DUAL MEMORY RETRIEVAL: BRIDGING CONTRASTIVE LEARNING WITH AND WITHOUT NEGATIVE PAIRS

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive learning builds self-supervision by comparing positive and negative samples. Methods such as SimSiam, DirectPred and DINO remove the negatives and resist collapse with a predictor or prototypes. Their contrast is carried jointly by the objective and these modules, rather than by the objective alone. We treat the negatives, the predictor and the prototypes alike as a memory of the representation distribution and recast contrast as retrieval from this memory, placing methods with and without negatives on one comparable axis. Revisiting contrastive learning from this perspective, we propose Dual Memory Retrieval (DMR). Its gradient is the difference between the retrievals of the two positive views at two temperatures in a softmax family or two exponents in a spectral family. We examine the resulting learning behavior on four image datasets. With a memory derived directly from the batch, the gap between the two temperatures or exponents is essential for training. With a memory learned by gradient steps, training fails when the positive enters the loss directly. Retrieving the positive through the memory lets the spectral family train with a learnable matrix, but not the softmax family without further mechanisms.

1 INTRODUCTION

Contrastive learning learns representations from unlabeled images through the pretext task of instance discrimination (Wu et al., 2018). Two augmented views of one image form a positive pair, and views of different images form negative pairs. The contrastive objective encourages positive pairs to be closer than negative pairs (Chen et al., 2020a; He et al., 2020). Later methods remove explicit negative pairs from the objective (Grill et al., 2020; Chen & He, 2021; Caron et al., 2020; 2021). These methods resist collapse through architectural designs such as a predictor, a stop-gradient, a momentum encoder, or trained prototypes.

Theoretical work explains how methods with and without negative pairs resist collapse (Wang & Isola, 2020; Tian et al., 2021; Zhang et al., 2022b), and several accounts relate their objectives through the spectrum of the representations (Balestriero & LeCun, 2022; Garrido et al., 2023). Among architectural designs, the predictor and the prototypes are modules that store information about the distribution of representations. Without negative pairs, the contrast is implemented jointly by the objective and the architecture, rather than by the objective alone.

Methods with and without negative pairs both locate each view within the representation distribution, and we treat this step as a retrieval from a memory. Following memory networks (Weston et al., 2015), a memory module records an estimate of this distribution in its memory state, which is derived from the current batch or learned by gradient steps. The view is the query, and the retrieval output is a weighted sum of the stored values. A gate sets the weights from the match between the query and the memory state. As the gate is adjusted continuously, for example through a temperature, the output can shift from retaining as much of the query as the memory state allows toward a shared output independent of the query. The amount of the query retained is the *level of self-retrieval*, and the path of this shift is the *retrieval trace* of the query.

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.

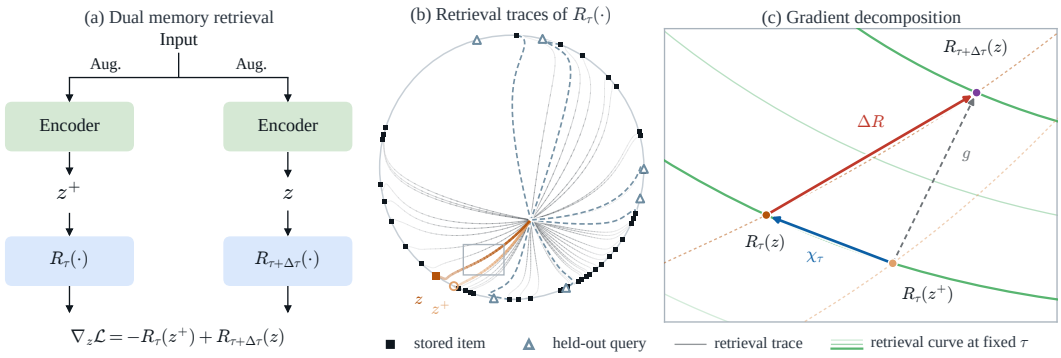


Figure 1: Dual memory retrieval. **(a)** The two views are retrieved at two levels of one memory, and the gradient is the difference of the two retrievals. **(b)** Retrieval traces of a softmax memory in two dimensions. Held-out queries lie outside the memory. As τ grows, each retrieval moves from the nearest stored item toward the centroid of all stored items. The box is enlarged in (c). **(c)** The gradient g splits into the positive-guided term χ_τ, along the retrieval curve at fixed τ, and the self-guided term ΔR, along the retrieval trace of z. Training moves z against the tangential part of g.

Revisiting the contrastive task from the retrieval perspective, we state two requirements that training should meet: *retrieval agreement* and *retrieval specificity*. Retrieval agreement requires positive pairs to retrieve similar outputs from the memory, which constrains both the representations and the memory state. A positive-guided term in the loss gradient serves this requirement by comparing the two views' retrievals at the same level. Retrieval agreement is also met by collapse, since identical representations retrieve identical outputs from any memory state. To resist collapse, retrieval specificity requires the retrieval to vary with the view along every direction across the batch. Collapse empties directions, so a direction should grow back once the representations begin to leave it empty. A self-guided term serves this requirement by comparing the view's own retrievals at two levels, so it acts even when the two views agree and the positive-guided term vanishes. Adding the two terms, we propose *dual memory retrieval*, whose loss gradient is the difference between the retrievals of the view and the positive at two levels of one memory state.

We instantiate memory retrieval as a softmax family and a spectral family. In the softmax family, the memory state stores items, such as the representations of the current batch or trained prototypes, and the gate is a softmax with an adjustable temperature. Raising the temperature moves the output from the nearest stored item to the centroid of all stored items. In the spectral family, the memory state is a symmetric matrix such as the second moment of the representations, and the gate raises its eigenvalues to an adjustable exponent. With eigenvalues below one, raising the exponent moves the output from the query itself to the origin. The spectral retrieval can also be written as a single matrix product. The two families relate the framework to existing methods and bridge methods with and without negative pairs. We then analyze which pairs of levels make an empty direction grow.

Contributions.

- We propose dual memory retrieval, a framework that models the representation distribution as a memory module and contrast as retrieval from this memory. The framework places classic methods such as SimCLR, SimSiam, DirectPred and DINO on a common axis for comparison, given by the pair of levels at which each method retrieves.
- We revisit contrastive learning from the retrieval perspective and propose two requirements, retrieval agreement and retrieval specificity. Building on these requirements, we construct the gradient and the loss of dual memory retrieval.
- Experiments on CIFAR-10, STL-10, ImageNet-100 and ImageNet-1k compare pairs of levels under matched settings, and diagnostics of the two requirements support our analysis of which pairs of levels resist collapse.

2 RELATED WORK

Contrastive learning. Methods with negative pairs estimate the representation distribution non-parametrically, contrasting each view with stored or in-batch samples under the InfoNCE objective (Wu et al., 2018; van den Oord et al., 2018; Chen et al., 2020a; He et al., 2020; Chen et al., 2020b). Variants add a second temperature (Zhang et al., 2022a) or replace the log-softmax with a quadratic objective (HaoChen et al., 2021). Methods without explicit negative pairs instead add a trained predictor (Grill et al., 2020; Chen & He, 2021). This predictor aligns with the eigenspaces of the correlation matrix of the representations, so DirectPred and DirectCopy set it directly from that matrix (Tian et al., 2021; Wang et al., 2021). Prototype methods compare views through their assignments to a set of prototypes (Caron et al., 2018; 2020; 2021). The components that resist collapse without negative pairs were found empirically, and a prototype head trains even without a momentum encoder or a stop-gradient (Amrani et al., 2022). With negative pairs, theory explains collapse avoidance through alignment, uniformity and temperature (Wang & Isola, 2020; Wang & Liu, 2021), and without them through the predictor and the stop-gradient (Halvagal et al., 2023; Richemond et al., 2023; Zhang et al., 2022b). Collapse can also be partial, confined to some directions or prototypes (Jing et al., 2022; Li et al., 2022; Govindarajan et al., 2024). Unifying accounts relate the objectives through spectral embeddings, the duality between contrastive and non-contrastive criteria, kernel losses and energy-based models (Balestriero & LeCun, 2022; Garrido et al., 2023; Koromilas et al., 2024; Sansone & Manhaeve, 2025).

Memory and retrieval. Memory networks separate what is stored, how it is updated and how it is read (Weston et al., 2015). End-to-end memory networks read with a softmax, which is our softmax family at a fixed temperature (Sukhbaatar et al., 2015). Hopfield networks retrieve stored patterns by descending an energy, quadratic in the classical network (Hopfield, 1982) and log-sum-exp in modern ones (Ramsauer et al., 2021; Millidge et al., 2022). Our spectral and softmax families correspond to these two energies. Instance discrimination used a memory bank as a store of samples (Wu et al., 2018), whereas we analyze the retrieval that carries the gradient.

3 DUAL MEMORY RETRIEVAL

3.1 MEMORY FOR THE REPRESENTATION DISTRIBUTION

Following memory networks (Weston et al., 2015), we describe the memory module by four components: memory state, gate, retrieval and memory update. The memory state records an estimate of the representation distribution, and a retrieval matches a query against this estimate.

3.1.1 MEMORY STATE, RETRIEVAL AND UPDATE

A retrieval has two steps (Sukhbaatar et al., 2015). *Addressing* turns the match between the query and the memory state into addressing weights through the gate. The *response* reweights the memory state with these weights and returns the retrieval output. At every training step, the memory update uses signals that are either derived directly from the queries or backpropagated from the retrieval. We analyze retrievals of the following form.

Definition 1 (Retrieval)**.** *Let the memory state M hold N keys $K \in \mathbb{R}^{N\times d}$, values $V \in \mathbb{R}^{N\times d}$, a bias $b \in \mathbb{R}^N$ and any strengths that the gate reads. Here d is the representation dimension and N the addressing dimension, one entry per key. For a query $q \in \mathcal{Z} := \mathbb{S}^{d-1}$ and $\tau > 0$, the retrieval consists of addressing and response,*

$$p_\tau := \hat{y}_\tau(Kq + b), \qquad R_\tau(q; M) := V^\top p_\tau. \tag{1}$$

The affinities $a := Kq + b$ score the match between the query and each key. The gate $\hat{y}_\tau$ converts them into the addressing weights p_τ at a scale set by τ.

We use two implementations of this retrieval, the *softmax family* and the *spectral family*. With zero bias, their retrievals are

$$R^{\mathrm{sm}}_\tau(q; M) = V^\top \mathrm{softmax}(Kq/\tau), \qquad R^{\mathrm{sp}}_\tau(q; M) = U^\top \,\mathrm{Diag}(\lambda^\tau)\, Uq = F^\tau q.$$

The spectral memory state is a symmetric matrix $F = U^\top \mathrm{Diag}(\lambda)\, U$, where the rows of U are $N \leq d$ orthonormal directions with strengths $\lambda \in (0, 1]^N$. The spectral keys and values are both U, and the spectral gate is $\mathrm{Diag}(\lambda^\tau)$, which is non-increasing in τ because every strength is at most one. Both retrievals have limits as $\tau \to 0$, and we write $\tau = 0$ for these limits, so that $\tau \in [0, \infty)$. On the softmax family the limit $\tau \to 0^+$ puts all weight on the keys of largest affinity, the nearest ones when keys and query have unit length. On the spectral family the limit is $F^0 = U^\top U$, the projection onto the stored directions, which is the identity when $N = d$.

When $V = K$, the response returns combinations of the stored keys themselves, and the retrieval is *auto-associative*, as in Hopfield networks (Hopfield, 1982). An auto-associative retrieval of either family is the gradient of a convex potential, $R_\tau(\cdot; M) = \nabla \Phi_\tau(\cdot; M)$. By default, retrievals in this work are auto-associative with zero bias.

3.1.2 Retrieval trace

The retrieval output is a weighted sum of the stored values. The output therefore reflects the memory state as well as the query and in general departs from the query. The gate controls how far the output departs. We first measure how much of the query the output retains and then follow the output as the gate varies. The *level of self-retrieval* is this amount, and the *self-retrieval score* measures it,

$$\rho_\tau(q; M) \;:=\; \langle q, R_\tau(q; M) \rangle. \tag{2}$$

Definition 2 (Retrieval trace)**.** *For a memory state M and a query q, the curve $\tau \mapsto R_\tau(q; M)$ over $\tau \geq 0$ is the* retrieval trace *of q.*

Figure 1(b) draws such traces for a softmax memory in two dimensions.

Proposition 1. *For both families, along the retrieval trace of any query q, the self-retrieval score $\rho_\tau(q; M)$ is non-increasing in $\tau \in [0, \infty)$ and decreases strictly wherever the retrieval moves.*

The score measures only the projection on the query, which differs from both the distance to the query and the retention along each direction. Proposition 1 orders the levels by τ. We therefore index levels by τ: a smaller τ is a higher level on every trace, though the score at one τ varies across queries. At $\tau = 0$ the retrieval is the *maximal self-retrieval*, which retains as much of the query as the memory state allows. As τ grows, the retrieval moves toward the *query-independent retrieval*, which returns one output shared by every query. The spectral family reaches it when every strength is below one. *Along the traces a temperature and an exponent point the same way, without a common scale*, because the level of self-retrieval orders both. For a fixed memory state and τ, the retrieval maps the sphere onto its image, which in two dimensions is a *retrieval curve*. Appendix C proves the statements of this section and gives the constructions behind them.

3.2 Revisiting contrastive learning from the retrieval perspective

Building on the memory module, we revisit the pretext task of contrastive learning from the retrieval perspective and build the loss by analyzing its gradient. For a view z with positive z^+, the gradient $g(z \mid z^+) := \nabla_z \mathcal{L}$ is taken through z alone, with the retrieval of z^+ and the memory state held fixed. We normalize each encoder output $\tilde{z}$ to $z = \tilde{z}/\|\tilde{z}\|$, so the encoder receives $\partial \mathcal{L}/\partial \tilde{z} = \Pi_z g/\|\tilde{z}\|$ with $\Pi_z := I - zz^\top$. Only the *effective gradient* $\Pi_z g$ therefore carries direction. We write g as the sum of two terms, one for each requirement on the retrievals.

3.2.1 Retrieval agreement

Retrieval agreement requires the two views of an image to give similar outputs when both are retrieved from the same memory state with the same gate. The requirement follows from the paired data and can be checked pair by pair. The gradient direction that serves it is the difference between the two retrievals,

$$\chi_\tau(z \mid z^+) := R_\tau(z; M) - R_\tau(z^+; M), \tag{3}$$

which we take as the *positive-guided term* of the gradient. This term compares the two views through the memory rather than directly. Since $R_\tau = \nabla \Phi_\tau$, the term is the gradient in z of the Bregman divergence of Φ_τ between z and z^+. The encoder receives $-\Pi_z \chi_\tau$, which draws z toward z^+ while the view difference stays below a threshold set by the conditioning of the retrieval's Jacobian.

Retrieval agreement constrains the memory state as well as the representations, since both retrievals read the memory state, whether derived from the batch or learned.

The positive-guided term vanishes on the diagonal $z^+ = z$, since $\chi_\tau(z \mid z) = 0$ at every level τ. To first order it is the view difference weighted by the Jacobian of the retrieval, so it acts only along the difference between the two views of an image. Within the range where it contracts the view difference, it drives its own signal to zero, and the encoder can reach that zero by collapsing, which makes the two views coincide. *Comparing each view only with its positive makes the positive-guided term serve retrieval agreement and leaves it blind to collapse.*

3.2.2 Retrieval specificity

The retrieval trace can supply what the positive-guided term lacks. Retrieval specificity requires the retrieval to vary with the view along every direction across the batch, so the representations have to occupy every direction. Representations that all coincide receive the same gradient from a shared memory state and stay together, so the requirement is stated as an instability. The variance along a direction is the uncentered second moment of the representations along it. The weaker grade asks that a direction left empty grow once every sample moves by a vanishing amount into it. The stronger grade also asks that this relative growth and the share of the variance common to the two views stay bounded away from zero as the direction weakens.

The self-guided term is the part of the gradient that a view receives on the diagonal $z^+ = z$. We ask three things of it. It should reach the encoder, which requires its tangential component to be non-zero somewhere. It should make a small transverse perturbation of a degenerate configuration grow. It should be built from retrieval outputs, the only signals the memory module provides. On the diagonal, these outputs are retrievals of the view itself. Retrievals from one memory state differ only in the gate, and along its traces they differ only in level. By Proposition 1, the retrieval retains less of the query at a larger τ wherever it moves. For two levels τ and $\tau + \Delta\tau$ with $\Delta\tau > 0$, the difference $R_\tau(z; M) - R_{\tau+\Delta\tau}(z; M)$ is therefore what the retrieval retains of the view at τ and loses at $\tau + \Delta\tau$. A step along this difference counteracts the loss of the query along the trace. A gradient step moves against the gradient, so we take the self-guided term to be

$$\Delta R(z) := R_{\tau+\Delta\tau}(z; M) - R_\tau(z; M). \tag{4}$$

Taken alone, the self-guided term redistributes variance among directions of the representations. A direction gains variance exactly when its own loss of self-retrieval between the two levels exceeds the average loss of the samples that occupy it. The average weights each sample by how much of it lies along the direction. The trace of the second moment is fixed by normalization to unit length, so a direction gains only what others lose. The weaker grade of retrieval specificity holds exactly when an empty direction wins this comparison.

3.2.3 The loss and its properties

Adding (3) and (4) cancels the common retrieval $R_\tau(z; M)$. Figure 1(c) draws the two terms for a softmax memory in two dimensions, the positive-guided term along the retrieval curve at fixed τ and the self-guided term along the retrieval trace of z.

Definition 3 (Gradient of dual memory retrieval)**.** *On the retrieval traces of one memory state, take the level at τ for the retrieval of z^+ and the level at $\tau + \Delta\tau$ for the retrieval of z. The two levels satisfy $\tau \geq 0$ and $\tau + \Delta\tau \geq 0$, and the interval $\Delta\tau$ is finite and non-zero. The sign of $\Delta\tau$ is not presumed. Put*

$$\begin{aligned} g(z \mid z^+) &:= \underbrace{R_{\tau+\Delta\tau}(z; M) - R_\tau(z; M)}_{\Delta R(z)} + \underbrace{R_\tau(z; M) - R_\tau(z^+; M)}_{\chi_\tau(z|z^+)} \\ &= R_{\tau+\Delta\tau}(z; M) - R_\tau(z^+; M). \end{aligned} \tag{5}$$

The definition gives one candidate construction, whose directional stability depends on the family and the pair of levels. The requirements leave room for others. Taking both levels on the traces of one memory state orders the two levels and gives them a common coordinate. This pair of retrievals from a single memory state is a *dual memory retrieval*, whose training pipeline Figure 1(a) shows.

We give the loss in closed form. The retrieval $R_\tau(z^+; M)$ is constant in z, and the auto-associative retrieval $R_{\tau+\Delta\tau}(z; M)$ is the gradient of its potential $\Phi_{\tau+\Delta\tau}(z; M)$. Then

$$\mathcal{L}_\Phi(z \mid z^+) = \Phi_{\tau+\Delta\tau}(z; M) - \big\langle z, R_\tau(z^+; M)\big\rangle. \tag{6}$$

If the maximal self-retrieval returns the positive itself, $R_0(z^+; M) = z^+$, the loss becomes

$$\mathcal{L} \;=\; \Phi_{\Delta\tau}(z; M) - \langle z, z^+\rangle. \tag{7}$$

With the softmax potential, loss (7) corresponds to InfoNCE (van den Oord et al., 2018; Chen et al., 2020a), and with the spectral potential at $\Delta\tau = 1$ it corresponds to the spectral contrastive loss (HaoChen et al., 2021). A weight $\beta > 0$ on the potential term gives $\mathcal{L}_\beta(z \mid z^+) = \beta\, \Phi_{\tau+\Delta\tau}(z; M) - \big\langle z, R_\tau(z^+; M)\big\rangle$, and $\beta = 1$ is loss (6). Loss (6) differs from (7) in its alignment target, which is the retrieval of z^+ rather than z^+ itself. The pull of z then points to this retrieval. It retains less of z^+ as τ grows and depends on how the memory state relates to the distribution of the representations. *For a learned memory state, however, the retrieval of z^+ also passes to the memory state a gradient that is essential to its training.* When the positive enters the loss as itself, it bypasses a learned memory state, which then receives gradient from the potential alone.

Appendix D proves the statements of Section 3.2.

3.3 Two instances and their relation to existing methods

3.3.1 The softmax family

On the softmax family with stored items $\{m_i\}$, the rows of $K = V$, loss (6) reads

$$\mathcal{L}_\Phi(z \mid z^+) = -\Big\langle z, \sum_i p_i^+\, m_i \Big\rangle + (\tau + \Delta\tau) \log \sum_i \exp \frac{\langle m_i, z\rangle}{\tau + \Delta\tau}, \tag{8}$$

where $p^+ := \mathrm{softmax}_i(\langle m_i, z^+\rangle/\tau)$ are the addressing weights of z^+ at temperature τ, one per stored item. Up to the factor $\tau + \Delta\tau$, the loss is the cross-entropy between p^+ and the addressing weights of z at temperature $\tau + \Delta\tau$, with p^+ held fixed.

The stored items can follow the representations or be learned. With $\{m_i\} = \{z_i\}$, the representations of the current batch, the memory state is recomputed at every step. With $\{m_i\}$ the rows of a linear layer normalized to unit length, the items are prototypes trained by gradient steps. A memory state with all prototypes equal is then a fixed point for any representations.

InfoNCE (van den Oord et al., 2018) is the loss at $\tau = 0$ with the batch representations as stored items and $\tau + \Delta\tau$ as its temperature, except that it also backpropagates through z^+ and $\{m_i\}$, which our gradient holds fixed. DINO (Caron et al., 2021) uses normalized prototypes as stored items and its teacher and student temperatures as the two levels, and adds a momentum teacher and a bias b that centers the teacher logits with a running mean over batches.

3.3.2 The spectral family

On the spectral family with memory state F, loss (6) reads

$$\mathcal{L}_\Phi(z \mid z^+) = -z^\top F^\tau z^+ + \tfrac{1}{2}\, z^\top F^{\tau+\Delta\tau} z. \tag{9}$$

A memory state that follows the representations is the second moment Σ of the current batch, positivized as $F = (\Sigma + \epsilon I)/(1 + d\epsilon)$ with a small $\epsilon > 0$, so that $\mathrm{tr}\, F = 1$ and every eigenvalue lies in $(0, 1)$. A learned memory state is a matrix W trained by gradient steps. At the integer levels $(1, 2)$ the loss then becomes $-z^\top W^\top z^+ + \frac{1}{2}\, z^\top W^\top W z$, the squared error $\frac{1}{2}\|Wz - z^+\|^2$ of a linear predictor up to a constant, which is loss (9) when W is symmetric. A point collapse together with any W that leaves it fixed is then a fixed point of the representations and the memory state.

For a memory state computed from the current batch, the gradient factors as $g = F^\tau(F^{\Delta\tau} z - z^+)$. At $\tau > 0$ the factor F^τ scales each direction by λ^τ, which is small on an empty direction, whereas at $\tau = 0$ it is the identity. The pair of levels therefore decides whether a direction that the representations leave empty grows.

Corollary 1. *Treat the representations as free points on the sphere that move along $-\Pi_z g$, with the memory state recomputed as the second moment of the current batch. Let the two views coincide and occupy n directions equally.* ***(i)*** *For $\tau = 0$ and $\Delta\tau > 0$, every empty direction is unstable for every n and every ϵ.* ***(ii)*** *For $\tau > 0$, $\Delta\tau > 0$ and $n \geq 2$, empty directions are stable once ϵ is small.* ***(iii)*** *For $\tau > 0$, $\Delta\tau < 0$, $\tau + \Delta\tau > 0$ and $n \geq 2$, empty directions are unstable once ϵ is small.*

An approximate model of the dynamics injects variance into a direction in proportion to its current variance. In this model, every level $\tau > 0$ sets a retention floor as $\epsilon \to 0$: the two views share the variance that spreads into a direction only above it.

At $\beta = 1$, Corollary 1(ii) keeps empty directions empty for $\tau > 0$ and $\Delta\tau > 0$ once ϵ is small and two or more directions are occupied. The weight β changes this outcome. On the spectral family the weighted loss is $\mathcal{L}_\beta(z \mid z^+) = -z^\top F^\tau z^+ + \frac{\beta}{2} z^\top F^{\tau+\Delta\tau} z$. For an aligned configuration with n equally occupied directions, $\tau > 0$ and $\Delta\tau > 0$, a weight $\beta > 1$ moves the stability boundary of an empty direction to $n = \beta^{1/\Delta\tau}$ as $\epsilon \to 0$. The direction is then unstable when $n < \beta^{1/\Delta\tau}$ and stable when $n > \beta^{1/\Delta\tau}$. At $\Delta\tau = 0$ the weight makes every empty direction unstable, and at $\tau = 0$ it leaves Corollary 1(i) unchanged.

The spectral contrastive loss (HaoChen et al., 2021) corresponds to levels $(0, 1)$ with the second moment of the batch as memory state, but it acts on features without ℓ_2 normalization. DirectPred (Tian et al., 2021) and DirectCopy (Wang et al., 2021) set a linear predictor to the square root and to the first power of a running second moment plus a small floor, which places them at levels $(1/2, 1)$ and $(1, 2)$. These three methods derive the memory state from the representations, although only the spectral contrastive loss uses negative pairs. BYOL (Grill et al., 2020) and SimSiam (Chen & He, 2021) instead train a two-layer predictor by gradient steps, as in the learned case above.

Appendix B places each existing objective at its pair of levels and analyzes a learned memory state, and Appendix E proves what this section states.

4 Experiments

Our experiments test the statements of Section 3 and do not aim at state-of-the-art accuracy.

4.1 Experiment setting

Datasets and models. We use CIFAR-10 (Krizhevsky, 2009), STL-10 (Coates et al., 2011), ImageNet-100 (IN-100) and ImageNet-1k (IN-1k) (Russakovsky et al., 2015). IN-100 consists of 100 of the first 151 classes of IN-1k by WordNet identifier, which differ from the subset of Tian et al. (2020) and are listed in the supplementary material. We use a ResNet-18 backbone (He et al., 2016) for CIFAR-10 and STL-10 and a ResNet-50 for ImageNet-100 and ImageNet-1k. Representations pass a three-layer projection head and are normalized to unit length before the loss.

