SPURIOUS STATES ARE LABELS: A MULTI-LABEL CODE IN THE HOPFIELD-RBM DUALITY

Anonymous authors
Paper under double-blind review

ABSTRACT

For forty years, the superpositions of stored patterns known as spurious states have been treated as the defining defect of the Hopfield model. We read them as an answer: a mixture of L patterns is a composite signal, and identifying its constituents is multi-label classification. A Hopfield configuration cannot represent a set of labels; its dual restricted Boltzmann machine (RBM) can. We show that the Hopfield-RBM duality analytically fixes this representation: across N visible neurons, active labels sit on the hidden layer at scale $\sqrt{N}$ with an amplitude matching the archetype-mixture overlap. Under this label assignment, the class averages of noisy training examples form an exact fixed point in mean of one-step supervised contrastive divergence. We prove a finite-size drift bound that decomposes into four vanishing error channels: thermal sampling, dataset noise, mini-batch consensus, and finite-dimensional fluctuations. For continuous mixtures combined prior to quantization (such as musical chords and hyperspectral pixels), this overlap amplitude transitions to an arcsine law via one-bit random projections. Across four benchmarks on three public datasets, the resulting configuration acts as an optimal analytical warm start: computed in a single pass without gradient updates, it surpasses random initializations by 0.08 to 0.43 in score, saving thousands of training steps. Finally, the induced parameter-free read-out rivals computationally intensive one-bit compressed sensing baselines at a fraction of their inference cost.

1 INTRODUCTION

The Hopfield model (Hopfield, 1982) stores K patterns by a Hebbian rule (Hebb, 1949) and retrieves them as minima of a quenched energy landscape (Amit et al., 1985b), which has minima the prescription did not ask for. Take an odd number L of stored patterns and superpose them site by site by majority: the result is a *spurious state*, metastable at low temperature (Amit et al., 1985b;a). The retrieval dynamics falls into these states and returns something that is not any of the stored patterns, which is why four decades of research have focused on suppressing them or escaping them through unlearning algorithms and competitive modular dynamics (Hertz et al., 1991; Agliari et al., 2026; 2025).

There is another way to read the same object. A mixture of L patterns is exactly what a composite signal looks like: a chord is a superposition of notes, a pixel of a hyperspectral scene a superposition of the materials under it. Ask the network not which single pattern the input resembles but which patterns it is made of, and the metastable state stops being a trap and becomes a query whose answer is a set of L labels. That is multi-label classification, and the Hopfield model has nowhere to write the answer: its state is a configuration, not a set. The machine that can write it is the restricted Boltzmann machine (Ackley et al., 1985), the same object under mild conditions (Barra et al., 2012; Agliari et al., 2012; Mézard, 2017; Tubiana & Monasson, 2017; Decelle & Furtlehner, 2021; Di Sarra et al., 2025; Smart & Zilman, 2021), whose hidden layer has one unit per pattern, so a set of labels is the set of units that light up. The correspondence is sharper in a supervised setting where only noisy labelled examples of the patterns are available: the class averages of the examples replace the patterns in Hebb's rule (Alemanno et al., 2023a;b), and for a single pattern at a time Leonelli et al. (2021) argue that those averages are asymptotically stationary under contrastive divergence.

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.

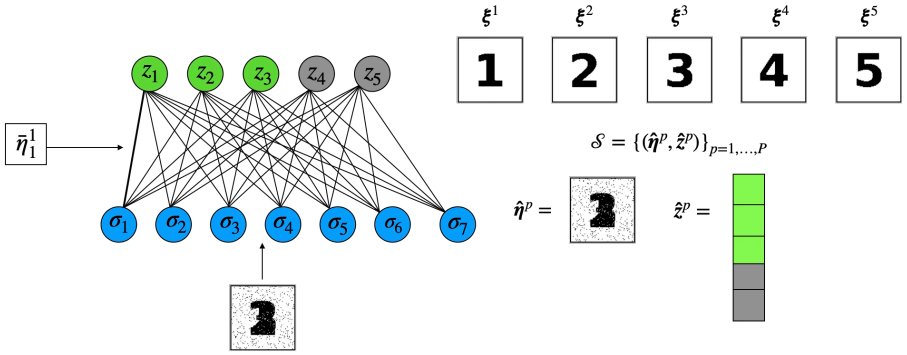


Figure 1: The machine and the task, with $N = 7$ visible neurons (blue) fully connected to $K = 5$ hidden neurons and the bold edge marking the weight $\bar{\eta}_1^1$. The weights are the class averages of noisy examples of the K archetypes, the visible layer is clamped to a noisy superposition of $L = 3$ of them, and the machine answers on the hidden layer: the three active units (green) name the constituents and the two silent ones (grey) the absent archetypes. The amplitude of the active units is what (13) fixes.

Three literatures have met parts of this object without asking this question of it. Dense associative memories and modern Hopfield networks (Krotov & Hopfield, 2016; Ramsauer et al., 2021) remove spurious states by sharpening the energy function; while effective for isolated retrieval, this non-linear sharpening suppresses mixtures and prevents multi-label decoding. We take the opposite path: preserving quadratic interactions precisely because physical signals add linearly, and tasking the dual hidden layer with reading their superpositions. In interpretability, the superposition hypothesis and the sparse autoencoders built on it (Elhage et al., 2022; Bricken et al., 2023) read an activation as a superposition of features and try to name them, which is our decoding problem with the dictionary unknown instead of given. And recovering which sources are present from measurements that have been reduced to their sign is one-bit compressed sensing (Boufounos & Baraniuk, 2008; Plan & Vershynin, 2013), whose central identity (Goemans & Williamson, 1995; Charikar, 2002) is also the bundling law of hyperdimensional computing (Plate, 1995; Kanerva, 2009) and reappears here as Lemma 2. What none of these frameworks supplies is the object required here: weights constructed from noisy finite-sample class averages, accompanied by an exact non-asymptotic characterization of how contrastive learning dynamics acts on them.

Our contributions are threefold. **(i)** We formalise the multi-label classification of spurious states in the dual RBM, proving that archetype labels sit on the hidden layer at scale $\sqrt{N}$ with amplitude c_L (the archetype-mixture overlap); clamped there, the empirical class averages are an exact fixed point in mean of supervised CD-1, with a non-asymptotic drift bound that decomposes into four vanishing channels, thermal, dataset, mini-batch and finite size (Theorem 1). **(ii)** For physical mixtures that combine linearly before quantisation we prove that the amplitude passes from c_L to an arcsine law $g_L = \frac{2}{\pi} \arcsin L^{-1/2}$ under one-bit random projections (Lemma 2), which carries the construction from Rademacher patterns to measured signals. **(iii)** We delineate the boundary between the two, on four settings from three public benchmarks (Engel et al., 2017; Zhu et al., 2014; Humphrey et al., 2018), and measure what survives it: the class averages start 0.08 to 0.43 above a random initialisation and save up to $1.6 \cdot 10^3$ CD-1 steps, reaching scores a random start does not reach within the budget; exact stationarity does not survive, the gradient rescaling the class averages and decorrelating the archetypes while barely changing the spectral content they store; and the parameter-free read-out $Z_\mu = N^{-1} \sum_i w_i^\mu \sigma_i$ matches or beats BIHT, OMP and LASSO on complex mixtures at $2NK$ flops against their iterative dictionary passes. Figure 1 shows the machine and the task; proofs, protocols and the complete numerical results are in the appendices.

2 The task, the machine and the fixed point

Let $\boldsymbol{\sigma} \in \{-1, +1\}^N$ be a configuration of N binary neurons and let $\boldsymbol{\xi}^\mu \in \{-1, +1\}^N$, $\mu = 1, \dots, K$, be K patterns with independent Rademacher entries, called *archetypes* below. The Hopfield model

assigns to a configuration the Hebbian cost

$$\mathcal{H}^{\mathrm{HM}}(\boldsymbol{\sigma};\boldsymbol{\xi}) := -\frac{1}{2N}\sum_{\mu=1}^{K}\sum_{i,j=1}^{N}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j = -\frac{N}{2}\sum_{\mu=1}^{K} m_\mu(\boldsymbol{\sigma}\mid\boldsymbol{\xi})^2, \quad m_\mu(\boldsymbol{\sigma}\mid\boldsymbol{\xi}) := \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\sigma_i, \tag{1}$$

with m_μ the Mattis magnetisation, and draws configurations from $\mathcal{P}^{\mathrm{HM}}(\boldsymbol{\sigma}) \propto e^{-\beta\mathcal{H}^{\mathrm{HM}}}$. We work in the low load regime $K = O(1)$, where the archetypes are mutually orthogonal up to $O(N^{-1/2})$. Given a set A of L archetypes with L odd, the *spurious state* built on A is their site by site majority,

$$\boldsymbol{s}(A) := \mathrm{sign}\Big(\sum_{\mu\in A}\boldsymbol{\xi}^\mu\Big), \tag{2}$$

a metastable minimum of (1) at low temperature. The parity restriction is what makes (2) well defined and it applies to this combinatorial branch only.[1] What matters here is not the stability of $\boldsymbol{s}(A)$ but how much it remembers of the archetypes that built it, and the following lemma answers that.

Lemma 1 (Mixture constant, Amit et al., 1985a)**.** *Let L be odd and let $\boldsymbol{s}(A)$ be the spurious state* (2)*. Then*

$$\mathbb{E}_{\boldsymbol{\xi}}\big[\xi^\mu s\big] = \begin{cases} c_L := \binom{L-1}{\frac{L-1}{2}}2^{-(L-1)}, & \mu\in A,\\ 0, & \mu\notin A, \end{cases} \tag{3}$$

with $c_1 = 1$, $c_3 = 1/2$, $c_5 = 3/8$ and $c_L \simeq \sqrt{2/(\pi L)}$ for large L.

The constant is the classical overlap of a symmetric mixture state with its constituents; we restate it because everything below is built on it and prove it in Appendix A for self-containment. A spurious state has not forgotten what it is made of: it keeps a macroscopic and explicitly computable overlap c_L with each constituent and none with the absent archetypes, so the K numbers m_μ already form a multi-label code, and the decay $c_L \simeq \sqrt{2/(\pi L)}$ keeps that code legible for rich mixtures.

The supervised setting, and the machine. In any realistic setting the archetypes are not available. What is available is a finite sample of corrupted copies, called *examples*, and the label saying which archetype each came from. Following Alemanno et al. (2023a) we generate M examples per archetype as

$$\eta_i^{\mu,m} := \xi_i^\mu\,\chi_i^{\mu,m}, \qquad \mathbb{P}(\chi_i^{\mu,m} = \pm 1) = \frac{1\pm r}{2}, \qquad m = 1,\dots,M, \tag{4}$$

with $r\in(0,1]$ the *quality* of the dataset, and we replace the unknown archetypes by the class averages

$$\bar{\boldsymbol{\eta}}^{\,\mu} := \frac{1}{rM}\sum_{m=1}^{M}\boldsymbol{\eta}^{\mu,m}, \qquad \mathbb{E}_{\boldsymbol{\chi}}\big[\bar\eta_i^\mu\big] = \xi_i^\mu, \qquad \mathbb{E}_{\boldsymbol{\chi}}\big[(\bar\eta_i^\mu - \xi_i^\mu)^2\big] = \rho := \frac{1-r^2}{Mr^2}. \tag{5}$$

The division by r makes the estimator unbiased, so that its whole imperfection is carried by the variance ρ; we call ρ the *dataset entropy* (Alemanno et al., 2023a), following the convention of that literature, but it is a variance and nothing informational is meant by the name. It is not privileged information, since it can be estimated from the examples alone.[2] Substituting (5) into (1) defines the *supervised Hopfield model*

$$\mathcal{H}^{\mathrm{HM}}(\boldsymbol{\sigma};\bar{\boldsymbol{\eta}}) := -\frac{1}{2N}\sum_{\mu=1}^{K}\sum_{i,j=1}^{N}\bar\eta_i^\mu\bar\eta_j^\mu\sigma_i\sigma_j. \tag{6}$$

The machine we train is a Bernoulli and Gaussian restricted Boltzmann machine on $N+K$ neurons: the visible layer carries the N binary neurons σ_i, the hidden layer K real valued neurons z_μ with a Gaussian prior, and neurons interact only across layers,

$$\mathcal{H}^{\mathrm{RBM}}(\boldsymbol{\sigma},\boldsymbol{z}\mid\boldsymbol{w}) = -\frac{1}{\sqrt{N}}\sum_{\mu=1}^{K}\sum_{i=1}^{N}w_i^\mu\,\sigma_i z_\mu + \frac{1}{2}\sum_{\mu=1}^{K}z_\mu^2. \tag{7}$$

[1] The physical superpositions of Section 3 are sums, they need no parity, and the hyperspectral experiments use $L = 2, 4, 6$ as freely as $L = 3, 5$.

[2] $\hat r^{\,2} := (M\widehat{\mathcal{R}} - 1)/(M-1)$ with $\widehat{\mathcal{R}} := (KN)^{-1}\sum_{\mu,i}(M^{-1}\sum_m \eta_i^{\mu,m})^2$ is unbiased for r^2. The estimators are not unit vectors, $\|\bar{\boldsymbol{\eta}}^{\,\mu}\|^2/N \to 1+\rho$, so conventions with unit couplings require $\beta\mapsto\beta(1+\rho)$.

Its conditional laws, derived in Barra et al. (2012) and Leonelli et al. (2021), are

$$\mathcal{P}(\boldsymbol{z} \mid \boldsymbol{\sigma}) = \prod_{\mu=1}^{K} \sqrt{\frac{\beta}{2\pi}}\, e^{-\frac{\beta}{2}(z_\mu - u_\mu(\boldsymbol{\sigma}))^2}, \qquad \mathcal{P}(\boldsymbol{\sigma} \mid \boldsymbol{z}) = \prod_{i=1}^{N} \frac{e^{\beta \sigma_i h_i(\boldsymbol{z})}}{2\cosh\big(\beta h_i(\boldsymbol{z})\big)}, \tag{8}$$

with local fields

$$u_\mu(\boldsymbol{\sigma}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i^\mu \sigma_i = \sqrt{N}\, m_\mu(\boldsymbol{\sigma} \mid \boldsymbol{w}), \qquad h_i(\boldsymbol{z}) = \frac{1}{\sqrt{N}} \sum_{\mu=1}^{K} w_i^\mu z_\mu. \tag{9}$$

The hidden layer is thus a noisy read-out of the overlaps between the visible configuration and the weight vectors, which is precisely the multi-label code of Lemma 1. Integrating the Gaussian hidden variables out of the partition function by a Hubbard and Stratonovich transformation gives

$$\mathcal{Z}^{\mathrm{RBM}}(\beta, \bar{\boldsymbol{\eta}}) = \sum_{\boldsymbol{\sigma}} \exp\Big\{ \frac{\beta}{2N} \sum_{\mu=1}^{K} \sum_{i,j=1}^{N} \bar{\eta}_i^\mu \bar{\eta}_j^\mu \sigma_i \sigma_j \Big\} = \mathcal{Z}^{\mathrm{HM}}(\beta, \bar{\boldsymbol{\eta}}), \tag{10}$$

so planting $w_i^\mu = \bar{\eta}_i^\mu$ turns the machine into the supervised Hopfield model (6) whose metastable states are exactly the queries it has to classify.

The dataset and the label code. Our training set is a collection of P items,

$$\mathcal{S} := \big\{ (\hat{\boldsymbol{\eta}}^{\,p}, \hat{\boldsymbol{z}}^{\,p}) \big\}_{p=1,\dots,P}, \tag{11}$$

where the p-th item carries a tuple $\hat{\boldsymbol{\mu}}^p$ of L distinct labels and its query is the noisy spurious state assembled from one fresh example of each constituent,

$$\hat{\eta}_i^p := \mathrm{sign}\Big(\sum_{\mu \in \hat{\boldsymbol{\mu}}^p} \eta_i^{\mu,p} \Big), \qquad i = 1, \dots, N. \tag{12}$$

The examples that build the weights and those that build the queries are kept distinct. We estimate the underlying distribution by the uniform measure over the dataset, $\mathcal{Q}(\boldsymbol{\sigma}, \boldsymbol{z}) := P^{-1} \sum_p \delta(\boldsymbol{\sigma} - \hat{\boldsymbol{\eta}}^{\,p}) \delta(\boldsymbol{z} - \hat{\boldsymbol{z}}^{\,p})$.

It remains to say how the label is written on the hidden layer, and the composite query leaves its mark here. By (8) a hidden configuration drawn at fixed visible state is its mean plus a fluctuation of order $\beta^{-1/2}$, so $z_\mu = \sqrt{N}\, m_\mu(\boldsymbol{\sigma} \mid \boldsymbol{w}) + O_{\mathcal{P}}(\beta^{-1/2})$. Asking $\boldsymbol{z}$ to be of order one would force $m_\mu = O(N^{-1/2})$, a visible layer with no macroscopic overlap with any archetype and hence a vanishing field h_i: nothing would be retrieved. Labels are forced to the scale $\sqrt{N}$ if retrieval is to be macroscopic, and their amplitude follows by evaluating the same identity where the duality holds, weights at the archetypes and visible layer at the clean mixture, where $m_\mu \to c_L \mathbf{1}\{\mu \in A\}$ by Lemma 1. The label code of the p-th item is therefore

$$\hat{z}_\mu^p = \sqrt{N}\, c_L\, \mathbf{1}\{\mu \in \hat{\boldsymbol{\mu}}^p\}, \qquad \mu = 1, \dots, K. \tag{13}$$

Since the label lists the constituents, L is known and c_L is an explicit function of it, so (13) uses nothing the machine is not told.

Remark 1. The factor c_L is fundamental: it reflects the exact overlap amplitude that the RBM reconstructs on its hidden units when driven by a spurious state on the visible layer. To achieve equilibrium in the contrastive divergence update (14), the positive phase must precisely mirror this quantity.

