \mathbb{R}^{k\times d}$ whose columns are the top d eigenvectors of C_X, $S = \mathrm{diag}(s_1, s_2, \ldots, s_d)$ contains the continuous, learnable weight coefficient for each feature, and Q is an orthogonal matrix.

Then the eigendecomposition over the feature matrix can be simplified as:

$$\begin{aligned} C_Z &= \frac{1}{n}Z^TZ = W^TC_XW \\ &= QS(V^{(\le d)})^T(V\Sigma^2V^T)V^{(\le d)}SQ^T \\ &= Q(S^2\Sigma^2_{\le d})Q^T = Q\Lambda Q^T \end{aligned} \tag{34}$$

Here, $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$ represents the eigenvalues of the embedding covariance matrix. Crucially, this exact decoupling demonstrates that each macroscopic embedding eigenvalue λ_k is strictly determined by the constant input variance σ_k^2 and its corresponding weight eigenvalue s_k^2 of the corresponding singular value of W:

$$\lambda_k = \sigma_k^2 s_k^2 \tag{35}$$

H GAMMA-INTEGRAL DERIVATION OF MMD-GAUSS

This appendix derives the engine-brake form of Definition 2 used in Section 4.4, and shows in what sense MMD-Gauss is an integral of BHEP statistics over the bandwidth. All three objectives of this paper are squared distances between the sample distribution and the isotropic Gaussian target, so each is a sum of three terms,

$$\underbrace{k(\|z_i - z_j\|_2^2)}_{\text{sample-sample}} - \underbrace{2 \cdot k(\|z_i - y\|_2^2)}_{\text{sample-target}} + \underbrace{k(\|y - y'\|_2^2)}_{\text{target-target}}, \tag{36}$$

with $z_i, z_j \sim C_Z$ and $y, y' \sim \mathcal{N}(0, I_d)$ which are independent to each other. The sample-sample term is the push potential P_1 of this paper, the target-target term is a constant C and is dropped from the derivation, and the sample-target term is the pull potential P_2.

H.1 INTEGRATING OUT THE GAUSSIAN TARGET

The pull term never needs target samples, because the Gaussian expectation of a Gaussian kernel is available in closed form. For one coordinate $y \sim \mathcal{N}(0, 1)$,

$$\mathbb{E}_y\big[e^{-a(z-y)^2}\big] = (1 + 2a)^{-1/2} \exp\Big(-\frac{a}{1 + 2a}\, z^2\Big), \tag{37}$$

and in d dimensions the factor becomes $(1+2a)^{-d/2}$ with the same shrunken exponent. Definition 4 is already written in this form: for the kernel $\exp(-\beta_{bw}^2\|z - y\|^2/2)$ the identity gives the prefactor $(1 + \beta_{bw}^2)^{-d/2}$ and the exponent $\beta_{bw}^2\|z\|^2/(2(1 + \beta_{bw}^2))$ of the BHEP pull term, and Appendix A collects them into the constants $c_1 = 2\beta_{bw}^2$, $c_2 = \beta_{bw}^2/(1 + \beta_{bw}^2)$, and $\gamma = 2(1 + \beta_{bw}^2)^{-d/2}$ of Theorem 1. The same step is how Zimmermann et al. (2025) obtain their MMD-Gauss estimator (their Theorem 8). Definition 2 applies the inverse-square-root approximation of the Kummer function before this integration and integrates each Gaussian slice exactly, whereas the original derivation integrates first and approximates afterwards. The latter is written in Appendix I.

H.2 MMD-GAUSS AS AN INTEGRAL OF BHEP STATISTICS

Applying the Gamma identity $Z^{-1/2} = \frac{1}{\sqrt{\pi}} \int_0^\infty t^{-1/2} e^{-Zt} dt$ to both terms of Definition 2 and integrating the target out of every slice with Eq. 37 gives Eq. 3, whose slice at t is the Gaussian-kernel statistic with $\beta_{bw}^2 = 2st$, $s = \frac{2}{2d-3}$. We therefore write each slice with the BHEP constants at that bandwidth,

$$c_1(t) = 4st, \qquad c_2(t) = \frac{2st}{1 + 2st}, \qquad \gamma_t = 2(1 + 2st)^{-d/2}, \tag{38}$$

so that Eq. 3 reads $\widehat{\text{MMD-Gauss}}^2 \approx \frac{1}{\sqrt{\pi}} \int_0^\infty t^{-1/2} e^{-t}\big[P_1(t) - \gamma_t P_2(t)\big] dt + C$ with the batch sums $P_1(t), P_2(t)$ of Eq. 3, the exponent of $P_2(t)$ being $c_2(t)\|z_i\|^2/2$. Under the Gaussian assumption of Appendix A the batch sums become determinants, and

$$\begin{aligned}
\widehat{\text{MMD-Gauss}}^2 &\approx \frac{1}{\sqrt{\pi}} \int_0^\infty t^{-1/2} e^{-t}\big[\det(I + c_1(t) C_z)^{-1/2} - \gamma_t \det(I + c_2(t) C_z)^{-1/2}\big] dt + C \\
&= \frac{1}{\sqrt{\pi}} \int_0^\infty t^{-1/2} e^{-t}\big[\prod_i (1 + c_1(t)\lambda_i)^{-1/2} - \gamma_t \prod_i (1 + c_2(t)\lambda_i)^{-1/2}\big] dt + C \\
&= \frac{1}{\sqrt{\pi}} \int_0^\infty t^{-1/2} e^{-t}\big[P_1(t) - \gamma_t P_2(t)\big] dt + C,
\end{aligned} \tag{39}$$

The bracket is the BHEP statistic of Definition 4 at bandwidth $\beta_{bw}^2 = 2st$ up to its constant, so MMD-Gauss is the $t^{-1/2}e^{-t}$-weighted integral of BHEP statistics over the bandwidth, and BHEP with $\beta_{bw}^2 = 1/d$ is the slice $t = \frac{2d-3}{4d} \approx \frac{1}{2}$, where the weight $t^{1/2}e^{-t}$ of the force integral (Eq. 41) becomes maximized over all bandwidth.

H.3 ENGINE-BRAKE FORM

For each fixed t the bracket of Eq. 39 is a BHEP loss, so the derivation of Appendix B applies slice by slice. With $\lambda_k = \sigma_k^2 s_k^2$ and $\dot{s}_k = -\partial L/\partial s_k$,

$$\dot{\lambda}_k = 2\sigma_k^2 \lambda_k \, F(\lambda_k), \qquad F(\lambda_k) = -2\frac{\partial L}{\partial \lambda_k},$$
$$F(\lambda_k) = \frac{1}{\sqrt{\pi}} \int_0^\infty t^{-1/2} e^{-t} \left[\frac{c_1(t) P_1(t)}{1 + c_1(t)\lambda_k} - \gamma_t \frac{c_2(t) P_2(t)}{1 + c_2(t)\lambda_k} \right] dt, \tag{40}$$

and the bracket is exactly the per-feature force of Theorem 1 at bandwidth $2st$, $u_{t,k}[\alpha_t^{\mathrm{BHEP}} - \beta_t^{\mathrm{BHEP}} \lambda_k]$ with $u_{t,k} = [(1 + c_1(t)\lambda_k)(1 + c_2(t)\lambda_k)]^{-1}$. Since $c_1(t)$ and $c_2(t)$ carry one factor of t each, we move it into the measure and define

$$F(\lambda_k) = \frac{1}{\sqrt{\pi}} \int_0^\infty t^{1/2} e^{-t} \, u_{t,k} \big[\alpha_t - \beta_t \lambda_k \big] dt, \tag{41}$$

$$\alpha_t = \frac{c_1(t)P_1(t) - \gamma_t c_2(t) P_2(t)}{t} = 4s \Big[P_1(t) - \frac{\gamma_t P_2(t)}{2(1 + 2st)} \Big], \tag{42}$$

$$\beta_t = \frac{c_1(t) c_2(t) \big(\gamma_t P_2(t) - P_1(t) \big)}{t} = \frac{8 s^2 t}{1 + 2st} \big[\gamma_t P_2(t) - P_1(t) \big].$$

For every t the bracket has the form of Theorem 1, and the t-dependence of $u_{t,k}$ is what prevents a single engine and brake after the integration. Since $u_{t,0} = 1$, the force on a dormant feature is the plain bandwidth average of the engine, $F(0) = \frac{1}{\sqrt{\pi}} \int_0^\infty t^{1/2} e^{-t} \alpha_t \, dt$, which is the quantity logged as $F(0)$ in the experiments.

H.4 TARGET EIGENVALUE T

For BHEP the target is the explicit ratio $T = \alpha/\beta$. For MMD-Gauss the target is the root of F, the eigenvalue at which the force vanishes. Since F is the same function for every feature, this root is common to all features, and we compute it by a one-dimensional root search where $F(T) = 0$ (estimated from the batch in the toy model and as epoch means over all batches on CIFAR-10, Figures 3 and 4).

The root exists whenever $F(0) > 0$. The slice brake β_t is not sign-definite: $\gamma_t P_2(t) < P_1(t)$ on the large-t slices while the spectrum lies below isotropy, so the bandwidth average of β_t may be negative early in training and a BHEP-style ratio of averaged coefficients is not a target. The sign of F at large λ_k is nevertheless fixed: since $u_{t,k} \sim (c_1(t)c_2(t)\lambda_k^2)^{-1}$, $F(\lambda_k) \approx -\lambda_k^{-1} \frac{1}{\sqrt{\pi}} \int_0^\infty t^{-1/2} e^{-t} \big(\gamma_t P_2(t) - P_1(t) \big) \, dt = -(C - \widehat{\mathrm{MMD\text{-}Gauss}}^2)/\lambda_k$ as $\lambda_k \to \infty$, with C the target-target constant of the loss. Since the loss is a squared MMD, $C - \widehat{\mathrm{MMD\text{-}Gauss}}^2$ is this bandwidth average of $\gamma_t P_2 - P_1$, and as long as the regularizer stays below its value at the collapsed initialization ($z \equiv 0$, where $P_1 = P_2 = 1$), we get $C - \widehat{\mathrm{MMD\text{-}Gauss}}^2 \geq \frac{1}{\sqrt{\pi}} \int_0^\infty t^{-1/2} e^{-t} (\gamma_t - 1) \, dt = \sqrt{2} - 1 + O(1/d) > 0$. So F is positive at 0 and negative for large λ_k, and a root exists.

On the equilibrium ladder of Appendix Q (m saturated features at a common level, the rest dormant, $d \to \infty$), the analogous endpoint signs in that level, together with a negative slope at every zero, give existence and uniqueness of the rung target $T(m)$ for every m.

H.5 ISOTROPIC FIXED POINT

At $\lambda_i = 1$ for all i, then $\gamma_t P_2(t) = 2(1 + 2st)^{-d/2} \Big(\frac{1+4st}{1+2st} \Big)^{-d/2} = 2(1 + 4st)^{-d/2} = 2P_1(t)$, so $\alpha_t = \beta_t = \frac{8s^2 t}{1+2st} P_1(t)$ and

$$F(\lambda_k) = (1 - \lambda_k) \frac{1}{\sqrt{\pi}} \int_0^\infty t^{1/2} e^{-t} \, u_{t,k} \, \beta_t \, dt. \tag{43}$$

Every bandwidth is at its own BHEP balance and the root is exactly 1, as for SIGReg and BHEP. The value of $F(0)$ at this point is computed in Appendix M.1.

I THE ABSORBED FORM OF MMD-GAUSS

Zimmermann et al. (2025) derive MMD-Gauss in the opposite order to Definition 2: the Gaussian target is integrated out of the exact kernel first, which yields Kummer's function with the prefactor $(1 + 2\gamma_{mmd}\sigma^2)^{-1/2} = 1/\sqrt{2}$ and the argument shrunk by the same factor, and the inverse-square-root approximation is applied afterwards. This replaces the pull term of Definition 2 by

$$-\frac{\sqrt{2}}{n}\sum_{i=1}^{n}\left(1+\frac{1}{2d-3}\|z_i\|_2^2\right)^{-1/2}, \tag{44}$$

and we call the resulting objective the absorbed form. We record how it expands and where it differs.

Under the Gamma identity the absorbed form becomes

$$\widehat{\text{MMD-Gauss}}^2_{\text{abs}} \approx \frac{1}{\sqrt{\pi}}\int_0^\infty t^{-1/2}e^{-t}\left[P_1(t) - \sqrt{2}\,P_2^{\text{abs}}(t)\right]dt + C, \tag{45}$$

with $P_2^{\text{abs}}(t) = \prod_i \left(1+st\lambda_i\right)^{-1/2}$, and the steps of Appendix H.3 give the same engine-brake form with $u_{t,k}^{\text{abs}} = [(1+4st\lambda_k)(1+st\lambda_k)]^{-1}$ and

$$\alpha_t^{\text{abs}} = s\left(4P_1(t) - \sqrt{2}P_2^{\text{abs}}(t)\right), \qquad \beta_t^{\text{abs}} = s^2 t\left(4\sqrt{2}P_2^{\text{abs}}(t) - 4P_1(t)\right). \tag{46}$$

Compared with Eq. 42, the slice constants $(1+2st)^{-d/2}$ and $c_2(t)$ have been replaced by their large-d limits e^{-t} and st (after the substitution $t \to t/2$ that produces the $\sqrt{2}$), so the slices are no longer individually balanced at $\lambda_i \equiv 1$: here $\beta_t^{\text{abs}} > 0$ for every t, and the target is the root of the integrated force as before.

Fixed point. On a uniform spectrum $\lambda_i \equiv \lambda$ the force vanishes when

$$\int_0^\infty t^{1/2}e^{-t}\left[4\big(1+4st\lambda\big)^{-\left(\frac{d}{2}+1\right)} - \sqrt{2}\big(1+st\lambda\big)^{-\left(\frac{d}{2}+1\right)}\right]dt = 0. \tag{47}$$

Expanding $\lambda = a + b/d$ in $1/d$, the leading order gives $4(2a+1)^{-3/2} = \sqrt{2}\,(\frac{a}{2}+1)^{-3/2}$, that is $a = 1$, and the $O(1/d)$ terms give $\frac{10}{3} - (2b+7) = \frac{5}{6} - (b+\frac{7}{2})$, that is $b = -1$. The absorbed form therefore settles at

$$\lambda^*_{\text{abs}} = 1 - \frac{1}{d}, \tag{48}$$

instead of the exact 1 of Definition 2: the $O(1/d)$ difference between the two orders of approximation shows up as an $O(1/d)$ shift of the fixed point. Tables 3 and 4 list the two forms side by side on the equilibrium ladder.

J THEORY-PRACTICE GAP ON THE TOY SETTING

To verify the steady-state predictions of Section 4.3 quantitatively, we directly compare the closed-form $\alpha(m)$, $\beta(m)$, and $\bar{\lambda}(m)$ against measurements from BHEP training on the 10-dim toy setting.

Figure 5 shows quantitative agreement between the analytical predictions (with the exact $\bar{\lambda}^*$ from Appendix F.1) and the empirical trajectory. In short, they are *accelerated* over theory, as empirical curves reflect simultaneous multi-feature activation in practice: when several eigenvalues grow concurrently, α-collapse and β-plateau both proceed faster than the one-by-one schedule assumed in the derivation.

K ASSUMPTIONS

As we wrote in main section, this paper's theory is based on several assumptions, and in this section we are going to address major assumptions: *invariance term saturation*, *Gaussian distribution*, *large n,d*, and *alignment*.

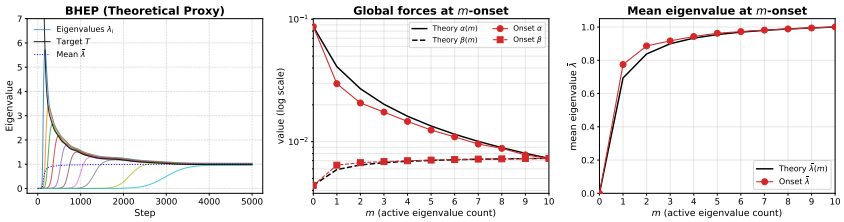


Figure 5: **Dynamics vs. number of active eigenvalues m for 10 dimension theoretical and empirical setting.** $\bar{\lambda}$ jumps toward 1, α collapses, and β plateaus after a few features (Used 0.01x learning rate warmup for first 1000 steps (epochs) to observe discrete eigenvalue emergence).

K.1 How augmentation of the invariance term affects learning

While deriving formulations, we assumed that invariance term does not affect learning dynamics: already saturated. However, in practical setting, it is highly unlikely, as augmentations are applied to promote model to learn invariant features. So, in this section, we are adding gradient over invariance term.

We have additional assumption that:

- Augmentation shares same SVD over input, that augmentation $A = U\Sigma_A V^T$ (Appendix G).
- Augmentation is also distributed over multivariate Gaussian.

Then each view is an augmented copy of the input:

$$X_v = X + A_v, \quad v = 1, 2, \qquad [A_v \sim C_A] \tag{49}$$

where $C_A = \frac{1}{n} A_v^T A_v = \frac{1}{n} V \Sigma_A^2 V^T$, and A_1, A_2 are independent of each other and of X. So the covariance matrix of either view is

$$\begin{aligned} C_{Z_v} &= \frac{1}{n} Z_v^T Z_v = W^T (C_X + C_A) W \\ &= Q(S^2(\Sigma^2_{\leq d} + \Sigma^2_{A:\leq d}))Q^T = Q\Lambda' Q^T \end{aligned} \tag{50}$$

Here, $\Lambda' = \text{diag}(\lambda'_1, \lambda'_2, \ldots, \lambda'_d) = \text{diag}((\sigma_1^2 + a_1^2)s_1^2, (\sigma_2^2 + a_2^2)s_2^2, \ldots, (\sigma_d^2 + a_d^2)s_d^2)$.