Training. For the spectral family we test four pairs of exponents $(\tau, \tau + \Delta\tau)$. For the softmax family the ratio $\tau/(\tau + \Delta\tau)$ is what matters, so we fix $\tau + \Delta\tau$ at a base temperature T, which is 0.5 on CIFAR-10 and 0.1 on the other three datasets. Within a dataset and a family every other training setting is held fixed. The entries marked ‡ are the exception. Their kNN accuracy stayed at or below its first evaluation for 30 epochs, so we stopped them there. The spectral runs of Table 1 weight the potential term of the loss with $\beta = 20$ on CIFAR-10 and STL-10 and $\beta = 40$ on ImageNet-100 and ImageNet-1k. Table 2 adds runs of the four spectral pairs at $\beta = 1$ on STL-10, one seed each, alongside the weighted runs of Table 1.

Metrics. *kNN and Linear-Probe (LP) Accuracy* are top-1 accuracies on frozen backbone features. *Effective Rank (ER)* is the exponential of the entropy of the normalized eigenvalues of the second moment of the representations. *Recovery (REC)* is the mean squared projection of the empty eigenvectors at one checkpoint onto the occupied eigenspace at the next, maximized over consecutive checkpoints. *Cosine (COS)* is the mean cosine similarity between the representations of positive pairs. *Shared Fraction (SF)* is $\mathbb{E}\langle Pz, Pz^+\rangle / \mathbb{E}\|Pz\|^2$ for the projection P onto the 2% of eigendirections with the least variance, and it is undefined when those directions carry no variance. ER and

Table 1: Mean accuracy by pair of levels, family best in bold. Spectral rows weight the potential term, and softmax levels are multiples of T. †One seed. ‡Stopped at epoch 30.

		CIFAR-10		STL-10		IN-100		IN-1k†	
Family	$(\tau, \tau + \Delta\tau)$	kNN	LP	kNN	LP	kNN	LP	kNN	LP
Spectral	$(0, 1)$	88.2	**88.8**	**82.9**	**86.5**	**62.2**	**67.3**	**42.5**	**51.7**
	$(1/2, 1)$	85.5	85.7	72.1	74.1	25.8	33.0	19.1	25.6
	$(1, 1)$	16.7	12.7	23.9‡	20.9‡	2.0‡	3.6‡	1.6‡	2.4‡
	$(1, 2)$	**89.2**	88.6	80.4	84.5	56.1	64.0	39.3	48.2
Softmax	$(0.01T, T)$	**89.7**	**91.1**	**80.8**	85.0	**57.8**	63.7	54.5	**64.3**
	$(0.1T, T)$	80.3	84.2	**80.8**	**85.1**	57.1	**64.0**	**54.9**	62.7
	(T, T)	10.0	10.0	28.6‡	25.4‡	8.3‡	9.3‡	0.1‡	0.1‡

Table 2: Diagnostics on STL-10: spectral family ($d = 2048$) at $\beta = 1$ (one seed) and $\beta = 20$, and softmax family ($d = 128$) with $T = 0.1$. A dash marks an undefined SF. ‡Stopped at epoch 30.

Family	$(\tau, \tau + \Delta\tau)$	ER	REC	COS	SF	kNN
Spectral, $\beta = 1$	$(0, 1)$	57	0.007	0.971	–	83.6
	$(1/2, 1)$	2	0.000	0.995	–	67.1
	$(1, 1)^{\ddagger}$	1	0.000	1.000	–	33.0
	$(1, 2)$	1	0.000	1.000	–	28.2
Spectral, $\beta = 20$	$(0, 1)$	587	0.071	0.949	0.794	82.9
	$(1/2, 1)$	1255	0.380	0.780	0.373	72.1
	$(1, 1)^{\ddagger}$	944	0.350	0.025	0.010	23.9
	$(1, 2)$	30	0.000	0.968	0.710	80.4
Softmax	$(0.01T, T)$	39	0.052	0.961	0.851	80.8
	$(0.1T, T)$	39	0.054	0.960	0.850	80.8
	$(T, T)^{\ddagger}$	1	0.002	1.000	0.053	28.6

REC track the weaker grade of retrieval specificity, SF tracks its stronger grade, and COS tracks retrieval agreement. Unless marked otherwise, every number is a mean over three seeds. Appendix F gives the details of the experiments.

4.2 Experiments with a derived memory update

A zero interval fails to train in both families. Every row with $\Delta\tau = 0$ falls far below the other rows of its family in Table 1. At $\beta = 1$ the gradient at a zero interval is the positive-guided term alone, which is blind to collapse. Both rows of Table 2 with $\Delta\tau = 0$ and $\beta = 1$ collapse to a single direction. With the weight, the spectral pair $(1, 1)$ keeps a high effective rank, because the weight makes every empty direction unstable at $\Delta\tau = 0$. Its positive-pair cosine stays near zero instead.

The smallest τ of each family is best or within one point of the best. By Proposition 1, how much of the positive the alignment target $R_\tau(z^+; M)$ retains is non-increasing in τ. At $\tau = 0$ a memory state derived from the batch returns the positive itself. Counting each family separately, the smallest τ gives the best entry in 12 of the 16 columns of Table 1. In the other four it comes within one point of the best. The spectral pair $(1, 2)$ is more accurate than $(1/2, 1)$ on every dataset, although it retrieves the positive at a lower level and also differs in $\Delta\tau$.

On the spectral family at $\beta = 1$, only the pair with $\tau = 0$ keeps an effective rank above two. By Corollary 1(i) every empty direction is unstable at $\tau = 0$, and in Table 2 only $(0, 1)$ has non-zero recovery at $\beta = 1$. Its single run is also the most accurate spectral run in Table 2. The pairs $(1/2, 1)$ and $(1, 2)$ collapse to one or two directions, with zero recovery. For two or more equally occupied directions, Corollary 1(ii) predicts zero recovery once ϵ is small. At $\beta = 1$ the softmax pairs with $\Delta\tau > 0$ have non-zero recovery as well.

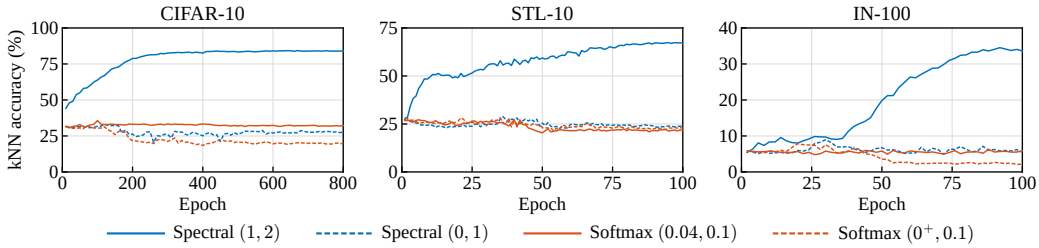


Figure 2: The curves show kNN accuracy during training with a learned memory update. On the softmax family, 0^+ marks the positive entering the loss as itself. One run per configuration.

With the weight, accuracy follows the shared fraction. The weight moves the stability boundary of an empty direction to $\beta^{1/\Delta\tau}$ occupied directions as $\epsilon \to 0$, and every weighted spectral pair ends with a higher effective rank than the same pair at $\beta = 1$. The shared fraction orders the four pairs as their accuracy does on STL-10, where both are measured.

4.3 Experiments with a learned memory update

Within the objectives compared here, a learned memory state needs the positive retrieved through the memory. When the positive enters the loss as itself, it bypasses a learned memory state. The memory state then receives gradient from the potential alone. Every sample pushes every prototype away from itself on the softmax family and W down along its own direction on the spectral family. Figure 2 tracks kNN accuracy during training for both families, with the positive entering the loss as itself or retrieved through the memory. In the first case neither family trains. Retrieving the positive through the memory adds the missing gradient through the alignment target. The argument also covers our unconstrained W and prototypes with trainable gains.

Retrieving the positive through the memory is insufficient. Even then, degenerate configurations remain fixed points in both families. The spectral run at $(1, 2)$ nevertheless trains on all three datasets. The softmax run at $(0.04, 0.1)$ uses the teacher and student temperatures of DINO on one set of prototypes and omits the momentum teacher and centering. Its kNN accuracy stays near or below its first evaluation on all three datasets. Published prototype methods add a mechanism that this run lacks. SwAV sets the bias b by Sinkhorn iterations (Caron et al., 2020), Self-Classifier normalizes the addressing weights over the batch (Amrani et al., 2022), and DINO sets b by centering and adds a momentum teacher (Caron et al., 2021).

5 Conclusion

We model the representation distribution as a memory and build dual memory retrieval on two requirements, retrieval agreement and retrieval specificity. Its positive-guided term serves retrieval agreement, and its self-guided term, along the retrieval trace, serves retrieval specificity. The framework places sample-contrastive methods such as SimCLR, DINO, SimSiam and DirectPred on one comparable axis. Experiments with a derived memory update show that the interval between the two levels is essential, since training with the positive-guided term alone collapses. Within the objectives compared here, training with a learned memory update needs the positive retrieved through the memory, but this is insufficient. Appendix G lists the questions the analysis leaves open, such as the joint dynamics of memory and encoder.

Reproducibility statement

Section 4.1 and Appendix F give the datasets, the training configurations, the evaluation protocol and the number of seeds of every experiment, and anonymized code for the experiments, with the class list of IN-100, is provided as supplementary material. For the theory, Appendix A lists every symbol and hypothesis, maps each result to the hypotheses it uses and locates the proof of every numbered statement of the main text.

AI USE STATEMENT

In this work, we used generative AI tools to assist in formulating and proving mathematical claims, in implementing methods and in analyzing experimental results. We have not used generative AI tools for developing the theoretical framework or proposing hypotheses, and generating synthetic data, cleaning datasets and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools for drafting and editing the paper and its figures. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Elad Amrani, Leonid Karlinsky, and Alex Bronstein. Self-supervised classification network. In *European Conference on Computer Vision (ECCV)*, volume 13691 of *Lecture Notes in Computer Science*, pp. 116–132. Springer, 2022. arXiv:2103.10994.

Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2205.11508.

Mathieu Blondel, André F. T. Martins, and Vlad Niculae. Learning with Fenchel–Young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision (ECCV)*, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

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 (ICCV)*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020a.

Xinlei Chen and Kaiming He. Exploring simple Siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

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 (AISTATS)*. JMLR Workshop and Conference Proceedings, 2011.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann LeCun. On the duality between contrastive and non-contrastive self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2206.02574.

Hariprasath Govindarajan, Per Sidén, Jacob Roll, and Fredrik Lindsten. On partial prototype collapse in the DINO family of self-supervised methods. In *British Machine Vision Conference (BMVC)*, 2024.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Manu Srinath Halvagal, Axel Laborieux, and Friedemann Zenke. Implicit variance regularization in non-contrastive SSL. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. arXiv:2212.04858.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. arXiv:2106.04156.

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)*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020.

John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2110.09348.

Panagiotis Koromilas, Giorgos Bouritsas, Theodoros Giannakopoulos, Mihalis Nicolaou, and Yannis Panagakis. Bridging mini-batch and asymptotic analysis in contrastive learning: From InfoNCE to kernel-based losses. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25276–25301, 2024.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Alexander C. Li, Alexei A. Efros, and Deepak Pathak. Understanding collapse in non-contrastive Siamese representation learning. In *European Conference on Computer Vision (ECCV)*. Springer, 2022.

André F. T. Martins and Ramón F. Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning (ICML)*. PMLR, 2016.

Beren Millidge, Tommaso Salvatori, Yuhang Song, Thomas Lukasiewicz, and Rafal Bogacz. Universal Hopfield networks: A general framework for single-shot associative memory models. In *International Conference on Machine Learning (ICML)*. PMLR, 2022.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K. Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2008.02217.

Pierre H. Richemond, Allison Tam, Yunhao Tang, Florian Strub, Bilal Piot, and Felix Hill. The edge of orthogonality: A simple view of what makes BYOL tick. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015.

Emanuele Sansone and Robin Manhaeve. Unifying self-supervised clustering and energy-based models. *Transactions on Machine Learning Research*, 2025.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2440–2448, 2015. arXiv:1503.08895.

Chenxin Tao, Honghui Wang, Xizhou Zhu, Jiahua Dong, Shiji Song, Gao Huang, and Jifeng Dai. Exploring the equivalence of Siamese self-supervised learning via a unified gradient framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14431–14440, 2022.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision (ECCV)*. Springer, 2020.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10268–10278. PMLR, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 2020.

Xiang Wang, Xinlei Chen, Simon S. Du, and Yuandong Tian. Towards demystifying representation learning with non-contrastive self-supervision. *arXiv preprint arXiv:2110.04947*, 2021.

Xi Weng, Yunhao Ni, Tengwei Song, Jie Luo, Rao Muhammad Anwer, Salman Khan, Fahad Shahbaz Khan, and Lei Huang. Modulate your spectrum in self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In *International Conference on Learning Representations*, 2015. arXiv:1410.3916.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Chaoning Zhang, Kang Zhang, Trung X. Pham, Axi Niu, Zhinan Qiao, Chang D. Yoo, and In So Kweon. Dual temperature helps contrastive learning without many negative samples: Towards understanding and simplifying MoCo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14441–14450, 2022a.

Chaoning Zhang, Kang Zhang, Chenshuang Zhang, Trung X. Pham, Chang D. Yoo, and In So Kweon. How does SimSiam avoid collapse without negative samples? a unified understanding with self-supervised contrastive learning. In *International Conference on Learning Representations (ICLR)*, 2022b. arXiv:2203.16262.

Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. ReSSL: Relational self-supervised learning with weak augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 2543–2555, 2021.

Symbol	Meaning	Defined in
$\mathcal{Z}$, Π_z	representation sphere; projection onto $z^\perp$	Def. 1, §3.2
μ_Z	distribution of the representations on $\mathcal{Z}$	App. A.3
z, z^+, $\tilde{z}$	view and positive; encoder output before normalization	§3.2
g	representation gradient; $\Pi_z g$ is effective	§3.2
M, m_i, $\hat{y}$	memory state (K, V, b, λ); stored item, a row of $K = V$; gate	Def. 1, §3.3.1
k_l, v_l	the lth key and value, rows of K and V	Def. 1
$\tilde{R}_{\tau+\Delta\tau}$	retrieval renormalized off the copies	App. B.2
b	affinity bias, zero by default	Def. 1
q, a, $p(z)$, N	query; affinity vector; addressing weights; number of keys	Def. 1
J_R, Φ	Jacobian of the retrieval; its potential	§3.1.1, Prop. 7
$\rho(z)$	self-retrieval score, $\langle z, R(z)\rangle$	Eq. (2)
$\hat{y}_\tau$, $\mathcal{I}$	gate family; index interval of a generator	Def. 1, 6
R_τ, τ, Γ_τ	retrieval; index of the gate; generator	Def. 1, 6
$\tau \to 0$	maximal self-retrieval, one end of the traces	§3.1.2
$\tau \to \infty$	query-independent retrieval, the other end	§3.1.2
$\bar{v}$	output of the softmax query-independent retrieval	Prop. 11
$\mathcal{A}$, Σ, ν_j, e_j	alignment set; second moment; its spectrum	Def. 7
χ_τ	positive-guided term, $R_\tau(z) - R_\tau(z^+)$	§3.2.1
ΔR	self-guided term, $g(z \mid z)$	Def. 9, Def. 3
τ, $\Delta\tau$	the two levels and the interval between them	Def. 3
D_τ, Ψ	Bregman divergence of Φ_τ; potential-surface difference	Prop. 19, 20
θ, $\dot{\square}$	flow time, since τ indexes the gate; derivative in θ	App. D.9
$\rho_j^{(\tau)}$, $\Delta\rho_j$	directional score; its gap	Def. 10
$\overline{\Delta\rho_j}$	occupancy average of the score gap	Prop. 23
C_j, f_j	cross-view second moment; contraction rate	Def. 11
δ, σ_j^2	view difference; its injection rate	Def. 11, App. A
L0–L3	grades of retrieval specificity	Def. 12
F, λ_j, ϵ	spectral memory state; its spectrum; positivization	Table 3, App. C.4
$\lambda_\varnothing$, λ_{occ}, $\lambda^\star$	empty and occupied eigenvalues; turning point	Prop. 28, Cor. 1
$h(\lambda)$, $\pi_k^{(j)}$	thrust spectrum; occupancy weights	Prop. 26
$\bar{h}$	mean thrust over directions	Prop. 28
p_S, $\hat{a}$, $\hat{p}$	copy mass; affinity and weights without direction j	§E.4
$\bar{p}_l$	marginal usage of a prototype	§B.3
$\bar{z}$	mean representation, $\mathbb{E}[z]$	Thm. 1
d, I, $\mathcal{N}$	representation dimension; identity; normalization onto $\mathcal{Z}$	Def. 1, Prop. 16
ψ_τ, ψ_τ^*, $\mathcal{C}$	potential of the gate; its conjugate; the gate's feasible set	App. C.2, Prop. 6
$\mathcal{U}$, $\mathcal{D}$	open domains of the gate and of the retrieval	Prop. 6, Prop. 18
B_τ	bilinear form of the generator	Prop. 10
Q, Z, H	assignment matrix, its logits and entropy, in SwAV	App. B.3.2
u_l, $\tilde{u}_i$	Sinkhorn factors of prototype l and of sample i, in SwAV	App. B.3.2
$\mathcal{L}$, η	the loss; the step size	Eq. (7), Prop. 16
$J_{\min}$, $J_{\max}$	the two ends of the conditioning bound	App. A.3
r, ϕ	an arbitrary vector; an arbitrary vector field	Prop. 22
x_j, ξ, κ_j	the coordinate $e_j^\top z$; its law; its kurtosis	App. D.12
$A_j^\times$, $\overline{A}_j^\times$	the two parts of the self-guided side of (34)	Prop. 24
U	eigenvectors of F as rows; the e_j under a current-batch memory state	§3.1
n	number of occupied directions	Cor. 1
P	projection onto the weakest 2% of eigendirections	§4.1
P_{occ}, $P_\varnothing$, ζ_j	occupied and empty projections; spread of the spectrum	App. E.1
$\mathcal{S}$	a subspace that contains the representations	Prop. 9
c, γ	constants of the window and of the weak-direction thrust	App. E.3
β, h_β	weight on the potential term; its thrust	§3.2.3, Prop. 29
$\tilde{p}$, Δa_l	weights on the remaining items; the part of the affinity from direction j	App. E.4
$\bar{a}$, $\bar{a}_0$	mean affinity off S, tilted and uniform	App. E.4
ϕ_1, ϕ_2, ω	the two maps of a pullback; the cosine factor	App. G.1
W, w	linear predictor, trained as a memory state or set from F; its eigenvalue	§3.3.2, App. B.4
M^{s}, M^{T}	the memory states retrieved from for z and z^+	App. B.3

	Softmax family	Spectral family
Gate $\hat{y}_\tau = \nabla\psi_\tau$	$\mathrm{softmax}(a/\tau)$	$\mathrm{Diag}(\lambda^\tau)\, a$
Retrieval ($V{=}K$, $b{=}0$)	$K^\top \mathrm{softmax}(Kz/\tau)$	$F^\tau z$
The memory state stores	copies of the representations	directions of F with their strengths
Original name of τ	temperature	exponent
Loss	$\Phi_{\tau+\Delta\tau}(z) - \langle z, R_\tau(z^+)\rangle$	$\frac{1}{2} z^\top F^{\tau+\Delta\tau} z - \langle z, F^\tau z^+\rangle$
Maximal self-retrieval returns	the nearest stored item	the query when $N = d$

Table 3: The two families, side by side.

A Notation and assumptions

A.1 Notation

Four hypotheses are stated in words wherever they are needed rather than labeled: auto-associative retrieval, the gate condition, the choice of family, and differentiability where the retrieval is used.

Table 3 sets the two families side by side, with the keys of each built from the batch. Softmax keys are unit-norm copies of the batch representations, and the softmax potential is $\Phi_\tau(z) = \tau \log \sum_l e^{k_l^\top z/\tau}$. Spectral keys are the eigenvectors of the positivized batch second moment F, whose eigenvalues λ_j lie in $(0, 1)$. Both families have a generator (Appendix C.5), and Appendix B relates each family to existing methods.

A.2 Kinds of hypothesis

Three kinds of hypothesis appear in this paper, with different standing.

Conventions fix what is being analyzed. They are stated in Appendix D.1.

Interface properties are asked of the memory state. Section 3.2 uses only three of them: unit scale, the conditioning bound and identity at the maximal self-retrieval. Its argument therefore holds for every memory state that has these three.

Statistical assumptions are about the data or about the flow. These are the ones that can fail, and Appendix A.5 says what would fail with them.

A fourth item is a *choice* rather than a hypothesis. The current-batch memory state of Section 3.3 lets the growth condition (34) be evaluated in closed form, and Appendix E.5 records what this choice gains and what it costs.

A.3 List of assumptions

Unit scale. The keys have unit-norm rows. Introduced with the softmax gate in Table 3. On the spectral family the corresponding requirement is $0 \prec F \preceq I$. Both bound the scale of the memory state by the scale of the representations, the first key by key and the second at once.

The conditioning bound. The spectrum of the retrieval Jacobian lies in a fixed interval $[J_{\min}, J_{\max}]$ with $J_{\min} > 0$ at every point of the unit ball. Stated in words in Section 3.2.1. By Appendix C.2 it is a statement about how the memory state covers the neighborhood of the queries.

Identity at the maximal self-retrieval. The maximal self-retrieval is the identity, $R_0 = \mathrm{id}$ almost everywhere under μ_Z. Stated in words in Section 3.2.3. Unit scale, the conditioning bound and identity at the maximal self-retrieval are logically independent.

Non-amplification. The retrieval is non-amplifying, $\rho(z) \leq 1$ everywhere. Stated as Proposition 5. It holds automatically on the softmax family and cuts a half-line on the spectral one.

Exchangeable views. The paired distribution is exchangeable, $(z, z^+) \overset{\mathrm{d}}{=} (z^+, z)$. Two views drawn from one augmentation distribution satisfy it. A teacher-student implementation satisfies it only approximately, and Appendix G.1 places such implementations outside the

analysis for a separate reason. Where only equality of the two second moments in each direction is needed, that is said in place.

The injection model. Augmentation injects view difference into direction e_j at a rate σ_j^2 fixed by the source and the augmentation, and independent of the training dynamics. Real augmentation departs from this idealization, and the analysis uses one qualitative consequence of it, that the rate is small on content directions and large on style directions. Proposition 32 takes the rate proportional to the variance instead, with the same ratio σ_j^2/ν_j in every direction.

The moment factorization. The occupancy weights factor, $\mathbb{E}[(e_j^\top z)^2(e_k^\top z)^2] = \nu_j\nu_k$ on occupied directions. For configurations of bounded kurtosis this is an $O(1/n)$ approximation, and concentrated configurations, such as samples at $\pm e_k$, break it. The diagonal term would need $\mathbb{E}[(e_j^\top z)^4] = \nu_j^2$, which on the sphere requires every coordinate to have a fixed magnitude. A Gaussian-like configuration gives $3\nu_j^2$ instead, so each direction underestimates its own occupancy weight by about $2\nu_j$ and the thresholds shift accordingly.

Independent coordinates. In the eigenbasis of Σ the coordinates are independent with zero mean. Direction j equals $\sqrt{\nu_j}\,\xi$ for a ξ whose law is the same in every direction. It is a statement about the data rather than about the memory state. On the sphere, $\sum_j (e_j^\top z)^2 = 1$ and independence give $\sum_j \mathrm{Var}\left((e_j^\top z)^2\right) = 0$. The assumption therefore holds exactly only for coordinates of fixed magnitude and random sign, where the kurtosis κ_j of Appendix E.4 equals one. For other laws, such as a normalized Gaussian, it is an approximation. Appendix E.4 records the class of weak directions it excludes, and uses the zero mean a second time for the sign of the escape rate.

The contraction estimate. On weak directions $f_j \asymp 2\lambda_j^\tau$. Introduced in Appendix E.3. At $\tau = 0$ it holds for small view differences, and it is consistent with the aggregate bound inside the recovery domain. Outside that domain it is unsupported.

A.4 Assumptions used by each result

A.5 Reliance on the statistical assumptions

The four statistical assumptions differ in how safe they are and in how much the results use them.

The contraction estimate is the weakest and is used by exactly one result, Proposition 32, and there only for the upper end of the window. The exclusion of the configurations with $\tau > 0$ and $\Delta\tau > 0$, used in its proof, comes from Corollary 1 and uses the current-batch memory state alone. So the exclusion survives a failure of the contraction estimate, and the value of $\tau_{\max}$ does not.

The injection model is used by Proposition 24(ii) and through it by the window. It fixes the injection rate and is an idealization, and the middle term of (33) is known to exist and is left uncomputed. The two together make the steady-state value of C_j an estimate.

The moment factorization is used by Proposition 28(i), and through it by the rank bound and the isotropy argument in the proof of Corollary 1. The statements about empty directions are independent of it. Its error on the diagonal term is one-sided, since $\mathbb{E}[(e_j^\top z)^4] \geq \nu_j^2$.

The assumption of independent coordinates is used only on the softmax family, by Proposition 34 and Corollary 6. The results on the spectral family are independent of it.

The two arguments that carry the conclusion of Appendix E.3 are independent of all four. That the thrust on weak directions is multiplied by λ^τ is Proposition 31, which follows from the definitions alone. That the tangential recovery threshold is $2(\lambda_{\min}/\lambda_{\max})^\tau$ follows from Proposition 14(ii) under the conditioning bound alone. Both statements hold without the injection model, the moment factorization, the assumption of independent coordinates and the contraction estimate. Both favor a small τ, and both admit a level with $\tau > 0$ at finite ϵ.

Two results carry the load of everything said about weak directions and are worth naming for a different reason. Proposition 8 is what identifies the weakest direction in variance with the weakest in spectrum. It holds because the memory state follows the configuration, independently of the gate.

