EB-gMCR: Energy-Based Generative Modeling for Signal Unmixing and Multivariate Curve Resolution

Anonymous authors
Paper under double-blind review

Abstract

A measurement of a mixture is the sum of a few component profiles, each weighted by a concentration, and multivariate curve resolution (MCR) recovers both from many samples. Classical MCR factorizes the data matrix, takes the component count as input, and leaves a rotational ambiguity. This paper instead models the generative process, generative MCR (gMCR): each sample selects a few components from a fixed set and is observed as their linear superposition plus noise. We prove that the data determine which decomposition a solver should reach. Under a spark condition on the component profiles, the decomposition of minimal usage is the true one, up to a relabeling of the components, and under noise the same holds up to a stated conjecture. Components recovered this way decode new samples of the same process. EB-gMCR, an energy-based solver, is one realization of gMCR: a gate selects components per sample and is trained toward minimal usage. With no count given, it recovers the component count to within a few percent on synthetic mixtures of up to 256 components, at a sample-to-component ratio of eight and a signal-to-noise ratio of 30 dB. On two public spectroscopy datasets it recovers the count exactly and reconstructs as well as the best method at that count. The trained model, frozen, decodes mixtures synthesized across the concentration simplex at the noise floor. gMCR applies to any sparse linear mixture of a fixed set of components. The results cover a solver whose candidate pool is at least as large as the true component count.

1 Introduction

A single measurement of a complex sample, such as a reaction mixture, a natural extract, or a tissue, records many chemical constituents at once. No part of the measurement belongs to one constituent alone. Determining which constituents are present, and at what concentrations, requires separating the measurement into one contribution per constituent. Each constituent, hereafter a component, contributes a fixed component profile. The profiles add in proportion to the concentrations, and the observed sample is a linear superposition of fixed component profiles. Unmixing recovers the composition from the measurement.

In chemistry this task is multivariate curve resolution (MCR) (de Juan & Tauler, 2021): given M samples with d features, $D \in \mathbb{R}^{M\times d}$, find N component profiles $S \in \mathbb{R}^{N\times d}$ and their concentrations $C \in \mathbb{R}^{M\times N}$ with

$$D = CS + \varepsilon, \tag{1}$$

where ε is measurement noise. Solved as an inverse problem, the factorization is ill-posed: the component profiles and the concentrations are both unknown, and any invertible T yields CT and $T^{-1}S$ with the same fit. Classical methods compute the factorization directly. They update C and S in turn, under constraints such as non-negativity, closure, and unimodality, until CS reproduces D (de Juan & Tauler, 2021). The factorization route has three limitations. First, the component count N is an input. The count is unknown and is often the quantity of interest. Second, the constraints describe the recovered profiles and concentrations, not how a mixture is made: which components each sample contains is never a variable of the factorization. Third, a factorization describes one data matrix. The feasible set depends on which samples the matrix contains (Tauler et al., 1995; de Juan & Tauler, 2021). A new sample from the same process requires a new factorization.

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.

This paper models how a mixture is made. Each sample contains few components, while the candidates number in the hundreds (Barcaru et al., 2017; Heinecke et al., 2021). The generative MCR model (gMCR, §3) fixes a pool of component profiles across samples. Each sample is the linear superposition of a few components drawn from the pool, plus noise. The contributions are:

- **An identifiable solution exists.** Under a spark condition and diverse occurrences, the decomposition of minimal usage, the one selecting the fewest components in total, is the truth up to a relabeling of the pool. Its pool decodes new samples (Theorems 1(a), 2(b), Corollary 1).
- **Conditions for reaching it.** A solver needs an objective that holds this solution under noise. With the usage penalty inside a stated window, each true support is a strict local minimizer with high probability. Fit and usage conditions certify a decomposition (Theorems 1(b) and 2).
- **A solver that realizes it.** EB-gMCR trains an energy-based gate toward minimal usage. The component count is an output of the solver, not an input (§5, §6).

2 Related Work

Prior work on the factorization of Eq. 1 differs in which of its unknowns are given. Sparse coding and dictionary learning are the cases of MCR in which the pool, or the pool's size, is given.

Sparse coding and dictionary learning. Sparse coding recovers a sparse concentration vector given the pool. Under the spark condition (no small set of pool profiles is linearly dependent) the sparsest vector is unique (Donoho & Elad, 2003). Convex minimization recovers it under incoherence (Donoho & Elad, 2003) and, with high probability, for random measurement ensembles (Candès et al., 2006; Donoho, 2006). Dictionary learning also recovers the pool, its size fixed. A sparsely used pool is identifiable (Aharon et al., 2006; Spielman et al., 2012). Identification is stable under noise (Garfinkle & Hillar, 2019). Sample-complexity bounds control how far the empirical objective deviates from its expectation (Gribonval et al., 2015).

Non-negative factorization and MCR. NMF takes the component count as an input. It is identifiable under separability (each component occurs alone in some sample or feature) or under sufficiently scattered factors (Donoho & Stodden, 2003; Fu et al., 2019). MCR is this factorization under the chemical constraints of §1, with the count given. The MCR solution set was first described for two components (Lawton & Sylvestre, 1971). Under non-negativity alone it is a continuum (Sawall et al., 2020). It reduces to a single factorization, up to scale, when every component has a selective window, samples or features where it alone is present (Tauler et al., 1993; 1995), or under local-rank conditions on sample windows (Tauler et al., 1995). For the gMCR model of §3, with the count unknown, Theorem 1 gives identifiability.

Model-order selection. Model-order selection estimates the count before the fit, from the eigenvalue spectrum of the data. The indicator function locates the number of significant factors in that spectrum (Malinowski, 1977). Evolving factor analysis repeats the eigenvalue analysis on submatrices that grow one sample at a time, and a new component is counted where a new significant eigenvalue appears (Maeder, 1987; Keller & Massart, 1991). ARD-NMF (automatic relevance determination) starts from an oversized factorization and reduces unused components to zero. The returned count depends on the hyperparameters of its prior, among them a noise dispersion that is fixed in advance (Tan & Févotte, 2013). The solvers compared in §6 take the count as input (Tauler et al., 1995; Lee & Seung, 1999; Hoyer, 2004; Schmidt et al., 2009; Comon, 1994).

Selection as a modeled variable. Three lines of work model selection as a binary variable per sample and component, inferring the number in use. Spike-and-slab sparse coding draws each coefficient as exactly zero (spike) or continuous (slab) (Goodfellow et al., 2012). Nonparametric Bayesian dictionary learning does not fix the pool size (Zhou et al., 2009). Bayesian library search in chemistry scores which known profiles a mixture contains (Barcaru et al., 2017).

3 THE GENERATIVE MCR MODEL

The gMCR formulation recasts MCR as a generative process: each sample selects a small set of components from a pool, draws their concentrations, and is observed as their linear superposition plus measurement noise. Only the samples are observed; the pool, the supports, and the concentrations are all latent and are the objects of inference. The model serves to decode samples, not to synthesize them. This paper analyzes linear superposition throughout.

The explicit modeling of the selection indicator δ, defined formally in the model below, is central to gMCR. Classical MCR implicitly encodes selection through concentration magnitude: a component is "absent" when its concentration is near zero. In gMCR, δ explicitly separates selection (which components participate) from intensity (how much each contributes), enabling discrete component discovery rather than soft shrinkage toward zero.

The gMCR model. A pool is a set of unit-norm component profiles $s_1, \ldots, s_N \in \mathbb{R}^d$, fixed across samples. A sample, indexed by a draw ω of the generative process, is

$$x(\omega) \;=\; \sum_{i \in J(\omega)} c_i(\omega)\, s_i \;+\; \varepsilon(\omega), \qquad |J(\omega)| \le k, \quad c_i(\omega) \in [c_{\min},\, c_{\max}] \;\text{ for } i \in J(\omega), \tag{2}$$

where $J(\omega) \subseteq [N] = \{1, \ldots, N\}$ is the sample's support ($\delta_i = \mathbf{1}[i \in J]$ recovers the indicator above) and the noise $\varepsilon(\omega)$ is sub-Gaussian with proxy σ (Appendix B). Each constraint has a physical reading: k caps how many components any single sample contains (few analytes per specimen); $c_{\min}$ is the detection floor, so that active means detectably active; $c_{\max}$ is the dynamic range; unit-norm profiles fix the scale ambiguity between a profile and its concentration; sub-Gaussian ε is instrument noise with controlled tails. This constraint set is the mathematical form of standard assumptions in spectroscopic unmixing: linear detector response, a detection floor, finite dynamic range, and few analytes per sample. Any domain satisfying them inherits every result of this paper.

Definition 1 (Spark condition with margin)**.** The pool satisfies the *spark condition with margin* $s_{\min} > 0$ if $\sigma_{\min}(S_A) \ge s_{\min}$ for every $A \subseteq [N]$ with $|A| \le 2k$, where S_A collects the profiles indexed by A as columns.

In words: no group of up to $2k$ profiles can nearly cancel; small subsets of the pool are not merely linearly independent but independent with a quantitative margin. This is the quantitative form of the spark condition of Donoho & Elad (2003, Definition 1). Here $\sigma_{\min}(S_A)$, the smallest singular value, is computed from the pool, whereas $s_{\min}$ is an assumed constant of the model: the condition asserts that the computed quantity never falls below the assumed margin. In particular the condition requires $2k \le d$; it constrains the per-sample sparsity against the feature dimension, not the pool size N, which may exceed d.

Proposition 1 (The decomposition is well-defined)**.** *Assume the spark condition with margin $s_{\min}$. Let $x_0 = \sum_{i \in J} c_i s_i$ with $|J| \le k$ and $c_i \ge c_{\min}$ on J. If a pair (J', c') with $|J'| \le k$ and $c'_i \ge c_{\min}$ on J' satisfies $\left\| x_0 - \sum_{i \in J'} c'_i s_i \right\| \le \epsilon$ for some $0 \le \epsilon < c_{\min}\, s_{\min}$, then $J' = J$ and $|c'_i - c_i| \le \epsilon / s_{\min}$ for every $i \in J$. In particular, at $\epsilon = 0$ the k-sparse decomposition of a noiseless sample is unique among decompositions with coefficients at least $c_{\min}$.*

The proof is in Appendix A. The tolerance ϵ is a deterministic number, distinct from the model noise $\varepsilon(\omega)$. The quantity $c_{\min} s_{\min}$, the weakest admissible concentration times the pool margin, is the threshold that governs every guarantee in this paper. Proposition 1 makes each sample's decomposition well-defined *given* the pool; the uniqueness of the pool itself is a separate question, taken up in §4.

Decoding a fresh sample over a known pool is exactly the setting of Proposition 1, which is what makes a recovered process reusable (Corollary 1); Appendix F works a toy example.

4 WHAT CAN BE KNOWN FROM MIXTURES: THE THEORY

In this section the data follow the gMCR model of §3 (Eq. 2) with the spark condition of Definition 1. The results below concern the model and its solution set, not any solver. The inverse problem has three unknowns: the pool, the support of each sample (the components that sample contains), and

the concentrations. Given the pool, each sample's decomposition is unique (Proposition 1). The remaining question is whether the samples determine the pool. Each support is a discrete set. The invertible changes of coordinates of §1 do not in general keep the supports or the constraints of Eq. 2. The continuum of factorizations they generate is not a continuum of gMCR solutions. Theorem 1 gives the quantity that identifies the truth: minimal usage. Theorem 2 gives the certificate: a returned decomposition that passes it is the truth up to a relabeling of the pool. Theorem 3 gives the limit as samples accumulate: the generative process is recovered, and some sample is decoded wrongly. One combinatorial statement remains open (Conjecture R, §4.2). Each theorem is stated under hypotheses given by reference. The appendices state every hypothesis in full and give the proofs.

4.1 A Criterion That Points at the Truth

The occurrence matrix $\delta \in \{0,1\}^{M\times N}$, with $\delta_{mj} = 1$ when component j serves sample m, encodes the supports. Total usage is $\sum_{m,j} \delta_{mj}$. Among all decompositions that reproduce the data exactly, the minimal-usage criterion selects the fewest occurrences. Compressed sensing rests on the same principle in a fixed basis. The sparsest representation is identifiable under the spark condition (Donoho & Elad, 2003, Corollary 1). It is findable from incomplete measurements (Candès et al., 2006; Donoho, 2006). gMCR poses the same question for a learned pool across many samples.

Noise makes exact fits impossible. The principle is expressed instead as a per-sample criterion. For a candidate selection $A \subseteq [N]$ and a sample x,

$$\ell_\lambda(A;\, x) \;=\; \min_{c\,\geq\, 0} \Big\| x - \sum_{j\in A} c_j s_j \Big\|^2 + \lambda\, |A|, \tag{3}$$

the best nonnegative fit plus a penalty λ per selected component. The penalty acts on the row of δ, not on the coefficient magnitudes: Eq. 3 is best-subset (ℓ_0-type) selection. Ridge and lasso (Hoerl & Kennard, 1970; Tibshirani, 1996) penalize $\|c\|$ with no cost on the selection count.

Theorem 1 (Identifiability at minimal usage)**.** *Assume the gMCR model (Eq. 2) with the spark condition (Definition 1).* (a) *For noiseless data, the true occurrence matrix $\delta^\star$ is the unique minimizer of total usage $\sum_{m,j} \delta_{mj}$ among all exact factorizations over the pool.* (b) *Set $L_\alpha = \sqrt{2\log(6NM/\alpha)}$. If $c_{\min} s_{\min} > 2\sigma L_\alpha$ and $\sigma^2 L_\alpha^2 < \lambda < (c_{\min} s_{\min} - \sigma L_\alpha)^2$, then with probability at least $1-\alpha$, simultaneously for every sample, the true support is a strict local minimizer of $\ell_\lambda(\,\cdot\,;\, x^{(m)})$ under single-component changes; at $\sigma = 0$ it is the unique global minimizer among selections with $|A| \leq k$.*

Theorem 1 says that an added component can improve the fit only by absorbing noise, a gain below λ. Removing a true component increases the fit error by more than λ. Between these two bounds the true support is a strict local minimizer, sample by sample. The formal statement and proof are in Appendix B.

4.2 The Structure of the Solution Set

Write $K^* = |\bigcup_m J_m|$ for the true component count, the number of distinct components that occur across the samples. The per-sample cap k is a different quantity. K^* is unknown. A count fixed at any other value excludes the true decomposition.

Three facts describe the set of decompositions the data admit. The data do not determine the order of the pool slots: renumbering the slots, and the columns of δ to match, produces the same samples. Components are recovered at best up to relabeling. When the supports rarely distinguish one component from another, the data admit a continuum of decompositions. Appendix F computes one such continuum. When occurrences are diverse, the set reduces to the truth up to relabeling.

The certificate is a pair of conditions on a decomposition:

$$\Big\| x^{(m)} - \sum_{j\in A_m} \hat{c}_j^{(m)} s_j \Big\| \leq \varepsilon_{\text{fit}} \;\text{ and }\; |A_m| \leq k \;\text{ for every } m, \qquad \sum_m |A_m| \;\leq\; \sum_m |J_m|. \tag{4}$$

The first condition is on fit, the second on usage.

Theorem 2 (Certified recovery)**.** (a) *Over a known pool, every solution passing (4) is the truth: $A_m = J_m$ for every sample, hence $|\bigcup_m A_m| = K^*$, with concentrations correct to the accuracy the noise permits.* (b) *Over a learned pool without noise, every solution passing the learned-pool check is the truth up to a relabeling π of the pool slots: $\hat{s}_{\pi(i)} = s_i$, $A_m = \pi(J_m)$, $\hat{c}^{(m)}_{\pi(i)} = c^{(m)}_i$; the count again equals K^*.* (c) *Under noise the same holds up to Conjecture R, stated below; one component per sample needs no conjecture.*

Part (b) of Theorem 2 adds conditions on the returned pool (Appendix C) and relies on a pinning step: a pair of samples sharing exactly one component pins that component to one pool slot. Under noise the step needs a further condition. The proportions in the two samples must differ enough that no coordinated error imitates the component across both. That requirement is the open statement. Fix $\gamma_c > 0$. Call a sample m *diversely served* if some samples $m_1, \ldots, m_{|J_m|}$ sharing its support and candidate set satisfy

$$\sigma_{\min}\Big(\big[\, c^{(m_1)} \;\cdots\; c^{(m_{|J_m|})} \,\big]\Big) \;\geq\; \gamma_c.$$

Conjecture R (isolating pairs)**.** *For some $\gamma_c > 0$: every active component i admits samples m, m' with $J_m \cap J_{m'} = \{i\}$, both diversely served at level γ_c.*

Whether the pairs exist is computed from a returned solution, from its supports and concentrations. When they are present, part (c) holds for that solution with nothing assumed. The fit condition is computed in the same way. The usage condition compares the total usage of the solution with that of the truth. The truth is unknown. No computation on the data and the solution can evaluate the usage condition. Theorem 1 applies to this condition: inside the λ-window the criterion favors minimal usage. The true decomposition passes the certificate. The formal version is in Appendix C.

Corollary 1 (Reuse)**.** *Let the pool be the components recovered by part (b), retained after training, and let $x = \sum_{i \in J} c_i s_i + \varepsilon$ be a fresh sample of the same process, with J among the recovered components. Then the check of (4), applied to x alone, again returns the truth: $A = J$, with concentrations correct to the accuracy the noise permits.*

Corollary 1 separates learning the process from fitting one dataset. The recovered pool is a component library. A new sample is decoded by itself, under the same check, with nothing re-trained. The proof is in Appendix C, with the corollary restated there.

The last result takes the sample count M to infinity.

Theorem 3 (Impossibility and consistency)**.** (a) *As the number of samples M grows, no estimator decodes every sample correctly: for every M and every measurable map from the samples to supports $(\hat{A}_1, \ldots, \hat{A}_M)$,*

$$\Pr\big(\hat{A}_m = J_m \text{ for all } m \leq M\big) \;\leq\; (1-q)^M, \qquad q \in (0,1) \text{ independent of } M.$$

(b) *With a known noise law and diverse occurrences, the estimated process $\hat{G}_M$ converges almost surely, up to relabeling, to the true process G^*, and the estimated count $\hat{N}_M$ is eventually exactly K^*:*

$$d_G\big([\hat{G}_M], [G^*]\big) \;\longrightarrow\; 0 \quad \text{a.s.}, \qquad \hat{N}_M = K^* \text{ for all sufficiently large } M \quad \text{a.s.},$$

where d_G is the distance between processes modulo relabeling of Appendix D.

Part (a) holds because the largest noise among M independent draws grows like $\sqrt{2 \log M}$. Eventually the noise on some sample exceeds its margin. Part (b) holds because, among the processes that reproduce the limiting data law, the usage-minimal one is the truth. Identification together with compactness gives convergence. The proofs are in Appendix D.

5 EB-gMCR: A Solver for the Generative Model

EB-gMCR is a realization of the gMCR model of §3. It infers the support, the component profiles, and the concentrations from the observed samples. The selection is computed by an energy-based gate. An energy is a scalar cost that a network assigns to a state, here to selecting or to not selecting a component for a sample. The gate is the map from the energies to the selection. §5.1 gives the gate. §5.2 gives the objective. §5.3 gives the rule for the states the solver returns. §5.4 gives the convergence of the selection during training. The algorithm and the hyperparameters are in Appendix H. The hyperparameters are fixed across every experiment.

5.1 EB-SELECT: HARD GATING FOR DISCRETE SELECTION

EB-select is the gate. It computes the discrete per-sample selection of Theorem 1: each entry of δ is one or zero. For each observed sample x, a neural network computes selection energies $E(x) \in \mathbb{R}^{N\times 2}$, one energy for selecting component j ($E_{j,1}$) and one for not selecting it ($E_{j,0}$). The selection is relaxed by the Gumbel-softmax reparameterization (Jang et al., 2017; Maddison et al., 2017). With $Z_{j,\ell} \sim \mathrm{Gumbel}(0,1)$ i.i.d.,

$$[f_e(x)]_j \;=\; \frac{\exp\big((Z_{j,1}-E_{j,1})/\tau\big)}{\sum_{\ell\in\{0,1\}}\exp\big((Z_{j,\ell}-E_{j,\ell})/\tau\big)} \;=\; \varsigma\Big(\frac{(E_{j,0}-E_{j,1})+(Z_{j,1}-Z_{j,0})}{\tau}\Big), \qquad (5)$$

the logistic function ς of the noisy energy gap. The temperature τ sets the sharpness of the gate. The relaxation $f_e(x) \approx \mathbb{E}[\delta \mid x]$ is differentiable in the energies. Training uses a positive τ. Evaluation uses a near-zero temperature. Component j is selected when its selection probability exceeds a fixed threshold. The threshold corresponds to a floor on the energy gap $E_{j,0} - E_{j,1}$. The threshold, the floor, and the gate's decision surface are reported in Appendix G. During training, Gaussian exploration noise is added to the energies before the gate:

$$E' \;=\; E + s_{\mathrm{ex}}\,\xi, \qquad \xi \sim \mathcal{N}(0, I), \qquad (6)$$

with a fixed scale s_{ex} (Appendix H). The noise acts on the selection distribution, not on the parameters.

5.2 OBJECTIVE: FIDELITY, PARSIMONY, DISTINCTNESS

Two further elements realize the remaining latent objects of Eq. 2. Profiles are learnable vectors $\mathbf{s}_i \in \mathbb{R}^d$, one for each of the N pool slots, normalized to unit norm. Concentrations are the non-negative outputs of a network $f_c : \mathbb{R}^d \to \mathbb{R}^N$ evaluated on the sample. The reconstruction $\hat{x}$ is the linear superposition of the selected profiles weighted by their concentrations.

The training loss is the differentiable form of the criterion of Eq. 3, with two regularizers:

$$\mathcal{L} \;=\; \|x-\hat{x}\|^2 + \lambda\, C(x) + \lambda_{\mathrm{se}}\|E\|_2^2 + \lambda_{\mathrm{amb}}\,\mathcal{R}(B), \qquad C(x) \;=\; \frac{1}{N}\sum_{i=1}^{N}[f_e(x)]_i. \qquad (7)$$

The usage term $C(x)$ is the selection count in soft form. As $\tau \to 0$ it tends to the number of selected components divided by N. The penalty $\lambda\, C(x)$ is the count penalty $\lambda|A|$ of Eq. 3 scaled by $1/N$. The weight λ is scheduled: it stays at zero until the normalized reconstruction error falls below a fixed level, then increases (Appendix H). The penalty serves Theorem 1. The schedule sets λ from the energy scale. The schedule does not place λ inside the λ-window of Theorem 1.

The term $\mathcal{R}(B)$ serves the spark condition. It penalizes similarity among the profiles in use in a minibatch B, the subset of samples processed in one optimizer step,

$$\mathcal{R}(B) \;=\; \frac{1}{|\mathcal{P}_B|}\sum_{\{j,k\}\in\mathcal{P}_B} K(\mathbf{s}_j, \mathbf{s}_k), \qquad (8)$$

with K a radial basis function (RBF) kernel, $\mathcal{I}_B$ the components selected for at least one sample of B, and $\mathcal{P}_B$ the pairs in $\mathcal{I}_B$. The term is modeled on the Hilbert–Schmidt independence criterion (Gretton et al., 2005b;a). Eq. 5 depends on each energy pair only through its gap. The regularizer $\lambda_{\mathrm{se}}\|E\|_2^2$ removes the free common shift of the pair, a shift of $E_{j,0}$ and $E_{j,1}$ by one common amount that leaves the gap unchanged (Appendix G).

5.3 CHECKPOINTING: THE CERTIFICATE

Under the gate's training noise, a run has no fixed point. The solver returns checkpoints. A checkpoint is a state of the model saved during training. A band is an interval of reconstruction quality measured by R^2. The bands are an input to the solver. Within each band the solver keeps the checkpoint with the smallest component count. The rule serves the usage condition of the certificate (Theorem 2). The fit condition is computed at every checkpoint. Non-negative concentrations and unit-norm profiles hold at every checkpoint by construction (§5.2). One run supplies a checkpoint for every band it has entered, whenever it is stopped. A known noise level sets the fit tolerance $\varepsilon_{\mathrm{fit}}$. The tolerance sets the band. Too strict a band returns no checkpoint. Too loose a band accepts states that have dropped true components. The bands used and the termination rule are in Appendix H.

5.4 Convergence: Stability After the Anchor

Proposition 2 states the convergence of a run from a step at which the gate has become decided (proof in Appendix E). The mean gap of component j is the dataset mean of its energy gap $E_{j,0} - E_{j,1}$ (§5.1), and $\hat{J}_t$ is the set of components with a positive mean gap at step t. The margin of a component is its mean gap, signed so that positive agrees with $\hat{J}_t$. A step at which every margin is at least $\gamma > 0$ is an *anchor*, computed from the gate's energies. The first phase of the proposition holds at any sample size, while the second holds under concentration conditions on the sample count.

Proposition 2 (Two-phase convergence)**.** *Let t_0 be an anchor at level γ, let $J_\sharp = \hat{J}_{t_0}$, and let the step size η be constant after t_0. Write θ_t for the parameters at step t and f for the objective on $J_\sharp$, with minimum $f^\star$. Under the conditions and with the constants of Appendix E, with $\rho = 1 - \mu\eta/2$ and probabilities conditional on the run up to t_0:*

(i) Inside the window. *For $T(\gamma) = \lfloor \gamma/(4L_\Delta G\eta) \rfloor$ steps after t_0, $\hat{J}_t = J_\sharp$ for every realization of the data and of the training noise, and for t in that window* $\mathbb{E}[f(\theta_t) - f^\star] \le \rho^{\,t-t_0}(f(\theta_{t_0}) - f^\star) + \Phi(\eta)$.

(ii) After the window. *At any fixed later step t,* $\Pr(\hat{J}_t \neq J_\sharp) \le p + \frac{1}{2}\,\rho^{\,t-t_0-T_{\mathrm{ctr}}}$, *with $p \le \frac{1}{2}$.*

6 Experiments

This section measures whether the solver of §5 reaches, on data, what §3 and §4 prove: the component count, the fit the certificate requires, and the reuse of a recovered process. Two quantities are reported throughout. The reconstruction R^2 summarizes the per-sample fit condition of the certificate over a dataset. The estimated component count (EC) is the number of components a solver returns, the quantity of Theorems 2 and 3(b). Every baseline except ARD-NMF takes the count as input. ARD-NMF chooses the count through its noise dispersion ϕ (§2).

6.1 Synthetic Mixtures

The synthetic mixtures (Appendix I) have known components: unit-norm non-negative profiles in $\mathbb{R}^{512}$, one to four active components per sample, a true count K^*, written N in this section, in $\{16, \ldots, 256\}$, $4N$ or $8N$ samples, at 20 and 30 dB. EB-gMCR starts from a pool of 1024 candidate components. The EB-gMCR count is read from one R^2 band (§5.3) per setting, the reported band, at which the baselines are read as well (Appendix J). Table 4 gives every band. Figure 1 reports the counts (a, b) and the reconstruction quality (c, d). Tables 2 and 3 give the numbers.

At 30 dB with $8N$ samples the count is within a few percent of the truth at every N (Table 4). With $4N$ samples the count is above the truth at intermediate N and below it at the smallest and largest N. At 20 dB the count exceeds the truth for $N \ge 96$ with $4N$ samples, and for $N \le 48$ and $N \ge 128$ with $8N$ samples. The window condition $c_{\min} s_{\min} > 2\sigma L_\alpha$ of Theorem 1 gives one reading of the two noise levels. At the edge of the window extra components fit noise, while no usage weight forbids them. At that edge the count can only exceed the truth. Whether either noise level is inside the window is not computed here. Runs at $8N$ end sooner (Table 6). Proposition 2 gives no rate in the sample count. Decoding further runs at $N \le 64$ ($4N$ samples) sample by sample gives their supports (Appendix K.6, Table 8). In four of the six settings the gate selects seven to eleven times the true usage at the reported band, with the true components inside the selection on 89 to 100% of the samples. In the other two the usage is 1.8 and 1.1 times the truth. Under the floor ϑ of Appendix C the usage is between 0.96 and 1.10 times the truth in every setting.

For the comparison each classical solver is given a search over counts near the truth (Appendix J). The classical solvers move away from the truth as N grows. Every classical solver has left the reported band or the true count by $N = 256$ (Fig. 1(a, b), Tables 2 and 3). With ϕ at the true noise variance, ARD-NMF returns the true count to within fifteen percent at 20 dB and 1.7 to 2.9 times the truth at 30 dB. The ARD-NMF counts on the same grid are analyzed in Appendix K.3. Runtimes are in Table 6.

Removing one loss term at a time at $N = 256$ (Appendix K.5, Table 7) measures what each term of §5 does. Without the usage penalty the count grows in all four settings, and without the energy

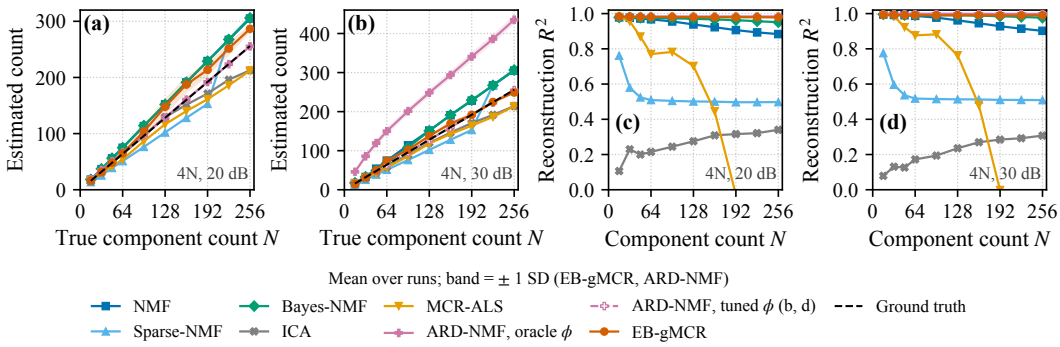


Figure 1: Synthetic mixtures, $4N$ samples. Estimated against true component count at (a) 20 dB and (b) 30 dB, and reconstruction R^2 against the true count at (c) 20 dB and (d) 30 dB.

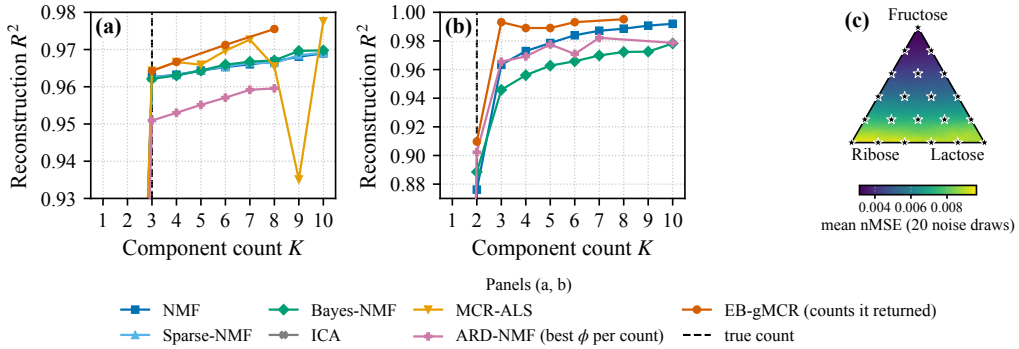


Figure 2: Real chemical spectra. Reconstruction R^2 against component count on (a) Carbs and (b) NIR; the black dashed line is the true component count of the dataset. (c) Frozen-model reuse on the Carbs simplex: mean nMSE at 861 mixtures, the 21 training mixtures starred.

regularizer the termination rule is never met. Removing the exploration noise or the ambiguity term leaves the count within eight percent.

6.2 Real Chemical Spectra

Two public spectroscopy datasets are used. **Carbs** (Engelsen, n.d.) is Raman: $M = 21$ mixtures of three carbohydrates. **NIR** (Liebmann et al., 2009) is first-derivative near-infrared: $M = 166$ samples with two analytes. EB-gMCR starts from a pool of 128 candidates. Metrics at the true count are in Table 10 (Appendix L) and Figure 2(a, b).

On both datasets the EC matches the true count, with no count given to the solver. On NIR the end state (Appendix H.2) is below the lowest band. No method reaches that band with two components (Table 10). Stricter R^2 bands of the same run return more components. A tighter fit tolerance requires more components. At the true count, EB-gMCR reconstructs as well as the best method given the count (Fig. 2(a, b)). On Carbs the EB-gMCR concentrations and profiles are close to those of MCR-ALS, the reference method in chemometrics (Table 10). ARD-NMF decomposes comparably at the true count. ARD-NMF reaches that count for ϕ in a range that a single run does not identify (Appendix L, Fig. 6).

The least-component Carbs checkpoint, selected with no use of the known count, was frozen. The frozen model then decoded mixtures synthesized from the reference profiles across the concentration simplex, with fresh noise (Appendix H.5). The model kept all three components on every mixture. The error tracks the noise floor, the nMSE of the added noise alone, over the whole simplex and not only near the training mixtures (Fig. 2(c); Appendix L, Fig. 7). Corollary 1 says that a model of the process that generated the mixtures decodes mixtures the model has never seen. The frozen

checkpoint is such a model. A fixed-count fit of the training mixtures returns concentrations for those mixtures; a new mixture needs a further fit. The frozen model decodes every mixture of the grid with no further fit.

7 DISCUSSION

Scope. The model covers any measurement that is a linear mixture from a fixed pool under the constraints of Eq. 2, whatever process produced it. Each sample selects a subset of the pool. Each selected component is a unit-norm profile of free sign scaled by a positive concentration. Signals that take negative values are covered: the first-derivative spectra of NIR (§6.2) were decomposed as measured, the concentrations kept positive and the profiles signed (Appendix L). A measurement resolved along an axis, in time or in position, is a set of samples, one per point, since the axis does not enter the selection. Retention time in chromatography is one such axis. When only part of a dataset has the form of Eq. 2, the model covers that part.

On real data the certificate is checked in its fit condition only. Of its two conditions, the fit condition is computed from the returned decomposition, whereas the usage condition compares the total usage of that decomposition with the total usage of the truth. The true supports are unknown, and the unknowns σ, $c_{\min}$ and $s_{\min}$ keep the λ-window of Theorem 1 from being placed. On such data the usage penalty is the pressure toward minimal usage, and a check of how close a run has come to minimal usage remains to be developed.

The condition of Conjecture R is a property of the sampled data: it is met by design when the samples are prepared, and is fixed by the measurement otherwise. It concerns which components occur together and in what proportions. Two components in fixed proportion cannot be separated by any method, and one further proportion restores the separation (Appendix F). Prepared samples that contain each component alone, and mixtures in varying proportions, meet the condition.

Limitations. The guarantees are stated in the cap k on the components per sample, through the spark condition on sets of $2k$ profiles, and in the concentration floor $c_{\min}$, through the product $c_{\min}s_{\min}$. A sample with many components at small concentrations stresses both at once, and the theory has no statement for that case. A component below $c_{\min}$ is outside Eq. 2, and what such a component does to the decomposition is not analyzed. On the solver side, Proposition 2 describes a run after an anchor, a step at which every component's margin is at least a chosen γ (Appendix E). Whether and when a run reaches an anchor is left open here: the selection is discrete, the solver relaxes it, and the relaxed run before an anchor is not analyzed. The objective of §5 is unchanged when the pool slots are permuted, so every minimum has relabeled copies, and which copy a run approaches is not described either. On the sample count, Theorem 3(b) says the component count is eventually exact as the number of samples grows. Its proof is a compactness argument and gives no rate. Appendix D has a concentration bound for the empirical measure. A rate would also need a bound on how the criterion grows with the distance from the truth, a quantitative identifiability statement that is not derived here.

Closing. This paper studies the structure of the solution set of a superposition problem, and EB-gMCR is one attempt at reaching a point of that set. The structure is set by the samples, and the size of the solver's pool decides only whether a solver can reach that point. Over any pool that contains the true profiles and satisfies the spark condition, minimal usage leaves the slots outside the truth unused (Theorem 1). Over a pool below the true count, a solver must merge or drop components, and the theorems do not say which. The sparse autoencoders used on network activations solve a problem of the same form (Bricken et al., 2023; Cunningham et al., 2024), in the opposite direction. There the features are held to outnumber the dimensions (Elhage et al., 2022), and the pool is taken to be smaller than the number of features. The same factorization in that direction may have a different structure, and what a solver returns there is left open.

AI USE STATEMENT

In this work, we used generative AI tools to refine the formulation of mathematical claims, to provide ingredients for proofs and to assist in writing them, to give feedback on the design of experiments,

to implement analysis and benchmark code, and to help interpret results. We have not used generative AI tools to generate synthetic datasets, to clean or reformat datasets, or for translation, and qualitative or thematic data analysis is not applicable to this work. Additionally, we used generative AI tools to draft and edit parts of the paper, to create and edit software code and figure scripts, to search for and check the related literature, to format references, and to suggest the structure of sections. The research questions, the model, and the solver originate with the authors. Every proof was checked line by line by the authors, and every citation was checked against its source. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

This work uses two public spectroscopy datasets and synthetic data. It involves no human subjects and no personal data.

Reproducibility statement

The synthetic data generation and the baseline search are described in Appendices I and J, and the solver's architecture, hyperparameters, and training procedure in Appendix H. Run counts, hardware, and wall-clock times are in Appendix K, and the real-data protocol in Appendix L. Every theorem is proved in full in the appendices. The supplementary material contains the code, the configuration files with every value used, and the result files behind each table and figure. The supplementary material also contains a Lean 4 formalization of parts of the proofs, with a file that maps each formalized statement to the corresponding result in the appendices.