Then for the original derivation for BHEP (Theorem 1), would be:

$$\dot{\lambda}_k = \underbrace{2(\sigma_k^2 + a_k^2)}_{\text{Local force}} \cdot \lambda_k \cdot \underbrace{u_k \cdot [\alpha - \beta\lambda_k]}_{\text{Global force}}. \tag{51}$$

K.1.1 GRADIENT FROM INVARIANCE TERM

Using the dimension-agnostic invariance term between the two views:

$$L_{\text{inv}} = \frac{1}{nd}\sum_{i=1}^{n}\|z_{i,1} - z_{i,2}\|_2^2 \tag{52}$$

$$= \frac{1}{nd}\|Z_1 - Z_2\|_F^2 \tag{53}$$

$$= \frac{1}{nd}\|(X_1 - X_2)W\|_F^2 \tag{54}$$

$$= \frac{1}{nd}\text{tr}\left(W^T(X_1 - X_2)^T(X_1 - X_2)W\right) \tag{55}$$

$$= \frac{1}{d}\text{tr}\left(W^T(2C_A)W\right) \quad \text{(in expectation, since } \tfrac{1}{n}A_1^T A_2 \to 0\text{)} \tag{56}$$

$$= \frac{2}{d}\sum_{i=1}^{d} a_i^2 s_i^2 \tag{57}$$

Since the gradient of a sum is the sum of the gradients, the invariance term adds to the weight dynamics the contribution

$$\dot{s}_{i,\text{inv}} = -\frac{\partial L_{\text{inv}}}{\partial s_i} = -\frac{4a_i^2 s_i}{d}, \tag{58}$$

and, with $\lambda_i = (\sigma_i^2 + a_i^2)s_i^2$ as defined above, the corresponding contribution to the eigenvalue dynamics is $\dot{\lambda}_{i,\text{inv}} = 2(\sigma_i^2 + a_i^2)s_i\dot{s}_{i,\text{inv}} = -\frac{8}{d}a_i^2\lambda_i$.

Then the unified loss term (invariant + BHEP) would be derived as:

$$\dot{\lambda}_k = 2(\sigma_k^2 + a_k^2)\cdot\lambda_k\cdot u_k\cdot[\alpha - \beta\lambda_k] - \frac{8}{d}a_k^2\cdot\lambda_k \tag{59}$$

K.1.2 AUGMENTATION: REMOVAL OF FEATURES

By linearizing the dynamics around $\lambda_k \approx 0$ (with $\lambda_k > 0$), we obtain the effective rate at which an eigenvalue grows as it leaves the origin. If this rate is negative, the eigenvalue cannot emerge at all, and the corresponding feature is effectively removed. This gives a simple criterion for predicting whether a feature survives or is pruned.

At $\lambda_k = 0$ we have $u_k \approx 1$, so the linearized rate reads

$$\dot{\lambda}_{k,0} = \left(2(\sigma_k^2 + a_k^2)\,\alpha - \frac{8}{d}\,a_k^2\right)\lambda_k. \tag{60}$$

The sign of the bracket is governed by the ratio of augmentation variance to total variance. Defining the augmentation ratio ν_k and the threshold τ as

$$\nu_k = \frac{a_k^2}{\sigma_k^2 + a_k^2}, \qquad \tau = \frac{d\alpha}{4}, \tag{61}$$

whether a feature collapses is determined by comparing ν_k with τ:

- if $\nu_k > \tau$, the rate is negative and the correlated feature collapses.
- if $\nu_k < \tau$, the rate is positive and the correlated feature survives.

The invariance term's force pushes down the ODE, and this accounts for two features of the practical runs. On the one hand the survivors settle above the isotropic target. Since α is a decreasing function of the saturated spectrum, a collapsed feature contributes a factor of 1 to P_1 and P_2 instead of suppressing them, so with fewer saturated features the engine, and with it the target $T = \alpha/\beta$, stays higher than in the fully saturated case. The surviving, augmentation-invariant features are therefore driven to a higher level, and the more features collapse, the higher the survivors sit: in Table 1 the survivor mean rises from 2.1 to 3.9 as the survivor count falls from 60 to 28. Each survivor itself

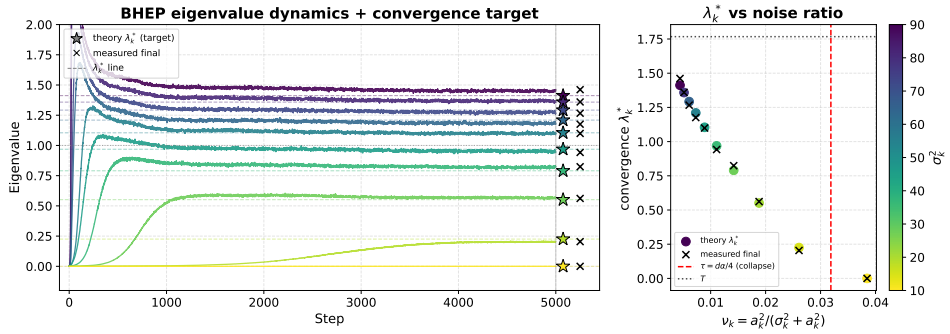


Figure 6: **Learning dynamics with augmentation.** Depending on input variance, features differ not only in their emergence times but also in the eigenvalues they converge to. The colorbar shows the block variance $10 \cdot \sigma_k^2$.

settles below that target, since the invariance term also brakes it: setting $\dot{\lambda}_k = 0$ in Equation 59 gives $\lambda_k = T - \frac{4\nu_k}{d\beta u_k}$, which would make survivors with different ν_k settle at different levels. On the other hand the mean eigenvalue $\bar{\lambda}$ goes down rather than up, because the count falls faster than the level rises: the product of survivor count and survivor mean, divided by d, approximately reproduces the measured $\bar{\lambda}$ for every encoder in the table (0.84 against a measured 0.85 for the MLP). A higher invariance loss thus comes with fewer survivors, a higher and more spread-out survivor band, and a lower mean, which is the pattern seen along the architecture axis of Table 1 and, qualitatively, along the dataset axis of Table 2.

Two caveats complicate a direct application of this criterion. First, consistent with the role of the invariance term, the model can substitute a more augmentation-invariant feature (one with smaller ν_k) for a collapsing one, so removal is not determined for each feature in isolation. Second, the criterion presupposes a saturated α, so it can only be evaluated once learning has fully saturated.

K.1.3 Adaptation to the Toy Model

We test the theory on a 10-dimensional toy model analogous to the one in the main section. The input variance is spread over $\sigma_k^2 \in [1.0^2, 3.0^2]$ while the augmentation variance is held fixed at $a_k^2 = 0.2^2$. As a result, as seen in figure 6, ν_k varies across features, and features whose ν_k exceeds the threshold collapse.

This mechanism accounts for two effects observed in practice:

- Individual convergence level drifts across eigenvalues rather than sitting at a fixed value, and
- "Stacked" eigenvalues that settle at different levels instead of collapsing onto a single line.

K.1.4 Augmentation dynamics on learning

As noted above, the encoder learns augmentation-invariant features over training, so the residual augmentation content of the learned features decreases. To mimic this in the toy model, we let the augmentation strength decrease during training. We use the toy setting of Figure 3 ($d = 32$) with the block-structured augmentation of the previous section, with a relative augmentation ratio $r_k = a_k/\sigma_k$ spread from 20% on the largest-variance block to 40% on the smallest, so that ν_k ranges over 0.04 to 0.14. The augmentation is multiplied by a factor $g(t)$ that decreases exponentially from 1 to 0 over training. MMD-Gauss is Definition 2, with the target drawn as the root of the per-feature force.

Shown in Figure 7, we can see several features that are observed in Figure 3 and figures in Appendix M:

- Increasing $\bar{\lambda}$: In MMD-Gauss and SIGReg, in ideal condition, $\bar{\lambda}$ slightly increases over its goal 1 and slides back towards its goal(Appendix L), unlike BHEP. The augmentation

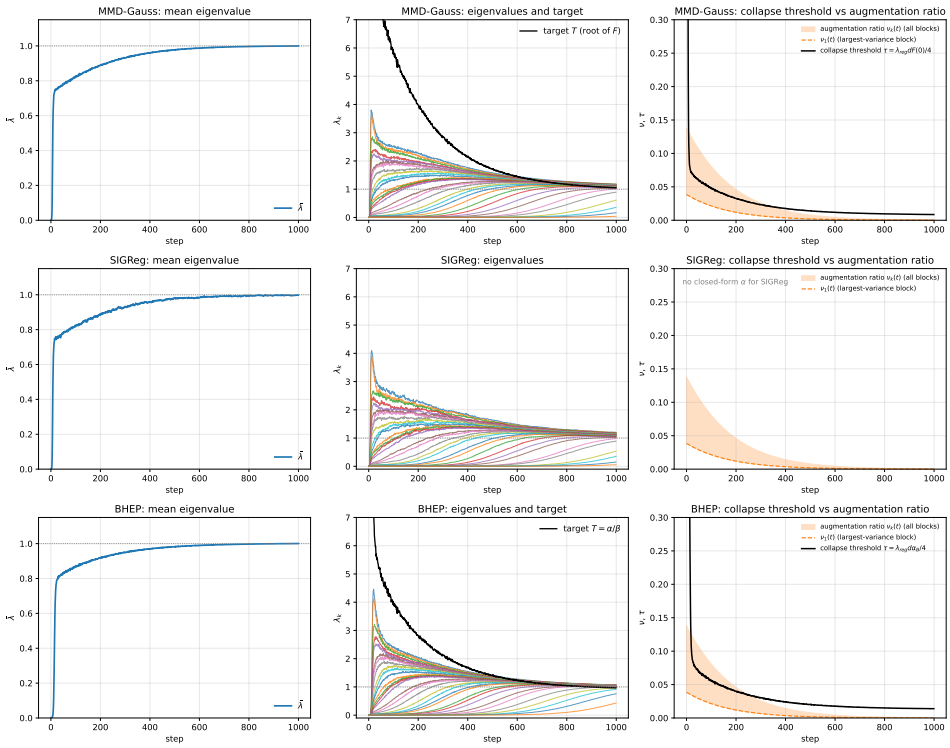


Figure 7: **Learning dynamics with decreasing augmentation.** Onto learning, we have applied decreasing augmentation over learning, which suppresses features scaling with the corresponding SNR ratio for each features (Used regularizer weight $\lambda_{\text{reg}} = 2.5$, which scales along the collapse threshold).

Table 1: **Architecture sweep on CIFAR-10** (end of training, representative single runs, standard setting). A higher invariance loss, indicating more residual augmentation in the learned features, comes with fewer surviving directions and higher, more spread-out survivor levels, for both BHEP and SIGReg, as the collapse criterion predicts, while the overall mean $\bar{\lambda}$ decreases.

		MLP (3-layer)	ViT-S	ResNet-18	ResNet-34
BHEP	invariance loss	0.46	0.21	0.12	0.11
	survivors ($\lambda > 0.5$)	40	55	65	66
	survivor mean	2.93	2.45	2.16	2.13
	survivor spread (std/mean)	0.58	0.48	0.40	0.38
	eigenvalue mean $\bar{\lambda}$	0.92	1.05	1.10	1.10
SIGReg	invariance loss	0.41	0.20	0.12	0.11
	survivors ($\lambda > 0.5$)	28	44	58	60
	survivor mean	3.85	2.78	2.20	2.12
	survivor spread (std/mean)	0.56	0.49	0.36	0.32
	eigenvalue mean $\bar{\lambda}$	0.85	0.96	1.00	1.00

force removes this jump, and all three rise toward 1 from below, as in the CIFAR-10 runs in Figure 3.

- Shimmering band: As the portion over augmentation becomes smaller (as the real-data model learns to acquire augmentation-invariant features, for empirical explanation), they slowly converge into a target line, and stack becomes thinner.

The toy model supports this attribution: with an augmentation that decreases over training all three objectives approaching upwards monotonically from below (Fig. 7).

K.1.5 Different architectures and datasets in the practical setting

We also test the mechanism on the CIFAR-10 setting of the main text, varying the encoder architecture: ResNet-18 and ResNet-34 (He et al., 2016), a 3-layer MLP (width 2048), and ViT-S (Dosovitskiy et al., 2021). The residual augmentation content is not observable feature by feature, but the invariance loss provides an aggregate indicator: a larger final invariance loss means the encoder leaves more augmentation-variant content in its features, that is, a larger ν_k on average.

At the end of training the ordering is $L_{\text{Res34}} \approx L_{\text{Res18}} < L_{\text{ViT}} < L_{\text{MLP}}$. The collapse criterion above then predicts that, moving from ResNet-18 to ViT-S to the MLP, more features should cross the threshold and collapse: fewer directions survive, and the survivors should settle at higher and more spread-out levels, as the less-suppressed engine pushes them further. Table 1 confirms this ordering for both BHEP and SIGReg.

Empirically, a higher invariance loss comes with fewer surviving features and a lower mean $\bar{\lambda}$ (Table 1). The direction is what the collapse criterion suggests, since pruned features neither contribute to the mean nor suppress the engine, but we do not have a quantitative prediction for $\bar{\lambda}$: the survivor count and levels depend on the augmentation ratios ν_k, which we do not model. We leave this to future work.

The same test can be run along the dataset axis. Keeping the ResNet-18 encoder and the CIFAR-10 recipe fixed, we train on STL-10 and CelebA (Appendix M for the trajectories). The final invariance loss is close on CIFAR-10 and STL-10 and less than half of it on CelebA, so the criterion predicts the most survivors, and the lowest survivor levels, on CelebA. Table 2 shows this for both objectives: the survivors rise from 65 to 77 to 128 for BHEP and from 58 to 67 to 113 for SIGReg, while the survivor mean falls toward 1 and the mean $\bar{\lambda}$ stays near 1. As for the architecture axis, these are single runs and the direction, not the size, is what the criterion predicts.

K.2 Non-Gaussian distribution

In this paper, we assumed a Gaussian *input distribution*. Because the weight itself is linear, the resulting embedding distribution is also Gaussian. However, for a nonlinear network the regularizer

Table 2: **Dataset sweep with ResNet-18** (end of training, representative single runs, standard setting). Same statistics as Table 1, whose ResNet-18 column is repeated here. A lower invariance loss comes with more surviving directions and lower survivor means, the same direction as along the architecture axis.

		CIFAR-10	STL-10	CelebA
BHEP	invariance loss	0.12	0.11	0.05
	survivors ($\lambda > 0.5$)	65	77	128
	survivor mean	2.16	1.84	1.00
	survivor spread (std/mean)	0.40	0.41	0.26
	eigenvalue mean $\bar{\lambda}$	1.10	1.11	1.00
SIGReg	invariance loss	0.12	0.11	0.05
	survivors ($\lambda > 0.5$)	58	67	113
	survivor mean	2.20	1.91	1.10
	survivor spread (std/mean)	0.35	0.37	0.31
	eigenvalue mean $\bar{\lambda}$	1.00	1.00	0.98

only sees the embedding distribution, not the input distribution. Below we quantify the error of this replacement and measure it on the trained embeddings: it is governed by a single number, the radial fourth cumulant κ_{rad}, which vanishes for every Gaussian.

K.2.1 The corrected transform

From Eq. 12, we can separate the original form as:

- $L_z(a) = \mathbb{E}[e^{-a\|z\|_2^2}]$: Empirical value using actual distribution over z
- $L_G(a) = \det(I + 2a\Sigma)^{-1/2}$: Theoretical value using covariance matrix over z

where $L_z(a) = L_G(a)$ in ideal Gaussian distribution used in this paper.

If the embedding is not Gaussian, the replacement must be corrected. To see how, let us go back to the identity that made it possible, and redo it without the Gaussian assumption. We will use one coordinate z with variance λ: as coordinates can be collected at the end.

The starting point is that the kernel itself is a Gaussian integral:

$$e^{-az^2} = \frac{1}{\sqrt{4\pi a}} \int e^{-\frac{t^2}{4a}} \cdot e^{itz} dt. \tag{62}$$

Taking the expectation over z on both sides, only e^{itz} depends on z, so we can push expectation inside the integral:

$$\mathbb{E}\big[e^{-az^2}\big] = \frac{1}{\sqrt{4\pi a}} \int e^{-\frac{t^2}{4a}} \mathbb{E}\big[e^{itz}\big] dt. \tag{63}$$

The remaining expectation is the characteristic function of z, and cumulants are, by definition, the Taylor coefficients of its logarithm:

$$\mathbb{E}\big[e^{itz}\big] = \exp\Big(-\frac{\lambda t^2}{2} + \frac{\kappa_4\, t^4}{24} + \cdots\Big), \tag{64}$$

where odd cumulants are omitted as they are odd in t and integrate to zero (and, $\kappa_2 = \lambda$).

Inserting Eq. 64 and merging the two Gaussian exponents,

$$-\frac{t^2}{4a} - \frac{\lambda t^2}{2} = -\frac{t^2}{4a}\big(1 + 2a\lambda\big), \tag{65}$$

and if the exponent stops here, the integral is purely Gaussian:

$$\frac{1}{\sqrt{4\pi a}} \int e^{-\frac{t^2}{4a}(1+2a\lambda)}\, dt = (1 + 2a\lambda)^{-1/2}, \tag{66}$$

which is the per-coordinate determinant factor: we have re-derived Eq. 12. For a non-Gaussian embedding, we expand the remaining factor to first order, $e^{\kappa_4 t^4/24} \approx 1 + \frac{\kappa_4 t^4}{24}$, and average t^4 under the merged Gaussian, whose variance $v = \frac{2a}{1+2a\lambda}$ (as $-\frac{t^2}{2v} = -\frac{t^2(1+2a\lambda)}{4a} \to v = \frac{2a}{1+2a\lambda}$):

$$\frac{\kappa_4}{24}\,\mathbb{E}\big[t^4\big] = \frac{\kappa_4}{24}\cdot 3v^2 = \frac{\kappa_4}{24}\cdot 3\Big(\frac{2a}{1+2a\lambda}\Big)^2 = \frac{\kappa_4}{8\,\sigma^4} \tag{67}$$

Where $\sigma^2 = \frac{1}{2a} + \lambda$.