Result	Hypotheses beyond the definitions
Prop. 27, 31	none
Prop. 1	auto-associative retrieval, zero bias (the generator of Prop. 10)
Lem. 1, Cor. 3	none
Cor. 2	the gradient-path convention
Prop. 16	one gradient map for the whole batch
Prop. 5	unit scale
Prop. 7, 10, 11	auto-associative retrieval, zero bias; every $\lambda_j < 1$ for the spectral limit
Prop. 12	the gate condition on the softmax family; unit scale, non-amplification and $F \neq I$ on the spectral family
Prop. 13	unit scale; a complete basis on the spectral family
Prop. 14(ii)	the conditioning bound
Cor. 4, Lem. 2, Prop. 15	none
Prop. 17	one memory state and one bias across evaluations
Prop. 18, 20	integrability at the relevant points
Prop. 19	auto-associative retrieval; the conditioning bound for strictness
Prop. 21, 22	none
Prop. 23	the configuration lies on $\mathcal{A}$; frozen basis
Prop. 24(i), Prop. 25	exchangeable views
Prop. 24(ii)	exchangeable views, the injection model
Prop. 8, 9	the positivization in the current-batch memory state; memory state is a functional of the configuration
Prop. 26, Cor. 5	current-batch memory state
Prop. 28, Cor. 1	current-batch memory state; the moment factorization only for Prop. 28(i), the rank bound and the isotropy argument
Prop. 29	current-batch memory state, coinciding views
Prop. 30	none; exchangeable views in the remark after it
Prop. 32	current-batch memory state, exchangeable views, the injection model with a rate proportional to the variance, a steady state with one term dropped, the contraction estimate
Prop. 34, Cor. 6	current-batch memory state, coinciding views, independent coordinates; for Cor. 6 also a finite temperature and two or more stored items that are not copies
Thm. 1	softmax gate, unit scale; $\mathbb{E}[z] \neq 0$ or $\mathrm{Cov}(z) \succ 0$ for the collapse statement
Prop. 2	softmax gate
Prop. 3	$W \succeq 0$ for (ii)
Prop. 4	exchangeable views
Prop. 33	none

Proposition 16 is what forces retrieval specificity into grades. Both are consequences of the setting and of the memory update, and both are independent of the four statistical assumptions.

B Relation to existing methods

Table 4 arranges existing objectives by family and by memory update, and this appendix treats one cell at a time. For each method it gives the pair of levels of Section 3.3 and the change that separates the published method from the loss at those levels. The changes fall into two groups. Changes of the first group keep the method at its pair of levels: a candidate set with one stored item removed, a stop-gradient, and a running estimate in place of the current batch. Changes of the second group let a method with a learned memory update train at its levels. Appendices C to E prove the results cited below. Four related methods have no cell in Table 4. UniGrad (Tao et al., 2022) unifies the gradients of Siamese self-supervised methods, and INTL (Weng et al., 2024) belongs to a family of spectral transformations of the embedding that includes whitening and other power functions. NNCLR (Dwibedi et al., 2021) takes the nearest neighbor of a view in a queue of past embeddings as its positive, and ReSSL (Zheng et al., 2021) matches similarity distributions over such a queue at two temperatures.

B.1 Existing objectives by family and memory update

Table 4 splits each family by memory update. Its first group of rows gives the methods, the memory state, the gate, the name of τ, and the retrieval and potential at a general level, the last four shared within a family. The second group gives the two levels, the thrust, the contraction rate and the grade of retrieval specificity that the pair reaches.

Table 4: The two families by memory update. Spanning entries hold for both updates of a family, T is a temperature, and SCL is the spectral contrastive loss.

	Softmax family		Spectral family	
	Learned	Derived	Learned	Derived
Methods	DINO	InfoNCE (SimCLR, MoCo)	predictors (BYOL, SimSiam)	SCL DirectPred DirectCopy
Memory state	trained prototypes	batch items or queue	trained matrix W	second moment F
Gate	softmax		spectral	
Name of τ	temperature		exponent	
R_τ	$K^\top \mathrm{softmax}(Kz/\tau)$		$F^\tau z$	
Φ_τ	$\tau \log \sum_l e^{k_l^\top z/\tau}$		$\frac{1}{2} z^\top F^\tau z$	
$(\tau, \tau+\Delta\tau)$	(τ_t, τ_s)	$(0, T)$	$(1, 2)$	$(0, 1)$ $(1/2, 1)$ $(1, 2)$
Thrust h	no closed form	weak directions only	from unshared variance, at most $\frac{1}{4}$	$\lambda^\tau - \lambda^{\tau+\Delta\tau}$
Rate f_j	no closed form	uniform, with copies	no closed form	λ^τ
Grade reached	undetermined	L1, first clause of L2, under App. E.4	undetermined	L1 at $\tau = 0$, below L1 at $\tau > 0$, both at $\beta = 1$

Assumptions behind each column. The learned softmax column is a proxy with one set of prototypes, zero bias and no centering or Sinkhorn step. Appendices B.3.1 and B.3.2 give the exact placement, which needs two memory states instead. The derived softmax column needs the current pair among the keys. Its grade is L1 and the first clause of L2 under the assumptions of Appendix E.4,

and the second clause of L2 holds only in the approximate model of Appendix E.3, with the uniform contraction rate of Appendix E.4. The positive key of MoCo comes from a momentum encoder, so the identity at the maximal self-retrieval holds there only approximately, and a queue of past steps holds no copy, which leaves the grade of MoCo undetermined. The learned spectral column trains a linear predictor by backpropagation, as BYOL and SimSiam train theirs. Its optimum fits the cross-view second moment, and Appendix B.7 gives its thrust and fixed points. Its grade is undetermined. A predictor followed by normalization, or a nonlinear one, has no generator. The derived spectral column assumes that DirectPred and DirectCopy set a linear predictor that shares an eigenbasis with F and has its spectrum in $(0, 1]$ (Appendix B.4).

The missing closed forms in the learned softmax column have one source. A trained prototype moves independently of the representations, so Proposition 8 fails for it. The weakest direction in variance can then differ from the weakest in spectrum, and the memory state holds no copy of the query. The thrust and the contraction rate therefore have no closed form in this column. By the note after Proposition 9, the representations can also leave a degenerate subspace. This column lies outside the current-batch memory state of Appendix C.4.

B.2 Softmax family with a derived memory: InfoNCE, SimCLR and MoCo

InfoNCE at the maximal self-retrieval. Substituting into Proposition 18 gives $\mathcal{L}_\Phi = \Phi_{\tau+\Delta\tau}(z) - \langle z, z^+\rangle$, which equals a temperature times the standard form of the family whose candidate set keeps both views of the pair. The two agree pointwise, and the variants that drop a view are treated below. Having $R_\tau = \mathrm{id}$ means τ sits at the maximal self-retrieval and the retrieval there returns the query itself. The contraction rate is then uniform across directions, which is what Appendix E.3 asks for. On this family the identity at the maximal self-retrieval needs the keys to hold copies of the pair.

Copies absorb addressing mass along the query. When the memory update includes the current step, both z and z^+ appear among the keys. Write S for those copies and p_S for the mass they receive at $\tau + \Delta\tau$. Write $\tilde{R}_{\tau+\Delta\tau}$ for the retrieval renormalized on the remaining stored items, which is again a softmax-gate retrieval. On the diagonal both kinds of copy equal z, so

$$R_{\tau+\Delta\tau}(z) = p_S z + (1 - p_S)\,\tilde{R}_{\tau+\Delta\tau}(z), \qquad \Pi_z g = (1 - p_S)\,\Pi_z \tilde{R}_{\tau+\Delta\tau}(z) \quad (z = z^+). \tag{10}$$

The tangential signal on the diagonal is the signal from the remaining items, scaled by $1 - p_S$. This is saturation of the gate. The copies have the highest affinities, so a smaller interval gives them more mass, and as $\Delta\tau \to 0$ we get $p_S \to 1$ and the signal dies. That is the positive-guided term of Section 3.2.1 returning. It is controlled by the interval and needs no separate constraint.

Departures of the implementations. Two departures separate the deployed objectives from the loss at these levels. The first is the candidate set. The usual form removes the query itself, and some variants remove the positive as well. By (10) this changes $|S|$ and with it how early the gate saturates at a small interval, and it leaves the levels unchanged. The construction of Section 3.2.3 keeps every stored item and admits either candidate set. Memory states made of parameters or of statistics hold no copy of any observation, so the prototype and the predictor methods lack this term. The second departure is a stop-gradient on z^+ and on the memory. Section 3.2 differentiates through the query alone, which is that routing, so the two agree on the gradient acting on z and differ in what reaches z^+ and the memory.

MoCo departs twice more. Its positive key comes from a momentum encoder, so the maximal self-retrieval returns a copy of the query only approximately. A queue filled at earlier steps holds only items of past steps. The grade in Table 4 therefore holds for InfoNCE on the current batch, and the grade of MoCo is undetermined.

B.3 Softmax family with learned prototypes: DINO, SwAV and Self-Classifier

DINO and SwAV retrieve from trained prototypes twice, once per view, at two temperatures. The published forms hide this structure, so the two subsections below carry out the decomposition. They end at the same place. The gradient is a difference of two retrievals, the retrieval of z^+ runs on a second memory state, and what separates that state from the one the query retrieves from is exactly the mechanism the method adds.

B.3.1 DINO AS A DIFFERENCE OF TWO RETRIEVALS

The query and the memory state. On top of the backbone, the head is a multilayer perceptron, an ℓ_2 normalization and a weight-normalized linear layer with N output units. The representation z of this paper is the output of the normalization, before that last layer. The columns $k_1, \dots, k_N$ of the weight-normalized layer are unit vectors and the logit of unit l is $k_l^\top z$, so those columns are a memory state in the sense of Section 3.1.1: N stored items, each a unit-norm pattern, carried across steps and moved by a gradient step. Reading the head as one module that produces a distribution hides this. Cutting it at the normalization is what makes the query and the memory state two separate objects.

The decomposition. Write z and z^+ for the student and teacher representations, K^{s} and K^{T} for their prototype matrices, and $\tau + \Delta\tau$ and τ for the student and teacher temperatures, where the superscripts mark the branch and not a value of τ. The teacher addressing weights carry a bias b^{T} in the sense of Definition 1, which DINO computes as a running mean of the teacher logits, negated:

$$p_{\tau+\Delta\tau}(z) = \mathrm{softmax}\big(K^{\mathrm{s}} z/(\tau+\Delta\tau)\big), \qquad p_\tau(z^+) = \mathrm{softmax}\big((K^{\mathrm{T}} z^+ + b^{\mathrm{T}})/\tau\big). \tag{11}$$

The loss is the cross-entropy $-\sum_l p_{\tau,l}(z^+) \log p_{\tau+\Delta\tau,l}(z)$. Expanding the logarithm splits each term into a logit and a log-partition function. The partition function does not depend on l and factors out because $p_\tau(z^+)$ sums to one. The query does not depend on l and factors out of the remaining sum. What is left is

$$\mathcal{L}_{\mathrm{DINO}} = \frac{1}{\tau+\Delta\tau}\Big[\Phi_{\tau+\Delta\tau}(z) - \big\langle z,\ \textstyle\sum_l p_{\tau,l}(z^+)\, k_l^{\mathrm{s}} \big\rangle\Big]. \tag{12}$$

Every step is algebra. Centering shapes the weights $p_\tau(z^+)$ and keeps the form.

The second memory state. The bracket in (12) is loss (6) with one change. Its second term weighs the *student* prototypes by addressing weights computed from the *teacher* prototypes and the centering vector. Let M^{s} hold keys K^{s}, values K^{s} and zero bias, and let M^{T} hold keys K^{T}, values K^{s} and bias b^{T}. Then $\sum_l p_{\tau,l}(z^+) k_l^{\mathrm{s}} = R_\tau(z^+; M^{\mathrm{T}})$ and

$$\nabla_z \mathcal{L}_{\mathrm{DINO}} = \frac{1}{\tau+\Delta\tau}\big[R_{\tau+\Delta\tau}(z; M^{\mathrm{s}}) - R_\tau(z^+; M^{\mathrm{T}})\big]. \tag{13}$$

This is an exact identity. The values in the retrieval of z^+ are the student keys because the chain rule puts them there, and the identity holds for any K^{s} and K^{T}. Its cost is the shared traces. The two memory states differ in their keys and in the bias, and a bias is memory state by Section 3.1.1. The two retrievals therefore lie on the traces of two memory states, outside Definition 3. Writing the gradient as a difference of two retrievals of one memory state fails: the true gradient uses K^{s} as values, a separate retrieval of z^+ would use K^{T}, and the two differ by $(K^{\mathrm{s}} - K^{\mathrm{T}})^\top p_\tau(z^+)$. They agree only when $K^{\mathrm{s}} = K^{\mathrm{T}}$. That sharing is the proxy that the learned softmax column of Table 4 records, and Appendix G.1 places it.

B.3.2 SWAV AS A DIFFERENCE OF TWO RETRIEVALS

SwAV keeps one prototype matrix K for both views, so (12) applies with $K^{\mathrm{T}} = K^{\mathrm{s}} = K$. The difference lies in the retrieval of z^+. Instead of a softmax of the teacher logits, the addressing weights are the code that the Sinkhorn step returns, and the decomposition shows that this code is again a softmax of affinities.

The Sinkhorn step maximizes $\mathrm{Tr}(Q^\top K Z) + \tau H(Q)$ over the matrices Q with prescribed row and column sums, where Z collects the batch and τ is the entropy weight. Its solution has the scaling form

$$Q^\star = \mathrm{Diag}(u)\ \exp(KZ/\tau)\,\mathrm{Diag}(\tilde{u}), \tag{14}$$

one positive factor u_l per prototype and one positive factor $\tilde{u}_i$ per sample. SwAV normalizes each column of $Q^\star$ before using it as the addressing weights of that sample, and that normalization divides the column by its own total, so $\tilde{u}_i$ cancels. The addressing weights of sample i are

$$p_{\tau,l}(z_i^+) = \frac{u_l \exp(k_l^\top z_i^+/\tau)}{\sum_{l'} u_{l'} \exp(k_{l'}^\top z_i^+/\tau)} = \mathrm{softmax}_l\Big(\frac{k_l^\top z_i^+ + \tau \log u_l}{\tau}\Big). \tag{15}$$

The code is a softmax of the affinities plus a bias, with the bias $b_l = \tau \log u_l$ carried by the prototype rather than by the sample. In the terms of Definition 1, SwAV retrieves z^+ from a memory state M^{T} whose keys and values are both K, as the query's own are, and whose bias is b. Its gradient is (13) with that M^{T}, and the entropy weight τ is the level at which z^+ is retrieved, while the student temperature is $\tau + \Delta\tau$.

SwAV is therefore closer to Definition 3 than DINO. The two retrievals share keys and values, and only the bias separates them. The bias is still memory state, so the pair lies on the traces of two memory states. The bias is also computed from the current batch, so it couples the samples of a step.

The role of the added mechanism. In both methods the gap between M^{T} and M^{s} is the mechanism: an exponential moving average and a centering vector in DINO, a per-prototype bias from the Sinkhorn step in SwAV. Self-Classifier instead normalizes the addressing weights over the batch (Amrani et al., 2022), which gives each prototype a bias computed from the batch. On this family the maximal self-retrieval returns the nearest prototype, and trained prototypes copy no observation. The retrieval of z^+ therefore differs from z^+ by the distance from the query to the memory state (Appendix C.5).

Each of the three acts on more than the marginal usage $\bar{p}_l := \mathbb{E}_{\mu_Z}[p_l(z)]$. At the fixed point with all prototypes equal every addressing weight is uniform, so the marginal usage is uniform there too, and a constraint that favors uniform usage already holds. A constraint that favors non-uniform usage breaks the fixed point but unbalances the usage, and alone it is insufficient for a non-degenerate memory state. A sufficient mechanism has to constrain the usage and each query's addressing weights together, the separation of the prototypes, or the update itself. Each of these lies outside the retrieval, and which of them suffices remains open.

B.4 Spectral family with a derived memory: the spectral contrastive loss, DirectPred and DirectCopy

The spectral contrastive loss at the maximal self-retrieval. Take the current-batch memory state on the spectral gate, retrieve z^+ at the maximal self-retrieval, and set $\Delta\tau = 1$. The loss (7) is then $\frac{1}{2} z^\top F z - \langle z, z^+ \rangle$ with F from (22). For z' drawn independently of z, twice its expectation is

$$2\,\mathbb{E}[\mathcal{L}] \;=\; -2\,\mathbb{E}\big[\langle z, z^+ \rangle\big] + \frac{\mathbb{E}\big[\langle z, z' \rangle^2\big] + \epsilon}{1 + d\,\epsilon}. \tag{16}$$

As $\epsilon \to 0$ the right side is the spectral contrastive loss of HaoChen et al. (2021), evaluated on normalized representations. The two gradients differ. The spectral contrastive loss differentiates its second term through both z and z', which doubles the contribution of that term, whereas Section 3.2 holds the memory state fixed within a step and differentiates through the query alone. With the path through F removed, the two gradients coincide up to the positivization.

DirectPred sets the predictor from a running second moment. One method in the predictor family sets the predictor directly rather than training it. It takes

$$W = F^{1/2} + \epsilon_{\mathrm{DP}} \lambda_1 I, \qquad \lambda_1 = \max_j \lambda_j, \tag{17}$$

where ϵ_{DP} is added to W and is a different quantity from the positivization ϵ of (22). The predictor (17) shares an eigenbasis with F. When every $w_j := \sqrt{\lambda_j} + \epsilon_{\mathrm{DP}} \lambda_1$ lies in $(0, 1]$, that is when $\sqrt{\lambda_1} + \epsilon_{\mathrm{DP}} \lambda_1 \le 1$, the family $\{W^\tau\}$ has a generator in its own right. Table 4 places the method in the coordinate of F, at $(\tau, \tau + \Delta\tau) = (1/2, 1)$, which is where $W = F^{1/2}$ would sit. The floor $\epsilon_{\mathrm{DP}} \lambda_1$ moves the predictor off that power of F, and on the traces of W itself the two retrievals sit at $\tau = 1$ and $\Delta\tau = 1$. Proposition 26 then gives $h(\lambda) = w(1 - w)$ with $w = \sqrt{\lambda} + \epsilon_{\mathrm{DP}} \lambda_1$, whose peak sits at $\lambda^\star = (\frac{1}{2} - \epsilon_{\mathrm{DP}} \lambda_1)^2$. By Proposition 31 the gradient is $W(Wz - z^+)$, the family with the positive at the maximal self-retrieval after preconditioning by W, and the added term puts a positive floor under that preconditioner. Three effects follow from that one cause. The turning point moves down, which relaxes the rank bound. The thrust on an empty direction rises from zero to $\epsilon_{\mathrm{DP}} \lambda_1 (1 - \epsilon_{\mathrm{DP}} \lambda_1)$, so such a direction can become unstable again. And the contraction rate on a zero-variance direction rises from zero to $\epsilon_{\mathrm{DP}} \lambda_1$, which lowers the retention floor. All three move toward the level at which Appendix E.3 finds the preconditioner equal to the identity. If some

$w_j > 1$, the family no longer satisfies the hypothesis of Proposition 10(ii). The score then loses its monotonicity along the family, the family loses its query-independent retrieval, non-amplification fails, and the thrust turns negative on the strongest mode.

The one change from the loss at $(1/2, 1)$ is the memory update. The second moment that sets W is an exponential moving average of the representations that supply z^+ rather than of the current batch, and z^+ itself comes from a momentum encoder. Both are derived memory updates in the sense of Section 3.1.1, a running estimate in place of a batch rewritten at every step. DirectCopy sets $W = F$ instead, which by the expansion below puts its two retrievals at $(\tau, \tau + \Delta\tau) = (1, 2)$.

B.5 Spectral family with a learned predictor: BYOL and SimSiam

The predictor family locks the interval. Expanding the squared error, $\frac{1}{2}\|Wz - z^+\|^2 = \frac{1}{2}z^\top W^\top Wz - \langle z, W^\top z^+\rangle + \frac{1}{2}\|z^+\|^2$, the first term is the potential of the retrieval of z. Taking $W = F^\tau$ puts the other level at 2τ, so $\Delta\tau = \tau$. The squared term locks the interval to τ itself. Setting $\tau = 0$ therefore also sets $\Delta\tau = 0$ and the two levels coincide, so this family can only occupy Corollary 1(ii). On configurations with two or more occupied directions, empty directions are stable once ϵ is small, and a lost direction stays lost. The contraction rate λ^τ vanishes on weak directions, so the retention floor is positive once ϵ is small. Neither clause of L2 holds. All of this is for a predictor set to a power of F. For a linear predictor trained by backpropagation the optimum fits the cross-view second moment instead, and Appendix B.7 gives its thrust and its fixed points.

Departures of BYOL and SimSiam. Three things separate the deployed methods from the loss at $(1, 2)$. The predictor is a two-layer perceptron with a nonlinearity rather than a power of F. The placement is the linear idealization. Even a linear predictor trained on the squared error at fixed representations reaches the least-squares fit of Appendix B.7, which is not a power of F in general. The normalization sits after the predictor rather than before it, so the deployed objective is not loss (6) at any pair of levels. And z^+ comes from a momentum encoder in BYOL and passes a stop-gradient in SimSiam. The first two leave the deployed predictor without a generator.

B.6 Learned prototypes on the softmax family

This subsection treats the stored items of the softmax family as prototypes trained by gradient steps. When the positive enters the loss as itself, the global optimum of the prototypes has all of them equal.

The convention of Appendix D.1 is lifted for this subsection only. We treat the memory state as an argument of the loss and study the sign structure of $\nabla_K \mathcal{L}$. The joint dynamics of memory and encoder remain outside the scope.

Definition 4 (One-sided and two-sided losses)**.** *Assume an auto-associative retrieval, a softmax gate and unit scale. Put*

$$\mathcal{L}_1 := -\langle z, z^+\rangle + \Phi_K(z), \qquad \mathcal{L}_2 := -\big\langle z,\ K^\top \mathrm{sg}[p_\tau(z^+)]\big\rangle + \Phi_K(z), \tag{18}$$

where p_τ are the addressing weights of the retrieval at τ and Φ_K is the potential at $\tau + \Delta\tau$ of the memory state with keys K. In $\mathcal{L}_1$ the memory state appears only in the potential. In $\mathcal{L}_2$ it appears in both terms. In $\mathcal{L}_1$ the positive enters the loss as itself, as in loss (7)*. It coincides with $\mathcal{L}_2$ at $\tau = 0$ when the maximal self-retrieval returns z^+, as for a memory state that holds copies of the pair. For trained prototypes the maximal self-retrieval returns the nearest prototype instead, and $\mathcal{L}_1$ is the loss whose alignment target bypasses the memory.*

Proposition 2 (Sign of the one-sided gradient)**.** *$\partial\mathcal{L}_1/\partial k_l = p_l(z)\, z$ with $p_l(z) > 0$. Every sample therefore pushes every prototype away from itself.*

Proof. In $\mathcal{L}_1$ the memory state appears only in Φ_K, and differentiating gives the expression. A softmax gate has full support, so $p_l > 0$. □

Theorem 1 (Global optimum of the one-sided loss)**.** *Put $\bar{z} := \mathbb{E}_{\mu_Z}[z]$. Then*

$$\mathbb{E}\big[\Phi_K(z)\big] \ \geq\ (\tau + \Delta\tau) \log \sum\nolimits_l e^{k_l^\top \bar{z}/(\tau+\Delta\tau)}, \tag{19}$$

with equality exactly when $k_l - k_{l'} \in \ker \mathrm{Cov}(z)$ for every pair, which for $\mathrm{Cov}(z) \succ 0$ means $k_1 = \cdots = k_N$. Under the unit-norm constraint, when $\bar{z} \neq 0$ the right side has the unique minimizer $k_l \equiv -\bar{z}/\|\bar{z}\|$, with value $(\tau + \Delta\tau)\log N - \|\bar{z}\|$, and that configuration attains equality in (19)*, so it is the unique global optimum. When $\bar{z} = 0$ the right side equals $(\tau + \Delta\tau)\log N$ for every K, and the global optima are the configurations meeting the equality condition. So whenever $\bar{z} \neq 0$ or $\mathrm{Cov}(z) \succ 0$, the global optima are configurations in which all prototypes coincide.*

Proof. As a function of the affinity vector, Φ_K is a log-sum-exp and hence convex, so Jensen's inequality gives $\mathbb{E}[\Phi_K] \geq \Phi_K(K\bar{z})$, which is (19). A log-sum-exp is affine along $\mathbf{1}$ and strictly convex on the orthogonal complement, so equality holds exactly when $K(z - \bar{z}) \in \mathrm{span}\,\mathbf{1}$ almost surely, that is when $(k_l - k_{l'})^\top (z - \bar{z}) = 0$ for every pair and μ_Z-almost every z. The support of $z - \bar{z}$ spans the range of $\mathrm{Cov}(z)$, so this is $k_l - k_{l'} \in \ker \mathrm{Cov}(z)$. For the minimization, the right side is strictly increasing in each $k_l^\top \bar{z}$, and on the unit sphere $k_l^\top \bar{z} \geq -\|\bar{z}\|$ with equality only at $k_l = -\bar{z}/\|\bar{z}\|$. A global optimum has to reach this minimum, so for $\bar{z} \neq 0$ it is that configuration. □

Collapse of the memory is therefore a global minimum, which a better initialization, a smaller learning rate or an earlier stop leaves in place. The statement is about K with the representation distribution held fixed. A collapsed memory gives every query the same addressing weights, so the retrieval at $\tau + \Delta\tau$ becomes a constant vector. If the memory stays collapsed and the representations move as free points, the self-guided term on $\mathcal{A}$ is a constant vector, and by Corollary 3 the flow on the sphere has a single attractor. That is a source of collapse risk. Whether memory and encoder evolving together must collapse the representations is a question about their joint dynamics, which Appendix G.1 places outside the analysis.

The two-sided loss has a different sign structure. $\partial\mathcal{L}_2/\partial k_l = (p_l(z) - p_{\tau,l}(z^+))\, z$, so $\sum_l \partial\mathcal{L}_2/\partial k_l = 0$. Each sample pulls the prototypes that $p_\tau(z^+)$ weights more heavily toward itself, and pushes away the prototypes that $p(z)$ weights more heavily. The net drift vanishes before the projection onto unit-norm prototypes.