References

Michal Aharon, Michael Elad, and Alfred M. Bruckstein. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Linear Algebra and its Applications*, 416(1):48–67, 2006. doi: 10.1016/j.laa.2005.06.035.

A. Barcaru, H. G. J. Mol, M. Tienstra, and G. Vivó-Truyols. Bayesian approach to peak deconvolution and library search for high resolution gas chromatography–mass spectrometry. *Analytica Chimica Acta*, 983:76–90, 2017. doi: 10.1016/j.aca.2017.06.044.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick 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. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015. doi: 10.1561/2200000050.

Emmanuel J. Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006. doi: 10.1109/TIT.2005.862083.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994. doi: 10.1016/0165-1684(94)90029-9.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK. arXiv:2309.08600.

A. de Juan and R. Tauler. Multivariate curve resolution: 50 years addressing the mixture analysis problem–a review. *Analytica Chimica Acta*, 1145:59–78, 2021. doi: 10.1016/j.aca.2020.10.051.

David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. doi: 10.1109/TIT.2006.871582.

David L. Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. *Proceedings of the National Academy of Sciences*, 100(5): 2197–2202, 2003. doi: 10.1073/pnas.0437847100.

David L. Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf (eds.), *Advances in Neural Information Processing Systems 16 (NIPS 2003)*, pp. 1141–1148. MIT Press, 2003. URL https://proceedings.neurips.cc/paper/2003/hash/1843e35d41ccf6e63273495ba42df3c1-Abstract.html.

R. M. Dudley. *Real Analysis and Probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 2nd edition, 2002.

Rick Durrett. *Probability: Theory and Examples*. Number 49 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 5th edition, 2019. ISBN 978-1-108-47368-2.

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. https://transformer-circuits.pub/2022/toy_model/index.html; arXiv:2209.10652.

Søren Balling Engelsen. Database on Raman spectra of carbohydrates. SPECARB, n.d. Data set; distributed as carbs in the R package mdatools; archived at https://web.archive.org/web/20140807034819/http://www.models.life.ku.dk/~specarb/specarb.html.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015. doi: 10.1007/s00440-014-0583-7.

Xiao Fu, Kejun Huang, Nicholas D. Sidiropoulos, and Wing-Kin Ma. Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications. *IEEE Signal Processing Magazine*, 36(2):59–80, 2019. doi: 10.1109/MSP.2018.2877582.

Charles J. Garfinkle and Christopher J. Hillar. On the uniqueness and stability of dictionaries for sparse representation of noisy signals. *IEEE Transactions on Signal Processing*, 67(23):5884–5892, 2019. doi: 10.1109/TSP.2019.2935914.

Ian J. Goodfellow, Aaron Courville, and Yoshua Bengio. Large-scale feature learning with spike-and-slab sparse coding. In *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, pp. 1387–1394, Edinburgh, Scotland, UK, 2012. Omnipress. arXiv:1206.6407.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory: 16th International Conference, ALT 2005*, volume 3734 of *Lecture Notes in Computer Science*, pp. 63–77. Springer, 2005a. doi: 10.1007/11564089_7.

Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(70):2075–2129, 2005b.

Rémi Gribonval, Rodolphe Jenatton, Francis Bach, Martin Kleinsteuber, and Matthias Seibert. Sample complexity of dictionary learning and other matrix factorizations. *IEEE Transactions on Information Theory*, 61(6):3469–3486, 2015. doi: 10.1109/TIT.2015.2424238.

A. Heinecke, L. Ye, M. De Iorio, and T. Ebbels. Bayesian deconvolution and quantification of metabolites from J-resolved NMR spectroscopy. *Bayesian Analysis*, 16(2):425–458, 2021. doi: 10.1214/20-BA1208.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2013. ISBN 978-0-521-83940-2.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321.

Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-softmax. In *International Conference on Learning Representations (ICLR)*, 2017.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19–23, 2016, Proceedings, Part I*, volume 9851 of *Lecture Notes in Computer Science*, pp. 795–811, Cham, 2016. Springer. doi: 10.1007/978-3-319-46128-1_50.

Alexander S. Kechris. *Classical Descriptive Set Theory*, volume 156 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995. ISBN 0-387-94374-9. doi: 10.1007/978-1-4612-4190-4.

H. R. Keller and D. L. Massart. Evolving factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 12(3):209–224, 1991. doi: 10.1016/0169-7439(92)80002-L.

W. H. Lawton and E. A. Sylvestre. Self modeling curve resolution. *Technometrics*, 13(3):617–633, 1971. doi: 10.1080/00401706.1971.10488823.

Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. doi: 10.1038/44565.

Sangkyun Lee and Stephen J. Wright. Manifold identification in dual averaging for regularized stochastic online learning. *Journal of Machine Learning Research*, 13(55):1705–1744, 2012.

Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Activity identification and local linear convergence of forward–backward-type methods. *SIAM Journal on Optimization*, 27(1):408–437, 2017. doi: 10.1137/16M106340X.

Bettina Liebmann, Anton Friedl, and Kurt Varmuza. Determination of glucose and ethanol in bioethanol production by near infrared spectroscopy and chemometrics. *Analytica Chimica Acta*, 642(1–2):171–178, 2009. doi: 10.1016/j.aca.2008.10.069. Data set `NIR` of the R package `chemometrics`.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The Concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations (ICLR)*, 2017. URL `https://openreview.net/forum?id=S1jE5L5gl`.

Marcel Maeder. Evolving factor analysis for the resolution of overlapping chromatographic peaks. *Analytical Chemistry*, 59(3):527–530, 1987. doi: 10.1021/ac00130a035.

Edmund R. Malinowski. Determination of the number of factors and the experimental error in a data matrix. *Analytical Chemistry*, 49(4):612–617, 1977. doi: 10.1021/ac50012a027.

Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparsity enforcing penalties. *Journal of Machine Learning Research*, 18(128):1–33, 2017.

K. R. Parthasarathy. *Probability Measures on Metric Spaces*. Academic Press, New York, 1967.

David Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, 9(1):135–140, 1981. doi: 10.1214/aos/1176345339.

Mathias Sawall, Henning Schröder, Denise Meinhardt, and Klaus Neymeyr. On the ambiguity underlying multivariate curve resolution methods. In *Comprehensive Chemometrics*, pp. 199–231. Elsevier, second edition, 2020. doi: 10.1016/B978-0-12-409547-2.14582-2.

Mikkel N. Schmidt, Ole Winther, and Lars Kai Hansen. Bayesian non-negative matrix factorization. In *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation (ICA 2009)*, Lecture Notes in Computer Science, pp. 540–547. Springer, 2009. doi: 10.1007/978-3-642-00599-2_68.

J. P. Smith, E. C. Holahan, F. C. Smith, V. Marrero, and K. S. Booksh. A novel multivariate curve resolution-alternating least squares (MCR-ALS) methodology for application in hyperspectral Raman imaging analysis. *Analyst*, 144(18):5425–5438, 2019. doi: 10.1039/C9AN00787C.

Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, volume 23 of *JMLR Workshop and Conference Proceedings*, pp. 37.1–37.18, 2012.

Vincent Y. F. Tan and Cédric Févotte. Automatic relevance determination in nonnegative matrix factorization with the β-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1592–1605, 2013. doi: 10.1109/TPAMI.2012.240.

Romà Tauler, Bruce Kowalski, and Sydney Fleming. Multivariate curve resolution applied to spectral data from multiple runs of an industrial process. *Analytical Chemistry*, 65(15):2040–2047, 1993. doi: 10.1021/ac00063a019.

Romà Tauler, Age Smilde, and Bruce Kowalski. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *Journal of Chemometrics*, 9(1):31–58, 1995. doi: 10.1002/cem.1180090105.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Michael E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

V. S. Varadarajan. On the convergence of sample probability distributions. *Sankhyā*, 19(1/2):23–26, 1958.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2018. ISBN 9781108415194.

Jean Ville. *Étude Critique de la Notion de Collectif*. PhD thesis, Faculté des Sciences de Paris, Paris, 1939. Published as Monographies des Probabilités, Fascicule III (ed. Émile Borel), Gauthier-Villars, Paris, 144pp.

Abraham Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 1949. doi: 10.1214/aoms/1177729952.

Mingyuan Zhou, Haojun Chen, John Paisley, Lu Ren, Guillermo Sapiro, and Lawrence Carin. Non-parametric Bayesian dictionary learning for sparse image representations. In *Advances in Neural Information Processing Systems 22 (NIPS 2009)*, pp. 2295–2303. Curran Associates, 2009. URL https://proceedings.neurips.cc/paper/2009/hash/cfecdb276f634854f3ef915e2e980c31-Abstract.html.

A Formal Statement and Proof of Proposition 1

Two remarks on the spark condition. The condition constrains the profiles themselves, not the selection indicators: which combinations of components occur plays no role in it, and the occurrence conditions appear separately in §4. Visually similar profiles are typically still linearly independent; what similarity degrades is the margin $s_{\min}$, and the margin, not bare independence, is what every guarantee consumes.

This appendix proves Proposition 1 in full. The proposition says: if any k-sparse combination with detectably active coefficients reconstructs a true sparse sample with error below the threshold $c_{\min} s_{\min}$, then it is the true decomposition, in support and in coefficients alike. The proof subtracts the two candidate representations and applies the spark condition to their difference: Steps 1–5 bound every coordinate of the difference strictly below $c_{\min}$; Step 6 turns that cap into support equality (a support disagreement would force a coordinate of magnitude at least $c_{\min}$); Step 7 reads off the coefficient error.

Notation.

$\|x\|$	Euclidean norm of $x \in \mathbb{R}^d$; d is the feature dimension
S_A	the $d \times \|A\|$ matrix with the profiles s_i, $i \in A$, as columns
$S_A v$	$= \sum_{i \in A} v_i s_i$, with v indexed by the elements of A
$\sigma_{\min}(S_A)$	$= \min_{\|u\|=1} \|S_A u\|$, the smallest singular value of S_A
$s_{\min},\ c_{\min}$	assumed constants of the model: spark margin, concentration floor
ϵ	deterministic reconstruction tolerance; distinct from the model noise ε

Two conventions. In the matrix notation of Eq. 1, profiles are *rows* of S, whereas S_A stacks the selected profiles as *columns*, so that $S_A v$ maps a coefficient vector to a synthetic sample in $\mathbb{R}^d$ and $\|S_A u\| \geq \sigma_{\min}(S_A)\,\|u\|$ for every u. $\sigma_{\min}(\cdot)$ always denotes a quantity *computed* from a matrix, while $s_{\min}$ (like $c_{\min}$) is an *assumed* constant of the model; the spark condition asserts that the computed quantity dominates the assumed one, $\sigma_{\min}(S_A) \geq s_{\min}$.

Proposition 1, restated. The main text states this result under hypotheses given by reference; the full statement follows. Fix $k \in \mathbb{N}$ and suppose the unit-norm profiles $s_1, \ldots, s_N \in \mathbb{R}^d$ satisfy the spark condition with margin $s_{\min} > 0$: $\sigma_{\min}(S_A) \geq s_{\min}$ for every $A \subseteq [N]$ with $|A| \leq 2k$. Consider supports $J, J' \subseteq [N]$ with $|J|, |J'| \leq k$ and coefficients $c_i \geq c_{\min} > 0$ for $i \in J$ and $c'_i \geq c_{\min}$ for $i \in J'$, and write $x_0 = \sum_{i \in J} c_i s_i$. If

$$\Big\| x_0 - \sum_{i \in J'} c'_i s_i \Big\| \;\leq\; \epsilon \qquad \text{for some } 0 \leq \epsilon < c_{\min}\, s_{\min},$$

then $J' = J$ and $|c_i - c'_i| \leq \epsilon / s_{\min}$ for every $i \in J$.

Proof. Step 1 (difference vector). Let $A = J \cup J'$ and define $v \in \mathbb{R}^{|A|}$, indexed by A, coordinate-wise:

$$v_i = \begin{cases} c_i, & i \in J \setminus J', \\ c_i - c'_i, & i \in J \cap J', \\ -c'_i, & i \in J' \setminus J. \end{cases}$$

This is the difference of the two coefficient vectors after each is extended by zero to all of A.

Step 2 (residual identity). By linearity of the matrix–vector product and Step 1,

$$S_A v \;=\; \sum_{i \in A} v_i s_i \;=\; \sum_{i \in J} c_i s_i - \sum_{i \in J'} c'_i s_i \;=\; x_0 - \sum_{i \in J'} c'_i s_i,$$

and therefore $\|S_A v\| \leq \epsilon$ by the hypothesis.

Step 3 (the spark condition applies). $|A| \leq |J| + |J'| \leq 2k$, so the spark condition gives $\sigma_{\min}(S_A) \geq s_{\min}$.

Step 4 (norm bound). By the definition of $\sigma_{\min}$ (Notation above) and Steps 2–3,

$$s_{\min} \|v\| \;\leq\; \sigma_{\min}(S_A)\,\|v\| \;\leq\; \|S_A v\| \;\leq\; \epsilon, \qquad \text{hence} \qquad \|v\| \;\leq\; \epsilon / s_{\min}.$$

Step 5 (coordinate bound). For every $i \in A$, $|v_i| \le \|v\|$ (a coordinate of a vector is bounded by its Euclidean norm), so by Step 4 and the hypothesis $\epsilon < c_{\min} s_{\min}$,

$$|v_i| \;\le\; \epsilon / s_{\min} \;<\; c_{\min}.$$

Step 6 (support equality). This step is the reason the gap between ϵ and $c_{\min} s_{\min}$ exists: a support disagreement forces some coordinate of v to reach magnitude at least $c_{\min}$, while Step 5 caps every coordinate strictly below $c_{\min}$. Concretely, suppose $i \in J \setminus J'$. Then $v_i = c_i \ge c_{\min}$ by Step 1 and the hypothesis $c_i \ge c_{\min}$, contradicting Step 5. Hence $J \setminus J' = \emptyset$. Symmetrically, $i \in J' \setminus J$ gives $|v_i| = c'_i \ge c_{\min}$, again contradicting Step 5, so $J' \setminus J = \emptyset$. Therefore $J' = J$.

Step 7 (coefficient bound). With $J' = J$, Step 1 gives $v_i = c_i - c'_i$ for every $i \in J$, and Steps 4–5 give $|c_i - c'_i| \le \|v\| \le \epsilon / s_{\min}$. □

The proof uses neither the unit-norm normalization of the profiles nor any upper bound $c_{\max}$ on the coefficients; the proposition therefore holds verbatim without these assumptions.

B Formal Statement and Proof of Theorem 1

This appendix proves Theorem 1 in full. Part (a) says that among all exact reproductions of noiseless data over the pool, the true occurrence matrix uses strictly fewer selections than any alternative. Part (b) says that under noise, for λ inside the explicit window, every true support beats every selection differing from it in one component, by a quantified gap, simultaneously across all samples with probability $1 - \alpha$; without noise the optimality is global among k-sparse selections. Part (a) is proved in three steps: uniqueness of sparse representations, a row-by-row spending bound, and summation. Part (b) runs in eight steps: a one-flip fit identity (Step 1), the costs of the two flips (Steps 2–3), dual vectors (Step 4), the noise event (Step 5), inactivity of the nonnegativity constraint at the truth (Step 6), the window verdict (Step 7), and the noiseless global case (Step 8).

Notation. The notation of Appendix A remains in force. New objects:

$X_0 = \{x_0^{(m)}\}_{m=1}^M$	noiseless data; the subscript is the digit zero (zero noise)
$x_0^{(m)}$	$= \sum_{i \in J_m} c_i^{(m)} s_i$; true support J_m, $\lvert J_m\rvert \le k$; $c_i^{(m)} \in [c_{\min}, c_{\max}]$
$x^{(m)}$	$= x_0^{(m)} + \varepsilon^{(m)}$, the observed data (only these are observed)
A_m	$= \{j : \delta_{mj} = 1\}$, the row supports of $\delta \in \{0,1\}^{M \times N}$
P_V, $\mathrm{dist}(x, V)$	orthogonal projection onto the subspace V; $\mathrm{dist}(x, V) = \|x - P_V x\|$
$\ell_\lambda(A; x)$	$= \min_{c \ge 0} \|x - \sum_{j \in A} c_j s_j\|^2 + \lambda \lvert A\rvert$ (Eq. 3)

A matrix δ with row supports A_m *admits an exact factorization* of X_0 over the pool if for every m there are real coefficients (of any sign) supported on A_m with $x_0^{(m)} = \sum_{j \in A_m} c'_j s_j$. The noise vectors $\varepsilon^{(m)}$ are independent, mean-zero, and sub-Gaussian with proxy σ (Vershynin, 2018, cf. §§2.5 and 3.4): for every fixed $a \in \mathbb{R}^d$ and every $t > 0$,

$$\Pr\Big(\big|\langle \varepsilon^{(m)}, a\rangle\big| > t\Big) \;\le\; 2\exp\Big(-\frac{t^2}{2\sigma^2 \|a\|^2}\Big).$$

Theorem 1, restated. The main text states this result under hypotheses given by reference; the full statement follows. Suppose the unit-norm profiles $s_1, \ldots, s_N \in \mathbb{R}^d$ satisfy the spark condition with margin $s_{\min}$ (Definition 1) and the data follow the model above.

(a) The matrix $\delta^\star$ with rows $\mathbf{1}_{J_m}$ is the unique minimizer of $\sum_{m,j} \delta_{mj}$ among all δ admitting an exact factorization of X_0 over the pool.

(b) Fix $\alpha \in (0,1)$ and set $L_\alpha = \sqrt{2\log(6NM/\alpha)}$. Assume the SNR condition and the window

$$c_{\min} s_{\min} \;>\; 2\sigma L_\alpha, \qquad \sigma^2 L_\alpha^2 \;<\; \lambda \;<\; \big(c_{\min} s_{\min} - \sigma L_\alpha\big)^2.$$

Then with probability at least $1-\alpha$ over the noise, simultaneously for every sample m and every $A \subseteq [N]$ obtained from J_m by adding or removing one component,

$$\ell_\lambda(A;\, x^{(m)}) - \ell_\lambda(J_m;\, x^{(m)}) \;\geq\; \min\Big\{\, \lambda - \sigma^2 L_\alpha^2,\;\; (c_{\min} s_{\min} - \sigma L_\alpha)^2 - \lambda \,\Big\} \;>\; 0.$$

At $\sigma = 0$: for every $0 < \lambda < c_{\min}^2 s_{\min}^2$, each J_m is the unique global minimizer of $\ell_\lambda(\,\cdot\,;\, x_0^{(m)})$ among selections with $|A| \leq k$.

Proof of part (a). Step 1 (two sparse representations coincide). Suppose

$$\sum_{i \in J} c_i s_i \;=\; \sum_{i \in J'} c'_i s_i, \qquad |J| \leq k, \quad |J'| \leq k,$$

with all listed coefficients nonzero. Set $A = J \cup J'$, so $|A| \leq |J| + |J'| \leq 2k$, and define $v \in \mathbb{R}^{|A|}$ coordinate-wise, exactly as in Step 1 of Appendix A:

$$v_i = \begin{cases} c_i, & i \in J \setminus J', \\ c_i - c'_i, & i \in J \cap J', \\ -\,c'_i, & i \in J' \setminus J. \end{cases}$$

By linearity of the matrix–vector product,

$$S_A v \;=\; \sum_{i \in J} c_i s_i - \sum_{i \in J'} c'_i s_i \;=\; 0.$$

If $(J, c) \neq (J', c')$ then $v \neq 0$, and the spark condition gives

$$0 \;=\; \|S_A v\| \;\geq\; \sigma_{\min}(S_A)\, \|v\| \;\geq\; s_{\min} \|v\| \;>\; 0,$$

a contradiction. Hence $J = J'$ and $c = c'$: a noiseless sample has exactly one representation using at most k pool components with nonzero coefficients.

Step 2 (every row spends at least the truth). Let δ admit an exact factorization with row supports A_m and realizing coefficients c', and let

$$A'_m \;=\; \{\, j \in A_m : c'_j \neq 0 \,\} \;\subseteq\; A_m$$

be the nonzero pattern on row m. Two cases.

Case $|A_m| \leq k$. Then $|A'_m| \leq k$, and sample m has the two representations

$$x_0^{(m)} \;=\; \sum_{i \in J_m} c_i^{(m)} s_i \;=\; \sum_{j \in A'_m} c'_j s_j,$$

each using at most k components with nonzero coefficients. Step 1 forces $A'_m = J_m$, hence $J_m \subseteq A_m$ and

$$\sum_j \delta_{mj} \;=\; |A_m| \;\geq\; |J_m|, \qquad \text{with equality exactly when } A_m = J_m.$$

Case $|A_m| > k$. Then directly

$$\sum_j \delta_{mj} \;=\; |A_m| \;>\; k \;\geq\; |J_m|.$$

Step 3 (summation). Summing the row bounds of Step 2 over m,

$$\sum_{m,j} \delta_{mj} \;\geq\; \sum_m |J_m| \;=\; \sum_{m,j} \delta^\star_{mj},$$

with equality exactly when $A_m = J_m$ for every m, that is, when $\delta = \delta^\star$. Hence $\delta^\star$ is the unique minimizer. □

Proof of part (b). Step 1 (one-flip fit identity). Every index set appearing in Steps 1–7 has at most $k+1 \le 2k$ elements, so by the spark condition every S_A below has full column rank. Write $V = \mathrm{span}(S_A)$; the *unconstrained* least-squares fit equals the squared distance to the span,

$$\min_{c\in\mathbb{R}^{|A|}} \Big\| x - \sum_{j\in A} c_j s_j \Big\|^2 \;=\; \|x - P_V x\|^2,$$

and we write, for the unconstrained counterpart of Eq. 3,

$$\ell^{\mathrm{u}}_\lambda(A;\, x) \;=\; \|x - P_V x\|^2 + \lambda|A|.$$

Let V' extend V by one component and let u be the unit vector spanning $V' \cap V^\perp$. Since $P_{V'}x = P_V x + \langle x, u\rangle u$ (projection onto an orthogonal decomposition of V'), the Pythagorean identity gives

$$\|x - P_{V'}x\|^2 \;=\; \|x - P_V x\|^2 - \langle x, u\rangle^2 :$$

one flip changes the unconstrained fit by exactly $\langle x, u\rangle^2$.

Step 2 (adding a spurious component). Fix a sample m and $j \notin J_m$; take $V = \mathrm{span}(S_{J_m})$ and $V' = \mathrm{span}(S_{J_m\cup\{j\}})$. Since $x_0^{(m)} \in V$ and $u \perp V$,

$$\langle x^{(m)}, u\rangle \;=\; \langle x_0^{(m)}, u\rangle + \langle \varepsilon^{(m)}, u\rangle \;=\; \langle \varepsilon^{(m)}, u\rangle,$$

and by Step 1 the objective change of the addition is

$$\ell^{\mathrm{u}}_\lambda(J_m \cup \{j\};\, x^{(m)}) - \ell^{\mathrm{u}}_\lambda(J_m;\, x^{(m)}) \;=\; \lambda - \langle \varepsilon^{(m)}, u\rangle^2.$$

Step 3 (removing a true component). Fix $i \in J_m$; take $V = \mathrm{span}(S_{J_m\setminus\{i\}})$ and $V' = \mathrm{span}(S_{J_m})$. Here u is the normalized component of s_i orthogonal to V, so

$$\langle x_0^{(m)}, u\rangle \;=\; \sum_{i'\in J_m} c_{i'}^{(m)} \langle s_{i'}, u\rangle \;=\; c_i^{(m)} \langle s_i, u\rangle, \qquad \langle s_i, u\rangle = \mathrm{dist}(s_i, V).$$

The distance is bounded below by the spark margin: for any coefficient vector w on $J_m \setminus \{i\}$,

$$\Big\| s_i - \sum_{j\in J_m\setminus\{i\}} w_j s_j \Big\| \;=\; \|S_{J_m} v\| \;\ge\; \sigma_{\min}(S_{J_m})\, \|v\| \;\ge\; s_{\min},$$

where v carries $-w$ on $J_m \setminus \{i\}$ and 1 at i, so $\|v\| \ge 1$; minimizing over w gives $\mathrm{dist}(s_i, V) \ge s_{\min}$. Combining, with $c_i^{(m)} \ge c_{\min}$,

$$\big|\langle x^{(m)}, u\rangle\big| \;\ge\; \big|\langle x_0^{(m)}, u\rangle\big| - \big|\langle \varepsilon^{(m)}, u\rangle\big| \;\ge\; c_{\min} s_{\min} - \big|\langle \varepsilon^{(m)}, u\rangle\big|,$$

and by Step 1 the objective change of the removal is

$$\ell^{\mathrm{u}}_\lambda(J_m \setminus \{i\};\, x^{(m)}) - \ell^{\mathrm{u}}_\lambda(J_m;\, x^{(m)}) \;=\; \langle x^{(m)}, u\rangle^2 - \lambda.$$

Step 4 (dual vectors). For each sample m and each $i \in J_m$, set $V_i = \mathrm{span}(S_{J_m\setminus\{i\}})$ and define

$$a_i^{(m)} \;=\; \frac{s_i - P_{V_i} s_i}{\|s_i - P_{V_i} s_i\|^2},$$

well defined since $\|s_i - P_{V_i}s_i\| = \mathrm{dist}(s_i, V_i) \ge s_{\min} > 0$ by Step 3. The numerator is orthogonal to V_i, so

$$\langle a_i^{(m)}, s_j\rangle = 0 \quad \text{for } j \in J_m \setminus \{i\}, \qquad \langle a_i^{(m)}, s_i\rangle = \frac{\big\langle s_i - P_{V_i}s_i,\, (s_i - P_{V_i}s_i) + P_{V_i}s_i\big\rangle}{\|s_i - P_{V_i}s_i\|^2} = 1,$$

and its norm is controlled by the same distance:

$$\|a_i^{(m)}\| \;=\; \frac{1}{\|s_i - P_{V_i}s_i\|} \;=\; \frac{1}{\mathrm{dist}(s_i, V_i)} \;\le\; \frac{1}{s_{\min}}.$$

Both s_i and $P_{V_i}s_i$ lie in $\mathrm{span}(S_{J_m})$, so $a_i^{(m)} \in \mathrm{span}(S_{J_m})$.

Step 5 (the noise event). The directions u of Steps 2–3 are at most $2NM$ fixed unit vectors (N additions and $k \le N$ removals per sample), and the dual vectors of Step 4 are at most $kM \le NM$ further fixed vectors; all are determined by S and $\{J_m\}$, never by the noise. Apply the sub-Gaussian tail (Notation) with $t = \sigma L_\alpha \|a\|$ to each of these at most $3NM$ fixed vectors:

$$\Pr\Big(\big|\langle \varepsilon^{(m)}, a\rangle\big| > \sigma L_\alpha \|a\|\Big) \;\le\; 2\exp\big(-L_\alpha^2/2\big) \;=\; 2\cdot\frac{\alpha}{6NM} \;=\; \frac{\alpha}{3NM}.$$

By the union bound over the at most $3NM$ events, with probability at least $1-\alpha$ the following event E holds:

$$E: \qquad \big|\langle \varepsilon^{(m)}, u\rangle\big| \le \sigma L_\alpha \;\text{ for every flip direction,}$$
$$\big|\langle \varepsilon^{(m)}, a_i^{(m)}\rangle\big| \le \frac{\sigma L_\alpha}{s_{\min}} \;\text{ for every } m,\; i \in J_m.$$

Steps 6–7 run on E.

Step 6 (the nonnegativity constraint is inactive at the truth). ℓ_λ constrains $c \ge 0$; Steps 1–3 used unconstrained fits. Write the projection of $x^{(m)}$ onto $V = \mathrm{span}(S_{J_m})$ in pool coordinates,

$$P_V x^{(m)} \;=\; \sum_{j\in J_m} \hat{c}_j^{(m)} s_j.$$

Pairing with the dual vectors of Step 4: for every $i \in J_m$,

$$\begin{aligned}\hat{c}_i^{(m)} \;&=\; \big\langle a_i^{(m)},\, P_V x^{(m)}\big\rangle \;=\; \big\langle a_i^{(m)},\, x^{(m)}\big\rangle \\ &=\; \big\langle a_i^{(m)},\, x_0^{(m)}\big\rangle + \big\langle a_i^{(m)},\, \varepsilon^{(m)}\big\rangle \;=\; c_i^{(m)} + \big\langle a_i^{(m)},\, \varepsilon^{(m)}\big\rangle,\end{aligned}$$

by, in order: biorthogonality (Step 4) applied to the sum; $a_i^{(m)} \in V$, so the residual $x^{(m)} - P_V x^{(m)} \perp a_i^{(m)}$ contributes nothing; $x^{(m)} = x_0^{(m)} + \varepsilon^{(m)}$; biorthogonality applied to $x_0^{(m)} = \sum_{j\in J_m} c_j^{(m)} s_j$. On E, entrywise,

$$\hat{c}_i^{(m)} \;\ge\; c_{\min} - \frac{\sigma L_\alpha}{s_{\min}} \;>\; c_{\min} - \frac{c_{\min}}{2} \;=\; \frac{c_{\min}}{2} \;>\; 0,$$

using the SNR condition in the form $\sigma L_\alpha / s_{\min} < c_{\min}/2$. The coefficients $\hat{c}^{(m)}$ are therefore feasible for the constraint $c \ge 0$ and attain the unconstrained fit value $\|x^{(m)} - P_V x^{(m)}\|^2$, so the two fits coincide at J_m:

$$\ell_\lambda(J_m;\, x^{(m)}) \;=\; \ell_\lambda^{\mathrm{u}}(J_m;\, x^{(m)}).$$

For every other selection A, the constrained minimum runs over a subset of coefficient vectors, so

$$\ell_\lambda(A;\, x^{(m)}) \;\ge\; \ell_\lambda^{\mathrm{u}}(A;\, x^{(m)}).$$

Step 7 (the window verdict). On E, chaining Step 6, Step 2, and Step 5, an addition costs

$$\ell_\lambda(J_m \cup \{j\};\, x^{(m)}) - \ell_\lambda(J_m;\, x^{(m)}) \;\ge\; \lambda - \langle \varepsilon^{(m)}, u\rangle^2 \;\ge\; \lambda - \sigma^2 L_\alpha^2 \;>\; 0$$

by the left edge of the window. Chaining Step 6, Step 3, and Step 5 (the SNR condition makes $c_{\min} s_{\min} - \sigma L_\alpha > 0$), a removal costs

$$\ell_\lambda(J_m \setminus \{i\};\, x^{(m)}) - \ell_\lambda(J_m;\, x^{(m)}) \;\ge\; \langle x^{(m)}, u\rangle^2 - \lambda \;\ge\; \big(c_{\min} s_{\min} - \sigma L_\alpha\big)^2 - \lambda \;>\; 0$$

by the right edge. Every one-flip change therefore costs at least

$$\min\Big\{\, \lambda - \sigma^2 L_\alpha^2,\;\; \big(c_{\min} s_{\min} - \sigma L_\alpha\big)^2 - \lambda \,\Big\},$$

the stated gap.

Step 8 (noiseless global optimality). Let $\sigma = 0$, so $x^{(m)} = x_0^{(m)}$ and the constrained fit at J_m is exact with the feasible coefficients $c^{(m)}$. Let $A \ne J_m$ with $|A| \le k$ and set $r = |J_m \setminus A|$. Two cases.

Case $r = 0$. Then $J_m \subsetneq A$, the fit is 0 at both selections, and

$$\ell_\lambda(A;\, x_0^{(m)}) - \ell_\lambda(J_m;\, x_0^{(m)}) \;=\; \lambda\big(|A| - |J_m|\big) \;\geq\; \lambda \;>\; 0.$$

Case $r \geq 1$. For any coefficient vector c supported on A (of any sign),

$$x_0^{(m)} - S_A c \;=\; S_{J_m \cup A}\, v,$$

where v carries $c_j^{(m)}$ at $j \in J_m \setminus A$, $c_j^{(m)} - c_j$ on $J_m \cap A$, and $-c_j$ on $A \setminus J_m$. Since $|J_m \cup A| \leq 2k$, the spark condition gives

$$\big\|x_0^{(m)} - S_A c\big\|^2 \;\geq\; s_{\min}^2 \|v\|^2 \;\geq\; s_{\min}^2 \sum_{j \in J_m \setminus A} \big(c_j^{(m)}\big)^2 \;\geq\; r\, c_{\min}^2 s_{\min}^2.$$

The penalty difference is bounded using $|A| \geq |J_m| - r$:

$$\lambda\big(|A| - |J_m|\big) \;\geq\; -\lambda r,$$

hence

$$\ell_\lambda(A;\, x_0^{(m)}) - \ell_\lambda(J_m;\, x_0^{(m)}) \;\geq\; r\, c_{\min}^2 s_{\min}^2 - \lambda r \;=\; r\big(c_{\min}^2 s_{\min}^2 - \lambda\big) \;>\; 0.$$

Both cases are strict, so J_m is the unique global minimizer among selections with $|A| \leq k$. □

Part (a) uses only the spark condition and the sparsity cap: neither the concentration floor, the ceiling, nor nonnegativity of the competing coefficients enters, so uniqueness of the minimal occurrence matrix holds against arbitrary sign patterns. The floor and nonnegativity enter part (b): the floor sets the removal cost $(c_{\min} s_{\min} - \sigma L_\alpha)^2$, and nonnegativity is neutralized in Step 6.

Relation to magnitude penalties. Eq. 3 charges the selection cardinality $|A| = \sum_j \delta_{mj}$, the per-sample summand of the total usage $\sum_{m,j} \delta_{mj}$ minimized in part (a); the inner minimization over c carries no penalty. A ridge or lasso penalty (Hoerl & Kennard, 1970; Tibshirani, 1996) instead charges $\|c\|_2^2$ or $\|c\|_1$ and perturbs every retained coefficient. Ridge sets no coefficient to zero, so its support must be fixed by thresholding outside the objective; the lasso does set coefficients to zero through its objective, but which ones is a by-product of magnitude shrinkage: the surviving support depends on coefficient scales and the regularization path, and neither objective prices the count $|A|$ itself. Relevance-determination methods (Tipping, 2001) likewise prune by magnitude, with no priced count. Appendix F computes the difference on data that admit infinitely many exact solutions: ℓ_1 and ℓ_2 each select a specific wrong member with certainty. On such data no criterion is guaranteed correct, the cardinality criterion included; what differs is the cause of failure, not the outcome. A magnitude penalty selects by built-in preference and is wrong with certainty; the cardinality criterion fails there only for lack of information, and is provably correct once the data supply it (in the toy, one further sample; in general, Theorem 1).

C Formal Statement and Proof of Theorem 2 (Certified Recovery)

This appendix states and proves Theorem 2 in full, together with Corollary 1. Its role in the argument is the second obstacle of §4: Theorem 1 makes the truth the optimum of a searchable criterion, but no optimizer is guaranteed to reach an optimum, and a returned solution cannot be compared against the unseen truth, so a search can succeed without anyone being able to know it. Theorem 2 removes the need to trust the search: it provides conditions on the returned solution (the certificate; their observability status is stated with their definitions below) under which any solution satisfying them, no matter how it was produced, is the truth. Every guarantee below therefore quantifies over certificate-passing solutions, never over the algorithm that found them.

Part (a) treats a known pool: on one named noise event of probability at least $1 - \alpha$, every solution that passes the certificate (each sample reconstructed within tolerance by at most k components; total usage no larger than the truth's) recovers every sample's support exactly, hence the exact component count, and every concentration to noise-floor accuracy; the guarantee quantifies over all

certificate-passing solutions at once, with no assumption on how they were found. Part (b) treats a learned pool on noiseless data: every certificate-passing solution recovers the active components exactly and correctly signed, up to an injective relabeling π, together with exact supports, component count, and concentrations. Part (c) treats the learned pool under noise and is conditional: its linear-algebraic content is proved in full below, and the one remaining step, a combinatorial statement about how samples distribute over candidate sets, is stated formally as Conjecture R and assumed, never claimed. Corollary 1 specializes part (a) to a single fresh sample decoded over the recovered pool.

The proofs are organized in five layers. Part (a): Layer I (Lemma 1) constructs the noise event E and extracts the single inequality (11) through which every later step uses E; Layer II (Lemmas 2 and 3) proves, deterministically on E, detection of missing components and usage counting; Layer III (Lemma 4) adds coefficient stability, and the assembly takes five steps. Part (b), Layer IV, runs through three lemmas (span intersection along an independent family, Lemma 5; tight usage with exact spans, Lemma 6; anchoring of each active component to a unique pool slot, Lemma 7) and a five-step assembly. Part (c), Layer V, proves three unconditional robust lemmas (a crude coefficient bound, Lemma 8; a perturbation-stable dimension count, Lemma 9; a two-subspace pinch, Lemma 10), one unconditional lemma quantifying why out-of-span errors cannot cancel across coefficient-diverse samples (Lemma 11), closes the 1-sparse case unconditionally (Lemma 12), and then proves the reduction to Conjecture R in six steps. Corollary 1 is part (a) at $M = 1$.