The learning rule. Training means matching the model distribution to $\mathcal{Q}$. Since (7) is linear in the weights, the gradient of the Kullback and Leibler divergence (Coolen et al., 2005) gives, with learning rate ϵ,

$$\Delta w_i^\mu = \frac{\epsilon\beta}{\sqrt{N}} \Big(\mathbb{E}_{\mathcal{Q}}\big[\sigma_i z_\mu\big] - \mathbb{E}_{\mathcal{P}}\big[\sigma_i z_\mu\big] \Big), \qquad \forall\, i, \mu. \tag{14}$$

The *positive phase* is an average over the data with both layers clamped to the item; the *negative phase* is an average over the equilibrium measure, intractable at large sizes, which we estimate by one step of contrastive divergence (Hinton, 2002; Hinton et al., 2006) using the two conditionals (8).

A single query is noisy, and clamping the visible layer to it would leave a residual no other parameter can control, so we present the machine with a consensus: given the tuple $\hat{\boldsymbol{\mu}}$, let $\mathcal{B}$ be a *mini-batch* of B items carrying that same tuple, B odd, and let

$$s_i(\mathcal{B}) := \operatorname{sign}\Big(\sum_{p\in\mathcal{B}} \hat{\eta}_i^p\Big) \in \{-1,+1\}, \qquad i = 1,\dots,N, \tag{15}$$

be the *cue*. It is binary, as the visible layer requires, and a cleaner replica of $\boldsymbol{s}(\hat{\boldsymbol{\mu}})$ than any of the items it is built from. The resulting update is Algorithm 1 of Appendix A.

Although written in terms of expectations, $\Delta\boldsymbol{w}$ is a random variable. Four sources of randomness are internal to the algorithm (the item, the mini-batch, the uniform variables that resample the visible layer, the Gaussian noise of the negative phase); the fifth is the generation of the data. We write $\mathbb{E}_{\rm alg}$ for the average over the first four at fixed data, $\mathbb{E}_{\mathcal{S}}$ for the average over the corruption variables, and $\mathbb{E} := \mathbb{E}_{\mathcal{S}}\mathbb{E}_{\rm alg}$. The archetypes are quenched throughout and the algorithm never accesses them.

Definition 1. A weight configuration $\boldsymbol{w}^\star$ is a *fixed point in mean* of supervised CD-1 if $\mathbb{E}[\Delta w_i^\mu] = 0$ for every $i = 1,\dots,N$ and every $\mu = 1,\dots,K$ when Algorithm 1 is run at $\boldsymbol{w} = \boldsymbol{w}^\star$.

The definition constrains the mean of one update and not its fluctuations, which stay of order one where the mean vanishes; at a finite learning rate the weights perform a stochastic approximation around $\bar{\boldsymbol{\eta}}$ with a stationary spread that shrinks with $\epsilon\beta$. Besides c_L, one more constant enters, measuring how well a single query reproduces the clean spurious state it stands for,

$$\vartheta_L(r) := r\,\mathbb{P}\big(V = V'\big), \qquad V, V' \text{ i.i.d. } \mathrm{Bin}\Big(\tfrac{L-1}{2}, \tfrac{1-r}{2}\Big), \tag{16}$$

which is $\vartheta_1(r) = r$ and decreases with L.

Theorem 1. *Consider the machine trained by the supervised CD-1 step of Algorithm 1, with weights planted at $\boldsymbol{w} = \bar{\boldsymbol{\eta}}$, the visible layer clamped to the cue and the hidden layer to the label code (13). Let ρ be as in (5), let $\vartheta := \vartheta_L(r)$, let T be the number of distinct label tuples carried by $\mathcal{S}$, and set $C := 20/r$ and $c_1(\theta) := (1-r^2)\theta^2/(2L^2)$. Then, for every $N \geq 4$, every odd $B \geq 1$ such that each tuple is carried by at least B items, every $\beta > 0$ and every $\gamma > 0$, with probability at least $1 - 2KTe^{-N\gamma^2/2}$ over the realisation of the archetypes,*

$$\max_{i,\mu} \big|\mathbb{E}[\Delta w_i^\mu]\big| \;\leq\; \epsilon\beta\Big[\inf_{\theta\in(0,1)} C\big(\underbrace{e^{-2(1-\theta)\beta c_L}}_{\textit{thermal}} + \underbrace{L\,e^{-c_1(\theta)/\rho}}_{\textit{dataset}}\big) + \underbrace{2e^{-B\vartheta^2/2}}_{\textit{mini-batch}} + \underbrace{\gamma}_{\textit{finite size}}\Big]. \tag{17}$$

The free θ trades the two exponential channels against each other; the symmetric $\theta = 1/2$ gives the readable exponents βc_L and $(1-r^2)/(8L^2\rho)$, and the infimum is never worse. Choosing $\gamma = \sqrt{2\ln(2KTN)/N}$, the bound holds with probability at least $1 - N^{-1}$, and its right-hand side vanishes as $\beta \to \infty$, $M \to \infty$, $B \to \infty$ and $N \to \infty$ with $\ln(KT) = o(N)$, in any order, so that $\bar{\boldsymbol{\eta}}$ is a fixed point in mean of supervised CD-1 in the sense of Definition 1.

Remark 2. The closest statement in the literature is that of Leonelli et al. (2021), who plant the plain sample mean in the weights and show, for the retrieval of one archetype at a time, that the mean update vanishes in the joint limit of large β, M and N. Three things separate Theorem 1 from it. The task here is a set of L labels rather than one, which is what forces the amplitude (13) and recovers their unit clamp only at $L = 1$, where $c_1 = 1$. The weights are debiased by r, so that a single number ρ carries the whole imperfection of the estimate. And the statement holds at finite N, M, B and β, which is what makes the four channels visible.

Appendix B verifies the four channels one at a time on Rademacher archetypes, where every hypothesis holds and the bound can be checked against its own constants, and shows the planted weights classifying there.

3 Superpositions that happen in the world

The queries in (12) are formed by majority vote over discrete binary configurations, whereas physical signals are continuous mixtures that add linearly in amplitude or power before any quantisation occurs. Consequently, real-world data violate three core assumptions of Theorem 1: the underlying archetypes

are not Rademacher vectors, prototype corruptions exhibit strong spatial or spectral correlations across sites, and mixtures do not arise from a vote. Rather than treating this discrepancy as a barrier, we investigate the robustness of the dual construction under this domain shift. Two foundational questions must be addressed: how to map continuous signals onto the binary visible layer without destroying their linear superposition, and which analytical constant governs the representation when c_L no longer applies.

From a measured signal to the visible layer. Let $\boldsymbol{x}(o) \in \mathbb{R}^d$ be a feature vector of a raw observation o. Of the feature map we require one property, that it be *linear in the physical quantity that adds*, acoustic power for audio and reflectance for a hyperspectral scene, so that up to cross terms that incoherent sources average away

$$\boldsymbol{x}\Big(\sum_{\mu\in A} a_\mu\, o^\mu\Big) \;=\; \sum_{\mu\in A} a_\mu\, \boldsymbol{x}(o^\mu), \qquad a_\mu > 0, \tag{18}$$

with $a_\mu > 0$ the amplitude with which source μ enters. The requirement is less restrictive than it looks: on the sensors of Section 4 it is not an assumption at all but the linear mixing model, the accepted physics of the instrument (Zhu et al., 2014), and it holds on the data we use.

The visible layer of the machine is binary, so a real-valued feature vector cannot be placed on it as it stands. We therefore fix a map from features to visible configurations, and call it the *encoder*:

$$\boldsymbol{\sigma}(\boldsymbol{x}) = \operatorname{sign}\big(\boldsymbol{\Omega}\,\boldsymbol{T}\,\boldsymbol{x}\big) \in \{-1,+1\}^N, \tag{19}$$

with $\boldsymbol{T}$ a linear whitening estimated on the prototype pool alone and $\boldsymbol{\Omega}$ a fixed random Gaussian matrix; neither is learnt from the labels, neither is fitted on a query (Appendix D.2). The composition $\boldsymbol{\Omega T}$ is linear, so by (18) the real valued projection of a mixture is exactly the sum of the projections of its sources. The sign does not commute with that sum, and the constant below is the price of that failure to commute. The network architecture is identical to Section 2, with weights set to the class averages of isolated prototype codes, and queries constructed from held-out sources. At inference, the hidden layer provides continuous matched-filter scores $Z_\mu = N^{-1}\sum_{i=1}^N w_i^\mu \sigma_i \in [-1,1]$. Because the active activation scale decays with the mixture order L, any fixed detection threshold would require order-dependent calibration. We instead employ a parameter-free *largest-gap rule*: sorting the activations $Z_{(1)} \le \cdots \le Z_{(K)}$, we place the decision threshold at the midpoint of the maximal gap between consecutive values and predict all classes above this cut (Appendix D.3). This decoding requires no hyperparameter tuning and operates without knowledge of L.

Lemma 2 (Additive mixtures)**.** *Let $\boldsymbol{x}^1,\dots,\boldsymbol{x}^L \in \mathbb{R}^d$, let $\boldsymbol{D} \in \mathbb{R}^{N\times d}$ have i.i.d. Gaussian entries, and encode a signal $\boldsymbol{v}$ as $\operatorname{sign}(\boldsymbol{D}\boldsymbol{v})$ componentwise. For the mixture $\boldsymbol{y} = \sum_\mu a_\mu \boldsymbol{x}^\mu$ with $a_\mu > 0$,*

$$\mathbb{E}_{\boldsymbol{D}}\big[\operatorname{sign}(\boldsymbol{D}\boldsymbol{x}^\mu)_i \operatorname{sign}(\boldsymbol{D}\boldsymbol{y})_i\big] = \frac{2}{\pi}\arcsin\kappa_\mu, \qquad \kappa_\mu = \frac{\langle \boldsymbol{x}^\mu, \boldsymbol{y}\rangle}{\|\boldsymbol{x}^\mu\|\,\|\boldsymbol{y}\|}, \tag{20}$$

for every i. If the sources are mutually orthogonal, of equal energy and equally weighted, then $\kappa_\mu = L^{-1/2}$ and the overlap is $g_L := \frac{2}{\pi}\arcsin L^{-1/2} = \frac{2}{\pi\sqrt{L}}(1+O(L^{-1}))$.

Equation (20) is Grothendieck's identity (Goemans & Williamson, 1995; Charikar, 2002) applied to $\boldsymbol{D} = \boldsymbol{\Omega T}$; what the lemma adds is its evaluation on a mixture, which is what produces the $L^{-1/2}$. Both c_L and g_L decay as $L^{-1/2}$ and differ only by the constant $\sqrt{2/\pi}$, from $c_3 = 0.500$ to $g_3 = 0.392$, and in the two places where c_L enters Section 2, the label amplitude (13) and the reconstruction margin of Lemma 8, g_L is the quantity the machine now returns. We do not claim that Theorem 1 directly extends under this substitution, as its formal proof relies on the discrete Rademacher structure and independent site noise. Instead, Lemma 2 formalizes a concrete *transfer hypothesis*: if the duality mechanism governs real-world representations, clamping hidden units at the arcsine amplitude g_L should still provide a near-stationary operating point. Specifically, we predict that: (i) the class averages remain an exceptional initialization with small residual drift; (ii) supervised contrastive divergence balances near g_L rather than naive binary clamps; and (iii) the relative read-out scale preserves the theoretical geometry across mixture orders.

While Lemma 2 characterizes individual sign bits, the network evaluates class averages over noisy prototype observations. Let $\kappa \in (0,1]$ denote the pre-sign correlation between a prototype and a held-out source of the same class entering the query. Applying (20) across the prototype ensemble

yields an active read-out of $Z_\mu = \hat{r}^{-1}\frac{2}{\pi}\arcsin(\kappa/\sqrt{L})$ on constituent units, while inactive units fluctuate around zero within an $O(N^{-1/2})$ envelope.

Crucially, normalizing the read-out relative to its isolated-source baseline ($L = 1$) eliminates both the empirical quality factor $\hat{r}$ and the latent correlation κ. This exposes a universal, parameter-free theoretical band:

$$R(L) := \frac{\arcsin(\kappa/\sqrt{L})}{\arcsin\kappa} \in \Big[g_L,\ L^{-1/2}\Big], \tag{21}$$

proved as Corollary 1 and tested on physical recordings in Appendix G.

4 What the duality is worth on measured signals

Three public datasets provide four empirical benchmarks, each probing the framework under a distinct departure from the ideal theoretical assumptions. *NSynth* (Engel et al., 2017) evaluates generalization across *timbre*: the archetypes are $K = 49$ semitone pitch classes, queries are waveform sums of L notes, and the 53 instruments are split into two disjoint halves (prototypes drawn from the first, queries from the second) so that the network never encounters the test timbres during initialization. *The Urban hyperspectral library* (Zhu et al., 2014) embeds the linear reflectance physics of optical sensors (18): each of the $K = 6$ materials supplies M labelled pixels of the scene, distributed as a spectral library, whose class average defines the archetype and whose spread constitutes the dataset noise variance. *The Samson image* evaluates unconstrained, measured scenes: its 9025 pixels are obtained directly from physical sensor measurements, its $K = 3$ prototypes are near-pure pixels of that same scene, selected at a ground-truth purity above 0.90, and ground-truth labels are defined by abundance thresholds. Finally, *OpenMIC-2018* (Humphrey et al., 2018) tests polyphonic music recorded in the wild (20 000 ten-second excerpts across $K = 20$ instruments), introducing both partial annotations and the complete absence of isolated stems.

Nothing is tuned on the test data: all four settings share the identical encoder (19), whose two integers are fixed on the prototype pool, and on OpenMIC on a validation split carved out of the training set, and initialize weights at class averages without gradient steps, using isolated sources on the first three benchmarks and polyphonic mixtures on OpenMIC. We ask whether this analytical configuration provides an effective warm start for supervised contrastive divergence, and whether its parameter-free matched filter remains competitive against computationally demanding sparse-recovery baselines. Table 1 summarizes the configurations; splits, seeds, and preprocessing details are provided in Appendix D.

Table 1: The four settings. Queries are built from sources held out from the class averages, except on Samson, whose library spectra are themselves pixels of the scene (Appendix D.5); the law error is the largest gap over the orders tested between the measurement of Appendix G and the prediction (20), on Samson the mean over pixel and material pairs, and on OpenMIC undefined (†) because the annotation is partial and the order only a lower bound. The last column is the primary score of Figure 2, at the planted weights and with no gradient step.

setting	a query is	K	features (d)	law error	score	planted
NSynth chords	sum of L recorded notes	49	log-freq. power (262)	$1.7 \cdot 10^{-3}$	micro-F1	0.481
Urban spectra	sum of L measured spectra	6	reflectance (162)	$1.7 \cdot 10^{-4}$	micro-F1	0.820
Samson pixels	a measured pixel	3	reflectance (156)	$4.5 \cdot 10^{-3}$	micro-F1	0.698
OpenMIC-2018	a 10 s excerpt of music	20	mel power (1024)	†	macro-AP	0.534

4.1 The initialisation the duality supplies

Figure 2 runs supervised CD-1 from the two initialisations, at every learning rate of a grid spanning one and a half decades, and the shape is the same in all four panels. The planted machine starts where a random one arrives only after hundreds or thousands of updates, and on the music excerpts it starts where a random one does not arrive at all within the budget we allow it. That is what the duality supplies for free: not a better optimum, but a place to begin, and the value of beginning there is the distance the other start has to travel.

On measured data, the class averages are not an exact stationary point: the gradient leaves them, and on three of the four settings it leaves them for the better. It gains most on the Samson pixels, with

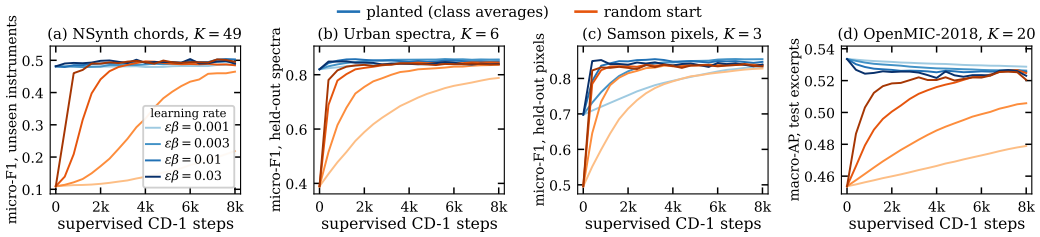


Figure 2: The first experiment: supervised CD-1 from the class averages (blue) and from random weights of the same per-entry scale (red), at four learning rates each, mean over three seeds on NSynth and OpenMIC and five on the hyperspectral panels. All runs are plotted without post-hoc selection. At step zero the two initialisations stand at 0.481 against 0.109 in *(a)*, 0.820 against 0.389 in *(b)*, 0.698 against 0.496 in *(c)* and 0.534 against 0.454 in *(d)*, where chance is 0.447; the random start reaches the planted step-zero score after $1.6 \cdot 10^3$, $1.2 \cdot 10^3$ and $4 \cdot 10^2$ steps in *(a)*–*(c)* and never in *(d)*, its closest approach over all rates and checkpoints being $6.9 \cdot 10^{-3}$ short. Over the run the gradient moves the planted machine by $+0.023$, $+0.038$, $+0.157$ and -0.005 to -0.011. The four planted curves lie within 0.025 of each other except in *(c)*, where they spread by 0.140. Scoring is on superpositions built from sources held out from the class averages, with the parameter-free largest-gap decoder, over $L \in \{1, 3, 5\}$ in *(a)* and $L \in \{1, 2, 3\}$ in *(b)* and *(c)*, and by macro average precision on the test split in *(d)*, where the annotation is partial. $N = 4096$ throughout; protocols in Appendix E.