Within the two losses of Definition 4, retrieving the positive through the memory is necessary but insufficient. When all prototypes coincide both sets of addressing weights are uniform and the gradient vanishes, so a degenerate configuration is still a fixed point.

B.7 A LEARNED PREDICTOR ON THE SPECTRAL FAMILY

The same question on the spectral family. The memory state is now a matrix W moved by gradient steps, a learned memory update, rather than set from F. The convention is lifted as in Appendix B.6: the representation distribution is held fixed, and only the gradient and the optimum in W are examined. Take the eigenbasis $\{e_j\}$ of Σ with x_j, x_j^+ and C_j as in Definition 11, and call a direction with $\nu_j > 0$ occupied.

Definition 5 (One-sided and two-sided losses on the spectral family)**.** *For $W \in \mathbb{R}^{d\times d}$ put*

$$\mathcal{L}_1^{\mathrm{sp}} := -\langle z, z^+\rangle + \tfrac{1}{2} z^\top W z \quad (W = W^\top), \qquad \mathcal{L}_2^{\mathrm{sp}} := \tfrac{1}{2}\|Wz - z^+\|^2 - \tfrac{1}{2}\|z^+\|^2. \tag{20}$$

In $\mathcal{L}_1^{\mathrm{sp}}$ the memory state appears only in the potential and z^+ is retrieved at the maximal self-retrieval, so the two levels are $(0, 1)$ with W in the role of F. $\mathcal{L}_2^{\mathrm{sp}}$ is the squared error of Appendix B.5 without its constant. It retrieves z^+ as $W^\top z^+$ and has the potential $\frac{1}{2} z^\top W^\top W z$, so the two levels are $(1, 2)$.

Proposition 3 (Gradient and optima of the one-sided spectral loss)**.** ***(i)*** *$\partial\,\mathbb{E}[\mathcal{L}_1^{\mathrm{sp}}]/\partial W = \frac{1}{2}\Sigma \succeq 0$, which is independent of W, and on one sample it is $\frac{1}{2} z z^\top$. Every sample pushes W down along its own direction.* ***(ii)*** *Let $W \succeq 0$. The global optima of $\mathbb{E}[\mathcal{L}_1^{\mathrm{sp}}]$ in W are exactly the $W \succeq 0$ with $W\Sigma = 0$. At any of them $Wz = 0$ almost surely, so the self-guided term on $\mathcal{A}$ is $Wz - z = -z$. It is radial, and L0 fails on every configuration. A point collapse at z_*, together with any such W, is a fixed point of representations and memory together.*

Proof. (i) $\mathbb{E}[\frac{1}{2} z^\top W z] = \frac{1}{2}\,\mathrm{tr}(W\Sigma)$, and the other term does not involve W. (ii) For $W, \Sigma \succeq 0$, $\mathrm{tr}(W\Sigma) = \|\Sigma^{1/2} W^{1/2}\|_F^2 \geq 0$, with equality exactly when $\Sigma W = 0$, and $W = 0$ attains it. Almost

every z lies in the range of Σ, on which W vanishes. At a point collapse $\Sigma = z_* z_*^\top$, so $Wz_* = 0$ and the self-guided term is $-z_*$. A gradient step pushes the eigenvalue of W along z_* below zero, and projecting back onto $W \succeq 0$ restores W. □

Without the constraint $W \succeq 0$, $\mathbb{E}[\mathcal{L}_1^{\mathrm{sp}}]$ is unbounded below in W and the gradient flow gives $W = W_0 - \frac{\theta}{2}\Sigma$. On an occupied direction the thrust is $-\Delta\rho_j = 1 - (W_0)_{jj} + \frac{\theta}{2}\nu_j$, and once θ is large the last term dominates. A direction with more variance then gets more thrust. Under a current-batch memory state the order is the reverse: at $\tau = 0$ the thrust $1 - \lambda^{\Delta\tau}$ of Proposition 26 decreases with the eigenvalue, which increases with the variance (Proposition 8).

Proposition 4 (Optimum of the two-sided spectral loss)**.** *Assume exchangeable views.* ***(i)*** $\mathbb{E}[\mathcal{L}_2^{\mathrm{sp}}]$ *is convex in* W*, with gradient* $W\Sigma - \mathbb{E}[z^+ z^\top]$*, and its global optima are the* W *with* $W\Sigma = \mathbb{E}[z^+ z^\top]$*. On an occupied direction* $We_j = \mathbb{E}[x_j z^+]/\nu_j$*. When* $\nu_j = 0$*,* $(W\Sigma - \mathbb{E}[z^+ z^\top])e_j = 0$*, so* We_j *changes only through its initial value and any weight decay.* ***(ii)*** *For any* W*,* $\rho_j^{(1)} = e_j^\top W e_j$ *and* $\rho_j^{(2)} = \|We_j\|^2$*. At an optimum, on an occupied direction,*

$$-\Delta\rho_j = \frac{C_j}{\nu_j}\Big(1 - \frac{C_j}{\nu_j}\Big) - \sum_{k \neq j}\Big(\frac{\mathbb{E}[x_k^+ x_j]}{\nu_j}\Big)^2 \;\le\; \frac{C_j}{\nu_j}\Big(1 - \frac{C_j}{\nu_j}\Big) \;\le\; \frac{1}{4}, \tag{21}$$

where the shared fraction C_j/ν_j *lies in* $[-1, 1]$*. The thrust can be positive only where* $0 < C_j/\nu_j < 1$*. When* W *shares an eigenbasis with* Σ *the sum vanishes, and the thrust has the form* $w(1-w)$ *of Appendix B.4 with* $w = C_j/\nu_j$*.* ***(iii)*** *On* $\mathcal{A}$*,* $\mathbb{E}[z^+ z^\top] = \Sigma$ *and an optimum has* $Wz = z$ *almost surely. The self-guided term* $W^\top(Wz - z)$ *then vanishes, and the two retrievals coincide on the data as they do at* $\Delta\tau = 0$*. Every aligned configuration together with its optimal memory is a fixed point of representations and memory together, and a point collapse at* z_* *with any* W *satisfying* $Wz_* = z_*$ *is one of them.*

Proof. (i) $\mathbb{E}[\frac{1}{2}\|Wz - z^+\|^2]$ is a convex quadratic in W, and differentiating gives the gradient. Multiplying it on the right by e_j gives $\nu_j We_j = \mathbb{E}[z^+ x_j]$, and when $\nu_j = 0$ both sides vanish because $x_j = 0$ almost surely. (ii) By Definition 10 and $\Sigma e_j = \nu_j e_j$, $\rho_j^{(1)} = (We_j)^\top \mathbb{E}[z x_j]/\nu_j = e_j^\top W e_j$ and $\rho_j^{(2)} = e_j^\top W^\top W e_j$. Substituting the column from (i) gives $e_j^\top W e_j = C_j/\nu_j$ and $\|We_j\|^2 = \sum_k (\mathbb{E}[x_k^+ x_j]/\nu_j)^2$. The bound holds because the sum is non-negative and $w(1-w) \le \frac{1}{4}$. Exchangeable views and the Cauchy–Schwarz inequality give $|C_j| \le \nu_j$, and $w(1-w) \le 0$ on $[-1, 0] \cup \{1\}$. (iii) With $\mathbb{E}[z^+ z^\top] = \Sigma$, part (i) gives $We_j = e_j$ on every occupied direction, so $Wz = \sum_j x_j We_j = z$. On $\mathcal{A}$ the positive-guided term $W^\top(z - z^+)$ vanishes as well, and so does the gradient $W\Sigma - \Sigma$ in W. □

The two-sided thrust comes from the unshared variance. The readings below assume that the memory stays near its optimum, which is the weakest link of this subsection. Part (i) of both propositions holds at any memory. Proposition 22 is linear in ϕ, so the thrust in (21) is the contribution of the self-guided term to $\dot\nu_j/\nu_j$ on every configuration. It can be positive only on directions the views share in part, and it is positive there when W shares an eigenbasis with Σ. As C_j/ν_j tends to one the thrust tends to zero, and on $\mathcal{A}$ it vanishes by (iii). Unlike the feedback of the current-batch memory state, it fades as the views align.

Empty directions keep their initial columns. Give a direction with $\nu_j = 0$ a vanishing variance while the memory stays at the optimum of an aligned configuration. By (iii), $\Delta\rho(z)$ vanishes on the data, and Proposition 23 gives the growth rate $2(e_j^\top W e_j - \|We_j\|^2)$, which depends only on the untrained column We_j. With unit columns at random initialization it is about -2, and the direction shrinks. With an initial column we_j and $0 < w < 1$ it is $2w(1-w) > 0$. With the memory held, L1 is decided by the initialization and the weight decay rather than by the loss. If the memory is refitted as the variance is injected, the column receives a gradient and tends to e_j by (iii), and the growth rate tends to zero.

Retrieving the positive through the memory is again necessary but insufficient. Within the two losses of Definition 5, the one-sided optimum makes the self-guided term radial on the data when $W \succeq 0$, and without that constraint it gives more thrust to directions with more variance. The two-sided memory has a finite optimum set by the configuration, but an aligned configuration with its optimal memory is a fixed point. The two families match here. On the softmax family the degenerate fixed point has all prototypes equal, and on the spectral family it has a memory that is the identity on the data, so that the two retrievals coincide. Their reach differs. With all prototypes equal both retrievals return the same vector for every query, so the gradients of $\mathcal{L}_2$ in K and in z vanish for every configuration. The spectral fixed points of Proposition 4(iii) are aligned configurations.

Relation to the idealization of Appendix B.5. Taking $W = F^\tau$ there describes a predictor set from an estimate of F, a derived memory update as in DirectPred and DirectCopy, and the conclusions there hold for it. Under a learned memory update the optimum of $\mathcal{L}_2^{\mathrm{sp}}$ is in general the solution of Proposition 4(i) rather than a power of F. Along an eigendirection of Σ the idealized eigenvalue λ_j^τ is set by the variance, while the fitted $e_j^\top W e_j = C_j/\nu_j$ is set by how much the two views share. The optimal memory is still a function of the configuration, but it follows the shared fraction rather than the variance, and directions that stay unoccupied keep their initial columns.

C Memory state, retrieval and retrieval trace

C.1 The memory state

Definition 1 (Retrieval)**.** *Let the memory state M hold N keys $K \in \mathbb{R}^{N\times d}$, values $V \in \mathbb{R}^{N\times d}$, a bias $b \in \mathbb{R}^N$ and any strengths that the gate reads. Here d is the representation dimension and N the addressing dimension, one entry per key. For a query $q \in \mathcal{Z} := \mathbb{S}^{d-1}$ and $\tau > 0$, the retrieval consists of addressing and response,*

$$p_\tau := \hat{y}_\tau(Kq + b), \qquad R_\tau(q; M) := V^\top p_\tau. \tag{1}$$

The affinities $a := Kq + b$ score the match between the query and each key. The gate $\hat{y}_\tau$ converts them into the addressing weights p_τ at a scale set by τ.

The memory module of Section 3.1 keeps four components apart: the memory state M, the gate, the retrieval R and the memory update. Two things follow. The gate family is fixed apart from the memory state, and on the spectral family the gate reads the strengths stored in it. One memory state therefore admits many retrievals, one for each τ. Any comparison of two families has to say whether it compares memory states or gates. The retrieval depends on history only through the current memory state, so it is a fixed map within one evaluation.

Two independent axes. A memory state is *single-step* when it depends only on the current observation and *cross-step* when it retains earlier ones. Independently of that, a memory state stores either *items*, vectors that can be addressed one by one, or a *statistic* of the representation distribution. All four combinations occur. A list of the representations in the current batch is single-step and stores items. A queue and a set of trained prototypes are cross-step and store items. A second moment estimated on the current batch is single-step and stores a statistic, and a running estimate of it is cross-step and stores a statistic.

Both axes are independent of size and of trainability. A cross-step second moment on $\mathbb{R}^d$ has $O(d^2)$ components, a queue persists across many steps and still stores items, and a set of trained prototypes is a cross-step memory state that stores items.

The cost of cross-step information. A cross-step memory state of items carries information from earlier steps. A queue holds encodings produced by earlier encoders, and trained prototypes hold parameters shaped by earlier gradients. Cross-step information is paid for in staleness. A cross-step statistic is refitted at every step by the memory update, which moves the cost onto the memory update. Single-step memory state has no staleness, since every stored item comes from the current encoder, and its cost is that it holds the current step alone. Which cost to pay is a choice of memory update, and it remains open here. Appendix E.5 records what each choice does to the results of Section 3.3.

Permutation invariance of the keys. Permuting the rows of K, the rows of V and the entries of b together leaves the retrieval unchanged. This needs the gate to be equivariant under coordinate permutation. When the gate is linear, any invertible refactorization of the product leaves the retrieval unchanged as well. Any comparison between two memory states must therefore be a comparison of permutation invariants.

C.2 THE RETRIEVAL AND ITS POTENTIAL

Definition 1 leaves the gate open. The results below that need a potential or a symmetric generator assume the *gate condition*: each gate is the gradient of a convex function on $\mathbb{R}^N$, $\hat{y}_\tau = \nabla\psi_\tau$. Both families meet it (Table 3).

Proposition 5 (Non-amplification)**.** *Assume unit scale, which on the spectral family reads $0 \prec F \preceq I$. Say that a retrieval is **non-amplifying** when $\rho(z) \leq 1$ for every query. The softmax gate satisfies it at every parameter value, a negative temperature included. When $F \neq I$ the spectral gate satisfies it exactly when $\tau \geq 0$, and when $F = I$ it satisfies it at every τ.*

Proof. On the softmax family $\rho = \sum_l p_l \langle z, k_l\rangle \leq \max_l \langle z, k_l\rangle \leq 1$, since the addressing weights form a convex combination and the keys have unit norm. On the spectral family $\rho = z^\top F^\tau z$ with every $\lambda_j \in (0, 1]$ by $0 \prec F \preceq I$, and this is at most 1 for every query exactly when $F^\tau \preceq I$, which for $F \neq I$ means $\tau \geq 0$. □

Non-amplification is automatic on one family and cuts a half-line on the other, and Proposition 12 uses it to fix the lower end of τ on the spectral family. The softmax family takes its lower end from the gate condition instead.

Proposition 6 (Equivalent forms of the gate condition)**.** *Let $\mathcal{U} \subseteq \mathbb{R}^N$ be open, convex and simply connected and let $\hat{y} \in C^1(\mathcal{U}; \mathbb{R}^N)$ be single-valued. The following are equivalent.*

1. *$\hat{y} = \nabla\psi$ for a convex $\psi \in C^2(\mathcal{U})$.*
2. *$\mathbf{J}_{\hat{y}}$ is symmetric and positive semidefinite on $\mathcal{U}$.*
3. *$\hat{y}(a) = \arg\max_p\{\langle p, a\rangle - \psi^*(p)\}$, with ψ^* the convex conjugate of ψ.*

Proof. (1)⇔(2) is the Poincaré lemma with the second-order characterization of convexity. (1)⇒(3) is conjugate duality, the maximiser of $\langle p, a\rangle - \psi^*(p)$ being $\nabla\psi(a)$. (3)⇒(1) holds because a single-valued argmax equals $\nabla\psi^{**}$, which is the gradient of a convex function. □

A convex gate chooses addressing weights that follow the affinities while paying a convex penalty, and τ sets the strength of that penalty (Martins & Astudillo, 2016; Blondel et al., 2020).

Proposition 7 (Integrability of auto-associative retrieval)**.** *For a retrieval of Definition 1 whose gate meets the gate condition, with $V = K$ and $b = 0$, $R(z) = K^\top\nabla\psi(Kz) = \nabla_z[\psi(Kz)]$, so $\Phi(z) = \psi(Kz)$, which is convex as a convex function composed with a linear map, and $J_R = \nabla^2\Phi = K^\top\mathbf{J}_{\hat{y}}K$. The potentials of the two families are the ones listed in Table 3.*

Proof. Differentiate $\psi(Kz)$ by the chain rule. The two potentials follow by substituting the gates of Table 3. □

Along the traces the two gate potentials move in opposite directions, ψ_τ increasing on the softmax family and decreasing on the spectral family, while the self-retrieval score $\rho_\tau = \langle a, \nabla\psi_\tau(a)\rangle$ decreases on both. Ordered by the gate potential, a temperature and an exponent point opposite ways. Ordered by the level of self-retrieval, as in Section 3.1.2, they point the same way.

The retrieval is the primitive object here and the potential is derived from it. The potential exists only where the retrieval is integrable, and Proposition 7 gives one sufficient condition for that. Representations are constrained to the sphere, so the gradient is taken in the ambient space first and projected afterwards by Lemma 1.

Different addressing weights can give one retrieval. Value aggregation maps the addressing weights from $\mathbb{R}^N$ back to $\mathbb{R}^d$. When $N > \operatorname{rank} K$ the map $K^\top$ has a non-trivial null space, so two different sets of addressing weights can give one retrieval. A softmax memory may store far more items than dimensions, and then this happens routinely. A spectral gate with an eigenbasis has $N \le d$, and its addressing weights determine the retrieval uniquely.

The conditioning bound is a coverage condition. By Proposition 7 the memory state enters every first-order quantity only through $K^\top \mathbf{J}_{\hat{y}} K$, and it passes two gates of its own on the way. A direction outside the row space of K is invisible to the affinities, and a direction in the null space of $\mathbf{J}_{\hat{y}}$ passes the aggregation unchanged. On the softmax family $\mathbf{J}_{\hat{y}} = (\operatorname{Diag} p - pp^\top)/\tau$ has $\mathbf{1}$ in its null space, so a direction equidistant from every key receives zero first-order restoring force. The conditioning bound excludes such directions near the queries. It is therefore a statement about how the memory state covers their neighborhood. On the spectral family it is just $F \succ 0$.

C.3 Two families of convex gates

A gate can be produced from a convex program. Take a closed convex function ψ^* and a non-empty closed convex set $\mathcal{C}$, and put $\hat{y}(a) := \arg\max_{p \in \mathcal{C}} \{\langle p, a\rangle - \psi^*(p)\}$. Strict convexity alone gives at most one maximizer. When ψ^* is strongly convex on $\mathcal{C}$ the maximizer exists and is unique for every a, and $\hat{y}$ is the gradient of a convex function with a Lipschitz gradient, so it is differentiable almost everywhere with a symmetric positive semidefinite Jacobian. Both gates below have strongly convex penalties and are smooth. A scale parameter enters as an overall positive multiple of ψ^*. The softmax gate is the case of negative entropy on the probability simplex. The spectral gate is the case of a weighted quadratic on all of $\mathbb{R}^N$.

On the spectral family the memory state stores directions and strengths. Keys and values are both U, whose rows are the eigenvectors of F with eigenvalues λ_j, and the gate is $\operatorname{Diag}(\lambda^\tau)$. So the memory state stores each eigenvector of F with its strength λ_j, the strength enters through the gate rather than through the affinities, and moving τ leaves the memory state unchanged. A linear retrieval F^τ admits other factorizations, and this paper uses this one throughout.

The simplex feasible set makes the gate non-separable. The constraint $\mathbf{1}^\top p = 1$ couples the coordinates of the addressing weights, so the gate couples the coordinates. The spectral gate is separable, with every key retrieved on its own, and Appendix C.1 explains why that suffices. When the memory state stores items, each stored item is a single observation, so distributional information enters through the gate at retrieval time, and the gate couples the items. A memory state that stores a statistic absorbs that information at update time, so its gate can act on each direction alone.

C.4 Current-batch memory state

The memory state is fixed by three clauses. **Derivation.** The memory state is derived from the current configuration. On the softmax family the keys are copies of the current batch of representations, including the pair being retrieved. On the spectral family the memory state F is built from the batch second moment. **Drawing.** Stored items that are not copies are drawn like the queries, independently of one another and of the pair. This treats the two views of any other image as independent draws. **Positivization.** The spectral memory state is positivized,

$$F \;:=\; \frac{\mathbb{E}_{\mu_Z}[zz^\top] + \epsilon I}{1 + d\,\epsilon}, \qquad \epsilon > 0. \tag{22}$$

Since $\|z\| = 1$ we have $\operatorname{tr} \mathbb{E}_{\mu_Z}[zz^\top] = 1$, so $\operatorname{tr} F = 1$ and $0 \prec F \preceq I$, and with $d \ge 2$ every eigenvalue lies below one. The bound $0 \prec F \preceq I$ is unit scale on the spectral family. Unit scale and identity at the maximal self-retrieval also hold on the softmax family, since the keys are unit-norm copies and the query is one of them.

Three roles of the positivization. A second moment estimated from a finite batch is generally rank deficient, and λ_j^τ is discontinuous at $\tau = 0$ on a null direction. A positive ϵ makes $F \succ 0$, so the maximal self-retrieval is a limit rather than a convention. It also gives every empty direction the

eigenvalue $\lambda_\varnothing = \epsilon/(1 + d\epsilon)$, which every endpoint comparison below uses. And the denominator $1 + d\epsilon$ is independent of τ, so both ends of the trace stay fixed. The first job belongs to the interface properties of Appendix A.2. The other two belong to this appendix.

Proposition 8 (Spectrum of the current-batch memory state)**.** *Under the positivization* (22)*, F is an affine function of Σ, so the two share an eigenbasis and the eigenvalue on e_j is $\lambda_j = (\nu_j+\epsilon)/(1+d\epsilon)$, a strictly increasing function of ν_j.*

Proof. Substitute (22). □

This proposition is the whole support for Corollary 5 below, and through it for every claim in Section 3.3 about weak directions. It is what turns the weakest direction in variance into the weakest direction in spectrum. Resistance to collapse on this family is therefore a feedback loop through the memory update. The gate only decides how that feedback is distributed across the spectrum.

Proposition 9 (Invariance of degenerate subspaces)**.** *Let the configuration lie in a subspace $\mathcal{S}$, and suppose either that the values lie in $\mathcal{S}$, or that the gate is a function of an operator having $\mathcal{S}$ as an invariant subspace. Then $\Delta R(\mathcal{S}) \subseteq \mathcal{S}$ and $\chi_\tau(\mathcal{S} \times \mathcal{S}) \subseteq \mathcal{S}$, so the flow preserves $\mathcal{S}$. In particular a direction with $\nu_j = 0$ stays at zero. Under a current-batch memory state the softmax family satisfies the first alternative and the spectral family the second.*

Proof. In the first case $R(z) = \sum_l \hat{y}_l(\cdot)v_l$ lies in the span of the values. In the second, functional calculus gives F^τ the same invariant subspace, so $R_\tau(z) = F^\tau z$ stays in $\mathcal{S}$. Normalization and tangential projection preserve $\mathcal{S}$. □

A trainable memory state escapes this: a trainable stored item placed in $\mathcal{S}^\perp$ lets ΔR leave $\mathcal{S}$, so the rank can grow. Point collapse remains the exception by Proposition 16, where one map carries every sample and the span stays one-dimensional. The same pair of levels therefore behaves differently with a memory state that stores a statistic and with trainable stored items.

Two further statistical assumptions of Appendix A.3 appear below, the moment factorization and the assumption of independent coordinates. The drawing clause of the current-batch memory state covers the memory state, so the second is a statement about the data alone.

C.5 The retrieval trace

Definition 2 (Retrieval trace)**.** *For a memory state M and a query q, the curve $\tau \mapsto R_\tau(q; M)$ over $\tau \geq 0$ is the* retrieval trace *of q.*

Definition 6 (Generator)**.** *Let $\mathcal{I} \subseteq (0, \infty)$ be an interval and let $\{R_\tau(\cdot; M)\}_{\tau\in\mathcal{I}}$ be retrievals from one memory state, with $\tau \mapsto R_\tau(q; M)$ differentiable for each q. Symmetric positive semidefinite matrices $\Gamma_\tau(q; M)$ are a* generator *of the retrievals when*

$$\partial_\tau R_\tau(q; M) \;=\; -\,\Gamma_\tau(q; M)\, q \qquad (\forall q \in \mathcal{Z},\ \forall \tau \in \mathcal{I}). \tag{23}$$

Both families have a generator (Proposition 10 below), and the proof of Proposition 1 uses only this property.

Proposition 1. *For both families, along the retrieval trace of any query q, the self-retrieval score $\rho_\tau(q; M)$ is non-increasing in $\tau \in [0, \infty)$ and decreases strictly wherever the retrieval moves.*

Proof. In $\rho_\tau(z) = \langle z, R_\tau(z)\rangle$ the query is independent of τ, so $\partial_\tau\rho_\tau(z) = \langle z, \partial_\tau R_\tau(z)\rangle$. Substituting Definition 6 gives $-z^\top\Gamma_\tau(z)z$, which is non-positive by positive semidefiniteness. Equality forces $z^\top\Gamma_\tau z = \|\Gamma_\tau^{1/2}z\|^2 = 0$, hence $\Gamma_\tau^{1/2}z = 0$ and $\Gamma_\tau z = \Gamma_\tau^{1/2}(\Gamma_\tau^{1/2}z) = 0$, which is $\partial_\tau R_\tau(z) = 0$. The converse is immediate. □

Only the projection is monotone. Proposition 1 governs the component of $R_\tau(z)$ along the fixed direction z rather than the retrieval trace itself. On the softmax family the retrieval is a convex combination of stored items. An intermediate τ can put the mass on stored items that point away from the query. The trace can turn sharply, and $\|R_\tau\|$ can fall and then rise. The maximal self-retrieval returns a stored item of unit norm and the query-independent retrieval returns $\bar{v}$. The residual norm

can also fall and rise, because $\|z - R_\tau(z)\|^2 = 1 - 2\rho_\tau(z) + \|R_\tau(z)\|^2$ has its second and third terms moving in opposite directions. Proposition 1 is therefore a statement about the projection, which a drawing of the trace leaves hidden.