Notation. The notation of Appendices A and B remains in force. New objects:

$J_{\mathrm{act}},\ K^*$	$J_{\mathrm{act}} = \bigcup_m J_m$, the active components; $K^* = \lvert J_{\mathrm{act}}\rvert$, their count
V_T	$= \mathrm{span}(s_i : i \in T)$, the span of the true profiles indexed by T
Q_V	$= I - P_V$, the projection onto the orthogonal complement of the subspace V
n_{2k}	$= \sum_{r=0}^{2k} \binom{N}{r}$, the number of pool subsets of size at most $2k$
$\varepsilon_{\mathrm{ns}}(\alpha)$	$= 2\sigma\sqrt{2\big(d\log 5 + \log(2M/\alpha)\big)}$, the noise-norm level
$\varepsilon_{\mathrm{prj}}(\alpha)$	$= 2\sigma\sqrt{2\big(2k\log 5 + \log(2Mn_{2k}/\alpha)\big)}$, the projected-noise level
$\varepsilon_{\mathrm{fit}}$	fit tolerance of the certificate; a constant chosen by the analyst
$\hat{S} = (\hat{s}_1, \ldots, \hat{s}_N)$	the learned pool, unit-norm columns $\hat{s}_j$ (parts (b) and (c))
$\hat{s}_{\min}$	assumed regularity margin of the learned pool
$\bar{\varepsilon}$	$= \varepsilon_{\mathrm{fit}} + \varepsilon_{\mathrm{ns}}(\alpha)$, the signal-level tolerance (part (c))
ϑ	significance floor on used coefficients (part (c))
γ_{c}	coefficient-diversity quantile (Conjecture R)
$\hat{C}$	$= (k\,c_{\max} + \bar{\varepsilon})/\hat{s}_{\min}$, the coefficient bound of Lemma 8
π	relabeling: an injective map $J_{\mathrm{act}} \to [N]$ (see below)
$\lVert\cdot\rVert_F,\ \lVert\cdot\rVert_{\mathrm{op}}$	Frobenius and operator norms of matrices

The noise vectors $\varepsilon^{(m)}$ are independent, mean zero, and sub-Gaussian with proxy σ in the moment-generating sense: $\mathbb{E}\, e^{\lambda\langle u, \varepsilon^{(m)}\rangle} \le e^{\lambda^2\sigma^2/2}$ for every unit $u \in \mathbb{R}^d$ and every $\lambda \in \mathbb{R}$ (Vershynin, 2018, §2.5). This implies the one-sided tail bound $\Pr(\langle u, \varepsilon^{(m)}\rangle \ge t) \le e^{-t^2/2\sigma^2}$ used in Lemma 1 and the two-sided tail form of Appendix B.

A *candidate solution over a known pool* is a family $(\{A_m\}_{m=1}^M, \{\hat{c}^{(m)}\}_{m=1}^M)$ of supports $A_m \subseteq [N]$ and coefficient vectors $\hat{c}^{(m)}$ supported on A_m. The *known-pool certificate*:

- *(C1)* $\lvert A_m\rvert \le k$ and $\big\lVert x^{(m)} - S_{A_m}\hat{c}^{(m)}\big\rVert \le \varepsilon_{\mathrm{fit}}$ for every $m \in [M]$;
- *(C2)* $\sum_m \lvert A_m\rvert \le \sum_m \lvert J_m\rvert$.

A *candidate solution with learned pool* is a triple $(\hat{S}, \{A_m\}, \{\hat{c}^{(m)}\})$ in which the pool $\hat{S}$ is itself estimated, with $\lVert\hat{s}_j\rVert = 1$ and every used coefficient strictly positive ($\hat{c}_j^{(m)} > 0$ for $j \in A_m$; zero-coefficient slots are dropped from A_m, which only lowers usage). In parts (b) and (c) the index set $[N]$ indexes the learned pool's slots, while the true supports J_m retain their meaning from the model; the true profiles enter only through $(s_i)_{i \in J_{\mathrm{act}}}$ and their spans V_T. The *noiseless certificate*:

- *(C3)* $|A_m| \le k$ and $\sum_{j \in A_m} \hat{c}_j^{(m)} \hat{s}_j = x_0^{(m)}$ for every m;
- *(C4)* $\sum_m |A_m| \le \sum_m |J_m|$;
- *(C5)* $\sigma_{\min}(\hat{S}_B) \ge \hat{s}_{\min} > 0$ for every $B \subseteq [N]$ with $|B| \le 2k$;
- *(C6)* for every $i \in J_{\text{act}}$ there exist samples m, m' with $J_m \cap J_{m'} = \{i\}$;
- *(C7)* $\dim\big(\operatorname{span}(\hat{S}_{A_m}) \cap V_{J_m}\big) = |J_m|$ for every m.

The *robust certificate*:

- *(C8)* $|A_m| \le k$, $\big\|x^{(m)} - \hat{S}_{A_m} \hat{c}^{(m)}\big\| \le \varepsilon_{\text{fit}}$, and $\hat{c}_j^{(m)} \ge \vartheta > 0$ for $j \in A_m$, for every m;
- *(C9)* $\sum_m |A_m| \le \sum_m |J_m|$;
- *(C10)* $\sigma_{\min}(\hat{S}_B) \ge \hat{s}_{\min} > 0$ for every $B \subseteq [N]$ with $|B| \le 2k$;
- *(C11)* $|A_m| \ge |J_m|$ for every m;

together with (C6). The conditions differ in what can be checked at run time, as follows. The fit conditions (C1), (C3), and (C8) are computable from data and solution alone. The usage-parsimony conditions (C2), (C4), and (C9) compare against the unknown truth and are not computable at run time: they are hypotheses on what the run achieved, promoted by the usage penalty inside the λ-window of Theorem 1(b), and they are not vacuous, because on the noise event below the true configuration itself passes the fit condition with equality in usage (final claim of part (a)). Conditions (C5) and (C10) are promoted by the ambiguity penalty; (C6) is an occurrence-richness assumption on the sampled supports; (C7) states that no candidate set represents a sample's span with fewer directions than the truth (its role is discussed after the proofs); (C11) is the cardinality form of (C7) under noise. As with $s_{\min}$ and $\sigma_{\min}(S_A)$ (Appendix A), $\sigma_{\min}(\hat{S}_B)$ is a quantity computed from the learned pool while $\hat{s}_{\min}$ is an assumed constant; (C5) and (C10) assert that the computed quantity dominates the assumed margin.

Finally, slot indices in a learned pool are bookkeeping, not structure: permuting the slots of $\hat{S}$ together with the columns of the occurrence matrix changes no generated sample, so no data can distinguish two pools that differ by a permutation. Recovery of components is therefore stated up to a *relabeling* π: an injective map $\pi : J_{\text{act}} \to [N]$ assigning to each active true component i the pool slot $\pi(i)$ that represents it. Exact recovery then reads $\hat{s}_{\pi(i)} = s_i$ and $A_m = \pi(J_m)$; the relabeling is the only ambiguity the model leaves.

Theorem 2, restated: parts (a) and (b). The main text states this result under hypotheses given by reference; the full statement follows. Fix $\alpha \in (0,1)$ and $\varepsilon_{\text{fit}} \ge 0$. Suppose the data follow the model of Appendix B: $x^{(m)} = x_0^{(m)} + \varepsilon^{(m)}$ with $x_0^{(m)} = \sum_{i \in J_m} c_i^{(m)} s_i$, $|J_m| \le k$, $c_i^{(m)} \in [c_{\min}, c_{\max}]$, and noise as in the Notation. Let E denote the event that

$$\big\|\varepsilon^{(m)}\big\| \le \varepsilon_{\text{ns}}(\alpha) \qquad \text{and} \qquad \big\|P_{V_B}\, \varepsilon^{(m)}\big\| \le \varepsilon_{\text{prj}}(\alpha)$$

for every $m \in [M]$ and every $B \subseteq [N]$ with $|B| \le 2k$.

(a) (Known pool.) Suppose the pool $s_1, \dots, s_N$ is known and satisfies the spark condition with margin $s_{\min}$ (Definition 1), and suppose the fixed-pool window

$$\varepsilon_{\text{ns}}(\alpha) \;\le\; \varepsilon_{\text{fit}}, \qquad \varepsilon_{\text{fit}} + \varepsilon_{\text{prj}}(\alpha) \;<\; c_{\min}\, s_{\min}. \tag{9}$$

Then $\Pr(E) \ge 1 - \alpha$, and on E, for every candidate solution $(\{A_m\}, \{\hat{c}^{(m)}\})$ that passes the certificate (C1)–(C2), the following hold: *(i)* $A_m = J_m$ for every $m \in [M]$; *(ii)* $\big|\bigcup_m A_m\big| = K^*$; *(iii)* $\|\hat{c}^{(m)} - c^{(m)}\| \le (\varepsilon_{\text{fit}} + \varepsilon_{\text{prj}}(\alpha))/s_{\min}$ for every m. Moreover, on E the true configuration $(\{J_m\}, \{c^{(m)}\})$ passes the certificate: (C1) holds, and (C2) holds with equality.

(b) (Learned pool, noiseless.) Suppose $\sigma = 0$, so $x^{(m)} = x_0^{(m)}$, and suppose the true pool satisfies the spark condition with margin $s_{\min}$ (Definition 1). Consider any candidate solution $(\hat{S}, \{A_m\}, \{\hat{c}^{(m)}\})$ that passes the certificate (C3)–(C7). Then there is an injective $\pi : J_{\text{act}} \to [N]$ such that: *(i)* $\hat{s}_{\pi(i)} = s_i$ for every $i \in J_{\text{act}}$; *(ii)* $A_m = \pi(J_m)$ for every m; *(iii)* $\big|\bigcup_m A_m\big| = K^*$; *(iv)* $\hat{c}_{\pi(i)}^{(m)} = c_i^{(m)}$ for every m and every $i \in J_m$.

Conjecture R (isolating pairs), restated. The main text states this result under hypotheses given by reference; the full statement follows. Fix $\gamma_{\mathrm{c}} > 0$. For a support pattern J with $|J| = \varkappa$ and a candidate set $A \subseteq [N]$ with $|A| = \varkappa$, call the pair (J, A) γ_{c}-*spanned* if there exist samples $m_1, \dots, m_\varkappa \in [M]$ with

$$J_{m_t} = J \quad \text{and} \quad A_{m_t} = A \qquad \text{for every } t \in [\varkappa],$$

whose coefficient matrix $C = [\, c^{(m_1)} \cdots c^{(m_\varkappa)} \,] \in \mathbb{R}^{\varkappa \times \varkappa}$ (rows indexed by J) satisfies $\sigma_{\min}(C) \geq \gamma_{\mathrm{c}}$. Conjecture R asserts: under the model of Appendix B, the spark condition (Definition 1), and the robust certificate (C8)–(C11) with (C6), for every active $i \in J_{\mathrm{act}}$ there exists a pair of samples (m, m') with $J_m \cap J_{m'} = \{i\}$ such that (J_m, A_m) and $(J_{m'}, A_{m'})$ are both γ_{c}-spanned.

Conjecture R is open for $k \geq 2$; it is a theorem for $k = 1$ (Lemma 12), and its role vanishes at $\bar{\varepsilon} = 0$ (closing remarks). Part (c) is stated and proved conditionally on it.

Theorem 2, restated: part (c). Retain the setup and the event E above.

(c) (Learned pool, noisy; conditional on Conjecture R.) Suppose the true pool satisfies the spark condition with margin $s_{\min}$ (Definition 1), and consider candidate solutions $(\hat{S}, \{A_m\}, \{\hat{c}^{(m)}\})$ passing the robust certificate (C8)–(C11) with (C6). Assume Conjecture R holds with quantile $\gamma_{\mathrm{c}} > 0$. Define

$$f \;=\; \frac{k\,\bar{\varepsilon}}{s_{\min}\gamma_{\mathrm{c}} - \sqrt{k}\,\bar{\varepsilon}}, \qquad f' \;=\; f + \frac{2\sqrt{2k}\,f}{s_{\min}}, \qquad \kappa_{\mathrm{c}} \;=\; \frac{\bar{\varepsilon} + \sqrt{k}\,\hat{C} f'}{s_{\min}},$$

and suppose the robust window

$$\begin{gathered} \sqrt{k}\,\bar{\varepsilon} < s_{\min}\gamma_{\mathrm{c}}, \qquad \sqrt{2k}\,f < \hat{s}_{\min}, \qquad f'\Big(1 + \frac{2}{s_{\min}}\Big) < 1, \\ \kappa_{\mathrm{c}} < c_{\min}, \qquad \frac{\bar{\varepsilon} + k\,c_{\max} f'/(1-f')}{\hat{s}_{\min}} < \vartheta. \end{gathered} \tag{10}$$

Then $\Pr(E) \geq 1 - \alpha$, and on E, every candidate solution satisfying the robust certificate admits an injective $\pi : J_{\mathrm{act}} \to [N]$ such that: *(i)* $\operatorname{dist}\big(\hat{s}_{\pi(i)}, \operatorname{span}(s_i)\big) \leq f'$ and $\langle \hat{s}_{\pi(i)}, s_i \rangle > 0$ for every $i \in J_{\mathrm{act}}$; *(ii)* $A_m = \pi(J_m)$ for every m; *(iii)* $\big|\bigcup_m A_m\big| = K^*$; *(iv)* $\big|\hat{c}^{(m)}_{\pi(i)} - c^{(m)}_i\big| \leq \kappa_{\mathrm{c}} + \hat{C} f'$ for every m and every $i \in J_m$.

Corollary 1 (Reuse), restated. The main text states this result under hypotheses given by reference; the full statement follows. Fix $\alpha \in (0,1)$ and $\varepsilon_{\mathrm{fit}} \geq 0$. Suppose the pool is $(s_i)_{i \in J_{\mathrm{act}}}$: either a known component library, or the K^* components recovered by part (b) and retained by the frozen model. Consider a fresh sample from the same generative process,

$$x \;=\; \sum_{i \in J} c_i s_i + \varepsilon, \qquad J \subseteq J_{\mathrm{act}}, \quad |J| \leq k, \quad c_i \in [c_{\min}, c_{\max}],$$

with ε sub-Gaussian with proxy σ (Notation). Write $\varepsilon^{(1)}_{\mathrm{ns}}(\alpha)$ and $\varepsilon^{(1)}_{\mathrm{prj}}(\alpha)$ for the levels of the Notation evaluated at $M = 1$, with N replaced by K^* in n_{2k}, and suppose the window (9) holds with these levels. Then with probability at least $1 - \alpha$ over ε, every decoding $(A, \hat{c})$ of x alone with

$$|A| \leq k, \qquad \|x - S_A \hat{c}\| \leq \varepsilon_{\mathrm{fit}}, \qquad |A| \leq |J|,$$

satisfies $A = J$ and $\|\hat{c} - c\| \leq (\varepsilon_{\mathrm{fit}} + \varepsilon^{(1)}_{\mathrm{prj}}(\alpha))/s_{\min}$.

The proofs follow in five layers; each part's supporting layers immediately precede its assembly. Layer I confines all randomness to one event. Layer II (detection and counting) proves part (a)'s support recovery, and Layer III adds coefficient stability. Layer IV (span geometry of a learned pool) proves part (b). Layer V (robust counterparts) proves part (c), and Corollary 1 closes.

Layer I: confining the randomness (Lemma 1)

This layer packs everything probabilistic into one event E of probability at least $1 - \alpha$: part (i) assembles E from a net bound and a union bound; part (ii) extracts the inner-product form in which the event is consumed. Every argument after this layer is deterministic on E.

Lemma 1 (The recovery noise event)**.** *Let E be the event: for every $m \in [M]$ and every $B \subseteq [N]$ with $|B| \le 2k$,*

$$\|\varepsilon^{(m)}\| \le \varepsilon_{\mathrm{ns}}(\alpha) \qquad \textit{and} \qquad \|P_{V_B}\,\varepsilon^{(m)}\| \le \varepsilon_{\mathrm{prj}}(\alpha).$$

Then: (i) $\Pr(E) \ge 1-\alpha$*;* (ii) *on E, for every $m \in [M]$, every $B \subseteq [N]$ with $|B| \le 2k$, and every coefficient vector v supported in B with $y = \sum_j v_j s_j$,*

$$\big|\langle \varepsilon^{(m)},\, y\rangle\big| \;\le\; \varepsilon_{\mathrm{prj}}(\alpha)\,\|y\|. \tag{11}$$

Proof. Step 1 (net). Fix a subspace $V \subseteq \mathbb{R}^d$ with $\dim V = d'$, chosen independently of the noise, and write ε for one sample's noise. Fix a $\frac{1}{2}$-net $\mathcal{N}$ of the unit sphere of V with

$$|\mathcal{N}| \;\le\; 5^{d'}$$

(Vershynin, 2018, §4.2).

Step 2 (net comparison). For any unit $u \in V$ pick $\bar{u} \in \mathcal{N}$ with $\|u-\bar{u}\| \le \frac{1}{2}$. Since $u-\bar{u} \in V$,

$$\langle u, \varepsilon\rangle \;=\; \langle \bar{u}, \varepsilon\rangle + \langle u-\bar{u},\, P_V\varepsilon\rangle \;\le\; \langle \bar{u}, \varepsilon\rangle + \tfrac{1}{2}\|P_V\varepsilon\|.$$

The supremum of $\langle u, \varepsilon\rangle$ over unit $u \in V$ equals $\|P_V\varepsilon\|$, so

$$\|P_V\varepsilon\| \;\le\; \max_{\bar{u}\in\mathcal{N}}\langle \bar{u}, \varepsilon\rangle + \tfrac{1}{2}\|P_V\varepsilon\|, \qquad \text{hence} \qquad \|P_V\varepsilon\| \;\le\; 2\max_{\bar{u}\in\mathcal{N}}\langle \bar{u}, \varepsilon\rangle.$$

Step 3 (tail at one net point). By the Chernoff bound from the moment-generating condition (Notation) (Vershynin, 2018, §2.5), for every $t > 0$,

$$\Pr\big(\langle \bar{u}, \varepsilon\rangle \ge t\big) \;\le\; e^{-t^2/2\sigma^2}.$$

Step 4 (projected noise on one fixed subspace). Set $t = \sigma\sqrt{2\log(5^{d'}/\delta)}$ for $\delta \in (0,1)$. By Steps 1 and 3,

$$\Pr\Big(\max_{\bar{u}\in\mathcal{N}}\langle \bar{u}, \varepsilon\rangle \ge t\Big) \;\le\; 5^{d'}\,e^{-\log(5^{d'}/\delta)} \;=\; \delta,$$

and on the complementary event Step 2 gives

$$\|P_V\varepsilon\| \;\le\; 2t \;=\; 2\sigma\sqrt{2\big(d'\log 5 + \log\tfrac{1}{\delta}\big)},$$

a bound nondecreasing in d'.

Step 5 (norm bounds). Apply Step 4 with $V = \mathbb{R}^d$ (P_V the identity, $d' = d$) and $\delta = \alpha/(2M)$, once for each m: each bound $\|\varepsilon^{(m)}\| \le \varepsilon_{\mathrm{ns}}(\alpha)$ fails with probability at most $\alpha/(2M)$.

Step 6 (projection bounds). For each pair (m, B), the subspace V_B is fixed: the pool is known, so V_B does not depend on the noise. Apply Step 4 with $V = V_B$, $d' = \dim V_B \le 2k$, and $\delta = \alpha/(2Mn_{2k})$: each bound $\|P_{V_B}\varepsilon^{(m)}\| \le \varepsilon_{\mathrm{prj}}(\alpha)$ fails with probability at most $\alpha/(2Mn_{2k})$.

Step 7 (union bound; part (i)). There are M events in Step 5 and at most Mn_{2k} in Step 6, so

$$\Pr(E^{\mathrm{c}}) \;\le\; M\cdot\frac{\alpha}{2M} + Mn_{2k}\cdot\frac{\alpha}{2Mn_{2k}} \;=\; \alpha.$$

Step 8 (inner-product form; part (ii)). On E, since $y \in V_B$,

$$\big|\langle \varepsilon^{(m)}, y\rangle\big| \;=\; \big|\langle P_{V_B}\,\varepsilon^{(m)},\, y\rangle\big| \;\le\; \|P_{V_B}\,\varepsilon^{(m)}\|\,\|y\| \;\le\; \varepsilon_{\mathrm{prj}}(\alpha)\,\|y\|$$

by Cauchy–Schwarz and the event bound. Inequality (11) is the only consequence of E used in Lemmas 2 and 4: the probabilistic layer ends here, and everything after it is deterministic on E. □

LAYER II: DETECTION AND COUNTING (LEMMAS 2 AND 3)

This layer proves part (a)'s mechanism: a candidate support missing a true component leaves a residual too large to hide (detection), and the certificate's usage condition then forces every support to match (counting). The usage condition is a hypothesis on the candidate solution, not on the data: it encodes that the search achieved the frugality the objective pursues. Theorem 1(a) makes it achievable (the truth satisfies it with equality), Theorem 1(b) makes it pursued (inside the window the penalized objective strictly prefers the true usage), and this layer shows it is exactly what converts detection into support equality.

Lemma 2 (Detection of missing components)**.** *On E: suppose a support $A \subseteq [N]$ with $|A| \le k$ misses at least one true component of sample m, that is, $J_m \setminus A \neq \emptyset$. Then for every coefficient vector $\hat{c}$ supported on A,*

$$\left\|x^{(m)} - S_A \hat{c}\right\| \;\ge\; c_{\min}\, s_{\min} - \varepsilon_{\mathrm{prj}}(\alpha).$$

Proof. Step 1 (residual inequality). For any $y, \varepsilon \in \mathbb{R}^d$ and $b \ge 0$ with $|\langle \varepsilon, y\rangle| \le b\|y\|$, it holds that $\|y+\varepsilon\| \ge \|y\| - b$: for $y = 0$ this reads $\|\varepsilon\| \ge -b$, true since $b \ge 0$; for $y \neq 0$, Cauchy–Schwarz and bilinearity give

$$\|y+\varepsilon\|\,\|y\| \;\ge\; \langle y+\varepsilon,\, y\rangle \;=\; \|y\|^2 + \langle \varepsilon, y\rangle \;\ge\; \|y\|^2 - b\|y\|,$$

and dividing by $\|y\| > 0$ yields the claim.

Step 2 (mixed coefficient vector). Define, for $j \in [N]$,

$$v_j \;=\; \mathbf{1}_{j\in J_m}\, c_j^{(m)} - \mathbf{1}_{j\in A}\, \hat{c}_j, \qquad y \;=\; \sum_j v_j s_j.$$

Then $y = x_0^{(m)} - S_A\hat{c}$, the vector v is supported in $B = J_m \cup A$, and $|B| \le |J_m| + |A| \le 2k$.

Step 3 (coordinate mass). Pick $j_0 \in J_m \setminus A$. Then $v_{j_0} = c_{j_0}^{(m)} \ge c_{\min}$, hence

$$\sum_j v_j^2 \;\ge\; v_{j_0}^2 \;\ge\; c_{\min}^2.$$

Step 4 (spark). By the spark condition on B ($|B| \le 2k$) and Step 3,

$$\|y\| \;\ge\; s_{\min}\sqrt{\textstyle\sum_j v_j^2} \;\ge\; s_{\min}\, c_{\min}.$$

Step 5 (noise). Lemma 1(ii) applied to (B, v) gives

$$\left|\langle \varepsilon^{(m)},\, y\rangle\right| \;\le\; \varepsilon_{\mathrm{prj}}(\alpha)\,\|y\|.$$

Step 6 (conclusion). Since $x^{(m)} - S_A\hat{c} = y + \varepsilon^{(m)}$, Step 1 with $b = \varepsilon_{\mathrm{prj}}(\alpha) \ge 0$ and Step 4 give

$$\left\|x^{(m)} - S_A\hat{c}\right\| \;\ge\; \|y\| - \varepsilon_{\mathrm{prj}}(\alpha) \;\ge\; c_{\min}\, s_{\min} - \varepsilon_{\mathrm{prj}}(\alpha). \qquad \square$$

Lemma 3 (Usage counting)**.** (i) *If $|J_m| \le |A_m|$ for every m and $\sum_m |A_m| \le \sum_m |J_m|$, then $|A_m| = |J_m|$ for every m.* (ii) *If moreover $J_m \subseteq A_m$ for every m, then $A_m = J_m$ for every m.*

Proof. Step 1 (cardinalities). Suppose $|J_{m_0}| < |A_{m_0}|$ for some m_0. Summing this strict inequality with $|J_m| \le |A_m|$ over all other m,

$$\sum_m |J_m| \;<\; \sum_m |A_m|,$$

contradicting the usage bound. Hence $|A_m| = |J_m|$ for every m.

Step 2 (sets). If $J_m \subseteq A_m$ and $|A_m| = |J_m|$, a containment between finite sets of equal cardinality is an equality: $A_m = J_m$. $\square$

LAYER III: COEFFICIENT STABILITY (LEMMA 4)

Once supports match, the spark condition pins every concentration to noise-floor accuracy; part (a)'s assembly follows.

Lemma 4 (Coefficient stability)**.** *On E: suppose sample m is fitted on the true support, $\|x^{(m)} - S_{J_m}\hat{c}^{(m)}\| \le \varepsilon_{\text{fit}}$. Then*

$$s_{\min}\,\|\hat{c}^{(m)} - c^{(m)}\| \;\le\; \varepsilon_{\text{fit}} + \varepsilon_{\text{prj}}(\alpha).$$

Proof. Step 1 (signal part of the residual). Set

$$v_j \;=\; \mathbf{1}_{j\in J_m}\big(c_j^{(m)} - \hat{c}_j^{(m)}\big), \qquad y \;=\; \sum_j v_j s_j \;=\; S_{J_m}\big(c^{(m)} - \hat{c}^{(m)}\big),$$

so that $x^{(m)} - S_{J_m}\hat{c}^{(m)} = y + \varepsilon^{(m)}$ and v is supported in J_m, $|J_m| \le k \le 2k$.

Step 2 (degenerate case). If $y = 0$, then the spark bound of Step 5 reads $s_{\min}\|\hat{c}^{(m)} - c^{(m)}\| \le \|y\| = 0$, and the conclusion holds because $\varepsilon_{\text{fit}} + \varepsilon_{\text{prj}}(\alpha) \ge 0$.

Step 3 (Cauchy–Schwarz with the fit). If $y \neq 0$,

$$\langle y + \varepsilon^{(m)},\, y\rangle \;\le\; \|y + \varepsilon^{(m)}\|\,\|y\| \;\le\; \varepsilon_{\text{fit}}\,\|y\|.$$

Step 4 (bilinearity and noise). By Lemma 1(ii) applied to (J_m, v),

$$\langle y + \varepsilon^{(m)},\, y\rangle \;=\; \|y\|^2 + \langle \varepsilon^{(m)}, y\rangle \;\ge\; \|y\|^2 - \varepsilon_{\text{prj}}(\alpha)\,\|y\|.$$

Combining with Step 3 and dividing by $\|y\| > 0$,

$$\|y\| \;\le\; \varepsilon_{\text{fit}} + \varepsilon_{\text{prj}}(\alpha).$$

Step 5 (spark conversion). By the spark condition on J_m,

$$s_{\min}\,\|\hat{c}^{(m)} - c^{(m)}\| \;=\; s_{\min}\sqrt{\textstyle\sum_j v_j^2} \;\le\; \|y\| \;\le\; \varepsilon_{\text{fit}} + \varepsilon_{\text{prj}}(\alpha). \qquad \square$$

Proof of part (a). Step 1 (the event). By Lemma 1(i), $\Pr(E) \ge 1 - \alpha$. Fix a realization in E; the remaining steps are deterministic and hold simultaneously for every candidate solution passing (C1)–(C2), since E does not depend on the candidate.

Step 2 (containment). Fix m and suppose $J_m \not\subseteq A_m$, so $J_m \setminus A_m \neq \emptyset$. By (C1), Lemma 2, and the right edge of (9),

$$\varepsilon_{\text{fit}} \;\ge\; \big\|x^{(m)} - S_{A_m}\hat{c}^{(m)}\big\| \;\ge\; c_{\min}\,s_{\min} - \varepsilon_{\text{prj}}(\alpha) \;>\; \varepsilon_{\text{fit}},$$

a contradiction. Hence $J_m \subseteq A_m$ for every m.

Step 3 (counting). Step 2 and (C2) are the hypotheses of Lemma 3(ii), so $A_m = J_m$ for every m: part (i). Taking unions,

$$\bigcup_m A_m \;=\; \bigcup_m J_m \;=\; J_{\text{act}}, \qquad \Big|\bigcup_m A_m\Big| \;=\; K^*:$$

part (ii).

Step 4 (coefficients). With $A_m = J_m$, condition (C1) is the hypothesis of Lemma 4, so

$$\|\hat{c}^{(m)} - c^{(m)}\| \;\le\; \frac{\varepsilon_{\text{fit}} + \varepsilon_{\text{prj}}(\alpha)}{s_{\min}}:$$

part (iii).

Step 5 (achievability). On E, by the left edge of (9),

$$\big\|x^{(m)} - S_{J_m}c^{(m)}\big\| \;=\; \|\varepsilon^{(m)}\| \;\le\; \varepsilon_{\text{ns}}(\alpha) \;\le\; \varepsilon_{\text{fit}},$$

and $|J_m| \le k$, so the true configuration passes (C1); its total usage equals $\sum_m |J_m|$, so it meets (C2) with equality. $\square$

LAYER IV: SPAN GEOMETRY OF A LEARNED POOL (LEMMAS 5–7)

This layer proves part (b)'s mechanism, noiseless: tight usage forces every used candidate component into the true spans, and singleton-intersection patterns pin each component individually, producing the relabeling π; non-negativity resolves the signs at the last step. Wherever a lemma below assumes linearly independent profiles, the assumption is supplied by the spark condition: the index sets involved have at most $2k$ elements, and a vanishing nontrivial combination over such a set would force $\sigma_{\min} = 0 < s_{\min}$, contradicting Definition 1.

Lemma 5 (Span intersection along an independent family)**.** *Let F, F' be finite index sets and let $(v_\iota)_{\iota \in F \cup F'}$ be linearly independent vectors in $\mathbb{R}^d$. Then*

$$\operatorname{span}(v_\iota : \iota \in F) \cap \operatorname{span}(v_\iota : \iota \in F') \;=\; \operatorname{span}(v_\iota : \iota \in F \cap F').$$

Proof. Step 1 ($\supseteq$). Monotonicity of the span: $F \cap F' \subseteq F$ and $F \cap F' \subseteq F'$.

Step 2 ($\subseteq$). Suppose

$$w \;=\; \sum_{\iota \in F} a_\iota v_\iota \;=\; \sum_{\iota \in F'} b_\iota v_\iota.$$

Subtracting,

$$\sum_{\iota \in F \setminus F'} a_\iota v_\iota + \sum_{\iota \in F \cap F'} (a_\iota - b_\iota) v_\iota - \sum_{\iota \in F' \setminus F} b_\iota v_\iota \;=\; 0,$$

a vanishing combination over $F \cup F'$, so independence forces $a_\iota = 0$ on $F \setminus F'$, $b_\iota = 0$ on $F' \setminus F$, and $a_\iota = b_\iota$ on $F \cap F'$. Hence $w = \sum_{\iota \in F \cap F'} a_\iota v_\iota$. □

Lemma 6 (Tight usage and exact spans)**.** *Assume $\sigma = 0$, the spark condition on the true pool (Definition 1), and (C3)–(C5) and (C7). Then for every m,*

$$|A_m| \;=\; |J_m|, \qquad \operatorname{span}\big(\hat{S}_{A_m}\big) \;=\; V_{J_m}.$$

In particular every used candidate profile lies exactly in the sample's true span, and the used profiles of sample m span it.

Proof. Step 1 (row-wise lower bound). For each m, by (C7),

$$|A_m| \;\geq\; \dim \operatorname{span}\big(\hat{S}_{A_m}\big) \;\geq\; \dim\big(\operatorname{span}(\hat{S}_{A_m}) \cap V_{J_m}\big) \;=\; |J_m|.$$

Step 2 (counting). Step 1 and (C4) are the hypotheses of Lemma 3(i), so $|A_m| = |J_m|$ for every m.

Step 3 (first rank squeeze). By (C5) the columns of $\hat{S}_{A_m}$ are linearly independent ($|A_m| \leq k \leq 2k$), so by Step 2 and (C7),

$$\dim \operatorname{span}\big(\hat{S}_{A_m}\big) \;=\; |A_m| \;=\; |J_m| \;=\; \dim\big(\operatorname{span}(\hat{S}_{A_m}) \cap V_{J_m}\big).$$

The intersection is a subspace of $\operatorname{span}(\hat{S}_{A_m})$ of the same finite dimension, hence equals it; therefore $\operatorname{span}(\hat{S}_{A_m}) \subseteq V_{J_m}$.

Step 4 (second rank squeeze). Independence of the true profiles on J_m (spark condition, $|J_m| \leq 2k$) gives $\dim V_{J_m} = |J_m|$, equal by Step 3 to the dimension of the contained subspace $\operatorname{span}(\hat{S}_{A_m})$; the containment of Step 3 is therefore an equality. □

Lemma 7 (Anchoring)**.** *Assume $\sigma = 0$, the spark condition on the true pool, and (C3)–(C7). Then for every active $i \in J_{\mathrm{act}}$ there is a pool index $\pi(i)$ with $\hat{s}_{\pi(i)} \in \operatorname{span}(s_i)$; writing $\hat{s}_{\pi(i)} = \tau_i s_i$, the scalar satisfies $\tau_i \in \{\pm 1\}$; and the map π is injective.*

Proof. Step 1 (the pair and the line). Fix i and take m, m' with $J_m \cap J_{m'} = \{i\}$ (C6). The true profiles indexed by $J_m \cup J_{m'}$ are independent ($|J_m \cup J_{m'}| \leq 2k$, spark condition), so Lemma 5 gives

$$V_{J_m} \cap V_{J_{m'}} \;=\; \operatorname{span}(s_i),$$

a subspace of dimension 1 ($s_i \neq 0$ by unit norm).

Step 2 (rank bookkeeping). Set $U = A_m \cup A_{m'}$. By Lemma 6, $|U| \le |A_m| + |A_{m'}| = |J_m| + |J_{m'}| \le 2k$, so (C5) makes the columns of $\hat{S}_U$ independent and

$$|U| \;=\; \dim \operatorname{span}\big(\hat{S}_U\big) \;=\; \dim\big(V_{J_m} + V_{J_{m'}}\big),$$

the second equality because $\operatorname{span}(\hat{S}_U) = \operatorname{span}(\hat{S}_{A_m}) + \operatorname{span}(\hat{S}_{A_{m'}}) = V_{J_m} + V_{J_{m'}}$ (Lemma 6). The dimension formula with Step 1 gives

$$\dim\big(V_{J_m} + V_{J_{m'}}\big) \;=\; |J_m| + |J_{m'}| - 1,$$

and hence, by inclusion–exclusion and Lemma 6,

$$|A_m \cap A_{m'}| \;=\; |A_m| + |A_{m'}| - |U| \;=\; 1 :$$

exactly one shared index j^*.

Step 3 (location and sign). The column $\hat{s}_{j^*}$ belongs to both spans, so by Lemma 6 and Step 1,

$$\hat{s}_{j^*} \;\in\; \operatorname{span}\big(\hat{S}_{A_m}\big) \cap \operatorname{span}\big(\hat{S}_{A_{m'}}\big) \;=\; V_{J_m} \cap V_{J_{m'}} \;=\; \operatorname{span}(s_i),$$

hence $\hat{s}_{j^*} = a\, s_i$ for a scalar a; unit norms give $|a| = 1$, and we set $\tau_i = a \in \{\pm 1\}$ and $\pi(i) = j^*$.

Step 4 (independence of the choice). The slot is unique: two distinct pool columns inside the line $\operatorname{span}(s_i)$ would form a linearly dependent pair, contradicting (C5) applied to that pair ($2 \le 2k$). Hence $\pi(i)$ does not depend on the pair (m, m') chosen in Step 1.

Step 5 (injectivity). If $\pi(i) = \pi(i')$ with $i \ne i'$, then $\tau_i s_i = \tau_{i'} s_{i'}$, that is, $s_{i'} = (\tau_{i'}^{-1}\tau_i)\, s_i$: a member of the pair $\{s_i, s_{i'}\}$, independent by the spark condition ($2 \le 2k$), written as a multiple of the other. This is impossible, so π is injective. □

Proof of part (b). If $J_{\mathrm{act}} = \emptyset$, then (C4) reads $\sum_m |A_m| \le 0$, every A_m is empty, and all claims hold with the empty relabeling; assume $J_{\mathrm{act}} \ne \emptyset$ from now on, whence $k \ge 1$. Let π and $\tau_i \in \{\pm 1\}$ be as in Lemma 7.

Step 1 (membership). Fix m and $i \in J_m$, and suppose $\pi(i) \notin A_m$. The pool subset $\{\pi(i)\} \cup A_m$ has cardinality $|A_m| + 1 \le k + 1 \le 2k$, hence its columns are independent by (C5). All of them lie in V_{J_m}: the columns of A_m by Lemma 6, and $\hat{s}_{\pi(i)} = \tau_i s_i \in V_{J_m}$ since $i \in J_m$. An independent family inside V_{J_m} has at most $\dim V_{J_m} = |J_m|$ members, so by Lemma 6,

$$|A_m| + 1 \;\le\; |J_m| \;=\; |A_m|,$$

which is absurd. Hence $\pi(i) \in A_m$ for every $i \in J_m$.

Step 2 (support identity). Step 1 gives $\pi(J_m) \subseteq A_m$, and by injectivity of π and Lemma 6,

$$|\pi(J_m)| \;=\; |J_m| \;=\; |A_m|,$$

so $A_m = \pi(J_m)$: part (ii).