$K = 3$ strongly overlapping materials, and least on chords, where $K = 49$ pitch classes are nearly orthogonal already; on the music excerpts, whose prototypes are averages of mixtures rather than of isolated sources, it drifts slightly the other way. Theorem 1 predicts exact stationarity where its hypotheses hold. What it does here is not a relearning of the spectra, which come back at a cosine of 0.99 or better: it rebalances the classes against each other and decorrelates the archetypes. Applying to the planted weights only the per-class gains that training produces recovers 0.013 of the 0.042 it gains on one Urban split, and a gain common to every class recovers nothing at all, the decoder being blind to it by construction (Appendix E).

The learning rate governs how fast the random start climbs, not whether it has to. Over the whole grid the random curves stay far apart at every checkpoint, while the planted ones lie almost on top of one another, the exception being Samson, where the rate matters precisely because there the planted curve is itself travelling.

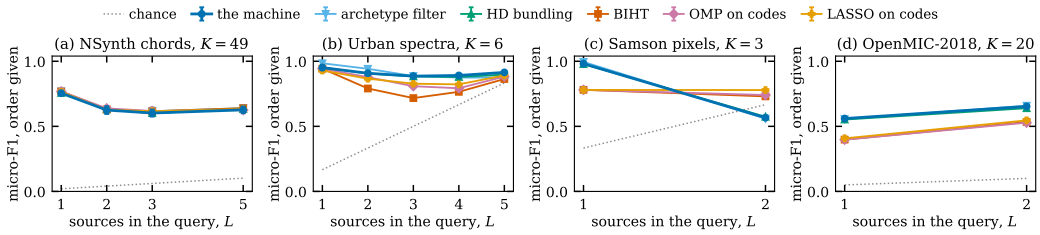


Figure 3: The second experiment: six read-outs at the same information budget, the three filters and the three reconstructions of (22). Each sees the same N-bit sign code of the same query and the same dictionary, and each is told how many sources to return. Against the best reconstruction the machine is ahead by 0.058 at $L = 3$ and 0.069 at $L = 4$ in *(b)*, by 0.201 at $L = 1$ in *(c)* and by 0.152 and 0.107 in *(d)*, and behind by 0.013 to 0.016 at every order in *(a)* and by 0.212 at $L = 2$ in *(c)*. Against the archetype filter it trails by 0.033 at $L = 1$ in *(b)* and leads by 0.039 at $L = 5$. Bars are standard deviations over the seeds, the dotted line the level of a random set of the right size; protocol, exact match and the parameter-free rule are in Appendix F.

4.2 The decoder those weights produce

The read-out of Section 3 is a matched filter on one-bit codes, so the question it has to answer is whether anything else does better with the same bits. Write $\boldsymbol{\sigma} \in \{-1, 1\}^N$ for the code of the query and $\boldsymbol{\Phi}$ for the matrix whose column μ is the projected archetype $\boldsymbol{\Omega T e}_\mu$. The six read-outs of Figure 3 fall in two families,

$$\text{filters:} \quad \tfrac{1}{N} \boldsymbol{V}^\top \boldsymbol{\sigma}, \ \ \boldsymbol{V} \in \{\boldsymbol{W}, \ \operatorname{sign} \boldsymbol{W}, \ \widehat{\boldsymbol{\Phi}}\}; \qquad \text{reconstructions:} \quad \widehat{\boldsymbol{a}} \ \text{ from } \ \boldsymbol{\sigma} \approx \operatorname{sign}(\boldsymbol{\Phi a}). \tag{22}$$

The filters keep one vector per class and correlate, and differ only in what they keep: the machine the class average $\boldsymbol{W}$ of the codes of the prototypes, hyperdimensional bundling its majority vote $\operatorname{sign} \boldsymbol{W}$ (Kanerva, 2009), the archetype filter the columns of $\boldsymbol{\Phi}$ normalised, which is the linear estimator of Plan & Vershynin (2013) and needs the prototypes before the sign, something a binary machine cannot build for itself. The reconstructions estimate the amplitudes instead: BIHT by iterative hard thresholding on the sign constraints (Jacques et al., 2013), OMP by greedy selection of a support of the right size, LASSO by an ℓ_1 penalty fixed per query. Every one sees the same $\boldsymbol{\sigma}$ and the same $\boldsymbol{\Phi}$ and is told how many sources to return, so what is compared is the ranking each produces and not the rule that turns a ranking into a set. The score is micro-F1 over all pairs of a query and a class; exact match and the parameter-free rule are in Appendix F.

The three reconstructions lose wherever the superposition is the one the theory describes: on the spectral library at every order, on the measured pixels when a single material is present, and on the music excerpts. They win in two places. On chords they are ahead by a margin the size of the seed-to-seed spread, which separates nothing. And on the pixels that two materials share they are ahead by a visible margin, for the reason Lemma 2 names: g_L is fixed by assuming equal abundances, real ones are far from equal, and a faint material then gives a smaller overlap than an absent one whose spectrum is close. Sparse recovery wins there by estimating the amplitudes our read-out assumes, and pays two orders of magnitude of arithmetic for them (Table 2).

Between our prototype and the pre-sign template the picture is a regime rather than a ranking, and on the spectral library, which reaches the deepest mixtures, what orders it is the depth: the template is the better match for a single source and the worse one as the order grows, and the two curves cross. Keeping the average of the codes keeps the spread of the class inside the prototype, and that spread is what a deep mixture needs. The crossing is a statement about the ranking with the order given and does not survive the parameter-free rule.

The experiments also fix the size of the departure from the theory. Each of the three hypotheses Section 3 gives up leaves a trace we can put a number on: an effective count of independent sites far below N, measured in Appendix D, and, in Appendix G, a read-out that leaves the band (21) where the abundances are furthest from equal and a balance point of the learning rule below the predicted amplitude. Appendix G.4 evaluates the bound at these parameters and finds it vacuous there.

5 Conclusion

In this work, we showed that the Hopfield-RBM duality naturally recasts spurious mixtures into a multi-label classification problem. By fixing the hidden unit representations at scale $\sqrt{N}$ with an amplitude matching the archetype-mixture overlap, the empirical class averages of noisy prototypes form an exact fixed point in mean of supervised CD-1, with an analytically bounded drift that vanishes along four distinct channels: thermal sampling, dataset noise, mini-batch consensus, and finite-size fluctuations. When extending this framework to continuous physical mixtures via one-bit random projections, the overlap amplitude transitions to an arcsine law. Across multiple empirical benchmarks, initializing the network at these class averages provides an effective warm start, consistently saving hundreds to thousands of contrastive divergence steps and yielding a parameter-free matched filter that rivals iterative compressed sensing baselines at a fraction of the computational cost.

Several natural directions emerge from these findings. First, the theoretical dataset channel should be reformulated using an effective dimension that reflects the spatial and spectral correlation structure of real-world signals rather than the nominal input size N. Second, the non-asymptotic drift bounds could be tightened by leveraging the empirical variance of the estimators rather than worst-case range inequalities. Finally, relaxing the equal-energy assumption by adaptively inferring non-uniform label amplitudes directly from the data represents a promising avenue to enhance robust multi-label decoding in real-world scenes.

Reproducibility statement

Every object entering the theorem is defined in Section 2 and every step of the proof in Appendices A and C. The datasets are public, and Appendices B and D to G give every parameter, protocol and split.

Declaration of Generative AI and AI-Assisted Technologies

During the preparation of this work, the author(s) used **Claude Opus 5 (Anthropic)** in order to improve the language, clarity, and readability of the manuscript. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985. doi: 10.1207/s15516709cog0901_7.

Elena Agliari, Adriano Barra, Andrea Galluzzi, Francesco Guerra, and Francesco Moauro. Multitasking associative networks. *Physical Review Letters*, 109(26):268101, 2012. doi: 10.1103/PhysRevLett.109.268101.

Elena Agliari, Andrea Alessandrelli, Paulo Duarte Mourao, and Alberto Fachechi. Multi-channel pattern reconstruction through l-directional associative memories. *arXiv preprint arXiv:2503.06274*, 2025.

Elena Agliari, Andrea Alessandrelli, Adriano Barra, Martino Centonze, and Federico Ricci-Tersenghi. Networks of hebbian networks: more is different. *Neural Networks*, 194:108181, 2026. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2025.108181.

Francesco Alemanno, Miriam Aquaro, Ido Kanter, Adriano Barra, and Elena Agliari. Supervised hebbian learning. *Europhysics Letters*, 141(1):11001, jan 2023a. doi: 10.1209/0295-5075/aca55f.

Francesco Alemanno, Luca Camanzi, Gianluca Manzan, and Daniele Tantari. Hopfield model with planted patterns: A teacher-student self-supervised learning model. *Applied Mathematics and Computation*, 458:128253, 2023b. ISSN 0096-3003. doi: https://doi.org/10.1016/j.amc.2023.128253.

D. J. Amit, H. Gutfreund, and H. Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, 55:1530–1533, 1985a. doi: 10.1103/PhysRevLett.55.1530.

Daniel J. Amit, Hanoch Gutfreund, and Haim Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007–1018, 1985b.

Adriano Barra, Alberto Bernacchia, Enrica Santucci, and Pierluigi Contucci. On the equivalence of hopfield networks and boltzmann machines. *Neural Networks*, 34:1–9, 2012. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2012.06.003.

Petros T. Boufounos and Richard G. Baraniuk. 1-bit compressive sensing. In *42nd Annual Conference on Information Sciences and Systems*, pp. 16–21, 2008. doi: 10.1109/CISS.2008.4558487.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, pp. 380–388, 2002. doi: 10.1145/509907.509965.

A C C Coolen, R Kühn, and P Sollich. *Theory of Neural Information Processing Systems*. Oxford University Press, 07 2005.

Aurélien Decelle and Cyril Furtlehner. Restricted Boltzmann machine: Recent advances and mean-field theory. *Chinese Physics B*, 30(4):040202, 2021. doi: 10.1088/1674-1056/abd160.

Giovanni Di Sarra, Barbara Bravi, and Yasser Roudi. The unbearable lightness of restricted Boltzmann machines: Theoretical insights and biological applications. *Europhysics Letters*, 149(2):21002, 2025. doi: 10.1209/0295-5075/ada636.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with WaveNet autoencoders. In *International Conference on Machine Learning*, pp. 1068–1077, 2017.

Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6):1115–1145, 1995. doi: 10.1145/227683.227684.

Donald O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York, 1949.

John Hertz, Anders Krogh, and Richard G. Palmer. *Introduction to the Theory of Neural Computation*. Santa Fe Institute Studies in the Sciences of Complexity. Addison-Wesley, Redwood City, CA, 1991.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554.

Eric J. Humphrey, Simon Durand, and Brian McFee. OpenMIC-2018: An open dataset for multiple instrument recognition. In *International Society for Music Information Retrieval Conference*, pp. 438–444, 2018. doi: 10.5281/zenodo.1432913.

Laurent Jacques, Jason N. Laska, Petros T. Boufounos, and Richard G. Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Transactions on Information Theory*, 59(4):2082–2102, 2013. doi: 10.1109/TIT.2012.2234823.

Pentti Kanerva. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation*, 1(2):139–159, 2009. doi: 10.1007/s12559-009-9009-8.

Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Francesca Elisa Leonelli, Elena Agliari, Linda Albanese, and Adriano Barra. On the effective initialisation for restricted boltzmann machines via duality with hopfield model. *Neural Networks*, 143:314–326, 2021. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2021.06.017.

Marc Mézard. Mean-field message-passing equations in the Hopfield model and its generalizations. *Physical Review E*, 95:022117, 2017. doi: 10.1103/PhysRevE.95.022117.

Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2013. doi: 10.1109/TIT.2012.2207945.

Tony A. Plate. Holographic reduced representations. *IEEE Transactions on Neural Networks*, 6(3): 623–641, 1995. doi: 10.1109/72.377968.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

Matthew Smart and Anton Zilman. On the mapping between Hopfield networks and restricted Boltzmann machines. In *International Conference on Learning Representations*, 2021.

Jérôme Tubiana and Rémi Monasson. Emergence of compositional representations in restricted Boltzmann machines. *Physical Review Letters*, 118:138301, 2017. doi: 10.1103/PhysRevLett.118.138301.

Feiyun Zhu, Ying Wang, Shiming Xiang, Bin Fan, and Chunhong Pan. Structured sparse method for hyperspectral unmixing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 88:101–118, 2014. doi: 10.1016/j.isprsjprs.2013.11.014.

Algorithm 1: Supervised CD-1 step

Input: training set $\mathcal{S}$ of eq. (11); examples $\{\eta^{\mu,m}\}$; inverse temperature β; mini-batch size B odd; learning rate ϵ with $\epsilon\beta \ll 1$
Output: weight update Δw_i^μ

```
/* initialisation                                                    */
```
$w_i^\mu \leftarrow \bar\eta_i^\mu = \frac{1}{rM}\sum_m \eta_i^{\mu,m}$ `// or` $w_i^\mu \sim \mathcal{N}(0,1)$
Draw an item from $\mathcal{S}$ and read its label tuple $\hat{\boldsymbol{\mu}}$
Draw a mini-batch $\mathcal{B}$ of B items of $\mathcal{S}$ bearing $\hat{\boldsymbol{\mu}}$
```
/* positive phase:  both layers clamped to the datum                 */
```
$(\sigma_i)' \leftarrow s_i(\mathcal{B})$ `// cue, eq. (15)`
$\hat z_\mu \leftarrow \sqrt{N}\, c_L \mathbf{1}\{\mu \in \hat{\boldsymbol{\mu}}\}$ `// label code, eq. (13)`
```
/* reconstruction, P(σ | z)                                          */
```
$h_i \leftarrow \frac{1}{\sqrt N}\sum_\mu w_i^\mu \hat z_\mu = c_L \sum_{\mu\in\hat{\boldsymbol{\mu}}} w_i^\mu$ `// only the` L `labelled units contribute`
$(\sigma_i)'' \leftarrow \operatorname{sign}\big([1+e^{-2\beta h_i}]^{-1} - \omega_i\big),\ \omega_i \sim \mathcal{U}(0,1)$
```
/* negative phase, P(z | σ)                                          */
```
$(z_\mu)'' \leftarrow u_\mu(\boldsymbol{\sigma}'') + \zeta_\mu,\ \zeta_\mu \sim \mathcal{N}(0,\beta^{-1})$
```
/* update                                                            */
```
$\Delta w_i^\mu \leftarrow \frac{\epsilon\beta}{\sqrt N}\big[(\sigma_i)'\hat z_\mu - (\sigma_i)''(z_\mu)''\big]$

A PROOFS

This appendix proves Theorem 1. It states the supervised CD-1 step as an algorithm, collects the auxiliary results the proof needs, each isolating one of the four sources of drift, and assembles them. It closes with the remark that averaging over the archetypes rather than conditioning on them replaces the finite-size term of (17) by $N^{-1/2}$.

Throughout, $\mathbb{E}_{\rm alg}$, $\mathbb{E}_{\mathcal{S}}$ and $\mathbb{E} = \mathbb{E}_{\mathcal{S}}\mathbb{E}_{\rm alg}$ are the averages introduced in Section 2, and $\mathbb{P}_{\boldsymbol{\xi}}$ is the law of the archetypes.

Since Algorithm 1 first draws an item and then acts on its tuple only, we condition on the tuple $\hat{\boldsymbol{\mu}}$ throughout and average over it at the very end; to lighten the notation we drop the item index and write A for the set of its L labels. We abbreviate

$$\Sigma_i := \sum_{\mu\in A} \xi_i^\mu, \qquad \Delta_i^\mu := \bar\eta_i^\mu - \xi_i^\mu, \tag{23}$$

and we set $\hat g_\mu := \hat z_\mu/\sqrt N = c_L \mathbf{1}\{\mu \in A\}$, so that $s_i = \operatorname{sign}(\Sigma_i)$ and, L being odd, $|\Sigma_i| \geq 1$ deterministically. At the planted weights the reconstruction field, the reconstructed magnetisation and the reconstruction error are

$$h_i := c_L \sum_{\mu\in A} \bar\eta_i^\mu, \qquad t_i := \tanh(\beta h_i), \qquad T_\mu := \frac{1}{N}\sum_{j=1}^N \bar\eta_j^\mu t_j, \qquad e_i := t_i - s_i. \tag{24}$$

We work with the prefactor-free update

$$\Delta y_i^\mu := \frac{\Delta w_i^\mu}{\epsilon\beta} = \frac{1}{\sqrt N}\Big[(\sigma_i)'\,\hat z_\mu - (\sigma_i)''(z_\mu)''\Big], \tag{25}$$

in terms of which Theorem 1 concerns $\mathbb{E}[\Delta y_i^\mu]$. Note that $|\bar\eta_i^\mu| \leq r^{-1}$ deterministically, since $\bar\eta_i^\mu$ is an average of M signs divided by r, and that $\mathbb{E}_{\mathcal{S}}[\Delta_i^\mu] = 0$.