In d dimensions, the same contraction collects all pairs of coordinates and replaces κ_4 by the radial excess

$$\kappa_{\mathrm{rad}} := \mathbb{E}\|z\|^4 - (\operatorname{tr}\Sigma)^2 - 2\operatorname{tr}(\Sigma^2), \tag{68}$$

the excess variance of $\|z\|^2$ over its Gaussian value, which vanishes for every Gaussian. Altogether, on the isotropic slice $\Sigma = \lambda I$, where λ is the common embedding eigenvalue:

$$L_z(a) = L_G(a)\Big[1 + \frac{\kappa_{\mathrm{rad}}}{8\,\sigma^4} + \cdots\Big], \qquad \sigma^2 = \frac{1}{2a} + \lambda, \tag{69}$$

where the dots collect higher cumulants (the sixth, and the squared third), which stay $O(d)$ for weakly dependent coordinates and are therefore negligible at $O(1/d)$.

For reference, the κ_{rad} for well-known distribution are (i.i.d. coordinates, standardized variance=1):

- Gaussian distribution: $\kappa_{\mathrm{rad}} = 0$
- Laplace distribution: $\kappa_{\mathrm{rad}} = 3d$
- Uniform distribution: $\kappa_{\mathrm{rad}} = -1.2d$
- Rademacher distribution: $\kappa_{\mathrm{rad}} = -2d$ (Lowest Possible)

K.2.2 What the error does to the dynamics

We have quantified the distribution error, but what can this error of relative size $O(\kappa_{\mathrm{rad}}/d^2)$, do to the overall dynamics?

As we said before, the error happens wherever an expectation is replaced by a determinant. This occurs in two terms: the push term (P_1) and the pull term (P_2). For ease, let's name their relative errors ϵ_1 and ϵ_2. Each follows from Eq. 69 by inserting the right kernel and argument. The push kernel has $\frac{1}{2a_1} = d$ and looks at $z_i - z_j$, whose per-coordinate variance is 2λ and whose cumulants are doubled, being a difference of *two* independent embeddings. The pull kernel has $\frac{1}{2a_2} = d + 1$ and looks at z alone, with variance λ. Hence

$$\epsilon_1 = \frac{2\kappa_{\mathrm{rad}}}{8\,(d+2\lambda)^2} = \frac{\kappa_{\mathrm{rad}}}{4\,(d+2\lambda)^2} \tag{70}$$

$$\epsilon_2 = \frac{\kappa_{\mathrm{rad}}}{8\,(d+1+\lambda)^2}. \tag{71}$$

The λ in the denominators is negligible for the size of each error. It becomes essential at the isotropic point $\lambda = 1$, where the two total variances becomes identical: $d + 2\lambda = d + 1 + \lambda = d + 2$. At that time, the relative error becomes constant, where $\epsilon_1 = 2\epsilon_2$ exactly. This is the seed of the cancellation shown next.

K.2.3 Effect on the engine and the brake, and the loss

Let's look at the α, β. At the isotropy, $\forall i \lambda_i = 1$, we can rewrite γP_2 as,

$$\gamma P_2 = 2\Big(1 + \tfrac{1}{d}\Big)^{-d/2}\Big(1 + \tfrac{1}{d+1}\Big)^{-d/2} = 2\Big(1 + \tfrac{2}{d}\Big)^{-d/2} = 2P_1, \tag{72}$$

Using this, the overall formula becomes much simple:

$$\frac{\Delta\alpha}{\alpha} = \frac{c_1 P_1 \epsilon_1 - c_2 \gamma P_2 \epsilon_2}{c_1 P_1 - c_2 \gamma P_2} = \frac{\frac{2}{d} P_1 \epsilon_1 - \frac{2}{d+1} P_1 \epsilon_2}{\frac{2}{d} P_1 - \frac{2}{d+1} P_1} = (d+1)\epsilon_1 - d\epsilon_2 = \frac{\kappa_{\text{rad}}}{8\,(d+2)} \tag{73}$$

$$\frac{\Delta\beta}{\beta} = \frac{c_1 c_2 (\gamma P_2 \epsilon_2 - P_1 \epsilon_1)}{c_1 c_2 (\gamma P_2 - P_1)} = \frac{2P_1\epsilon_2 - P_1\epsilon_1}{P_1} = 2\epsilon_2 - \epsilon_1 = 0 \tag{74}$$

So the engine α elevates or decreases, relative to cumulant, while brake β remains unchanged.

For the loss, $L = P_1 - \gamma P_2 + C$, the shift of the loss would be:

$$\Delta L = P_1\epsilon_1 - \gamma P_2 \epsilon_2 = P_1(\epsilon_1 - 2\epsilon_2) = 0 \tag{75}$$

therefore canceling out the loss shift, unchanged at the 0, as β did.

However, although it stays at 0, it does not mean it would be the optimal point (as Gaussian distribution was), as α has changed while β unchanged, moving the overall target T. (The fixed point moves much less than the target does: the relative shift of $T = \alpha/\beta$ at $\lambda = 1$ is $\kappa_{\text{rad}}/(8(d+2))$, while the shift of the stationary point computed next is smaller by a factor of order $2/d$, because T itself falls steeply as λ increases and a small move of λ absorbs the change in α).

To calculate the new stationary point, made by non-Gaussianity, we would make up the formula using gradient and curvature (with respect to λ):

$$\Delta\lambda^* = -\frac{\Delta L'(1)}{L''(1)}, \tag{76}$$

to calculate how much $\Delta\lambda^*$ is needed to negate the gradient caused by shift.

Using $f' = f \cdot (\ln f)'$ (as $P_1, P_2, \epsilon_1, \epsilon_2 > 0$), we can use:

$$\Delta L'(1) = P_1\epsilon_1((\ln P_1)' + (\ln \epsilon_1)') - \gamma P_2 \epsilon_2((\ln P_2)' + (\ln \epsilon_2)') \tag{77}$$

$$= P_1\epsilon_1 \left(-\frac{d}{2} \cdot \frac{\frac{2}{d}}{1+\frac{2}{d}} - \frac{4}{d+2}\right) - 2P_1\epsilon_2 \left(-\frac{d}{2} \cdot \frac{\frac{1}{d+1}}{1+\frac{1}{d+1}} - \frac{2}{d+2}\right) \tag{78}$$

$$= P_1\epsilon_2 \left(-\frac{2d}{d+2} - \frac{8}{d+2} + \frac{d}{d+2} + \frac{4}{d+2}\right) \tag{79}$$

$$= -P_1\epsilon_2 \left(\frac{d+4}{d+2}\right) \tag{80}$$

For curvature L'', we can differentiate the original formula 2 times:

$$L' = P_1' - \gamma P_2' \tag{81}$$

$$= P_1((\ln P_1)') - \gamma P_2((\ln P_2)') \tag{82}$$

$$= P_1 \left(-\frac{d}{2} \cdot \frac{\frac{2}{d}}{1+\frac{2}{d}\lambda}\right) - \gamma P_2 \left(-\frac{d}{2} \cdot \frac{\frac{1}{d+1}}{1+\frac{1}{d+1}\lambda}\right) \tag{83}$$

$$= -P_1 \left(\frac{d}{d+2\lambda}\right) + \gamma P_2 \left(\frac{d}{2(d+1+\lambda)}\right) \tag{84}$$

$$L''(1) = -P_1' \left(\frac{d}{d+2\lambda}\right) - P_1 \left(-\frac{2d}{(d+2\lambda)^2}\right) \tag{85}$$

$$+ \gamma P_2' \left(\frac{d}{2(d+1+\lambda)}\right) + \gamma P_2 \left(-\frac{d}{2(d+1+\lambda)^2}\right) \tag{86}$$

$$= \frac{P_1}{(d+2)^2} \left(d^2 + 2d - \frac{1}{2}d^2 - d\right) \tag{87}$$

$$= \frac{dP_1}{2(d+2)} \tag{88}$$

Finally, the resulting shift would be:

$$\Delta\lambda^* = -\frac{\Delta L'(1)}{L''(1)} = -\frac{-P_1\epsilon_2\left(\frac{d+4}{d+2}\right)}{\frac{dP_1}{2(d+2)}} = \frac{(d+4)\kappa_{\mathrm{rad}}}{4d(d+2)^2} \tag{89}$$

so a leptokurtic embedding ($\kappa_{\mathrm{rad}} > 0$) settles slightly above the Gaussian prediction, and a platykurtic ($\kappa_{\mathrm{rad}} < 0$) one slightly below.

K.2.4 What does this do in practical situations?

We measured κ_{rad} on the trained embeddings of the runs in Appendix M, from checkpoints at epochs 100, 200, 250, and the final one, on 20,000 training images passed through the network in training mode, so that the embedding is the one the regularizer sees (Appendix O). It is significantly positive in all runs but two, the MLP under the small initialization with SIGReg and with MMD-Gauss, where $|\kappa_{\mathrm{rad}}| < 0.5d$, and it is far larger for BHEP than for the other two objectives: up to about $26d$ for MMD-Gauss and SIGReg, and up to about $90d$ under the default and $1200d$ under the small initialization for BHEP, where in most runs it keeps growing until the end of training. For BHEP these values lie outside the range where the first-order expansion above is quantitative, since $\kappa_{\mathrm{rad}}/d^2$ is not small, so we use the correction only for its sign: a positive cumulant moves the fixed point upward, which is the direction of the overshoot of $\bar{\lambda}$ observed for BHEP, while the invariance term moves $\bar{\lambda}$ as well (Appendix K.1). We do not attempt a quantitative attribution of the overshoot.

Also, the asymmetry between the engine and the brake is visible in how the forces are measured. In our experiments α and β are computed from the empirical push and pull terms P_1, P_2, that is, from $L_z(a)$. Evaluating the same formulas on the measured eigenvalue spectrum instead, which implicitly treats the embedding as Gaussian, leaves β essentially unchanged while α comes out lower, which is the direction of $\Delta\alpha/\alpha > 0$ and $\Delta\beta = 0$ above.

K.2.5 Why is the cumulant so much larger for BHEP?

We offer a hypothesis. A Gaussian-kernel MMD compares the characteristic function of the embedding with that of the Gaussian target, and the bandwidth decides at which frequencies the comparison is made: the frequencies drawn under the kernel concentrate, in high dimension, on a thin shell whose radius is set by the bandwidth. MMD-Gauss integrates over bandwidths and SIGReg over frequencies along every projection, so both compare the two characteristic functions over a whole band of frequencies. BHEP uses a single bandwidth ($\frac{1}{d}$), so it compares them on one shell only.

For a spherically symmetric embedding the characteristic function depends only on the frequency magnitude, and a comparison on one shell is a single scalar constraint on the radial distribution. A non-Gaussian radial spread changes the value on that shell, but a single constraint can be restored by a single knob: raising the mean eigenvalue $\bar{\lambda}$ moves the value on the shell back to its Gaussian level, so the residual BHEP penalty for the spread can become small. MMD-Gauss and SIGReg constrain the whole band, which no single scalar can restore, so the spread itself has to shrink.

If this picture is right, then whatever drives the embedding away from Gaussianity, BHEP does not resist it but trades it for a higher $\bar{\lambda}$, whereas MMD-Gauss and SIGReg push back. This would tie the large cumulant of BHEP and its overshoot of $\bar{\lambda}$ to one cause, and it agrees in sign with the first-order shift derived above. It is a hypothesis: we have not tested it, and nothing here explains what drives the non-Gaussianity in the first place.

K.3 Not-large n, d

Our derivation invokes "large n, d" at several distinct points, and the consequences of violating each differ. We list them separately.

(i) Covariance estimation (n). The determinant identity $\mathbb{E}_x[\exp(-\frac{1}{2}x^\top Mx)] = \det(I + \Sigma M)^{-1/2}$ replaces the batch covariance $\hat{C}_Z$ with its population counterpart. For finite n the sample spectrum is dispersed relative to the true one, so the interaction potentials P_1, P_2, and hence α, β, and the target $T = \alpha/\beta$, carry a finite-sample error that grows as the aspect ratio d/n increases. This affects the measured values of the global forces, not the form of the ODE.

(ii) Rank deficiency ($n \geq d$). When $n < d$, $\hat{C}_Z$ cannot be full-rank, so at least $d - n$ eigenvalues are structurally zero. Since $P_1 = \prod_i (1 + c_1\lambda_i)^{-1/2}$ is a product over the whole spectrum, these zeros contribute a factor of 1 and therefore fail to suppress the engine α.

(iii) Empirical difference of KerJEPA (n). Following Zimmermann et al. (2025), the bias between SIGReg and MMD-Gauss is $|\mathbb{E}[\widehat{\text{MMD-Gauss}}^2_{\kappa_d}(\hat{P}_n) - \widehat{\text{SIGReg}}(\hat{P}_n)]| \leq O(1/n)$ and is thus controlled explicitly rather than assumed away.

(iv) Kernel substitution ($d > 20$). The MMD-Gauss form of Definition 2 relies on the inverse-root approximation to Kummer's function, valid for $d > 20$ (Knop et al., 2020). In Appendix L, we have empirically observed that the error between SIGReg and MMD-Gauss scales with $1/d$.

(v) U-statistic for BHEP (n). Definition 4 is the V-statistic, and the statistic we optimize and log is its U-statistic form (Appendix A), that is, the unbiased estimator of Rustamov (2021): the diagonal terms $j = k$ are constants that do not depend on the representation, and the population push term is an expectation over two independent representations. The two forms differ by a constant and a factor $\frac{n}{n-1}$.

K.4 Misalignment

None of our experiments, toy or CIFAR-10, enforces the alignment between the weights and the input principal components that the derivation assumes. Every encoder starts from a random initialization (Appendix O) that has no relation to the input principal components, and whatever alignment appears is produced by training itself. Here we drop the assumption and track the eigenvalues of C_Z for an arbitrary $W \in \mathbb{R}^{k\times d}$, with $Z = XW$ and $C_Z = W^T C_X W$ as in Appendix G.

Under the Gaussian assumption of Appendix A the regularizer loss depends on C_Z only through its spectrum, so its gradient with respect to the covariance, $G = \partial L/\partial C_Z$, is a symmetric matrix that shares the eigenvectors h_k of C_Z and has eigenvalues $\partial L/\partial\lambda_k$. Gradient flow on W then reads

$$\dot{W} = -\frac{\partial L}{\partial W} = -2C_X W G, \quad \dot{C}_Z = -2(G\mathcal{M} + \mathcal{M}G), \quad \mathcal{M} = W^T C_X^2 W, \tag{90}$$

where G encodes the direction in which the loss wants C_Z to move and $\mathcal{M}$ encodes how a change of W is transmitted to C_Z (the input covariance enters twice, once through $\partial L/\partial W$ and once through $\dot{C}_Z$). For a simple eigenvalue (for a degenerate one, after choosing within its eigenspace the basis that diagonalizes the restriction of $\dot{C}_Z$), $\dot{\lambda}_k = h_k^T \dot{C}_Z h_k = -4\frac{\partial L}{\partial \lambda_k}\mathcal{M}_k$ with $\mathcal{M}_k = h_k^T \mathcal{M} h_k = \|C_X W h_k\|^2 \geq 0$, and inserting $\frac{\partial L}{\partial \lambda_k} = -\frac{1}{2}u_k(\alpha - \beta\lambda_k)$ from Appendix B gives

$$\dot{\lambda}_k = 2\mathcal{M}_k u_k(\alpha - \beta\lambda_k). \tag{91}$$

Under perfect alignment, $Wh_k = s_k v_k$ with v_k the k-th input eigenvector, so $C_X W h_k = \sigma_k^2 s_k v_k$ and $\mathcal{M}_k = \sigma_k^4 s_k^2 = \sigma_k^2 \lambda_k$, which recovers the ODE of Theorem 1.

In general, misalignment enters through the single positive scalar $\rho_k = \mathcal{M}_k/\lambda_k$. Writing $w = Wh_k$ in the input eigenbasis, $w = \sum_j w_j v_j$,

$$\lambda_k = \sum_j \sigma_j^2 w_j^2, \qquad \mathcal{M}_k = \sum_j \sigma_j^4 w_j^2, \qquad \rho_k = \frac{\sum_j \sigma_j^4 w_j^2}{\sum_j \sigma_j^2 w_j^2}, \tag{92}$$

so ρ_k is a weighted mean of the input variances with weights $\sigma_j^2 w_j^2$, and at every instant

$$\min_{j\in\text{supp}(w)} \sigma_j^2 \leq \rho_k \leq \max_{j\in\text{supp}(w)} \sigma_j^2, \tag{93}$$

where $\text{supp}(w)$ is the set of input directions with nonzero weight in h_k. Perfect alignment is the case of a single direction, $\rho_k = \sigma_k^2$.

Since $\mathcal{M}_k \geq 0$, the sign of $\dot{\lambda}_k$, the fixed point, and the engine-brake form do not depend on alignment. Misalignment replaces the constant rate σ_k^2 by a positive, in general time-dependent factor

$\rho_k(t)$ that stays within the range of the input variances on the support of h_k, and does nothing else. What does depend on it is the identification of the k-th emerging feature with the k-th input principal component. Our conclusions do not use that identification: α and β are symmetric functions of the whole spectrum, so relabeling which mode carries which eigenvalue leaves them unchanged.