Step 3 (coefficient extraction). Reindex the fit (C3) over $A_m = \pi(J_m)$ (π injective) and substitute $\hat{s}_{\pi(i)} = \tau_i s_i$:

$$\sum_{i \in J_m} \big(\hat{c}^{(m)}_{\pi(i)}\, \tau_i\big)\, s_i \;=\; x_0^{(m)} \;=\; \sum_{i \in J_m} c_i^{(m)}\, s_i.$$

The family $(s_i)_{i \in J_m}$ is independent (spark condition), so the coefficients coincide:

$$\hat{c}^{(m)}_{\pi(i)}\, \tau_i \;=\; c_i^{(m)} \qquad \text{for every } i \in J_m.$$

Step 4 (sign resolution). For each active i pick a sample m with $i \in J_m$ (the (C6) pair provides one). There $c_i^{(m)} > 0$, and $\hat{c}^{(m)}_{\pi(i)} > 0$ because $\pi(i) \in A_m$ (Step 1) and used coefficients are strictly positive. Step 3 then forces $\tau_i = +1$, so $\hat{s}_{\pi(i)} = s_i$: part (i); and Step 3 becomes $\hat{c}^{(m)}_{\pi(i)} = c_i^{(m)}$ for every m and $i \in J_m$: part (iv). Only positivity of the coefficients is used; $c_{\min}$ plays no quantitative role.

Step 5 (count). By Step 2 and injectivity of π,

$$\bigcup_m A_m \;=\; \pi\Big(\bigcup_m J_m\Big) \;=\; \pi\big(J_{\mathrm{act}}\big), \qquad \Big|\bigcup_m A_m\Big| \;=\; K^* :$$

part (iii). □

LAYER V: ROBUST COUNTERPARTS (LEMMAS 8–12)

This layer proves part (c): each exact-geometry fact of Layer IV degrades continuously with the fit error (a crude coefficient bound, a Weyl dimension count, a two-subspace pinch, and span proximity from coefficient-diverse fits). Conjecture R enters exactly once, to supply the span-proximity hypothesis for every active component; the 1-sparse case closes without it.

Lemma 8 (Crude coefficient bound)**.** *Assume (C8) and (C10), and suppose the signal-level fit* $\|x_0^{(m)} - \hat{S}_{A_m}\hat{c}^{(m)}\| \le \bar{\varepsilon}$ *holds for sample* m*. Then* $\|\hat{c}^{(m)}\| \le \hat{C} = (k\,c_{\max} + \bar{\varepsilon})/\hat{s}_{\min}$.

Proof. By (C10) with $|A_m| \le k \le 2k$, the triangle inequality, and $\|x_0^{(m)}\| \le \sum_{i\in J_m} c_i^{(m)}\|s_i\| \le k\,c_{\max}$,

$$\hat{s}_{\min}\,\|\hat{c}^{(m)}\| \;\le\; \sigma_{\min}\big(\hat{S}_{A_m}\big)\,\|\hat{c}^{(m)}\| \;\le\; \big\|\hat{S}_{A_m}\hat{c}^{(m)}\big\| \;\le\; \|x_0^{(m)}\| + \bar{\varepsilon} \;\le\; k\,c_{\max} + \bar{\varepsilon}. \qquad \square$$

Lemma 9 (Robust dimension count)**.** *Let* $B \subseteq [N]$ *with* $\sigma_{\min}(\hat{S}_B) \ge \hat{s}_{\min}$*, let* $W \subseteq \mathbb{R}^d$ *be a subspace of dimension* D*, and suppose every column satisfies* $\operatorname{dist}(\hat{s}_j, W) \le \eta$ *for* $j \in B$*. If* $\sqrt{|B|}\,\eta < \hat{s}_{\min}$*, then* $|B| \le D$.

Proof. Step 1 (rank). Suppose $|B| > D$. The projected matrix $P_W\hat{S}_B$ has rank at most $D < |B|$, so its smallest singular value (of $|B|$ many) vanishes: $\sigma_{\min}(P_W\hat{S}_B) = 0$.

Step 2 (column-wise error).

$$\big\|\hat{S}_B - P_W\hat{S}_B\big\|_F^2 \;=\; \sum_{j\in B}\|\hat{s}_j - P_W\hat{s}_j\|^2 \;=\; \sum_{j\in B}\operatorname{dist}(\hat{s}_j, W)^2 \;\le\; |B|\,\eta^2.$$

Step 3 (Weyl). Singular values are perturbed by at most the operator norm of the difference (Horn & Johnson, 2013, §7.3), so by Steps 1–2,

$$\hat{s}_{\min} \;\le\; \sigma_{\min}\big(\hat{S}_B\big) \;\le\; \sigma_{\min}\big(P_W\hat{S}_B\big) + \big\|\hat{S}_B - P_W\hat{S}_B\big\|_{\mathrm{op}} \;\le\; 0 + \sqrt{|B|}\,\eta \;<\; \hat{s}_{\min},$$

a contradiction. Hence $|B| \le D$. $\square$

Lemma 10 (Two-subspace pinch)**.** *Let* u *be a unit vector with* $\operatorname{dist}(u, V_J) \le \delta_1$ *and* $\operatorname{dist}(u, V_{J'}) \le \delta_2$*, where* $J \cap J' = \{i\}$*,* $|J \cup J'| \le 2k$*, and* $\sigma_{\min}(S_{J\cup J'}) \ge s_{\min}$*. Then*

$$\operatorname{dist}\big(u,\ \operatorname{span}(s_i)\big) \;\le\; \delta_1 + \sqrt{2k}\,\frac{\delta_1 + \delta_2}{s_{\min}}.$$

Proof. Step 1 (two decompositions). Take the least-squares decompositions

$$u \;=\; S_J a + r_1, \qquad u \;=\; S_{J'} b + r_2, \qquad \|r_1\| \le \delta_1, \quad \|r_2\| \le \delta_2.$$

Step 2 (difference through the spark). Subtracting, the coefficient vector v on $J \cup J'$ (entries a_j on $J \setminus \{i\}$, $a_i - b_i$ at i, $-b_j$ on $J' \setminus \{i\}$) satisfies $S_{J\cup J'}\,v = r_2 - r_1$, so

$$s_{\min}\,\|v\| \;\le\; \big\|S_{J\cup J'}\,v\big\| \;\le\; \delta_1 + \delta_2, \qquad \text{hence} \qquad \big\|a_{J\setminus\{i\}}\big\| \;\le\; \|v\| \;\le\; \frac{\delta_1 + \delta_2}{s_{\min}}.$$

Step 3 (drop the off-i part). By Step 1, $u - a_i s_i = S_{J\setminus\{i\}}\,a_{J\setminus\{i\}} + r_1$. A matrix with at most $2k$ unit columns has operator norm at most $\sqrt{2k}$, so by Step 2,

$$\operatorname{dist}\big(u, \operatorname{span}(s_i)\big) \;\le\; \|u - a_i s_i\| \;\le\; \sqrt{2k}\,\big\|a_{J\setminus\{i\}}\big\| + \delta_1 \;\le\; \delta_1 + \sqrt{2k}\,\frac{\delta_1 + \delta_2}{s_{\min}}. \qquad \square$$

Lemma 11 (Span proximity from coefficient-diverse fits)**.** *Fix a pattern* J *with* $|J| = \varkappa \le k$ *and* $\sigma_{\min}(S_J) \ge s$ *for some* $s > 0$*, and a candidate set* $A \subseteq [N]$ *with* $|A| = \varkappa$*. Suppose samples* $m_1, \dots, m_\varkappa$ *share the pattern and the candidate set,* $J_{m_t} = J$ *and* $A_{m_t} = A$ *for every* t*; that their coefficient matrix* $C = [\,c^{(m_1)} \cdots c^{(m_\varkappa)}\,]$ *has* $\sigma_{\min}(C) \ge \gamma > 0$*; and that each meets the signal-level fit* $\|S_J c^{(m_t)} - \hat{S}_A\hat{c}^{(m_t)}\| \le \bar{\varepsilon}$*. If* $\sqrt{\varkappa}\,\bar{\varepsilon} < s\,\gamma$*, then for every* $j \in A$*,*

$$\operatorname{dist}\big(\hat{s}_j,\ V_J\big) \;\le\; \frac{\varkappa\,\bar{\varepsilon}}{s\,\gamma - \sqrt{\varkappa}\,\bar{\varepsilon}}.$$

Proof. Step 1 (stacked fit). Write $\hat{C} = [\,\hat{c}^{(m_1)} \cdots \hat{c}^{(m_\varkappa)}\,]$ (rows indexed by A). Summing the squared fits,

$$\big\|S_J C - \hat{S}_A \hat{C}\big\|_F^2 \;=\; \sum_{t=1}^{\varkappa} \big\|S_J c^{(m_t)} - \hat{S}_A \hat{c}^{(m_t)}\big\|^2 \;\le\; \varkappa\,\bar{\varepsilon}^2.$$

Step 2 (projection off the span). Write $Q_J = I - P_{V_J}$. Since $Q_J S_J C = 0$ and $\|Q_J\|_{\mathrm{op}} \le 1$,

$$\big\|Q_J \hat{S}_A \hat{C}\big\|_F \;=\; \big\|Q_J\big(\hat{S}_A \hat{C} - S_J C\big)\big\|_F \;\le\; \sqrt{\varkappa}\,\bar{\varepsilon}.$$

Step 3 ($\hat{C}$ is well conditioned). First,

$$\sigma_{\min}\big(S_J C\big) \;\ge\; \sigma_{\min}(S_J)\,\sigma_{\min}(C) \;\ge\; s\,\gamma,$$

and singular values are perturbed by at most the operator norm of the difference (Horn & Johnson, 2013, §7.3), so by Step 1,

$$\sigma_{\min}\big(\hat{S}_A \hat{C}\big) \;\ge\; \sigma_{\min}\big(S_J C\big) - \big\|S_J C - \hat{S}_A \hat{C}\big\|_{\mathrm{op}} \;\ge\; s\,\gamma - \sqrt{\varkappa}\,\bar{\varepsilon} \;>\; 0.$$

Second, since $\hat{S}_A$ has $\varkappa$ unit columns, $\|\hat{S}_A\|_{\mathrm{op}} \le \sqrt{\varkappa}$ and

$$\sigma_{\min}\big(\hat{S}_A \hat{C}\big) \;\le\; \|\hat{S}_A\|_{\mathrm{op}}\,\sigma_{\min}\big(\hat{C}\big) \;\le\; \sqrt{\varkappa}\,\sigma_{\min}\big(\hat{C}\big).$$

Combining, $\hat{C}$ is invertible with

$$\big\|\hat{C}^{-1}\big\|_{\mathrm{op}} \;=\; \frac{1}{\sigma_{\min}(\hat{C})} \;\le\; \frac{\sqrt{\varkappa}}{s\,\gamma - \sqrt{\varkappa}\,\bar{\varepsilon}}.$$

Step 4 (column bound). For each $j \in A$, with e_j the corresponding coordinate vector, by Steps 2–3,

$$\begin{aligned}\mathrm{dist}\big(\hat{s}_j, V_J\big) \;=\; \big\|Q_J \hat{S}_A e_j\big\| \;&=\; \big\|\big(Q_J \hat{S}_A \hat{C}\big)\big(\hat{C}^{-1} e_j\big)\big\| \\ &\le\; \big\|Q_J \hat{S}_A \hat{C}\big\|_{\mathrm{op}}\,\big\|\hat{C}^{-1}\big\|_{\mathrm{op}} \;\le\; \frac{\varkappa\,\bar{\varepsilon}}{s\,\gamma - \sqrt{\varkappa}\,\bar{\varepsilon}}.\end{aligned}$$

This is the quantitative reason cancellation cannot persist: a candidate set can cancel its out-of-span parts against one coefficient vector, but not against $\varkappa$ linearly independent ones; the stacked constraint of Step 2 controls the whole matrix, hence each column. □

Lemma 12 (Conjecture R holds for $k = 1$)**.** *Let $k = 1$, let the candidate solution pass (C8)–(C11) with (C6), and suppose the signal-level fit $\|x_0^{(m)} - \hat{S}_{A_m}\hat{c}^{(m)}\| \le \bar{\varepsilon}$ holds with $\bar{\varepsilon} < c_{\min}$. Then the conclusion of Conjecture R holds with quantile $\gamma_{\mathrm{c}} = c_{\min}$; indeed*

$$\mathrm{dist}\big(\hat{s}_j,\, V_{J_m}\big) \;\le\; \frac{\bar{\varepsilon}}{c_{\min} - \bar{\varepsilon}} \qquad \textit{for every } m \textit{ and the single } j \in A_m.$$

Proof. Step 1 (tight cardinalities). (C11) and (C9) are the hypotheses of Lemma 3(i), so $|A_m| = |J_m| = 1$ for every m with $J_m \neq \emptyset$.

Step 2 (each sample is its own witness). Fix m with $J_m = \{i\}$. The pair (J_m, A_m) is $c_{\min}$-spanned by the single sample m itself: the coefficient matrix is the 1×1 matrix $C = (c_i^{(m)})$ with $\sigma_{\min}(C) = c_i^{(m)} \ge c_{\min}$, and $\sigma_{\min}(S_{J_m}) = \|s_i\| = 1$. In particular, for every active i, any (C6) pair (m, m') satisfies the conclusion of Conjecture R with $\gamma_{\mathrm{c}} = c_{\min}$.

Step 3 (span proximity). Lemma 11 with $\varkappa = 1$, $s = 1$, $\gamma = c_{\min}$, and the hypothesis $\bar{\varepsilon} < c_{\min}$ gives, for the single $j \in A_m$,

$$\mathrm{dist}\big(\hat{s}_j, V_{J_m}\big) \;\le\; \frac{\bar{\varepsilon}}{c_{\min} - \bar{\varepsilon}}.$$ □

Proof of part (c). Step 1 (signal-level fit on the event). By Lemma 1(i), $\Pr(E) \ge 1 - \alpha$; fix a realization in E. Only the norm bounds $\|\varepsilon^{(m)}\| \le \varepsilon_{\mathrm{ns}}(\alpha)$ of E are used in this part: the learned

directions depend on the noise, so the fixed-subspace bounds of E do not apply to them. By (C8) and the triangle inequality, for every m,

$$\big\|x_0^{(m)} - \hat{S}_{A_m}\hat{c}^{(m)}\big\| \;\le\; \big\|x^{(m)} - \hat{S}_{A_m}\hat{c}^{(m)}\big\| + \|\varepsilon^{(m)}\| \;\le\; \varepsilon_{\mathrm{fit}} + \varepsilon_{\mathrm{ns}}(\alpha) \;=\; \bar{\varepsilon}.$$

Step 2 (tight cardinalities). (C11) and (C9) are the hypotheses of Lemma 3(i), so $|A_m| = |J_m|$ for every m.

Step 3 (anchoring). Fix active i and let (m, m') be the pair provided by Conjecture R, so that (J_m, A_m) and $(J_{m'}, A_{m'})$ are γ_{c}-spanned. For (J_m, A_m), with $\varkappa = |J_m| = |A_m|$ (Step 2): the witnessing samples obey the Step-1 fit, the spark condition gives $\sigma_{\min}(S_{J_m}) \ge s_{\min}$, and the first condition of (10) gives $\sqrt{\varkappa}\,\bar{\varepsilon} \le \sqrt{k}\,\bar{\varepsilon} < s_{\min}\gamma_{\mathrm{c}}$. Lemma 11 (with $s = s_{\min}$, $\gamma = \gamma_{\mathrm{c}}$) therefore yields, for every $j \in A_m$,

$$\mathrm{dist}\big(\hat{s}_j, V_{J_m}\big) \;\le\; \frac{\varkappa\,\bar{\varepsilon}}{s_{\min}\gamma_{\mathrm{c}} - \sqrt{\varkappa}\,\bar{\varepsilon}} \;\le\; \frac{k\,\bar{\varepsilon}}{s_{\min}\gamma_{\mathrm{c}} - \sqrt{k}\,\bar{\varepsilon}} \;=\; f,$$

the second inequality because the bound is nondecreasing in $\varkappa$ and $\varkappa \le k$; likewise $\mathrm{dist}(\hat{s}_j, V_{J_{m'}}) \le f$ for every $j \in A_{m'}$.

Set $U = A_m \cup A_{m'}$ and $W = V_{J_m \cup J_{m'}}$. Every column indexed by U lies within f of W (as $V_{J_m}, V_{J_{m'}} \subseteq W$), and independence of the true profiles on $J_m \cup J_{m'}$ (spark condition, $|J_m \cup J_{m'}| \le 2k$) with Lemma 5 gives $V_{J_m} \cap V_{J_{m'}} = \mathrm{span}(s_i)$ and

$$\dim W \;=\; |J_m| + |J_{m'}| - 1 \;=:\; D.$$

By (C10), $\sigma_{\min}(\hat{S}_U) \ge \hat{s}_{\min}$ ($|U| \le 2k$), and the second condition of (10) gives $\sqrt{|U|}\,f \le \sqrt{2k}\,f < \hat{s}_{\min}$; Lemma 9 with $\eta = f$ yields $|U| \le D$. Inclusion–exclusion with Step 2 then gives

$$|A_m \cap A_{m'}| \;=\; |A_m| + |A_{m'}| - |U| \;\ge\; |J_m| + |J_{m'}| - D \;=\; 1.$$

Pick $j^* \in A_m \cap A_{m'}$; it lies within f of both V_{J_m} and $V_{J_{m'}}$, so Lemma 10 (with $\delta_1 = \delta_2 = f$) gives

$$\mathrm{dist}\big(\hat{s}_{j^*},\, \mathrm{span}(s_i)\big) \;\le\; f + \sqrt{2k}\,\frac{2f}{s_{\min}} \;=\; f'.$$

Set $\pi(i) = j^*$.

Step 4 (injectivity). Suppose $\pi(i) = \pi(i') = j$ with $i \ne i'$. Step 3 provides scalars a, a' with $\|\hat{s}_j - a\,s_i\| \le f'$ and $\|\hat{s}_j - a'\,s_{i'}\| \le f'$, so $\|a\,s_i - a'\,s_{i'}\| \le 2f'$, and the spark condition on the pair $\{i, i'\}$ gives

$$s_{\min}\sqrt{a^2 + a'^2} \;\le\; 2f', \qquad \text{hence} \qquad |a| \;\le\; \frac{2f'}{s_{\min}}.$$

Then, by the third condition of (10),

$$1 \;=\; \|\hat{s}_j\| \;\le\; |a| + \|\hat{s}_j - a\,s_i\| \;\le\; \frac{2f'}{s_{\min}} + f' \;=\; f'\Big(1 + \frac{2}{s_{\min}}\Big) \;<\; 1,$$

a contradiction. Hence π is injective.

Step 5 (supports by floor exclusion). Fix m. For $i \in J_m$ write the orthogonal decomposition

$$\hat{s}_{\pi(i)} \;=\; a_i s_i + e_i, \qquad e_i \perp s_i, \qquad \|e_i\| = \mathrm{dist}\big(\hat{s}_{\pi(i)}, \mathrm{span}(s_i)\big) \le f',$$

so $a_i^2 = 1 - \|e_i\|^2$ and $|a_i| \ge \sqrt{1 - f'^2} \ge 1 - f' > 0$ (the third condition of (10) forces $f' < 1$). The sign of a_i is fixed only in Step 6 and is not used here. Substituting $s_i = (\hat{s}_{\pi(i)} - e_i)/a_i$,

$$x_0^{(m)} \;=\; \hat{S}_{\pi(J_m)}\,\tilde{c} + r, \qquad \tilde{c}_i = \frac{c_i^{(m)}}{a_i}, \qquad r = -\sum_{i \in J_m} \tilde{c}_i\, e_i, \qquad \|r\| \;\le\; \frac{k\,c_{\max} f'}{1 - f'} \;=:\; \bar{r}.$$

With Step 1 and the triangle inequality,

$$\big\|\hat{S}_{A_m}\hat{c}^{(m)} - \hat{S}_{\pi(J_m)}\,\tilde{c}\big\| \;\le\; \bar{\varepsilon} + \bar{r}.$$

The difference is $\hat{S}_B v$ on $B = A_m \cup \pi(J_m)$, $|B| \le 2k$ (Steps 2 and 4), where v carries $\hat{c}_j^{(m)}$ on $A_m \setminus \pi(J_m)$, $\hat{c}_j^{(m)} - \tilde{c}_{\pi^{-1}(j)}$ on the intersection, and $-\tilde{c}_{\pi^{-1}(j)}$ on $\pi(J_m) \setminus A_m$. By (C10),

$$\|v\| \;\le\; \frac{\bar{\varepsilon} + \bar{r}}{\hat{s}_{\min}}.$$

If some $j \in A_m \setminus \pi(J_m)$ existed, its entry would give, by the (C8) floor and the fifth condition of (10),

$$\vartheta \;\le\; \hat{c}_j^{(m)} \;=\; |v_j| \;\le\; \|v\| \;\le\; \frac{\bar{\varepsilon} + k\, c_{\max} f'/(1-f')}{\hat{s}_{\min}} \;<\; \vartheta,$$

a contradiction. Hence $A_m \subseteq \pi(J_m)$, and $|A_m| = |J_m| = |\pi(J_m)|$ (Steps 2 and 4) forces $A_m = \pi(J_m)$: part (ii). Taking unions, by injectivity,

$$\bigcup_m A_m \;=\; \pi\big(J_{\mathrm{act}}\big), \qquad \Big|\bigcup_m A_m\Big| \;=\; K^* :$$

part (iii).

Step 6 (coefficients and signs). Fix m. Substituting $A_m = \pi(J_m)$ and $\hat{s}_{\pi(i)} = a_i s_i + e_i$ into the Step-1 fit,

$$x_0^{(m)} - \hat{S}_{A_m} \hat{c}^{(m)} \;=\; S_{J_m}\big(c^{(m)} - \tilde{a}\big) - \sum_{i \in J_m} \hat{c}_{\pi(i)}^{(m)}\, e_i, \qquad \tilde{a}_i = a_i\, \hat{c}_{\pi(i)}^{(m)}.$$

By Step 1, the triangle inequality, Cauchy–Schwarz, and Lemma 8,

$$\big\|S_{J_m}\big(c^{(m)} - \tilde{a}\big)\big\| \;\le\; \bar{\varepsilon} + \sum_{i \in J_m} \hat{c}_{\pi(i)}^{(m)}\, \|e_i\| \;\le\; \bar{\varepsilon} + \sqrt{k}\, \big\|\hat{c}^{(m)}\big\|\, f' \;\le\; \bar{\varepsilon} + \sqrt{k}\, \hat{C} f',$$

so the spark condition on J_m gives, entrywise,

$$\big|c_i^{(m)} - a_i\, \hat{c}_{\pi(i)}^{(m)}\big| \;\le\; \big\|c^{(m)} - \tilde{a}\big\| \;\le\; \frac{\bar{\varepsilon} + \sqrt{k}\, \hat{C} f'}{s_{\min}} \;=\; \kappa_{\mathrm{c}}.$$

By the fourth condition of (10),

$$a_i\, \hat{c}_{\pi(i)}^{(m)} \;\ge\; c_i^{(m)} - \kappa_{\mathrm{c}} \;\ge\; c_{\min} - \kappa_{\mathrm{c}} \;>\; 0,$$

and $\hat{c}_{\pi(i)}^{(m)} \ge \vartheta > 0$, so $a_i > 0$; with Step 5, $a_i \in [1 - f', 1]$, and $\langle \hat{s}_{\pi(i)}, s_i \rangle = a_i > 0$ together with Step 3 proves part (i). Finally, using $\hat{c}_{\pi(i)}^{(m)} \le \|\hat{c}^{(m)}\| \le \hat{C}$ (Lemma 8) and $1 - a_i \le f'$,

$$\big|\hat{c}_{\pi(i)}^{(m)} - c_i^{(m)}\big| \;\le\; (1 - a_i)\, \hat{c}_{\pi(i)}^{(m)} + \big|a_i\, \hat{c}_{\pi(i)}^{(m)} - c_i^{(m)}\big| \;\le\; \hat{C} f' + \kappa_{\mathrm{c}} :$$

part (iv). □

Proof of Corollary 1. Step 1 (the retained pool inherits the spark condition). When the pool is the recovered one, part (b)(i) gives $\hat{s}_{\pi(i)} = s_i$ exactly for every $i \in J_{\mathrm{act}}$, so the retained columns are the true active profiles; when the pool is a known component library this holds by definition. Every $B \subseteq J_{\mathrm{act}}$ with $|B| \le 2k$ is a subset of the full true pool, so $\sigma_{\min}(S_B) \ge s_{\min}$ by Definition 1: the retained pool satisfies the spark condition with the same margin.

Step 2 (part (a) at $M = 1$). The fresh sample obeys the model of part (a) with $M = 1$ over the retained pool, and the stated conditions on $(A, \hat{c})$ are exactly (C1)–(C2) at $M = 1$ (there, $\sum_m |A_m| = |A|$ and $\sum_m |J_m| = |J|$). The window (9) holds with the $M = 1$ levels by hypothesis, so part (a) applies: with probability at least $1 - \alpha$ over ε, every such decoding satisfies

$$A \;=\; J, \qquad \|\hat{c} - c\| \;\le\; \frac{\varepsilon_{\mathrm{fit}} + \varepsilon_{\mathrm{prj}}^{(1)}(\alpha)}{s_{\min}}.$$ □

Three remarks close the appendix. *What each part consumes.* Part (a) nowhere uses nonnegativity of the candidate coefficients: the spark condition and parsimony alone carry the known-pool result,

which therefore holds against sign-unconstrained decoders. Part (b) consumes less than its statement offers: its proof uses only positivity of the used coefficients ($c_{\min}$ enters nowhere) and only linear independence of the subsets named in (C5) and Definition 1 (neither $\hat{s}_{\min}$ nor $s_{\min}$ enters quantitatively); positivity acts exactly once, in Step 4, to fix signs, and without it part (b)(i) weakens to $\hat{s}_{\pi(i)} = \pm s_i$. The quantitative margins return in part (c), where every bound is explicit in $\bar{\varepsilon}$. Part (c) also pays a $\sqrt{d}$-order noise constant through $\varepsilon_{\mathrm{ns}}(\alpha)$: the learned directions depend on the noise, so the fixed-direction projection bounds that sharpen part (a) (where detection costs only $\varepsilon_{\mathrm{prj}}(\alpha)$, of order $\sigma\sqrt{k \log N + \log M}$) are unavailable there.

The role of (C7) and (C5). Without (C7), recovery fails outright: with $N \geq M$ pool slots, the candidate pool $\hat{s}_m = x_0^{(m)}/\|x_0^{(m)}\|$ with $A_m = \{m\}$ satisfies (C3) and (C4) at total usage $M \leq \sum_m |J_m|$ yet recovers nothing; (C7) excludes exactly this compression of several true components into one bespoke slot. A counting argument makes (C7) plausible rather than proving it: a subspace of dimension below $|J|$ can contain the coefficient vectors of at most $|J| - 1$ same-pattern samples in general position, against at most $\sum_{r \leq k} \binom{N}{r}$ candidate subsets, so sufficient per-pattern multiplicity rules compression out; (C7) is stated here as an assumption. Condition (C5) also protects the component count, not only the geometry: solutions that represent overlapping true spans by rotated bases, or trade usage through a single combined profile inside a two-dimensional span, place more unit columns inside a span than its dimension and are excluded by (C5) through Lemmas 6 and 7.

Status of Conjecture R. Three proven facts calibrate the conjecture. Its mechanism is unconditional: Lemma 11 shows that out-of-span errors cannot cancel across $\varkappa$ coefficient-diverse samples. Its 1-sparse case is closed: Lemma 12 derives its conclusion from the certificate alone, so part (c) is unconditional in the 1-sparse regime (the spark condition applied to a singleton gives $s_{\min} \leq 1$, so the first condition of (10) at $\gamma_{\mathrm{c}} = c_{\min}$ implies the lemma's hypothesis $\bar{\varepsilon} < c_{\min}$). And in the noiseless limit its role vanishes: at $\bar{\varepsilon} = 0$, exact span membership follows from tight usage alone (Lemma 6), $f \to 0$, and the conclusions of part (c) reduce to those of part (b). What remains open is purely combinatorial: ruling out adaptive fragmentation, the configuration in which, for some active component, the samples of every singleton-intersection pattern are split among candidate sets so that no single candidate set accumulates $|J|$ same-pattern samples with γ_{c}-independent coefficients. A pigeonhole over the at most N^k candidate sets produces recurring sets whenever patterns recur; the missing step is that a recurring set's coefficient matrix inherits genericity even though the solver assigns samples to candidate sets after seeing the data. Alternatively, the conjecture can be bypassed operationally: its conclusion is checkable on a returned solution, and adding it to the certificate trades the open lemma for a runtime check (§4.2).

What this appendix establishes. The three parts return the answer promised at the head of the appendix: success is now recognizable. A returned solution can be certified as the truth with no claim about the search that produced it. With a known pool, the certificate and one noise event yield every support, the exact component count, and noise-floor concentrations (part (a)). With a learned pool, the same interface delivers the components themselves, up to the relabeling the data provably cannot resolve (part (b)), and persists under noise up to one named combinatorial statement whose conclusion is itself checkable on the solution (part (c)). Corollary 1 turns recovery into reuse: a certified pool decodes fresh samples one at a time. Everything downstream stands on this quantification over certificate-passing solutions: the solver section builds an instrument whose objective pursues the certificate, the experiments measure the quantities the certificate names, and Theorem 3 (Appendix D) takes over where finite-sample certification ends, stating what accumulating samples must deliver.

D Formal Statement and Proof of Theorem 3

This appendix proves Theorem 3 in full. The theorem states what accumulating samples deliver, and why the framework's asymptotic target is the generative process rather than the per-sample decodings. Certification (Theorem 2) is a finite-sample notion: it covers the M samples of a fixed dataset. Part (a) shows this finiteness is intrinsic, not a defect of the certificate: at fixed SNR, as M grows, no estimator (the pool known, the latent law supplied, unlimited computation) recovers every sample's support; all-sample correctness has probability decaying at the rate $(1 - q)^M$, and

the noise norm exceeds any fixed tolerance on infinitely many samples. The claim is made because of the paper's object of inference: samples are instantiations of the process, and under noise the instantiations cannot all be solved at any number of samples, so what can be acquired exactly in the limit is the process, not its instantiations. Part (b) states that it is acquired, exactly: estimated by near-minimizing a penalized minimum-distance criterion (the model's data law matched to the empirical measure, with the expected usage entering at vanishing weight as a tie-breaker), the generative process converges almost surely, in the quotient metric that identifies processes up to relabeling, to the true process, and the estimated component count is eventually exactly correct.

Part (a) is proved in three steps: noise recurrence (Step 1), posterior factorization across samples (Step 2), and strict positivity of the per-sample posterior error mass q (Step 3). Part (b) rests on three lemmas, organized in three layers, and runs the compactness scheme of Wald-type consistency proofs (Wald, 1949; Pollard, 1981). Layer I (Lemma 13) constructs the space of processes modulo relabeling: the map sending a process to the law of its anonymous configurations separates exactly the relabeling classes and makes the quotient a compact metric space. Layer II (Lemma 14) shows the data law determines the latent structure: known noise divides out, and a pool decodes its own mixtures uniquely, with a measurable decoder. Layer III (Lemma 15) is the population identification theorem: any process that reproduces the true noiseless law at no greater expected usage is the true process up to relabeling; "population" here and below means: at the level of the exact laws rather than a finite sample. The proof of part (b) then runs in five steps (uniform control, value comparison, limit points, identification, quotient convergence) followed by a four-step count argument that reads the component count off the quotient limit through small balls around the true component profiles.

Notation. The notation of Appendices A and B remains in force. New objects:

φ	the noise law of ε on $\mathbb{R}^d$; in part (a), $\mathcal{N}(0, \sigma^2 I_d)$
Ω_k	$= \bigsqcup_{A\subseteq[N],\,\lvert A\rvert\le k}\{A\}\times[c_{\min}, c_{\max}]^A$, the latent space
$\mathcal{S}$	pools: $S \in (\mathbb{S}^{d-1})^N$ with $\sigma_{\min}(S_B) \ge s_{\min}$ for all $\lvert B\rvert \le 2k$
$G = (S, \mu)$	a generative process: pool $S \in \mathcal{S}$, latent law $\mu \in \mathcal{P}(\Omega_k)$
Π	$= \mathcal{S} \times \mathcal{P}(\Omega_k)$, the class of processes; $G^* = (S^*, \mu^*)$ the truth
Λ_S	mixing map $(A, c) \mapsto \sum_{j\in A} c_j S_j$
ν_G	$= (\Lambda_S)_{\#}\mu$, the noiseless law of the process
G_x	$= \nu_G * \varphi$, the data law
$U(G)$, $\rho_j(G)$	expected usage $\mathbb{E}_\mu\lvert A\rvert$; activation probability $\mu(j \in A)$
$\mathrm{act}(G)$, K^*	active set $\{j : \rho_j(G) > 0\}$; $K^* = \lvert\mathrm{act}(G^*)\rvert$, the true count
$\approx$, $[G]$	relabeling equivalence (below); the class of G in $\Pi/\!\approx$
Ξ, $T(G)$	anonymous configuration $\{(S_j, c_j) : j \in A\}$; its law under μ
β	bounded-Lipschitz metric between laws
d_G	$= \beta\big(T(\cdot), T(\cdot)\big)$, the quotient metric on $\Pi/\!\approx$
$\mathbb{D}_M$, β_M	empirical measure of $x^{(1)}, \dots, x^{(M)}$; $\beta_M = \beta(\mathbb{D}_M, G_x^*)$
F_M, $\hat{G}_M$	criterion $\beta(G_x, \mathbb{D}_M) + \lambda_M U(G)$; an ε_M-minimizer of F_M
$\hat{N}_M$	$= \#\{j : \rho_j(\hat{G}_M) > \rho_{\min}/2\}$, the thresholded component count

Eight conventions and facts. First, the gMCR model (Eq. 2) with the spark condition (Definition 1) is exactly sampling from a process $G^* \in \Pi$: draw $(J, c) \sim \mu^*$, observe $x = \Lambda_{S^*}(J, c) + \varepsilon$; the observed samples are i.i.d. with law G_x^*. Second, Ω_k is a compact metric space whose pieces $\{A\} \times [c_{\min}, c_{\max}]^A$ are clopen; $\mathcal{S}$ is compact ($\sigma_{\min}(S_B)$ is continuous in S, so $\mathcal{S}$ is a closed subset of a product of spheres); $\mathcal{P}(\Omega_k)$ is compact metrizable in the weak topology by Prokhorov's theorem (Parthasarathy, 1967, Theorems II.6.2 and II.6.4); hence Π is compact metrizable. The functionals U and ρ_j are weakly continuous, because $\lvert A\rvert$ and $\mathbf{1}[j \in A]$ are bounded and continuous on Ω_k (locally constant on the clopen pieces). Third, the relabeling equivalence: the symmetric group $\mathrm{Sym}(N)$ acts on latents by $R_\pi(A, c) = (\pi(A),\, c \circ \pi^{-1})$, with pushforward $\pi_{\#}\mu = (R_\pi)_{\#}\mu$, and

$$G \approx G' \;:\Longleftrightarrow\; \exists\, \pi \in \mathrm{Sym}(N) :\;\; \pi_{\#}\mu = \mu' \;\text{ and }\; S'_{\pi(j)} = S_j \;\text{ for every } j \in \mathrm{act}(G).$$

This is an equivalence relation: the identity witnesses reflexivity, and the identity $\rho_{\pi(j)}(G') = \rho_j(G)$ (immediate from $\pi_{\#}\mu = \mu'$) shows π maps $\mathrm{act}(G)$ onto $\mathrm{act}(G')$, so π^{-1} and composites witness symmetry and transitivity. Fourth, the configuration space: on $X := \mathbb{S}^{d-1} \times [c_{\min}, c_{\max}]$ with the metric $d\big((s, c), (s', c')\big) = \max\{\lVert s - s'\rVert,\, \lvert c - c'\rvert\}$, let $\mathcal{K}_k$ be the space of subsets of X of cardinality

at most k: the nonempty ones under the Hausdorff metric, the empty configuration adjoined as an isolated point. $\mathcal{K}_k$ is a compact metric space, and $T(G)$ is the pushforward of μ under the assembly map $(A, c) \mapsto \{(S_j, c_j) : j \in A\}$ (continuous on each clopen piece for fixed S). Fifth, β denotes the bounded-Lipschitz metric, $\beta(\mu, \nu) = \sup\{\, |\int f\, d\mu - \int f\, d\nu| : \|f\|_\infty + \mathrm{Lip}(f) \le 1 \,\}$; on a separable metric space it metrizes weak convergence (Dudley, 2002, §11.3). Sixth, part (a): an estimator is a measurable map $(x^{(1)}, \ldots, x^{(M)}) \mapsto (\hat{A}_1, \ldots, \hat{A}_M)$ into supports (the known pool S may enter as a parameter), the posterior $\Pr(J = A \mid x)$ is taken under the joint law of one draw (J, c, ε) of the model, and

$$q \;:=\; 1 - \mathbb{E}\Big[\max_{A} \Pr\big(J = A \mid x\big)\Big].$$

Seventh, two shorthands of Lemma 15: the *class spark* condition is the spark condition that every pool in $\mathcal{S}$ satisfies by membership in the configuration space; the *true spark* condition is the same condition for the true pool S^*. Eighth, the symbol K^*: Appendix C uses the dataset-level count $|\bigcup_m J_m|$, while here $K^* = |\mathrm{act}(G^*)|$ is the process-level count; the two agree almost surely for all sufficiently large M, since every $j \in \mathrm{act}(G^*)$ has $\rho_j(G^*) > 0$ and therefore eventually appears in some support (second Borel–Cantelli), while components outside $\mathrm{act}(G^*)$ never appear. Each appendix uses the count of its own setting.