A.1 AUXILIARY RESULTS

Lemma 3 (Hoeffding)**.** *Let $X_1,\dots,X_n$ be independent with $\mathbb{P}(a_k \leq X_k \leq b_k) = 1$ for every k, and let $S = \sum_k X_k$. Then, for every $t > 0$, $\mathbb{P}\big(|S - \mathbb{E}[S]| \geq t\big) \leq 2\exp\big(-2t^2/\sum_k (b_k - a_k)^2\big)$.*

Proof of Lemma 1. Fix $\mu \in A$ and condition on $\xi_i^\mu = +1$; by the flip symmetry $\mathbb{E}[\xi^\mu s] = \mathbb{E}[\operatorname{sign}(1+V)]$ with $V = \sum_{\nu\in A,\nu\neq\mu}\xi^\nu$ a sum of $L-1$ independent Rademacher variables. For L odd, $L-1$

is even, so V is even and $\mathbb{E}[\mathrm{sign}(1+V)] = \mathbb{P}(V \geq 0) - \mathbb{P}(V \leq -2) = \mathbb{P}(V = 0) = c_L$ by the symmetry $V \mapsto -V$. For $\mu \notin A$ the archetype $\boldsymbol{\xi}^\mu$ is independent of $\boldsymbol{s}$ and centred. The convergence of the Mattis magnetisation is the law of large numbers applied to the independent bounded variables $\xi_j^\mu s_j$, and the energy density follows from (1), since the $K - L$ silent overlaps contribute $O(N^{-1})$. Finally $c_L = \binom{L-1}{(L-1)/2} 2^{-(L-1)} \simeq \sqrt{2/(\pi L)}$ by Stirling, so $L c_L^2 \to 2/\pi$. □

Lemma 4 (Class average)**.** *Let $\Delta_i^\mu = \bar{\eta}_i^\mu - \xi_i^\mu$. Then $\mathbb{E}_{\mathcal{S}}[\Delta_i^\mu] = 0$, $\mathbb{E}_{\mathcal{S}}[(\Delta_i^\mu)^2] = \rho$, $|\Delta_i^\mu| \leq 1 + r^{-1}$, and for every $\tau > 0$*

$$\mathbb{P}\big(|\Delta_i^\mu| \geq \tau\big) \leq 2\exp\Big(-\frac{(1-r^2)\tau^2}{2\rho}\Big) = 2\exp\Big(-\frac{M r^2 \tau^2}{2}\Big). \tag{26}$$

Proof. Fix μ and i and condition on ξ_i^μ. The variables $X_m := (\eta_i^{\mu,m} - r\xi_i^\mu)/r$ are independent and centred, since $\mathbb{E}_{\mathcal{S}}[\eta_i^{\mu,m}] = r\xi_i^\mu$. Each takes two values: $\eta_i^{\mu,m} = \xi_i^\mu$ with probability $\frac{1+r}{2}$, giving $X_m = r^{-1}(1-r)\xi_i^\mu$, and $\eta_i^{\mu,m} = -\xi_i^\mu$ with probability $\frac{1-r}{2}$, giving $X_m = -r^{-1}(1+r)\xi_i^\mu$. The two values are separated by $2/r$ irrespective of the sign of ξ_i^μ, so $b_m - a_m = 2/r$ and $\sum_m (b_m - a_m)^2 = 4M/r^2$. Since $M\Delta_i^\mu = \sum_m X_m$, Lemma 3 with $t = M\tau$ gives $\mathbb{P}(|\Delta_i^\mu| \geq \tau) \leq 2\exp(-2M^2\tau^2 r^2/4M)$, which is (26) once $Mr^2 = (1-r^2)/\rho$ is used. The variance is $\mathrm{Var}[\eta_i^{\mu,m}]/(r^2 M) = (1-r^2)/(r^2 M) = \rho$, and the deterministic bound follows from $|\bar{\eta}_i^\mu| \leq r^{-1}$. □

The next lemma is the counterpart, for a composite query, of the statement that a single noisy example of an archetype reproduces it with reliability r. It is where the mixture order first enters.

Lemma 5 (Single query)**.** *Let L be odd and let $\hat{\eta}_i^p$ be a query (12) of the tuple A. Then*

$$\mathbb{E}_{\mathcal{S}}\big[\hat{\eta}_i^p \mid \boldsymbol{\xi}\big] = s_i\, \vartheta(|\Sigma_i|), \tag{27}$$

where $\vartheta(k) := \mathbb{E}[\mathrm{sign}(T) \mid \Sigma_i = k]$, with $T := \sum_{\mu \in A} \xi_i^\mu \chi_i^{\mu,p}$, is non-decreasing along the odd integers $1 \leq k \leq L$ and satisfies

$$\vartheta(1) = \vartheta_L(r) = r\,\mathbb{P}(V = V'), \tag{28}$$

with V, V' independent $\mathrm{Bin}\big(\frac{L-1}{2}, q\big)$ variables and $q := \frac{1-r}{2}$. In particular $\vartheta_1(r) = r$ and $\vartheta(k) \geq \vartheta_L(r) > 0$ for every odd $k \geq 1$.

Proof. Since $\eta_i^{\mu,p} = \xi_i^\mu \chi_i^{\mu,p}$, one has $\hat{\eta}_i^p = \mathrm{sign}(T)$ with $T = \sum_{\mu \in A} \xi_i^\mu \chi_i^{\mu,p}$ a sum of L independent ± 1 variables: the $n_+ := (L + \Sigma_i)/2$ summands attached to $\xi_i^\mu = +1$ have mean r, the $n_- := (L - \Sigma_i)/2$ attached to $\xi_i^\mu = -1$ have mean $-r$. As L is odd, $T \neq 0$ almost surely and $\mathbb{E}[\mathrm{sign}(T)]$ is well defined. Flipping every ξ_i^μ, $\mu \in A$, maps Σ_i into $-\Sigma_i$ and T into $-T$, whence (27).

For the monotonicity, write $\chi = 1 - 2u$ with $u \sim \mathrm{Bernoulli}(q)$, so that

$$T = \Sigma_i - 2D, \qquad D := U_+ - U_-, \tag{29}$$

with $U_+ \sim \mathrm{Bin}(n_+, q)$ and $U_- \sim \mathrm{Bin}(n_-, q)$ independent. Increasing k by two turns one summand of mean $-r$ into one of mean $+r$, which increases T stochastically; hence ϑ is non-decreasing.

For the value at $k = 1$ we have $n_+ = \frac{L+1}{2} = n_- + 1$ and $n_- = \frac{L-1}{2}$. By (29), $T > 0$ if and only if $D \leq 0$ and $T < 0$ if and only if $D \geq 1$, so that $\vartheta(1) = 2\mathbb{P}(D \leq 0) - 1$. Split $U_+ = V + u$ with $V \sim \mathrm{Bin}(n_-, q)$ and $u \sim \mathrm{Bernoulli}(q)$ independent, and set $D' := V - U_-$, which is symmetric because V and U_- are independent and identically distributed. Writing $p_0 := \mathbb{P}(D' = 0)$ one has $\mathbb{P}(D' \leq 0) = \frac{1+p_0}{2}$ and $\mathbb{P}(D' \leq -1) = \frac{1-p_0}{2}$, whence

$$\mathbb{P}(D \leq 0) = (1-q)\mathbb{P}(D' \leq 0) + q\,\mathbb{P}(D' \leq -1) = \frac{1 + (1-2q)p_0}{2}, \tag{30}$$

and therefore $\vartheta(1) = (1-2q)p_0 = r\,\mathbb{P}(V = V')$, which is (28). For $L = 1$ the two binomials are degenerate and $p_0 = 1$, so $\vartheta_1(r) = r$. Positivity is clear since $r > 0$ and $p_0 > 0$. □

Lemma 6 (Majority vote)**.** *Let B be odd and let $\mathcal{B}$ be a mini-batch of B items of $\mathcal{S}$ bearing the tuple A, so that the cue (15) is the majority vote of B independent queries. Then*

$$\mathbb{E}_{\mathcal{S}}\big[s_i(\mathcal{B}) \mid \boldsymbol{\xi}\big] = s_i\, \vartheta_B(|\Sigma_i|), \tag{31}$$

with $\vartheta_B(k) := 2\mathbb{P}\big(\mathrm{Bin}(B, \frac{1+\vartheta(k)}{2}) > \frac{B}{2}\big) - 1$. Moreover ϑ_B is non-decreasing along the odd values of B and

$$0 \;\le\; 1 - \vartheta_B(k) \;\le\; 2e^{-B\vartheta(k)^2/2} \;\le\; 2e^{-B\vartheta_L(r)^2/2}. \tag{32}$$

Proof. The items of $\mathcal{S}$ are built from independent examples, so conditionally on the archetypes the B queries in $\mathcal{B}$ are independent and identically distributed, with $\mathbb{P}(\hat{\eta}_i^p = s_i) = \frac{1+\vartheta(|\Sigma_i|)}{2}$ by Lemma 5. Let $W := \#\{p \in \mathcal{B} : \hat{\eta}_i^p = s_i\} \sim \mathrm{Bin}(B, \frac{1+\vartheta}{2})$; since B is odd, $\sum_p \hat{\eta}_i^p = s_i(2W - B) \neq 0$ and the vote equals s_i if and only if $W > B/2$, which gives (31).

For (32), note that $\mathbb{E}[W] = B\frac{1+\vartheta}{2}$, so that $\{W < B/2\} \subseteq \{W - \mathbb{E}[W] \le -B\vartheta/2\}$. The one-sided form of Lemma 3, applied to the sum of B independent Bernoulli variables of unit range, gives $\mathbb{P}(W - \mathbb{E}[W] \le -t) \le e^{-2t^2/B}$, and with $t = B\vartheta/2$ this is $e^{-B\vartheta^2/2}$. Since $1 - \vartheta_B = 2\mathbb{P}(W < B/2)$, the claim follows, the last inequality being Lemma 5. Non-negativity is clear from $\vartheta_B \le 1$. The monotonicity in B is the standard fact that adding two further votes changes the outcome only when the tally is ± 1, and does so in favour of the majority because $\frac{1+\vartheta}{2} > \frac{1-\vartheta}{2}$. □

It is convenient to record the empirical counterpart of (31). Averaging the cue over the choice of the mini-batch at *fixed* data defines

$$\bar{s}_i := \mathbb{E}_{\mathrm{alg}}\big[s_i(\mathcal{B})\big], \tag{33}$$

a deterministic function of the data obeying $|\bar{s}_i| \le 1$ and, by Lemma 6 and the tower property, $\mathbb{E}_{\mathcal{S}}[\bar{s}_i] = s_i\vartheta_B(|\Sigma_i|)$. We stress that $\bar{s}_i$ and $\bar{\eta}_i^\mu$ are built from two disjoint pools of examples and are therefore independent, which is what the two-pool construction of Section 2 buys.

Lemma 7 (One-step reduction)**.** *For any weights $\boldsymbol{w}$, any realised data and every i, μ,*

$$\mathbb{E}_{\mathrm{alg}}\big[\Delta y_i^\mu\big] = \bar{s}_i\, \hat{g}_\mu - t_i T_\mu - \frac{w_i^\mu\big(1 - t_i^2\big)}{N}, \tag{34}$$

with $\bar{s}_i$ the empirical majority-vote mean (33).

Proof. Consider the two phases of (25) in turn. The label code is deterministic and $\hat{z}_\mu/\sqrt{N} = \hat{g}_\mu$, so the first term is linear in the presented cue and averaging over the mini-batch gives $\mathbb{E}_{\mathrm{alg}}[(\sigma_i)']\,\hat{g}_\mu = \bar{s}_i\hat{g}_\mu$. The field h_i that drives the reconstruction is built from the label and the weights only, so it does not depend on which mini-batch was drawn, and the two phases decouple under $\mathbb{E}_{\mathrm{alg}}$.

For the negative phase, insert $(z_\mu)'' = N^{-1/2}\sum_j w_j^\mu(\sigma_j)'' + \zeta_\mu$ and collect the two factors $N^{-1/2}$,

$$\frac{(\sigma_i)''(z_\mu)''}{\sqrt{N}} = \frac{1}{N}\sum_{j=1}^{N} w_j^\mu(\sigma_i)''(\sigma_j)'' + \frac{(\sigma_i)''\zeta_\mu}{\sqrt{N}}. \tag{35}$$

In the Gibbs chain the noise ζ is drawn after the visible layer has been reconstructed and independently of it; being also centred, it gives $\mathbb{E}_{\mathrm{alg}}[(\sigma_i)''\zeta_\mu] = 0$. As for the first term, once the data are fixed the field h_i is deterministic and the second law of (8) factorises over the sites, so the reconstructed spins are conditionally independent with $\mathbb{E}_{\mathrm{alg}}[(\sigma_i)''] = t_i$, whence $\mathbb{E}_{\mathrm{alg}}[(\sigma_i)''(\sigma_j)''] = t_i t_j + [1 - t_i^2]\delta_{ij}$, the correction being needed only for the single term $j = i$, where $[(\sigma_i)'']^2 = 1$. Substituting gives (34). □

The next lemma controls the sign of the local field, and hence the quality of the reconstruction. It is stated for a one-parameter family of events: the parameter θ fixes how the worst-case signal is split between the noise of the data and the margin left to the thermal fluctuations.

Lemma 8 (Margin)**.** *Let $\theta \in (0, 1)$ and define*

$$\tau_\theta := \frac{\theta}{L}, \qquad \delta_\theta := c_L(1 - \theta), \qquad \mathcal{E}_i := \Big\{ \max_{\mu \in A} |\Delta_i^\mu| \le \tau_\theta \Big\}. \tag{36}$$

Then, on the event $\mathcal{E}_i$ and at the planted weights, $s_i h_i \geq \delta_\theta > 0$, so that $\operatorname{sign}(h_i) = s_i$ and $|h_i| \geq \delta_\theta$. Moreover

$$\pi_\theta := \mathbb{P}(\mathcal{E}_i^c) \leq 2L\, e^{-c_1(\theta)/\rho}, \qquad c_1(\theta) := \frac{(1-r^2)\theta^2}{2L^2}. \tag{37}$$

Proof. At $\boldsymbol{w} = \bar{\boldsymbol{\eta}}$ the field of (24) is a combination of the planted weights of the L presented archetypes only, $h_i = c_L \sum_{\mu\in A}(\xi_i^\mu + \Delta_i^\mu) = c_L(\Sigma_i + \sum_{\mu\in A}\Delta_i^\mu)$. Multiplying by $s_i \in \{\pm 1\}$ and using $s_i\Sigma_i = |\Sigma_i|$,

$$s_i h_i = c_L \underbrace{|\Sigma_i|}_{\textit{signal}} + c_L\, s_i \underbrace{\sum_{\mu\in A} \Delta_i^\mu}_{\textit{data noise}}. \tag{38}$$

Since L is odd, $|\Sigma_i|$ is an odd integer and is therefore at least one; the worst case $|\Sigma_i| = 1$ is attained on the sites where the constituents split as evenly as they can, which is a positive fraction of them, so the signal in (38) cannot be improved beyond c_L. By the triangle inequality the second term is bounded in modulus by $c_L L \max_{\mu\in A}|\Delta_i^\mu|$, and on $\mathcal{E}_i$ this is at most $c_L L\tau_\theta = c_L\theta$, that is a fraction θ of the worst-case signal. Combining, $s_i h_i \geq c_L(1-\theta) = \delta_\theta$. Note that c_L multiplies signal and noise alike, so it cancels in the threshold τ_θ and survives only in the margin δ_θ. For (37), a union bound over the L presented archetypes and Lemma 4 give $\mathbb{P}(\mathcal{E}_i^c) \leq \sum_{\mu\in A}\mathbb{P}(|\Delta_i^\mu| > \tau_\theta) \leq 2L\exp(-(1-r^2)\tau_\theta^2/2\rho)$, and substituting $\tau_\theta = \theta/L$ identifies the exponent with $c_1(\theta)/\rho$. □

Lemma 9 (Reconstruction error)**.** *With $e_i = t_i - s_i$ one has $|e_i| \leq 2$ always, and*

$$\mathbb{E}_{\mathcal{S}}\big[|e_i|\big] \leq \varepsilon_\beta(\theta) + 2\pi_\theta, \qquad \varepsilon_\beta(\theta) := 2e^{-2\beta\delta_\theta}, \tag{39}$$

with δ_θ and π_θ as in Lemma 8. Moreover $\mathbb{P}\big((\sigma_i)'' \neq s_i\big) \leq e^{-2\beta\delta_\theta} + \pi_\theta$.

Proof. On $\mathcal{E}_i$, Lemma 8 guarantees that the field points in the direction of s_i and that its modulus is at least δ_θ; since $\operatorname{sign}(h_i) = s_i$ there, we may write

$$|e_i| = \big|\tanh(\beta h_i) - \operatorname{sign}(h_i)\big| = \frac{2}{e^{2\beta|h_i|}+1} \leq 2e^{-2\beta|h_i|} \leq \varepsilon_\beta(\theta), \tag{40}$$

where we used $1 - \tanh x = 2/(e^{2x}+1) \leq 2e^{-2x}$ and then $|h_i| \geq \delta_\theta$. On the complementary event we only use the deterministic bound $|e_i| \leq |t_i| + |s_i| \leq 2$. Splitting the expectation accordingly gives $\mathbb{E}_{\mathcal{S}}|e_i| \leq \varepsilon_\beta(\theta)\mathbb{P}(\mathcal{E}_i) + 2\mathbb{P}(\mathcal{E}_i^c)$, which is (39). For the last assertion, the logistic rule in (8) gives $\mathbb{P}\big((\sigma_i)'' \neq \operatorname{sign}(h_i) \mid h_i\big) = (1+e^{2\beta|h_i|})^{-1} \leq e^{-2\beta\delta_\theta}$ on $\mathcal{E}_i$; adding $\mathbb{P}(\mathcal{E}_i^c) \leq \pi_\theta$ concludes. □

Lemma 10 (Overlaps)**.** *Let $Q^\mu := N^{-1}\sum_j \bar{\eta}_j^\mu s_j$ and $c^\mu := N^{-1}\sum_j \xi_j^\mu s_j$. Then $|Q^\mu| \leq r^{-1}$ almost surely and $\mathbb{E}_{\mathcal{S}}[Q^\mu] = c^\mu$, while, over the archetypes, for every $t > 0$*

$$\mathbb{P}_{\boldsymbol{\xi}}\Big(\max_{\mu\leq K}\big|c^\mu - c_L\mathbf{1}\{\mu\in A\}\big| \geq t\Big) \leq 2K\, e^{-Nt^2/2}, \tag{41}$$

and consequently $\mathbb{E}_{\boldsymbol{\xi}}\big|c^\mu - c_L\mathbf{1}\{\mu\in A\}\big| \leq N^{-1/2}$.