This assumption is tied to the saturated-invariance assumption of Appendix K.1, because the invariance term is what decides which directions the weights align to. It favors augmentation-invariant directions, so the weights keep re-aligning toward them while the regularizer acts, and the coefficients ρ_k drift until convergence. The augmentation ratios ν_k of the emerging features change along this re-alignment as well, since the invariance term drives them down, so the rates, although always positive and bounded, cannot be predicted in advance. The direction of motion can: it is fixed by the sign of $F(\lambda_k)$, which depends on the spectrum alone.

L Relation between BHEP, MMD-Gauss, and SIGReg

Throughout this paper, we have proposed BHEP as an analytical proxy for SIGReg, motivated by their similar qualitative behavior. This appendix asks how much of the picture the two objectives, together with the intermediate MMD-Gauss, actually share.

Summary They share almost every qualitative feature shown in this paper. For instance:

- Asymmetric decrease of the engine: α for BHEP, the dormant-feature force $F(0)$ for MMD-Gauss and SIGReg
- Decrease over target eigenvalue T
- Stepwise delay over emergence of later features
- Overshooting / Shrinking of early features

The one difference is the path over the mean eigenvalue $\bar{\lambda}$, in two-group spectrum used in main section 4.3.2. Unlike BHEP, which monotonically increases to match $\bar{\lambda} = 1$, MMD-Gauss and SIGReg *jump above* the goal $\bar{\lambda}^* = 1$, then slowly decrease to match 1.

L.1 Exact per-feature forces for all three objectives

A simple closed form of the force is available only for BHEP, because its per-feature forces are elementary products. The other two need extra machinery. For MMD-Gauss, the Gamma representation leaves a t-integral, which we evaluate here by Gauss-Kronrod quadrature[2].

Definition 2 is the form used in every experiment and every proof of this paper. The absorbed form of Appendix I is listed alongside for comparison: Tables 3 and 4 show how large the difference between the two orders of approximation is on the same ladder, up to 7% at $d = 32$ and about 2% at $d = 128$, with the absorbed form settling at $\bar{\lambda} = 1 - 1/d$ instead of 1. For SIGReg, no closed form is available at all, so we replace the random projections by nested Beta expectations, which reduce the spherical average exactly.

All forces are evaluated on the two-group spectrum of Section 4.3.2: m features have emerged and saturated at a common level, while the other $d - m$ features stay at 0. $\Phi(\lambda_A)$ denotes the force on one selected feature (the probe) whose eigenvalue is moved to λ_A, while the m' other saturated features are held fixed ($m' = m - 1$ when the probe is one of the emerged features, $m' = m$ when the probe is dormant) at their common level, written λ_{-A}, and the global potentials P_1, P_2 (and their per-slice versions $P_1(t), P_2(t)$) are taken over this full spectrum.

Throughout this appendix the force is normalized as in Theorem 1, $\Phi(\lambda_A) = u_A(\alpha - \beta\lambda_A) = -2\,\partial L/\partial\lambda_A$, so that $\dot{\lambda}_A = 2\sigma_A^2\lambda_A\Phi(\lambda_A)$ for all three objectives, and α, β carry the scale of Theorem 1.

[2]The experiments use a fixed Gauss-Laguerre rule (Section O), since the gradient has to propagate through the quadrature. In this appendix no gradient is needed, so we use an adaptive rule and control the error explicitly.

The exact formulas are:

$$\Phi_{\text{BHEP}}(\lambda_A) = \frac{c_1 P_1}{1 + c_1\lambda_A} - \gamma \frac{c_2 P_2}{1 + c_2\lambda_A}$$

$$\Phi_{\text{MMD-Gauss (Appendix I)}}(\lambda_A) = \frac{1}{\sqrt{\pi}} \int_0^\infty t^{1/2} e^{-t} \cdot s \left[\frac{4P_1(t)}{1 + 4st\lambda_A} - \frac{\sqrt{2} P_2^{\text{abs}}(t)}{1 + st\lambda_A} \right] dt$$

$$\Phi_{\text{MMD-Gauss (Def. 2)}}(\lambda_A) = \frac{1}{\sqrt{\pi}} \int_0^\infty t^{1/2} e^{-t} \cdot 4s \left[\frac{P_1(t)}{1 + 4st\lambda_A} - \frac{\tilde{P}_2(t)}{1 + 2st(\lambda_A + 1)} \right] dt$$

$$\Phi_{\text{SIGReg}}(\lambda_A) = 2\,\mathbb{E}_{\xi \sim S^{d-1}}[g(v)\,\xi_A^2], \quad v = \sum_j \lambda_j \xi_j^2,$$

$$g(v) = (1 + 2v)^{-3/2} - (2 + v)^{-3/2},$$

where $P_j = \prod_k (1 + c_j\lambda_k)^{-1/2}$ with $c_1 = 2/d$, $c_2 = 1/(d+1)$ and $\gamma = 2(1 + \frac{1}{d})^{-d/2}$ for BHEP, $P_1(t) = \prod_k (1 + 4st\lambda_k)^{-1/2}$ and $P_2(t) = \prod_k (1 + c_2(t)\lambda_k)^{-1/2}$ are the per-slice products of Appendix H.3, $\tilde{P}_2(t) = \prod_k (1 + 2st(1 + \lambda_k))^{-1/2} = \gamma_t P_2(t)/2$ absorbs the Gaussian factor of the pull term, $P_2^{\text{abs}}(t) = \prod_k (1 + st\lambda_k)^{-1/2}$ is the pull product of the absorbed form (Appendix I), and for SIGReg the random projection ξ is uniform on the sphere, so that ξ_j^2 is the weight with which the variance λ_j enters the projected variance v (Appendix C).

Reduction of the SIGReg average. With ξ uniform on S^{d-1}, the squared coordinates $(\xi_1^2, \ldots, \xi_d^2)$ follow a $\text{Dirichlet}(\frac{1}{2}, \ldots, \frac{1}{2})$ law, and grouping coordinates preserves the Dirichlet family. On the two-group spectrum, v therefore depends on ξ only through two scalars: the share of the probe, $p = \xi_A^2$, and the fraction q of the remaining mass $1 - p$ that falls on the m' other saturated features, all of which sit at the fixed level λ_{-A}. These are independent, with $p \sim \text{Beta}(\frac{1}{2}, \frac{d-1}{2})$ and $q \sim \text{Beta}(\frac{m'}{2}, \frac{d-1-m'}{2})$. The dormant features carry $\lambda = 0$ and drop out, so $v = p\lambda_A + (1-p)q\lambda_{-A}$ and

$$\Phi_{\text{SIGReg}}(\lambda_A) = 2\,\mathbb{E}_{p,q}\big[g\big(p\lambda_A + (1-p)q\lambda_{-A}\big)\,p\big]\,.$$

Only λ_A is differentiated: λ_{-A} enters as a parameter. For $m' = 0$ there are no other saturated features and the q-average is absent, and for $m' = d - 1$ it is $q \equiv 1$.

Per-feature quantities. From these forces we compute:

- Target T: $\lambda_{-A} = T$ where $\Phi(T) = 0$ (the probe is itself one of the m emerged features, so the other $m' = m - 1$ saturated eigenvalues inside P_1, P_2 are also set to T)
- Eigenvalue mean $\bar{\lambda} = \frac{m}{d} \cdot T$
- Brake β_A: $-\Phi'(T) = -\frac{-\Phi(T+2h) + 8\Phi(T+h) - 8\Phi(T-h) + \Phi(T-2h)}{12h}, \quad h = 10^{-4} \times \max(1, T)$
- Engine α_A: $\Phi(T) = \alpha_A - \beta_A \cdot T = 0 \to \alpha_A = \beta_A \cdot T$
- Dormant-feature force $\alpha_0 = \Phi(0)$, with the m emerged features held at T. This is the quantity written $F(0)$ in Section 4.4: since the state factor equals 1 at $\lambda_A = 0$, $\Phi(0) = \alpha$ exactly for BHEP, $\Phi(0) = F(0)$ for MMD-Gauss, and the corresponding spherical average for SIGReg.

Three remarks on these conventions. First, we report the engine and the brake at two different states, the saturated level T and the not-yet-emerged level 0, because the force felt by a feature depends on its current eigenvalue. For BHEP, this state dependence is captured in closed form by the factor u_k of Theorem 1, while for the other two objectives no such closed decomposition is available, so the tangent construction above is the uniform choice across the three.

Second, we do not report a brake for the not-yet-emerged features (β_0): near $\lambda = 0$ the brake term contributes only at order λ, and the finite difference there mainly measures the curvature of the engine rather than a meaningful brake.

Third, note that α_A, β_A differ from the theoretical α, β of Theorem 1 (whose coefficients include u_k): the finite difference also passes through the probe's own λ inside the global potentials P_1, P_2,

Table 3: **Per-feature quantities of the three objectives** ($d = 32$). All four rows report the force on a single feature under the same convention, so absolute values are comparable across objectives. Computation: BHEP in closed form, the two MMD-Gauss rows by the same t-integral for Definition 2 and for the absorbed form of Appendix I, SIGReg by exact reduction of the spherical average.

	m	1	2	3	5	10	24	32
SIGReg	α_A	2.20e-3	2.37e-3	2.26e-3	1.93e-3	1.32e-3	6.80e-4	5.31e-4
	α_0	1.30e-2	7.21e-3	4.86e-3	2.87e-3	1.38e-3	5.53e-4	–
	β_A	6.69e-5	1.38e-4	1.98e-4	2.87e-4	4.03e-4	5.08e-4	5.31e-4
	T	32.873	17.173	11.409	6.733	3.284	1.339	1.000
	$\bar{\lambda}$	1.027	1.073	1.070	1.052	1.026	1.004	1.000
MMD-Gauss (Appendix I)	α_A	2.21e-3	2.37e-3	2.26e-3	1.92e-3	1.31e-3	6.73e-4	5.24e-4
	α_0	1.34e-2	7.33e-3	4.90e-3	2.87e-3	1.37e-3	5.47e-4	–
	β_A	6.91e-5	1.42e-4	2.04e-4	2.95e-4	4.14e-4	5.19e-4	5.42e-4
	T	31.965	16.664	11.059	6.520	3.177	1.295	0.967
	$\bar{\lambda}$	0.999	1.041	1.037	1.019	0.993	0.972	0.967
MMD-Gauss (Definition 2)	α_A	2.16e-3	2.34e-3	2.25e-3	1.92e-3	1.32e-3	6.75e-4	5.26e-4
	α_0	1.36e-2	7.41e-3	4.95e-3	2.89e-3	1.38e-3	5.50e-4	–
	β_A	6.43e-5	1.35e-4	1.95e-4	2.84e-4	4.00e-4	5.04e-4	5.26e-4
	T	33.640	17.399	11.509	6.766	3.290	1.340	1.000
	$\bar{\lambda}$	1.051	1.087	1.079	1.057	1.028	1.005	1.000
BHEP	α_A	4.70e-3	4.75e-3	4.42e-3	3.68e-3	2.48e-3	1.26e-3	9.84e-4
	α_0	1.20e-2	7.96e-3	5.95e-3	3.96e-3	2.16e-3	9.48e-4	–
	β_A	2.20e-4	3.67e-4	4.77e-4	6.23e-4	8.00e-4	9.51e-4	9.84e-4
	T	21.343	12.921	9.257	5.903	3.095	1.327	1.000
	$\bar{\lambda}$	0.667	0.808	0.868	0.922	0.967	0.995	1.000

which makes both come out larger by a common factor. For BHEP, the factor is explicit: differentiating the ODE gives $\alpha_A = \frac{3}{2}\, u_k\, \alpha$ and $\beta_A = \frac{3}{2}\, u_k\, \beta$, because the probe's own factor $(1 + c\lambda_A)^{-1/2}$ inside P_1, P_2 raises the exponent of the global force by $1/2$. For MMD-Gauss the force is not affine in λ_k, so no closed-form factor is available and we do not separate the two contributions. The ratio $T = \alpha_A/\beta_A$ cancels this factor, and α_0 involves no derivative, so both are convention-free, and the phenomena discussed in this section (asymmetric decrease of α, collapsing target T, stepwise delay, overshoot/shrinking) are unaffected.

Note the small rise of α_A at $m = 2$, shared by all three objectives. For BHEP it is the state factor: $\alpha_A = \frac{3}{2} u_k \alpha$ and $u_k(T)$ grows as T drops sharply between $m = 1$ and $m = 2$, faster than α itself decreases. We expect the same mechanism for the other two, whose weights $u_{t,k}$ play the same role.

Reading the tables confirms the summary. For all three objectives, the dormant engine α_0 collapses by more than an order of magnitude across the sweep, while the brake β_A rises by less than one: this is the engine-brake asymmetry. For BHEP the rise of β_A is mostly the state factor $u_k(T)$ of Theorem 1, which grows from about 0.25 to 0.9 as T falls, while β itself stays within 30%. We expect the same mechanism behind the larger rise for SIGReg and MMD-Gauss, but cannot separate the two factors there. The target T likewise collapses for all three, with SIGReg and MMD-Gauss on nearly the same ladder and BHEP starting lower but joining them at moderate m. The ratio α_0/T of the SIGReg and MMD-Gauss rows stays within $[1.00, 1.07]$ times $2s^2/3^{3/2}$ on every rung of Table 4, the band drawn in Figure 10. The remaining two items of the summary are consequences of these two decreases: since the emergence rate of a dormant feature is governed by α_0, its collapse produces the stepwise delay, and the collapse of T means that early features first saturate at a high target and are then dragged down as later features emerge, which is the per-feature overshoot-then-shrink. Finally, the $\bar{\lambda}$ row records the one difference announced above. As m increases, BHEP approaches 1 from below, while SIGReg and MMD-Gauss rise above their fixed point at small m and come back down to it.

Between Table 3 ($d = 32$) and Table 4 ($d = 128$) the relative gap between SIGReg and MMD-Gauss, in either form, shrinks by a factor of about 4 in every tabulated quantity, the same factor by

Table 4: **Same construction and conventions as Table 3** ($d = 128$). The qualitative structure replicates: the engine collapses while the brake varies mildly, the targets of SIGReg and MMD-Gauss share nearly the same ladder with BHEP joining at moderate m, and $\bar{\lambda}$ approaches 1 from above for SIGReg and MMD-Gauss but from below for BHEP.

	m	1	2	4	8	32	96	128
SIGReg	α_A	5.32e-4	5.71e-4	5.05e-4	3.65e-4	1.27e-4	4.58e-5	3.47e-5
	α_0	3.20e-3	1.76e-3	8.76e-4	4.20e-4	9.83e-5	3.21e-5	–
	β_A	3.99e-6	8.20e-6	1.47e-5	2.18e-5	3.14e-5	3.43e-5	3.47e-5
	T	133.469	69.614	34.348	16.730	4.044	1.335	1.000
	$\bar{\lambda}$	1.043	1.088	1.073	1.046	1.011	1.001	1.000
MMD-Gauss (Appendix I)	α_A	5.33e-4	5.71e-4	5.05e-4	3.65e-4	1.26e-4	4.56e-5	3.46e-5
	α_0	3.22e-3	1.77e-3	8.77e-4	4.19e-4	9.81e-5	3.20e-5	–
	β_A	4.02e-6	8.26e-6	1.48e-5	2.20e-5	3.15e-5	3.45e-5	3.49e-5
	T	132.575	69.113	34.088	16.600	4.012	1.324	0.992
	$\bar{\lambda}$	1.036	1.080	1.065	1.037	1.003	0.993	0.992
MMD-Gauss (Definition 2)	α_A	5.30e-4	5.70e-4	5.04e-4	3.65e-4	1.27e-4	4.57e-5	3.46e-5
	α_0	3.24e-3	1.77e-3	8.79e-4	4.20e-4	9.82e-5	3.21e-5	–
	β_A	3.95e-6	8.16e-6	1.47e-5	2.18e-5	3.13e-5	3.42e-5	3.46e-5
	T	134.242	69.845	34.406	16.743	4.044	1.335	1.000
	$\bar{\lambda}$	1.049	1.091	1.075	1.046	1.011	1.001	1.000
BHEP	α_A	1.15e-3	1.16e-3	9.87e-4	7.00e-4	2.41e-4	8.69e-5	6.58e-5
	α_0	2.94e-3	1.94e-3	1.16e-3	6.44e-4	1.76e-4	5.97e-5	–
	β_A	1.37e-5	2.28e-5	3.47e-5	4.63e-5	6.09e-5	6.52e-5	6.58e-5
	T	84.183	51.005	28.478	15.106	3.954	1.332	1.000
	$\bar{\lambda}$	0.658	0.797	0.890	0.944	0.988	0.999	1.000

which d grows. Thus, beyond the fixed-point deviation $1-1/d$ of the absorbed form Appendix I, the remaining differences between the objectives also decrease in proportion as d grows, at least over the range we computed ($d = 32$ to 128). We do not have a proof of this rate.

L.2 Monotone decrease of the engine α and the target T

The statistics above are snapshots at fixed m. To turn them into a statement about the whole trajectory, Appendix Q proves that along the equilibrium ladder, on which m features are saturated at the self-consistent target $T(m)$ and the remaining ones are dormant, both the dormant engine $\alpha_0(m)$ and the target $T(m)$ are strictly decreasing in m. The two decreases come from the same object: α_0 is a rung of the balance integral one step below the equilibrium, and T is the equilibrium level itself, so neither is the cause of the other. The proof is carried out for the MMD-Gauss of Definition 2 in the limit $d \to \infty$, where the balance condition reduces to a one-parameter family, and it covers every $m \geq 1$: the certified constants of Table 6 settle $\frac{3}{2} \leq \theta \leq 5$ and an elementary tail bound settles the rest. For BHEP the closed-form ladder of Appendix F shows the same decrease over $T(m)$ directly, and for SIGReg we rely on the exact numerical evaluation of Table 3, for which we have no proof.