The role of positive semidefiniteness. The sign in Proposition 1 needs a positive semidefinite generator. Take the softmax family, write a for the affinities and p_τ for the addressing weights. For an auto-associative retrieval with zero bias, $\partial_\tau \rho_\tau = -\frac{1}{\tau^2} \operatorname{Var}_{p_\tau}(k^\top z)$, which is non-positive whatever the memory state and wherever the query sits. Decoupling values from keys turns it into $-\frac{1}{\tau^2} \operatorname{Cov}_{p_\tau}(v^\top z,\, k^\top z)$. That is a covariance between the overlaps of two different sets of vectors with the query, whose sign is indefinite. Adding a bias turns it into $-\frac{1}{\tau^2}[\operatorname{Var}_{p_\tau}(k^\top z) + \operatorname{Cov}_{p_\tau}(k^\top z,\, b)]$, whose second term can be negative and can dominate. An auto-associative retrieval with zero bias therefore guarantees a generator on the softmax family, and dropping either can remove it. A method that adds a running statistic to the affinities loses this guarantee. Treating it as if it had a generator is an idealization that has to be declared.

Proposition 10 (Generators of the two families)**.** *Let $b = 0$ and $V = K$, so that $a = Kz$ and $R_\tau(z) = K^\top \hat{y}_\tau(a)$. Suppose the gate family has a generator for the addressing weights, meaning symmetric positive semidefinite $B_\tau(a)$ with $\partial_\tau \hat{y}_\tau(a) = -B_\tau(a)a$. Then the retrievals $\{R_\tau\}$ have the generator $\Gamma_\tau(z) = K^\top B_\tau(a) K$. Two families satisfy the hypothesis. **(i)** A scale family $\hat{y}_\tau(a) = \hat{y}(a/\tau)$ has $B_\tau(a) = \mathbf{J}_{\hat{y}_\tau}(a)/\tau$ and $\Gamma_\tau = J_{R_\tau}/\tau$, and the softmax family is of this kind. **(ii)** A spectral power family $\hat{y}_\tau(a) = \operatorname{Diag}(\lambda^\tau)a$ has $B_\tau = \operatorname{Diag}(\lambda^\tau \log \frac{1}{\lambda})$ and $\Gamma_\tau = F^\tau \log(F^{-1})$.*

Proof. $\partial_\tau R_\tau(z) = K^\top \partial_\tau \hat{y}_\tau(a) = -K^\top B_\tau(a) K z$, which is the form of Definition 6. The matrix $K^\top B_\tau K$ is symmetric, and $r^\top K^\top B_\tau K r = (Kr)^\top B_\tau(Kr) \geq 0$ for every r. For (i), $\mathbf{J}_{\hat{y}_\tau}(a) = \nabla \hat{y}(a/\tau)/\tau$ and $\partial_\tau \hat{y}_\tau(a) = \nabla \hat{y}(a/\tau) \cdot (-a/\tau^2) = -\frac{1}{\tau}\mathbf{J}_{\hat{y}_\tau}(a)a$, and the symmetry and positive semidefiniteness of $\mathbf{J}_{\hat{y}_\tau}$ follow from the gate condition by Proposition 6. For (ii), $\partial_\tau \hat{y}_\tau(a) = \operatorname{Diag}(\lambda^\tau \log \lambda)a$, the matrix is diagonal and so symmetric, and $\lambda \leq 1$ gives $\log(1/\lambda) \geq 0$. □

Congruence carries symmetry and positive semidefiniteness from the addressing weights to the retrieval, so the three demands of Definition 6 are inherited one by one.

Proposition 11 (Query independence at the lowest level)**.** *On the softmax family the addressing weights tend to uniform as $\tau \to \infty$, so $R_\tau(z) \to \bar{v} := \frac{1}{N}\sum_l k_l$. On the spectral family, if every $\lambda_j < 1$, then $R_\tau(z) = F^\tau z \to 0$. Both limits are independent of the query.*

Proof. Take the limits pointwise. □

The two query-independent retrievals differ in value and agree in property. The spectral limit is the origin because every eigenvalue tends to zero together. If some $\lambda_j = 1$, the spectral retrieval tends instead to the projection onto that eigenspace, and that end depends on the query. A current-batch memory state with $d \geq 2$ gives every $\lambda_j < 1$ (Appendix C.4).

Proposition 12 (Lower bound on τ)**.** *On both families $\tau \geq 0$, and the maximal self-retrieval $\tau = 0$ lies in the domain as a limit. The two lower ends have different sources. On the softmax family a negative temperature makes the Jacobian of the softmax negative semidefinite, so the map violates the gate condition, while non-amplification holds at every $\tau \neq 0$. On the spectral family $\operatorname{Diag}(\lambda^\tau)$ meets the gate condition at every τ, and non-amplification excludes $\tau < 0$ when $F \neq I$.*

Proof. On the softmax family $\mathbf{J}_{\hat{y}_\tau}(a) = (\operatorname{Diag} p - pp^\top)/\tau$ with $\operatorname{Diag} p - pp^\top \succeq 0$, so a negative τ gives a negative semidefinite Jacobian and violates the gate condition. The addressing weights still lie on the simplex and the retrieval is still a convex combination of unit keys, so $\rho \leq 1$ holds regardless. On the spectral family apply Proposition 5. There unit scale gives $F \succ 0$, which makes $\tau \mapsto F^\tau$ continuous on $[0, \infty)$, so $F^0 = I$ arrives as a limit. □

The domain is derived rather than imposed, from a different condition on each family. Definition 6 places $\mathcal{I}$ inside $(0, \infty)$ only so that the generator is defined everywhere, and Proposition 12 excludes $\tau < 0$. On both families a negative τ still gives a formally valid family. On the softmax family its generator for the addressing weights, $(\operatorname{Diag} p - pp^\top)/\tau^2$, is still positive semidefinite, but the map

violates the gate condition. On the spectral family with $F \neq I$, F^τ exceeds the identity along every direction with $\lambda_j < 1$, and the retrieval amplifies the query instead of returning it.

Proposition 13 (Score of the maximal self-retrieval)**.** *On the softmax family under unit scale, as $\tau \to 0^+$ the addressing weights concentrate on the keys of largest affinity, so $\rho \to \max_l \langle z, k_l \rangle = 1 - \frac{1}{2} \min_l \|z - k_l\|^2$. On the spectral family at $\tau = 0$ with a complete basis, $R_0 = \mathrm{id}$ and $\rho \equiv 1$ for every query.*

Proof. Take the limits pointwise. On the sphere $\|z - k_l\|^2 = 2 - 2\langle z, k_l \rangle$. □

The shortfall is the distance from the query to the memory state. On the softmax family the maximal self-retrieval returns only stored items and pays that distance. On the spectral family with $N = d$ the maximal self-retrieval returns every query itself, because $F^0 = I$ bypasses the memory state. Which end of a trace is which is fixed by the trace. What the maximal self-retrieval returns is fixed in addition by where the query sits relative to the memory state.

The order of levels and the content of a retrieval. Proposition 1 compares the retrievals of one query at two levels, and the content of the maximal self-retrieval is a separate matter. A query outside the support of the memory state can be met there by a value pointing almost the other way, and Proposition 13 says how far.

Concentration and the order of levels. The gate acts on the addressing weights, and concentrated weights retrieve one stored item. Whether that item moves with the query depends on the memory state. When the memory state stores items, the item of largest affinity follows the query. Concentrating is then retrieving the query itself, and a lower temperature sits nearer the maximal self-retrieval. When it stores a statistic, the leading direction is fixed, so concentrating sends every query to one direction, and a higher exponent sits nearer the query-independent retrieval. One notion of concentration thus produces opposite effects on the two families, because concentration ignores the query. The level of self-retrieval is measured relative to the query, and Proposition 1 treats the two families alike. That is why one orientation covers both.

C.6 Retrieval traces and retrieval curves

In two dimensions, as in Figure 1(b) and (c), view the image of $(z, \tau) \mapsto R_\tau(z)$ as a surface carrying two families of curves. Fixing τ gives $R_\tau(\mathcal{Z})$, the *retrieval curve at fixed τ*, whose points come from different queries at one level. Fixing z gives the retrieval trace of that query. The two families form a coordinate system on the surface, and the two terms of Definition 3 lie along its two coordinate directions. The positive-guided term is a secant of a retrieval curve at fixed τ: both z and z^+ lie in $\mathcal{Z}$, so their retrievals at τ lie on one such curve. The self-guided term is a secant of the retrieval trace of z. Both statements are exact, and in higher dimensions they hold with the image $R_\tau(\mathcal{Z})$ in place of the retrieval curve.

A retrieval curve at fixed τ differs from a level set: the score and the norm both vary along it, and only τ is constant. The level sets of Φ_τ are different objects, and the retrieval is everywhere orthogonal to those.

The two ends are visible in this picture. On the spectral family with $N = d$ the retrieval curve at $\tau = 0$ is $\mathcal{Z}$ itself. On the softmax family it degenerates to the N stored items.

D Retrieval requirements and the loss

D.1 Conventions and the effective gradient

Two conventions are in force. Each excludes something, and the two exclusions are of different kinds.

One gradient path per representation. A representation carries gradient to the encoder only where it acts as the query. Elsewhere it is held at a stop-gradient. This is what makes $R_\tau(z^+)$ a

constant vector in z. Propositions 18 and 19 both use that, and the split of Definition 9 needs it to be a split of a gradient in z.

The convention constrains the path rather than the outcome, and the memory update, rather than the loss, decides how many paths a representation has. When the memory state is a list of current representations there are two, the query path and the memory path. A parameterized memory state has only the first. Closing the memory path puts the two memory updates inside one analysis, at the cost of differing from the exact gradient of a current-batch implementation. The attraction term is the same either way, and the repulsion term differs.

Closing the gradient path through the retrieval of z^+ is a different matter, and it has a consequence worth recording. Without the stop-gradient, (25) gives $-\Pi_{z^+}\nabla_{z^+}\mathcal{L}_\Phi = +\Pi_{z^+}J_{R_\tau}(z^+)^\top z$, which pulls z^+ toward z for an auto-associative retrieval, and it is the only update z^+ receives. The potential at $\tau + \Delta\tau$ involves z alone, so the repulsion acts on z alone. After symmetrizing the loss each representation is attracted twice and repelled once.

Fixed memory state within one gradient computation. The memory state stays fixed while the gradient is taken.

Lemma 1 (Tangential projection of the gradient)**.** *Let $\tilde{z}$ be the unnormalized encoder output and $z = \tilde{z}/\|\tilde{z}\|$. Then $\partial z/\partial\tilde{z} = \Pi_z/\|\tilde{z}\|$ with $\Pi_z = I - zz^\top$, so $\partial\mathcal{L}/\partial\tilde{z} = \Pi_z g/\|\tilde{z}\|$.*

Proof. $\partial\tilde{z}/\partial\tilde{z} = I$ and $\partial\|\tilde{z}\|/\partial\tilde{z} = \tilde{z}^\top/\|\tilde{z}\|$, and the quotient rule gives $(I - zz^\top)/\|\tilde{z}\|$. □

The radial component is annihilated exactly. The factor $1/\|\tilde{z}\|$ is a step size introduced by the normalization and only rescales the step.

Corollary 2 (Equivalence of distance and inner-product losses)**.** *Let ξ be a vector independent of z. The losses $\frac{1}{2}\|z - \mathrm{sg}[\xi]\|^2$ and $-\langle z, \mathrm{sg}[\xi]\rangle$ deliver the same effective gradient at every point.*

Proof. The first gives $\nabla_z = z - \xi$ and $\Pi_z(z - \xi) = -\Pi_z\xi$. The second gives $\nabla_z = -\xi$ and $\Pi_z(-\xi) = -\Pi_z\xi$. □

So on normalized representations a comparison against a stopped alignment target has one effective gradient, whichever of the two forms is written, and the results of this paper hold for both.

Corollary 3 (Attractor of a constant gradient)**.** *Let $\nabla_z\mathcal{L} = -\xi$ with $\xi \neq 0$ independent of z. The flow $\dot{z} = \Pi_z\xi$ on the sphere has exactly two fixed points, $\pm\xi/\|\xi\|$, the first stable and the second unstable.*

Proof. $\Pi_z\xi = 0$ holds exactly when ξ is parallel to z. Linearizing at the two points gives the stability. □

This is what Appendix D.4 lifts to a whole configuration. On a point-collapsed configuration every self-guided term takes one value at every sample, so the update is a rigid motion and the collapse survives it.

D.2 Configurations

Definition 7 (Configuration, alignment set, degeneracy)**.** *A batch of paired representations $\{(z_i, z_i^+)\}$ is a* configuration*. The* alignment set *$\mathcal{A}$ collects the configurations with $z_i = z_i^+$ for every i. The second moment of a configuration is $\Sigma := \mathbb{E}[zz^\top]$, with eigenpairs (ν_j, e_j) and $\sum_j \nu_j = 1$. A configuration is* degenerate *when Σ has rank below d, and* point collapsed *when all z_i agree. On the sphere the first is weaker than the second.*

An uncentered second moment. A direction can therefore be occupied by an offset that every sample shares. This is a property of the memory state rather than a defect of the measure. Discounting a shared offset takes more than centering the memory state, because the maximal self-retrieval is the identity and still returns z. The query and the retrieval of the positive have to move to centered coordinates as well, which gives an affine retrieval. That is a different retrieval construction rather

than a different measure of the same Σ. Mixing the two, a centered memory state with an uncentered retrieval of the positive, is an error.

Rank one versus point collapse. On the sphere $\operatorname{rank}\Sigma = 1$ gives only $z_i \in \{\pm e_1\}$. A rigid-body argument, in which all samples are carried by one map, needs point collapse rather than rank one. On an antipodal configuration a gradient that takes the same value at every sample already raises the rank from one to two.

D.3 The positive-guided term

Definition 8 (Positive-guided term)**.** *For $\tau \geq 0$, put $\chi_\tau(z \mid z^+) := R_\tau(z) - R_\tau(z^+)$.*

Proposition 14 (Local properties of the positive-guided term)**.** *Take χ_τ of Definition 8 and write $\delta := z - z^+$. **(i)** $\chi_\tau(z \mid z) = 0$ for every z and every τ. If J_{R_τ} is locally Lipschitz then $\chi_\tau = J_{R_\tau}(z)\delta + O(\|\delta\|^2)$, and under C^1 alone the remainder is only $o(\|\delta\|)$. **(ii)** What reaches the encoder is $-\Pi_z\chi_\tau$, and*

$$\big\langle \delta,\ \Pi_z\chi_\tau \big\rangle = \big\langle \delta, \chi_\tau \big\rangle - \tfrac{1}{2}\|\delta\|^2 \big\langle z, \chi_\tau \big\rangle. \tag{24}$$

Under the conditioning bound the left side is positive whenever $\|\delta\| < 2J_{\min}/J_{\max}$.

Proof. (i) Evaluating at a single point cancels the two terms. The expansion is Taylor's theorem. (ii) Π_z is symmetric, and $\|z\| = \|z^+\| = 1$ gives $\langle \delta, z\rangle = \frac{1}{2}\|\delta\|^2$, which is (24). The segment from z^+ to z lies in the unit ball and $\chi_\tau = \int_0^1 J_{R_\tau}(z^+ + s\delta)\,\mathrm{d}s\ \delta$, so the conditioning bound gives $\langle \delta, \chi_\tau\rangle \geq J_{\min}\|\delta\|^2$ and $|\langle z, \chi_\tau\rangle| \leq \|\chi_\tau\| \leq J_{\max}\|\delta\|$, so the right side of (24) is at least $\|\delta\|^2(J_{\min} - \frac{J_{\max}}{2}\|\delta\|)$. □

Corollary 4 (Vanishing on the alignment set)**.** *At every level, the positive-guided term vanishes pairwise on every configuration in $\mathcal{A}$. Point collapse is one such configuration.*

Proof. Proposition 14(i). □

Convexity and the projected gradient. For an auto-associative retrieval Φ_τ is convex, so $\langle z - z^+, \chi_\tau\rangle \geq 0$. That inequality holds in the ambient space. By Lemma 1 the encoder receives the projected term instead, and the second term of (24) can take either sign. Beyond the bound of Proposition 14(ii), contraction is not guaranteed. The positive-guided term supplies the pull that retrieval agreement asks for, and a bound on the view difference keeps that pull in what reaches the encoder. Appendix E.2 evaluates $2J_{\min}/J_{\max}$ on the spectral gate, where it becomes $2(\lambda_{\min}/\lambda_{\max})^\tau$.

Vanishing on the whole alignment set. Every positive-guided term vanishes when the views coincide, by Definition 9, so that property alone distinguishes no candidate. Corollary 4 adds that χ_τ vanishes on the whole alignment set, spread out or collapsed alike, and at every level.

D.4 The self-guided term

Definition 9 (Self-guided term and positive-guided term)**.** *Let the training signal give a representation gradient $g(z \mid z^+)$ at a fixed memory state, with the diagonal $\{(z, z)\}$ in its domain. Put $\Delta R(z) := g(z \mid z)$. The* self-guided term *is ΔR, and the* positive-guided term *is $g - \Delta R$.*

The memory state generally moves with the configuration. The split is taken at one evaluation, with M given, so $\Delta R(z) = g(z \mid z; M)$. Wherever we say that a quantity depends on the query alone, we mean at a given memory state.

Lemma 2 (Uniqueness of the split)**.** *If $g = g_1 + g_2$ with g_1 independent of z^+ and g_2 vanishing on the diagonal, then $g_1 = \Delta R$ and $g_2 = g - \Delta R$.*

Proof. Evaluate on the diagonal: $g(z \mid z) = g_1(z) + g_2(z \mid z) = g_1(z)$. □

Proposition 15 (Restriction to the alignment set)**.** *On a configuration in $\mathcal{A}$, the gradient takes the value $\Delta R(z_i)$ at each sample, independently of the positive-guided term. Two training signals with the same self-guided term therefore induce the same flow on $\mathcal{A}$.*

Proof. $z_i = z_i^+$ gives $g(z_i \mid z_i^+) = g(z_i \mid z_i) = \Delta R(z_i)$. □

Scope of the split. Lemma 2 is one line, and the statement built on it needs more. Reading retrieval specificity as a statement about the self-guided term needs the endpoint to lie on $\mathcal{A}$, which needs retrieval agreement to be met and the retrieval to be injective. The self-guided term can drive representations along directions outside the control of the positive-guided term. The endpoint then leaves $\mathcal{A}$, and retrieval agreement has failed. Proposition 25 tests this condition.

Proposition 16 (Invariance of the point-collapsed set)**.** *Consider the update of the representations with one gradient map g for the whole batch, $z_i \mapsto \mathcal{N}\big(z_i - \eta\, \Pi_{z_i} g(z_i \mid z_i^+)\big)$, where $\mathcal{N}$ normalizes. If the configuration is point collapsed, with $z_i = z_*$ for every i, then all z_i remain equal after the update and the second moment stays rank one. The conclusion holds for every memory state shared across the batch and every self-guided term.*

Proof. On a point-collapsed configuration $z_i = z_i^+ = z_*$ for every i, so $g(z_i \mid z_i^+) = \Delta R(z_*)$ is the same for every i. The update map is then the same map for every sample, and it has one image point. □

Invariance and motion. The proposition makes the set invariant while the common point z_* may keep travelling on the sphere. A self-guided term therefore keeps a point-collapsed configuration collapsed, and Corollary 3 is the single-point version of this. The obstacle is structural and holds for every memory state and every self-guided term.

Invariance and transverse stability. An invariant set can be unstable under transverse perturbation, and resistance to collapse asks exactly for that. The origin of a scalar flow with a positive growth rate is invariant and unstable. Proposition 16 excludes one phrasing of the requirement, an outward velocity at the degenerate point itself, and the requirement survives in another. The test moves to a neighborhood or to the limit of a vanishing direction, as in Definition 12.

D.5 The three constraints

C1, reaching the encoder. What reaches the encoder is $-\Pi_z \Delta R$. By Lemma 1 only the tangential part arrives, so the requirement is that $\Pi_z \Delta R$ is non-zero somewhere on $\mathcal{A}$, a statement about $\Pi_z \Delta R$ rather than about ΔR. Definition 12 grades whether degenerate configurations receive such a signal.

C2, transverse instability. After one update $z_i' \in \operatorname{span}\{z_i, \Delta R(z_i)\}$, so the rank grows only if the self-guided term has a component outside the current span, which differs from asking the term to vary across samples. A term that is constant across samples but points outside the span raises the rank, and a term that varies across samples inside the span keeps it. On a point-collapsed configuration the span stays one-dimensional by Proposition 16, so the constraint is stated in graded form. Degenerate configurations are asked to be linearly unstable in the transverse directions. That instability is asked to persist as the configuration degenerates further. Definition 12 makes this quantitative.

C3, built from retrieval outputs. The self-guided term has to be assembled from what the retrieval provides. A term drawn from elsewhere enters through an interface outside this framework.

A constraint on the positive-guided side. Proposition 15 assumes that the configuration lies on $\mathcal{A}$, which needs retrieval agreement and an injective retrieval. The positive-guided term then has to restore the view difference in every direction, so every direction exerts a restoring force on it. That constraint and C2 fall on the same directions, and Proposition 25 compares them.

D.6 The gradient of dual memory retrieval

Definition 3 (Gradient of dual memory retrieval)**.** *On the retrieval traces of one memory state, take the level at τ for the retrieval of z^+ and the level at $\tau + \Delta\tau$ for the retrieval of z. The two levels*

satisfy $\tau \geq 0$ and $\tau + \Delta\tau \geq 0$, and the interval $\Delta\tau$ is finite and non-zero. The sign of $\Delta\tau$ is not presumed. Put

$$g(z \mid z^+) := \underbrace{R_{\tau+\Delta\tau}(z; M) - R_\tau(z; M)}_{\Delta R(z)} + \underbrace{R_\tau(z; M) - R_\tau(z^+; M)}_{\chi_\tau(z|z^+)} \\ = R_{\tau+\Delta\tau}(z; M) - R_\tau(z^+; M). \tag{5}$$

Here τ records only the level at which z^+ is retrieved, and $\Delta\tau$ records only the interval between the two levels. A positive $\Delta\tau$ retrieves z at a lower level, nearer the query-independent retrieval, and a negative one at a higher level, nearer the maximal self-retrieval. The analysis fixes neither sign in advance. In the approximate model of Appendix E.3, Proposition 32 settles the sign on the spectral family as $\epsilon \to 0$ and leaves it open at finite ϵ. A generic level is also written τ. The displacement ΔR is a finite difference over a finite interval rather than a tangent vector, so the identities below are exact rather than first-order expansions.

Proposition 17 (Reduction to the gate on the diagonal)**.** *Let ΔR be built from retrieval values of this layer, with every evaluation sharing one memory state M and one affinity bias. Then on the diagonal each retrieval acts on the same query, the same memory state and the same bias, so the only remaining argument is the gate. Hence ΔR is a function of the values taken by one query under several gates.*

Proof. By Definition 1 a retrieval has query, memory state, bias and gate as its arguments. The first three agree on the diagonal. □

Degeneracy of the self-guided term. The displacement vanishes as a useful signal in two situations, and each is blocked by a different constraint. If both levels sit at the query-independent retrieval then R_τ is a constant map and the displacement is identically zero. If the memory state degenerates until the retrieval is independent of the gate, the displacement is zero as well. A memory state that stores copies of the current configuration does this under point collapse, because all keys coincide. A memory state that stores a statistic behaves differently. Under point collapse it retains one direction and the retrieval still depends on the gate. But the displacement is then parallel to the query, and C1 removes it. So C2 blocks one kind of memory state and C1 blocks the other, which agrees with Proposition 16.

Weights on the two terms. A gate multiplied by a positive number is again a gate, so weighting one side replaces its gate and leaves the memory state unchanged. Separate coefficients on the positive-guided and self-guided terms therefore fall outside Definition 3 while still meeting the three constraints. Proposition 29 gives the effect of the weight β of Section 3.2.3 on the spectral family.

The radial half of the displacement. By Proposition 1, $\langle z, \Delta R(z)\rangle$ equals $\rho_{\tau+\Delta\tau}(z) - \rho_\tau(z)$, which is non-zero for a moving retrieval and has the sign opposite to $\Delta\tau$. Lemma 1 removes this radial half. C1 asks the displacement for a component transverse to the query, which Section 3.1 leaves open. The radial part still matters: Proposition 23 shows that it is the common average in the evolution of each variance, the quantity whose sign is fixed and whose spread is at issue.

D.7 The loss and its Bregman form

Proposition 18 (Existence of a scalar loss)**.** *Let $\mathcal{D} \subseteq \mathbb{R}^d$ be open and simply connected with $R_{\tau+\Delta\tau} \in C^1(\mathcal{D}; \mathbb{R}^d)$. Since $R_\tau(z^+)$ is a constant vector in z, we have $J_z g = J_{R_{\tau+\Delta\tau}}$, so g is the gradient of a scalar if and only if $R_{\tau+\Delta\tau}$ is integrable. When $R_{\tau+\Delta\tau} = \nabla\Phi_{\tau+\Delta\tau}$ the antiderivatives are*

$$\mathcal{L}_\Phi(z \mid z^+) = \Phi_{\tau+\Delta\tau}(z) - \big\langle z, R_\tau(z^+)\big\rangle + \text{const}(z^+). \tag{25}$$

Proof. The z^+ term is constant in z, so the Jacobian of g reduces to $J_{R_{\tau+\Delta\tau}}$. Apply the integrability criterion of Appendix C.2. □

Realizability constrains the retrieval at $\tau + \Delta\tau$ alone. The retrieval at τ may be non-integrable, of another form, or even discontinuous. Auto-associative retrievals on the traces are integrable in any case.