Proof. The deterministic bound follows from $|\bar{\eta}_j^\mu| \leq r^{-1}$ and $|s_j| = 1$, and the mean from $\mathbb{E}_{\mathcal{S}}[\bar{\eta}_j^\mu] = \xi_j^\mu$, the two pools being independent. For (41), fix μ: the variables $\xi_j^\mu s_j$ are independent across j, since s_j depends on the archetypes only through site j, they take values in $[-1,1]$, and by Lemma 1 their common mean is $c_L\mathbf{1}\{\mu\in A\}$. Lemma 3 applied to $Nc^\mu = \sum_j \xi_j^\mu s_j$ with $b_j - a_j = 2$ and $t \mapsto Nt$ gives $\mathbb{P}(|c^\mu - c_L\mathbf{1}\{\mu\in A\}| \geq t) \leq 2e^{-2N^2t^2/(4N)} = 2e^{-Nt^2/2}$, and the uniform bound follows from a union bound over the K hidden units. The last claim is Jensen's inequality, the variance of an average of N independent variables bounded by one being at most N^{-1}. □

Remark 3. The $N^{-1/2}$ of Lemma 10 is not an artefact of the algorithm: it is the residual overlap that distinct archetypes have at finite N, which the negative phase inevitably sees. It is the only contribution to Theorem 1 that does not vanish exponentially, and it survives both $\beta \to \infty$ and $M \to \infty$. Being a random function of the quenched disorder, it must be kept explicit in the bound and converted into a statement about $\boldsymbol{\xi}$ only at the very end, which is why Theorem 1 is formulated with a confidence level γ.

A.2 Proof of the Theorem and of the Corollary

Proof of Theorem 1. Fix $\theta \in (0,1)$, set $\boldsymbol{w} = \bar{\boldsymbol{\eta}}$ and condition on the tuple A drawn from $\mathcal{Q}$. Taking the average $\mathbb{E}_{\mathcal{S}}$ of the exact identity (34),

$$\mathbb{E}\big[\Delta y_i^{\mu} \mid A\big] = \underbrace{\mathbb{E}_{\mathcal{S}}[\bar{s}_i]\,\hat{g}_{\mu}}_{=:P} - \underbrace{\mathbb{E}_{\mathcal{S}}\big[t_i T_{\mu}\big]}_{=:T} - \underbrace{\frac{1}{N}\mathbb{E}_{\mathcal{S}}\big[\bar{\eta}_i^{\mu}(1-t_i^2)\big]}_{=:D}. \tag{42}$$

Step 1: the positive phase. By Lemma 6 and (33), $\mathbb{E}_{\mathcal{S}}[\bar{s}_i] = s_i \vartheta_B(|\Sigma_i|)$, so that

$$P = \vartheta_B(|\Sigma_i|)\, s_i\, c_L \mathbf{1}\{\mu \in A\}. \tag{43}$$

The majority vote is therefore what the positive phase contributes: an image of the clean spurious state, faithful up to the factor ϑ_B of (32). Had the visible layer been clamped to a single query, this factor would have been $\vartheta_L(r)$, and no choice of the mini-batch size could have removed it.

Step 2: the diagonal correction is negligible. Writing $t_i = s_i + e_i$ and using $s_i^2 = 1$ gives $1 - t_i^2 = -2s_i e_i - e_i^2$, so that $|1 - t_i^2| \le 2|e_i| + |e_i|^2 \le 4|e_i|$ by $|e_i| \le 2$. Since $|\bar{\eta}_i^{\mu}| \le r^{-1}$, Lemma 9 gives

$$|D| \;\le\; \frac{4}{rN}\,\mathbb{E}_{\mathcal{S}}|e_i| \;\le\; \frac{4\big(\varepsilon_{\beta}(\theta) + 2\pi_{\theta}\big)}{rN}. \tag{44}$$

Step 3: expanding T, and the cancellation. Substituting $t_i = s_i + e_i$ into (24) gives $T_{\mu} = Q^{\mu} + N^{-1}\sum_j \bar{\eta}_j^{\mu} e_j$, and expanding the product $t_i T_{\mu}$ produces four contributions,

$$t_i T_{\mu} = \underbrace{s_i Q^{\mu}}_{\text{(I)}} + \underbrace{\frac{s_i}{N}\sum_j \bar{\eta}_j^{\mu} e_j}_{\text{(II)}} + \underbrace{e_i Q^{\mu}}_{\text{(III)}} + \underbrace{\frac{e_i}{N}\sum_j \bar{\eta}_j^{\mu} e_j}_{\text{(IV)}}, \tag{45}$$

which we treat separately.

(I), *the leading term.* By Lemma 10, $\mathbb{E}_{\mathcal{S}}[(\mathrm{I})] = s_i c^{\mu}$. Subtracting (43), adding and removing $s_i c_L \mathbf{1}\{\mu \in A\}$ and using $c_L \le 1$,

$$\big|P - \mathbb{E}_{\mathcal{S}}[(\mathrm{I})]\big| = \big|s_i\big(\vartheta_B c_L \mathbf{1}\{\mu \in A\} - c^{\mu}\big)\big| \le\; c_L\big(1 - \vartheta_B\big) + \max_{\nu \le K}\big|c^{\nu} - c_L \mathbf{1}\{\nu \in A\}\big|. \tag{46}$$

This is the cancellation the theorem rests on, and it is worth reading it once more in words: the label code asks the hidden layer for the value $\sqrt{N}c_L$ on the constituents, and the machine, having reconstructed the mixture on its visible layer, returns exactly that, the weights being centred on the archetypes and not on a shrunken copy of them. The two agree once multiplied by the visible activity, which the majority vote supplies with amplitude ϑ_B: the residual mismatch $1 - \vartheta_B$ is the price of the fact that a cue is a binary vector, and it is the mini-batch channel of (17). This is also where the two ingredients of the protocol are seen to be inseparable. With biased weights both sides would have carried a factor r and would have cancelled; with unbiased weights the positive phase must be cleaned as well, and the majority vote is what cleans it.

(II). Using $|s_i| = 1$, $|\bar{\eta}_j^{\mu}| \le r^{-1}$ and Lemma 9 at each site, $\big|\mathbb{E}_{\mathcal{S}}[(\mathrm{II})]\big| \le (rN)^{-1}\sum_j \mathbb{E}_{\mathcal{S}}|e_j| \le r^{-1}(\varepsilon_{\beta}(\theta) + 2\pi_{\theta})$, since $\mathbb{E}_{\mathcal{S}}|e_j| \le \varepsilon_{\beta}(\theta) + 2\pi_{\theta}$ uniformly in j and the N terms are divided by N.

(III). Here e_i does not depend on the summation index, and the deterministic bound $|Q^{\mu}| \le r^{-1}$ of Lemma 10 gives $\big|\mathbb{E}_{\mathcal{S}}[(\mathrm{III})]\big| \le r^{-1}\mathbb{E}_{\mathcal{S}}|e_i| \le r^{-1}(\varepsilon_{\beta}(\theta) + 2\pi_{\theta})$.

(IV). Bounding one factor by $|e_i| \le 2$ and averaging the other, $\big|\mathbb{E}_{\mathcal{S}}[(\mathrm{IV})]\big| \le 2(rN)^{-1}\sum_j \mathbb{E}_{\mathcal{S}}|e_j| \le 2r^{-1}(\varepsilon_{\beta}(\theta) + 2\pi_{\theta})$.

Step 4: collecting the estimates. Inserting the four contributions and (44) into (42), the leading terms cancel by (46) and we are left, for $N \ge 4$ so that $4/N \le 1$, with

$$\big|\mathbb{E}[\Delta y_i^{\mu} \mid A]\big| \le c_L\big(1 - \vartheta_B\big) + \max_{\nu \le K}\big|c^{\nu} - c_L \mathbf{1}\{\nu \in A\}\big| + \frac{5}{r}\big(\varepsilon_{\beta}(\theta) + 2\pi_{\theta}\big). \tag{47}$$

Substituting $\varepsilon_\beta(\theta) = 2e^{-2\beta c_L(1-\theta)}$ from Lemma 9, $\pi_\theta \le 2Le^{-c_1(\theta)/\rho}$ from Lemma 8 and $c_L(1-\vartheta_B) \le 2e^{-B\vartheta^2/2}$ from Lemma 6 together with $c_L \le 1$, we obtain

$$\left|\mathbb{E}[\Delta y_i^\mu \mid A]\right| \le \frac{20}{r}\left(e^{-2(1-\theta)\beta c_L} + L\, e^{-c_1(\theta)/\rho}\right) + 2e^{-B\vartheta^2/2} + \max_{\nu \le K}\left|c^\nu - c_L \mathbf{1}\{\nu \in A\}\right|, \quad (48)$$

where we used $10 \le 20$ to write the two exponential channels with a single constant. Since $\theta \in (0,1)$ was arbitrary, (48) holds for every θ and therefore for the infimum over θ, which is the form (17) takes, with $C = 20/r$. The symmetric choice $\theta = 1/2$ gives $2(1-\theta) = 1$ and $c_1(1/2) = (1-r^2)/(8L^2)$, the two exponents in the readable form quoted after (17).

Step 5: averaging over the item and over the archetypes. The right hand side of (48) depends on the drawn tuple only through the last term. Averaging (48) over the item drawn from $\mathcal{Q}$ therefore gives the same expression with that term replaced by its average over the T distinct tuples carried by $\mathcal{S}$, which is at most its maximum over them. By the uniform tail (41) and a union bound over the K hidden units and the T tuples, for every $\gamma > 0$ the event on which that maximum is smaller than γ has $\mathbb{P}_{\boldsymbol{\xi}}$-probability at least $1 - 2KTe^{-N\gamma^2/2}$, and on it the right hand side may be maximised over i and μ. Recalling $\Delta w_i^\mu = \epsilon\beta\, \Delta y_i^\mu$ gives (17). □

Proof of the limit statement of Theorem 1. The choice $\gamma = \sqrt{2\ln(2KTN)/N}$ makes $2KTe^{-N\gamma^2/2} = N^{-1}$, which gives the stated confidence level. The first term of (17) vanishes as $\beta \to \infty$, the second as $M \to \infty$, that is $\rho \to 0$, the third as $B \to \infty$ and the fourth as $N \to \infty$ provided $\ln(KT) = o(N)$; the four channels being independent, the order of the limits is immaterial. Finally, taking $\mathbb{E}_{\boldsymbol{\xi}}$ of (48) at fixed i, μ and using $\mathbb{E}_{\boldsymbol{\xi}}|c^\mu - c_L\mathbf{1}\{\mu \in A\}| \le N^{-1/2}$ from Lemma 10 replaces the last term by $N^{-1/2}$. □

Remark 4. Theorem 1 is a statement about the mean of a single update, not about its fluctuations, which remain of order one even where the mean vanishes: the positive phase is built on one mini-batch and the negative one on a single Gibbs sweep, and neither is averaged. At a finite learning rate the weights therefore perform a stochastic approximation around $\bar{\boldsymbol{\eta}}$ rather than sitting exactly on it, with a stationary spread that shrinks with $\epsilon\beta$. This is the usual situation of stochastic gradient methods, and it is why the fixed point is to be understood in the sense of Definition 1.

B THE MODEL ITSELF: THE BOUND AND THE TASK ON RADEMACHER PATTERNS

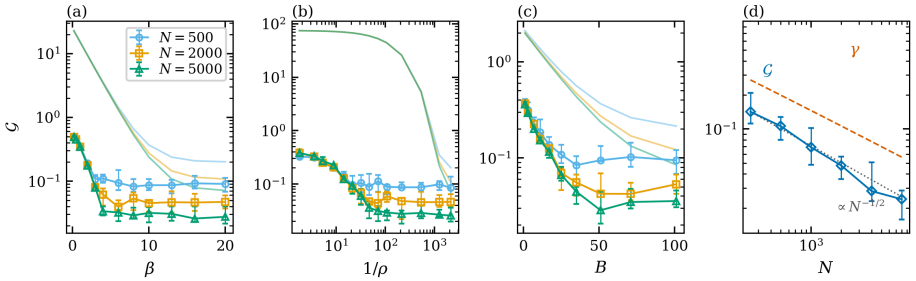


Figure 4: Numerical verification of Theorem 1. Residual drift $\mathcal{G}$ at the planted weights $\boldsymbol{w} = \bar{\boldsymbol{\eta}}$, for $K = 20$ archetypes, $L = 3$ and a single label tuple, so $T = 1$ in (17). Markers are the median over 6 samples of the quenched archetypes and bars span their range; thin lines are the right hand side of (17) with its own constants, one per system size and in the matching colour. Each panel switches off every channel but one: *(a)* thermal, against β; *(b)* dataset, against $1/\rho$; *(c)* mini-batch, against B; *(d)* finite size, against N. The plateau of the first three panels is that fourth channel, which no parameter removes and which orders the curves as $N^{-1/2}$. Parameters in Appendix B.

Everything in this appendix lives inside the model of Section 2: archetypes drawn i.i.d. Rademacher, examples generated by (4), queries built at test time from examples independent of those that entered the class averages. This is the only place in the paper where the hypotheses of Theorem 1 all hold, so

it is the only place where the theorem can be checked rather than probed, and two things are checked: whether each of the four channels of the bound behaves as its parametric form says, and whether the planted weights do the task the paper is named after.

The residual drift of Figure 4 is evaluated through the closed-form reduction of Lemma 7, so the three averages internal to a CD-1 step carry no Monte Carlo error and the only sampling is over the data and over the archetypes: forty realisations for the dataset average, and the median over six samples of the quenched archetypes with the bars spanning their range. Each panel holds three channels at their saturated value and sweeps the fourth, at $K = 20$, $L = 3$, a single label tuple, and $N \in \{500, 2000, 5000\}$. The thermal panel sweeps β from 0.2 to 20 at $r = 0.8$, $M = 10^4$ and $B = 101$; the dataset panel sweeps M from 1 to 1200 at $r = 0.8$ and $\beta = 40$, with the cue at its large-batch limit, so that $1/\rho$ runs from 1.8 to $2.1 \cdot 10^3$; the mini-batch panel sweeps B from 1 to 101 at $r = 0.5$, $M = 5001$ and $\beta = 40$; and the finite-size panel sweeps N from 250 to 8000 at $r = 0.8$, $M = 10^4$, $B = 101$ and $\beta = 40$. The plateau visible in the first three panels is the finite-size channel, which no parameter removes.

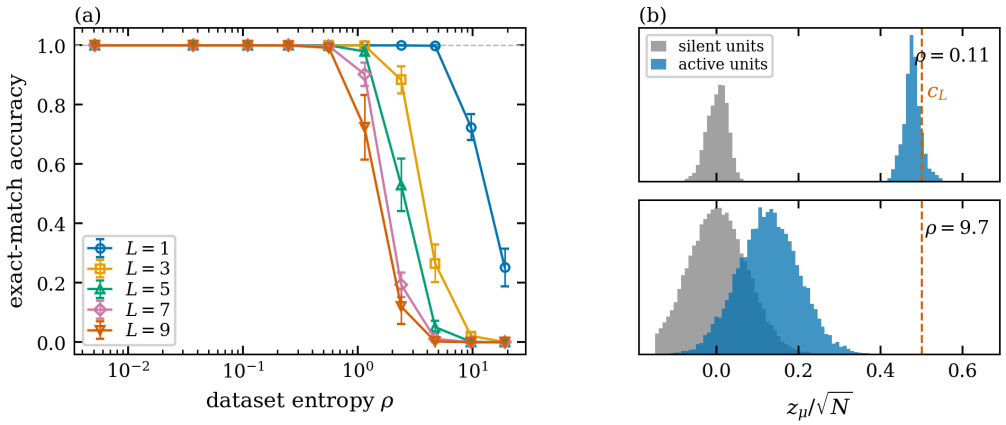


Figure 5: Multi-label classification of spurious states at the planted weights, with no training at all. At each query a tuple of L archetypes is drawn at random and is not revealed, the read-out $\mathbb{E}[z_\mu]$ is thresholded by the parameter-free largest-gap rule, and the prediction is scored by exact match, so that a single wrong bit invalidates the whole label vector. *(a)* Exact-match accuracy against the dataset entropy ρ, tuned through the quality r, for mixture orders $L \in \{1, 3, 5, 7, 9\}$: classification is perfect up to a threshold entropy that decreases with L, as the shrinking margin c_L of Theorem 1 predicts. *(b)* Distribution of the read-out of the active and of the silent units at $L = 3$, in the regime where the machine classifies (top) and in the one where it does not (bottom); the dashed line marks c_L. The failure is the collapse of the gap: the silent population stays centred on zero, while the active one both broadens and slides down to $\mathbb{E}[\xi^\mu s\, \vartheta_B(|\Sigma|)]$, the two effects contributing comparably. Markers are means over 12 realisations and bars denote one standard deviation; parameters in Appendix B.

Figure 5 turns the same weights into a classifier and sweeps the one quantity the dataset channel depends on. At $N = 2000$, $K = 20$, $M = 21$ with $B = M$, $\beta = 25$ and $L \in \{1, 3, 5, 7, 9\}$, the dataset entropy is tuned through the quality r, which runs from 0.05 to 0.95 and takes ρ from 19 down to $5.1 \cdot 10^{-3}$; each point is the mean over twelve realisations of three hundred queries each, and the two histograms are taken at $r = 0.55$ and $r = 0.07$, that is at $\rho = 0.11$ and $\rho = 9.7$. The decision rule is the parameter-free largest-gap threshold of Appendix D.3.