The phenomena of this appendix follow from these two monotonicities. A dormant feature emerges at a rate proportional to α_0, so each new feature starts under a weaker engine than the previous one, which is the stepwise delay. A saturated feature sits at T, so as T decreases with further emergence, the early features that saturated at a high target are pulled back down, which is the overshoot followed by shrinking. We claim only the direction here: the proof gives strict decrease, not a rate, and the size of the delays reported in the main text comes from the rate $1/\alpha_0$ of Appendix E, not from the proof.

M EXPERIMENTS ACROSS DATASETS AND ARCHITECTURES

To check that the reported dynamics are not an artifact of one dataset or one backbone, we repeat the training runs over five settings (CIFAR-10 (Krizhevsky, 2009) with an MLP, ViT-S (Dosovitskiy et al., 2021), and ResNet-34, and ResNet-18 (He et al., 2016) on STL-10 (Coates et al., 2011) and CelebA (Liu et al., 2015)), two initializations (default and small), and all three objectives (SIGReg, MMD-Gauss with Gauss-Laguerre quadrature, BHEP), for a total of 30 runs. All remaining hyperparameters follow the template config of the released code. Figures 8 and 9 show the full covariance eigenvalue trajectories of every run.

Three observations. First, the qualitative picture of the main text carries over: features emerge sequentially and the emerged band saturates, for every backbone, dataset, and objective. Second, along the dataset axis the survivor count follows the invariance level in the same direction as along the architecture axis (Table 2). Third, on CelebA the top band descends over training while later features continue to emerge, which is the closest any of our real-data runs comes to the shrinking phase of the two-phase description.

M.1 THE DORMANT-FEATURE FORCE $F(0)$ AND THE TARGET T MOVE TOGETHER

Observation. For MMD-Gauss the two quantities that survive the bandwidth integration, the dormant-feature force $F(0)$ and the root target T, are logged in every run of this appendix, and both are tabulated on the equilibrium ladder of Tables 3 and 4. In both places their ratio is nearly constant although each of the two falls by more than an order of magnitude: on the rungs of Table 4 it stays within $[1.00,\ 1.07]$ times $2s^2/3^{3/2}$ for SIGReg and both MMD-Gauss forms, and in the twelve runs of Figure 10 it settles after the first few epochs between 1.03 and 1.10 times the same value with about 1% after the first 20 epochs. The reference value $2s^2/3^{3/2}$ is not fitted: it is the value the ratio takes at the isotropic fixed point, which we derive next.

The constant. For BHEP the engine, the brake, and the target are tied by $\alpha/T = \beta$. MMD-Gauss has no single brake, but the same ratio has a definite value at the isotropic state $\lambda_i \equiv 1$, which is visible slice by slice (Appendix H). At $\lambda_i = 1$ for all i, $\tilde{P}_2(t) = P_1(t)$ for every t (Appendix H.5), hence $\alpha_t = \beta_t$ (Eq. 41), and

$$F(\lambda_k) = \frac{1}{\sqrt{\pi}} \int_0^\infty t^{1/2} e^{-t}\, u_{t,k}\, (\alpha_t - \beta_t \lambda_k)\, dt = (1 - \lambda_k)\, \frac{1}{\sqrt{\pi}} \int_0^\infty t^{1/2} e^{-t}\, u_{t,k}\, \beta_t\, dt. \tag{94}$$

The root is therefore $T = 1$, and the dormant-feature force is the bandwidth average of the brake,

$$\begin{aligned} \frac{F(0)}{T} = F(0) &= \frac{1}{\sqrt{\pi}} \int_0^\infty t^{1/2} e^{-t}\, u_{t,0}\, \beta_t\, dt = \frac{1}{\sqrt{\pi}} \int_0^\infty t^{1/2} e^{-t}\, \frac{8s^2 t\, P_1(t)}{1 + 2st}\, dt \\ &\approx \frac{8s^2}{\sqrt{\pi}} \int_0^\infty t^{3/2} e^{-3t}\, dt = \frac{2s^2}{3^{3/2}} = \hat{\beta}, \end{aligned} \tag{95}$$

where the second line keeps the leading order in $1/d$ ($P_1(t) = (1 + 4st)^{-d/2} \to e^{-2t}$). The absorbed form of Appendix I reaches the same value after the bandwidth integration, its slices not being individually balanced, and a direct evaluation on the uniform spectrum at the fixed point of either form agrees with $\hat{\beta}$ to 0.4% at $d = 128$. So $\hat{\beta}$ plays for MMD-Gauss the role that β plays for BHEP at the fixed point, and it is the constant that scales the right axis of the right panel of Figure 4.

What is not derived. The derivation holds at the fixed point only. The ladder value is a Gaussian, equilibrium quantity and the logged ratio is a sample quantity of a trained network, yet both stay near $\hat{\beta}$ along the whole path, where the tangent approximation above no longer applies. We read this as an empirical regularity: along these runs a single force $F(0)$ carries the same information as the target, as the engine α does for BHEP. We do not have a derivation of this constancy away from the fixed point.

N LIMITATIONS

While our framework provides the first closed-form analysis of LeJEPA's feature learning dynamics, several theoretical and practical limitations should be acknowledged. In particular, given the gap between these idealized assumptions and large-scale training, the engine-brake decomposition should

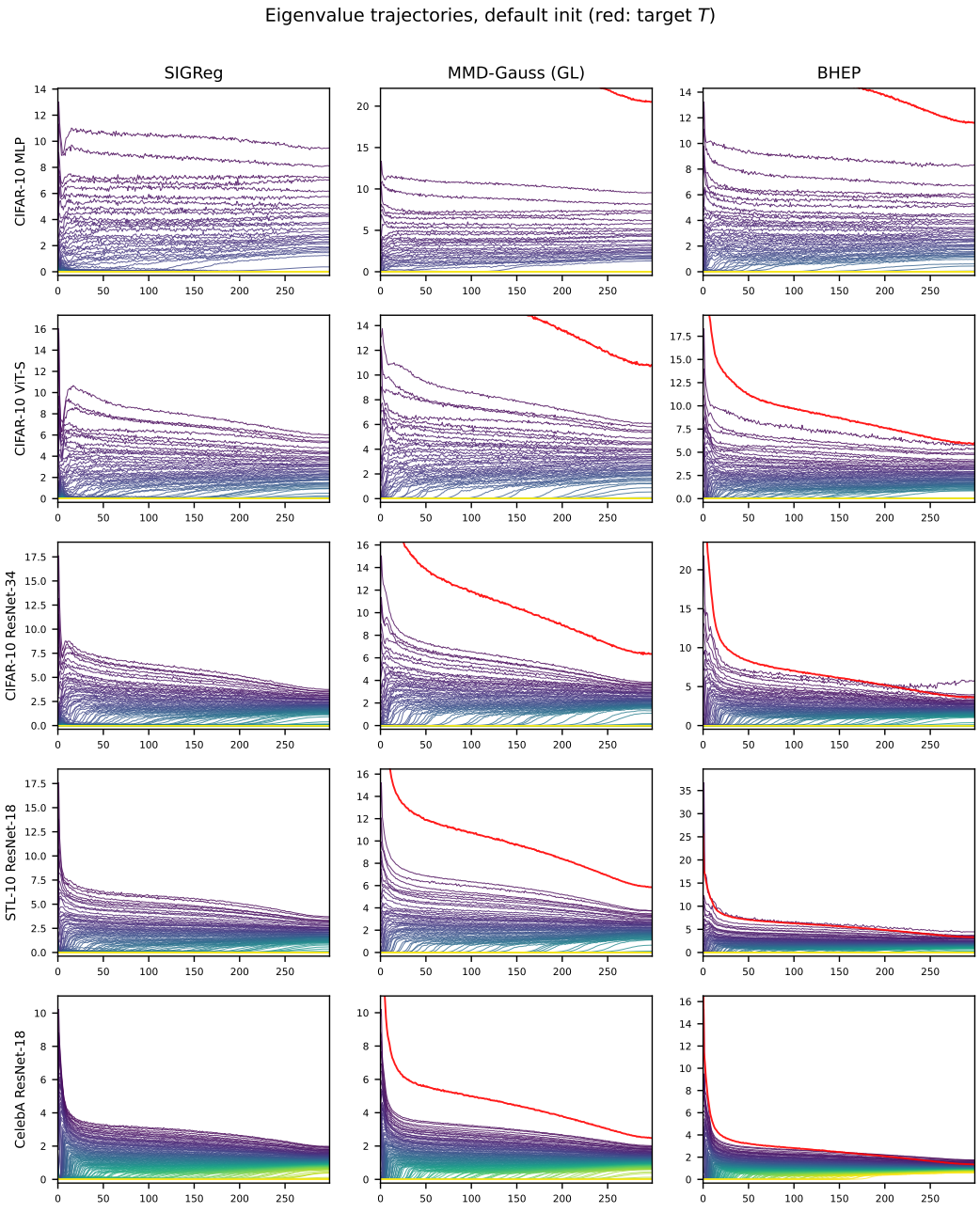


Figure 8: Eigenvalue trajectories for the five settings (rows) and three objectives (columns) under the default initialization. Colors index the eigenvalue rank, and red line indicates the target T, same as that of Figure 3.

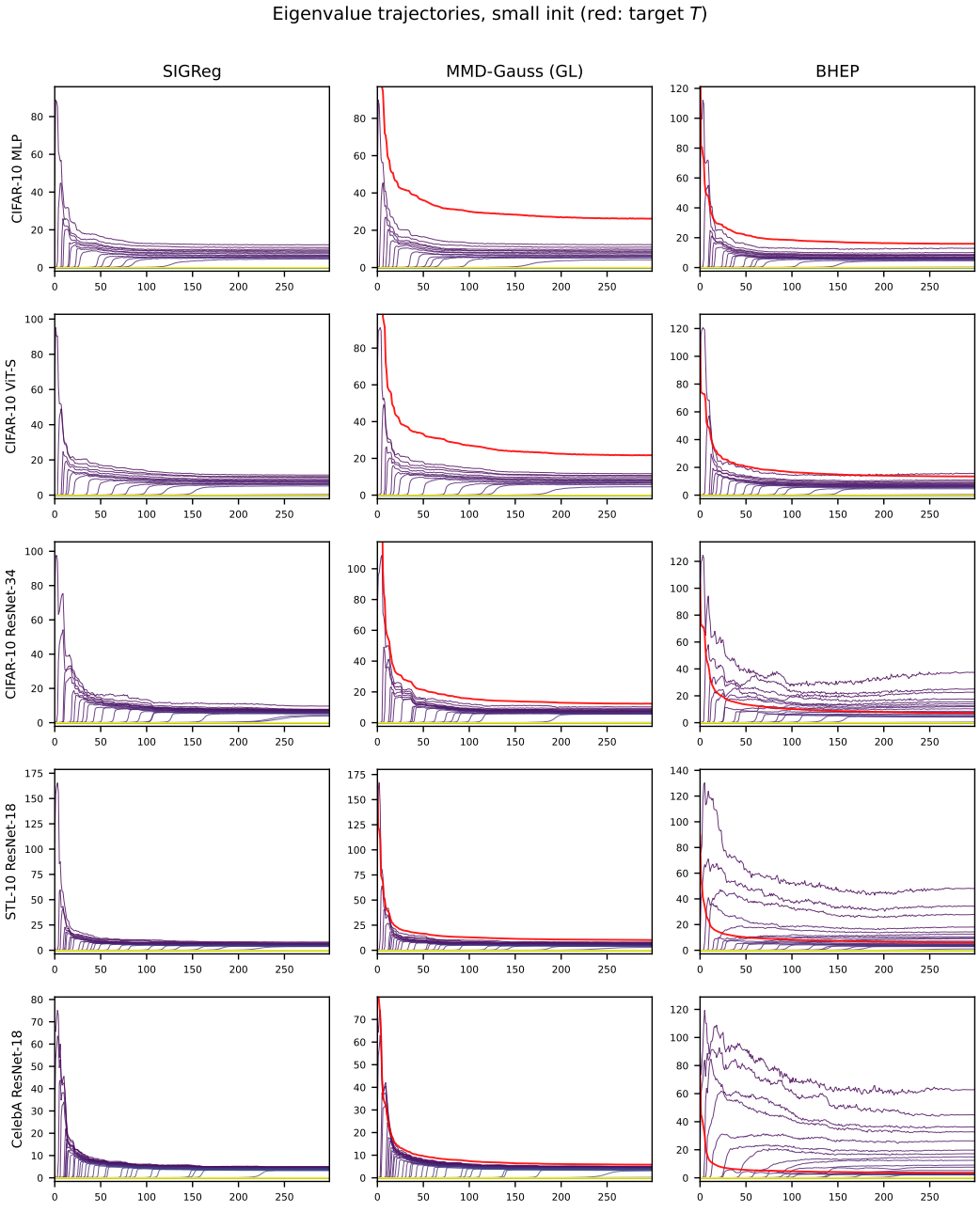


Figure 9: Same as Figure 8 under the small initialization, where the stepwise emergence of features is most visible.

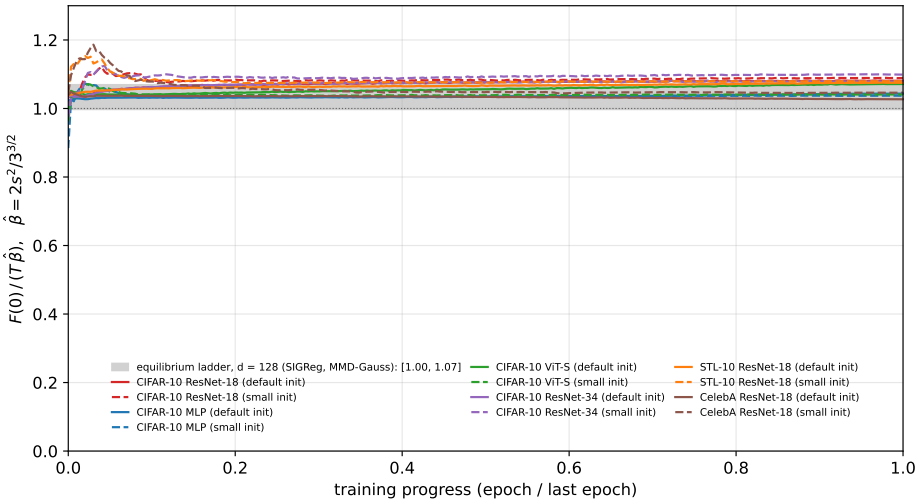


Figure 10: $F(0)/T$ of MMD-Gauss along training, divided by $\hat{\beta} = 2s^2/3^{3/2}$ (Eq. 95), for the ten MMD-Gauss runs of Figures 8 and 9 and the two CIFAR-10 ResNet-18 runs of Figure 4 (solid: default, dashed: small initialization). The gray band is the range of the same ratio on the equilibrium ladder of Table 4.

be read as a plausible mechanistic explanation of the expansion, overshoot, and shrinking dynamics observed under SIGReg, rather than as a definitive or complete account.

Theoretical Assumptions. To ensure mathematical tractability and derive the macroscopic ODEs, our analysis relies on several idealized assumptions:

- We assume high dimensionality d and a sufficiently large sample size n ($n > d$) to support the asymptotic expansions.
- Following the linear network framework, we assume that the weight matrix W structurally aligns with the top-d principal components of the input covariance matrix during training.
- The projected input features are assumed to follow a multivariate Gaussian distribution, which provides a necessary proxy for deriving the energy landscape.
- We assume that invariance terms (e.g., $\|z - z'\|_2^2$) have already saturated, allowing the model to focus purely on the regularization dynamics of non-trivial features.

We have discussed the impact over each assumption violation in Appendix K.

Computational Complexity. The theoretical proxies scale differently from SIGReg in the batch size. While the original SIGReg regularizer is highly efficient at $\mathcal{O}(ndp)$ through stochastic projections, the BHEP and MMD-Gauss formulations used for our analysis require pairwise comparisons, resulting in $\mathcal{O}(n^2 d)$ complexity. Furthermore, evaluating the MMD-Gauss integral via Gauss-Laguerre quadrature adds a knot-dependent term, for a total complexity of $\mathcal{O}(n^2(d + k))$. At the batch size and embedding dimension of our runs ($n = 256$, $d = 128$) this term is negligible: on the same GPU the wall-clock time and the peak memory of the three objectives agreed within run-to-run variation (about 1.4 hours and 2.3 to 2.6 GB for ResNet-18 on CIFAR-10), so the quadratic scaling would matter only for the much larger batches used at scale or for tracking the dynamics there.

Structural Gap. Lastly, our proxy-based approach introduces minor structural discrepancies. The fundamental difference between BHEP and SIGReg prevents the derivation of certain closed-form statistics and direct application of α, β terms to MMD-Gauss. Nevertheless, we provide various non-closed derivations and statistics achievable via the Gamma integration framework in Appendix H, Appendix L, and appendix M.1.

O EXPERIMENTAL DETAILS

Every figure and table of the paper is produced by a script of the released code that carries the same number, with the exact hyperparameters. This appendix records the settings needed to read the results. Every training curve is a single run with a fixed seed: the figures illustrate the predicted dynamics and are not statistical validation.