Standing assumptions of part (b). Throughout part (b) the observed samples are i.i.d. with law G_x^* for a true process $G^* \in \Pi$; membership in Π is the model (Eq. 2) together with the spark condition (Definition 1) on the true pool. Four further assumptions:

- **(D1) Known noise.** The noise law φ is known and its characteristic function $\widehat{\varphi}$ vanishes nowhere. Gaussian noise qualifies, and the assumption is a boundary rather than a convenience: with fully unknown noise the concentration law is not identifiable, since concentration spread and noise can be exchanged along a component ray.
- **(D2) Richness.** For every $i \in \mathrm{act}(G^*)$ there are supports B_1, B_2 with $\mu^*\big(\{B_l\} \times [c_{\min}, c_{\max}]^{B_l}\big) > 0$ for $l = 1, 2$ and $B_1 \cap B_2 = \{i\}$. Two supports overlapping in exactly one component pin that component profile to a single pool slot (Lemma 15, Step 3).
- **(D3) Smooth concentrations.** For every support A of positive μ^*-mass, the conditional law $\mu^*(\cdot \mid A)$ of the concentrations is absolutely continuous on $[c_{\min}, c_{\max}]^A$. This is what makes every proper-subspace event null in Lemma 15.
- **(D4) Schedule and near-minimization.** The estimators $\hat{G}_M$ are measurable with $F_M(\hat{G}_M) \le \inf_\Pi F_M + \varepsilon_M$, where $\lambda_M \downarrow 0$, $\varepsilon_M \downarrow 0$, and almost surely $(\beta_M + \varepsilon_M)/\lambda_M \to 0$. Near-minimization over the nonconvex Π is an assumption about what the run achieved, not a certified runtime property; the schedule condition is nonvacuous, since $\beta \le W_1$ and the concentration bounds of Fournier & Guillin (2015, Theorem 2) combined with Borel–Cantelli give an almost-sure rate $\beta_M = O(M^{-\gamma})$ for an explicit $\gamma(d) > 0$, so any $\lambda_M = M^{-a}$ with $0 < a < \gamma$ and $\varepsilon_M = o(\lambda_M)$ qualifies.

Theorem 3, restated. The main text states this result under hypotheses given by reference; the full statement follows. *(a)* Fix $\sigma > 0$ and a pool of unit-norm component profiles S. Suppose the samples follow Eq. 2 with $\varepsilon^{(m)} \sim \mathcal{N}(0, \sigma^2 I_d)$ i.i.d., the latent pairs $(J_m, c^{(m)})$ i.i.d. under a law that assigns positive probability to at least two distinct supports, and latents independent of the noise. Then:

(i) for every fixed $\epsilon > 0$, almost surely $\|\varepsilon^{(m)}\| > \epsilon$ for infinitely many m;

(ii) for every M and every measurable map from $(x^{(1)}, \ldots, x^{(M)})$ to supports $(\hat{A}_1, \ldots, \hat{A}_M)$,

$$\Pr\big(\hat{A}_m = J_m \text{ for all } m \le M\big) \;\le\; (1 - q)^M, \qquad q \in (0, 1),$$

where q (Notation) depends only on S, the latent law, and σ, and not on M.

(b) Under the standing assumptions (D1)–(D4), with

$$\rho_{\min} \;:=\; \min_{i \in \mathrm{act}(G^*)} \rho_i(G^*) \;>\; 0 : \qquad d_G\big([\hat{G}_M],\, [G^*]\big) \;\longrightarrow\; 0 \quad \text{almost surely,}$$

and almost surely $\hat{N}_M = K^*$ for all sufficiently large M.

What the three theorems say together. Read together, the theorems tell a user what to expect. Per-sample decodings are trustworthy at the scale of the dataset in hand, where Theorem 2 certifies them. What improves as data accumulate is the process, and above all the component count, the quantity to trust most. Failure has known addresses, not a general shadow. With too little noise margin, the window of Theorem 1 closes and over-selection follows. With too little occurrence diversity, isolating pairs disappear and components merge. With an unknown noise law, the concentration *law* is out of reach, while supports, profiles, and the count are not. Away from these conditions, the inverse problem is not merely solvable: a returned solution can be checked.

Proof of part (a). Step 1 (noise recurrence). Fix $\epsilon > 0$ and set

$$p_0 \;:=\; \Pr\big(\|\varepsilon\| > \epsilon\big) \;>\; 0,$$

which is positive because $\|\varepsilon\|^2/\sigma^2$ is a χ^2_d variable, whose support is all of $[0,\infty)$. The events $\{\|\varepsilon^{(m)}\| > \epsilon\}$ are independent with constant probability p_0, so

$$\sum_{m=1}^{\infty} \Pr\big(\|\varepsilon^{(m)}\| > \epsilon\big) \;=\; \sum_{m=1}^{\infty} p_0 \;=\; \infty,$$

and the second Borel–Cantelli lemma (Durrett, 2019, §2.3) gives $\|\varepsilon^{(m)}\| > \epsilon$ for infinitely many m, almost surely. This proves (i).

Step 2 (posterior factorization). Write $\mathcal{D} = (x^{(1)}, \ldots, x^{(M)})$. The triples $(J_m, c^{(m)}, \varepsilon^{(m)})$ are independent across m and $x^{(m)}$ is a function of the m-th triple alone, so conditionally on $\mathcal{D}$ the supports $J_1, \ldots, J_M$ are independent, with

$$\Pr\big(J_m = A \mid \mathcal{D}\big) \;=\; \Pr\big(J_m = A \mid x^{(m)}\big) \qquad \text{for every } A.$$

Each $\hat{A}_m$ is $\mathcal{D}$-measurable, hence a constant given $\mathcal{D}$, and therefore

$$\begin{aligned}
\Pr\big(\hat{A}_m = J_m \;\forall m \le M\big) &= \mathbb{E}\Big[\Pr\big(J_1 = \hat{A}_1, \ldots, J_M = \hat{A}_M \;\big|\; \mathcal{D}\big)\Big] \\
&= \mathbb{E}\Big[\textstyle\prod_{m=1}^{M} \Pr\big(J_m = \hat{A}_m \;\big|\; x^{(m)}\big)\Big] \\
&\le \mathbb{E}\Big[\textstyle\prod_{m=1}^{M} \max_A \Pr\big(J = A \;\big|\; x^{(m)}\big)\Big] \\
&= \prod_{m=1}^{M} \mathbb{E}\Big[\max_A \Pr\big(J = A \;\big|\; x^{(m)}\big)\Big] \;=\; (1-q)^M,
\end{aligned}$$

by, in order: the tower property; conditional independence of the supports given $\mathcal{D}$ (display above); bounding each factor by the maximum over supports; independence of the $x^{(m)}$; their identical distribution and the definition of q.

Step 3 (q lies in $(0,1)$). The supports range over the finite family $\mathcal{A} = \{A \subseteq [N] : |A| \le k\}$, and the posteriors sum to one, so pointwise

$$\max_A \Pr\big(J = A \mid x\big) \;\ge\; \frac{1}{|\mathcal{A}|}, \qquad \text{hence} \qquad q \;\le\; 1 - \frac{1}{|\mathcal{A}|} \;<\; 1.$$

For $q > 0$: pick distinct supports $A_1 \neq A_2$ with prior masses $p_l := \Pr(J = A_l) > 0$. Conditionally on $J = A_l$, the sample $x = S_{A_l} c + \varepsilon$ has density

$$f_l(x) \;=\; \int \varphi_\sigma\big(x - S_{A_l} c\big)\, d\mathrm{Law}\big(c \mid J = A_l\big) \;>\; 0 \qquad \text{for every } x \in \mathbb{R}^d,$$

where $\mathrm{Law}(c \mid J = A_l)$ denotes the conditional distribution of the concentration vector given the event $J = A_l$ (a probability measure on the coefficient space): the integral averages the Gaussian density over the concentrations the model produces with support A_l, and is strictly positive because φ_σ is strictly positive everywhere. The marginal density is $f = \sum_A \Pr(J = A)\, f_A$, and Bayes' rule gives, for every x,

$$\Pr\big(J = A_l \mid x\big) \;=\; \frac{p_l\, f_l(x)}{f(x)} \;>\; 0, \qquad l = 1, 2.$$

Any maximizer of $A \mapsto \Pr(J = A \mid x)$ differs from at least one of A_1, A_2, and the posteriors are nonnegative and sum to one, so for every x,

$$\max_A \Pr\big(J = A \mid x\big) \;\le\; 1 - \min\Big\{\Pr\big(J = A_1 \mid x\big),\; \Pr\big(J = A_2 \mid x\big)\Big\} \;<\; 1.$$

The variable $1 - \max_A \Pr(J = A \mid x)$ is a measurable function of x (a finite maximum over the posterior kernel's components), bounded in $[0, 1]$, and strictly positive; its expectation is therefore strictly positive (a nonnegative variable with zero expectation vanishes almost surely):

$$q \;=\; \mathbb{E}\Big[1 - \max_A \Pr\big(J = A \mid x\big)\Big] \;>\; 0.$$

Finally q is defined from the law of a single draw (J, c, ε) and the pool, so it does not depend on M. This proves (ii). □

The remaining material proves part (b), in three layers and an assembly: Layer I constructs the space in which the limit lives, Layer II shows the observed law determines the latent structure, Layer III identifies the process from its population law, and the assembly runs the compactness scheme and reads off the count.

Layer I: the space of processes (Lemma 13)

Convergence up to relabeling needs a space in which it is plain convergence. This layer builds it: anonymous configurations separate exactly the relabeling classes, and the resulting quotient is a compact metric space.

Lemma 13 (Quotient metric)**.** *T is continuous and $\approx$-invariant, and $T(G) = T(G')$ if and only if $G \approx G'$. Consequently $d_G\big([G], [G']\big) := \beta\big(T(G), T(G')\big)$ is a metric on $\Pi/\approx$, the space $(\Pi/\approx, d_G)$ is compact, and d_G metrizes the quotient topology.*

Proof. Step 1 (slot separation). Fix $G = (S, \mu) \in \Pi$ and $j \neq j'$. The Gram matrix of the pair $(S_j, S_{j'})$ is $\left(\begin{smallmatrix} 1 & t \\ t & 1 \end{smallmatrix}\right)$ with $t = \langle S_j, S_{j'}\rangle$, with eigenvalues $1 \pm t$; the spark condition at $|B| = 2$ gives $\sigma_{\min}(S_{\{j,j'\}})^2 = 1 - |t| \ge s_{\min}^2$, hence

$$\|S_j - S_{j'}\|^2 \;=\; 2 - 2t \;\ge\; 2\, s_{\min}^2 \;=:\; \Delta^2, \qquad \Delta = \sqrt{2}\, s_{\min} \;>\; 0.$$

Every process in Π therefore has pairwise Δ-separated component profiles. In particular the points of $\Xi = \{(S_j, c_j) : j \in A\}$ have pairwise distinct first coordinates, $|\Xi| = |A|$, and for fixed S the assembly map $(A, c) \mapsto \Xi$ is injective on Ω_k.

Step 2 (invariance). First,

$$\mu\big(A \not\subseteq \mathrm{act}(G)\big) \;\le\; \sum_{j \notin \mathrm{act}(G)} \mu(j \in A) \;=\; \sum_{j \notin \mathrm{act}(G)} \rho_j(G) \;=\; 0,$$

so $A \subseteq \mathrm{act}(G)$ for μ-a.e. latent. Let $G \approx G'$ with witness π. For any latent (A, c) with $A \subseteq \mathrm{act}(G)$, the configuration of the relabeled latent under the pool S' is

$$\big\{\big(S'_{j'},\, (c \circ \pi^{-1})_{j'}\big) : j' \in \pi(A)\big\} \;=\; \big\{\big(S'_{\pi(j)},\, c_j\big) : j \in A\big\} \;=\; \big\{(S_j, c_j) : j \in A\big\},$$

the configuration of (A, c) under S, using $S'_{\pi(j)} = S_j$ on $\mathrm{act}(G)$. Pushing $\mu' = \pi_{\#}\mu$ through the assembly therefore gives $T(G') = T(G)$.

Step 3 (continuity). Let $G_n = (S_n, \mu_n) \to G = (S, \mu)$ in Π and let f on $\mathcal{K}_k$ satisfy $\|f\|_\infty + \mathrm{Lip}(f) \le 1$. Then

$$\Big|\int f\, dT(G_n) - \int f\, dT(G)\Big| \;\le\; \sup_{(A,c)\in\Omega_k} \big|f\big(\Xi_{S_n}(A, c)\big) - f\big(\Xi_S(A, c)\big)\big| + \Big|\int (f \circ \Xi_S)\, d\mu_n - \int (f \circ \Xi_S)\, d\mu\Big|,$$

where Ξ_S denotes the assembly map under S. The first term is controlled by the Hausdorff distance: matching the points of the two configurations slot by slot,

$$d_H\big(\Xi_{S_n}(A,c),\,\Xi_S(A,c)\big)\;\le\;\max_{j\in A}\|S_{n,j}-S_j\|\;\longrightarrow\;0\qquad\text{uniformly over }\Omega_k,$$

so the supremum tends to 0 (f is Lipschitz). The second term tends to 0 by weak convergence $\mu_n\rightharpoonup\mu$, because $f\circ\Xi_S$ is bounded and continuous on Ω_k (continuous on each clopen piece). Taking the supremum over such f: $\beta\big(T(G_n),T(G)\big)\to 0$.

Step 4 (T determines the class). Suppose $T(G)=T(G')$ with $G=(S,\mu)$, $G'=(S',\mu')$. For a unit vector s, put $K_s:=\{s\}\times[c_{\min},c_{\max}]$, a compact subset of X. By Step 1 at most one slot carries the profile s, so

$$T(G)\big(\{\Xi:\Xi\cap K_s\neq\emptyset\}\big)\;=\;\mu\big(\exists\, j\in A: S_j=s\big),$$

which equals $\rho_j(G)$ when $s=S_j$ for the (unique) slot j, and 0 when no slot carries s. The set $\{S_j:j\in\mathrm{act}(G)\}$ is therefore recoverable from $T(G)$ as the profiles carried with positive probability, together with their activation masses. Equality of the two laws gives $\{S_j:j\in\mathrm{act}(G)\}=\{S'_{j'}:j'\in\mathrm{act}(G')\}$ as subsets of the sphere, so there is a bijection $\pi:\mathrm{act}(G)\to\mathrm{act}(G')$ with $S'_{\pi(j)}=S_j$, unique by pairwise distinctness on each side; extend π to a permutation of $[N]$ arbitrarily.

Next, the match-back map. For $B\subseteq\mathrm{act}(G)$ define

$$D_B\;:=\;\Big\{\Xi\in\mathcal{K}_k:\;\Xi\cap K_{S_j}\neq\emptyset\;\forall j\in B,\;\;\Xi\subseteq\textstyle\bigcup_{j\in B}K_{S_j},\;\;|\Xi|=|B|\Big\}.$$

Each defining condition is Borel in the Hausdorff topology on $\mathcal{K}_k$ (Kechris, 1995, cf. §4.F): $\{\Xi:\Xi\cap K\neq\emptyset\}$ is closed for compact K, $\{\Xi:\Xi\subseteq F\}$ is closed for closed F (its complement is $\{\Xi:\Xi\cap F^c\neq\emptyset\}$, open since F^c is open), and $\{\Xi:|\Xi|\le n\}$ is closed. The sets D_B are pairwise disjoint, and a configuration in D_B carries exactly one point over each S_j, $j\in B$. On $\{\Xi:\Xi\cap K_{S_j}\neq\emptyset\}$ define

$$g_j(\Xi)\;:=\;\min\big\{t\in[c_{\min},c_{\max}]:(S_j,t)\in\Xi\big\};$$

g_j is Borel, because $\{\Xi:g_j(\Xi)\le a\}=\{\Xi:\Xi\cap(\{S_j\}\times[c_{\min},a])\neq\emptyset\}$ is closed. The match-back map

$$m(\Xi)\;:=\;\big(B,\;(g_j(\Xi))_{j\in B}\big)\qquad\text{on }D_B,\quad B\subseteq\mathrm{act}(G),$$

is Borel on $\bigcup_B D_B$, and on the μ-full event $\{A\subseteq\mathrm{act}(G)\}$ it inverts the assembly: $m\big(\Xi_S(A,c)\big)=(A,c)$, since the assembled configuration lies in D_A and g_j reads off c_j. Since $T(G)\big(\bigcup_B D_B\big)=\mu\big(A\subseteq\mathrm{act}(G)\big)=1$ (Step 2),

$$\mu\;=\;m_{\#}\,T(G).$$

Build m' from S' over $\mathrm{act}(G')$ the same way. A configuration $\Xi\in D_B$ has first coordinates $\{S_j:j\in B\}=\{S'_{\pi(j)}:j\in B\}$, so it lies in the S'-set $D'_{\pi(B)}$, and $g'_{\pi(j)}(\Xi)=g_j(\Xi)$ (the same slice of Ξ); hence $m'=R_\pi\circ m$ on $\bigcup_B D_B$, and

$$\mu'\;=\;m'_{\#}\,T(G')\;=\;m'_{\#}\,T(G)\;=\;(R_\pi)_{\#}\,m_{\#}\,T(G)\;=\;\pi_{\#}\,\mu.$$

Together with $S'_{\pi(j)}=S_j$ on $\mathrm{act}(G)$ this is precisely $G\approx G'$. With Step 2, $T(G)=T(G')$ if and only if $G\approx G'$.

Step 5 (metric). d_G is well defined on classes by Step 2, symmetric with the triangle inequality inherited from β, and $d_G([G],[G'])=0$ if and only if $T(G)=T(G')$ (β is a metric) if and only if $G\approx G'$ (Step 4), that is, $[G]=[G']$.

Step 6 (compactness and the quotient topology). The map $[G]\mapsto T(G)$ is a well-defined isometry from $(\Pi/\!\approx,d_G)$ onto $(T(\Pi),\beta)$; $T(\Pi)$ is compact as the continuous image (Step 3) of the compact Π, so $(\Pi/\!\approx,d_G)$ is compact. The quotient topology is compact as the image of Π under the (continuous) quotient map, and the identity from $\Pi/\!\approx$ with the quotient topology to $(\Pi/\!\approx,d_G)$ is continuous (T is continuous and constant on classes, so d_G-open sets pull back to open sets); a continuous bijection from a compact space to a Hausdorff space is a homeomorphism, so d_G metrizes the quotient topology. □

LAYER II: FROM THE DATA LAW TO THE LATENT LAW (LEMMA 14)

This layer shows the observation determines the structure: the known noise divides out of the data law, and a pool reads its own mixtures uniquely, with a measurable decoder; the pool and the noiseless law together determine the latent law.

Lemma 14 (From the data law to the latent law)**.** *Let $S \in \mathcal{S}$. Then:* (i) *under (D1), if ν, ν' are finite Borel measures on $\mathbb{R}^d$ with $\nu * \varphi = \nu' * \varphi$, then $\nu = \nu'$; in particular $G_x = G'_x$ implies $\nu_G = \nu_{G'}$;* (ii) *Λ_S is injective on Ω_k;* (iii) *Λ_S is a homeomorphism onto its range $R_S := \Lambda_S(\Omega_k)$, and $\Lambda_S^{-1} : R_S \to \Omega_k$ is Borel, as are the images $\Lambda_S(B)$ of Borel sets $B \subseteq \Omega_k$;* (iv) *for every $G = (S, \mu) \in \Pi$,*

$$\mu \;=\; \big(\Lambda_S^{-1}\big)_{\#}\, \nu_G :$$

the pool and the noiseless law determine the latent law.

Proof. Step 1 (Fourier factorization). For every $t \in \mathbb{R}^d$, the characteristic function of a convolution is the product:

$$\widehat{\nu * \varphi}\,(t) \;=\; \widehat{\nu}(t)\,\widehat{\varphi}(t), \qquad \widehat{\nu' * \varphi}\,(t) \;=\; \widehat{\nu'}(t)\,\widehat{\varphi}(t).$$

Step 2 (division). The hypothesis of part (i) equates the left sides, and $\widehat{\varphi}(t) \neq 0$ for every t by (D1), so

$$\widehat{\nu}(t) \;=\; \widehat{\nu'}(t) \qquad \text{for every } t \in \mathbb{R}^d.$$

Step 3 (uniqueness; part (i)). Finite Borel measures on $\mathbb{R}^d$ with equal characteristic functions coincide (Durrett, 2019, Theorem 3.10.4). Hence $\nu = \nu'$; applying this to $\nu_G, \nu_{G'}$ gives the second claim of part (i).

Step 4 (injectivity; part (ii)). Suppose $\Lambda_S(A, c) = \Lambda_S(A', c')$. Define v on $A \cup A'$ (with $|A \cup A'| \leq 2k$) coordinate-wise, exactly as in Step 1 of Appendix A:

$$v_j = \begin{cases} c_j, & j \in A \setminus A', \\ c_j - c'_j, & j \in A \cap A', \\ -c'_j, & j \in A' \setminus A. \end{cases}$$

Then $S_{A \cup A'}\, v = \Lambda_S(A, c) - \Lambda_S(A', c') = 0$, and the spark condition gives

$$0 \;=\; \|S_{A \cup A'}\, v\| \;\geq\; s_{\min}\, \|v\|, \qquad \text{hence} \qquad v = 0.$$

On $A \setminus A'$ and $A' \setminus A$ the entries of v have magnitude at least $c_{\min} > 0$, so both differences are empty: $A = A'$; and then $c - c' = v = 0$.

Step 5 (Borel inverse; part (iii)). Λ_S is continuous on Ω_k (linear in c on each clopen piece), Ω_k is compact, and $\mathbb{R}^d$ is Hausdorff, so Λ_S is a closed map and, being injective by Step 4, a homeomorphism onto R_S; its inverse is continuous on R_S, hence Borel. (In general, an injective Borel map between standard Borel spaces has Borel image and Borel inverse by the Lusin–Suslin theorem, Kechris, 1995, Theorem 15.1; compactness makes the elementary route sufficient here.) For Borel $B \subseteq \Omega_k$, the image $\Lambda_S(B) = (\Lambda_S^{-1})^{-1}(B)$ is Borel in the compact R_S, hence Borel in $\mathbb{R}^d$.

Step 6 (measure determination; part (iv)). For every Borel $B \subseteq \Omega_k$, using $\nu_G = (\Lambda_S)_{\#}\mu$ and injectivity,

$$\big(\Lambda_S^{-1}\big)_{\#}\nu_G\,(B) \;=\; \nu_G\big(\Lambda_S(B)\big) \;=\; \mu\big(\Lambda_S^{-1}\big(\Lambda_S(B)\big)\big) \;=\; \mu(B).$$

□

LAYER III: POPULATION IDENTIFICATION (LEMMA 15)

This layer is the identification engine: a usage-minimal reproduction of the true noiseless law is the true process up to relabeling. At the population level the assumptions lighten rather than accumulate: the coefficient-genericity and no-compression conditions of the finite-sample theory become theorems here, because absolutely continuous concentration laws (D3) make every proper-subspace event null. Wherever linear independence of profiles is used below, it is supplied by the class or true spark condition on index sets of at most $2k$ elements.

Lemma 15 (Population identification)**.** *Let $G^* \in \Pi$ obey (D2)–(D3), and let $G = (S, \mu) \in \Pi$ satisfy*

$$\nu_G \;=\; \nu_{G^*} \qquad \textit{and} \qquad U(G) \;\le\; U(G^*).$$

Then $G \approx G^*$.

Proof. Write $\nu := \nu_{G^*} = \nu_G$ and $V_B := \operatorname{span}(s_i^* : i \in B)$ for $B \subseteq [N]$.

Step 0 (decoding both pools; the degenerate case). By Lemma 14 applied to S and to S^*, for every x in the ν-full sets R_S and R_{S^*} there are unique decodings

$$x \;=\; \Lambda_S\big(\alpha(x),\, c(x)\big) \;=\; \Lambda_{S^*}\big(\alpha^*(x),\, c^*(x)\big),$$

with Borel maps α, c, α^*, c^*; the change of variables $\nu = (\Lambda_S)_{\#}\mu$ gives

$$\mathbb{E}_\nu|\alpha| \;=\; \mathbb{E}_\mu|A| \;=\; U(G), \qquad \mathbb{E}_\nu|\alpha^*| \;=\; U(G^*),$$

and, for any Borel set $\Gamma \subseteq \Omega_k$, $\nu\big((\alpha, c) \in \Gamma\big) = \mu(\Gamma)$, likewise for the starred pair with μ^*. If $\operatorname{act}(G^*) = \emptyset$ then $U(G^*) = 0$, so $U(G) = \mathbb{E}_\mu|A| = 0$, both μ and μ^* put full mass on the empty support, and the identity permutation witnesses $G \approx G^*$. Assume henceforth $\operatorname{act}(G^*) \neq \emptyset$; by (D2) some support of positive mass is nonempty, so $k \ge 1$.

Step 1 (full-rank service is automatic). Fix a pair of supports (A, B) and suppose $W_{A,B} := \operatorname{span}(S_A) \cap V_B \subsetneq V_B$. On the event $\{\alpha = A,\ \alpha^* = B\}$ the decodings place

$$x \;\in\; \operatorname{span}\big(S_{\alpha(x)}\big) \cap V_{\alpha^*(x)} \;=\; W_{A,B}.$$

The event $\{\alpha^* = B\}$ is, up to ν-null sets, the image of the latent event $\{A' = B\}$, so conditioning on it, $x = S_B^*\, c$ with $c \sim \mu^*(\cdot \mid B)$ absolutely continuous by (D3); the preimage $(S_B^*)^{-1}(W_{A,B})$ is a proper linear subspace of $\mathbb{R}^B$ (the map S_B^* is injective by the true spark condition, and $W_{A,B} \subsetneq V_B$), hence Lebesgue-null, so

$$\nu\big(\alpha = A,\ \alpha^* = B\big) \;\le\; \mu^*\big(A' = B\big) \cdot \mu^*\big(c \in (S_B^*)^{-1}(W_{A,B}) \;\big|\; B\big) \;=\; 0.$$

There are finitely many pairs (A, B), so for ν-a.e. x,

$$V_{\alpha^*(x)} \;\subseteq\; \operatorname{span}\big(S_{\alpha(x)}\big).$$

Step 2 (tight service, a.e.). For ν-a.e. x, by Step 1 and $\dim V_B = |B|$ (true spark),

$$|\alpha(x)| \;\ge\; \dim \operatorname{span}\big(S_{\alpha(x)}\big) \;\ge\; \dim V_{\alpha^*(x)} \;=\; |\alpha^*(x)|.$$

Integrating against ν and using the hypothesis,

$$U(G^*) \;\ge\; U(G) \;=\; \mathbb{E}_\nu|\alpha| \;\ge\; \mathbb{E}_\nu|\alpha^*| \;=\; U(G^*),$$

so $\mathbb{E}_\nu\big[|\alpha| - |\alpha^*|\big] = 0$ with a ν-a.e. nonnegative integrand, forcing $|\alpha(x)| = |\alpha^*(x)|$ for ν-a.e. x. The columns $S_{\alpha(x)}$ are linearly independent ($|\alpha(x)| \le k \le 2k$, class spark), so

$$\dim \operatorname{span}\big(S_{\alpha(x)}\big) \;=\; |\alpha(x)| \;=\; |\alpha^*(x)| \;=\; \dim V_{\alpha^*(x)},$$

and with the containment of Step 1 (equal finite dimensions),

$$\operatorname{span}\big(S_{\alpha(x)}\big) \;=\; V_{\alpha^*(x)} \qquad \text{for } \nu\text{-a.e. } x.$$

Step 3 (anchoring). Fix $i \in \operatorname{act}(G^*)$ and the (D2) supports B_1, B_2 with $B_1 \cap B_2 = \{i\}$. The events $\{\alpha^* = B_l\}$ have positive ν-mass, so each contains a point of the a.e.-good set of Steps 1–2: pick such x_1, x_2 and set $A_1 := \alpha(x_1)$, $A_2 := \alpha(x_2)$; by Step 2,

$$\operatorname{span}\big(S_{A_l}\big) = V_{B_l}, \qquad |A_l| = |B_l|, \qquad l = 1, 2.$$

The true family on $B_1 \cup B_2$ is linearly independent ($|B_1 \cup B_2| \le 2k$, true spark), and spans over an independent family intersect exactly along shared indices: a member of $V_{B_1} \cap V_{B_2}$ has a representation over each of B_1 and B_2, the difference of the two representations is a vanishing combination over $B_1 \cup B_2$, and independence forces every coefficient off $B_1 \cap B_2 = \{i\}$ to vanish. Hence

$$V_{B_1} \cap V_{B_2} \;=\; \operatorname{span}\big(s_i^*\big), \qquad \dim\big(V_{B_1} \cap V_{B_2}\big) = 1.$$

Set $U := A_1 \cup A_2$, of cardinality $|U| \le |A_1| + |A_2| = |B_1| + |B_2| \le 2k$, so the columns S_U are independent (class spark) and, by the dimension formula for a sum of subspaces,

$$|U| \;=\; \dim \operatorname{span}(S_U) \;=\; \dim\big(V_{B_1} + V_{B_2}\big) \;=\; |B_1| + |B_2| - 1.$$

Inclusion–exclusion then yields exactly one shared index:

$$|A_1 \cap A_2| \;=\; |A_1| + |A_2| - |U| \;=\; 1, \qquad A_1 \cap A_2 =: \{j^*\}.$$

The shared column lies in both spans, so

$$S_{j^*} \;\in\; \operatorname{span}\big(S_{A_1}\big) \cap \operatorname{span}\big(S_{A_2}\big) \;=\; V_{B_1} \cap V_{B_2} \;=\; \operatorname{span}\big(s_i^*\big),$$

hence $S_{j^*} = a\, s_i^*$ for a scalar a, and unit norms give $|a| = 1$: write $S_{j^*} = \sigma_i\, s_i^*$ with $\sigma_i \in \{\pm 1\}$. The slot is unique: a second pool column in the line $\operatorname{span}(s_i^*)$ would form with S_{j^*} a linearly dependent pair, contradicting the class spark at $|B| = 2$; so j^* does not depend on the chosen pair of generic points, and we set $\pi(i) := j^*$. The map π is injective: $\pi(i) = \pi(i')$ with $i \ne i'$ would give $\sigma_i s_i^* = \sigma_{i'} s_{i'}^*$, writing one member of the independent true pair $\{s_i^*, s_{i'}^*\}$ as a multiple of the other, which is impossible.

Step 4 (signs, supports, and coefficients, a.e.). Fix a point x of the a.e.-good set and $i \in \alpha^*(x)$, and suppose $\pi(i) \notin \alpha(x)$. The pool subset $\{\pi(i)\} \cup \alpha(x)$ has cardinality $|\alpha(x)| + 1 \le k + 1 \le 2k$, hence its columns are independent (class spark); yet all of them lie in $V_{\alpha^*(x)}$: the columns of $\alpha(x)$ by Step 2, and $S_{\pi(i)} = \sigma_i s_i^* \in V_{\alpha^*(x)}$ since $i \in \alpha^*(x)$. An independent family in $V_{\alpha^*(x)}$ has at most $\dim V_{\alpha^*(x)} = |\alpha^*(x)| = |\alpha(x)|$ members, so

$$|\alpha(x)| + 1 \;\le\; |\alpha(x)|,$$

absurd. Hence $\pi(\alpha^*(x)) \subseteq \alpha(x)$, and $|\pi(\alpha^*(x))| = |\alpha^*(x)| = |\alpha(x)|$ (π injective; Step 2) upgrades the containment to

$$\alpha(x) \;=\; \pi\big(\alpha^*(x)\big).$$

Substituting $S_{\pi(i)} = \sigma_i s_i^*$ into the two decodings of the same point,

$$\sum_{i \in \alpha^*(x)} \big(c(x)_{\pi(i)}\, \sigma_i\big)\, s_i^* \;=\; x \;=\; \sum_{i \in \alpha^*(x)} c^*(x)_i\, s_i^*,$$

and coefficient extraction along the independent true family on $\alpha^*(x)$ gives

$$c(x)_{\pi(i)}\, \sigma_i \;=\; c^*(x)_i \qquad \text{for every } i \in \alpha^*(x).$$

Both $c(x)_{\pi(i)} \ge c_{\min} > 0$ and $c^*(x)_i \ge c_{\min} > 0$ (coefficients live in the box), so $\sigma_i = +1$ whenever some good point has $i \in \alpha^*(x)$; the (D2) event $\{\alpha^* = B_1\}$ with $B_1 \ni i$ has positive mass, so this holds for every $i \in \operatorname{act}(G^*)$:

$$S_{\pi(i)} \;=\; s_i^* \qquad \text{for every } i \in \operatorname{act}(G^*),$$

and $c(x)_{\pi(i)} = c^*(x)_i$ for ν-a.e. x.

Step 5 (assembly). Extend π to a permutation of $[N]$ (arbitrarily on inactive indices; immaterial, because $\mu^*(A \subseteq \operatorname{act}(G^*)) = 1$ by the computation of Lemma 13, Step 2). On the μ^*-full event $\{A \subseteq \operatorname{act}(G^*)\}$, Step 4's profile identity gives

$$\Lambda_S\big(R_\pi(A, c)\big) \;=\; \sum_{j \in \pi(A)} (c \circ \pi^{-1})_j\, S_j \;=\; \sum_{i \in A} c_i\, S_{\pi(i)} \;=\; \sum_{i \in A} c_i\, s_i^* \;=\; \Lambda_{S^*}(A, c),$$

so, by injectivity of Λ_S (Lemma 14(ii)), $\Lambda_S^{-1} \circ \Lambda_{S^*} = R_\pi$ holds μ^*-almost everywhere. Chaining pushforwards (Lemma 14(iv)),

$$\mu \;=\; \big(\Lambda_S^{-1}\big)_{\#}\nu \;=\; \big(\Lambda_S^{-1}\big)_{\#}\big(\Lambda_{S^*}\big)_{\#}\mu^* \;=\; \big(\Lambda_S^{-1} \circ \Lambda_{S^*}\big)_{\#}\,\mu^* \;=\; (R_\pi)_{\#}\,\mu^* \;=\; \pi_{\#}\,\mu^*.$$

Together with $S_{\pi(i)} = s_i^*$ on $\operatorname{act}(G^*)$, the permutation π witnesses $G^* \approx G$; the relation is symmetric (Notation), so $G \approx G^*$. □

Proof of part (b). All statements are pathwise. By Varadarajan's theorem (Varadarajan, 1958), almost surely $\mathbb{D}_M \rightharpoonup G_x^*$, and β metrizes weak convergence, so $\beta_M \to 0$ almost surely; intersect this event with the almost-sure (D4) event $(\beta_M + \varepsilon_M)/\lambda_M \to 0$, fix one point of the intersection, and argue deterministically. Steps 1–5 prove the quotient convergence; Steps 6–9 prove the count claim.

Step 1 (uniform control). For every $G \in \Pi$, by the triangle inequality for β,

$$\big|\beta(G_x, \mathbb{D}_M) - \beta(G_x, G_x^*)\big| \;\le\; \beta(\mathbb{D}_M, G_x^*) \;=\; \beta_M :$$

the minimum-distance design makes the uniform law of large numbers a single line.

Step 2 (value comparison). Near-minimization against the candidate G^*, then Step 1 at G^* ($\beta(G_x^*, \mathbb{D}_M) \le \beta_M$) and $U(G^*) = \mathbb{E}_{\mu^*}|A| \le k$:

$$F_M(\hat{G}_M) \;\le\; F_M(G^*) + \varepsilon_M \;=\; \beta(G_x^*, \mathbb{D}_M) + \lambda_M\, U(G^*) + \varepsilon_M \;\le\; \beta_M + \lambda_M k + \varepsilon_M.$$

Both terms of F_M are nonnegative, so each is bounded by the whole. The fit term, moved to the true law by Step 1,

$$\rho(\hat{G}_M) \;:=\; \beta\big((\hat{G}_M)_x,\, G_x^*\big) \;\le\; \beta\big((\hat{G}_M)_x,\, \mathbb{D}_M\big) + \beta_M \;\le\; 2\beta_M + \lambda_M k + \varepsilon_M \;\longrightarrow\; 0,$$

and the usage term, divided by $\lambda_M > 0$,

$$U(\hat{G}_M) \;\le\; U(G^*) + \frac{\beta_M + \varepsilon_M}{\lambda_M} \;\longrightarrow\; U(G^*)$$

by the (D4) schedule condition.

Step 3 (limit points). The functional $G \mapsto \rho(G) = \beta(G_x, G_x^*)$ is continuous on Π. Indeed, for $\|f\|_\infty + \mathrm{Lip}(f) \le 1$, the smoothed function $g(y) := \int f(y+e)\, d\varphi(e)$ satisfies $\|g\|_\infty + \mathrm{Lip}(g) \le 1$, and

$$\Big|\int f\, dG_x - \int f\, dG'_x\Big| \;=\; \Big|\int g\, d\nu_G - \int g\, d\nu_{G'}\Big|, \qquad \text{hence} \quad \beta(G_x, G'_x) \le \beta(\nu_G, \nu_{G'});$$

and $G \mapsto \nu_G$ is weakly continuous by the computation of Lemma 13, Step 3, with Λ_S in place of the assembly map and the uniform bound

$$\big\|\Lambda_{S_n}(A,c) - \Lambda_S(A,c)\big\| \;\le\; \sum_{j \in A} c_j\, \|S_{n,j} - S_j\| \;\le\; k\, c_{\max} \max_j \|S_{n,j} - S_j\| \;\longrightarrow\; 0.$$

U is continuous (Notation). Π is compact, so every subsequence of $(\hat{G}_M)$ has a convergent sub-subsequence; for any subsequential limit G_∞, Step 2 and continuity give

$$\rho(G_\infty) \;=\; \lim \rho(\hat{G}_{M_j}) \;=\; 0, \qquad U(G_\infty) \;=\; \lim U(\hat{G}_{M_j}) \;\le\; U(G^*),$$

that is, $(G_\infty)_x = G_x^*$ and $U(G_\infty) \le U(G^*)$.