C Additive mixtures: proofs of Lemma 2 and Corollary 1

Proof of Lemma 2. Fix a row $\boldsymbol{p}$ of $\boldsymbol{D}$, so that $\boldsymbol{p}$ has i.i.d. centred Gaussian entries and the pair $(\langle \boldsymbol{p}, \boldsymbol{x}^\mu \rangle, \langle \boldsymbol{p}, \boldsymbol{y} \rangle)$ is centred bivariate Gaussian with correlation $\kappa_\mu = \langle \boldsymbol{x}^\mu, \boldsymbol{y} \rangle / (\|\boldsymbol{x}^\mu\| \|\boldsymbol{y}\|)$. For a centred bivariate Gaussian (X, Y) with correlation κ, Grothendieck's identity gives $\mathbb{E}[\mathrm{sign}(X)\mathrm{sign}(Y)] = 1 - 2\mathbb{P}(XY < 0) = \frac{2}{\pi} \arcsin \kappa$, which is (20); the rows are independent and identically distributed, so the value does not depend on i, and the common scale of the rows is irrelevant because the sign is scale invariant.

If $\langle \boldsymbol{x}^\mu, \boldsymbol{x}^\nu \rangle = 0$ for $\mu \neq \nu$ and $\|\boldsymbol{x}^\mu\| = \|\boldsymbol{x}^\nu\|$, then with equal weights $a_\mu \equiv a$ one has $\langle \boldsymbol{x}^\mu, \boldsymbol{y} \rangle = a\|\boldsymbol{x}^\mu\|^2$ and $\|\boldsymbol{y}\| = a\sqrt{L}\,\|\boldsymbol{x}^\mu\|$, so $\kappa_\mu = L^{-1/2}$ and the overlap is g_L; the expansion follows from $\arcsin u = u + O(u^3)$. □

Remark 5. Comparing with Lemma 1, $c_L \simeq \sqrt{2/(\pi L)}$ and $g_L \simeq 2/(\pi\sqrt{L})$ have the same $L^{-1/2}$ decay and differ by the constant $\sqrt{2/\pi} \simeq 0.798$. Numerically $(c_L, g_L) = (0.500, 0.392)$, $(0.375, 0.295)$, $(0.313, 0.247)$, $(0.273, 0.216)$ at $L = 3, 5, 7, 9$. Which of the two governs a given experiment is decided by *where the superposition happens*: before the sign, as in (18), it is g_L; after it, as in (2), it is c_L. On synthetic Gaussian sources, where both regimes can be realised at will, the dichotomy is verified directly.

In (19) the role of $\boldsymbol{D}$ is played by the composition $\boldsymbol{\Omega T}$: $\boldsymbol{T}$ is a fixed linear map and $\boldsymbol{\Omega}$ has i.i.d. Gaussian rows, so conditionally on $\boldsymbol{T}$ the rows of $\boldsymbol{\Omega T}$ are i.i.d. centred Gaussian vectors with covariance $\boldsymbol{T}^\top \boldsymbol{T}$, and Lemma 2 applies with the inner products of (20) taken in the metric induced by $\boldsymbol{T}$. This is the sense in which whitening matters: it is what makes those inner products the ordinary ones on the whitened features, and Appendix D.2 returns to the point.

Corollary 1 (Read-out of a physical mixture)**.** *Under the hypotheses of Lemma 2, let $\kappa \in (0, 1]$ be the correlation, before the sign, between the projection of a prototype example of class μ and that of the fresh source of the same class entering the query, and let it be the same for all prototypes. Then the constituent hidden units of the machine of Section 3 read*

$$Z_\mu = \frac{\mathbb{E}[z_\mu \mid \boldsymbol{\sigma}]}{\sqrt{N}} = \frac{1}{\hat{r}}\,\frac{2}{\pi}\arcsin\frac{\kappa}{\sqrt{L}}, \qquad \mu \in A, \tag{49}$$

and the silent ones read zero up to $O(N^{-1/2})$. In units of its own value at $L = 1$, which removes $\hat{r}$ and κ alike from the leading behaviour, the ratio $R(L) = \arcsin(\kappa/\sqrt{L})/\arcsin\kappa$ decreases with κ and therefore obeys the parameter-free band (21), $R(L) \in [g_L, L^{-1/2}]$.

Proof of Corollary 1. Write $\boldsymbol{a}^{\mu,m} = \boldsymbol{\Omega T}\boldsymbol{x}^{\mu,m}$ for the projection of the m-th prototype example of class μ and $\boldsymbol{a}^{\mathrm{mix}} = \sum_{\nu\in A} \boldsymbol{a}^{\nu,q}$ for that of the query, built from sources held out from the prototypes; the second equality is exact by (18) and the linearity of $\boldsymbol{\Omega T}$. By hypothesis the sources are orthogonal and of equal energy, and the prototype is correlated with the fresh source of its own class alone, so

$$\langle \boldsymbol{a}^{\mu,m}, \boldsymbol{a}^{\mathrm{mix}} \rangle = \langle \boldsymbol{a}^{\mu,m}, \boldsymbol{a}^{\mu,q} \rangle = \kappa\,\|\boldsymbol{a}^{\mu,m}\|\,\|\boldsymbol{a}^{\mu,q}\|, \qquad \|\boldsymbol{a}^{\mathrm{mix}}\| = \sqrt{L}\,\|\boldsymbol{a}^{\mu,q}\|, \tag{50}$$

whence the correlation entering (20) is $\kappa/\sqrt{L}$. Applying Lemma 2 bit by bit and averaging over the M_μ prototypes and over i gives $\frac{2}{\pi}\arcsin(\kappa/\sqrt{L})$ for the raw class average, and the division by $\hat{r}$ in the class averages of Section 3 gives (49). For $\mu \notin A$ the two projections are uncorrelated, so the expectation vanishes and the empirical average is $O_{\mathbb{P}}(N^{-1/2})$ by Lemma 10.

For (21) set $c := L^{-1/2} \in (0, 1]$ and $h(u) := \arcsin(u)/u = \sum_{k\geq 0} a_k u^{2k}$ with $a_k = \binom{2k}{k} 4^{-k}(2k+1)^{-1} > 0$, a power series with non-negative coefficients convergent on $[0, 1]$. Then $R(L) = c\,h(c\kappa)/h(\kappa)$ and

$$\kappa\,\frac{d}{d\kappa}\log R = \Phi(c\kappa) - \Phi(\kappa), \qquad \Phi(u) := u\,\frac{h'(u)}{h(u)} = \frac{\sum_k 2k\,a_k u^{2k}}{\sum_k a_k u^{2k}}. \tag{51}$$

$\Phi(u)$ is the expectation of $2k$ under the probability weights proportional to $a_k u^{2k}$, which shift stochastically towards larger k as u grows, so Φ is non-decreasing; since $c\kappa \leq \kappa$ we get $\frac{d}{d\kappa}\log R \leq 0$. Hence R decreases on $(0, 1]$ from $R(0^+) = c = L^{-1/2}$ to $R(1) = \arcsin(L^{-1/2})/(\pi/2) = g_L$, which is (21). □

Remark 6 (the three quality parameters). The paper carries three measures of how good the prototypes are, and they are one quantity read in three places. Applying (20) to a prototype and a fresh source of the same class turns the pre-sign correlation κ into the bit-level quality $r = \frac{2}{\pi}\arcsin\kappa$ of (4), and r then enters the dataset entropy $\rho = (1 - r^2)/(Mr^2)$ of (5). Large κ, large r and small ρ are the same statement.

This is also why the class-level prediction is the band (21) and not a point. A point would need κ, and κ is not an observable of the machine: what the $L = 1$ level gives is the marginal correlation of a

single prototype, $\kappa \simeq 0.79$ on NSynth, which is not the effective correlation a class average inherits, since the examples of a class do not differ from it independently across sites. The band is insensitive to that distinction. The value $R(3) = 0.520$ that the marginal κ would give and the measured 0.397 lie both inside it, and Corollary 1 holds either way.

D Datasets, encoder and read-out

This appendix fixes everything that is common to the four settings and does not depend on which experiment is run on them: the map from a physical signal to a feature vector, the map from a feature vector to the visible layer, the rule that turns a hidden read-out into a set of labels, and then, dataset by dataset, what a query is, how the two pools are split and how the weights are built. The two experiments of Section 4 are described in Appendices E and F, and what the theory is worth on these signals in Appendix G.

D.1 The physical front end

This appendix fixes the feature map $\boldsymbol{x}(\cdot)$ of (18), domain by domain, at the level of detail needed to recompute it from the raw files. Only one property is required of it, linearity in the quantity that physically adds, and it is that property which forbids a logarithm: $\log(a+b)$ is not $\log a + \log b$, so a log-mel spectrogram or any decibel scale would stop the superposition being a sum before the encoder sees it, and Lemma 2 would no longer describe anything. Keeping the linearity has a price, paid in Appendix D.6 as the gap between the mel-power and the VGGish representations.

Amplitude normalisation. Every audio file is normalised to unit root-mean-square energy before anything else,

$$\tilde{w}(t) = \frac{w(t)}{\sqrt{T^{-1}\sum_t w(t)^2 + \varepsilon}}, \qquad \varepsilon = 10^{-12}, \tag{52}$$

and mixtures are sums of normalised waveforms. This is not cosmetic: the equal energy hypothesis of Lemma 2 is exactly the statement that the L sources enter the sum with the same power, and (52) is how it is enforced. Appendix D.5 reports what happens when it is deliberately violated.

Audio, wide band (OpenMIC-2018). Sampling rate $f_s = 44.1$ kHz. Short-time Fourier transform with a Hann window of $N_{\text{fft}} = 1024$ samples (23.2 ms) and hop 512 (11.6 ms, 50% overlap), with no artificial padding and no boundary extension. The linear power spectrogram $S(f,t) = |Z(f,t)|^2$ is passed through a bank of 128 triangular filters equispaced on the mel scale of O'Shaughnessy,

$$m(f) = 2595 \log_{10}\Big(1 + \frac{f}{700}\Big), \qquad f_{\min} = 50 \text{ Hz}, \quad f_{\max} = f_s/2 = 22.05 \text{ kHz}, \tag{53}$$

and the frame axis is cut into $N_{\text{seg}} = 8$ contiguous blocks of equal length whose frames are averaged arithmetically, giving $\boldsymbol{x} \in \mathbb{R}^d$ with $d = 128 \times 8 = 1024$. Both operations after the spectrogram are linear, so (18) is inherited from the additivity of the power spectrum of incoherent sources.

Audio, pitch resolved (NSynth). Sampling rate $f_s = 16$ kHz, files of 4.0 s at 16 bit. Only the sustained part of the note is kept, the interval $[T_0, T_1] = [0.15, 2.60]$ s (samples 2400 to 41 600), which removes both the percussive attack and the final decay and leaves the stationary harmonic spectrum that carries the timbre. Short-time Fourier transform with a Hann window of $N_{\text{fft}} = 8192$ samples (512 ms, for resolution at low frequency) and hop 2048 (128 ms); the power spectrum is averaged over the frames. The filter bank is logarithmic in frequency, with 3 bins per semitone between MIDI 21 ($A_0 \simeq 27.5$ Hz) and MIDI 108 ($C_8 \simeq 4186$ Hz), that is $(108-21)\times 3 + 1 = 262$ adjacent triangular filters centred at $f_c(m) = 440 \cdot 2^{(m-69)/12}$ and each normalised to unit weight sum; $d = 262$.

Hyperspectral data. The features are the reflectance spectra delivered by the sensor, with no filtering and no logarithm: $d = 162$ bands over 0.4–2.5 μm for Urban and 156 for Samson. Here (18) is not an approximation to be verified but the linear mixing model, the accepted physics of the instrument, and the Samson image lets us measure how well the instrument obeys it. Writing $\boldsymbol{y}_p$ for

the measured reflectance of pixel p, $\boldsymbol{e}_\mu$ for the ground-truth endmembers and $a_{\mu p}$ for the ground-truth abundances, the cosine between $\boldsymbol{y}_p$ and $\sum_\mu a_{\mu p}\boldsymbol{e}_\mu$ is 0.9989 on average over the 9025 pixels and 0.9876 at the worst pixel of the scene, and it is 0.99915, 0.99877 and 0.99754 on the pixels that mix one, two and three materials at the presence threshold 0.10. The two vectors differ by a per-pixel factor of order 0.38, which is illumination and topography rather than a failure of additivity, and the encoder (19) is blind to it: $\operatorname{sign}(\boldsymbol{\Omega T x})$ is unchanged under $\boldsymbol{x} \mapsto c\boldsymbol{x}$ for every $c > 0$.

D.2 The encoder

The encoder (19) is a whitening followed by a Gaussian projection and a sign. Let $X_{\text{proto}} \in \mathbb{R}^{n\times d}$ collect the features of the prototype pool, the pool that builds the weights, never the queries, and let $X_{\text{proto}} = USV^\top$ be its singular value decomposition. With two integers $\text{skip} \geq 0$ and $q \geq 1$,

$$\boldsymbol{T} = \operatorname{diag}\big(s_{\text{skip}+1}, \dots, s_{\text{skip}+q}\big)^{-1} V^\top_{:,\,\text{skip}+1:\text{skip}+q} \in \mathbb{R}^{q\times d}, \qquad \Omega_{ij} \sim \mathcal{N}(0, q^{-1}) \text{ i.i.d.}, \tag{54}$$

$\boldsymbol{\Omega} \in \mathbb{R}^{N\times q}$ drawn once and frozen. Nothing here uses the labels and nothing is fitted on a query.

The two integers. With $\boldsymbol{\Omega}$ Gaussian the pre-sign pair is jointly Gaussian for every fixed $\boldsymbol{T}$, so Lemma 2 holds whatever the front end and whitening is not what makes (20) true. On Urban the law is accurate to 2.7 to $3.2 \cdot 10^{-3}$ on the raw bands and to 5.3 to $6.4 \cdot 10^{-3}$ on the whitened ones, over $L \in \{2, 3, 5\}$ and evaluated at the correlation each metric reports. What $\boldsymbol{T}$ decides is the metric in which κ_μ is read, and with it how close to orthogonal the archetypes are. Raw spectral bands are strongly correlated with each other and very unequal in variance: the leading singular direction of the Urban pool carries 0.909 of its energy, and the K class codes built on the raw features have mean $|\cos|$ equal to 0.639 against 0.024 after whitening. What whitening buys is therefore the low-load hypothesis $K = O(1)$ of Section 2 and the band of Corollary 1, not the law.

The leading singular direction is discarded ($\text{skip} = 1$) for OpenMIC and for the hyperspectral libraries. It is the direction along which the pool varies most, the loudness of the recording or the mean albedo of the material; whitening rescales it to the variance of the others without removing it, and every class sits on the same side of it, so the class average built on it describes the size of the observation as much as its shape. Discarding it does not make the archetypes more orthogonal. Removing a component common to K vectors pushes their pairwise correlations towards $-1/(K-1)$, and that is what is measured: on Urban the $K = 6$ class codes have mean cosine $+0.017$ at $\text{skip} = 0$ and -0.091 at $\text{skip} = 1$, thirteen of the fifteen pairs negative against the -0.2 a fully shared component would force, with mean $|\cos|$ going from 0.024 to 0.114. Near-orthogonality is what the whitening buys; skip chooses what the class average describes, and we report it as the choice it is. NSynth is the exception, with $\text{skip} = 0$: the per-note normalisation (52) has already removed the loudness carried by that direction, and with $K = 49$ classes and only $q = 50$ retained directions the pool cannot spare one. The choice was made once, on the prototype pool, and no test quantity entered it.

Parameter values. $q = 50$ (NSynth), $q = 40$ (hyperspectral), $q = 400$ (OpenMIC on mel-power) and $q = 100$ (OpenMIC on VGGish). On NSynth and on the hyperspectral libraries q is fixed on the off-diagonal mass of the Gram matrix of the K archetype codes, and not by any accuracy on the task. That mass has no interior minimum: on Urban at $\text{skip} = 1$ it reads 0.240, 0.122, 0.122, 0.114, 0.104, 0.095 and 0.090 at $q = 5, 10, 20, 40, 80, 120, 160$, decreasing throughout. The rule is the smallest q past which it stops falling appreciably, here $q = 40$; the remaining 0.023 costs a factor four in q, and the directions that would buy it are the tail of the spectrum, where the whitening divides by singular values the pool estimates from few spectra. Below $q = 10$ the codes are visibly correlated. On OpenMIC it is chosen by validation macro-AP on a split carved out of the training set (Appendix D.6). The number of bits is $N = 16384$ except in the contrastive-divergence experiments of Figures 2 and 6(c) and in the read-out comparison of Figure 3(d), which use $N = 4096$ for cost; $N = 16384$ puts the finite-size floor $N^{-1/2}$ at $7.8 \cdot 10^{-3}$, an order of magnitude below every effect reported in Sections 3 and 4.

D.3 From a read-out to a label set

Given a visible configuration $\boldsymbol{\sigma}$, the field arriving on hidden unit μ is $u_\mu(\boldsymbol{\sigma}) = N^{-1/2}\sum_i w_i^\mu \sigma_i$ by (9), and the conditional mean of (8) makes the observable read-out the empirical overlap

$$Z_\mu := \frac{\mathbb{E}[z_\mu \mid \boldsymbol{\sigma}]}{\sqrt{N}} = \frac{1}{N}\sum_{i=1}^{N} w_i^\mu \sigma_i \ \in [-1, 1]. \tag{55}$$

A multi-label prediction is a binary vector $\hat{\boldsymbol{y}} \in \{0,1\}^K$, and producing one from $\boldsymbol{Z}$ requires a threshold. Since the plateau of the active units is g_L and therefore moves with L, a fixed threshold would have to be tuned per order; instead we use the parameter-free *largest-gap* rule: sort $Z_{(1)} \le \cdots \le Z_{(K)}$, let

$$j^\star = \arg\max_{1\le j\le K-1} \big(Z_{(j+1)} - Z_{(j)}\big), \qquad \theta^\star = \tfrac{1}{2}\big(Z_{(j^\star)} + Z_{(j^\star+1)}\big), \qquad \hat{y}_\mu = \mathbf{1}\{Z_\mu > \theta^\star\}, \tag{56}$$

so that the rule adapts to the plateau on its own and returns, as a by-product, the estimated order $\hat{L} = K - j^\star$. It is never told L. Since $j^\star$ ranges over $1, \dots, K-1$, the rule returns between one and $K-1$ labels: it cannot answer "none" and it cannot answer "all". That costs nothing where K is large, and it costs something on Samson, where $K = 3$ and the 293 pixels that mix all three materials, 3.2% of the labelled image, cannot be got exactly right; micro-F1 on them is capped at 0.8, and on the 84 of them that fall in the held-out set of Appendix E it measures 0.743.