O.1 TOY EXPERIMENTS

Common setting. The encoder is a single bias-free linear layer $z = W^\top x$, $W \in \mathbb{R}^{m\times d}$ with $m = 100$, initialized as $W_{ij} \sim \mathcal{N}(0, 0.001^2)$ so that every feature starts from $\lambda_k \approx 0$. The input is block-structured: the m coordinates are split into blocks, each block is a rank-one signal whose standard deviation σ_k differs across blocks. The two views are the same signal with independent noise. The loss is the mean squared invariance term between the two views plus $\lambda_{\text{reg}}\mathcal{L}_{\text{reg}}$, minimized by full-batch SGD without momentum. The regularizers are SIGReg with 256 random projections and a 17-point trapezoid rule on $t \in [0, 3]$ (the integrand is even in t), BHEP with $\beta_{bw}^2 = 1/d$, and MMD-Gauss in the form of Definition 2 with $K = 10$ Gauss-Laguerre knots. Unless stated otherwise $\lambda_{\text{reg}} = 1$ for SIGReg and $\sqrt{2\pi} \approx 2.5$ for BHEP and MMD-Gauss, which matches their scale to that of SIGReg (Appendix C).

Per-figure settings.

- **Figure 1** (BHEP): $d = 32$, $n = 400$, 50 blocks with σ_k linearly spaced in $[1, 2]$, view noise of standard deviation 0.01 on every coordinate, learning rate 1.0, $1{,}000$ steps.
- **Figure 3, top row** (all three objectives): as above with $n = 1{,}000$, learning rate 1.2 for SIGReg and MMD-Gauss and 0.8 for BHEP, chosen so that the three runs saturate on the same horizon. The target line is $T = \alpha/\beta$ for BHEP and the root of the per-feature force for MMD-Gauss, both computed from the batch potentials of the current step.
- **Figure 5** (BHEP): $d = 10$, $n = 1{,}000$, 20 blocks with the geometric profile $\sigma_k = 1.2^{-k}$, no view noise, learning rate 0.67 for $5{,}000$ steps with a linear warmup from 1% of the learning rate over the first $1{,}000$ steps, so that the features saturate one at a time. The theory lines are the ladder of Appendix F.1.
- **Figure 6** (BHEP): $d = 10$, $n = 1{,}000$, 10 blocks with σ_k linearly spaced in $[1, 3]$, and instead of isotropic view noise an augmentation of standard deviation $a_k = 0.2$ along the signal direction of each block, so that the augmentation ratio $\nu_k = a_k^2/(\sigma_k^2 + a_k^2)$ differs across features. Learning rate 0.033, $5{,}000$ steps, $\lambda_{\text{reg}} = 1$.
- **Figure 7** (all three objectives): the setting of Figure 3 with the augmentation of the previous item, its relative strength $r_k = a_k/\sigma_k$ spread from 0.2 on the largest-variance block to 0.4 on the smallest, multiplied by $g(t) = e^{-3t/t_{\max}}$ over the $1{,}000$ steps, with the learning rates of Figure 3.

O.2 REAL-DATA EXPERIMENTS

Common setting. All runs use the LeJEPA recipe of the EB-JEPA library (Terver et al., 2026) with the SIGReg term replaced by the objective under study and everything else unchanged: a projector with hidden width $2{,}048$ and output dimension $d = 128$, batch size 256, 300 epochs, weight decay 10^{-4}, the library's warmup-cosine learning-rate schedule with 10 warmup epochs, its default augmentations, and mixed precision. The regularizers are SIGReg with $4{,}096$ projections and the library's 10 knots at $\lambda_{\text{reg}} = 10$, and BHEP ($\beta_{bw}^2 = 1/d$) and MMD-Gauss (Definition 2, $K = 100$ Gauss-Laguerre knots) at $\lambda_{\text{reg}} = 10\sqrt{2\pi} \approx 25.066$. Two initializations are used throughout.

- **Default**: the library defaults, learning rate 0.3 and the PyTorch default weight initialization. The projector keeps the last Batch Normalization layer of the library, which sits before the output layer.
- **Small**: to stretch the stepwise emergence, which otherwise happens within the first epoch, the learning rate is reduced to 0.003, every Conv and Linear weight is initialized with standard deviation 0.005, and the last Batch Normalization of the projector is removed so that the embedding variance grows only through the regularizer.

Every run logs, once per epoch, the eigenvalues of the embedding covariance and the global forces: for BHEP the engine α, the brake β, and $T = \alpha/\beta$ composed from the epoch means of the push and pull terms, and for MMD-Gauss the dormant-feature force $F(0)$ and the root target T composed from the epoch means of the slice potentials (Appendix H.4). All pairwise terms are U-statistics ($i \neq j$), in the loss and in the logged forces.

Settings and runs. Table 5 lists the six dataset and backbone settings. Each is trained with the three objectives and the two initializations, so every setting contributes six runs and no other training runs enter the paper. The CIFAR-10 ResNet-18 runs are those of Figures 3 and 4, and the other five settings form the 30-run sweep of Figures 8 and 9. The tables and the remaining figures read these runs: Table 1 reads the default-initialization BHEP and SIGReg runs of the four CIFAR-10 backbones at the last epoch, Table 2 the default-initialization ResNet-18 runs of the three datasets, with surviving directions defined as eigenvalues above 0.5, and Figure 10 the logged $F(0)$ and T of the twelve MMD-Gauss runs. The radial cumulant κ_{rad} of Appendix K.2 is measured on the saved checkpoints of these same runs (epochs 100, 200, 250, and the final one) from the embeddings of $20{,}000$ training images under the training augmentation, with the network in training mode (BatchNorm on batch statistics, batches of 256) so that the embedding is the one the regularizer sees, and is compared with the value of a Gaussian sample of the same size and spectrum.

Table 5: **Real-data settings.** Each row is trained with SIGReg, MMD-Gauss, and BHEP under the default and the small initialization.

dataset	backbone	image size	shown in
CIFAR-10	ResNet-18	32	Figures 3, 4, Tables 1, 2
CIFAR-10	3-layer MLP (width 2,048)	32	Figures 8, 9, Table 1
CIFAR-10	ResNet-34	32	same
CIFAR-10	ViT-S (patch size 2)	32	same
STL-10	ResNet-18	96	Figures 8, 9, Table 2
CelebA	ResNet-18	128	same

For STL-10 the 96×96 images are used at their native size, and for CelebA the augmentation takes random resized crops of 128×128 directly from the original images, with the ResNet-18 stem adapted to the image size.

O.3 Calculation-only results

Figure 2 evaluates the closed-form BHEP coefficients on the ladder of Appendix F.1 at $d = 128$. Tables 3 and 4 are computed as described in Appendix L: BHEP in closed form, MMD-Gauss by adaptive quadrature of the t-integral with a root search for T, and SIGReg by the nested Beta reduction of the spherical average. Table 6 is produced by interval arithmetic with outward rounding, and the released script re-derives every enclosure of the table and reports whether all checks pass. The band of Figure 10 is the range of $F(0)/T$ over the rungs of Table 4.

P Some Notations

Symbols are grouped by where they are introduced. Several letters are reused with different meanings (β, σ, t, γ, s, T), and each such case is flagged in its group. The notation of the α-decrease proof is self-contained in Appendix Q and is not repeated here.

Model and spectrum (Sections 3 and 4).

$d \in \mathbb{N}$	(embedding dimension)
$n \in \mathbb{N}$	(batch size)
$x_i \in \mathbb{R}^m$, $z_i \in \mathbb{R}^d$	(i-th input and its embedding, $Z = XW$)
$W \in \mathbb{R}^{m \times d}$	(linear encoder, $W = V^{(\le d)} S Q^T$ under the alignment assumption, Appendix G)
C_X, $C_Z = \frac{1}{n} Z^T Z$	(input and embedding covariance matrices)
σ_k^2	(k-th eigenvalue of C_X, the input variance of feature k)
s_k	(learnable scale of feature k, diagonal of S)
$\lambda_k = \sigma_k^2 s_k^2$	(k-th eigenvalue of C_Z; "feature k" means this eigen-direction)
$\bar{\lambda} = \frac{1}{d} \sum_i \lambda_i$	(mean eigenvalue)
m	(number of emerged and saturated features, Section 4.3.2)
$\lambda^* = 1$	(fixed point of every feature for BHEP and MMD-Gauss, Section 4.4, Appendix H.5; the absorbed form of Appendix I settles at $1 - 1/d$ instead)

BHEP dynamics (Theorem 1, Definition 5).

$c_1 = \frac{2}{d}$, $c_2 = \frac{1}{d+1}$	(structural constants of the BHEP kernel at $\beta_{bw}^2 = 1/d$)
$\gamma = 2(1 + \frac{1}{d})^{-d/2} \approx \frac{2}{\sqrt{e}}$	(Gaussian factor of the pull term)
$P_1 = \prod_i (1 + c_1 \lambda_i)^{-1/2}$	(push potential, from the pairwise term)
$P_2 = \prod_i (1 + c_2 \lambda_i)^{-1/2}$	(pull potential, from the target term)
$\alpha = c_1 P_1 - \gamma c_2 P_2$	(engine: the force on a dormant feature)
$\beta = c_1 c_2 (\gamma P_2 - P_1)$	(brake)
$T = \alpha / \beta$	(target eigenvalue, the root of the per-feature force)
$u_k = [(1 + c_1 \lambda_k)(1 + c_2 \lambda_k)]^{-1}$	(state factor of feature k, in $(0, 1]$, equal to 1 at $\lambda_k = 0$)

MMD-Gauss dynamics (Definition 2, Section 4.4, Appendix H).

$s = \frac{2}{2d-3}$	(scale inside the inverse-square-root kernel; *not* the weight s_k)
C	(target-target constant of Definition 2)
$t \in [0, \infty)$	(Gamma-integral variable, one bandwidth per t; unrelated to the frequency t of Definition 1)
$c_1(t) = 4st$, $c_2(t) = \frac{2st}{1+2st}$	(slice constants: the BHEP constants at $\beta_{bw}^2 = 2st$)
$\gamma_t = 2(1 + 2st)^{-d/2}$	(slice Gaussian factor)
$P_1(t)$, $P_2(t)$	(slice potentials $\prod_i (1 + c_j(t) \lambda_i)^{-1/2}$; batch sums in Eq. 3)
$\tilde{P}_2(t) = \gamma_t P_2(t)/2$	(pull potential with the Gaussian factor absorbed, $\prod_i (1 + 2st(1 + \lambda_i))^{-1/2}$; Appendices L, Q)
α_t, β_t	(slice engine and brake: the BHEP coefficients at bandwidth $2st$ divided by t, Eq. 42)
$u_{t,k}$	(slice state factor $[(1 + c_1(t)\lambda_k)(1 + c_2(t)\lambda_k)]^{-1}$)
$F(\lambda_k)$	(per-feature force, Eq. 10: $\frac{1}{\sqrt{\pi}} \int_0^\infty t^{1/2} e^{-t} u_{t,k} [\alpha_t - \beta_t \lambda_k] dt$, with $\dot{\lambda}_k = 2\sigma_k^2 \lambda_k F(\lambda_k)$)
$F(0)$	(dormant-feature force, the counterpart of α)
T	(root of $F(T) = 0$, the counterpart of α/β; Appendix H.4)
K	(number of Gauss-Laguerre knots: $k = 10$ in the toy and $k = 100$ in the real-data experiments)

Every BHEP quantity has a slice counterpart here: BHEP with the constants above is the slice $t = \frac{2d-3}{4d} \approx \frac{1}{2}$.

Absorbed form (Appendix I only).

$\gamma_{mmd} = \frac{1}{2}, \sigma^2 = 1$	(kernel exponent and variance in Zimmermann et al. (2025); *not* the BHEP γ, *not* the input variance σ_k^2)
$P_2^{\text{abs}}(t)$	(pull potential $\prod_i (1 + st\lambda_i)^{-1/2}$ of the absorbed form, Appendix L)
$\lambda_{\text{ideal}} = a + b/d$	(two-scale ansatz for its fixed point, solved as $a = 1$, $b = -1$)

Loss definitions (Section 3).

$t \in \mathbb{R}$	(frequency variable of the Epps-Pulley test)
$w(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$	(Gaussian window; the factor $\frac{1}{\sqrt{2\pi}}$ is dropped in the implementation of Balestriero & LeCun (2025))
$\hat{\phi}_X(t)$, $\phi(t) = e^{-t^2/2}$	(empirical and target Gaussian characteristic functions)
β_{bw}^2	(BHEP kernel bandwidth: $1/d$ in Definition 4, $2st$ on a slice; *not* the brake β)
$\xi \sim S^{d-1}$	(random projection direction of SIGReg, Appendix C)

All BHEP and MMD-Gauss statistics in the paper use the U-statistic convention (diagonal terms dropped) and are divided by n, and their regularization coefficients are multiplied by $\sqrt{2\pi}$ to match the scale of SIGReg (Appendix C).

Forces on the equilibrium ladder (Appendices L and M.1).

λ_A, λ_{-A}	(eigenvalue of the probe feature; common level of the other saturated features)
m'	(number of saturated features other than the probe, $m - 1$ or m)
$\Phi(\lambda_A)$	(force on the probe, normalized as $u_A(\alpha - \beta\lambda_A) = -2\,\partial L/\partial \lambda_A$)
$\alpha_0 = \Phi(0)$	(dormant-feature force with m features held at T; equals α for BHEP and $F(0)$ for MMD-Gauss)
$\beta_A = -\Phi'(T)$, $\alpha_A = \beta_A T$	(local brake and engine at the target, measured numerically; Tables 3 and 4)
$v = \sum_j \lambda_j \xi_j^2$, $g(v)$	(projected variance and kernel derivative of SIGReg)

Assumption violations (Appendix K).

A_v, $C_A = \frac{1}{n} A_v^T A_v$	(augmentation of view v and its covariance, with eigenvalues a_k^2)
$\nu_k = \frac{a_k^2}{\sigma_k^2 + a_k^2}$	(augmentation ratio of feature k)
$\tau = \frac{d\alpha}{4}$	(collapse threshold: feature k collapses when $\nu_k > \tau$; Appendix K.1)
invariance loss	(invariance term of LeJEPA at the end of training, the aggregate indicator of residual augmentation; Tables 1 and 2)
surviving directions	(eigenvalues with $\lambda > 0.5$ at the end of training)
κ_{rad}	(radial fourth cumulant, zero for every Gaussian: $\mathbb{E}\|z\|^4 - (\operatorname{tr}\Sigma)^2 - 2\operatorname{tr}(\Sigma^2)$; Appendix K.2)
ϵ_1, ϵ_2	(relative errors of P_1 and P_2 to first order in κ_{rad})

Q FULL DERIVATION OVER α DECREASE

Q.1 PRELIMINARY

We use the MMD-Gauss of Definition 2 for theoretical analysis, assuming d is large. All the α below indicates $\alpha_0 = F(0)$.

In this form, BHEP is literally a slice, so the fixed point is shared. The limit $d \to \infty$ replaces the baseline factor of the dormant features by e^{-t}.

We can simplify the overall derivation into a single equation:

$$|\Delta\alpha_{decr}| > |\Delta\alpha_{incr}|, \tag{96}$$

where the decrease is caused by emergence of a single additional feature, and increase is caused by shrink of other m features to match decreasing T. Our goal is to prove that the absolute amount of decrease is bigger than that of increase.

Q.2 Derivation Form

Using Eq. 39, excluding the coefficient $\frac{1}{\sqrt{\pi}}$,

$$\widehat{\text{MMD-Gauss}}^2 \propto \int_0^\infty t^{-1/2}e^{-t}\left[\prod_{i=1}^{d}(1+4st\lambda_i)^{-1/2} - 2\prod_{i=1}^{d}(1+2st(\lambda_i+1))^{-1/2}\right]dt$$

The second product is $2\tilde{P}_2(t) = \gamma_t P_2(t)$ in the notation of Appendices H.2 and L.

At equilibrium, when m features are saturated, at eigenvalue $\lambda = L$, where $\sigma = sL = \frac{2L}{2d-3}$ the force Φ would be (this σ is unrelated to the input variance σ_k^2 of the main text):

$$\begin{aligned}
\Phi(m) &= \int_0^\infty t^{-1/2}e^{-t}2t\left[(1+4t\sigma)^{-(m+2)/2} - (1+2st)^{-(d-m)/2}(1+2t(\sigma+s))^{-(m+2)/2}\right]dt \\
&= 2\int_0^\infty t^{1/2}e^{-t}\left[(1+4t\sigma)^{-(m+2)/2} - (1+2st)^{-(d+2)/2}\left(1+\frac{2t\sigma}{1+2st}\right)^{-(m+2)/2}\right]dt \\
&\xrightarrow{d\to\infty} 2\int_0^\infty t^{1/2}e^{-t}\left[(1+4t\sigma)^{-(m+2)/2} - e^{-t}\cdot(1+2t\sigma)^{-(m+2)/2}\right]dt \\
&= 2\int_0^\infty t^{1/2}e^{-t}\left[\hat{A}(t)^{-(m+2)/2} - e^{-t}\cdot\hat{B}(t)^{-(m+2)/2}\right]dt = 0
\end{aligned}$$

As this assumes equilibrium at that state, nullifying the force.

Here $\hat{A}(t) = \hat{B}(2t)$, so the pull term of $\Phi(m)$ (whose weight e^{-2t} is the measure's e^{-t} times the dormant baseline e^{-t}) can be re-read over the base $\hat{A}$ by substituting $t = 2t'$ ($\theta := \frac{m+2}{2}$):

$$\int_0^\infty t^{1/2}e^{-2t}\hat{B}(t)^{-\theta}dt = \int_0^\infty \sqrt{2}\,t'^{1/2}e^{-4t'}\hat{B}(2t')^{-\theta}\cdot 2dt' = 2\sqrt{2}\int_0^\infty t'^{1/2}e^{-4t'}\hat{A}(t')^{-\theta}dt'.$$

Using this, we can merge the formula into single form (Removed coefficient 2, $\hat{A}(t) \to \hat{A}$, $\varphi(t) \to \varphi$ for simplicity):

$$\Phi(\theta) = \int_0^\infty t^{1/2}[e^{-t} - 2\sqrt{2}e^{-4t}]\hat{A}(t)^{-\theta}dt = \int_0^\infty t^{1/2}\varphi\hat{A}^{-\theta}dt = 0 \tag{97}$$

Where $\varphi(t) = e^{-t} - 2\sqrt{2}e^{-4t}$.