Proposition 19 (Bregman form of the positive-guided term)**.** *Put* $D_\tau(z\|z^+) := \Phi_\tau(z) - \Phi_\tau(z^+) - \langle R_\tau(z^+), z - z^+\rangle$*. Then* $\chi_\tau = \nabla_z D_\tau$*, and* D_τ *is a Bregman divergence, so* $D_\tau \geq 0$ *with* $D_\tau(z\|z) = 0$*. Under the conditioning bound it vanishes only at* $z = z^+$*.*

Proof. Differentiating gives $\nabla_z D_\tau = R_\tau(z) - R_\tau(z^+)$. By the integrability criterion $\nabla^2\Phi_\tau = J_{R_\tau} = K^\top \mathbf{J}_{\hat{y}} K \succeq 0$, so Φ_τ is convex. □

Proposition 20 (Potential-surface difference)**.** *Let the retrievals at both levels have potentials and put* $\Psi(z) := \Phi_{\tau+\Delta\tau}(z) - \Phi_\tau(z)$*. Then*

$$\mathcal{L}_\Phi(z \mid z^+) = D_\tau(z\|z^+) + \Psi(z) + \text{const}(z^+), \qquad \nabla\Psi = \Delta R. \tag{26}$$

Proof. Substitute $\Phi_\tau(z) = D_\tau(z\|z^+) + \Phi_\tau(z^+) + \langle R_\tau(z^+), z - z^+\rangle$ into (25) and collect the terms that depend on z^+ alone. □

Equation (26) is Definition 9 written one layer up. The divergence carries the positive-guided term and the potential-surface difference carries the self-guided term. The divergence takes its reference point at z^+, the view that receives no gradient.

The Bregman divergence and the self-guided side. Every configuration in $\mathcal{A}$ makes $\mathbb{E}[D_\tau]$ attain its global minimum of zero, collapse included. The divergence therefore identifies the positive-guided side and leaves the self-guided side to the potential-surface difference.

D.8 Alternative constructions

Three constraints were placed on the self-guided term in Appendix D.5, and Definition 3 is one way of meeting them. Others exist, and a few of them show which parts of the paper depend on the choice.

More than two gates. Take three levels of one gate and combine their retrievals with coefficients summing to zero. For a generic choice the result is non-zero on the diagonal, depends on the query alone, and is built from retrieval values, so C1 and C3 hold. C2 is a property of the dynamics and is checked through Definition 12, as for two levels. It combines three levels, so a single interval no longer indexes it.

Two gates from different families. C1 to C3 allow the two retrievals to come from different gates. A softmax gate for one view and a spectral gate for the other meet C1 and C3, provided the memory state admits both. C2 is again checked through Definition 12. What is lost is the order of the two levels, their common coordinate and the generator, which is why Section 3.2.3 chooses the traces of one memory state explicitly and separately.

Unequal weights on the two terms. They replace one gate, as shown in Appendix D.6.

Terms beyond differences of retrievals. C3 asks the self-guided term to be built from retrieval values, in any form. Any function of the values taken by one query under several gates qualifies.

The growth condition (34) is the same for all of them, because Propositions 23 and 25 hold for an arbitrary self-guided term. The choice in Section 3.2.3 gains two things. The positive-guided term, the existence of a scalar loss and the two constraints become separately checkable. And Section 3.3 gains a coordinate to evaluate along.

D.9 The flow on a fixed basis

Flow time is θ, because τ already indexes the gate, and a dot denotes the derivative in θ. Representations are treated as free points on the sphere and the flow is $\dot{z} = -\Pi_z g(z \mid z^+)$. Since $\Pi_z g \perp z$, the continuous flow preserves $\|z\| = 1$ exactly and needs no renormalization. A discrete update departs from the sphere by $O(\eta^2)$ in the step size.

All rates are taken in the eigenbasis of Σ at the current configuration, with the basis held fixed while the derivative is taken. The choice matters for some quantities, so it is stated once here.

For $\nu_j = e_j^\top \Sigma e_j$ the fixed basis gives $\dot\nu_j = e_j^\top \dot\Sigma e_j$ with no further argument. When ν_j is a simple eigenvalue this also equals the derivative of that eigenvalue, because the motion of the eigenvector contributes only at second order. When ν_j is degenerate the two readings differ. The derivatives of the eigenvalues are then the eigenvalues of $\dot\Sigma$ compressed to the eigenspace, and they agree with the fixed-basis rates only when the basis diagonalizes that block. For $\Sigma = \nu I_2$ and a $\dot\Sigma$ with off-diagonal entries 1, both fixed-basis rates vanish while the eigenvalues move at ± 1. We take the fixed-basis reading throughout. Quantities independent of the basis inside an eigenspace, its trace and the spread of the variances within it, have derivatives that use the fixed-basis rates only, and Proposition 28 is stated in them. The results below hold without simplicity.

For a quadratic that depends on the basis in another way, such as $\mathbb{E}[(e_j^\top \delta)^2]$, this coincidence fails. Differentiating along a moving basis adds $2\dot e_j^\top \mathbb{E}[\delta\delta^\top] e_j$. Fixing the basis removes that term for every quantity at once. That is why the convention is stated for the whole section rather than per proposition.

D.10 The directional self-retrieval score

Definition 10 (Directional score)**.** *Let $\{e_j\}$ be an orthonormal basis and put $\nu_j := e_j^\top \Sigma e_j = \mathbb{E}[(e_j^\top z)^2] > 0$. The directional score of the retrieval at level τ in direction e_j, which measures its level of self-retrieval along e_j, is*

$$\rho_j^{(\tau)} := \frac{\mathbb{E}\big[(e_j^\top R_\tau(z))\,(e_j^\top z)\big]}{\nu_j}. \tag{27}$$

We write $\Delta\rho_j := \rho_j^{(\tau+\Delta\tau)} - \rho_j^{(\tau)}$ for the gap between the two levels in direction e_j, and $\Delta\rho(z) := \rho_{\tau+\Delta\tau}(z) - \rho_\tau(z) = \langle z, \Delta R(z)\rangle$ for the same gap on one sample. The subscript marks only that distinction.

Proposition 21 (Decomposition of the score along an orthonormal basis)**.** *For any orthonormal basis $\{e_j\}$, $\sum_j \nu_j\, \rho_j^{(\tau)} = \mathbb{E}[\rho_\tau(z)]$, and consequently $\sum_j \nu_j\, \Delta\rho_j = \mathbb{E}[\Delta\rho(z)]$.*

Proof. $\sum_j \mathbb{E}[(e_j^\top R_\tau)(e_j^\top z)] = \mathbb{E}[\langle R_\tau(z), z\rangle]$. □

So the directional scores decompose the average score, weighted by the variances. Proposition 1 fixes the sign of that average, and Appendix D.6 identifies it as the radial half of the displacement, the half that Lemma 1 removes. What Proposition 23 asks about is the spread of the directional gaps around it.

D.11 The variance along each direction

Proposition 22 (Evolution of a second moment)**.** *Along $\dot z = -\Pi_z \phi(z)$, for any fixed unit vector r and any vector field ϕ,*

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\, r^\top \Sigma r = -2\,\mathbb{E}\big[(r^\top\phi)(r^\top z)\big] + 2\,\mathbb{E}\big[(r^\top z)^2 \langle z, \phi\rangle\big]. \tag{28}$$

Proof. $\frac{\mathrm{d}}{\mathrm{d}\theta}(r^\top z)^2 = 2(r^\top z)(r^\top \dot z)$. Substitute $\dot z = -\Pi_z\phi = -\phi + z\langle z,\phi\rangle$ and take expectations. □

Proposition 23 (Variance along a fixed basis)**.** *Let the configuration lie on $\mathcal{A}$, let $\nu_j > 0$, and let the basis be held fixed as above. Along the flow $\dot z = -\Pi_z \Delta R(z)$,*

$$\frac{\dot\nu_j}{\nu_j} \;=\; 2\big[\,\overline{\Delta\rho}_j - \Delta\rho_j\,\big], \qquad \overline{\Delta\rho}_j := \frac{\mathbb{E}\big[(e_j^\top z)^2\,\Delta\rho(z)\big]}{\nu_j}. \tag{29}$$

Proof. The basis is fixed, so e_j is a fixed vector and $\dot\nu_j = e_j^\top \dot\Sigma e_j$. Put $r = e_j$ and $\phi = \Delta R$ in Proposition 22. By Definition 10 the first term is $-2\nu_j\Delta\rho_j$ and the second is $2\nu_j\overline{\Delta\rho}_j$. □

Generality of the identity. It holds for any self-guided term, including terms beyond differences of two retrievals, so the growth condition (34) is independent of the construction of Section 3.2.3. It holds for operators with any eigenbasis, so it is available on both families. It also holds for a degenerate ν_j, because the proof uses a fixed e_j only, and any orthonormal basis of a degenerate eigenspace will do. Read as the motion of eigenvalues, it is exact for simple eigenvalues only. Corollary 1 uses a configuration with n equally occupied directions, which is degenerate, and states its stability through the basis-free quantities of Appendix D.9.

The positive-guided term can be kept. Equation (28) is linear in ϕ, so putting $\phi = \Delta R + (g - \Delta R)$ gives a version that keeps the positive-guided term. The extra bracket vanishes on $\mathcal{A}$ and Proposition 23 is that special case. Discarding the positive-guided term on the variance side is one treatment. Making it the only driver on the other side is a different one. Doing both is incompatible. Appendix D.13 avoids the issue by writing both sides into a single identity on a cross-view quantity.

D.12 The cross-view second moment

Definition 11 (Cross-view second moment and contraction rate)**.** *Write $\delta := z - z^+$, $\delta_j := e_j^\top \delta$, $x_j := e_j^\top z$ and $x_j^+ := e_j^\top z^+$. The cross-view second moment in direction e_j is $C_j := \mathbb{E}[x_j x_j^+]$. The contraction rate f_j is the exponential rate at which the positive-guided term acting alone shrinks $\mathbb{E}[\delta_j^2]$, that is*

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}[\delta_j^2]\Big|_{\textit{positive-guided term}} =: -2 f_j\, \mathbb{E}[\delta_j^2]. \tag{30}$$

The rate is defined by the coefficient it occupies in (30) rather than by a formula, because a formula moves with the symmetrization convention and with the tangential correction while the place in the equation stays fixed. Proposition 24(ii) gives the formula, and Appendix E.2 evaluates it on the spectral gate.

Alternatives to the cross-view second moment. Assume exchangeable views, or in fact only that the two views have the same second moment in each direction. Then

$$\mathbb{E}[\delta_j^2] = \mathbb{E}[x_j^2] + \mathbb{E}[(x_j^+)^2] - 2C_j = 2(\nu_j - C_j), \qquad \text{so} \qquad C_j = \nu_j - \tfrac{1}{2}\mathbb{E}[\delta_j^2]. \tag{31}$$

The shared part carries a factor one half, and two natural-looking alternatives fail because of it.

The ratio $\mathbb{E}[\delta_j^2]/\nu_j$ equals $2(1 - C_j/\nu_j)$ by (31). Cauchy–Schwarz gives $|C_j| \le \nu_j$, so the ratio ranges over $[0, 4]$. It lies in $[0, 1]$ exactly when $C_j \ge \nu_j/2$, that is when the cross-view correlation is at least one half. Early in training the correlation is lower, and a quantity read as a share while exceeding one corrupts whatever is built on it.

The difference $\nu_j - \mathbb{E}[\delta_j^2]$ equals $2C_j - \nu_j$, which is shared minus residual rather than shared. It can be negative while the shared part is positive, so its sign is uninformative about sharing.

Definition 11 avoids both. It needs no normalization, and its sign is informative: a negative value says the two views are anti-correlated along e_j, so variance placed there is unshared. On $\mathcal{A}$ it equals ν_j, and $\sum_j C_j = \mathbb{E}\langle z, z^+\rangle$ is the mean cosine between the two views.

D.13 The growth condition for the shared second moment

Both views act as queries, so the flow moves both views: $\dot z = -\Pi_z g(z \mid z^+)$ and $\dot{z^+} = -\Pi_{z^+} g(z^+ \mid z)$. The positive-guided term is antisymmetric, $\chi_\tau(z^+ \mid z) = -\chi_\tau(z \mid z^+)$, and we write χ_τ for $\chi_\tau(z \mid z^+)$.

Proposition 24 (Evolution of the cross-view moment and the view difference)**.** *Put $A_j^\times := \mathbb{E}[(e_j^\top \Delta R(z))\, x_j^+]$ and $\overline{A}_j^\times := \mathbb{E}[x_j x_j^+ \, \Delta\rho(z)]$.*

***(i)** Assume exchangeable views. Then*

$$\dot C_j = 2\big[\overline{A}_j^\times - A_j^\times\big] + \mathbb{E}\big[(e_j^\top \chi_\tau)\, \delta_j\big] + \mathbb{E}\big[x_j x_j^+ \, \langle \delta, \chi_\tau\rangle\big]. \tag{32}$$

On $\mathcal{A}$ we have $\delta = 0$, $C_j = \nu_j$, $A_j^\times = \nu_j \Delta\rho_j$ and $\overline{A}_j^\times = \nu_j \overline{\Delta\rho}_j$, the last two terms vanish, and (32) returns to Proposition 23.

(ii) *Assume the injection model. Then*

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}[\delta_j^2] = -2f_j\mathbb{E}[\delta_j^2] - 2\,\mathbb{E}\Big[\delta_j\, e_j^\top\big(\Pi_z\Delta R(z) - \Pi_{z^+}\Delta R(z^+)\big)\Big] + \sigma_j^2, \tag{33}$$

where $f_j\mathbb{E}[\delta_j^2] = \mathbb{E}[\delta_j\, e_j^\top(\Pi_z + \Pi_{z^+})\chi_\tau]$. *Dropping the middle term gives the steady state* $\mathbb{E}[\delta_j^2]^* = \sigma_j^2/(2f_j)$, *hence* $C_j = \nu_j - \sigma_j^2/(4f_j)$.

Proof. Expand the flow of z. With $\Pi_z\Delta R = \Delta R - z\langle z, \Delta R\rangle$ and $\Delta\rho(z) = \langle z, \Delta R(z)\rangle$,

$$\dot{x}_j = -\big(e_j^\top\Delta R(z)\big) + x_j\Delta\rho(z) - \big(e_j^\top\chi_\tau\big) + x_j\langle z, \chi_\tau\rangle,$$

and for z^+, using the antisymmetry of χ_τ,

$$\dot{x}_j^+ = -\big(e_j^\top\Delta R(z^+)\big) + x_j^+\Delta\rho(z^+) + \big(e_j^\top\chi_\tau\big) - x_j^+\langle z^+, \chi_\tau\rangle.$$

(i) Now $\dot{C}_j = \mathbb{E}[\dot{x}_j x_j^+] + \mathbb{E}[x_j\dot{x}_j^+]$. In the second expectation exchangeability lets the roles be exchanged, giving $\mathbb{E}[x_j\, e_j^\top\Delta R(z^+)] = A_j^\times$ and $\mathbb{E}[x_j x_j^+\Delta\rho(z^+)] = \overline{A}_j^\times$, so the self-guided part contributes $2[\overline{A}_j^\times - A_j^\times]$ in total. The positive-guided part contributes $\mathbb{E}[(e_j^\top\chi_\tau)(x_j - x_j^+)] + \mathbb{E}[x_j x_j^+\langle z - z^+, \chi_\tau\rangle]$, which is (32). (ii) Subtracting the two flows,

$$\dot{\delta} = -\big[\Pi_z\Delta R(z) - \Pi_{z^+}\Delta R(z^+)\big] - \big(\Pi_z + \Pi_{z^+}\big)\chi_\tau,$$

and pairing with δ_j, taking expectations and adding the injection gives (33). The steady state follows by setting the left side to zero. □

Proposition 25 (Growth condition for the shared second moment)**.** *Assume exchangeable views and keep the basis fixed. Write* $A_j^\times := \mathbb{E}[(e_j^\top\Delta R(z))\, x_j^+]$ *and* $\overline{A}_j^\times := \mathbb{E}[x_j x_j^+\,\Delta\rho(z)]$. *Then the cross-view second moment in direction* e_j *grows if and only if*

$$\underbrace{2\big[\overline{A}_j^\times - A_j^\times\big]}_{\textit{self-guided side}} + \underbrace{\mathbb{E}\big[(e_j^\top\chi_\tau)\,\delta_j\big] + \mathbb{E}\big[x_j x_j^+\,\langle\delta, \chi_\tau\rangle\big]}_{\textit{positive-guided side}} > 0. \tag{34}$$

Proof. Proposition 24(i). □

The two tangential projections. The positive-guided term delivers $-(\Pi_z + \Pi_{z^+})\chi_\tau$ to the two views rather than $-2\chi_\tau$. Treating it as the unprojected difference misplaces f_j by a factor of order one. Appendix E.2 computes that factor on the spectral gate, where it tends to one on weak directions. The growth condition (34) is evaluated on weak directions, so the conclusions there hold, and the definition stays the tangential one. This is the second appearance of the caution of Appendix D.3, now on the positive-guided side.

Unnormalized configuration terms. When $C_j \neq 0$ the two configuration terms are $C_j\Delta\rho_j^\times$ and $C_j\overline{\Delta\rho}_j^\times$ for the obvious cross-view analogues. The growth condition (34) has to pass through $C_j = 0$, the boundary between shared and unshared directions, and dividing by C_j puts a singularity exactly there, so the terms stay unnormalized.

The signs on the positive-guided side. Both terms on the positive-guided side can take either sign, each for its own reason. An auto-associative retrieval gives $\langle\delta, \chi_\tau\rangle \geq 0$, which is a statement about the sum over directions. The constraint of Appendix D.5 asks for the per-direction statement $\mathbb{E}[(e_j^\top\chi_\tau)\delta_j] \geq 0$, which the sum leaves open. The second term is $x_j x_j^+$ reweighted by the non-negative random weight $\langle\delta, \chi_\tau\rangle$, so $C_j > 0$ leaves its sign open. Negative values of $x_j x_j^+$ may sit on exactly the samples where the restoring force is strongest. The term is non-negative when $x_j x_j^+ \geq 0$ almost surely.

The middle term of (33)**.** The self-guided term takes different values at the two views, so it produces view difference while it acts. Part of what it sends into a direction arrives as view difference rather than as shared variance. That is the concern of retrieval agreement, and keeping the term puts it in an equation rather than in a remark. It stays unevaluated, and Appendix D.15 records what that costs.

D.14 GRADES OF RETRIEVAL SPECIFICITY

Definition 12 (Grades of retrieval specificity)**.** *For a given memory state and a given pair of levels:*

L0 $\Pi_z \Delta R$ *is non-zero somewhere on* $\mathcal{A}$*.*

L1 *On every degenerate aligned configuration, a new-direction perturbation grows: moving every sample by a vanishing amount into an empty direction yields* $\lim_{\nu \to 0^+} \dot{\nu}/\nu > 0$*.*

L2 *L1 holds, and the relative growth and the shared fraction stay bounded away from zero as a direction weakens. The self-guided side of* (34) *divided by* ν_j *has a positive lower bound on weak directions, and so does the shared fraction* C_j/ν_j*.*

L3 *The degenerate configurations attract a null set of initial conditions.*

Both clauses of L2 are ratios. On $\mathcal{A}$ the self-guided side of (34) equals $2\nu_j(\overline{\Delta\rho}_j - \Delta\rho_j) = \dot{\nu}_j$ by Proposition 23, so it vanishes with ν_j for a bounded retrieval. Under exchangeable views, Cauchy–Schwarz gives $|C_j| \le \sqrt{\mathbb{E}[x_j^2]\mathbb{E}[(x_j^+)^2]} = \nu_j$, so $C_j \to 0$ whenever $\nu_j \to 0$. Both quantities therefore vanish on empty directions, and the quantities that can carry a floor are the ratios $\dot{\nu}_j/\nu_j$ and C_j/ν_j. Appendices E.3 and E.4 work with these ratios.

Why the requirement is graded. Proposition 16 forbids the phrasing in which every degenerate point carries an outward velocity. The test therefore moves to a neighborhood, or to the limit of a vanishing direction, and the grades record how far a given configuration gets. L1, the weaker grade of Section 3.2.2, asks that a new direction grow. L2, the stronger grade, adds that the relative growth and the shared fraction stay bounded away from zero as the direction weakens, which is where the two constraints of Appendix D.5 become quantitative. L3 is a statement about basins and remains open here.

D.15 GAPS IN EVALUATING THE GROWTH CONDITION

The growth condition (34) and Definition 12 are defined on either family, and this appendix treats both alike. Evaluating the two quantities inside the growth condition is a different matter, since $A_j^\times$ and f_j both depend on the gate. That is Section 3.3.

Two gaps are worth naming here rather than later. The term $\overline{A}_j^\times$ contains a cross-view fourth moment, which this paper leaves unevaluated on both families. Appendix E.3 therefore works with the alignment-set form of the self-guided side rather than the exact one. And the injection model is an idealization, while the middle term of (33) is known to exist and is left uncomputed. Both make the steady-state value of C_j an estimate rather than a consequence. That is why Proposition 32 is presented as an approximate selection criterion rather than a theorem.

E EVALUATING THE GROWTH CONDITION ON THE TWO FAMILIES

E.1 THE THRUST SPECTRUM

Proposition 26 (Directional score on the spectral family)**.** *Under a current-batch memory state the retrieval at level* τ *is* $R_\tau(z) = F^\tau z$*, and its score along* e_j *is* λ_j^τ*. On a configuration whose views coincide, the variance of each direction evolves as*

$$\frac{\dot{\nu}_j}{\nu_j} = 2\Big[h(\lambda_j) - \sum\nolimits_k \pi_k^{(j)}\, h(\lambda_k)\Big], \qquad h(\lambda) := \lambda^\tau - \lambda^{\tau+\Delta\tau}, \tag{35}$$

where the occupancy weights $\pi_k^{(j)} := \mathbb{E}[(e_j^\top z)^2 (e_k^\top z)^2]/\nu_j$ *are non-negative and sum to one.*

Proof. $e_j^\top F^\tau z = \lambda_j^\tau (e_j^\top z)$, which gives $\rho_j^{(\tau)} = \lambda_j^\tau$ and hence $\Delta\rho_j = -h(\lambda_j)$. On one sample, $\Delta\rho(z) = z^\top (F^{\tau+\Delta\tau} - F^\tau) z = -\sum_k h(\lambda_k)(e_k^\top z)^2$. Substituting both into Proposition 23 gives the stated form, and $\sum_k \pi_k^{(j)} = \mathbb{E}[(e_j^\top z)^2 \|z\|^2]/\nu_j = 1$. □

The function h is the *thrust spectrum*, and $h(\lambda_j)$ is the score lost along e_j between the two levels.

Proposition 27 (Thrust as the integrated generator)**.** *By Definition 6,* $\Delta R(z) = -\int_{\tau}^{\tau+\Delta\tau} \Gamma_{\tau'}(z)z\,\mathrm{d}\tau'$*. On the spectral gate* $\Gamma_\tau = F^\tau \log F^{-1}$*, so*

$$h(\lambda) = \int_{\tau}^{\tau+\Delta\tau} \lambda^{\tau'} \log \tfrac{1}{\lambda}\,\mathrm{d}\tau' = \lambda^{\tau} - \lambda^{\tau+\Delta\tau}. \tag{36}$$

Proof. Direct integration. □

This is where the symmetry demanded of the generator in Definition 6 is spent. Reading how the pull is distributed across directions means reading the spectrum of Γ_τ, and (36) is that reading. The right side is an integral over a finite interval rather than a first-order expansion in $\Delta\tau$.

Corollary 5 (Weakest and strongest directions)**.** *Assume a current-batch memory state and let* h *be strictly decreasing on* $(0, 1]$*. Then the variance of the weakest direction is non-decreasing and that of the strongest direction is non-increasing. Both are strict whenever the occupancy weight of the direction reaches some* $\lambda_k \neq \lambda_j$*. If* h *is strictly increasing the two conclusions swap.*

Proof. By Proposition 8 the direction of least variance is the direction of least λ. Call it j. A decreasing h gives $h(\lambda_j) \geq h(\lambda_k)$ for every k, so $\sum_k \pi_k^{(j)} h(\lambda_k) \leq h(\lambda_j)$ and Proposition 26 gives $\dot\nu_j \geq 0$. Equality holds exactly when $\pi^{(j)}$ sits on directions with $\lambda_k = \lambda_j$. The strongest direction is symmetric. □

This is the accurate form of resistance to collapse on this family. It concerns the weakest surviving direction, which recovers, while Proposition 16 settles the degenerate point. The step from least variance to least spectrum is Proposition 8 and belongs to the memory update.