This method is strictly harder than the oracle that is told L and takes the top L activations; both are reported in Appendix F so that the cost of not knowing L is visible.

Scores are the exact match, which counts a query as correct only if the whole label vector is right and is therefore severe at large K, and the micro-F1 over all (query, class) pairs. When the annotation is partial, as on OpenMIC, both are restricted to the annotated pairs and the primary metric becomes the average precision per class of the continuous score Z_μ, macro-averaged, which is the standard choice for partially labelled multi-label data and needs no threshold at all. Exact match is reported alongside micro-F1 throughout because at $K = 49$ it collapses for $L \ge 3$, where a single wrong entry among 49 invalidates the vector and even a decoder told L stays below 0.18 (Table 3); micro-F1 alone would flatter the machine.

D.4 NSynth: chords of recorded notes

Which split. The law and read-out measurements of this subsection use the pool of the official valid and test splits together, $M = 91$ prototype notes per pitch at $N = 16384$, which is the protocol behind Table 1's law error. Figures 2(a) and 3(a) use the test split alone, $M \simeq 15$ notes per pitch, because rendering a chord requires the waveforms and only that split's audio was available for those runs; the 53 instruments are the same in the two splits, so the instrument split that defines the task is unchanged and what shrinks is the number of examples behind each class average. The estimated quality is $\hat{r} = 0.77$, and the planted micro-F1 averaged over $L \in \{1, 3, 5\}$ is 0.481.

Task and split. The archetypes are the $K = 49$ semitone classes from MIDI 36 (C_2) to MIDI 84 (C_6). The filter bank of Appendix D.1 runs wider, from MIDI 21 to 108, and the classes are restricted to the central four octaves for three reasons. Below C_2 the spacing between adjacent semitones, 3.9 Hz at MIDI 36, is no larger than the half-width of the main lobe of the analysis window, 3.9 Hz, and it halves every octave down, so the classes stop being resolved. At C_6 only seven harmonics fall below the Nyquist frequency and above it fewer still, three by MIDI 96, so the timbre the class average is supposed to carry is largely cut away. And the pool itself thins out at the edges: inside the range every pitch has between 189 and 246 notes, outside it the count falls as low as one, which would leave M_μ neither balanced nor usable. The 53 instruments are partitioned at random into two disjoint halves with seed 0: the prototypes are notes played by the instruments of the first half, the queries are chords built by summing the unit-RMS waveforms of notes played by the instruments of the second, truncated to their common length. The generalisation is therefore across *timbre*, and no recording that enters the weights ever appears in a query.

Weights. The codes are $\boldsymbol{\sigma}(\boldsymbol{x}^{\mu,m})$ of (19) with $N = 16384$, $q = 50$, skip $= 0$; the raw class averages $W^{\mathrm{raw}}_{:,\mu} = M_\mu^{-1} \sum_m \boldsymbol{\sigma}(\boldsymbol{x}^{\mu,m})$ are made unbiased by the estimator of footnote 2,

$$\widehat{\mathcal{R}} = \frac{1}{KN} \sum_{\mu=1}^{K} \sum_{i=1}^{N} \left(W^{\mathrm{raw}}_{i\mu}\right)^2, \qquad \hat{r} = \sqrt{\max\left(\frac{M\widehat{\mathcal{R}} - 1}{M - 1}, 10^{-12}\right)}, \qquad W = \frac{W^{\mathrm{raw}}}{\hat{r}}, \tag{57}$$

which returns $\hat{r} = 0.7825$ with $M = 91$ prototype notes per pitch, hence a nominal dataset entropy $\rho = (1 - \hat{r}^2)/(M\hat{r}^2) = 0.0070$. That number overstates what the examples carry. The corruption of (4) is independent across sites and the corruption of a real recording is not, two instruments playing the same pitch differing by whole bands of harmonics rather than by scattered bits: the measured variance of the overlap between an archetype code and the code of one of its examples is $2.6 \cdot 10^3$ times the value the model predicts, so the effective number of independent sites is $N_{\mathrm{eff}} = 6.4$ and the effective dataset entropy $\rho_{\mathrm{eff}} = \rho N / N_{\mathrm{eff}} = 17.8$. Sweeping M over two decades moves micro-F1 by 0.087 at $L = 3$, which is the channel saturating. Writing the dataset channel of Theorem 1 in terms of an effective dimension is the first of the three directions the conclusion names. Where the classes are of unequal size the estimator is evaluated at $M = \min_\mu M_\mu$, which biases $\hat{r}$ down and the weights up: on Urban, whose classes hold 34 to 82 spectra, it gives 0.3435 against the 0.3561 obtained by pooling the per-class estimates at their own M_μ, a difference of 3.5%. The conservative direction is the one taken, and by (56) the scale of the weights does not reach the decoder in any case. After the division the read-out (55) lives in $[-\hat{r}^{-1}, \hat{r}^{-1}]$ rather than in $[-1, 1]$.

D.5 Hyperspectral: the Urban library and the Samson image

The linear mixing model of remote sensing states that the reflectance of a pixel *is* the abundance-weighted sum of the endmember spectra,

$$\boldsymbol{y} = \sum_{\mu=1}^{K} a_\mu \boldsymbol{e}_\mu + \boldsymbol{\varepsilon}, \qquad a_\mu \geq 0, \quad \sum_{\mu=1}^{K} a_\mu = 1, \tag{58}$$

so (18) needs no verification here. More than that, the spectral libraries distributed with these images have exactly the structure Section 2 postulates: each material comes with M measured spectra whose class average is the archetype and whose spread is the dataset entropy. Those spectra are pixels of the scene itself, gathered into a library by the people who published the data, and not measurements made elsewhere: each of the 651 Urban spectra is an exact copy of a cube pixel up to the documented factor 10^3, and each of the 105 Samson spectra is an exact copy of a near-pure one. The prototype half therefore carries the scene, and what a query holds out is the mixing.

Protocol. Each library is split at random into two disjoint halves with seed 91: half the spectra of every material build the prototypes, the other half the queries. A query is the abundance-weighted sum of one held-out spectrum per material, with equal abundances unless stated. The encoder is (19) with $N = 16384$, $q = 40$, skip $= 1$, and $N_Q = 400$ queries per order for the law and band measurements of Appendix G, $N_Q = 300$ for the read-out comparison of Appendix F and for the held-out evaluation of Appendix E. Urban has $K = 6$ materials, 162 bands, 69 to 164 library spectra per class, $M = 34$ in the smallest class, $\hat{r} = 0.3435$ and $\rho = 0.220$ at $N = 16384$, 0.3422 and 0.222 at the $N = 4096$ of Appendix E. Samson has $K = 3$, 156 bands and 30, 30 and 45 spectra per class, so $M = 30$ in the smallest; there (57) gives $\hat{r} = 0.1317$ at $N = 4096$ and the weights are divided by it as everywhere else. The value is small because the library spectra of a class are nearly uncorrelated after whitening, and by (56) the scale does not reach the decoder in any case.

Measured pixels: the Samson image. The mixtures in this test are not assembled by us. The Samson cube is $95 \times 95 = 9025$ measured pixels of a real scene distributed with ground-truth endmembers $\boldsymbol{e}_\mu \in \mathbb{R}^{156}$ and ground-truth abundance maps $a_\mu(p)$. The prototypes are the 105 spectra of the library distributed with the scene, and those spectra are pixels of the cube itself: each is an exact copy of one pixel, taken at a ground-truth purity of at least 0.901 for soil and 0.98 for tree and water. The weights are built from near-pure pixels of the same image, not from an independent source, and the claim this setting supports is that the machine reads mixtures it has not seen, not that it was blind to the scene. Of the 105 pixels, 25 fall in the held-out set, all of them at $L = 1$: removing them moves the step-zero micro-F1 at $L = 1$ from 0.901 to 0.900 and leaves the average over the three orders unchanged. The queries are the measured pixels; the labels are the abundances thresholded at a presence level, 0.10 unless stated, which leaves $\bar{L} = 1.56$ materials per pixel.

D.6 OpenMIC-2018: music nobody mixed for us

NSynth and the spectral libraries answer the question of whether the theory describes physical superposition; they do not answer the objection that the superpositions, though physical, were assembled by the experimenter. OpenMIC-2018 is 20 000 ten-second excerpts of Creative Commons music from the Free Music Archive, recorded and mixed by musicians, and crowd-annotated for the presence or absence of 20 instruments.

Partial labels. Only 41 268 (excerpt, instrument) pairs are annotated, that is 10.3% of the total: the annotation says which instruments are present *among those that were asked about*, and is silent on the rest. This has three consequences. Every metric is computed on the annotated pairs only, through the mask Y_{mask}. The primary metric is the macro average precision, whose chance level on this test split is 0.447. And there is no exact mixture order L: the number of annotated positives is a lower bound on it, which is why Table 1 reports no law error for OpenMIC.

Prototypes without isolated sources. There are no isolated stems, so the construction of Section 3 has to be adapted, and we report both adaptations rather than picking one silently. The `solo` prototypes use only the training excerpts whose sole annotated-present instrument is the one in question (10 863 of the 20 000 excerpts have a single annotated positive; on the training split the median is 361 per instrument): this is the closest surrogate for an isolated example and the setting the theory describes. The `all` prototypes use every training excerpt in which the instrument is annotated present (median 643 per instrument), accepting that other instruments are also playing and relying on their incoherent cancellation in the class average. The experiments reported here, both the training of Figure 2(d) and the read-out of Figure 3(d), use `solo`, falling back to `all` for any instrument with fewer than 20 such excerpts, which never occurs: the smallest solo pool holds 71 on the split the prototypes are built from, 89 on the training split as a whole. It is the conservative choice, since `all` measured on its own at $N = 16384$ and $q = 400$ gives macro-AP 0.5465 against 0.5351 for `solo`. The value in Table 1 is the $N = 4096$ protocol of Figure 2. The weights are divided by the root mean square of the raw class averages rather than by $\hat{r}$ of (57): with prototypes that are mixtures the debiasing has nothing to debias, and by (56) and the invariance of average precision under a positive rescaling the choice is unobservable in the score. It is not unobservable in Figure 2(d), where it fixes the per-entry scale at which the random initialisation is drawn to match.

Hyper-parameters. The official partition `split01` gives 14 915 training and 5 085 test excerpts. A validation split is carved out of the training set by taking every fifth index, q is selected on it, and the test excerpts are used once. The features are the mel-power spectra of Appendix D.1 at $q = 400$.

E The first experiment: supervised CD-1 from the two initialisations

This appendix gives the protocol behind Figure 2, panel by panel. The datasets themselves, the encoder and the decoder are those of Appendix D; what is added here is how the training items are drawn, how the random initialisation is scaled and what is held out.

All four panels use $N = 4096$ visible neurons so that $8 \cdot 10^3$ CD-1 steps are affordable, $\beta = 8$, a score taken every 400 steps with the largest-gap decoder of Appendix D.3, and the learning rate grid $\epsilon\beta \in \{1, 3, 10, 30\} \cdot 10^{-3}$ run for *both* initialisations, so that no rate is chosen for one of them. The random start is drawn i.i.d. Gaussian with the per-entry standard deviation of the planted weights, so the two differ in structure and not in scale. Five independent splits are used on the hyperspectral panels and three seeds on NSynth and OpenMIC, each seed drawing its own random projection $\mathbf{\Omega}$.

What the update does to the weights. On Urban the spectra the weights stand for come through the $8 \cdot 10^3$ steps almost untouched: the cosine between the band-space archetype before and after stays above 0.99 over all six classes and all four rates. What changes is how large the archetypes are and how they sit relative to each other. Their column norms move by a factor between 0.49 and 1.21, shrinking at the three smaller rates and holding at $\epsilon\beta = 3 \cdot 10^{-2}$; their mean $|\cos|$ falls from 0.118 at $N = 4096$ to 0.073, 0.048, 0.048 and 0.054 across the grid; and the fraction of the squared

net displacement lying along $\boldsymbol{W}_0$ itself is 0.808, 0.773, 0.534 and 0.234, so at the smaller rates the update is largely a change of length and at the largest it is not.

The decoder sees only part of that. On one of the five splits, at $\epsilon\beta = 3 \cdot 10^{-3}$, the planted weights score 0.8227 pooled over the held-out queries of all three orders, and training takes them to 0.8651. Multiplying the planted weights by the single mean gain 0.587 leaves the score at 0.8227 exactly, as (56) requires; multiplying them by the six per-class gains, which run from 0.490 to 0.650, gives 0.8361. So about a third of what training gains is the classes being rebalanced against each other, and the rest is the archetypes being decorrelated. This is the decomposition Section 4 reports.

NSynth. $K = 49$, 2500 training chords drawn with $L \in \{1, 3, 5\}$ from the prototype instruments, evaluation on 200 held-out chords per order.

Urban. The 651 library spectra of the $K = 6$ materials are split in two disjoint halves. The first builds the class averages ($\hat{r} = 0.3422$, $M = 34$ spectra in the smallest class, per-entry standard deviation 1.105), the second the queries. Training uses 2500 mixtures of $L \in \{1, 2, 3\}$ materials drawn from the prototype half, evaluation 300 mixtures per order from the held-out half, with equal abundances.

Samson. The prototypes come from the spectral library distributed with the scene and the queries are the measured pixels themselves ($\hat{r} = 0.1317$). Those library spectra are themselves near-pure pixels of the cube, as Appendix D.5 sets out, so what is held out at step zero is the mixing and not the scene. The pixels whose ground-truth abundances exceed 0.10 in at least one material are split $70/30$ into a training and a held-out set; the held-out set carries 1291 pixels at $L = 1$, 1333 at $L = 2$ and 84 at $L = 3$, of which 25, all at $L = 1$, are library pixels.

OpenMIC. $q = 400$. The weights are the class averages of the excerpts in which an instrument is annotated present, restricted where possible to the excerpts whose only annotated-present instrument is that one, which is the closest thing this dataset has to an isolated source; the training items are the training-split excerpts with at least one annotated positive, and the label tuple of an item is its set of annotated positives, so L varies from item to item and the clamp uses g_L at that L. The score is the macro average precision on the test split, over the annotated pairs only, whose prevalence is 0.447. This is the only domain where supervised CD-1 makes the planted machine worse rather than better, by 0.007 of macro-AP over $8 \cdot 10^3$ steps, small but systematic across the three seeds, 0.008 on average over the four rates. Theorem 1 says nothing about the sign or size of that drift, and we offer only a conjecture: the prototypes here are averages of mixtures in which the class happens to be present rather than of isolated sources, so the positive phase is clamped to a label the visible layer only partly supports, and the two phases balance at weights that are not the ones we planted.

F THE SECOND EXPERIMENT: THE ONE-BIT READ-OUTS IN FULL

Figure 3 of Section 4 shows the micro-F1 of the six read-outs when each is told how many sources to return; Table 3 gives that comparison with exact match as well, and Table 4 repeats it with the parameter-free largest-gap rule for the three read-outs that can run without the order. Means over three seeds on NSynth and OpenMIC and five on the hyperspectral settings; the largest standard deviation over all cells of Table 3 is 0.044.

Every read-out sees the same queries and the same N-bit codes. Write $\boldsymbol{\sigma} \in \{-1, 1\}^N$ for the code of the query and $\boldsymbol{\Phi} \in \mathbb{R}^{N \times K}$ for the sensing matrix restricted to the class archetypes, so that column μ is $\boldsymbol{\Omega}\boldsymbol{T}\boldsymbol{e}_\mu$ and a noiseless equal-energy mixture of A has code $\operatorname{sign}(\boldsymbol{\Phi}\mathbb{1}_A)$. The machine scores $\boldsymbol{W}^\top\boldsymbol{\sigma}/N$ with $\boldsymbol{W}$ the class average of the codes of the prototypes; hyperdimensional bundling scores $\operatorname{sign}(\boldsymbol{W})^\top\boldsymbol{\sigma}/N$, the majority vote a hypervector memory stores (Kanerva, 2009); the archetype filter scores $\boldsymbol{\Phi}^\top\boldsymbol{\sigma}/N$ with the columns normalised, which is the linear estimator of Plan & Vershynin (2013) and needs the prototypes before the sign, something the binary machine cannot build for itself. BIHT iterates $\boldsymbol{a} \leftarrow H_L(\boldsymbol{a} + \frac{2}{N}\boldsymbol{\Phi}^\top(\boldsymbol{\sigma} - \operatorname{sign}(\boldsymbol{\Phi}\boldsymbol{a})))$ with renormalisation, 120 steps from $\boldsymbol{a} = 0$, which is the scheme of Jacques et al. (2013); its first step is the linear estimator, so it is that estimator plus the correction the one-bit literature adds to it. OMP and LASSO run on $(\boldsymbol{\Phi}, \boldsymbol{\sigma}/\|\boldsymbol{\sigma}\|)$, OMP with the support size set to the order and LASSO with a penalty equal to 0.1 times the smallest one that

zeroes every coefficient, computed per query so that nothing is tuned on the queries being scored. On Samson the pixels of each order are subsampled to 1200 per seed and on OpenMIC the excerpts to 1200, which is what makes the iterative methods affordable at $N = 16384$ on the hyperspectral settings; the OpenMIC read-outs run at $N = 4096$, as the contrastive-divergence experiments do. On OpenMIC the annotation is partial, so each excerpt is scored on its annotated entries only, the target is the set of annotated positives, and only excerpts with at least two annotated negatives enter, which is what makes a set prediction well posed. One control is run outside the one-bit budget, to check that the budget is not what decides the comparison: replacing the sign code by the real whitened features, with the same class means as prototypes, moves the archetype filter by at most 0.011 of micro-F1 over the four settings and at most 0.003 at every order on Urban.