The engine is a rung of the same function. A dormant feature's own push factor is $1 + 4st \cdot 0 = 1$, so dividing the global product by it once more leaves the exponent one unit lower in θ, while its pull factor $1 + 2st$ tends to 1 as $d \to \infty$. So the engine is

$$\alpha(\theta) = \Phi(\theta - 1).$$

Equivalently, the engine at m saturated features equals the balance integral of a system with $m - 2$ features at the same eigenvalue level.

Having written the force as $\Phi(\theta)$, where θ makes it vanish, we now regard it as a function of two parameters: $\Phi(\omega; \sigma)$, where equilibrium is $\omega = \theta$ ($\theta - 1$ for engine α) and $\sigma = \sigma(\theta)$. This is because they are the 2 parameters that is responsible for α change. (To avoid confusion: θ is the constant at which the equilibrium holds, and ω is the variable that moves around θ and in which we differentiate.)

The equilibrium condition $\Phi(\theta;\sigma) = 0$ defines a curve $\theta \mapsto \sigma(\theta)$, which we call the *equilibrium ladder*: as training saturates features one by one, the system climbs it rung by rung, while conserving the equilibrium, by updating corresponding $\sigma(\theta)$. We write $\Phi(\omega;\sigma)$ with two slots, where ω is a free variable in which we differentiate and θ is the value at which the ladder closes ($\theta - 1$ for the engine).

Q.3 Bracket form

Let $a(t) := \ln \hat{A}(t)$. On the ladder we abbreviate integrals against the (sign-mixed) weight by a bracket:

$$\langle f(a(t))\rangle := \int_0^\infty f(a(t))\, t^{1/2}\varphi(t)\hat{A}(t)^{-\theta}dt, \tag{98}$$

where the letter a inside a bracket merely names the argument of the function. Since $\hat{A}^{-\omega} = e^{(\theta-\omega)a}\hat{A}^{-\theta}$, neighboring rungs of Φ carry geometric weights:

$$\Phi(\theta-1) = \langle e^a\rangle = \alpha(\theta), \qquad \Phi(\theta) = \langle 1\rangle = 0, \qquad \Phi(\theta+1) = \langle e^{-a}\rangle. \tag{99}$$

Q.4 Partial derivation over ω and σ

When θ varies, Φ changes through two slots at once: the exponent ω directly, and the level σ through the constraint $\Phi(\theta;\sigma(\theta)) = 0$. We therefore prepare the two partial derivatives. Differentiation under the integral sign is justified throughout, since every integrand is dominated by e^{-t} times polynomial growth.

Lemma 1. *For every real ω and $\sigma > 0$,*

$$\partial_\omega\Phi(\omega;\sigma) = -\int_0^\infty t^{1/2}\varphi(t)a(t)\hat{A}(t)^{-\omega}dt, \qquad \partial_\sigma\Phi(\omega;\sigma) = -\frac{\omega}{\sigma}\big[\Phi(\omega;\sigma) - \Phi(\omega+1;\sigma)\big]. \tag{100}$$

Proof. For the first identity, $\hat{A}^{-\omega} = e^{-\omega a}$ gives $\partial_\omega\hat{A}^{-\omega} = -a\hat{A}^{-\omega}$.

For the second, $\partial_\sigma\hat{A} = 4t$ gives $\partial_\sigma\hat{A}^{-\omega} = -4\omega t\hat{A}^{-\omega-1}$, hence

$$\partial_\sigma\Phi(\omega;\sigma) = -4\omega\int_0^\infty t^{1/2}\varphi\cdot t\cdot\hat{A}^{-\omega-1}dt.$$

The extra factor of t takes the integrand out of the family Φ. Since $\hat{A}$ is affine in t we can trade it back, $t = \frac{\hat{A}-1}{4\sigma}$, so that

$$\partial_\sigma\Phi(\omega;\sigma) = -\frac{\omega}{\sigma}\int_0^\infty t^{1/2}\varphi(t)\big(\hat{A}^{-\omega} - \hat{A}^{-\omega-1}\big)dt = -\frac{\omega}{\sigma}\big[\Phi(\omega;\sigma) - \Phi(\omega+1;\sigma)\big].$$

□

As we've achieved a general form over derivation, we can use that for 2 different settings. First, for equilibrium setting, as it's always $\Phi(\theta,\sigma(\theta)) = 0$ regardless of θ, so the partial derivatives with respect to ω and $\sigma(\theta)$ cancel each other out:

$$0 = \partial_\omega\Phi(\theta;\sigma(\theta))\cdot\frac{d\omega}{d\theta} + \partial_\sigma\Phi(\theta;\sigma(\theta))\cdot\frac{d\sigma}{d\theta} \tag{101}$$

$$= \partial_\omega\Phi(\theta;\sigma(\theta))\cdot 1 + \partial_\sigma\Phi(\theta;\sigma(\theta))\cdot\frac{d\sigma}{d\theta} \tag{102}$$

$$\therefore \frac{d\sigma}{d\theta} = -\frac{\partial_\omega\Phi(\theta;\sigma(\theta))}{\partial_\sigma\Phi(\theta;\sigma(\theta))} \tag{103}$$

And for engine α, the form would be:

$$\begin{aligned}\frac{d\alpha}{d\theta} &= \partial_\omega\Phi(\theta-1;\sigma(\theta))\cdot\frac{d\omega}{d\theta} + \partial_\sigma\Phi(\theta-1;\sigma(\theta))\cdot\frac{d\sigma}{d\theta}\\ &= \partial_\omega\Phi(\theta-1;\sigma(\theta)) + \partial_\sigma\Phi(\theta-1;\sigma(\theta))\cdot\left(-\frac{\partial_\omega\Phi(\theta;\sigma(\theta))}{\partial_\sigma\Phi(\theta;\sigma(\theta))}\right)\end{aligned}$$

This formula mirrors our initial proposal in Equation 96: the first term corresponds to the decrease ($\Delta\alpha_{decr}$) and the second to the increase ($\Delta\alpha_{incr}$).

Using the derivatives of Equation 100 together with the ladder $\Phi(\theta;\sigma(\theta)) = 0$, it can be simplified as:

$$\frac{d\alpha}{d\theta} = -\langle ae^a\rangle + \left(-\frac{\theta-1}{\sigma(\theta)}\langle e^a\rangle\right)\cdot\left(\frac{\langle a\rangle}{\frac{\theta}{\sigma(\theta)}\langle e^{-a}\rangle}\right) = -\langle ae^a\rangle - \frac{(\theta-1)\langle a\rangle\langle e^a\rangle}{\theta\langle e^{-a}\rangle} = -Q + \frac{(\theta-1)AP}{\theta R} \tag{104}$$

Where each of the term corresponds to (all of them are positive, is shown after Lemma 2):

- $P = -\partial_\omega\Phi(\theta;\sigma(\theta)) = \langle a\rangle$: Perturbation caused by ω (activated feature count)
- $R = -\Phi(\theta+1;\sigma(\theta)) = -\langle e^{-a}\rangle$: Stiffness over next force
- $A = \Phi(\theta-1;\sigma(\theta)) = \langle e^a\rangle = \alpha$: Engine itself, we need to know
- $Q = -\partial_\omega\Phi(\theta-1;\sigma(\theta)) = \langle ae^a\rangle$: Perturbation caused by ω, acting on the engine α

Altogether, we need to show $\frac{d\alpha}{d\theta} < 0$. In other words, we need to show:

$$Q > \frac{(\theta-1)AP}{\theta R}. \tag{105}$$

Q.5 One more bracket with tail mass G

To use the bracket form we face an obstacle: φ changes sign at $t_0 = \frac{1}{2}\ln 2$, so $\langle\cdot\rangle$ is not an integral against a positive weight, and none of the signs of P, A, R, Q are obvious. Two structural facts repair this: the weight changes sign only once, and its total mass is the equilibrium itself, $\langle 1\rangle = \Phi(\theta;\sigma(\theta)) = 0$. They let us trade the sign-mixed weight for a positive one.

For a threshold $x > 0$ define the tail mass

$$G(x) := \int_{t_x}^{\infty} t^{1/2}\varphi(t)\hat{A}(t)^{-\theta}dt, \qquad t_x := a^{-1}(x) = \frac{e^x-1}{4\sigma}. \tag{106}$$

Lemma 2. *$G \geq 0$ on $(0,\infty)$, with $G > 0$ on $(0,a_0]$, and for every C^1 function f of at most exponential growth ($|f| + |f'| \leq Ce^{c|x|}$):*

$$\langle f\rangle = \int_0^{\infty} f'(x)G(x)dx =: \langle\langle f'\rangle\rangle. \tag{107}$$

Proof. For $x \geq a_0 = a(t_0)$, where the sign changes, the integrand of $G(x)$ is positive on (t_x,∞), since $t_x \geq t_0$. For $x \leq a_0$, zero total mass flips the tail into the head:

$$G(x) = -\int_0^{t_x} t^{1/2}\varphi\hat{A}^{-\theta}dt = \int_0^{t_x} t^{1/2}|\varphi|\hat{A}^{-\theta}dt \geq 0,$$

which only collects negative mass with its sign removed.

For the identity, the fundamental theorem of calculus gives $f(a(t)) - f(0) = \int_0^{\infty} f'(x)\mathbf{1}\{a(t) > x\}dx$. Multiplying by $t^{1/2}\varphi\hat{A}^{-\theta}$, integrating in t, and swapping the two integrals (justified by exponential decay),

$$\langle f\rangle - f(0)\langle 1\rangle = \int_0^{\infty} f'(x)G(x)dx,$$

and $\langle 1\rangle = 0$ removes the constant term. □

Then, we can rewrite the predetermined P, R, A, Q as

- $P = \langle a\rangle = \langle\langle 1\rangle\rangle$
- $R = -\langle e^{-a}\rangle = \langle\langle e^{-a}\rangle\rangle$
- $A = \langle e^a\rangle = \langle\langle e^a\rangle\rangle$

- $Q = \langle ae^a \rangle = \langle\!\langle (a+1)e^a \rangle\!\rangle$

Each function inside $\langle\!\langle \cdot \rangle\!\rangle$ above is pointwise positive, and $G \geq 0$ with $G > 0$ on $(0, a_0]$, so all four quantities are positive. Moreover

$$P - R = \langle\!\langle 1 - e^{-a} \rangle\!\rangle > 0, \qquad A - P = \langle\!\langle e^a - 1 \rangle\!\rangle > 0,$$

hence $R < P < A$. Positivity of R also settles two debts of the previous subsection. At any root of the ladder, $\partial_\sigma \Phi(\theta; \sigma) = -\frac{\theta}{\sigma} R < 0$, so the root $\sigma(\theta)$ is unique and C^1 by the implicit function theorem, legitimizing the division that produced $\frac{d\sigma}{d\theta}$. Existence follows from the endpoint signs: $\Phi(\theta; \sigma) \to \Gamma(\frac{3}{2})(1 - \frac{\sqrt{2}}{4}) > 0$ as $\sigma \to 0^+$, while $(4\sigma)^{3/2}\Phi(\theta; \sigma) \to (1 - 2\sqrt{2})B(\frac{3}{2}, \theta - \frac{3}{2}) < 0$ for $\theta > \frac{3}{2}$ (with $B(\cdot, \cdot)$ the Beta function). At the first rung $\theta = \frac{3}{2}$ the certified anchor of Table 6, $\Phi(\frac{3}{2}; \frac{Z}{3}) < 0$, supplies the negative value directly.

The four quantities package into a single function, the moment generating function of G:

$$\psi(v) := \langle\!\langle e^{va} \rangle\!\rangle = \int_0^\infty e^{vx} G(x) dx. \tag{108}$$

The integral converges for every v, since $G(x)$ decays superexponentially in x (Since $t_x \propto e^x$ and integration contains e^{-t}). Taking $f(y) = e^{vy}$ in the Lemma gives $\langle e^{va} \rangle = \langle\!\langle ve^{va} \rangle\!\rangle = v\psi(v)$, while $\hat{A}^{-(\theta - v)} = e^{va}\hat{A}^{-\theta}$ identifies the left side as a rung of Φ. Hence, on the ladder,

$$\Phi(\theta - v; \sigma(\theta)) = v\psi(v). \tag{109}$$

Then, we can also rewrite as:

- $P = \langle\!\langle 1 \rangle\!\rangle = \psi(0)$
- $R = \langle\!\langle e^{-a} \rangle\!\rangle = \psi(-1)$
- $A = \langle\!\langle e^a \rangle\!\rangle = \psi(1)$
- $Q = \langle\!\langle (a+1)e^a \rangle\!\rangle = \psi(1) + \psi'(1) = (v\psi)'(1)$

Two properties of ψ drive everything that follows. First, ψ is positive and strictly increasing:

$$\psi'(v) = \int_0^\infty x e^{vx} G(x) dx > 0,$$

recovering the ordering $R = \psi(-1) < \psi(0) = P < \psi(1) = A$ obtained above. Second, $\ln \psi$ is convex: by the Cauchy-Schwarz inequality using $G(x)dx \geq 0$,

$$\begin{aligned} \psi\Big(\frac{v_1 + v_2}{2}\Big) &= \int_0^\infty e^{v_1 x/2} e^{v_2 x/2} G(x) dx \\ &\leq \left(\int_0^\infty e^{v_1 x} G(x) dx\right)^{1/2} \left(\int_0^\infty e^{v_2 x} G(x) dx\right)^{1/2} = \sqrt{\psi(v_1)\psi(v_2)}. \end{aligned}$$

This convexity is the only probabilistic ingredient of the whole proof.

Q.6 Reduction to a single scalar

The criterion $Q > \frac{(\theta-1)AP}{\theta R}$ compares four separate integrals. We now compress it into a statement about a single number. Since $R > 0$, the criterion is equivalent to

$$\theta QR > (\theta - 1)PA.$$

All four quantities are values of ψ, so it is natural to read this through $\ell := \ln \psi$. Convexity of ℓ means that ℓ' is nondecreasing, and the single number we keep is its value at the origin,

$$\mu := \ell'(0) = \frac{\psi'(0)}{\psi(0)} = \frac{\langle\!\langle a \rangle\!\rangle}{\langle\!\langle 1 \rangle\!\rangle} = \frac{\int_0^\infty x G(x) dx}{\int_0^\infty G(x) dx}, \tag{110}$$

the barycenter of the positive weight G.

Convexity now delivers the two sides of the criterion. For the pair (Q, A): since $Q = (v\psi)'(1) = \psi(1)\big(1 + \ell'(1)\big)$ and $\ell'(1) \geq \ell'(0) = \mu$, we get

$$Q \geq (1 + \mu)A.$$

For the pair (P, R): since $\ln \frac{P}{R} = \ell(0) - \ell(-1) = \int_{-1}^{0} \ell'(v)dv \leq \ell'(0) = \mu$, we get

$$R \geq Pe^{-\mu}.$$

Multiplying the two bounds and comparing with the criterion,

$$\theta QR \geq \theta(1+\mu)e^{-\mu}PA, \qquad \text{so} \qquad (1+\mu)e^{-\mu} > 1-\frac{1}{\theta} \implies \theta QR > \theta\Big(1-\frac{1}{\theta}\Big)PA = (\theta-1)PA, \tag{111}$$

that is, $\frac{d\alpha}{d\theta} < 0$.

The function $(1 + x)e^{-x}$ equals 1 at $x = 0$ and strictly decreases for $x > 0$, so the condition in Equation 111 says exactly that the barycenter stays below an explicit threshold: $\mu < \mu^*(\theta)$, where $\mu^*(\theta)$ is the unique positive solution of $(1 + \mu^*)e^{-\mu^*} = 1 - \frac{1}{\theta}$. The threshold decreases in θ, since the right side increases in θ while the left side decreases in μ^*, and $\mu^*(\theta) \approx \sqrt{2/\theta}$ for large θ. The remaining task is therefore an explicit upper bound on μ, which the next subsections provide.

Q.7 Boundary over level combination z

Before moving onto next section, we would discuss the boundary over:

$$z = 2\sigma\theta. \tag{112}$$

Monotone decrease First, z decreases monotonically as θ increases. This can be simply shown using $\frac{d\sigma}{d\theta} = -\frac{\sigma P}{\theta R}$ from Equation 101:

$$\frac{d\ln z}{d\theta} = \frac{1}{z}\cdot\frac{dz}{d\theta} = \frac{1}{2\sigma\theta}\cdot\frac{d(2\sigma\theta)}{d\theta} = \frac{1}{2\sigma\theta}\cdot\left(2\sigma + 2\theta\cdot -\frac{\sigma P}{\theta R}\right) = \frac{1}{\theta}\cdot\left(1-\frac{P}{R}\right) < 0, \tag{113}$$

Since $P > R$.