Step 4 (identification). Lemma 14(i) upgrades $(G_\infty)_x = G_x^*$ to $\nu_{G_\infty} = \nu_{G^*}$, and Lemma 15 then gives $G_\infty \approx G^*$: every subsequential limit of $(\hat{G}_M)$ lies in the single class $[G^*]$.

Step 5 (quotient convergence). Suppose $d_G([\hat{G}_M], [G^*]) \not\to 0$: some $\eta > 0$ and a subsequence keep $d_G \ge \eta$. Extract a convergent sub-subsequence $\hat{G}_{M_j} \to G_\infty$ (compactness); by Steps 3–4, $T(G_\infty) = T(G^*)$, and by continuity of T (Lemma 13),

$$d_G\big([\hat{G}_{M_j}], [G^*]\big) \;=\; \beta\big(T(\hat{G}_{M_j}),\, T(G^*)\big) \;\longrightarrow\; \beta\big(T(G_\infty),\, T(G^*)\big) \;=\; 0,$$

contradicting $d_G \ge \eta$. Hence $d_G([\hat{G}_M], [G^*]) \to 0$ on the fixed almost-sure event, which is the first claim.

Step 6 (slot separation and the radius). By Lemma 13, Step 1, every process in Π has pairwise Δ-separated component profiles, $\Delta = \sqrt{2}\, s_{\min}$; in particular the active true profiles s_i^*, $i \in \mathrm{act}(G^*)$, are pairwise Δ-separated ($S^* \in \mathcal{S}$). Fix a radius r with

$$0 \;<\; r \;<\; \frac{\Delta}{2}, \qquad \text{so that} \qquad 2r \;<\; \Delta \;\le\; \min_{\substack{i,i' \in \mathrm{act}(G^*) \\ i \ne i'}} \big\|s_i^* - s_{i'}^*\big\|.$$

Step 7 (balls are continuity sets). For each $i \in \mathrm{act}(G^*)$ put $K_i := \bar{B}(s_i^*, r) \times [c_{\min}, c_{\max}]$ and

$$E_i \;:=\; \big\{\Xi \in \mathcal{K}_k : \Xi \cap K_i \neq \emptyset\big\}.$$

E_i is closed, and its interior contains $\{\Xi : \Xi \cap (B(s_i^*, r) \times [c_{\min}, c_{\max}]) \neq \emptyset\}$ (an open set: the open ball times the full coefficient range is open in X), so every boundary configuration meets the closed ball but not the open one: it carries a point whose profile is at distance exactly r from s_i^*. Under $T(G^*)$ every carried profile is some $s_{i'}^*$, at distance 0 or at least $2r > r$ from s_i^*, so

$$T(G^*)\big(\partial E_i\big) \;=\; 0, \qquad T(G^*)(E_i) \;=\; \mu^*(i \in A) \;=\; \rho_i(G^*).$$

By Step 5, $T(\hat{G}_M) \rightharpoonup T(G^*)$ (β metrizes weak convergence), so the continuity-set form of the portmanteau theorem (Durrett, 2019, Theorem 3.2.11) gives

$$T(\hat{G}_M)(E_i) \;\longrightarrow\; \rho_i(G^*) \;\geq\; \rho_{\min} \qquad \text{for every } i \in \mathrm{act}(G^*).$$

Step 8 (matched slots). The ball $\bar{B}(s_i^*, r)$ has diameter $2r < \Delta$, so at most one slot of $\hat{G}_M$ has its profile inside it (Step 6); when such a slot j exists,

$$T(\hat{G}_M)(E_i) \;=\; \hat{\mu}_M\big(\exists\, j' \in A : S_{j'}(\hat{G}_M) \in \bar{B}(s_i^*, r)\big) \;=\; \hat{\mu}_M(j \in A) \;=\; \rho_j(\hat{G}_M),$$

and when no such slot exists, $T(\hat{G}_M)(E_i) = 0$. Since $T(\hat{G}_M)(E_i) \to \rho_i(G^*) \geq \rho_{\min} > 0$, eventually a (unique) matched slot $j_i(M)$ exists for every active i, with

$$\rho_{j_i(M)}(\hat{G}_M) \;=\; T(\hat{G}_M)(E_i) \;>\; \frac{\rho_{\min}}{2} \qquad \text{eventually.}$$

The K^* balls are pairwise disjoint ($2r$ below the minimal distance between active profiles), so the matched slots are distinct across i: eventually at least K^* slots clear the threshold.

Step 9 (unmatched slots and the verdict). Let

$$C \;:=\; \Big\{(s, c) \in X : \min_{i \in \mathrm{act}(G^*)} \|s - s_i^*\| \geq r\Big\}, \qquad E_{\mathrm{out}} \;:=\; \big\{\Xi : \Xi \cap C \neq \emptyset\big\},$$

a closed set of configurations. Under $T(G^*)$ every carried profile is an active s_i^*, at distance 0 from itself, so $T(G^*)(E_{\mathrm{out}}) = 0$, and the closed-set half of the portmanteau theorem (Durrett, 2019, Theorem 3.2.11) gives

$$\limsup_M \; T(\hat{G}_M)\big(E_{\mathrm{out}}\big) \;\leq\; T(G^*)\big(E_{\mathrm{out}}\big) \;=\; 0.$$

Call a slot j of $\hat{G}_M$ unmatched if its profile is at distance greater than r from every active s_i^*; then the event $\{j \in A\}$ forces Ξ to carry the point $(S_j(\hat{G}_M),\, c_j) \in C$, so

$$\rho_j(\hat{G}_M) \;\leq\; T(\hat{G}_M)\big(E_{\mathrm{out}}\big) \;\longrightarrow\; 0,$$

uniformly over unmatched slots: eventually every unmatched slot has activation below $\rho_{\min}/2$. Every slot is matched to exactly one active i or unmatched (the balls are disjoint), so eventually the slots with $\rho_j(\hat{G}_M) > \rho_{\min}/2$ are exactly the K^* matched ones:

$$\hat{N}_M \;=\; K^* \qquad \text{for all sufficiently large } M,$$

on the fixed almost-sure event. This is the second claim. □

Two boundaries of part (b), stated exactly. First, the criterion: identification rests on the distance term alone; the usage term enters with vanishing weight as a tie-breaker among processes reproducing the data law, so no claim is made that the true process optimizes a fit functional at any fixed M, and (D4) near-minimization is an assumption about what the optimization run achieved, not a property certified at runtime. Second, the count: the threshold $\rho_{\min}/2$ is defined from the unknown G^*, so $\hat{N}_M$ is an oracle-threshold count, stated this way to isolate the identification content; a data-driven threshold $\tau_M \to 0$ decaying sufficiently slowly yields the same conclusion, a separate statement not pursued here. Without any threshold the raw count of slots with positive activation is only lower-semicontinuous along d_G-convergence, and over-counting by slots of vanishing activation cannot be excluded: this matches the empirically observed over-selection at low SNR (Fig. 4).

What this appendix establishes. The two parts return the answer promised at the head of the appendix: the object of inference is the right one. Part (a) shows that per-sample decoding is a finite-sample notion by necessity: under noise, no estimator decodes every instantiation, at any data size. Part (b) shows the process itself is delivered without residue: the estimated process converges almost surely to the truth up to relabeling, and the estimated component count is eventually exactly the true count. Together the theorems close the framework's account of what can be known from mixtures: existence (§3), findability (Theorem 1), recognizability at finite samples (Theorem 2), and the exact asymptotic deliverable (Theorem 3). The component count the experiments report is the quantity part (b) proves eventually exact, and reuse of a recovered process (Corollary 1) is meaningful precisely because the process is the object that accumulating data pin down.

E Formal Statement and Proof of Proposition 2 (Two-Phase Convergence)

This appendix states and proves Proposition 2 in full. Its subject is the one object about which Theorems 1–3 are deliberately silent: the running solver itself. During training, the solver revises which components it selects and the concentrations it assigns to them; early in a run the selected component set changes often, and at some point it stops changing. The practical question is whether a stable selection can be trusted. Proposition 2 identifies, from quantities the solver already computes, a condition under which it can: once the condition holds, the selected components provably remain fixed for a computable number of steps while the reconstruction error decreases toward an explicit floor, and a checkpoint taken inside this period records a stable selection rather than a transient one.

In the vocabulary of the model, the condition is an *anchor event*: at some step t_0, every component's empirical energy margin (the dataset mean of the gate's per-sample energy gap, signed so that positive means agreement with the selection made at t_0) clears a level γ. Two phases follow. In the first, the anchor opens a *support-stability window* of explicit length $T(\gamma)$: for that many steps the selected support cannot change, for any realization of the data and of the training noise, because Lipschitz continuity of the energies and the bounded optimizer step bound how fast margins of size γ can decrease. In the second, the objective with the gate frozen at the anchored support contracts linearly, in expectation, to an explicit noise floor. Once the objective gap falls below an explicit basin level $\varepsilon^\star$, the small gap forces every margin positive again, so the support becomes self-sustaining. The gate re-selects the anchored support with quantified failure probability, at each fixed time and uniformly over any finite horizon.

Theorems 1–3 state what a returned solution means; this proposition explains why the solver is a dependable way to reach one. What a practitioner observes in a run, selections stabilizing, the fit descending, the stabilized selection persisting, is proved here to be the mechanism's expected behavior: after an observed anchor event, the selected support holds for an explicit horizon, the objective contracts toward an explicit floor, and the support re-selects itself with quantified risk. The guarantee is conditional and finite-horizon by design: it activates at the anchor (whether a run reaches one is read off the run itself), it rests on stated assumptions about the local geometry of the objective, and whether the stabilized support is the true one is decided by Theorem 2's certificate, with the variational content supplied by Theorem 1(b) through Lemma 16. The proposition is the reason the checkpointing rule of §5.3 samples a meaningful object: after an anchor, checkpoints within the window all record the same support, and the risk of a later support flip is quantified per unit horizon. Even where the conditions for exact recovery fail, the run reaches a stable, checkable object rather than a transient fit.

The statement's two-phase shape follows the finite-identification literature, where proximal methods provably reach the correct support and then converge linearly on it (Liang et al., 2017; Lee & Wright, 2012). Its trigger follows the certificate tradition of safe screening, where an observable quantity yields a certified conclusion (Ndiaye et al., 2017). Three elements are new, each with a consequence for the MCR task. The trigger is an event the practitioner observes, not an existence claim about the run. The guarantee therefore attaches to a real unmixing run at a moment its log identifies, with no unverifiable condition on how training began. The window protects the component set the solver actually selected, correct or not. Stability and truth are therefore separate verdicts: stability from this analysis, truth from the certificate of Theorem 2. The persistence bound treats the whole horizon

as one event. A single number therefore bounds the risk that the selected set drifts while the run continues past a checkpoint.

Three lemmas carry the proofs. Lemma 16 grounds the margin levels the assumptions demand and confines all dependence on the sample to one event: the flip gap of Theorem 1(b) yields per-sample energy margins at some parameter, under an expressivity hypothesis (part (i)); an aggregation argument transfers them to the dataset-mean gate whenever component prevalences avoid an explicit intermediate band (part (ii)); and a net argument makes the deviation of the empirical gaps uniform over the training region (part (iii)), with a scope remark recording at which scales its constants bind. Lemma 17 proves the certified stability window. Lemma 18 proves linear contraction of the frozen objective with bias, down to a noise floor, under the Polyak–Łojasiewicz inequality (part (i)), converts a small objective gap into selection margins (part (ii)), bounds the per-time-point risk (part (iii)), and makes the risk bound uniform in time by a maximal inequality (part (iv)). The assembly takes five steps.

Notation. The notation of Appendices A–D remains in force. New objects of the gate and the window:

$\theta_t,\ \Theta,\ R,\ D$	training iterates; the region they visit, $\Theta \subseteq B(\theta_{\mathrm{ref}}, R)$, a ball of radius R fixed in advance; D the parameter dimension of the gate
$e_j^{\mathrm{sel}}(x;\theta),\ e_j^{\mathrm{rej}}(x;\theta)$	the gate's two selection energies for component j at sample x (row j of E, §5.1)
$g_j(x;\theta)$	$= e_j^{\mathrm{rej}}(x;\theta) - e_j^{\mathrm{sel}}(x;\theta)$, the per-sample energy gap
$\hat{\Delta}_j(\theta),\ \Delta_j(\theta)$	$= \frac{1}{M}\sum_m g_j(x^{(m)};\theta)$ and $= \mathbb{E}_{x\sim\mathcal{D}}\, g_j(x;\theta)$, the mean gaps
$\hat{J}_t$	$= \{j \in [N] : \hat{\Delta}_j(\theta_t) > 0\}$, the evaluation support at step t
$J_\sharp,\ t_0,\ \gamma$	anchored support $\hat{J}_{t_0}$; anchor time; anchor level
$m_j(\theta),\ \hat{m}_j(\theta)$	signed margins: $\Delta_j(\theta)$, $\hat{\Delta}_j(\theta)$ for $j \in J_\sharp$, their negatives for $j \notin J_\sharp$
$U,\ L_e,\ L_\Delta$	energy bound; energy Lipschitz constant; $L_\Delta = 2L_e$
$G,\ \eta_t,\ \eta$	update-displacement constant; step sizes (constant η after t_0)
$T(\gamma)$	$= \sup\{T \in \mathbb{N} : L_\Delta G \sum_{s=t_0}^{t_0+T-1} \eta_s \le \gamma/4\}$, the stability horizon
E_{gap}	the concentration event of Lemma 16(iii)
$\mathcal{F}_t$	the σ-algebra of all randomness up to step t (the data included)

Objects of the descent analysis and of the margin grounding:

$f,\ f^\star,\ \mathcal{X}^\star,\ \mathcal{U}$	objective with the gate frozen at $J_\sharp$; its infimum over a neighborhood $\mathcal{U} \subseteq \Theta$; its minimizer set in $\mathcal{U}$
$L,\ \mu$	gradient-Lipschitz and Polyak–Łojasiewicz constants of f on $\mathcal{U}$
$\beta,\ \nu^2$	gradient bias bound; gradient variance bound
$\Phi(\eta)$	$= 2\beta^2/\mu + L\eta(\nu^2 + 2\beta^2)/\mu$, the noise floor
$\gamma^\star,\ \varepsilon^\star$	minimizer margin level; $\varepsilon^\star = \mu(\gamma^\star)^2/(8L_\Delta^2)$, the basin level
$z_t,\ c_\eta$	objective gap $f(\theta_t) - f^\star$; per-step injection $\eta\beta^2 + \frac{L}{2}\eta^2(\nu^2 + 2\beta^2)$
T_{ctr}	$= \left\lceil \frac{2}{\mu\eta} \log\big(2(f(\theta_{t_0}) - f^\star)/\varepsilon^\star\big) \right\rceil$, the contraction length
g_{gap}	flip gap of Theorem 1(b): $\min\{\lambda - \sigma^2 L_{\alpha_{\mathrm{w}}}^2,\ (c_{\min}s_{\min} - \sigma L_{\alpha_{\mathrm{w}}})^2 - \lambda\}$
$\theta^\dagger,\ a$	margin-realizing parameter (Lemma 16(i)); margin level $a = g_{\mathrm{gap}}/2$
$p_j,\ \hat{\rho}_j$	prevalence $\Pr_{x\sim\mathcal{D}}(j \in J(x))$, $J(x)$ the generative support of x; its empirical counterpart

The uniform deviation level of the empirical gaps, established in Lemma 16(iii), is

$$r_M(\alpha) \;=\; 2U\sqrt{\frac{2\big(\log\frac{2N}{\alpha} + D\log\big(3 + \frac{2RL_\Delta\sqrt{M}}{U}\big)\big)}{M}} \;+\; \frac{2U}{\sqrt{M}}. \tag{12}$$

Four conventions. First, symbol reuse. Within this appendix D is the parameter dimension of the gate network; the data matrix of Eq. 1 does not appear. Likewise μ, β, U, G, and L denote the objects

of the tables above, unrelated to the latent law, metric, usage functional, process, and criterion of Appendix D; the noise event E of Appendix C is not used; and T, $T(\gamma)$, T_{ctr} are step counts, not index sets. The letter g follows its subscript: $g_j(x;\theta)$, with component subscript, is a gate output, and g_t, with time subscript, is the stochastic update direction of (E6); the deterministic gradient is always written ∇f. Second, observability. A hat marks a quantity computable from the data and the current iterate at run time ($\hat{\Delta}_j$, $\hat{m}_j$, $\hat{J}_t$, $\hat{\rho}_j$); unhatted counterparts are population quantities. The anchor event is a condition on hatted quantities only, and $T(\gamma)$ is computable from architecture-level constants and the step schedule; by contrast $\gamma^\star$, μ, L, β, ν^2, $f^\star$, and therefore $\varepsilon^\star$, $\Phi(\eta)$, and T_{ctr}, are properties of the objective and its geometry, not computable at run time. Third, the gap sign matches the gate: the evaluation gate selects component j exactly when the selected state has the lower energy, so positive g_j favors selection, and $\hat{m}_j(\theta_t) > 0$ for all j implies $\hat{J}_t = J_\sharp$ (ties $\hat{\Delta}_j = 0$ break the converse; only this direction is used). Fourth, sums over an empty index range are zero, so $T(\gamma) \geq 0$ always; $T(\gamma) = \infty$ occurs for summable step schedules and only strengthens every conclusion, and for non-constant schedules $T(\gamma)$ depends on t_0.

Seven standing assumptions.

- *(E1)* (sampling) $x^{(1)}, \ldots, x^{(M)}$ are i.i.d. draws from a distribution $\mathcal{D}$ on $\mathbb{R}^d$.
- *(E2)* (bounded, Lipschitz energies) There are finite $U, L_e > 0$ and a ball $B(\theta_{\mathrm{ref}}, R)$ whose center and radius are fixed in advance, independently of the data and of the training noise, such that all training iterates lie in $\Theta \subseteq B(\theta_{\mathrm{ref}}, R)$ surely, and for all $j \in [N]$, both energy states $s \in \{\mathrm{sel}, \mathrm{rej}\}$, all x in the support of $\mathcal{D}$, and all $\theta, \theta' \in \Theta$,
$$\left|e_j^s(x;\theta)\right| \leq U, \qquad \left|e_j^s(x;\theta) - e_j^s(x;\theta')\right| \leq L_e\|\theta - \theta'\|.$$
Consequently $|g_j| \leq 2U$ and $g_j(x;\cdot)$, $\hat{\Delta}_j(\cdot)$, $\Delta_j(\cdot)$ are L_Δ-Lipschitz on Θ with $L_\Delta = 2L_e$.
- *(E3)* (bounded update displacement) $\|\theta_{t+1} - \theta_t\| \leq \eta_t\, G$ for every training step t.
- *(E4)* (anchor certificate) There exist a time t_0 and a level $\gamma > 0$ with $\hat{m}_j(\theta_{t_0}) \geq \gamma$ for all $j \in [N]$, the margins signed relative to $J_\sharp = \hat{J}_{t_0}$.
- *(E5)* (geometry after the anchor) The objective f with the gate frozen at $J_\sharp$ (reconstruction on the selected components plus the smooth regularizers of Eq. 7; the usage penalty is constant once the gate is frozen and is omitted) has L-Lipschitz gradient on a neighborhood $\mathcal{U} \subseteq \Theta$ containing the post-anchor iterates, satisfies there the Polyak–Łojasiewicz inequality $\frac{1}{2}\|\nabla f(\theta)\|^2 \geq \mu(f(\theta) - f^\star)$ with $\mu > 0$ and $\mathcal{X}^\star \neq \emptyset$, and satisfies the quadratic-growth bound $f(\theta) - f^\star \geq \frac{\mu}{2}\operatorname{dist}(\theta, \mathcal{X}^\star)^2$ on $\mathcal{U}$.
- *(E6)* (stochastic gradient model after the anchor) For $t \geq t_0$ the update is $\theta_{t+1} = \theta_t - \eta_t g_t$ with
$$\mathbb{E}[g_t \mid \mathcal{F}_t] = \nabla f(\theta_t) + b_t, \qquad \|b_t\| \leq \beta, \qquad \mathbb{E}\big[\|g_t - \mathbb{E}[g_t \mid \mathcal{F}_t]\|^2 \,\big|\, \mathcal{F}_t\big] \leq \nu^2.$$
- *(E7)* (margin regularity of minimizers) There is $\gamma^\star > 0$ such that every $\bar{\theta} \in \mathcal{X}^\star$ satisfies $m_j(\bar{\theta}) \geq \gamma^\star$ for all $j \in [N]$.

The seven conditions differ in status: some hold by construction, some are observable during a run, and some are assumptions about the run. The comments below state which is which. First, (E4) is the only condition on the run that is observable: $\hat{\Delta}_j(\theta_t)$ is computable from the forward passes the solver already performs. The implemented checkpointing records R^2, nMSE, and usage rather than margin means (Appendix H.2), so realizing the anchor certificate takes one additional logged quantity; (E4) is then a logging predicate, not an unverifiable assumption. The analysis does not prove that training reaches the anchor; it proves what follows once the anchor is observed.

Second, (E1) is a modeling assumption, and it is the only distributional one: no tail condition on the data is needed anywhere in this appendix, because the boundedness in (E2) already yields Hoeffding-type concentration regardless of the data law; the gMCR model of Eq. 2 enters only through the grounding results of Lemma 16.

Third, (E2) and (E3) are verifiable from the architecture: the gate is a three-layer $\tanh$ network, so its outputs are bounded and Lipschitz on any bounded weight set, with the constants controlled in practice by the energy regularizer $\lambda_{\mathrm{se}}\|E\|_2^2$ and by weight decay; and an Adam-type update moves

each coordinate by at most η_t (its stabilized update ratio is bounded by one), so (E3) holds deterministically with $G = \sqrt{D}$; weight decay enlarges G slightly.

Fourth, (E5)–(E7) are assumptions about the run, in force for all $t \geq t_0$: the analysis models post-anchor training as stochastic descent on the frozen objective throughout, an idealization of the implemented sampler whose recognized cost is the bias term β in (E6). That bias is the gap between the relaxed-gate gradient at temperature $\tau > 0$ and the gradient of f; it vanishes as $\tau \to 0$. ν^2 collects mini-batch, Gumbel-resampling, and energy-exploration noise. In the global setting the Polyak–Łojasiewicz inequality implies quadratic growth (Karimi et al., 2016, Theorem 2); on a neighborhood the implication additionally requires $\mathcal{U}$ to contain the relevant sublevel set, so (E5) hypothesizes the quadratic-growth bound directly.

Fifth, the gate itself is idealized: the evaluation support $\hat{J}_t$ reads the sign of dataset-mean gaps, whereas the implemented gate thresholds per-sample selection probabilities at τ_{thres}, which at $\tau_{\mathrm{eval}} = 0.01$ is a per-sample energy-gap floor of about 0.14 (Appendix G). Lemma 16 states exactly what grounds this idealization; whether training reaches the grounding parameter remains assumed.

Reading Proposition 2. Phase (i) of the main-text statement is parts (i) and (ii) of the full statement below, phase (ii) is part (iii), and part (iv) is a horizon bound stated here alone. In words, L_Δ bounds how fast a margin changes with the parameters, G bounds the displacement of one optimizer step, μ is the curvature of the objective on the support $J_\sharp$, $\Phi(\eta)$ is the noise floor at step size η, and $p \leq \frac{1}{2}$ is the floor of the failure probability. The analysis uses the run from the observed anchor onward. The parts describe one run in sequence. The window of part (i) is the interval $[\,t_0,\ t_0 + T(\gamma)\,]$, with $T(\gamma) = \lfloor \gamma/(4L_\Delta G\eta) \rfloor$: the margin divided by four times the largest change one optimizer step can make to any margin. Part (ii) is linear convergence, in the optimization sense, to a noise floor: $\mathbb{E}[f(\theta_t) - f^\star] \leq (1 - \mu\eta/2)^{t-t_0}(f(\theta_{t_0}) - f^\star) + \Phi(\eta)$ for t in the window, and the contraction completes inside the window, after T_{ctr} steps. Parts (iii) and (iv) concern the run after the window. The single-time bound of part (iii), $\Pr(\hat{J}_t \neq J_\sharp) \leq p + \frac{1}{2}(1 - \mu\eta/2)^{t-t_0-T_{\mathrm{ctr}}}$, says that at any one later step the selection has left $J_\sharp$ with probability at most p plus a term that decays geometrically. The horizon bound of part (iv) covers every step of a horizon of length T at once: the probability that the selection leaves $J_\sharp$ at any time in the horizon is at most $\frac{1}{2} + p + p\,\frac{\mu\eta}{2}\,T$. That bound starts at $\frac{1}{2} + p$, grows with T, and is below one for horizons shorter than about $1/(p\mu\eta)$ steps. For longer horizons the single-time bound is the statement that applies. Parts (iii) and (iv) need the concentration conditions above, which require more samples than the experiments of §6 use, so on those data parts (i) and (ii) apply and parts (iii) and (iv) do not.

Proposition 2, restated. The main text states this result under hypotheses given by reference; the full statement follows. Fix $\alpha \in (0, 1)$ and $p \in (0, \frac{1}{2}]$. Suppose (E1)–(E3) hold; suppose (E4) holds with anchor level γ at time t_0 and anchored support $J_\sharp = \hat{J}_{t_0}$; suppose (E5)–(E7) hold for all $t \geq t_0$, with every iterate θ_t, $t \geq t_0$, and every segment $[\theta_t, \theta_{t+1}]$ contained in $\mathcal{U}$, and with a constant step $\eta \leq \frac{1}{2L}$ for $t \geq t_0$. All expectations and probabilities over the training randomness below are conditional on $\mathcal{F}_{t_0}$. Suppose the compatibility conditions

$$r_M(\alpha) \ \leq\ \frac{\gamma}{4}, \tag{13}$$

$$r_M(\alpha) \ \leq\ \frac{\gamma^\star}{4}, \tag{14}$$

$$\Phi(\eta) \ \leq\ p\,\varepsilon^\star, \tag{15}$$

$$T(\gamma) \ \geq\ T_{\mathrm{ctr}}, \qquad \text{with } T_{\mathrm{ctr}} = 0 \text{ when } f(\theta_{t_0}) - f^\star \leq \varepsilon^\star/2. \tag{16}$$

Let E_{gap} be the event of Lemma 16(iii) at level α, so $\Pr(E_{\mathrm{gap}}) \geq 1 - \alpha$ over the sample. Then:

(i) (Certified block.) Deterministically, for every realization of the sample and of the training noise,

$$\hat{J}_t = J_\sharp \qquad \text{and} \qquad \hat{m}_j(\theta_t) \geq \frac{3\gamma}{4} \qquad \text{for all } j \in [N],\ t \in [t_0,\ t_0 + T(\gamma)];$$

on E_{gap}, under (13), also $m_j(\theta_t) \geq \gamma/2$ for all such j, t.

(ii) (Contraction.) For $0 \leq k \leq T(\gamma)$,

$$\mathbb{E}\big[f(\theta_{t_0+k}) - f^\star\big] \ \leq\ \Big(1 - \frac{\mu\eta}{2}\Big)^k \big(f(\theta_{t_0}) - f^\star\big) + \Phi(\eta),$$

and in particular $\mathbb{E}[f(\theta_{t_0+T_{\rm ctr}}) - f^\star] \le (\frac{1}{2} + p)\,\varepsilon^\star \le \varepsilon^\star$.

(iii) (Self-sustaining support.) For every fixed $t \ge t_0 + T_{\rm ctr}$,

$$\Pr\big(f(\theta_t) - f^\star > \varepsilon^\star\big) \;\le\; p + \frac{1}{2}\Big(1 - \frac{\mu\eta}{2}\Big)^{t - t_0 - T_{\rm ctr}},$$

and on every sample realization in $E_{\rm gap}$, under (14), the same bound holds for $\Pr\big(\hat{J}_t \ne J_\sharp\big)$.

(iv) (Uniform permanence.) For every horizon $T \in \mathbb{N}$,

$$\Pr\Big(\exists\, t \in [\,t_0 + T_{\rm ctr},\; t_0 + T_{\rm ctr} + T\,] : f(\theta_t) - f^\star \ge \varepsilon^\star\Big) \;\le\; \frac{1}{2} + p + p\,\frac{\mu\eta}{2}\,T,$$

and on every sample realization in $E_{\rm gap}$, under (14), the same bound holds for $\Pr\big(\exists\, t \in [\,t_0 + T_{\rm ctr},\;\; t_0 + T_{\rm ctr} + T\,] : \hat{J}_t \ne J_\sharp\big)$. For every $\delta \in (p, 1)$, replacing $T_{\rm ctr}$ by $T_{\rm ctr}(\delta) = \big\lceil \frac{2}{\mu\eta} \log_+\big((f(\theta_{t_0}) - f^\star)/((\delta - p)\,\varepsilon^\star)\big)\big\rceil$, with $\log_+ = \max\{\log, 0\}$, replaces $\frac{1}{2} + p$ by δ.

Lemma 16 grounds the margin levels and confines the sample randomness; Lemma 17 proves part (i); Lemma 18 proves parts (ii)–(iv); the assembly closes.

The first lemma connects the dynamics to the identifiable target of Theorem 1 and confines all dependence on the sample to one event: the flip gap of Theorem 1(b) (Appendix B) produces per-sample energy margins at some parameter $\theta^\dagger$ under an expressivity hypothesis, and prevalence separation transfers them to the dataset-mean gate that defines $\hat{J}_t$, so the levels demanded by (E4) and (E7) are attainable, not merely posited; whether training reaches such a parameter is exactly the anchor event, and is not proved. The uniformity of the deviation bound over the whole training region is a necessity, not a refinement: the iterates θ_t are functions of the same M samples that define $\hat{\Delta}_j$, so concentration applied at θ_t pointwise would be circular, and a union bound over time would tie the sample size to the window length; a net argument over Θ removes both defects at once, with a bound that holds at every point of every trajectory contained in Θ.

Lemma 16 (The empirical gate and the data)**.** (i) (From flip gap to per-sample margins.) *Assume the gMCR model (Eq. 2) with the spark condition (Definition 1). Fix $\alpha_{\rm w} \in (0,1)$, set $L_{\alpha_{\rm w}} = \sqrt{2\log(6NM/\alpha_{\rm w})}$, and assume the SNR condition and window of Theorem 1(b): $c_{\min} s_{\min} > 2\sigma L_{\alpha_{\rm w}}$ and $\sigma^2 L_{\alpha_{\rm w}}^2 < \lambda < (c_{\min} s_{\min} - \sigma L_{\alpha_{\rm w}})^2$, so that $g_{\rm gap} > 0$. Define, for each sample m and each $j \in [N]$, the ideal signed gaps*

$$\Delta_j^{\rm idl}\big(x^{(m)}\big) \;=\; \begin{cases} +\big[\,\ell_\lambda\big(J_m \setminus \{j\};\, x^{(m)}\big) - \ell_\lambda\big(J_m;\, x^{(m)}\big)\,\big], & j \in J_m, \\ -\big[\,\ell_\lambda\big(J_m \cup \{j\};\, x^{(m)}\big) - \ell_\lambda\big(J_m;\, x^{(m)}\big)\,\big], & j \notin J_m, \end{cases}$$

with ℓ_λ as in Eq. 3. On the event of Theorem 1(b), of probability at least $1 - \alpha_{\rm w}$ over the noise: (a) *$\Delta_j^{\rm idl}(x^{(m)}) \ge g_{\rm gap}$ for $j \in J_m$ and $\Delta_j^{\rm idl}(x^{(m)}) \le -g_{\rm gap}$ for $j \notin J_m$, for every m;* (b) *if some $\theta^\dagger \in \Theta$ satisfies the expressivity hypothesis*

$$\big|g_j\big(x^{(m)};\, \theta^\dagger\big) - \Delta_j^{\rm idl}\big(x^{(m)}\big)\big| \;\le\; \frac{g_{\rm gap}}{2} \qquad \text{for every } m \in [M],\; j \in [N], \tag{17}$$

then $g_j(x^{(m)}; \theta^\dagger) \ge g_{\rm gap}/2$ for $j \in J_m$ and $g_j(x^{(m)}; \theta^\dagger) \le -g_{\rm gap}/2$ for $j \notin J_m$, for every m.

(ii) (Mean-gate aggregation.) *In the setting of part (i), with (E1) and (E2), write $a = g_{\rm gap}/2$ and let $p_j = \Pr_{x\sim\mathcal{D}}(j \in J(x))$ be the prevalence of component j, where $J(x)$ is the generative support of x. Fix $\gamma > 0$ and $\kappa > 0$, and suppose every $j \in [N]$ satisfies either*

$$p_j \;\ge\; \frac{2U + \gamma}{a + 2U} + \kappa \quad \text{(the prevalent set } S^\star\text{)} \qquad \text{or} \qquad p_j \;\le\; \frac{a - \gamma}{a + 2U} - \kappa. \tag{18}$$

The prevalent alternative is satisfiable precisely when $\gamma + \kappa(a + 2U) \le a$. Then with probability at least $1 - \alpha_{\rm w} - 2N\exp(-2M\kappa^2)$ over the joint draw of the samples and their noise,

$$\hat{\Delta}_j(\theta^\dagger) \ge \gamma \quad (j \in S^\star), \qquad \hat{\Delta}_j(\theta^\dagger) \le -\gamma \quad (j \notin S^\star);$$

in particular the evaluation gate at $\theta^\dagger$ returns exactly $S^\star$, and the anchor certificate (E4) holds at $\theta^\dagger$ with level γ and reference support $S^\star$, i.e., (E4) is satisfied at any step whose iterate is $\theta^\dagger$.

(iii) (Uniform deviation of empirical gaps.) *Let (E1) and (E2) hold. For any $\rho \in (0, R]$ and $\alpha \in (0, 1)$, with probability at least $1 - \alpha$ over the sample,*

$$\sup_{\theta\in\Theta} \max_{j\in[N]} \left|\hat{\Delta}_j(\theta) - \Delta_j(\theta)\right| \;\le\; 2U\sqrt{\frac{2\big(\log\frac{2N}{\alpha} + D\log\big(1+\frac{2R}{\rho}\big)\big)}{M}} \;+\; 2L_\Delta\,\rho.$$

The choice $\rho^\star = \min\{R,\ U/(L_\Delta\sqrt{M})\}$ bounds the right side by $r_M(\alpha)$ of Eq. 12. E_{gap} denotes the event of the display at $\rho = \rho^\star$.

Proof. Part (i).

Step 1 (gap signs). On the event of Theorem 1(b), every selection obtained from J_m by adding or removing one component costs at least g_{gap}:

$$\ell_\lambda\big(A;\, x^{(m)}\big) - \ell_\lambda\big(J_m;\, x^{(m)}\big) \;\ge\; g_{\mathrm{gap}} \qquad \text{whenever } |A \triangle J_m| = 1.$$

For $j \in J_m$ the set $A = J_m \setminus \{j\}$ satisfies $|A \triangle J_m| = 1$, so

$$\Delta_j^{\mathrm{idl}}\big(x^{(m)}\big) \;=\; \ell_\lambda\big(J_m \setminus \{j\};\, x^{(m)}\big) - \ell_\lambda\big(J_m;\, x^{(m)}\big) \;\ge\; g_{\mathrm{gap}};$$

for $j \notin J_m$ the set $A = J_m \cup \{j\}$ satisfies $|A \triangle J_m| = 1$, so

$$\Delta_j^{\mathrm{idl}}\big(x^{(m)}\big) \;=\; -\big[\,\ell_\lambda\big(J_m \cup \{j\};\, x^{(m)}\big) - \ell_\lambda\big(J_m;\, x^{(m)}\big)\,\big] \;\le\; -g_{\mathrm{gap}}.$$

Step 2 (realization). For $j \in J_m$, by the triangle inequality, (17), and Step 1,

$$g_j\big(x^{(m)};\, \theta^\dagger\big) \;\ge\; \Delta_j^{\mathrm{idl}}\big(x^{(m)}\big) - \frac{g_{\mathrm{gap}}}{2} \;\ge\; g_{\mathrm{gap}} - \frac{g_{\mathrm{gap}}}{2} \;=\; \frac{g_{\mathrm{gap}}}{2},$$

and for $j \notin J_m$, by the same three facts,

$$g_j\big(x^{(m)};\, \theta^\dagger\big) \;\le\; \Delta_j^{\mathrm{idl}}\big(x^{(m)}\big) + \frac{g_{\mathrm{gap}}}{2} \;\le\; -g_{\mathrm{gap}} + \frac{g_{\mathrm{gap}}}{2} \;=\; -\frac{g_{\mathrm{gap}}}{2}.$$

Part (ii). Write $\chi_j^{(m)} = \mathbf{1}[\,j \in J_m\,]$ and $\hat{\rho}_j = \frac{1}{M}\sum_m \chi_j^{(m)}$.