The six do not cost the same. Table 2 measures the wall clock of each per query at $N = 16384$, $K = 6$ and $L = 3$, single threaded and in double precision, and counts the arithmetic each performs. The three filters are one inner product per class, $2NK$ operations; BIHT is two such products per iteration for 120 iterations; OMP is one correlation pass per selected atom and LASSO one per coordinate-descent sweep.

Table 2: Cost per query of the six read-outs of Figure 3, at $N = 16384$, $K = 6$, $L = 3$, median over three repetitions of a batch of 200 queries on one core. The three filters are evaluated as a single matrix product over the batch, while OMP and LASSO call a solver once per query, so their wall clock carries an overhead their arithmetic does not; BIHT is batched like the filters and its factor is arithmetic.

read-out	operations per query	wall clock (μs)	relative
the machine	$2NK = 1.97 \cdot 10^5$	17.9	1.0
HD bundling	$2NK$	19.3	1.1
archetype filter	$2NK$	18.9	1.1
BIHT, 120 steps	$480NK = 4.72 \cdot 10^7$	9979	557
OMP, support L	$2NKL = 5.90 \cdot 10^5$	1378	77
LASSO on codes	$2NK$ per sweep	1103	62

BIHT and OMP take the number of sources as an input, so the largest-gap rule is not defined for them; LASSO's penalty does not, but we keep it with the other two reconstructions, and Table 4 covers the three filters alone. It says two things. Choosing the cardinality costs more where K is largest, taking the machine on NSynth from 0.627 to 0.397 at $L = 5$ against 0.009 on Samson at $L = 2$, and it is what sinks the filters below the trivial level wherever that level is high, which on these four settings means wherever L is a large fraction of K: 0.558 against 0.800 on Samson at $L = 2$, 0.790 against 0.800 on Urban at $L = 4$ and 0.776 against 0.909 on Urban at $L = 5$, where five sources out of six make "return everything" almost optimal by construction. Told the order, the machine is ahead of every other read-out at both Urban orders in Table 3, and the ranking is the same one the gap rule cuts, so the whole of the difference is the cost of choosing the cardinality. And the ordering between the three filters is not the one of Table 3: under the largest-gap rule the archetype filter is ahead of the machine at every order on Urban in micro-F1, though not in exact match, so the crossing of Section 4 is a statement about the ranking a read-out produces and not about the sets a parameter-free rule extracts from it.

G THE THEORY ON MEASURED SIGNALS

None of the hypotheses of Theorem 1 holds on a chord or on a hyperspectral pixel, so what the theory is worth here is a matter of measurement. This appendix measures it in three places, which Figure 6 collects: the law of Lemma 2 at bit level, the band of Corollary 1 at class level, and the amplitude at which the learning rule balances.

G.1 THE LAW AT BIT LEVEL

Panel (a) of Figure 6 confronts (20) with the measurement. Evaluated at the correlation of the recordings themselves it is accurate to $1.7 \cdot 10^{-3}$ on NSynth and $1.7 \cdot 10^{-4}$ on Urban, one to two decades better than the idealised g_L, which is the upper envelope of the data because real sources are neither orthogonal nor of equal energy and (20) absorbs the difference through κ_μ.

On NSynth the measurement uses $N_Q = 250$ queries per order, seed 21 and $L \in \{1, 3, 5, 7, 9\}$; against the idealised g_L the largest gap is $2.3 \cdot 10^{-2}$, an order of magnitude worse than the parameter-

Table 3: Order given to every read-out: micro-F1 and exact match. The three reconstructions are behind the machine on Urban, on Samson at $L = 1$ and on OpenMIC, and ahead of it on NSynth, where at least two of the three lead at every order, and on Samson at $L = 2$; the archetype filter is ahead of the machine while the query is shallow and behind once it is deep on Urban, most visibly in the exact match, and ahead of it at every order on NSynth.

	NSynth, $K = 49$				Urban, $K = 6$					Samson, $K = 3$		OpenMIC, $K = 20$	
L	1	2	3	5	1	2	3	4	5	1	2	1	2
micro-F1													
the machine	.755	.625	.602	.627	.951	.908	.886	.891	.915	.981	.567	.560	.653
archetype filter	.768	.627	.616	.640	.985	.942	.886	.875	.875	.995	.559	.556	.659
HD bundling	.757	.630	.602	.632	.953	.906	.884	.879	.897	.979	.574	.552	.639
BIHT	.768	.630	.616	.640	.943	.791	.717	.765	.864	.781	.730	.398	.533
OMP on codes	.768	.638	.618	.625	.943	.878	.808	.793	.883	.780	.741	.398	.528
LASSO on codes	.765	.628	.616	.632	.929	.867	.828	.823	.891	.780	.779	.408	.546
exact match													
the machine	.755	.332	.142	.035	.951	.819	.663	.575	.573	.981	.133	.560	.405
archetype filter	.768	.342	.153	.045	.985	.885	.660	.503	.377	.995	.118	.556	.404
HD bundling	.757	.347	.142	.035	.953	.815	.658	.531	.487	.979	.147	.552	.384
BIHT	.768	.348	.153	.045	.943	.591	.205	.133	.319	.781	.461	.398	.236
OMP on codes	.768	.393	.178	.050	.943	.792	.575	.386	.413	.780	.481	.398	.224
LASSO on codes	.765	.378	.175	.053	.929	.765	.625	.464	.457	.780	.558	.408	.232

Table 4: The three filters with the parameter-free largest-gap rule, which is not told the order, and the trivial decoder that returns every class, whose micro-F1 is $2L/(K + L)$; on OpenMIC it returns every annotated class, since only the annotated entries are scored, which is why its level there is higher than that formula. BIHT and OMP do not appear because the order is an input to both.

	NSynth, $K = 49$				Urban, $K = 6$					Samson, $K = 3$		OpenMIC, $K = 20$	
L	1	2	3	5	1	2	3	4	5	1	2	1	2
micro-F1													
the machine	.648	.571	.500	.397	.895	.824	.792	.790	.776	.913	.558	.547	.623
archetype filter	.672	.593	.529	.432	.952	.847	.801	.817	.835	.915	.580	.549	.636
HD bundling	.649	.557	.504	.417	.886	.838	.808	.817	.811	.914	.569	.545	.629
predict every class	.040	.078	.115	.185	.286	.500	.667	.800	.909	.500	.800	.435	.623
exact match													
the machine	.662	.098	.023	.002	.891	.445	.259	.201	.211	.810	.034	.351	.197
archetype filter	.623	.107	.025	.003	.943	.556	.241	.141	.175	.815	.033	.339	.203
HD bundling	.663	.085	.023	.003	.889	.521	.287	.207	.233	.814	.041	.337	.177

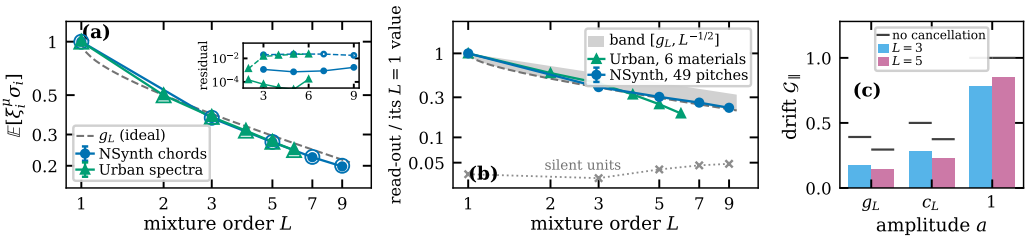


Figure 6: Physical superpositions, with nothing fitted. *(a)* Overlap between the code of a source and that of a mixture containing it: filled markers measured, open markers (20) evaluated at the correlation of the recordings, dashed line the idealised g_L; *inset*, the residual of the two predictions. *(b)* Class-level read-out of a constituent unit in units of its own $L = 1$ value, against the band (21). *(c)* Residual drift (59) at the three candidate label amplitudes of Remark 1; the horizontal marks are the value the drift takes when the negative phase cancels nothing.

free prediction, which is the content of the inset. The same measurement on real chords, major and minor and the sevenths, ninths and thirteenths, whose constituents share harmonics and whose archetypes are therefore not orthogonal, gives levels of the same size. On the Samson image, on mixtures nobody assembled, (20) reproduces the measurement over all $3 \times 9025 = 27\,075$ pixel and material pairs with mean absolute error $4.5 \cdot 10^{-3}$ and correlation 0.999. Table 5 gives the law and the band together on Urban.

Table 5: Urban, $K = 6$. Left: the arcsine law at bit level, measured against the parameter-free prediction (20) and against the idealised g_L. Right: the class-level read-out ratio of Corollary 1 against the lower edge of the band (21); the measurement leaves the band from below at $L \geq 4$, marginally at $L = 4$ and by 28% at $L = 6$, which is the signature of six strongly non-orthogonal endmembers.

L	measured	Eq. (20)	g_L	$R(L)$	g_L
1	1.0000	1.0000	1.0000	1.000	1.000
2	0.4987	0.4988	0.5000	0.583	0.500
3	0.3770	0.3770	0.3918	0.442	0.392
4	0.3153	0.3153	0.3333	0.324	0.333
5	0.2729	0.2730	0.2952	0.248	0.295
6	0.2450	0.2451	0.2677	0.193	0.268

Both hypotheses of Lemma 2 that a real scene violates can be violated on purpose, and the law survives both while the margin does not: drawing the abundances from a Dirichlet(1) distribution instead of keeping them equal takes the Urban micro-F1 at $L = 3$ from 0.806 to 0.651, and Gaussian sensor noise at 20 dB takes it to 0.663, in both cases by eating the separation between active and silent read-outs and not by moving the law.

G.2 The band at class level

Panel (b) moves to the hidden layer, where the band (21) replaces the point prediction. The NSynth read-out sits on the lower edge, within 5.1% of g_L, the silent units staying between $+0.024$ and $+0.036$ against an active read-out of 0.743 at $L = 1$ and 0.167 at $L = 9$. The Urban read-out is inside the band up to $L = 3$ and falls below it at $L \geq 4$, six material spectra in 162 bands being far from orthogonal. Binning the Samson pixels by the number of materials present gives ratios 1.000, 0.216, 0.319 at $L = 1, 2, 3$, which fall faster than g_L and leave the band as soon as $L = 2$, because the abundances of a real pixel are very far from equal.

The two panels separate what the encoder does from what the archetypes are. The per-pixel prediction (20), which uses the actual abundances through κ_μ, needs no hypothesis on them and shows no discrepancy at all; the band needs equal energy, and where that fails it is the band that fails. This is the failure Section 4 reads on the decoder side in Figure 3(c): the equal-energy hypothesis of Lemma 2 is what fixes the constant, and where it is worst the decoder built on the constant is worst too.

G.3 The amplitude at which the rule balances

Panel (c) measures the object Theorem 1 is about, on measured chords, through the component the theory predicts to vanish, the projection of the mean update on the presented cue,

$$\mathcal{G}_{\parallel} := \Big\langle \frac{s_i \, \mathbb{E}[\Delta w_i^{\mu}]}{\epsilon \beta} \Big\rangle_{i,\ \mu \in A}, \tag{59}$$

which, being signed, locates the amplitude at which the two phases balance rather than only bounding the distance from it. It uses the closed-form reduction of Lemma 7 at $\beta = 20$ and $N = 4096$, averaged over 150 measured chords per order, with the positive phase clamped to the code of the real chord and the label to $\hat{z}_\mu = \sqrt{N}\, a\, \mathbf{1}\{\mu \in A\}$, and it has to be read against the value it takes when the negative phase cancels nothing, which is the label amplitude a itself.

At $L = 3$ the measured drift is 0.171 at the clamp $g_L = 0.392$, 0.279 at $c_L = 0.5$ and 0.779 at the binary clamp, that is 56%, 44% and 22% of the positive phase cancelled, and at $L = 5$ the same ordering with 51%, 40% and 15%. Of the three candidates g_L is therefore the one that leaves the smallest residual, which is the ranking Section 4 reports. The rule does not balance exactly at it: scanning a over a grid containing the three candidates and the measured read-out gives a single zero crossing per order, at $a = 0.219$ for $L = 3$, against a read-out level of 0.307 on those same chords, and that gap is the third of the three deviations Section 4 measures. The same measurement on Urban spectra, with the random-weight reference computed rather than argued, gives $\mathcal{G}_{\parallel} = 0.450$ against 0.632 at zero weights, a cancellation of 29% with g_L against 27% with c_L and 18% with the binary clamp, each read against its own label amplitude and in the same order as on chords; the cancelled

fractions are smaller than on NSynth because $\hat{r} = 0.34$ on Urban against 0.78, so the class averages are noisier estimates and the reconstruction they drive is correspondingly worse.

Two remarks on why this is the statistic reported. The maximum $\mathcal{G} = \max_{i,\mu} |\mathbb{E}[\Delta w_i^\mu]|/(\epsilon\beta)$, which is the statistic Theorem 1 bounds and Figure 4 verifies, is taken over NK entries and therefore follows the largest single fluctuation rather than the systematic part of the update; two initialisations that differ in the systematic part are separated by the projection on the cue. And the amplitude cannot be read off the score instead: along $8 \cdot 10^3$ CD-1 steps the three planted runs are indistinguishable in micro-F1, within 0.005 of each other at every checkpoint. Standard deviations quoted in this appendix are population values over the runs.

The proof of Theorem 1 assumes, through Lemma 9, that the reconstruction driven by the label reproduces the presented cue up to the vote reliability ϑ_B, and on real chords it does not: the agreement $\langle \boldsymbol{s}, \tanh(\beta \boldsymbol{h}) \rangle / N$ is 0.657 at $L = 1$, 0.443 at $L = 3$ and 0.386 at $L = 5$. The balance point moves with it. The zero crossing above, at 0.219 against a read-out level of 0.307, is a ratio of 0.71 that acts as an effective ϑ for measured audio, and it is the same partial reconstruction measured in a second place. Closing the identity between the two needs the cross term $N^{-1} \sum_j \bar{\eta}_j^\mu \tanh(\beta h_j)$ of Lemma 7, which we do not measure here. The class averages therefore remain a fixed point of supervised CD-1 on real data, but an approximate one, and the residual is neither noise nor a pure gain change: rescaling the weights does not remove it.

G.4 What the bound is worth numerically

The left-hand side of (17) is $O(\epsilon\beta)$ by construction: every entry of the update is a difference of two terms bounded by r^{-1} times the label amplitude, so a right-hand side above $\epsilon\beta$ says nothing about the size of the drift. Table 6 evaluates the four channels at the parameters of the runs reported in the paper, with $\gamma = \sqrt{2 \ln(2KTN)/N}$ at the nominal $N = 16384$ of Appendix D.2.

Table 6: The four channels of (17) in units of $\epsilon\beta$, each at the θ that minimises their sum. The dataset channel dominates everywhere, because $c_1(\theta)/\rho$ is of order one at every ρ these experiments run at. The infimum matters where the two exponential channels are comparable: on NSynth it halves the total against the symmetric $\theta = 1/2$, which gives 37.5 there, 170.1 on Urban and 4.9 on the fifth row. The last block shows where the bracket falls below one; with the infimum this happens at $\rho = 10^{-3}$ rather than 10^{-4}, and it still needs both a clean estimator and a mini-batch, neither of which is reached on measured data.

setting	thermal	dataset	mini-batch	finite size	total	$\theta^\star$
NSynth, $r = 0.78$, $L = 3$, $\beta = 8$, $\rho = 0.007$, $B = 1$	6.07	9.52	1.64	0.054	17.3	0.82
NSynth with the effective $\rho = 17.8$ of App. D.4	0.01	76.92	1.64	0.054	78.6	$\to 0$
Urban, $r = 0.34$, $L = 3$, $\beta = 8$, $\rho = 0.222$, $B = 1$	6.27	157.34	1.96	0.043	165.6	0.72
illustrative point, $r = 0.9$, $\beta = 20$, $\rho = 0.01$, $B = 11$	2.71	28.64	0.052	0.089	31.5	0.90
$r = 0.9$, $L = 3$, $\beta = 20$, $\rho = 10^{-3}$, $B = 11$	0.14	0.18	0.052	0.089	0.47	0.75
$r = 0.9$, $L = 3$, $\beta = 20$, $\rho = 10^{-4}$, $B = 11$	0.00	0.00	0.052	0.089	0.14	0.38
$r = 0.9$, $L = 3$, $\beta = 20$, $\rho = 10^{-4}$, $B = 101$	0.00	0.00	0.000	0.089	0.09	0.38

Two consequences, both stated in Section 2. First, the content of Theorem 1 is the decomposition and the parametric form of the four channels, which Figure 4 switches off one at a time and finds correct, and not a numerical certificate on the drift. Second, the constants are not tight: the measured drift on that grid is one to two decades below the bound, and only the finite-size channel is captured closely, because it is the only one that comes from a concentration inequality applied to the quantity that actually fluctuates rather than to a chain of worst cases. Sharpening the dataset channel, which means replacing the Hoeffding step of Lemma 4 by a bound that uses the variance ρ rather than the range of $\bar{\eta}_i^\mu$, is the first thing a tighter version of the theorem needs, and the effective dimension measured in Appendix D.4 has to enter it as well.