Lower bound Since we have established the monotonic decrease of the strictly positive variable z, the lower limit must be $z_\infty \geq 0$.

$$0 = \int_0^\infty t^{1/2}\varphi(t)\Big(1+\frac{2z(\theta)t}{\theta}\Big)^{-\theta}dt \xrightarrow{\theta\to\infty} \int_0^\infty t^{1/2}\varphi(t)e^{-2z_\infty t}dt.$$

Using $\varphi(t) = e^{-t} - 2\sqrt{2}e^{-4t}$,

$$\begin{aligned}
&= \int_0^\infty t^{1/2}[e^{-(1+2z_\infty)t} - 2\sqrt{2}e^{-(4+2z_\infty)t}]dt \\
&= \Gamma\left(\frac{3}{2}\right)\cdot\left[\frac{1}{(1+2z_\infty)^{3/2}} - \frac{2\sqrt{2}}{(4+2z_\infty)^{3/2}}\right] \\
&= 0
\end{aligned}$$

Solving this equation gives the lower bound for z, $z_\infty = 1$. Since z strictly decreases to this limit, $z(\theta) > 1$ for every finite θ.

Upper bound Suppose a single sign is certified at some θ_0: $\Phi\big(\theta_0; \frac{Z}{2\theta_0}\big) < 0$. Since $\Phi(\theta_0; \sigma)$ is positive as $\sigma \to 0^+$ and crosses zero only once, the root must lie below, $\sigma(\theta_0) < \frac{Z}{2\theta_0}$, that is $z(\theta_0) < Z$. Monotone decrease then spreads the bound to everything on the right: $z(\theta) < Z$ for every $\theta \geq \theta_0$. One certified sign per interval fences the whole half-line.

A first payoff: the target T decrease On the ladder the saturated features sit at the target, $L = T$, so the level tracks the target itself: $\sigma = sT$, that is $T = \frac{z}{2s\theta}$. Both factors now work downward. z decreases strictly toward 1, and $\frac{1}{\theta}$ dilutes as features emerge. This proves the monotone decrease of the target along the equilibrium ladder. It also explains why the early drop is steep: at small θ (= small m) the collapse of z from its largest values toward 1 compounds the dilution, while at large θ only the dilution remains, $T \simeq \frac{1}{2s\theta}$.

Q.8 UPPER BOUND FOR μ

The threshold μ^* is tractable: it solves an explicit scalar equation, it decreases in θ, and it can be bounded from below by evaluating $(1+x)e^{-x}$ at candidate numbers. However, unlike μ^*, μ does not have any closed form, as it sits on ladder $\sigma(\theta)$, which is already highly intractable. We don't even know whether it monotonically decreases over θ. They cannot be compared directly.

Instead, we construct an explicit function $\mu_{hi}(\theta)$ with two properties: it dominates the barycenter pointwise, $\mu \le \mu_{hi}(\theta)$, and it monotonically decreases in θ. On an interval $[\theta_0, \theta_1]$ the comparison then collapses to the endpoints: μ_{hi} is largest at the left end while μ^* is smallest at the right end, so the single numeric check $\mu_{hi}(\theta_0) < \mu^*(\theta_1)$ certifies $\mu < \mu^*$ on the whole interval.

The construction takes three moves. First we rewrite μ exactly as the mean of a positive measure (folding the signs). Then we push the t-mean of that measure to a frozen benchmark $\mathcal{T}(\theta_0)$ (a correlation inequality). Finally concavity of a converts the a-mean into the t-mean (Jensen), leaving a closed formula.

Here, as $\langle 1\rangle = 0$, we can manually add or subtract any constant, so we can use

$$\langle a\rangle = \langle a - a_0\rangle, \qquad \langle a^2\rangle = \langle a^2 - a_0^2\rangle = \langle (a-a_0)(a+a_0)\rangle.$$

In other words, since both $a - a_0$ and φ change sign at the same point and in the same direction ($- \to +$), we can use:

$$(a-a_0)\varphi = |a-a_0||\varphi|.$$

By the Lemma with $f = a^2/2$ and $f = a$, the barycenter reads $\mu = \frac{\langle\langle a\rangle\rangle}{\langle\langle 1\rangle\rangle} = \frac{\langle a^2\rangle/2}{\langle a\rangle} = \frac{\langle a^2 - a_0^2\rangle/2}{\langle a - a_0\rangle}$.

Defining the positive measure $d\eta := t^{1/2}|\varphi(t)||a-a_0|\hat{A}^{-\theta}dt$, the two brackets become $\langle a - a_0\rangle = \int d\eta$ and $\langle (a-a_0)(a+a_0)\rangle = \int (a+a_0)d\eta$, hence exactly

$$\mu = \frac{\frac{1}{2}\int (a+a_0)d\eta}{\int d\eta} = \frac{\mathbb{E}_\eta[a] + a_0}{2}. \tag{114}$$

The sign obstacle is now fully behind us: μ is half the sum of the η-mean of a and the fixed number a_0, and it only remains to bound $\mathbb{E}_\eta[a]$.

Once more, however, the η is hard to analyze, as it still moves along θ, which makes derivation difficult. So, we can divide $d\eta = h \cdot d\xi_0$ (Unrelated to the ξ used in SIGReg definition):

$$h = \underbrace{\frac{|a-a_0|}{|t-t_0|}}_{h_1} \cdot \underbrace{\hat{A}^{-\theta}(1+\tfrac{2t}{\theta_0})^{\theta_0}}_{h_2}, \qquad d\xi_0 = t^{1/2}|\varphi||t-t_0|(1+2t/\theta_0)^{-\theta_0}dt \tag{115}$$

With a reference point θ_0 with $\frac{3}{2} \le \theta_0 \le \theta$. It plays the role of the left endpoint of a certification interval: everything below holds for every $\theta \ge \theta_0$, and the assembly will instantiate θ_0 at finitely many concrete values, one per interval.

The main reason for the division is to show h decreases monotonically over t. For that, we show that both h_1 and h_2 decrease monotonically.

For h_1, note first that $a - a_0$ and $t - t_0$ always share the same sign (a is increasing), so the absolute values can be dropped together:

$$h_1(t) = \frac{a(t)-a_0}{t-t_0} = \int_0^1 a'\big(t_0 + \tau(t-t_0)\big)d\tau,$$

the average of a' along the segment from t_0 to t, with $h_1(t_0) = a'(t_0)$. Since $a(t) = \ln(1+4\sigma t)$ is concave, a' is decreasing. As t grows, every point $t_0 + \tau(t-t_0)$ of the segment moves to the right, so the integrand decreases pointwise, and h_1 decreases on all of $(0,\infty)$, through t_0 included.

For h_2, using $\hat{A} = 1 + 4t\sigma$,

$$\frac{d}{dt}\ln h_2(t) = -\frac{2z\theta}{\theta + 2zt} + \frac{2\theta_0}{\theta_0 + 2t} = -\frac{2\big[\theta\theta_0(z-1) + 2zt(\theta-\theta_0)\big]}{(\theta+2zt)(\theta_0+2t)} \le 0.$$

The last expression is ≤ 0 since $z > 1$ by the previous subsection and $\theta \geq \theta_0$, so h_2 decreases (as $h_2(t) > 0$).

As we have shown that h decreases, the reweighting by h favors small t, so it can only drag the mean of t down. This is the correlation inequality: for a probability measure π with f increasing and g decreasing, $\mathbb{E}_\pi[fg] \leq \mathbb{E}_\pi[f]\mathbb{E}_\pi[g]$, since $\frac{1}{2}\mathbb{E}\big[(f(t) - f(t'))(g(t) - g(t'))\big] = \mathbb{E}_\pi[fg] - \mathbb{E}_\pi[f]\mathbb{E}_\pi[g]$ over independent copies has a pointwise nonpositive integrand. Applying it with π the normalized ξ_0, $f = t$, $g = h$:

$$\mathbb{E}_\eta[t] = \frac{\mathbb{E}_{\xi_0}[t \cdot h]}{\mathbb{E}_{\xi_0}[h]} \leq \mathbb{E}_{\xi_0}[t] = \mathcal{T}(\theta_0). \tag{116}$$

Where

$$\mathcal{T}(\theta_0) = \frac{\int_0^\infty t \cdot d\xi_0}{\int_0^\infty d\xi_0} = \frac{\int_0^\infty t^{3/2}|\varphi|\,|t - t_0|\,(1 + 2t/\theta_0)^{-\theta_0} dt}{\int_0^\infty t^{1/2}|\varphi|\,|t - t_0|\,(1 + 2t/\theta_0)^{-\theta_0} dt} \tag{117}$$

As $a(t) = \ln(1 + 4\sigma t)$ is concave, we can use Jensen's inequality:

$$\mathbb{E}_\eta[a] \leq \ln(1 + 4\sigma\mathbb{E}_\eta[t]) \leq \ln(1 + 4\sigma\mathcal{T}(\theta_0)) \tag{118}$$

Combining with Equation 114 and $a_0 = \ln(1 + 4\sigma t_0)$, which is exact,

$$\mu \leq \frac{\ln(1 + 4\sigma\mathcal{T}(\theta_0)) + \ln(1 + 4\sigma t_0)}{2}.$$

Finally we replace the level by its fences. Writing $4\sigma = \frac{2z}{\theta}$ and using $z < Z$ on the interval (both logarithms increase in σ),

$$\mu \leq \mu_{hi}(\theta) := \frac{1}{2}\ln\Big[\Big(1 + \frac{2Z\mathcal{T}(\theta_0)}{\theta}\Big)\Big(1 + \frac{2Zt_0}{\theta}\Big)\Big]. \tag{119}$$

This is the explicit upper bound promised at the start of the subsection. For fixed Z and $\mathcal{T}(\theta_0)$, both factors decrease in θ, so μ_{hi} visibly decreases, which is exactly the property the endpoint reduction needs.

Q.9 Computational validation over boundary

The pieces assemble as follows (with Z and $\mathcal{T}(\theta_0)$ frozen on the interval):

- For certain θ_0 and θ_1 that satisfy $\frac{3}{2} \leq \theta_0 \leq \theta_1$,
- If $\mu_{hi}(\theta_0) < \mu^*(\theta_1)$,
- Then $\mu_{hi}(\theta') < \mu^*(\theta'), \quad \theta' \in [\theta_0, \theta_1]$ as they both decrease,
- And since $\mu(\theta) \leq \mu_{hi}(\theta)$,
- Then $\mu(\theta') < \mu^*(\theta'), \quad \theta' \in [\theta_0, \theta_1]$, proving our initial α decrease.

Each interval check needs three certified ingredients: the anchor sign $\Phi\big(\theta_0; \frac{Z}{2\theta_0}\big) < 0$, which fences $z < Z$ on $[\theta_0, \infty)$, an upper bound on $\mathcal{T}(\theta_0)$, and a lower bound on $\mu^*(\theta_1)$, obtained by evaluating $(1 + x)e^{-x}$ at a candidate number x^* and checking $(1 + x^*)e^{-x^*} > 1 - \frac{1}{\theta_1}$. Since the first two are definite integrals that do not close in elementary functions, we certify them by interval arithmetic: every operation is carried out on intervals with outward rounding, integrals are enclosed between Riemann lower and upper sums on a graded grid with an analytic tail bound, so the resulting enclosures are proofs rather than floating-point estimates.[3] No infinite family is checked case by case: the theory above reduces the claim $d\alpha/d\theta < 0$ to the finitely many explicit inequalities of Table 6, and the computation settles only those. For comparison the check is written on the $e^{2\mu}$ scale, where the endpoint inequality $\mu_{hi}(\theta_0) < \mu^*(\theta_1)$ reads

$$\Big(1 + \frac{2Z\mathcal{T}(\theta_0)}{\theta_0}\Big)\Big(1 + \frac{2Zt_0}{\theta_0}\Big) < e^{2\mu^*(\theta_1)}. \tag{120}$$

The three finite intervals pass with margins 1.204, 1.085, and 1.170, so $\frac{d\alpha}{d\theta} < 0$ on $[\frac{3}{2}, 5]$.

[3] The single remaining assumption is the correctly rounded (1 ulp) implementation of the standard library functions exp, log, sqrt, and pow.

Table 6: **Certified constants for the interval checks.** The anchor column shows the certified enclosure of $\Phi\big(\theta_0; \frac{Z}{2\theta_0}\big)$, whose negative sign fences $z < Z$. The last three columns compare the two sides of Equation 120: the left side uses the certified upper bounds on $\mathcal{T}(\theta_0)$, the right side uses the certified lower bound on $\mu^*(\theta_1)$ (column $\mu^*(\theta_1) >$). All bounds are rounded to the digits shown.

interval	Z	anchor $\Phi \times 10^3$	$\mathcal{T}(\theta_0) \leq$	$\mu^*(\theta_1) >$	lhs $\leq$	rhs $\geq$	margin
$[\frac{3}{2}, 2]$	3.3	$[-1.3931, -1.3845]$	1.882759	1.67	23.4417	28.2191	1.204
$[2, 3]$	2.5	$[-5.1507, -5.1417]$	1.723148	1.1875	9.9068	10.7510	1.085
$[3, 5]$	2.0	$[-10.1434, -10.1341]$	1.510483	0.82	4.4067	5.1552	1.170
$[5, \infty)$	1.6	$[-11.5436, -11.5339]$	1.293015	closed in elementary terms below			

The tail $\theta \geq 5$ Beyond the last anchor the fences persist ($Z = 1.6$ and $\mathcal{T}(5) \leq 1.293015$), and the check closes without further numbers. Expanding the product in Equation 120,

$$e^{2\mu_{hi}(\theta)} \leq 1 + \frac{c_1}{\theta} + \frac{c_2}{\theta^2}, \qquad c_1 := 2Z\big(\mathcal{T}(5) + t_0\big) \leq 5.246683, \qquad c_2 := 4Z^2\mathcal{T}(5)t_0 \leq 4.588798.$$

On the other side, $(1+\mu)e^{-\mu} \geq 1 - \frac{\mu^2}{2}$ gives $\mu^*(\theta) \geq \sqrt{2/\theta}$, and $e^x \geq 1 + x + \frac{x^2}{2}$ gives $e^{2\mu^*(\theta)} \geq 1 + \frac{2\sqrt{2}}{\sqrt{\theta}} + \frac{4}{\theta}$. The comparison therefore reduces to

$$2\sqrt{2}\sqrt{\theta} > c_1 - 4 + \frac{c_2}{\theta},$$

whose left side increases and right side decreases in θ, and at $\theta = 5$ it reads $6.3246 > 2.1645$. So the criterion holds on all of $[5, \infty)$.

The same fences also settle the target: the endpoint signs and the slope $\partial_\sigma \Phi = -\frac{\theta}{\sigma} R < 0$ established with the tail mass G give existence and uniqueness of the rung root $\sigma(\theta)$, hence of $T(m)$ for every m, which is the ladder counterpart of the root argument of Appendix H.4.

Combining the three interval checks with the tail, $\frac{d\alpha}{d\theta} < 0$ on all of $[\frac{3}{2}, \infty)$: the engine α_0 strictly decreases in the number of saturated features, which is the monotone decrease announced at the start of this section.

Q.10 Notations

The proof manipulates one two-slot function and a small cast of quantities read off from it. All of them are evaluated on the ladder defined below.

- $\varphi(t) = e^{-t} - 2\sqrt{2}e^{-4t}$: the universal weight of the family. It changes sign exactly once, at $t_0 = \frac{\ln 2}{2}$, from negative to positive.
- $\Phi(\omega; \sigma) = \int_0^\infty t^{1/2}\varphi(t)(1+4\sigma t)^{-\omega}dt$: the master function. The slot ω counts suppression (one unit of ω corresponds to two emerged features) and the slot σ is the common level of the saturated features.
- $\theta = \frac{m+2}{2}$: the position on the ω axis of the balance equation when m features are saturated. The offset 2 comes from the probe's denominator.
- $\sigma(\theta)$: the ladder, defined by $\Phi(\theta; \sigma(\theta)) = 0$, that is, the level at which the saturated features feel zero net force. Every quantity below is a snapshot taken at this level. Not the input eigenvalue σ_k^2 nor the MMD kernel variance.
- $z = 2\sigma\theta$: level times mode count. It decreases strictly and tends to 1, and the anchor constants of Table 6 certify its upper bounds.
- $A = \Phi(\theta - 1; \sigma(\theta))$: the engine itself ($\alpha_0 \propto A$), the force that wakes the next dormant feature.
- $Q = -\partial_\omega \Phi(\theta - 1; \sigma(\theta))$: the direct effect, how fast the engine drops when suppression deepens while all eigenvalues are frozen.
- $P = -\partial_\omega \Phi(\theta; \sigma(\theta))$: the kick, how far the balance of the saturated features is pushed off zero by the same deepening.

- $R = -\Phi(\theta + 1; \sigma(\theta))$: the stiffness, the depth of over-suppression one rung past the balance. It controls how strongly the balance restores itself when the level moves, through $\sigma' = -\sigma P/(\theta R)$.
- $a = \ln(1 + 4\sigma t)$: the log-suppression variable, in which neighboring rungs of Φ differ by a factor $e^{\pm a}$.
- $G(x)$: the tail mass of the sign-mixed integration weight. On the ladder $G \geq 0$, which turns all four quantities above into moments of a genuinely positive weight.
- $\psi(v) = \int_0^\infty e^{vx} G(x)dx$: the moment generating function of G. It packages the cast as $P = \psi(0)$, $A = \psi(1)$, $R = \psi(-1)$, $Q = (v\psi)'(1)$, and its log-convexity is the only probabilistic tool of the proof.
- $\mu = \psi'(0)/\psi(0)$: the barycenter of G. The whole monotonicity criterion reduces to $(1 + \mu)e^{-\mu} > 1 - \frac{1}{\theta}$, that is, to μ staying below an explicit threshold $\mu^*(\theta)$.
- $\mathcal{T}(\theta_0)$: an explicit comparison integral that bounds the mean of t under the reweighted measure, certified at finitely many θ_0 together with the anchors. Not related to the target eigenvalue $T = \alpha/\beta$.