Proposition 28 (Stability of a uniform configuration)**.** *Assume a current-batch memory state. Let the configuration have equal variance on* n *directions and zero elsewhere, so that* $\lambda_{\mathrm{occ}} = (1/n + \epsilon)/(1 + d\epsilon)$ *and* $\lambda_\varnothing = \epsilon/(1 + d\epsilon)$*, and let* P_{occ} *and* $P_\varnothing$ *project onto the span of the* n *largest eigenvalues of* Σ *and onto the rest. The configuration is degenerate, so both statements use quantities independent of the basis inside an eigenspace.* ***(i)*** *Assume also the moment factorization, so that* $\sum_k \pi_k^{(j)} h(\lambda_k) = \bar h := \sum_k \nu_k h(\lambda_k)$ *and Proposition 26 takes the replicator form* $\dot\nu_j/\nu_j = 2[h(\lambda_j) - \bar h]$*. Under a variance-preserving perturbation inside the occupied span, the spread* $\sum_j \zeta_j^2$ *of the occupied variances about their mean, with* $\zeta_j := \nu_j - \frac{1}{n}\operatorname{tr}(P_{\mathrm{occ}}\Sigma)$*, evolves at first order as* $\exp\big(4h'(\lambda_{\mathrm{occ}})\theta/[n(1 + d\epsilon)]\big)$*, so the configuration is linearly stable against it exactly when* $h'(\lambda_{\mathrm{occ}}) < 0$*.* ***(ii)*** *Under a perturbation that moves every sample by a vanishing amount into the empty span, the total variance* $\operatorname{tr}(P_\varnothing\Sigma)$ *and the variance along each eigenvector in that span grow at rates tending to* $2[h(\lambda_\varnothing) - h(\lambda_{\mathrm{occ}})]$*, so an empty direction grows exactly when* $h(\lambda_\varnothing) > h(\lambda_{\mathrm{occ}})$*. This part holds without the moment factorization.*

Proof. Hold the current eigenbasis fixed. Both quantities involve only the fixed-basis rates $\dot\nu_j = e_j^\top \dot\Sigma e_j$, and both hold for any $\dot\Sigma$ on a degenerate block. The trace $\operatorname{tr}(P_\varnothing\Sigma)$ is a sum of eigenvalues separated from the others by a gap, so its derivative is $\sum_{j\in\varnothing} \dot\nu_j$. The spread is $\sum_j \zeta_j^2$ in this basis with $\sum_j \zeta_j = 0$, so its derivative is $2\sum_j \zeta_j\dot\nu_j$, and ζ_j is constant on each eigenspace, so every basis inside one gives the same value. (i) Put $\nu_j = 1/n + \zeta_j$, so $\lambda_j = \lambda_{\mathrm{occ}} + \zeta_j/(1 + d\epsilon)$. The first-order terms of $\bar h$ cancel, leaving $\bar h = h(\lambda_{\mathrm{occ}}) + O(\zeta^2)$, hence $\dot\nu_j = 2h'(\lambda_{\mathrm{occ}})\zeta_j/[n(1 + d\epsilon)] + O(\zeta^2)$, and the spread grows at the rate $4h'(\lambda_{\mathrm{occ}})/[n(1 + d\epsilon)]$. The factor $1/(1 + d\epsilon)$ is positive and keeps the sign. (ii) By Proposition 26, $\dot\nu_j/\nu_j = 2[h(\lambda_j) - \sum_k \pi_k^{(j)} h(\lambda_k)]$ with $\sum_k \pi_k^{(j)} = 1$. As the displacement vanishes, $\lambda_j \to \lambda_\varnothing$ on the empty span, every occupied eigenvalue tends to the same λ_{occ}, and $\sum_{k\in\varnothing} \pi_k^{(j)} = \mathbb{E}[(e_j^\top z)^2\|P_\varnothing z\|^2]/\nu_j \leq \max\|P_\varnothing z\|^2 \to 0$. Hence $\sum_k \pi_k^{(j)} h(\lambda_k) \to h(\lambda_{\mathrm{occ}})$, and summing over the empty span gives the rate of the trace. □

The eigenvalues are kept in the form $(\nu + \epsilon)/(1 + d\epsilon)$ rather than written as ν. The empty direction carries $\lambda_\varnothing > 0$ rather than zero, and the distinction matters in Corollary 1(iii) below. The familiar limits $\lambda_\varnothing \to 0$ and $\lambda_{\mathrm{occ}} \to 1/n$ are the $\epsilon \to 0$ case of what is written here.

Corollary 1. *Treat the representations as free points on the sphere that move along $-\Pi_z g$, with the memory state recomputed as the second moment of the current batch. Let the two views coincide and occupy n directions equally.* ***(i)*** *For $\tau = 0$ and $\Delta\tau > 0$, every empty direction is unstable for every n and every ϵ.* ***(ii)*** *For $\tau > 0$, $\Delta\tau > 0$ and $n \geq 2$, empty directions are stable once ϵ is small.* ***(iii)*** *For $\tau > 0$, $\Delta\tau < 0$, $\tau + \Delta\tau > 0$ and $n \geq 2$, empty directions are unstable once ϵ is small.*

Proof. The three statements substitute into Proposition 28(ii), which holds without the moment factorization. The derivative is $h'(\lambda) = \lambda^{\tau-1}\big(\tau - (\tau + \Delta\tau)\lambda^{\Delta\tau}\big)$, which changes sign at $\lambda^\star := \big(\tau/(\tau+\Delta\tau)\big)^{1/\Delta\tau}$. (i) At $\tau = 0$ we have $h = 1-\lambda^{\Delta\tau}$, strictly decreasing for $\Delta\tau > 0$, so $\lambda_\varnothing < \lambda_{\mathrm{occ}}$ gives $h(\lambda_\varnothing) > h(\lambda_{\mathrm{occ}})$ for every n and every ϵ. Under the moment factorization every occupied direction carries a positive occupancy weight, so by the equality case of Corollary 5 a stationary configuration has equal occupied eigenvalues. Among those, only the isotropic one has no unstable empty direction. (ii) Let $\tau > 0$, $\Delta\tau > 0$ and $n \geq 2$. As $\epsilon \to 0$, $h(\lambda_\varnothing) = \lambda_\varnothing^\tau(1 - \lambda_\varnothing^{\Delta\tau})$ tends to zero, while $h(\lambda_{\mathrm{occ}})$ tends to $h(1/n) = n^{-\tau}(1 - n^{-\Delta\tau}) > 0$. So $h(\lambda_\varnothing)$ falls below $h(\lambda_{\mathrm{occ}})$ once ϵ is small. (iii) For $\tau > 0$ and $\Delta\tau < 0$ with $\tau + \Delta\tau > 0$, the thrust is negative on $(0, 1)$ and tends to zero at both ends, so the comparison depends on where $\lambda_\varnothing$ sits relative to $\lambda^\star$. For $n \geq 2$ the limit $h(1/n)$ is negative while $h(\lambda_\varnothing)$ tends to zero, so the empty direction is unstable once ϵ is small. When $\tau + \Delta\tau = 0$ the thrust is $\lambda^\tau - 1$, strictly increasing, and the empty direction is stable. □

The rank bound. For $\tau > 0$ and $\Delta\tau > 0$, Proposition 28(i) adds a condition inside the occupied span. Under the moment factorization the uniform configuration is stable against variance-preserving perturbations exactly when $\lambda_{\mathrm{occ}} > \lambda^\star$, that is $n < [\lambda^\star(1 + d\epsilon) - \epsilon]^{-1}$, which becomes $n < 1/\lambda^\star$ as $\epsilon \to 0$. For configurations of bounded kurtosis the factorization is an $O(1/n)$ approximation, so the bound is least accurate at small n.

The rank-one configuration. At $n = 1$ both eigenvalues approach zeros of the thrust, since $\lambda_\varnothing \to 0$ and $\lambda_{\mathrm{occ}} = (1 + \epsilon)/(1 + d\epsilon) \to 1$. Proposition 28(ii) still decides the outcome, and the order in which the two thrusts vanish settles it as $\epsilon \to 0$. For $\tau > 0$ and $\tau + \Delta\tau > 0$,

$$h(\lambda_\varnothing) = \mathrm{sgn}(\Delta\tau)\,\epsilon^{\min(\tau,\tau+\Delta\tau)}\big(1 + o(1)\big), \qquad h(\lambda_{\mathrm{occ}}) = \Delta\tau\,(d-1)\,\epsilon\,\big(1 + o(1)\big).$$

For $\Delta\tau > 0$ an empty direction is therefore unstable once ϵ is small when $\tau < 1$, and stable when $\tau > 1$. At $\tau = 1$ it is unstable when $\Delta\tau(d-1) < 1$ and stable when $\Delta\tau(d-1) > 1$. For $\Delta\tau < 0$ the two outcomes swap, with $\tau + \Delta\tau$ in place of τ. The empty direction is then stable when $\tau + \Delta\tau < 1$ and unstable when $\tau + \Delta\tau > 1$. At $\tau + \Delta\tau = 1$ it is unstable when $|\Delta\tau|(d-1) > 1$ and stable when $|\Delta\tau|(d-1) < 1$. For $\Delta\tau > 0$ and $\tau < 1$ the sign changes near $\epsilon = (\Delta\tau(d-1))^{-1/(1-\tau)}$. At $(1/2, 1)$ with $d = 2048$ this is about 10^{-6}, so at $\epsilon = 10^{-4}$ the rank-one configuration is stable at $(1/2, 1)$ and $(1, 2)$.

The condition in outcome (iii). Two ends tending to zero leave an endpoint comparison undecided. The thrust falls and then rises, and an empty eigenvalue below $\lambda^\star$ but not close enough to zero reverses the conclusion. Take $\tau = 1$, $\Delta\tau = -0.8$, $d = 3$, $\epsilon = 0.2$ and $n = 2$. Then $\lambda_\varnothing = 0.125$ and $\lambda_{\mathrm{occ}} \approx 0.44$, while $\lambda^\star \approx 0.13$, and $h(0.125) \approx -0.53$ falls below $h(0.44) \approx -0.41$. The empty direction is stable, which is the opposite of the generic reading. Outcome (i) by contrast holds for every $\epsilon \geq 0$ and needs no limit.

Proposition 29 (A weight on the potential term)**.** *Assume a current-batch memory state and a configuration whose views coincide. A weight $\beta > 0$ on the potential term replaces the thrust of Proposition 26 by $h_\beta(\lambda) := \lambda^\tau - \beta\lambda^{\tau+\Delta\tau}$. For the uniform configuration of Proposition 28 let $\tau > 0$ and $\beta > 1$. Once ϵ is small, an empty direction is unstable when $n < \beta^{1/\Delta\tau}$ and stable when $n > \beta^{1/\Delta\tau}$ if $\Delta\tau > 0$, and unstable for every n if $\Delta\tau = 0$. At $\tau = 0$ the comparison of Proposition 28(ii) is multiplied by β and keeps its sign.*

Proof. On a configuration whose views coincide the weighted loss has the gradient $\Delta R_\beta(z) := \beta F^{\tau+\Delta\tau} z - F^\tau z$. Proposition 23 holds for any self-guided term. With ΔR_β in place of ΔR, the proof of Proposition 26 gives $\Delta\rho_j = -h_\beta(\lambda_j)$ and $\Delta\rho(z) = -\sum_k h_\beta(\lambda_k)(e_k^\top z)^2$, so (35) holds with h_β. Proposition 28(ii) uses the thrust only through $h(\lambda_\varnothing) - h(\lambda_{\mathrm{occ}})$, so an empty direction grows exactly when $h_\beta(\lambda_\varnothing) > h_\beta(\lambda_{\mathrm{occ}})$. For $\tau > 0$ the first term $\lambda_\varnothing^\tau(1 - \beta\lambda_\varnothing^{\Delta\tau})$ tends to zero as

$\epsilon \to 0$, and the second tends to $n^{-\tau}(1-\beta n^{-\Delta\tau})$. For $\Delta\tau > 0$ the limit of the difference is positive when $\beta n^{-\Delta\tau} > 1$, that is $n < \beta^{1/\Delta\tau}$, and negative when $n > \beta^{1/\Delta\tau}$. At $n = \beta^{1/\Delta\tau}$ the limit is zero, and to leading order the difference is $\epsilon^\tau - n^{-\tau}\Delta\tau\,(d-n)\,\epsilon$, which decides the sign as on the rank-one configuration. For $\Delta\tau = 0$ the thrust $(1-\beta)\lambda^\tau$ is negative and strictly decreasing, so the difference is positive for every n. At $\tau = 0$ the difference is $\beta(\lambda_{\mathrm{occ}}^{\Delta\tau} - \lambda_{\varnothing}^{\Delta\tau})$, which is β times its value at $\beta = 1$. □

The weighted loss is the loss $\mathcal{L}_\beta$ of Section 3.2.3, the case of unequal weights in Appendix D.6. It falls outside Definition 3 and still meets the three constraints. At $\beta = 1$ the boundary $\beta^{1/\Delta\tau}$ equals one, so $n \geq 2$ gives outcome (ii) of Corollary 1, and $n = 1$ is the rank-one configuration above.

Order of levels and the outcome. Outcomes (i) and (iii) both destabilize empty directions, and their intervals have opposite signs. The order of the two levels therefore leaves the conclusion open. What fixes it is the monotonicity of h together with where $h(\lambda_\varnothing)$ sits relative to $h(\lambda_{\mathrm{occ}})$. On configurations with $n \geq 2$, outcome (iii) passes the self-guided side once ϵ is small. The rank-one configuration of a reverse order with $\tau + \Delta\tau < 1$ keeps its empty directions stable once ϵ is small, so such a reverse order fails L1. Appendix E.3 excludes reverse orders with $\tau \geq \tau_{\max}$, which covers every fixed reverse order as $\epsilon \to 0$.

Outcome (ii) is excluded by the self-guided side, and the positive-guided side judges the remaining pairs. At the maximal self-retrieval and for small view differences, the positive-guided term shrinks the view difference at the same rate in every direction, which is what retrieval agreement asks for, and away from it that rate has no closed form per direction (Appendix E.2). The preconditioning identity $g_{(\tau,\Delta\tau)} = F^\tau(F^{\Delta\tau}z - z^+)$ explains why both sides degrade together: every pair of levels is the pair with the positive at the maximal self-retrieval, reweighted by F^τ, which lowers weak directions by λ^τ on both sides at once.

E.2 The contraction rate

Proposition 30 (Contraction rate on the spectral family)**.** *Write $\Pi_\pm := \Pi_z + \Pi_{z^+}$. The positive-guided term is $F^\tau\delta$. **(i)** At $\tau = 0$ the rate is exact. Since $z^\top\delta = \frac{1}{2}\|\delta\|^2$ and $(z^+)^\top\delta = -\frac{1}{2}\|\delta\|^2$, we have $\Pi_\pm\delta = (2 - \frac{1}{2}\|\delta\|^2)\delta$, so δ is an eigenvector of $\Pi_\pm$ and*

$$f_j = \frac{\mathbb{E}\big[\delta_j^2\,(2 - \frac{1}{2}\|\delta\|^2)\big]}{\mathbb{E}[\delta_j^2]} \in [0, 2], \tag{37}$$

*which tends to 2 uniformly in the direction as $\|\delta\| \to 0$. **(ii)** For general τ only an aggregate is available,*

$$\sum_j f_j \mathbb{E}[\delta_j^2] = \mathbb{E}\big[(2 - \tfrac{1}{2}\|\delta\|^2)\,\delta^\top F^\tau\delta\big] \in \big[\lambda_{\min}^\tau,\ \lambda_{\max}^\tau\big] \cdot \mathbb{E}\big[(2 - \tfrac{1}{2}\|\delta\|^2)\|\delta\|^2\big]. \tag{38}$$

***(iii)** There is no closed form per direction.*

Proof. (i) Expand $\Pi_\pm = 2I - zz^\top - z^+(z^+)^\top$ and use the two inner products. (ii) $\Pi_\pm$ is symmetric with δ an eigenvector, so $\Pi_\pm$ moves to the left factor, and $\lambda_{\min}^\tau\|\delta\|^2 \leq \delta^\top F^\tau\delta \leq \lambda_{\max}^\tau\|\delta\|^2$. (iii) Expanding $F^\tau\delta = \sum_k \lambda_k^\tau\delta_k e_k$ and $e_j^\top(\Pi_z + \Pi_{z^+})e_k = 2\cdot\mathbf{1}[j=k] - x_jx_k - x_j^+x_k^+$ gives $f_j\mathbb{E}[\delta_j^2] = 2\lambda_j^\tau\mathbb{E}[\delta_j^2] - \sum_k \lambda_k^\tau\mathbb{E}[\delta_j\delta_k(x_jx_k + x_j^+x_k^+)]$, whose $k \neq j$ terms bring strong directions in through the tangential projection. □

Closing (iii) needs the view difference to be treated as dependent on the query. Under exchangeable views the identity $\mathbb{E}[\delta_j x_j] = \frac{1}{2}\mathbb{E}[\delta_j^2]$ holds, and it is non-zero whenever the views differ. An independence assumption made to split the fourth moment would contradict the definition rather than approximate it. The mixing term in (iii) and the recovery domain of Appendix D.3 are one phenomenon. On weak directions their ratio is about $(\lambda_{\max}/\lambda_{\min})^\tau\|\delta\|$, which the threshold $2(\lambda_{\min}/\lambda_{\max})^\tau$ controls. Inside the domain the two terms are comparable and outside it the strong directions dominate. At $\tau = 0$ both degenerate, the mixing term closes into the $\|\delta\|^2$ correction of (i), and (i) is exact.

Proposition 31 (Preconditioning identity). $g_{(\tau,\Delta\tau)} = F^{\tau+\Delta\tau}z - F^{\tau}z^{+} = F^{\tau}\big(F^{\Delta\tau}z - z^{+}\big)$, *so* $h_{(\tau,\Delta\tau)}(\lambda) = \lambda^{\tau}\, h_{(0,\Delta\tau)}(\lambda)$.

Proof. Functional calculus. □

The whole two-parameter family is therefore the family with the positive at the maximal self-retrieval, preconditioned by F^{τ}. Moving the positive away from the maximal self-retrieval is a positive-definite reweighting that lowers weak directions by λ^{τ}. Thrust and aggregate contraction rate fall together there, and weak directions are what both requirements need to keep.

The reweighting preserves the zeros of the unprojected gradient only, since F^{τ} is invertible. On the sphere the fixed-point condition is $F^{\tau}(F^{\Delta\tau}z - z^{+})$ parallel to z, which differs from $F^{\Delta\tau}z - z^{+}$ parallel to z, so on the tangential flow a preconditioner changes fixed points as well as rates.

E.3 A WINDOW OF LEVELS IN AN APPROXIMATE MODEL

The contraction estimate. On weak directions we take $f_j \asymp 2\lambda_j^{\tau}$. At $\tau = 0$ it holds for small view differences by (37), and it is consistent with (38) inside the recovery domain. Outside that domain it is unsupported. The next proposition depends on it, and its first part holds without it.

Proposition 32 (A window for τ in an approximate model)**.** *Assume a current-batch memory state, exchangeable views, view differences injected in proportion to the variance with the same ratio* σ^2/ν *in every direction, a steady state with one term dropped, and the estimate of the contraction rate on weak directions stated in Appendix E.3. Let the ratio be small enough for* $\tau = 0$ *to pass the steady-state test. Every usable pair of levels then has* τ *in an interval* $[0, \tau_{\max})$*, whose upper end Appendix E.3 gives in closed form. The pairs with* $\tau = 0$ *and* $\Delta\tau > 0$ *are usable. Each* $\tau \geq \tau_{\max}$ *sets a retention floor: the two views share the variance that spreads into a direction only above it. As* $\epsilon \to 0$ *the interval shrinks to a point, and every* $\tau > 0$ *then sets a retention floor.*

Proof. By the proof of Corollary 1, the pairs with $\tau > 0$ and $\tau + \Delta\tau = 0$ fail L1, since their uniform configurations with $n \geq 2$ have stable empty directions. Once ϵ is small the pairs with $\tau > 0$ and $\Delta\tau > 0$ fail it for the same reason. That leaves $\tau = 0$ and reverse order. In the steady state of Proposition 24(ii) with the middle term dropped, $C_j > 0$ holds exactly when $f_j > \sigma_j^2/(4\nu_j)$, and the contraction estimate turns this into $\lambda_j^{\tau} > \sigma_j^2/(c\,\nu_j)$ with a constant c fixed by the symmetrization convention and the tangential correction. The convention used here gives $c = 8$. With the same ratio σ^2/ν in every direction, the test at $\tau = 0$ reads $c\nu > \sigma^2$, which is the assumption on the ratio. Solving $\lambda_{\varnothing}^{\tau} > \sigma^2/(c\nu)$ and taking logarithms of two negative quantities gives $\tau_{\max} = \log(c\,\nu/\sigma^2)/\log(1/\lambda_{\varnothing})$. A pair with $\tau \geq \tau_{\max}$ fails the test on its weakest directions. The pairs with $\tau = 0$ and $\Delta\tau > 0$ pass it, and they pass L1 by Corollaries 1(i) and 5. As $\epsilon \to 0$ we have $\lambda_{\varnothing} \to 0$ and $\tau_{\max} \to 0$. □

Injection at a fixed rate. The injection model of Appendix A.3 fixes σ_j^2 while ν_j varies. The same test then reads $\nu_j > \sigma_j^2/(c\,\lambda_j^{\tau})$. It sets the floor σ_j^2/c already at $\tau = 0$, and every $\tau > 0$ raises the floor by the factor $\lambda_j^{-\tau}$. Proportional injection is the case in which the floor at $\tau = 0$ vanishes.

Arguments behind the window. The window is an approximate selection criterion. It rests on the injection model, on dropping the middle term of (33), and on the contraction estimate. Two other arguments favor the same end and are independent of all three. The thrust on weak directions is multiplied by λ^{τ} by Proposition 31, and weak directions can least afford it. The tangential recovery threshold of Appendix D.3 is $2(\lambda_{\min}/\lambda_{\max})^{\tau}$ on this family, which equals 2 at $\tau = 0$ and contracts exponentially afterwards. Appendix A.5 sets out how each result relies on the statistical assumptions. At $\tau = 0$ every pair except an exactly antipodal one lies inside the domain. The conclusion rests mainly on these two, and both admit $\tau > 0$ at finite ϵ.

The window at finite ϵ. With $\epsilon > 0$ we have $\lambda_{\varnothing} > 0$, so the contraction rate has a positive lower bound at every τ, and every level remains admissible. The sharper phrasing holds only at $\epsilon = 0$, and Appendix C.4 argues that ϵ must be positive. What survives, under proportional injection, is a threshold $(\sigma^2/(c\,\nu))^{1/\tau}$ on the eigenvalue of a direction, which vanishes at $\tau = 0$ and rises with τ.

At finite ϵ it lies below every eigenvalue for $0 < \tau < \tau_{\max}$. It becomes a strictly positive retention floor once $\tau \geq \tau_{\max}$, where it rises above $\lambda_\varnothing$, and every $\tau > 0$ sets one as $\epsilon \to 0$. The constant c depends on convention, while the start of the window at zero and its shrinking to a point as $\epsilon \to 0$ hold for every convention.

Reverse order at finite ϵ. The window constrains τ rather than the sign of $\Delta\tau$. On a configuration with $n \geq 2$, a reverse order with $0 < \tau < \tau_{\max}$ passes the steady-state test. Once ϵ is small it passes the self-guided side as well, by Corollary 1(iii). With $d = 2048$, $n = 100$, $\epsilon = 10^{-4}$, $\sigma^2/\nu = 1$ and $c = 8$ we get $\tau_{\max} \approx 0.22$, and $(\tau, \Delta\tau) = (0.196, -0.086)$ passes both sides. The rank-one configuration at the same pair, d and ϵ keeps its empty directions stable (Appendix E.1), so the pair fails L1. At these values every reverse order on a grid of 361 across the window fails there as well. The pairs with $\tau = 0$ and $\Delta\tau > 0$ remain. As $\epsilon \to 0$, where $\tau_{\max} \to 0$, every fixed reverse order also falls outside the window, and there it spreads variance into empty directions that the two views do not share.

The sign and the size of the interval. In the limit $\epsilon \to 0$, $\tau = 0$ and $\tau + \Delta\tau \geq 0$ make the interval positive, so the retrieval of z sits nearer the query-independent retrieval. At finite ϵ this argument leaves the sign open. The size controls only the spread, since $h = 1 - \lambda^{\Delta\tau}$ tends to a constant at both extremes. Against n equally occupied directions the escape rate of an empty direction is $2[\lambda_{\mathrm{occ}}^{\Delta\tau} - \lambda_\varnothing^{\Delta\tau}]$, which is largest at $\Delta\tau = \log(\log\lambda_\varnothing/\log\lambda_{\mathrm{occ}})/\log(\lambda_{\mathrm{occ}}/\lambda_\varnothing)$. That value describes this one escape rate rather than the spread of the whole spectrum. At $\tau = 0$ loss (7) is $\frac{1}{2}z^\top F^{\Delta\tau} z - \langle z, z^+\rangle$.

Merged retrievals.

Definition 13 (Merged retrievals and merge classes)**.** *Two positives are said to have* merged retrievals *when $R_\tau(z^+) = R_\tau(z^{+\prime})$. The equivalence classes of that relation are the* merge classes*. Two queries whose positives share a class receive the same signal from the retrieval at τ in that step.*

Proposition 33 (Origin of the merge classes)**.** *The merge classes are produced by the choice of the retrieval at τ rather than by the data.*

Proof. A paired observation is a relation on the observation space, and its type involves the observation space alone, so it leaves open which representations ought to share a retrieval. By Definition 13 the classes are determined by R_τ alone. □

Whether the assertion is correct depends on whether the queries in one class do need to predict similar things. That is a question about the downstream task and remains open here. One consequence is immediate. Every class is a singleton wherever the retrieval at τ is injective, and then the retrieval asserts no merge. Injectivity holds at every finite τ on the spectral family with $F \succ 0$, and at $\tau = 0$ with an identity retrieval. Classes with more than one member come from a many-to-one retrieval, such as the nearest-neighbor quantizer of the softmax family at $\tau = 0$ with trained prototypes. The merge classes therefore separate quantizing retrievals from injective ones. Under a current-batch memory state, the choice among levels rests on the two arguments above.

E.4 The softmax family

Write $x_j := e_j^\top z$ and $x_j^{(l)} := e_j^\top z_l$ for the stored items. Let S be the set of stored items that are copies of the pair and $p_S := \sum_{l \in S} p_l$ the addressing mass they receive. Let $\tilde{p}$ be the addressing weights renormalized on the remaining items. Let $\hat{p}$ be the addressing weights computed with the contribution of direction j removed from the affinities, and $\hat{p}_S$ its mass on S. Throughout this subsection z^+ is retrieved at the maximal self-retrieval, so the retrieval of z uses the gate $\hat{y}_{\Delta\tau}$.

The contraction rate with and without copies. The maximal self-retrieval of this family is a nearest-neighbor quantizer. Suppose the memory state holds copies of the pair. Each query then retrieves its own copy, so the positive-guided term is the view difference itself. The rate f_j is uniform across directions up to the tangential correction of order ν_j. That is the same situation as $\tau = 0$ on the spectral gate. Without copies the retrieval is piecewise constant. The positive-guided term vanishes inside a cell, and the rate is carried entirely by events crossing a cell boundary. What controls it is

the probability that the two views fall in different cells, multiplied by the distance between the two retrieved items. Both are governed by how well the memory state covers the neighborhood of the query. What separates the two memory states is whether the maximal self-retrieval is the identity, and the rate differs between them in presence rather than in degree. A lower bound in the second case remains open.

Proposition 34 (Thrust on weak directions)**.** *Assume a current-batch memory state, a configuration whose views coincide and independent coordinates, let z^+ be retrieved at the maximal self-retrieval, and let the memory state hold copies of the pair. Then*

$$\rho_j^{(\Delta\tau)} = \mathbb{E}[p_S] + \frac{\nu_j}{\Delta\tau}\Big[(\kappa_j - 1)\mathbb{E}\big[\hat{p}_S(1-\hat{p}_S)\big] + \mathbb{E}[1-\hat{p}_S] - \mathbb{E}\|\hat{p}\|^2_{\notin S}\Big] + O\big((\nu_j/\Delta\tau)^2\big), \quad (39)$$

where $\kappa_j := \mathbb{E}[x_j^4]/\nu_j^2$. The zero-order term is the same for every direction, and the bracket is non-negative.