Step 3 (pointwise sandwich). On the event of part (i), for every m and j,

$$(a + 2U)\,\chi_j^{(m)} - 2U \;\le\; g_j\big(x^{(m)};\, \theta^\dagger\big) \;\le\; (a + 2U)\,\chi_j^{(m)} - a.$$

Both cases check directly. For $\chi_j^{(m)} = 1$ the claim reads

$$a \;\le\; g_j\big(x^{(m)};\, \theta^\dagger\big) \;\le\; 2U,$$

the left side by part (i)(b) and the right by (E2); for $\chi_j^{(m)} = 0$ it reads

$$-2U \;\le\; g_j\big(x^{(m)};\, \theta^\dagger\big) \;\le\; -a,$$

the left side by (E2) and the right by part (i)(b). Averaging over m,

$$\hat{\rho}_j\,(a + 2U) - 2U \;\le\; \hat{\Delta}_j(\theta^\dagger) \;\le\; \hat{\rho}_j\,(a + 2U) - a.$$

Step 4 (deterministic gate). If $\hat{\rho}_j \ge (2U + \gamma)/(a + 2U)$, then by Step 3

$$\hat{\Delta}_j(\theta^\dagger) \;\ge\; \hat{\rho}_j\,(a + 2U) - 2U \;\ge\; (2U + \gamma) - 2U \;=\; \gamma,$$

and if $\hat{\rho}_j \le (a - \gamma)/(a + 2U)$, then by Step 3

$$\hat{\Delta}_j(\theta^\dagger) \;\le\; \hat{\rho}_j\,(a + 2U) - a \;\le\; (a - \gamma) - a \;=\; -\gamma.$$

Step 5 (prevalence concentration). For fixed j the indicators $\chi_j^{(m)}$, $m \in [M]$, are i.i.d. $\{0, 1\}$-valued with mean p_j by (E1), so Hoeffding's inequality (Hoeffding, 1963, Theorem 2) gives

$$\Pr\big(|\hat{\rho}_j - p_j| \ge \kappa\big) \;\le\; 2\exp(-2M\kappa^2).$$

Step 6 (union bound). Off the union of the N events of Step 5 and the complement of the event of part (i) (the failure probabilities add), $|\hat{\rho}_j - p_j| < \kappa$ for every j simultaneously. For $j \in S^\star$, by (18),

$$\hat{\rho}_j \;>\; p_j - \kappa \;\geq\; \frac{2U+\gamma}{a+2U},$$

and for $j \notin S^\star$, by (18),

$$\hat{\rho}_j \;<\; p_j + \kappa \;\leq\; \frac{a-\gamma}{a+2U},$$

so Step 4 applies to every j. Since $\gamma > 0$, the sign pattern of $\hat{\Delta}_j(\theta^\dagger)$ is exactly $S^\star$, which is the gate identity, and the margins clear γ, which is (E4) at $\theta^\dagger$.

Part (iii).

Step 7 (net). Fix a ρ-net $\mathcal{N}_\rho$ of $B(\theta_{\mathrm{ref}}, R)$ with

$$|\mathcal{N}_\rho| \;\leq\; \Big(1+\frac{2R}{\rho}\Big)^D$$

(Vershynin, 2018, §4.2). The net is fixed in advance, because the ball is data-independent by (E2).

Step 8 (tail at one net point). Fix $(j,\bar{\theta}) \in [N] \times \mathcal{N}_\rho$. By (E2), $g_j(x;\bar{\theta}) \in [-2U, 2U]$; by (E1) the summands of $\hat{\Delta}_j(\bar{\theta})$ are i.i.d.; and $\bar{\theta}$ is a fixed, data-independent point. Hoeffding's inequality (Hoeffding, 1963, Theorem 2) therefore applies:

$$\Pr\Big(\big|\hat{\Delta}_j(\bar{\theta}) - \Delta_j(\bar{\theta})\big| \geq u\Big) \;\leq\; 2\exp\Big(-\frac{2Mu^2}{(4U)^2}\Big) \;=\; 2\exp\Big(-\frac{Mu^2}{8U^2}\Big).$$

Step 9 (union bound and level). Set

$$u \;=\; 2U\sqrt{\frac{2\big(\log\frac{2N}{\alpha} + \log|\mathcal{N}_\rho|\big)}{M}}, \qquad \text{so} \qquad 2\exp\Big(-\frac{Mu^2}{8U^2}\Big) \;=\; \frac{\alpha}{N|\mathcal{N}_\rho|},$$

and a union bound over the $N|\mathcal{N}_\rho|$ pairs makes the total failure probability at most α.

Step 10 (Lipschitz transfer). On the complementary event, every $\theta \in \Theta$ has a net point $\bar{\theta}$ with $\|\theta - \bar{\theta}\| \leq \rho$, and since both $\hat{\Delta}_j$ and Δ_j are L_Δ-Lipschitz on Θ by (E2) (the empirical bound holds surely for every realization of the sample),

$$\big|\hat{\Delta}_j(\theta) - \Delta_j(\theta)\big| \;\leq\; \big|\hat{\Delta}_j(\theta) - \hat{\Delta}_j(\bar{\theta})\big| + \big|\hat{\Delta}_j(\bar{\theta}) - \Delta_j(\bar{\theta})\big| + \big|\Delta_j(\bar{\theta}) - \Delta_j(\theta)\big| \;\leq\; u + 2L_\Delta\rho.$$

Step 11 (canonical choice). If $\rho^\star = U/(L_\Delta\sqrt{M}) \leq R$, the transfer term and the net argument satisfy

$$2L_\Delta\rho^\star \;=\; \frac{2U}{\sqrt{M}}, \qquad 1+\frac{2R}{\rho^\star} \;=\; 1+\frac{2RL_\Delta\sqrt{M}}{U} \;\leq\; 3+\frac{2RL_\Delta\sqrt{M}}{U}.$$

If $\rho^\star = R$, then $R \leq U/(L_\Delta\sqrt{M})$, so

$$2L_\Delta\rho^\star \;=\; 2L_\Delta R \;\leq\; \frac{2U}{\sqrt{M}}, \qquad 1+\frac{2R}{\rho^\star} \;=\; 3 \;\leq\; 3+\frac{2RL_\Delta\sqrt{M}}{U}.$$

In both regimes the right side of the display is at most $r_M(\alpha)$ of Eq. 12. □

Two remarks on the hypothesis (17). First, the approximation target in (17) is data-defined, not chosen: the ideal gaps are computed at the true supports J_m, and by Theorem 1(a) and Proposition 1 the true supports are the unique minimal selections of the noiseless samples. Second, (17) is an expressivity assumption on the gate network, stated at the observed samples because Lemma 16(ii) applies the assumption at exactly those samples. Under this assumption, part (i) turns the existence of margins from a postulate into a consequence of identifiability plus expressivity, and fixes the scale at which margins can be demanded, namely $g_{\mathrm{gap}}/2$.

What the aggregation grounds, and what it does not. Under (18) the per-sample margins ground the dataset-mean gate with an explicit finite-sample failure probability: the mean-gate idealization of

$\hat{J}_t$ is discharged rather than assumed. What the aggregation does not ground is equally explicit. As $\gamma \to 0$ the band excluded by (18) tends to $\big(g_{\mathrm{gap}}/(g_{\mathrm{gap}}+4U),\ 4U/(g_{\mathrm{gap}}+4U)\big)$, a wide band when $U \gg g_{\mathrm{gap}}$, and a component whose prevalence falls inside the band genuinely has a mean gate of ambiguous sign: a real component active in few samples mixes a few positive per-sample margins with many negative ones. For such a component the mean gate is the wrong object, not a gap in the analysis, and the implemented gate is the better object: the implemented gate thresholds the per-sample selection probability at $\tau_{\mathrm{thres}} = 0.9999994$, which at $\tau_{\mathrm{eval}} = 0.01$ is a per-sample energy-gap floor of about 0.14 (Appendix G), and per-sample selection is the semantics that Lemma 16(i) grounds without any prevalence condition. Finally, the same sandwich taken in expectation reads $m_j(\theta^\dagger) \geq \gamma + \kappa(a+2U)$ for $j \in S^\star$ under (18), which is the scale demanded of $\gamma^\star$ in (E7): the expectation sandwich calibrates the scale of $\gamma^\star$. The calibration is not a certificate for $\gamma^\star$, because it treats the per-sample margins as holding for $\mathcal{D}$-almost every sample, while Lemma 16(i) provides them only on a finite-sample noise event.

Scope of the constants (the $D\log$ term). The parameter count D is a property of the solver's gate network, not of the inference problem: the pool, the supports, and the concentrations live in $\mathbb{R}^d$, and no statement of Theorems 1–3 involves D. D enters in exactly one place, the uniform concentration over the parameter ball in part (iii), where the $D\log(\cdot)$ factor counts the net that covers the full parameter ball, a deliberately coarse account of the search space rather than a measured property of the gate network. For the implemented architecture D is the full parameter count, of order 10^4 to 10^6, while the experiments use M between 10^3 and 10^4. At these scales the conditions (13)–(14) fail numerically: $r_M(\alpha)$ exceeds $U/2$, while any usable margin has $\gamma \leq 2U$ because each $\hat{\Delta}_j$ is an average of $[-2U, 2U]$-valued terms, so the requirement $r_M(\alpha) \leq \gamma/4 \leq U/2$ cannot hold. The failure is confined accordingly. The population rider of part (i) and the support conversions of parts (iii)–(iv) are, at experimental scale, conditional on hypotheses that fail numerically; the sure window of part (i) and the objective-gap bounds of parts (ii)–(iv) contain no r_M and stand as stated. The bound is read for the mechanism and the direction of the M-dependence: larger M tightens every margin uniformly over Θ, the mechanism behind the faster stabilization observed at $8N$ versus $4N$ samples (§6); the bound does not give certified constants at experimental scale.

The second lemma proves part (i) of the proposition: after the observed anchor, Lipschitz continuity of the energies and the bounded optimizer step alone confine every margin for an explicit number of steps. The window is sure, needing neither probability nor sample size; the concentration event enters only to upgrade empirical margins to population ones.

Lemma 17 (The anchor window)**.** *Let (E2)–(E4) hold with anchor level γ at time t_0, and recall the stability horizon*

$$T(\gamma) \;=\; \sup\Big\{T \in \mathbb{N} \;:\; L_\Delta\, G \sum_{s=t_0}^{t_0+T-1} \eta_s \;\leq\; \frac{\gamma}{4}\Big\} \qquad \Big(\textit{for constant } \eta:\;\; T(\gamma) = \Big\lfloor \frac{\gamma}{4L_\Delta G \eta} \Big\rfloor\Big). \tag{19}$$

Then: (i) (Margins persist.) *Deterministically, for every realization of the sample and of the training noise,*

$$\hat{m}_j(\theta_t) \geq \frac{3\gamma}{4} \qquad \textit{for all } j \in [N],\ t \in [\,t_0,\ t_0 + T(\gamma)\,].$$

(ii) (Constant support.) *On the same realizations, $\hat{J}_t = J_\sharp$ for all $t \in [\,t_0,\ t_0 + T(\gamma)\,]$.* (iii) (Population rider.) *If in addition (E1) and (13) hold, then on E_{gap} also $m_j(\theta_t) \geq \gamma/2$ for all such j and t.*

Proof. Step 1 (path length). By the triangle inequality and (E3),

$$\|\theta_t - \theta_{t_0}\| \;\leq\; \sum_{s=t_0}^{t-1} \|\theta_{s+1} - \theta_s\| \;\leq\; G \sum_{s=t_0}^{t-1} \eta_s.$$

Step 2 (Lipschitz). By (E2), Δ_j and $\hat{\Delta}_j$ are L_Δ-Lipschitz on Θ (the empirical statement holds surely), so Step 1 gives, for any $t_0 \leq t$,

$$\max_{j\in[N]} \big|\Delta_j(\theta_t) - \Delta_j(\theta_{t_0})\big| \;\leq\; L_\Delta\, G \sum_{s=t_0}^{t-1} \eta_s,$$

and the same bound holds for $\hat{\Delta}_j$, for every realization of the sample.

Step 3 (sure empirical route). By (E4), $\hat{m}_j(\theta_{t_0}) \geq \gamma$ for every j. The bound of Step 2 applies to the empirical gaps for every sample realization, so for $t \in [t_0,\, t_0 + T(\gamma)]$, by (19),

$$\hat{m}_j(\theta_t) \;\geq\; \hat{m}_j(\theta_{t_0}) - L_\Delta G \sum_{s=t_0}^{t-1} \eta_s \;\geq\; \gamma - \frac{\gamma}{4} \;=\; \frac{3\gamma}{4} \;>\; 0.$$

This is part (i).

Step 4 (support equality). Positivity of every signed empirical margin is exactly the statement that the sign pattern of $\hat{\Delta}_j(\theta_t)$ agrees with $J_\sharp$ (the gap-sign convention of the Notation paragraph), that is, $\hat{J}_t = J_\sharp$, which is part (ii).

Step 5 (population rider). On E_{gap}, by Lemma 16(iii) and (13), for every $\theta \in \Theta$ simultaneously,

$$\big|\hat{\Delta}_j(\theta) - \Delta_j(\theta)\big| \;\leq\; r_M(\alpha) \;\leq\; \frac{\gamma}{4};$$

in particular the bound holds at every iterate of the trajectory, with no independence between θ_t and the data required. Hence

$$m_j(\theta_t) \;\geq\; \hat{m}_j(\theta_t) - \frac{\gamma}{4} \;\geq\; \frac{3\gamma}{4} - \frac{\gamma}{4} \;=\; \frac{\gamma}{2}. \qquad \square$$

The rest of the proposition takes place inside this block: the descent of part (ii) runs on it, and the permanence of parts (iii)–(iv) starts from the state it certifies. The two routes of the lemma serve separate needs of that program. The sure route involves no probability: $T(\gamma)$ is computed from the Lipschitz constant L_Δ, the displacement bound G, and the step sizes alone, so an observed anchor certifies a support-stable window whose length is set by run parameters, at any sample size. The population rider upgrades the empirical margins to margins for the latent law on E_{gap}; the sample-size condition (13) enters the proposition only there. The two hypotheses respond to separate controls of a run: a larger M reduces $r_M(\alpha)$, easing (13); a smaller step or a smaller $L_\Delta G$ lengthens $T(\gamma)$. Both controls are observable in a run.

On the certified block the evaluation support is constant, so the objective the checkpoint sees reduces to f, the objective with the gate frozen at $J_\sharp$; the third lemma proves that f contracts linearly, in expectation, to an explicit noise floor, the standard Polyak–Łojasiewicz descent (Karimi et al., 2016, Theorems 1 and 4) carried out with a bias term, the recognized cost of training through a relaxed gate at $\tau > 0$. It then supplies the mechanism of self-sustenance: quadratic growth converts a small objective gap into parameter proximity to the minimizer set, the minimizer margins of (E7) and Lipschitz continuity convert proximity into margin positivity, so a small gap forces the gate to re-select $J_\sharp$ with no separate assumption, and Markov's inequality bounds the risk at each fixed time. The per-time bound cannot be summed over a long horizon, a union over T time points costing $\approx Tp$; treating the whole horizon as a single event, by compensating the objective gap into a supermartingale and applying a maximal inequality, makes the failure probability grow only at the noise-injection rate per step.

Lemma 18 (Descent, basin, and permanence)**.** (i) (One-step contraction with bias and noise.) *Let (E5) and (E6) hold, let $\eta_t \leq \frac{1}{2L}$, and suppose the segment $[\theta_t, \theta_{t+1}]$ lies in $\mathcal{U}$ almost surely given $\mathcal{F}_t$. Then*

$$\mathbb{E}\big[f(\theta_{t+1}) - f^\star \,\big|\, \mathcal{F}_t\big] \;\leq\; \Big(1 - \frac{\mu\eta_t}{2}\Big)\big(f(\theta_t) - f^\star\big) \;+\; \eta_t\beta^2 \;+\; \frac{L}{2}\eta_t^2\big(\nu^2 + 2\beta^2\big). \tag{20}$$

With a constant step $\eta \leq \frac{1}{2L}$, unrolling from t_0 gives, for every $k \geq 0$,

$$\mathbb{E}\big[f(\theta_{t_0+k}) - f^\star\big] \;\leq\; \Big(1 - \frac{\mu\eta}{2}\Big)^k\big(f(\theta_{t_0}) - f^\star\big) \;+\; \Phi(\eta), \qquad \Phi(\eta) = \frac{2\beta^2}{\mu} + \frac{L\eta(\nu^2 + 2\beta^2)}{\mu}. \tag{21}$$

(ii) (Margin from objective gap.) *Let (E2), (E5), and (E7) hold, and recall $\varepsilon^\star = \mu(\gamma^\star)^2/(8L_\Delta^2)$. Then every $\theta \in \mathcal{U}$ with $f(\theta) - f^\star \leq \varepsilon^\star$ satisfies*

$$m_j(\theta) \;\geq\; \frac{\gamma^\star}{2} \qquad \textit{for all } j \in [N].$$

If moreover (E1) and (14) hold, then on E_{gap} additionally $\hat{m}_j(\theta) \geq \gamma^\star/4 > 0$, so the evaluation gate at θ selects $J_\sharp$.

(iii) (Per-time-point permanence.) *In the setting of parts (i) and (ii), with constant step $\eta \leq \frac{1}{2L}$, fix $p \in (0,1)$ and suppose (15), $\Phi(\eta) \leq p\,\varepsilon^\star$. Let $t_1 \geq t_0$ be a time with $\mathbb{E}[f(\theta_{t_1}) - f^\star] \leq \varepsilon^\star/2$. Then for every fixed $t \geq t_1$, with iterates in $\mathcal{U}$,*

$$\Pr\big(f(\theta_t) - f^\star > \varepsilon^\star\big) \;\leq\; \frac{1}{2}\Big(1 - \frac{\mu\eta}{2}\Big)^{t-t_1} + p,$$

and on every sample realization in E_{gap}, under (14), the same bound holds for $\Pr(\hat{J}_t \neq J_\sharp)$. The probabilities are over the training randomness only; the data is $\mathcal{F}_{t_1}$-measurable, so the bound and the entry condition are both read per data realization.

(iv) (Uniform-in-time permanence.) *In the setting of parts (i) and (ii), with constant step $\eta \leq \frac{1}{2L}$, write $z_t = f(\theta_t) - f^\star \geq 0$ and*

$$c_\eta \;=\; \eta\beta^2 + \frac{L}{2}\eta^2\big(\nu^2 + 2\beta^2\big) \;=\; \frac{\mu\eta}{2}\,\Phi(\eta),$$

the per-step injection of (20). Then for every time $t_1 \geq t_0$, every horizon $T \in \mathbb{N}$, and every level $\varepsilon > 0$, with iterates in $\mathcal{U}$,

$$\Pr\Big(\exists\, t \in [\,t_1,\; t_1 + T\,] :\; z_t \geq \varepsilon\Big) \;\leq\; \frac{\mathbb{E}[z_{t_1}] + Tc_\eta}{\varepsilon}. \tag{22}$$

Under (15) (applied below at the basin level $\varepsilon^\star$) the right side is at most $\mathbb{E}[z_{t_1}]/\varepsilon^\star + p\,\frac{\mu\eta}{2}\,T$. On the complementary event every t in the horizon has gap below $\varepsilon^\star$, so on every sample realization in E_{gap}, under (14), part (ii) yields

$$\hat{J}_t = J_\sharp \qquad \textit{and} \qquad \hat{m}_j(\theta_t) \geq \frac{\gamma^\star}{4} \qquad \textit{for all } t \in [\,t_1,\; t_1 + T\,] \textit{ simultaneously.}$$

(Expectations conditional on $\mathcal{F}_{t_0}$; conditioning on $\mathcal{F}_{t_1}$ instead replaces $\mathbb{E}[z_{t_1}]$ by the observed gap z_{t_1}.)

Proof. Part (i).

Step 1 (smoothness). By the L-Lipschitz gradient of (E5) and the quadratic upper bound it implies along the segment (Bubeck, 2015, Lemma 3.4),

$$f(\theta_{t+1}) \;\leq\; f(\theta_t) - \eta_t\big\langle \nabla f(\theta_t),\, g_t\big\rangle + \frac{L}{2}\eta_t^2\|g_t\|^2.$$

Step 2 (inner product with bias). Taking $\mathbb{E}[\,\cdot \mid \mathcal{F}_t]$ and writing $\nabla f = \nabla f(\theta_t)$, by (E6) and Cauchy–Schwarz,

$$\mathbb{E}\big[\langle \nabla f, g_t\rangle \,\big|\, \mathcal{F}_t\big] \;=\; \|\nabla f\|^2 + \langle \nabla f, b_t\rangle \;\geq\; \|\nabla f\|^2 - \beta\|\nabla f\|,$$

and since $0 \leq \big(\frac{1}{2}\|\nabla f\| - \beta\big)^2$ expands to the bound on the left below,

$$\beta\|\nabla f\| \;\leq\; \frac{1}{4}\|\nabla f\|^2 + \beta^2, \qquad \text{so} \qquad \mathbb{E}\big[\langle \nabla f, g_t\rangle \,\big|\, \mathcal{F}_t\big] \;\geq\; \frac{3}{4}\|\nabla f\|^2 - \beta^2.$$

Step 3 (quadratic term). By the variance bound of (E6) and $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$,

$$\mathbb{E}\big[\|g_t\|^2 \,\big|\, \mathcal{F}_t\big] \;\leq\; \nu^2 + \|\nabla f + b_t\|^2 \;\leq\; \nu^2 + 2\|\nabla f\|^2 + 2\beta^2.$$

Step 4 (collect). Substituting Steps 2 and 3 into Step 1,

$$\mathbb{E}\big[f(\theta_{t+1}) \,\big|\, \mathcal{F}_t\big] \;\leq\; f(\theta_t) - \eta_t\Big(\frac{3}{4} - L\eta_t\Big)\|\nabla f\|^2 + \eta_t\beta^2 + \frac{L}{2}\eta_t^2\big(\nu^2 + 2\beta^2\big).$$

Step 5 (Polyak–Łojasiewicz). With $\eta_t \leq \frac{1}{2L}$,

$$\frac{3}{4} - L\eta_t \;\geq\; \frac{3}{4} - \frac{1}{2} \;=\; \frac{1}{4},$$

and the inequality of (E5) gives

$$-\frac{\eta_t}{4}\|\nabla f\|^2 \;\le\; -\frac{\mu\eta_t}{2}\big(f(\theta_t)-f^\star\big)$$

(Karimi et al., 2016, Theorem 4). Subtracting $f^\star$ from both sides of Step 4 yields (20).

Step 6 (unroll). Take total expectations in (20) and iterate. The constants satisfy $\mu \le L$: L-smoothness gives $\|\nabla f(\theta)\|^2 \le 2L\big(f(\theta)-f^\star\big)$ (the descent lemma evaluated at a step of length $1/L$), the Polyak–Łojasiewicz inequality gives $\|\nabla f(\theta)\|^2 \ge 2\mu\big(f(\theta)-f^\star\big)$, and dividing the two bounds at any θ with $f(\theta) > f^\star$ yields $\mu \le L$ (if no such θ exists, every bound below is trivial). So with $\eta \le \frac{1}{2L}$,

$$\frac{\mu\eta}{2} \;\le\; \frac{\mu}{4L} \;\le\; \frac{1}{4}, \qquad \text{hence} \qquad 1-\frac{\mu\eta}{2} \in [0,1).$$

The accumulated injections are bounded by the geometric series

$$\sum_{i\ge 0}\Big(1-\frac{\mu\eta}{2}\Big)^i\Big(\eta\beta^2+\frac{L}{2}\eta^2(\nu^2+2\beta^2)\Big) \;\le\; \frac{2}{\mu\eta}\Big(\eta\beta^2+\frac{L}{2}\eta^2(\nu^2+2\beta^2)\Big) \;=\; \Phi(\eta),$$

which is (21).

Part (ii).

Step 7 (proximity). By the quadratic-growth bound of (E5),

$$\operatorname{dist}\big(\theta,\mathcal{X}^\star\big) \;\le\; \sqrt{\frac{2(f(\theta)-f^\star)}{\mu}} \;\le\; \sqrt{\frac{2\varepsilon^\star}{\mu}} \;=\; \frac{\gamma^\star}{2L_\Delta}.$$

Step 8 (margin transfer). Fix $\epsilon > 0$ and pick $\bar\theta \in \mathcal{X}^\star$ with $\|\theta-\bar\theta\| \le \operatorname{dist}(\theta,\mathcal{X}^\star)+\epsilon$. By (E7) and the Lipschitz property of (E2),

$$m_j(\theta) \;\ge\; m_j(\bar\theta) - L_\Delta\|\theta-\bar\theta\| \;\ge\; \gamma^\star - L_\Delta\Big(\frac{\gamma^\star}{2L_\Delta}+\epsilon\Big) \;=\; \frac{\gamma^\star}{2} - L_\Delta\epsilon,$$

and letting $\epsilon \downarrow 0$ gives $m_j(\theta) \ge \gamma^\star/2$.

Step 9 (empirical rider). On E_{gap} with (14),

$$\hat m_j(\theta) \;\ge\; m_j(\theta) - r_M(\alpha) \;\ge\; \frac{\gamma^\star}{2} - \frac{\gamma^\star}{4} \;=\; \frac{\gamma^\star}{4} \;>\; 0,$$

and positivity of every signed empirical margin is the gate identity (the gap-sign convention of the Notation paragraph).

Part (iii).

Step 10 (unroll from t_1). By (21) started at t_1 and the entry condition,

$$\mathbb{E}\big[f(\theta_t)-f^\star\big] \;\le\; \Big(1-\frac{\mu\eta}{2}\Big)^{t-t_1}\frac{\varepsilon^\star}{2} + \Phi(\eta).$$

Step 11 (Markov). By Markov's inequality at level $\varepsilon^\star$ and (15),

$$\Pr\big(f(\theta_t)-f^\star > \varepsilon^\star\big) \;\le\; \frac{(1-\frac{\mu\eta}{2})^{t-t_1}\,\frac{\varepsilon^\star}{2} + p\,\varepsilon^\star}{\varepsilon^\star} \;=\; \frac{1}{2}\Big(1-\frac{\mu\eta}{2}\Big)^{t-t_1} + p.$$

Step 12 (conversion). On the complementary event $\{f(\theta_t)-f^\star \le \varepsilon^\star\}$, intersected with E_{gap}, part (ii) gives $\hat J_t = J_\sharp$; the support-failure event is therefore contained in the gap-exceedance event, on E_{gap}.

Part (iv).

Step 13 (drift bound). Since $z_t \ge 0$ and $1-\frac{\mu\eta}{2} \le 1$, the one-step recursion (20) implies

$$\mathbb{E}\big[z_{t+1}\,\big|\,\mathcal{F}_t\big] \;\le\; \Big(1-\frac{\mu\eta}{2}\Big)z_t + c_\eta \;\le\; z_t + c_\eta.$$

Step 14 (compensated supermartingale). Define, for $t \in [\,t_1,\ t_1+T\,]$,

$$Y_t \;=\; z_t + \big(T-(t-t_1)\big)\,c_\eta,$$

the gap plus a deterministic compensator that absorbs the remaining injections in advance. By Step 13,

$$\mathbb{E}\big[Y_{t+1}\,\big|\,\mathcal{F}_t\big] \;=\; \mathbb{E}\big[z_{t+1}\,\big|\,\mathcal{F}_t\big] + \big(T-(t+1-t_1)\big)c_\eta \;\le\; z_t + c_\eta + \big(T-(t-t_1)\big)c_\eta - c_\eta \;=\; Y_t,$$

so (Y_t) is a supermartingale on the horizon, and $Y_t \ge z_t \ge 0$ there.

Step 15 (maximal inequality). Let $\tau = \min\{t \ge t_1 : Y_t \ge \varepsilon^\star\}$, with $\tau = \infty$ if no such t exists in the horizon. For $t_1 \le n < t_1 + T$,

$$Y_{\tau\wedge(n+1)} - Y_{\tau\wedge n} \;=\; \mathbf{1}[\,\tau > n\,]\big(Y_{n+1} - Y_n\big),$$

and since $\{\tau > n\} \in \mathcal{F}_n$, taking $\mathbb{E}[\,\cdot \mid \mathcal{F}_n]$ and using Step 14,

$$\mathbb{E}\big[Y_{\tau\wedge(n+1)}\,\big|\,\mathcal{F}_n\big] - Y_{\tau\wedge n} \;=\; \mathbf{1}[\,\tau > n\,]\Big(\mathbb{E}\big[Y_{n+1}\,\big|\,\mathcal{F}_n\big] - Y_n\Big) \;\le\; 0.$$

Iterating over n and taking total expectations,

$$\mathbb{E}\big[Y_{\tau\wedge(t_1+T)}\big] \;\le\; \mathbb{E}\big[Y_{t_1}\big].$$

On $\{\tau \le t_1 + T\}$, $Y_{\tau\wedge(t_1+T)} = Y_\tau \ge \varepsilon^\star$, and $Y \ge 0$ everywhere on the horizon, so

$$\varepsilon^\star\,\Pr\big(\tau \le t_1 + T\big) \;\le\; \mathbb{E}\big[Y_{\tau\wedge(t_1+T)}\big] \;\le\; \mathbb{E}\big[Y_{t_1}\big],$$

the maximal inequality of Ville (1939) for nonnegative supermartingales, bounding the running maximum by the initial expectation (in its modern time-uniform form, Howard et al., 2020, Lemma 1). This is the step that replaces the per-time-point Markov bound of part (iii): the submartingale inequality of Doob bounds by the terminal value and cannot deliver uniformity from an initial condition.

Step 16 (assembly). On the horizon $Y_t \ge z_t$, so the z-exceedance event is contained in $\{\tau \le t_1 + T\}$, and

$$\mathbb{E}\big[Y_{t_1}\big] \;=\; \mathbb{E}\big[z_{t_1}\big] + Tc_\eta,$$

which is (22). Under (15),

$$\frac{Tc_\eta}{\varepsilon^\star} \;=\; \frac{\mu\eta}{2}\,\frac{T\,\Phi(\eta)}{\varepsilon^\star} \;\le\; p\,\frac{\mu\eta}{2}\,T.$$

The support statement follows by applying part (ii) at every t of the complementary event, on E_{gap}. □

With the three lemmas in place, the proposition follows by composition, and each step of the proof completes one stage of the solver's stabilization. Step 1 sets the conditioning: every statement is read from the run's log at the anchor. Step 2 gives part (i): by Lemma 17, bounded optimizer steps cannot move any margin across zero for $T(\gamma)$ steps, so the selection holds fixed. Step 3 gives part (ii): with the selection frozen, the objective the run descends is a single smooth function, and Lemma 18(i) contracts it to its noise floor, with (16) guaranteeing that the contraction finishes inside the window. Steps 4 and 5 give parts (iii) and (iv): the contracted gap bounds the flip risk at each fixed time through Markov's inequality, Lemma 18(ii) converts a small gap into re-selection of $J_\sharp$, and Lemma 18(iv), applied at the entry time, treats the whole horizon as a single event. Every population statement holds on the concentration event E_{gap} of Lemma 16.

The probabilistic language of the proof describes an ordinary training run of the EB-gMCR solver of §5; the run in every statement below is that solver running. The symbol $\mathcal{F}_t$, defined in the notation table, stands for the training log of the run up to step t: the dataset together with every random choice the solver has made through step t. A quantity is $\mathcal{F}_{t_0}$-measurable when the log up to the anchor time determines it; the iterate θ_{t_0} and the anchored support $J_\sharp$ are the two examples the proof uses. Conditioning on $\mathcal{F}_{t_0}$ fixes the observed history and lets only the randomness after the anchor vary, so every probability below is a statement about how the solver's run continues, given what its log already shows at the anchor.

Proof of Proposition 2. Step 1 (conditioning). All expectations and probabilities over the training randomness are conditional on $\mathcal{F}_{t_0}$. The anchor time is a stopping time, but the anchor event and $f(\theta_{t_0}) - f^\star$ are $\mathcal{F}_{t_0}$-measurable; decomposing over the events $\{t_0 = s\} \in \mathcal{F}_s$ and iterating on each, the one-step recursion of Lemma 18(i) applies conditionally on $\mathcal{F}_{t_0}$, and conditions (15)–(16) are $\mathcal{F}_{t_0}$-measurable. The conditioning is suppressed in the notation below. The sample is $\mathcal{F}_{t_0}$-measurable as well, so every statement made on E_{gap} is a statement per sample realization.

Step 2 (part (i)). This is Lemma 17 verbatim: the block support and the empirical margins hold surely, and the population margins follow on E_{gap} under (13).

Step 3 (part (ii)). On the certified block the evaluation support is constant by part (i), the usage penalty is constant once the gate is frozen, and Lemma 18(i) applies to f; unrolling (21) for $0 \le k \le T(\gamma)$ gives the display, and (16) guarantees $T_{\mathrm{ctr}} \le T(\gamma)$, so the contraction completes within the block. By the definition of T_{ctr} and $\log(1-x) \le -x$ for $x \in [0,1)$,

$$\Big(1 - \frac{\mu\eta}{2}\Big)^{T_{\mathrm{ctr}}} \big(f(\theta_{t_0}) - f^\star\big) \;\le\; \exp\Big(-\frac{\mu\eta}{2}\, T_{\mathrm{ctr}}\Big)\big(f(\theta_{t_0}) - f^\star\big) \;\le\; \frac{\varepsilon^\star}{2},$$

so, by (15) and $p \le \frac{1}{2}$,

$$\mathbb{E}\big[f(\theta_{t_0+T_{\mathrm{ctr}}}) - f^\star\big] \;\le\; \frac{\varepsilon^\star}{2} + \Phi(\eta) \;\le\; \Big(\frac{1}{2} + p\Big)\varepsilon^\star \;\le\; \varepsilon^\star.$$

This is exactly the in-particular bound of part (ii).

Step 4 (part (iii)). Lemma 18(iii) cannot be applied at $t_1 = t_0 + T_{\mathrm{ctr}}$, because Step 3 delivers entry level $(\frac{1}{2}+p)\varepsilon^\star$ rather than the $\varepsilon^\star/2$ its hypothesis requires; instead unroll (21) once from t_0, without restarting: for $t \ge t_0 + T_{\mathrm{ctr}}$, by Step 3 and (15),

$$\mathbb{E}\big[f(\theta_t) - f^\star\big] \;\le\; \Big(1 - \frac{\mu\eta}{2}\Big)^{t-t_0}\big(f(\theta_{t_0}) - f^\star\big) + \Phi(\eta) \;\le\; \Big(1 - \frac{\mu\eta}{2}\Big)^{t-t_0-T_{\mathrm{ctr}}} \frac{\varepsilon^\star}{2} + p\,\varepsilon^\star,$$

and Markov's inequality at level $\varepsilon^\star$ gives the gap bound of part (iii), per sample realization. On realizations in E_{gap}, under (14), Lemma 18(ii) converts $\{f(\theta_t) - f^\star \le \varepsilon^\star\}$ into $\{\hat{J}_t = J_\sharp\}$, which is the support form.

Step 5 (part (iv)). Apply Lemma 18(iv) at $t_1 = t_0 + T_{\mathrm{ctr}}$. By Step 3 the entry satisfies $\mathbb{E}[z_{t_1}] \le (\frac{1}{2} + p)\,\varepsilon^\star$ (the floor does not contract away, whence the $\frac{1}{2} + p$), and $T c_\eta/\varepsilon^\star \le p\,\frac{\mu\eta}{2}\,T$ by (15). The right side of (22) at level $\varepsilon^\star$ is therefore at most $\frac{1}{2} + p + p\,\frac{\mu\eta}{2}\,T$, which is exactly the bound of part (iv); on every sample realization in E_{gap}, under (14), the support form over the same horizon is the conversion stated in Lemma 18(iv). With the deeper entry $T_{\mathrm{ctr}}(\delta)$, by the argument of Step 3,

$$\Big(1 - \frac{\mu\eta}{2}\Big)^{T_{\mathrm{ctr}}(\delta)} \big(f(\theta_{t_0}) - f^\star\big) \;\le\; (\delta - p)\,\varepsilon^\star,$$

so, adding the floor and using (15),

$$\mathbb{E}\big[z_{t_0+T_{\mathrm{ctr}}(\delta)}\big] \;\le\; (\delta - p)\,\varepsilon^\star + \Phi(\eta) \;\le\; \delta\,\varepsilon^\star,$$

replacing $\frac{1}{2} + p$ by δ. □

Concluding remarks. Taking the proposition part by part:

- *Part (i), the window.* The conclusion is sure, at any sample size, and it transfers to the implemented solver as proved: Lemma 17 consumes only the Lipschitz continuity of (E2) and the bounded update of (E3), and the normalized AdamW step satisfies (E3) by construction.
- *Part (ii), the contraction.* The rate is proved for the SGD-form update of (E6); its AdamW counterpart is open. The exploration noise of Eq. 6 sits inside (E6), reaching the parameters through the gradient with its effect carried by β and ν^2, and annealing τ reduces the bias β and with it the floor $\Phi(\eta)$, matching the implemented schedule. Smoothness of f is idealized: the ReLU layers of the concentration network break it at kink sets, a softplus reparametrization restores it at a larger L, and the window needs only Lipschitz continuity. The analysis also holds the usage weight and the temperature fixed after the anchor, while the implementation continues to update both; the window is unaffected, and the contraction rate inherits the open status it has under AdamW.

- *Parts (iii) and (iv), permanence.* After the contraction, the per-time failure probability decays geometrically to p, and the uniform bound accrues failure at p per $2/(\mu\eta)$ steps rather than p per time point queried, an improvement by the factor $2/(\mu\eta) \gg 1$; its leading term reflects entering the basin only $\varepsilon^\star/2$-deep in expectation. The linear growth in T is sharp: a Gaussian noise model consistent with (E6) leaves the basin $\{z_t \le \varepsilon^\star\}$ with per-step probability bounded away from zero, so any uniform bound must grow with the horizon. The support forms consume the concentration conditions, whose status at implemented scale is the scope remark after Lemma 16 (a norm-based or energy-head covering would reduce the $D\log(\cdot)$ factor; none is shown sufficient here), and containment of long-horizon segments in $\mathcal{U}$ remains assumed, with projection onto a convex $\mathcal{U}$ (Karimi et al., 2016, Theorem 5) or truncation of the injected noise, which only reduces ν^2, the enforcing modifications.