Proof. Split the affinity as $\langle z, z_l\rangle/\Delta\tau = \hat{a}_l + \Delta a_l$ with $\hat{a}_l := \Delta\tau^{-1}\sum_{k\neq j} x_k x_k^{(l)}$ and $\Delta a_l := x_j x_j^{(l)}/\Delta\tau$, and put $\hat{p} := \mathrm{softmax}(\hat{a})$. Expanding the softmax to first order gives $p_l = \hat{p}_l(1 + \Delta a_l - \langle \Delta a\rangle_{\hat{p}}) + O(\Delta a^2)$, hence

$$\sum_l p_l x_j^{(l)} = \sum_l \hat{p}_l x_j^{(l)} + \frac{x_j}{\Delta\tau}\,\mathrm{Var}_{\hat{p}}\left(x_j^{(\cdot)}\right) + O(\Delta a^2).$$

By independent coordinates and the drawing clause of the current-batch memory state, the addressing weights $\hat{p}$ depend only on the coordinates $k \neq j$, so they are independent of x_j and of $\{x_j^{(l)}\}_{l\notin S}$. On copies $x_j^{(l)} = x_j$. Pairing the first term with x_j gives $\nu_j\mathbb{E}[\hat{p}_S]$, the non-copy part vanishing because $\mathbb{E}[x_j^{(l)}] = 0$. Pairing the second gives $\Delta\tau^{-1}\big[\mathbb{E}[x_j^4]\mathbb{E}[\hat{p}_S(1-\hat{p}_S)] + \nu_j^2(\mathbb{E}[1-\hat{p}_S] - \mathbb{E}\|\hat{p}\|^2_{\notin S})\big]$. The anchor is then moved. The same expansion gives $\mathbb{E}[p_S - \hat{p}_S] = \mathbb{E}[\hat{p}_S(1-\hat{p}_S)]\nu_j/\Delta\tau$, and substituting it produces (39) with zero-order term $\mathbb{E}[p_S]$. Non-negativity of the bracket follows from $\kappa_j \geq 1$ by Jensen and from $\mathbb{E}\|\hat{p}\|^2_{\notin S} \leq \mathbb{E}[1-\hat{p}_S]$. □

The anchor has to be direction-free. The natural anchor is $\mathbb{E}[\hat{p}_S]$, and it is the wrong one. The weights $\hat{p}$ are computed with direction j removed. So $\mathbb{E}[\hat{p}_S]$ carries its own $O(\nu_j/\Delta\tau)$ dependence on the direction, of the same order as the first-order term. Comparing directions against it is then meaningless. Moving the anchor to the full addressing mass $\mathbb{E}[p_S]$ costs one substitution and changes κ_j to $\kappa_j - 1$ in the bracket. The two changes come as a pair.

The role of the zero mean. The assumption of independent coordinates excludes a direction on which every point takes the same non-zero value. There $\mathbb{E}[x_j]^2 = \nu_j$, the zero-order term survives, and the thrust is exactly zero. A vanishing variance forces $\mathbb{E}[x_j] \to 0$ while $\mathbb{E}[x_j]^2/\nu_j$ can stay at one, as $x_j \equiv \sqrt{\nu_j}$ shows. A class of weak directions is therefore excluded, which is the price of the uncentered second moment recorded in Appendix D.2. Corollary 6 uses the zero mean a second time, to make the conditional mean of the non-copy affinities vanish.

The replicator form without the moment factorization. On the spectral gate, replacing the occupancy weights by ν_k requires the moment factorization. Here the same simplification follows from independent coordinates directly, because a weak direction is nearly independent of the addressing weights. The price is that the resulting form is asymptotic in $\nu_j/\Delta\tau$ rather than exact.

Corollary 6 (L1 and the first clause of L2 on the softmax family)**.** *Under the setting of Proposition 34, assume further a finite $\Delta\tau$, at least two stored items that are not copies, and non-copy affinities that are not almost surely all equal.* ***(i)*** *On directions with $\nu_j \ll \Delta\tau$ the thrust is $h_j = (1 - \mathbb{E}[p_S]) - \gamma\,\nu_j/\Delta\tau + O((\nu_j/\Delta\tau)^2)$, where γ is the bracket of* (39) *evaluated with the full addressing weights p. Then $\gamma > 0$ is the same for every direction, and to leading order the thrust decreases strictly in ν_j.* ***(ii)*** *The escape rate of an empty direction is*

$$\lim_{\nu_j\to 0}\frac{\dot{\nu}_j}{\nu_j} = 2\,\mathbb{E}\big[(1-p_S)\,\bar{a}(z)\big] > 0, \qquad \bar{a}(z) := \sum\nolimits_{l\notin S}\tilde{p}_l\,\langle z, z_l\rangle, \quad (40)$$

so an empty direction is unstable, and the rate stays bounded away from zero as the direction weakens.

Proof. (i) Combine Proposition 34 with $h_j = 1 - \rho_j^{(\Delta\tau)}$. The zero-order term is direction-free, so directions may be compared. Replacing $\hat{p}$ by p in the bracket changes it by $O(\nu_j/\Delta\tau)$, which the remainder absorbs. The full addressing weights and, under independent coordinates, the kurtosis κ_j are the same for every j, so γ is the same for every direction. Moreover $\kappa_j \geq 1$ and $\|p\|^2_{\notin S} \leq (1-p_S)^2$ give $\gamma \geq \mathbb{E}[p_S(1-p_S)] > 0$. (ii) As $\nu_j \to 0$ the thrust tends to $\mathbb{E}[1-p_S]$, while $\rho_\Delta\tau(z) = p_S + (1-p_S)\bar{a}(z)$ gives $\bar{h} = \mathbb{E}[(1-p_S)(1-\bar{a})]$. Subtracting gives (40). For the sign, write $a_l := \langle z, z_l\rangle$ for the non-copy affinities and $\bar{a}_0(z)$ for their uniform average. The quantity $\bar{a}$ averages the same affinities under the distribution tilted by $e^{a_l/\Delta\tau}$, so $\bar{a} - \bar{a}_0 \geq 0$, strictly when the affinities are not all equal, and $1 - p_S > 0$. Hence $\mathbb{E}[(1-p_S)(\bar{a}-\bar{a}_0)] > 0$. For the remaining term $\mathbb{E}[(1-p_S)\bar{a}_0]$, condition on z. On the diagonal the copies have affinity 1, so $1-p_S$ and $\bar{a}_0$ are both non-decreasing in each a_l. By the drawing clause of the current-batch memory state the a_l are independent given z, and the Harris inequality gives $\mathbb{E}[(1-p_S)\bar{a}_0 \mid z] \geq \mathbb{E}[1-p_S \mid z]\,\mathbb{E}[\bar{a}_0 \mid z]$, where $\mathbb{E}[\bar{a}_0 \mid z] = \langle z, \mathbb{E}z_l\rangle = 0$ by independent coordinates. Adding the two terms gives strict positivity. □

The two limits of the interval. As $\Delta\tau \to 0$ the addressing weights concentrate on the copies, $p_S \to 1$, and the escape rate goes to zero. As $\Delta\tau \to \infty$ the addressing weights become uniform, the retrieval tends to $\bar{v}$, and the rate goes to zero on an isotropic configuration. Spread is available only at a finite interval, which matches the behaviour of $h = 1 - \lambda^{\Delta\tau}$ at the two extremes on the spectral gate.

Strong directions remain open. The expansion needs $\nu_j \ll \Delta\tau$. On directions with $\nu_j \gtrsim \Delta\tau$ there is no closed form for $\rho_j^{(\Delta\tau)}$, so conclusions requiring the whole spectrum are unavailable on this family. A limiting rank and the turning point of h are two of them. The grades of Definition 12 ask about weak directions only, so they hold under the assumptions of this subsection.

The sign of the interval on the softmax family. On the spectral family, in the approximate model of Appendix E.3, the interval is positive in the limit $\epsilon \to 0$, because every level $\tau > 0$ scales a weak direction e_j by the factor λ_j^τ, which vanishes with ν_j as $\epsilon \to 0$. The positive-guided term then contracts the view difference along e_j only at $\tau = 0$, so any placement that retrieves z^+ at a level $\tau > 0$ leaves the variance spread into e_j unshared, in forward and in reverse order alike, and $\tau = 0$ together with $\tau + \Delta\tau \geq 0$ and $\Delta\tau \neq 0$ gives $\Delta\tau > 0$. The softmax family has no such factor. With copies of the pair in the memory, the score of a weak direction tends to the addressing mass of those copies at every temperature, so the retrievals do not fade along e_j and the first step of the argument fails. Without copies, the maximal self-retrieval returns the nearest stored item rather than the query, and a retrieval trace far from the stored vectors can bend sharply, as in Figure 1(b), so the difference of two retrievals on such a trace can point away from the query and from e_j. Both cases leave the sign of the interval open on this family.

E.5 Open questions on the two families

Every value in this appendix is computed under a current-batch memory state. Three departures from it behave differently and share one cause. A queue makes the stored items stale, so copies are present but were produced by an earlier encoder and the maximal self-retrieval is the identity only approximately. A trainable prototype breaks Proposition 8, after which the thrust and the contraction rate lose their closed forms. Proposition 9 fails as well, so a degenerate subspace can be left. With a learned memory update, the one-sided optimum is a collapsed memory on the softmax family and, under $W \succeq 0$, a memory that vanishes on the data on the spectral family. On the spectral family the two-sided optimum makes the two retrievals coincide on the data once the views align (Appendices B.6 and B.7). The common cause is that the memory state no longer follows the configuration through its variance. That cuts the feedback loop which Appendix C.4 identifies as the source of resistance to collapse, or leaves it acting only while the views disagree. Closing this belongs to the memory update.

Appendix G lists the remaining open questions, from the encoder parameters to reverse order at finite ϵ.

One quantitative caveat is worth separating from the rest. Two configurations with $\tau > 0$ differ in overall scale by a power of λ, by Proposition 31. A static criterion compares directions of flow. An experiment compares endpoints after a fixed number of steps, so the configuration with the smaller scale may simply have traveled less far.

F EXPERIMENTAL DETAILS

Datasets. CIFAR-10 contains $60{,}000$ images of 32×32 pixels in 10 classes. STL-10 contains $13{,}000$ labeled and $100{,}000$ unlabeled images of 96×96 pixels in 10 classes. ImageNet-1k contains 1.28 million training images in $1{,}000$ classes.

Augmentation. Augmentation uses random resized crops, horizontal flips, colour jitter, random grayscale and Gaussian blur, with blur omitted on CIFAR-10. On STL-10 the blur has probability one on the first view and 0.1 on the second, and the second view also carries solarization with probability 0.2.

	CIFAR-10	STL-10	ImageNet-100	ImageNet-1k
Backbone	ResNet-18	ResNet-18	ResNet-50	ResNet-50
Effective batch	512	256, 512[§]	256	2048
Optimizer	AdamW, SGD[§]	AdamW	AdamW	AdamW, LARS[§]
Base learning rate	10^{-3}, 0.03[§]	10^{-3}	10^{-3}	10^{-3}, 0.3[§]
Weight decay	0.04, 5×10^{-4}[§]	0.04, 10^{-4}[§]	0.04	0.04, 10^{-6}[§]
Warmup epochs	10, 0[§]	10	10	10
Total epochs	800, 1000[§]	100	100	200
Schedule	cosine	cosine	cosine	cosine
Softmax family				
Projection, hidden / out	1024 / 128	512 / 128	512 / 128	2048 / 128
Temperature $T = \tau + \Delta\tau$	0.5	0.1	0.1	0.1
Spectral family				
Projection, hidden / out	2048 / 2048	2048 / 2048	2048 / 2048	2048 / 2048
Weight β on the potential term	20	20, 1[¶]	40	40

Table 5: Training configurations. [§]Where the two families differ, the spectral value comes first. [¶]The runs of Table 2 at $\beta = 1$.

Memory state. The spectral memory state follows (22) with $\epsilon = 10^{-4}$. On the softmax family the stored items are the representations of both views of the batch, copies of z and z^+ among them, and they carry no gradient.

Runs with a learned memory update. On the spectral family the memory state is an unconstrained 128×128 matrix. On the softmax family it holds $8{,}192$ prototypes in 128 dimensions, each normalized to unit length and then scaled by its own trainable gain. The projection head has widths 512, 512 and 128. The optimizer is AdamW with a warmup of 10 epochs and a cosine schedule. Training lasts 800 epochs on CIFAR-10 and 100 elsewhere. The effective batch is 512 on CIFAR-10, 256 on ImageNet-100, and on STL-10 512 for the spectral family and 256 for the softmax family. The learning rate is 10^{-3} times the batch divided by 256. The weight decay is 10^{-4} on CIFAR-10 and STL-10 and 0.04 on ImageNet-100, and the memory state is exempt from it. Figure 2 plots one run per configuration at every evaluation: every epoch on STL-10, every 2 on ImageNet-100 and every 10 on CIFAR-10.

Evaluation. The kNN classifier uses 20 neighbors and is evaluated on the test split with the training split as its bank. The linear probe trains for 100 epochs with SGD, momentum 0.9 and no weight decay, with learning rate 30 for the spectral rows of Table 1 on CIFAR-10 and STL-10 and 0.1 otherwise. Table 6 gives the softmax family on ImageNet-1k.

Half ranges. In Tables 7 and 8, a half range of zero means the seeds agree to the digits shown.

τ	kNN	Top-1	Top-5
$0.01T$	54.52	64.34	86.35
$0.1T$	54.85	62.65	85.62
$T^{\ddagger}$	0.10	0.10	0.50

Table 6: ImageNet-1k, softmax family, one seed. Linear-probe Top-1 and Top-5, and kNN on the same checkpoints. ‡Stopped at epoch 30.

Table 7: Table 1 without its one-seed ImageNet-1k column, with the half range over seeds beside each mean. ‡Stopped at epoch 30.

		CIFAR-10		STL-10		IN-100	
Family	$(\tau, \tau+\Delta\tau)$	kNN	LP	kNN	LP	kNN	LP
Spectral	$(0,1)$	$88.2_{\pm 0.1}$	$88.8_{\pm 0.3}$	$82.9_{\pm 0.1}$	$86.5_{\pm 0.3}$	$62.2_{\pm 0.2}$	$67.3_{\pm 0.4}$
	$(1/2,1)$	$85.5_{\pm 0.4}$	$85.7_{\pm 0.0}$	$72.1_{\pm 2.2}$	$74.1_{\pm 2.2}$	$25.8_{\pm 0.6}$	$33.0_{\pm 0.8}$
	$(1,1)$	$16.7_{\pm 0.2}$	$12.7_{\pm 1.1}$	$23.9_{\pm 0.8}{}^{\ddagger}$	$20.9_{\pm 5.6}{}^{\ddagger}$	$2.0_{\pm 0.3}{}^{\ddagger}$	$3.6_{\pm 0.1}{}^{\ddagger}$
	$(1,2)$	$89.2_{\pm 0.1}$	$88.6_{\pm 1.1}$	$80.4_{\pm 0.0}$	$84.5_{\pm 0.3}$	$56.1_{\pm 0.3}$	$64.0_{\pm 0.3}$
Softmax	$(0.01T, T)$	$89.7_{\pm 0.2}$	$91.1_{\pm 0.0}$	$80.8_{\pm 0.2}$	$85.0_{\pm 0.3}$	$57.8_{\pm 0.2}$	$63.7_{\pm 0.2}$
	$(0.1T, T)$	$80.3_{\pm 0.2}$	$84.2_{\pm 0.1}$	$80.8_{\pm 0.2}$	$85.1_{\pm 0.3}$	$57.1_{\pm 0.1}$	$64.0_{\pm 0.2}$
	(T, T)	$10.0_{\pm 0.0}$	$10.0_{\pm 0.0}$	$28.6_{\pm 0.5}{}^{\ddagger}$	$25.4_{\pm 0.8}{}^{\ddagger}$	$8.3_{\pm 0.6}{}^{\ddagger}$	$9.3_{\pm 0.9}{}^{\ddagger}$

Table 8: The three-seed rows of Table 2 with the half range beside each mean. ‡Stopped at epoch 30, where the row is read.

Family	$(\tau, \tau+\Delta\tau)$	ER	REC	COS	SF	kNN
Spectral, $\beta = 20$	$(0,1)$	$587_{\pm 1}$	$0.071_{\pm 0.001}$	$0.949_{\pm 0.000}$	$0.794_{\pm 0.002}$	$82.9_{\pm 0.1}$
	$(1/2,1)$	$1255_{\pm 653}$	$0.380_{\pm 0.253}$	$0.780_{\pm 0.004}$	$0.373_{\pm 0.311}$	$72.1_{\pm 2.2}$
	$(1,1)^{\ddagger}$	$944_{\pm 196}$	$0.350_{\pm 0.098}$	$0.025_{\pm 0.005}$	$0.010_{\pm 0.005}$	$23.9_{\pm 0.8}$
	$(1,2)$	$30_{\pm 0}$	$0.000_{\pm 0.000}$	$0.968_{\pm 0.000}$	$0.710_{\pm 0.009}$	$80.4_{\pm 0.0}$
Softmax	$(0.01T, T)$	$39_{\pm 0}$	$0.052_{\pm 0.000}$	$0.961_{\pm 0.000}$	$0.851_{\pm 0.002}$	$80.8_{\pm 0.2}$
	$(0.1T, T)$	$39_{\pm 0}$	$0.054_{\pm 0.005}$	$0.960_{\pm 0.000}$	$0.850_{\pm 0.002}$	$80.8_{\pm 0.2}$
	$(T, T)^{\ddagger}$	$1_{\pm 0}$	$0.002_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.053_{\pm 0.003}$	$28.6_{\pm 0.5}$

Diagnostic statistics. The diagnostics use a fixed set of $20{,}000$ images from the unlabeled split of STL-10. The second moment Σ, with eigenvalues ν_j, pools the normalized representations of both views. A direction counts as empty when $\nu_j / \operatorname{tr}\Sigma$ falls below the first percentile of the same ratio for $40{,}000$ uniform points on the sphere in the same dimension. Each checkpoint is thresholded on its own spectrum. For recovery, a random subspace would give the fraction of directions occupied at the next checkpoint. The analyzed checkpoints are every 20 epochs on the spectral family and every 10 on the softmax family, to epoch 100, or to epoch 30 for the rows marked ‡. Recovery compares rows only within a block of Table 2. At $(1/2, 1)$ with the weight the three seeds end near effective ranks of 450, 1540 and 1760.

G LIMITATIONS

Twelve questions remain open. Each is argued where it arises and collected here, and Section G.1 below collects the methods outside the framework.

Encoder parameters. Corollary 1 and Appendix D.9 treat the representations as free points on the sphere. A real update acts on encoder parameters and moves every representation at once, so Propositions 16 and 9 can both fail there. The results describe the flow this layer induces rather than the trajectory training follows. See Appendix E.5.

Memory states other than the current batch. Every stability statement in Section 3.3 is derived under a current-batch memory state. A queue, trained prototypes and a learned predictor each break it differently. The common reason is that the memory state no longer follows the configuration through its variance. See Appendices E.5, B.6 and B.7.

Batch estimates of the second moment. The analysis uses the second moment of the representation distribution, while training estimates it on the current batch. On the spectral family $d = 2048$, so a batch of B images leaves at least $d - 2B$ directions with zero batch second moment: 1024 on CIFAR-10 and 1536 on STL-10 and ImageNet-100. How this affects the stability statements remains open.

Constraints on a learned memory state. The analysis of a learned memory update rests on constraints that the runs of Section 4.3 do not impose. Appendix B.6 assumes unit-norm prototypes, while each run scales its normalized prototypes by trainable gains. Appendix B.7 assumes a symmetric W in the one-sided loss and $W \succeq 0$ in Proposition 3(ii), while each run trains an unconstrained W, which need not stay symmetric or positive semidefinite. Section 4.3 draws on Propositions 2, 3(i) and 4, whose proofs use neither constraint. How the rest of the analysis extends to an unconstrained memory state remains open. See Appendix F.

The spectral rows away from the maximal self-retrieval. Section 4.2 tests Corollary 1 on the rows of Table 2 at $\beta = 1$, which come from STL-10 with one seed. At $\tau > 0$ the weight on the potential term changes the shape of the thrust, and at a zero interval it leaves a non-zero self-guided term. Proposition 29 gives the minimum number of directions that the weight sets, and Section 4.2 reads the weighted rows through it. More seeds and the other datasets at $\beta = 1$ remain to be run. See Appendix E.5.

Strong directions on the softmax family. The expansion of Proposition 34 needs $\nu_j \ll \Delta\tau$, so a limiting rank and the turning point of the thrust are unavailable there. The grades of Definition 12 ask about weak directions only and hold under the assumptions of Appendix E.4.

The contraction rate without copies. When the memory state holds no copy of the pair, the rate is carried by events crossing a cell boundary. A lower bound on it remains open. See Appendix E.4.

From L2 to L3. In the approximate model of Proposition 32, the pairs with $\tau = 0$ and $\Delta\tau > 0$ reach L2. Under the moment factorization isotropy is the only candidate endpoint at $\tau = 0$, and the basins of attraction remain unproved. The runs at $(0, 1)$ on STL-10 end with effective ranks of 57 at $\beta = 1$ and 587 at $\beta = 20$, far below the isotropic value $d = 2048$. Whether they approach isotropy slowly or stop at another endpoint remains open.

The injection model. Proposition 32 takes the injection proportional to the variance, and the middle term of (33) is known to exist and is left uncomputed. Together they make $\tau_{\max}$ an estimate

rather than a consequence, which is why Proposition 32 is stated as an approximate result. A rate fixed apart from the variance sets a retention floor at $\tau = 0$ as well.

Reverse order at finite ϵ. The window of Proposition 32 bounds τ rather than the sign of $\Delta\tau$. On a configuration with $n \geq 2$, a reverse order with $0 < \tau < \tau_{\max}$ can pass both sides when ϵ is small but positive. Its rank-one configuration keeps empty directions stable once ϵ is small. At $d = 2048$ and $\epsilon = 10^{-4}$ this excludes every reverse order tested across the window. Whether some finite ϵ admits a reverse order that passes on every degenerate configuration remains open. As $\epsilon \to 0$ the pairs with $\tau = 0$ and $\Delta\tau > 0$ are the ones left. See Appendix E.3.

Exchangeable views on STL-10. The two views of STL-10 use different blur and solarization probabilities (Appendix F), so they follow different distributions. Propositions 24 and 25 and the bound $|C_j| \leq \nu_j$ on the shared fraction assume exchangeable views, and the diagnostics of Table 2 come from STL-10.

Runs stopped early. The runs marked ‡ stop at epoch 30, since their kNN accuracy had stayed at or below its first evaluation until then, so whether they recover later is untested. Runs with the reverse order $\Delta\tau < 0$ were also stopped before completion, so that order awaits an experimental test.

G.1 Methods outside the framework

Five kinds of method lie outside the analysis, each for a stated reason.

Distribution estimates used without a retrieval. Methods that represent the distribution as a cross-correlation matrix or as per-dimension variances, and penalize that object directly, have no representation-space retrieval to analyze. The score is undefined for them, and so are the directional quantities of Appendix D.10.

Gradients beyond differences of two retrievals. A general training signal is a pullback. Send z through a map $\phi_1 \in C^1$ and z^+ through a map ϕ_2, and take $\ell(\phi_1(z), \mathrm{sg}[\phi_2(z^+)])$. Then $\nabla_z \mathcal{L} = J_{\phi_1}(z)^\top \nabla_1 \ell$, and the Jacobian entangles the two views. A nonlinear predictor belongs here.

The deployed cosine variants of the predictor methods belong here as well. Normalizing after the predictor gives

$$\nabla_z \mathcal{L} \;\propto\; \omega(z, z^+)\, W^\top W z - W^\top z^+, \qquad \omega = \frac{\langle Wz, z^+ \rangle}{\|Wz\|^2}, \tag{41}$$

which is still a difference of two retrievals. But a gain ω now multiplies the retrieval at $\tau + \Delta\tau$ and varies with both views. The self-guided term is still the one Definition 9 gives, and it is no longer a displacement between two fixed gates. The positive-guided term also picks up $[\omega(z, z^+) - \omega(z, z)]\, W^\top W z$. The linear form placed on the traces in Appendix B is an interpretable special case of (41) and of the pullback above. The three are distinct.

Biased affinities and hetero-associative retrieval. An auto-associative retrieval with zero bias guarantees a generator, and Appendix C.5 shows how dropping either can remove it. Centering in DINO adds a running statistic to the teacher's affinities, and Sinkhorn normalization in SwAV likewise changes the affinities rather than the gate. Retrievals in that family therefore lose that guarantee, and Appendix B places their zero-bias form. Teacher and student sharing one set of prototypes is an idealization as well, since an exponential moving average makes the two close rather than equal.

Joint dynamics of memory and encoder. The representation distribution is taken as given here, and the joint dynamics of the memory state and the encoder remain outside the analysis.

A shared parameterization of addressing and response. The retrieval of Definition 1 consists of addressing, which reads the keys, and response, which reads the values. We keep one parameterization for both, $V = K$, for three reasons.

Decoupling inside one retrieval generally breaks integrability, since Proposition 7 relies on $V = K$ to make the Jacobian $V^\top \mathbf{J}_{\hat{y}} K$ symmetric. With it go the potential, the order of the levels, and everything in Sections 3.2 and 3.3 that rests on those.

The deviation that occurs in practice is a different one: two retrievals address different memory states, as with teacher and student keys or with a centered teacher. That is a mismatch between two memory states rather than a freedom inside one retrieval. Each deviation already has a closed form, the bias in Appendix C.5 and the key mismatch in Appendix B.

In that family the value matrix is the Jacobian of the affinity map, fixed by the chain rule rather than chosen. Treating it as a free choice of values would present a chain-rule factor as a modeling decision. Its accurate place is the pullback above with ϕ_1 the addressing map, where the loss lives on the addressing weights rather than on the retrieval, which is also why it lacks a thrust spectrum.