Two facts stand outside the parts. Nothing here proves that training reaches the anchor: the selection dynamics under the scheduled λ_t, the annealed τ_t, and the Gumbel sampling are not analyzed. And whether the anchored support is the true one is a variational fact, not a dynamical one: inside the SNR window of Theorem 1(b) the true selection is the strict optimum with gap g_{gap}, and Lemma 16 shows margins at the demanded scale exist at some parameter, but whether the run's $J_\sharp$ is that support remains an assumption about the run.

What this appendix establishes. After an observed anchor, every checkpoint taken within the certified window records the same support (part (i)); by the end of the window the frozen objective has contracted to its noise floor (part (ii)); afterwards the support maintains itself, at a risk quantified per time point (part (iii)) and per horizon (part (iv)). No other result of the paper consumes this proposition. The reading of the mechanism, its scope at implemented scale, and its consequences for the MCR task are stated in §5.4.

F A Toy Example: The Model, Identifiability, and Reuse

This appendix works the model of §3 on three samples and two components, small enough to check by hand. It shows four things in order. Without a sample in which one component occurs apart from the other, the data admit wrong decompositions that no method is guaranteed to reject. One further sample removes them. With all three samples, minimal usage still selects a wrong decomposition if the spark condition is dropped. A recovered pool decodes new samples. Two of the samples contain one component alone, so the truth of the toy can be read by inspection.

Setup. The pool has two unit-norm component profiles in $\mathbb{R}^4$,

$$s_1 = [0.6,\ 0.8,\ 0,\ 0], \qquad s_2 = [0,\ 0,\ 0.8,\ 0.6].$$

The two profiles are orthogonal, so $\sigma_{\min}([s_1\ s_2]) = 1$. The constants of Eq. 2 are $k = 2$, $c_{\min} = 1$, and $c_{\max} = 10$, and the spark condition is assumed with margin $s_{\min} = 1/2$, which the pool meets. There are three noiseless samples:

$$\begin{aligned} x_1 &= 5\,s_1 = [3,4,0,0], & J_1 &= \{1\},\\ x_2 &= 5\,s_2 = [0,0,4,3], & J_2 &= \{2\},\\ x_3 &= 5\,s_1 + 5\,s_2 = [3,4,4,3], & J_3 &= \{1,2\}. \end{aligned}$$

The concentration matrix is $[[5,0],[0,5],[5,5]]$, the occurrence matrix δ has rows $[1,0],[0,1],[1,1]$, and the total usage is 4. By Proposition 1, each sample has exactly one decomposition over this pool with at most two components and concentrations of at least $c_{\min}$.

Without x_1, minimal usage prefers a wrong decomposition. Consider x_2 and x_3 alone. Component 1 then never occurs apart from component 2, so it has no isolating pair. An isolating pair is a pair of samples sharing exactly one component, and the pinning step of Theorem 2(b) relies on one (§4.2). Two kinds of wrong decomposition reproduce both samples exactly within the constraints of Eq. 2, and both satisfy the spark condition. What fails here is the data, and not the spark condition. The first kind replaces component 1 by the whole of x_3:

$$s' = \frac{x_3}{\|x_3\|} \approx [0.42,\ 0.57,\ 0.57,\ 0.42], \qquad x_2 = 5\,s_2, \qquad x_3 = \sqrt{50}\,s'.$$

The pool $\{s_2, s'\}$ has $\sigma_{\min} \approx 0.54 \ge s_{\min}$ and total usage 2, below the usage 3 of the truth on these two samples, so minimal usage prefers it to the truth.

A continuum of decompositions has the usage of the truth. The second kind is a continuum. Let $a \in [c_{\min}, 5]$ be the concentration of s_2 assigned to x_3, and set

$$r(a) = x_3 - a\, s_2, \qquad \hat{s}_1(a) = \frac{r(a)}{\|r(a)\|}, \qquad x_2 = 5\, s_2, \qquad x_3 = \|r(a)\|\, \hat{s}_1(a) + a\, s_2.$$

For every a the decomposition is exact, has usage 3, keeps its concentrations inside $[c_{\min}, c_{\max}]$, and has $\sigma_{\min} \geq 0.6 \geq s_{\min}$. The value $a = 5$ is the truth, and every other a gives a wrong profile, for instance

$$a = 2.5: \qquad \hat{s}_1(2.5) \approx [0.54,\ 0.72,\ 0.36,\ 0.27], \qquad x_3 = 5.59\, \hat{s}_1(2.5) + 2.5\, s_2.$$

A penalty on the size of the concentrations, the lasso and ridge penalties of §4.1, selects a wrong member of the continuum:

$$\ell_1: \quad \sum |c| = 5 + \|r(a)\| + a, \qquad \text{minimum at } a = 1;$$
$$\ell_2: \quad \sum c^2 = 25 + \|r(a)\|^2 + a^2 = 75 - 10a + 2a^2, \qquad \text{minimum at } a = 2.5.$$

From x_2 and x_3 alone, no method is guaranteed to select the truth.

One further sample removes the wrong decompositions. Now include x_1. Since $\hat{s}_1(a) = (5s_1 + (5-a)s_2)/\|r(a)\|$, the only way to write x_1 over the pool $\{\hat{s}_1(a), s_2\}$ is

$$x_1 = \|r(a)\|\, \hat{s}_1(a) - (5-a)\, s_2.$$

The concentration of s_2 is negative for every $a \neq 5$, which Eq. 2 forbids. The same holds for the pool $\{s_2, s'\}$ of the first kind, which is the case $a = 0$. Adding a third profile in the plane of s_1 and s_2 does not rescue these pools, since three unit profiles in one plane are linearly dependent, which violates Definition 1. With x_1 present, the truth is the only decomposition within the constraints among pools in that plane. Samples x_1 and x_3 share exactly component 1, and they are the isolating pair that was missing.

The spark condition is needed. With all three samples, a decomposition with lower usage than the truth exists:

$$x_1 = 5\, s_1, \qquad x_2 = 5\, s_2, \qquad x_3 = \sqrt{50}\, s_3 \quad \text{with } s_3 = x_3/\|x_3\| = s',$$

at total usage $3 < 4$. The profile s_3 is a combination of s_1 and s_2, so

$$\sigma_{\min}\big([\, s_1 \ \ s_2 \ \ s_3\,]\big) = 0,$$

and the pool $\{s_1, s_2, s_3\}$ violates Definition 1. With the spark condition dropped, minimal usage selects this pool. With the spark condition and all three samples, minimal usage selects the truth, as Theorem 2(b) states for a learned pool without noise.

Reuse. Two new samples of the same process, $x_4 = 10\, s_1 = [6, 8, 0, 0]$ and $x_5 = 2\, s_1 + 7\, s_2 = [1.2,\ 1.6,\ 5.6,\ 4.2]$, are decoded over the recovered pool with nothing re-trained (Corollary 1). The profiles are orthonormal, so the decoding reduces to the inner products $x \cdot s_1$ and $x \cdot s_2$, which give $(10, 0)$ for x_4 and $(2, 7)$ for x_5. The support of each sample is the set of components with a concentration of at least $c_{\min}$, which is $\{1\}$ and $\{1, 2\}$. A factorization of the first three samples gives no such decoding, since the two new samples are not in its data matrix.

G GATE NUMERICS

This appendix gives the four things §5 points here for: the selection threshold of the gate, the floor on the energy gap that the threshold corresponds to, the common shift that the energy regularizer removes, and the gate's decision surface.

Threshold and floor. At evaluation the gate runs at the near-zero temperature $\tau_{\text{eval}} = 0.01$ and selects component j when its selection probability exceeds $\tau_{\text{thres}} = 0.9999994$. That value is the probability that a Gaussian variable lies within five standard deviations of its mean. Setting the Gumbel noise of Eq. 5 to zero and inverting the logistic function turns the threshold into a floor on the energy gap:

$$E_{j,0}(x) - E_{j,1}(x) \;>\; \tau_{\text{eval}} \log \frac{\tau_{\text{thres}}}{1 - \tau_{\text{thres}}} \;\approx\; 0.14. \tag{23}$$

Without noise, a component is selected when not selecting it costs more than 0.14 energy units above selecting it. The solver samples the Gumbel noise in every pass, at evaluation as in training. The rule it applies is that the gap plus the difference of two $\text{Gumbel}(0, 1)$ draws exceeds 0.14. That difference is logistic with unit scale, so a gap many energy units from the floor gives the same decision in almost every draw. In the thirty runs decoded in §K.6, at every retained checkpoint, 99% of the gaps lie more than ten energy units from the floor, and at least 99.76% of the decisions are the same in all twenty repeated decodes.

The common shift. The gate's decision depends on the two energies of a component through their gap alone. Adding one constant to $E_{j,0}$ and $E_{j,1}$ leaves Eq. 5 unchanged, which is the free common shift of §5.2. The energy regularizer $\lambda_{\text{se}}\|E\|_2^2$, with $\lambda_{\text{se}} = 10^{-10}$, removes the shift. At a fixed gap the regularizer is smallest when $E_{j,0}$ is plus half the gap and $E_{j,1}$ is minus half the gap.

Decision surface. Figure 3 plots the selection probability of one component against its two energies in panels (a) and (b), and against the energy gap and the temperature in panels (c) and (d). Panel (a) shows the expected selection probability under Gumbel noise at $\tau = 0.1$. It changes gradually across the diagonal $E_{j,0} = E_{j,1}$, as a logistic of about unit scale in the energy gap. Panel (b) is an illustration without noise at τ_{eval}. The probability there rises from 0.01 to 0.99 over about 0.09 in the gap, and the dashed line is the floor of Eq. 23, where the probability is within 6×10^{-7} of one. Panels (c) and (d) vary the temperature. Without noise (c), the width of the rise is proportional to τ. With Gumbel noise (d), the expected probability tends to $\varsigma(E_{j,0} - E_{j,1})$ as $\tau \to 0$, since the difference of two $\text{Gumbel}(0, 1)$ draws is logistic with unit scale. Under Gumbel noise the expected rise does not become narrower than that logistic, whatever the temperature. The margin of §5.4 is the dataset mean of the energy gap plotted here, signed to agree with the selection.

H REPRODUCIBILITY DETAILS

This appendix gives what is needed to reproduce the solver of §5. It covers the training configuration, the per-epoch evaluation with the checkpoint rule of §5.3, the termination rule, the training loop as pseudocode (Algorithm 1), and the decoding used in the reuse experiment. Table 1 collects the constants, which are the same in every experiment. Four settings are stated per experiment: the pool size N (1024 on the synthetic data and 128 on the real data), the lower edge of the R^2 bands, the epoch cap, and whether the profiles are constrained to be non-negative (they are on the synthetic data and on Carbs, and not on NIR).

H.1 TRAINING CONFIGURATION

Networks. The energy network, which computes the selection energies $E(x)$ of §5.1, has three linear layers with tanh activations, hidden width $4N$, and output of size $N \times 2$. The concentration network f_c has three linear layers with ReLU activations and hidden width $2N$, and its output passes through an absolute value. The pool is a learnable $N \times d$ matrix. Each row is normalized to unit ℓ_2 norm, after passing through an absolute value when the profiles are constrained to be non-negative. The weights of the energy network are initialized from $\mathcal{N}(0, 0.05^2)$, all other weights from $\mathcal{N}(0, 0.01^2)$, and the profiles from $|\mathcal{N}(0, 0.1^2)|$.

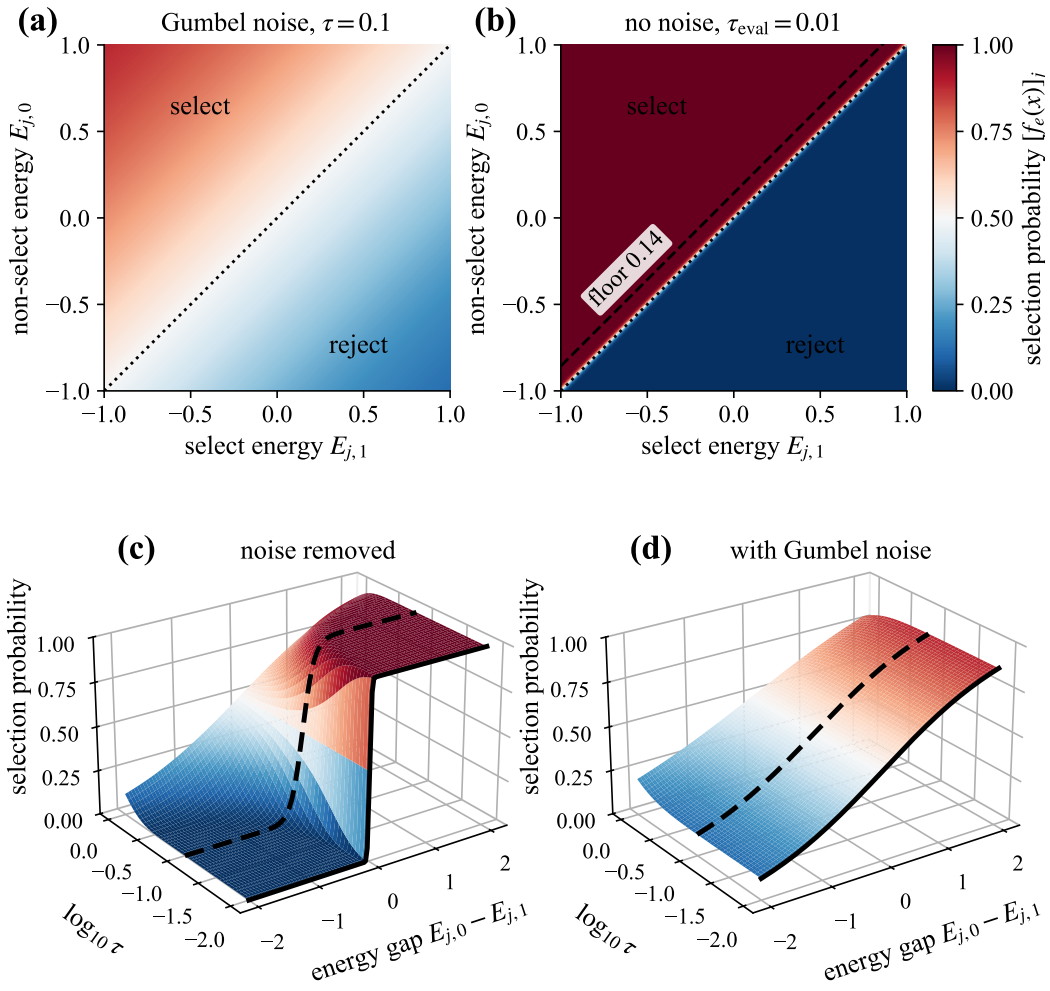


Figure 3: Selection probability of the gate for one component. (a) Expected probability at $\tau = 0.1$ under Gumbel noise, over the energy for selecting, $E_{j,1}$, and the energy for not selecting, $E_{j,0}$. (b) An illustration without noise at $\tau_{\text{eval}} = 0.01$. The solver samples the noise at evaluation. In (a) and (b) the dotted diagonal is where the probability is $1/2$, and in (b) the dashed line is the floor $E_{j,0} - E_{j,1} = 0.14$ of Eq. 23. (c, d) Probability over the energy gap and $\log_{10}\tau$, without (c) and with (d) Gumbel noise. The black curves are the cuts at $\tau_{\text{eval}} = 0.01$ (solid) and $\tau = 0.1$ (dashed).

Optimizer. AdamW (Loshchilov & Hutter, 2019) with learning rate 5×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.995$, weight decay 10^{-3}, and $\epsilon = 10^{-8}$. Minibatches hold 64 samples and are reshuffled every epoch; a synthetic run with fewer than 256 samples uses minibatches of a quarter of the sample count.

Temperature and gate output. Training starts at $\tau_0 = 1$. After every 100 epochs τ decreases by 0.01, down to $\tau_{\min} = 0.4$. Training uses the gate output of Eq. 5 as it is, a probability between zero and one, with no rounding to zero or one.

Exploration noise. The exploration noise of Eq. 6 is Gaussian with scale $s_{\text{ex}} = 0.1\sqrt{32} \approx 0.57$ per energy entry. The implementation draws it as the sum of 32 Gaussian draws of scale 0.1. It also

Table 1: EB-gMCR constants, the same in every experiment.

Parameter	Value	Role
Gate (Eq. 5; Appendix G)		
τ_0	1	Training temperature at the start
$\tau_{\min}$	0.4	Lowest training temperature
Temperature step	0.01 per 100 epochs	Decrease of τ during training
τ_{eval}	0.01	Temperature at evaluation
τ_{thres}	0.9999994	Selection threshold at evaluation, a probability (Appendix G)
s_{ex}	$0.1\sqrt{32} \approx 0.57$	Scale of the exploration noise (Eq. 6), training only
Loss (Eq. 7)		
λ	scheduled	Weight of the usage penalty; its target is zero while $\text{nMSE} \geq 0.005$
$(\beta_\lambda,\ \kappa)$	$(0.95,\ 1)$	Smoothing factor and target ratio of the λ schedule
λ_{se}	10^{-10}	Weight of the energy regularizer
λ_{amb}	10^{-10}	Weight of the ambiguity term (Eq. 8), RBF kernel of bandwidth 1
In-use level	0.5	Selection probability above which a component counts as in use for the ambiguity term
Optimization		
Optimizer	AdamW	Learning rate 5×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.995)$, weight decay 10^{-3}, $\epsilon = 10^{-8}$
Batch size	64	Minibatches reshuffled every epoch
Checkpointing and termination (§5.3)		
R^2 bands	width 0.005	Up to $R^2 = 1$; the lower edge is set per experiment
nMSE tolerance	0.01	Level of the termination rule
Window	100 epochs	Length of the rolling mean of nMSE in the termination rule
Energy ratio	0.25	Level of $\bar{E}_t/\bar{E}_{\text{init}}$ in the termination rule

subtracts one constant from both energies of every component, which leaves the gate output of Eq. 5 unchanged. The noise is applied in training only.

Loss. Eq. 7 is evaluated per minibatch B. The fit term is the mean of the squared entries of the minibatch minus its reconstruction. The usage term is the average of $C(x)$ over the minibatch. The energy regularizer is the mean of the squared energies, after the exploration noise is added, over the samples, the components, and the two energies of each component. Both terms are means where Eq. 7 writes sums, so the weights of Table 1 are weights of the means. The ambiguity term is the RBF kernel $K(\mathbf{s}_j, \mathbf{s}_k) = \exp(-\|\mathbf{s}_j - \mathbf{s}_k\|^2/2)$ averaged over the pairs of components in use, as in Eq. 8. In training a component counts as selected for a sample when its selection probability exceeds 0.5, and the components in use, the set $\mathcal{I}_B$ of Eq. 8, are those selected for at least one sample of the minibatch.

Schedule of the usage weight. The weight λ is updated once per epoch from the evaluation pass of Appendix H.2. It uses three quantities of that pass: the normalized error nMSE, the mean squared error MSE, and the mean selection energy $\bar{E}_t$, the average of all selection energies over the samples of the last training minibatch of the epoch, the components, and the two energies of each component. In an epoch with $\text{nMSE} \geq 0.005$ the target is zero. In an epoch with $\text{nMSE} < 0.005$ the target is $\lambda_{\text{target}} = \kappa\,\text{MSE}\,N/\bar{E}_t$ with $\kappa = 1$. The weight follows the target through an exponential moving average with smoothing factor $\beta_\lambda = 0.95$, $\lambda \leftarrow 0.95\,\lambda + 0.05\,\lambda_{\text{target}}$, starting from $\lambda = 0$. This is the schedule that §5.2 describes as setting λ from the energy scale.

H.2 Evaluation and Checkpointing

Evaluation mode. At the end of every epoch the model is run in evaluation mode over the training set. In evaluation mode the exploration noise is off, the gate runs at $\tau_{\text{eval}} = 0.01$, and component

j is selected when its selection probability exceeds τ_{thres}, that is, $\delta_j = \mathbf{1}[\,[f_e(x)]_j > \tau_{\text{thres}}\,]$. The Gumbel noise of Eq. 5 is still sampled (Appendix G). The reconstruction is $\hat{x} = \sum_j \delta_j\,[f_c(x)]_j\,\mathbf{s}_j$.

Metrics. With D the data matrix, $\hat{D}$ its reconstruction, and $\bar{D}$ the matrix of column means, $R^2 = 1 - \|D - \hat{D}\|_F^2 / \|D - \bar{D}\|_F^2$ and $\text{nMSE} = \|D - \hat{D}\|_F^2 / \|D\|_F^2$, the normalized reconstruction error of §5.2. MSE is the mean of the squared entries of $D - \hat{D}$. The estimated component count (EC) of §6 is the number of components selected for at least one sample.

Bands. The bands are consecutive intervals of width 0.005 in R^2, closed on the left: $[0.970, 0.975), \ldots, [0.995, 1.000)$ in the synthetic experiments, and starting from $[0.950, 0.955)$ in the real-data experiments. One checkpoint is kept per band.

Rule. When the R^2 of an epoch falls in a band that holds no checkpoint, the current model is stored. When the band holds a checkpoint, the current model replaces it if the current component count is smaller, or if the count is equal and the current R^2 is at least as high. All kept checkpoints are written out when the run ends. The model as it stands at that point, the *end state*, is written out as well.

What the rule gives a user. No computation certifies minimal usage on data whose composition is unknown (§4.2), so a run has no epoch at which the solver can declare the answer reached. The checkpoint rule is designed for that situation. At every epoch, each band holds the model with the smallest component count found so far at that reconstruction quality. A user who monitors a run can stop it at any epoch and still obtain, for every band the run has entered, the most parsimonious model found at that quality. The termination rule and the epoch cap are two ways of stopping a run, and neither changes what the bands hold. Choosing among the bands is the user's statement about the noise level (§5.3).

H.3 Termination

The rule. Let t_{in} be the first epoch with $\text{nMSE} < 0.01$, and let $\bar{E}_{\text{init}}$ be the average of the mean selection energy $\bar{E}_t$ over the epochs up to t_{in}. A run terminates at the first epoch t at which three conditions hold.

- *Entry.* The epoch t_{in} has occurred.
- *Oscillation.* The mean of nMSE over some window of 100 consecutive epochs, all after t_{in}, exceeded 0.01. Once this has been observed the condition stays satisfied. A run therefore continues for at least 100 epochs past t_{in}.
- *Energy.* $\bar{E}_t \le 0.25\,\bar{E}_{\text{init}}$.

The conditions in the synthetic runs. After a synthetic run first reaches $\text{nMSE} < 0.01$, its reconstruction error leaves and re-enters the tolerance repeatedly while the component count is still high. The oscillation condition is met when that has happened. In the synthetic runs the ratio $\bar{E}_t / \bar{E}_{\text{init}}$ is above 0.25 when the oscillation begins and later falls to about 0.05.

Epoch caps and how runs end. Every run also has an epoch cap: 10^5 epochs on the synthetic data, 5×10^5 on Carbs, and 3×10^6 on NIR. The synthetic runs end at the termination rule, except the ablation runs without the energy regularizer, which end at the cap (§K.5). On the real datasets the oscillation does not always occur, and the Carbs run ends at its cap. With 21 samples and a batch size of 64, one epoch on Carbs is one optimizer step.

H.4 Algorithm

Algorithm 1 summarizes the training procedure. The per-epoch steps are in the order of the implementation.

Algorithm 1 EB-gMCR: training and checkpointing

Require: Samples D; pool size N; R^2 bands; epoch cap; constants of Table 1

1: Initialize the pool, the energy network, and the concentration network f_c; $\lambda \leftarrow 0$; no checkpoint in any band
2: **repeat**
3: **Train one epoch:** for each minibatch, compute the gate output with the exploration noise (Eqs. 5 and 6), reconstruct (Eq. 2), and take an optimizer step on the loss of Eq. 7
4: **Evaluate:** decode D in evaluation mode; record R^2, nMSE, MSE, the estimated component count, and the mean selection energy $\bar{E}_t$
5: **Test termination** (Appendix H.3); the remaining steps of the epoch still run
6: **Update λ:** move λ toward its target, which is zero while $\mathrm{nMSE} \geq 0.005$ (Appendix H.1)
7: **Checkpoint:** if R^2 falls in a band, keep the current model for that band when the band is empty, when its component count is smaller than the stored one, or when the count is equal and R^2 is at least as high
8: **Lower** τ on its schedule
9: **until** the termination test has succeeded or the epoch cap is reached
10: **return** the checkpoint of every band that holds one, and the end state

H.5 Decoding and Reuse

A sample is decoded by running the frozen model in the evaluation mode of Appendix H.2. This is the decoding used in the per-epoch evaluation and in reuse (Corollary 1; §6.2). The reuse experiment of §6.2 proceeds as follows.

Choice of the checkpoint. Every kept checkpoint of the Carbs run is decoded on the 21 training mixtures. The frozen model is the checkpoint with the fewest components across the bands, with ties broken by the higher training R^2. That checkpoint is in the band $[0.960, 0.965)$ and has three components. The known component count is not used.

Test mixtures. Concentration triples are placed on the simplex $c_1 + c_2 + c_3 = 1$ with step 0.025, which gives 861 points. Each clean spectrum is $c^\top S$ with S the reference profiles. Noise is added to every spectral coordinate, drawn uniformly from $[0,\ 0.03\, I_{\max}]$, where $I_{\max}$ is the largest entry of CS, with C and S the reference concentrations and profiles of the 21 training mixtures. With 20 noise draws per point there are $17{,}220$ test samples.

Scoring. Each test sample is decoded as above. The per-sample nMSE is $\|x - \hat{x}\|^2 / \|x\|^2$, and the component count of a sample is the number of j with $\delta_j = 1$. The results are in Appendix L.

I Construction of the Synthetic Mixtures

Each synthetic mixture of §6.1 is a noisy sum of one to four profiles from a random non-negative pool (Algorithm 2). Component profiles are vectors in $\mathbb{R}^d$ with $d = 512$. The pool starts as a random $N \times d$ matrix with orthonormal rows, drawn by the orthogonal initialization of PyTorch. The element-wise absolute value is taken, and each row is normalized to unit ℓ_2 norm. The experiments use $N \in \{16, 32, 48, 64, 96, 128, 160, 192, 224, 256\}$ and $M = 4N$ or $8N$ samples. For every sample, a number n of components is drawn uniformly from $\{1, \ldots, 4\}$, and n components are then drawn uniformly at random, without replacement, from the pool. The cap of Eq. 2 is therefore $k = 4$. Each concentration is drawn uniformly from $[1, 10)$, so the concentrations lie in $[c_{\min}, c_{\max}]$ with $c_{\min} = 1$ and $c_{\max} = 10$. The noiseless sample is the sum of the selected profiles weighted by their concentrations. Zero-mean Gaussian noise is added to the nonzero entries of the sample, which in these data are all of its entries, at a signal-to-noise ratio (SNR) of 20 or 30 dB. The variance of the noise is set per sample. It is the mean of the squared nonzero entries divided by $10^{\mathrm{SNR}/10}$. The noisy sample is clamped from below at 10^{-6}, which keeps every entry positive.

Algorithm 2 Generation of the synthetic mixtures

Require: Pool $S \in \mathbb{R}^{N \times d}$ with non-negative rows of unit ℓ_2 norm, number of samples M, cap k, concentration range $[c_{\min}, c_{\max})$, SNR in dB

for $m = 1$ **to** M **do**
  Draw the number of components n_m uniformly from $\{1, \ldots, k\}$
  Draw the support $J_m \subset \{1, \ldots, N\}$ with $|J_m| = n_m$, uniformly at random without replacement
  for each $j \in J_m$ **do**
    Draw c_j uniformly from $[c_{\min}, c_{\max})$
  end for
  $x_m = \sum_{j \in J_m} c_j \, S_{j,:}$
  $\sigma_m^2 \leftarrow \big(\text{mean of } x_{m,l}^2 \text{ over the entries } l \text{ with } x_{m,l} \neq 0\big) / 10^{\mathrm{SNR}/10}$
  Add $\mathcal{N}(0, \sigma_m^2)$ noise to the nonzero entries of x_m
  $x_m \leftarrow \max(x_m,\, 10^{-6})$ element-wise
end for
return the data matrix $D \in \mathbb{R}^{M \times d}$ with rows x_m

J Count Search for the Baseline Solvers

Every baseline except ARD-NMF needs the component count as input. On the synthetic data the true component count is known, so each of those baselines is given a search over counts near it (Algorithm 3). The true component count K^* is the N of §6.1. Each combination of sample count and noise level has a reported band, the R^2 band at which results are reported. The reported bands are $[0.98, 0.985)$ for $4N$ samples at 20 dB, $[0.975, 0.98)$ for $8N$ samples at 20 dB, and $[0.99, 0.995)$ at 30 dB for both sample counts. The reported band is the R^2 band from which the EB-gMCR count of the same combination is read (§6.1, Table 4). The baselines and EB-gMCR are thereby compared at the same reconstruction quality. The search tries $K = \lfloor sK^* \rfloor$ for s from 1.20 down to 0.80 in steps of 0.05, and computes the R^2 of the baseline's reconstruction of D at each K. It returns the smallest K whose R^2 is inside the reported band, together with that R^2. When no K has its R^2 inside the band, the search returns the K with the highest R^2.

Algorithm 3 Count search for a baseline solver

Require: Data D, baseline solver, true component count K^*, ratios $\{1.20, 1.15, \ldots, 0.80\}$ in descending order, reported band $[r_{\text{lo}}, r_{\text{hi}})$

```
Initialize: R²_sel ← −∞, K_sel ← none, hit ← false
for each ratio s, in descending order do
    K_try ← ⌊s · K*⌋
    R² ← the R² of the reconstruction of D by the baseline solver with K_try components
    if r_lo ≤ R² < r_hi then
        hit ← true, R²_sel ← R², K_sel ← K_try {s descends, so the last hit is the smallest K in the band}
    else if not hit and R² > R²_sel then
        R²_sel ← R², K_sel ← K_try
    end if
end for
return K_sel and its R²_sel
```

K Synthetic-Data Evaluation

This appendix collects the results on the synthetic mixtures of Appendix I: the component counts against the baselines (§K.1), the counts in every band (§K.2), ARD-NMF on the same grid (§K.3), wall-clock times (§K.4), the ablation (§K.5), and the per-sample supports (§K.6). EB-gMCR is run five times per setting and the baselines one hundred times. The run counts follow the compute budget, since one EB-gMCR run at $N = 256$ takes two to four hours (Table 6). An entry written with $\pm$ is a mean $\pm$ one standard deviation over runs.

K.1 Component Count Against the Baselines

Tables 2 and 3 report, for each true component count N and noise level, the fit rate, the reconstruction R^2, and the estimated component count (EC) of every solver on $4N$ samples. These two tables hold the numbers behind Figure 1 and the count statements of §6.1.

Fit rate. The fit rate is the fraction of runs whose R^2 reaches the lower edge of the reported band, 0.98 at 20 dB or 0.99 at 30 dB. It measures the fit and not the count. ARD-NMF with ϕ at the true noise variance has a fit rate of one at 30 dB while returning 1.7 to 2.9 times the true count. The fit rate is not defined for EB-gMCR, whose count is read from the band, and its cells show a dash.

The baselines. The five baselines that take the count as input receive it through the search of Appendix J, so their EC is the count the search returned. ARD-NMF chooses its own count (§K.3). Its tuned setting was not run at 20 dB, and those cells show a dash.

The EB-gMCR cells. The EC of EB-gMCR is read from the reported band (Appendix J), $[0.98, 0.985)$ at 20 dB and $[0.99, 0.995)$ at 30 dB, and every other band is in Table 4. The EC of a cell is the mean over the runs that hold a checkpoint in the reported band. All five runs hold one except in seven cells of these tables: four runs at $N = 128$ to 224 and three at $N = 256$ at 20 dB, and four runs at $N = 192$ and 224 at 30 dB. The R^2 of a cell is the mean over all five runs. A run without a checkpoint in the reported band contributes the R^2 of the highest band it holds. The R^2 of some of those cells is below the lower edge of the reported band for that reason.

K.2 Counts in Every Band

Table 4 lists the EC of EB-gMCR in every band that holds checkpoints, for $4N$ and $8N$ samples at both noise levels, and Figure 4 plots the same numbers. Each entry is a mean over the runs that hold a checkpoint in that band.

The count rises with the band. A band of higher R^2 holds a checkpoint with more components. With $4N$ samples the EC rises with the band at every N. With $8N$ samples at 30 dB it rises with

Table 2: Comparison of EB-gMCR with six baseline solvers on $4N$ synthetic samples, $N = 16$ to 96. ARD-NMF (oracle ϕ) sets ϕ to the true noise variance of the draw, and ARD-NMF (tuned) uses $\phi = 0.03$ (§K.3). The fit rate counts the R^2 criterion only. Columns are defined in the text.

Components	Method	20dB (fit rate: $R^2 \geq 0.98$)			30dB (fit rate: $R^2 \geq 0.99$)		
		Fit	R^2	EC	Fit	R^2	EC
16	NMF	0.33	0.978±0.002	18.0±1.1	1.00	0.995±0.002	17.0±1.4
	Sparse-NMF	0.00	0.761±0.034	13.7±2.0	0.00	0.776±0.031	13.4±1.9
	Bayes-NMF	0.18	0.978±0.002	18.3±0.9	1.00	0.996±0.002	17.6±1.4
	ICA	0.00	0.106±0.172	14.7±1.9	0.00	0.079±0.199	13.8±1.4
	MCR-ALS	0.80	0.981±0.003	16.4±0.9	0.92	0.994±0.008	16.1±1.4
	EB-gMCR	-	0.983±0.001	16.2±1.1	-	0.993±0.002	15.4±0.5
	ARD-NMF (oracle ϕ)	0.94	0.983±0.002	18.4±1.2	1.00	0.999±0.000	45.9±4.2
	ARD-NMF (tuned ϕ=0.03)	–	–	–	1.00	0.998±0.000	16.1±0.3
32	NMF	0.01	0.976±0.002	36.4±1.6	1.00	0.992±0.001	32.7±1.2
	Sparse-NMF	0.00	0.580±0.034	25.5±1.8	0.00	0.596±0.034	25.4±1.5
	Bayes-NMF	0.08	0.978±0.002	37.3±1.3	1.00	0.993±0.002	33.1±2.1
	ICA	0.00	0.230±0.123	30.4±3.5	0.00	0.131±0.188	29.1±2.9
	MCR-ALS	0.12	0.956±0.048	29.5±2.8	0.52	0.986±0.015	30.3±2.1
	EB-gMCR	-	0.982±0.001	31.4±0.9	-	0.992±0.001	31.4±2.2
	ARD-NMF (oracle ϕ)	1.00	0.983±0.001	35.1±1.5	1.00	0.999±0.000	85.6±6.0
	ARD-NMF (tuned ϕ=0.03)	–	–	–	1.00	0.998±0.000	32.1±0.4
48	NMF	0.00	0.972±0.003	55.8±1.6	0.83	0.991±0.001	53.9±2.7
	Sparse-NMF	0.00	0.524±0.019	38.8±2.2	0.00	0.536±0.019	38.9±3.0
	Bayes-NMF	0.03	0.977±0.002	56.2±1.4	1.00	0.992±0.001	49.5±1.8
	ICA	0.00	0.200±0.141	47.8±5.9	0.00	0.126±0.180	43.8±3.6
	MCR-ALS	0.00	0.870±0.139	41.5±3.7	0.11	0.923±0.108	42.1±3.2
	EB-gMCR	-	0.981±0.001	48.0±1.2	-	0.991±0.001	50.0±4.2
	ARD-NMF (oracle ϕ)	1.00	0.983±0.001	50.6±1.3	1.00	0.999±0.000	119.5±7.2
	ARD-NMF (tuned ϕ=0.03)	–	–	–	1.00	0.998±0.000	48.0±0.3
64	NMF	0.00	0.967±0.003	74.7±2.0	0.09	0.987±0.002	74.4±2.3
	Sparse-NMF	0.00	0.509±0.019	51.0±0.0	0.00	0.518±0.018	51.1±0.6
	Bayes-NMF	0.00	0.976±0.002	74.9±1.8	1.00	0.992±0.001	67.9±2.3
	ICA	0.00	0.215±0.114	63.3±8.3	0.00	0.172±0.143	58.5±4.7
	MCR-ALS	0.00	0.770±0.226	54.2±4.1	0.00	0.876±0.183	56.5±4.9
	EB-gMCR	-	0.981±0.001	64.8±2.6	-	0.991±0.001	74.4±5.0
	ARD-NMF (oracle ϕ)	1.00	0.983±0.001	66.3±1.4	1.00	0.999±0.000	150.0±8.7
	ARD-NMF (tuned ϕ=0.03)	–	–	–	1.00	0.998±0.000	64.0±0.4
96	NMF	0.00	0.955±0.004	113.3±2.9	0.00	0.977±0.003	113.7±2.4
	Sparse-NMF	0.00	0.504±0.015	76.3±1.6	0.00	0.515±0.015	76.1±1.0
	Bayes-NMF	0.00	0.974±0.002	114.2±1.9	1.00	0.991±0.001	108.0±3.8
	ICA	0.00	0.245±0.077	96.0±12.0	0.00	0.194±0.100	89.5±7.8
	MCR-ALS	0.00	0.782±0.237	86.4±6.9	0.00	0.882±0.041	85.9±6.2
	EB-gMCR	-	0.981±0.001	103.8±12.0	-	0.990±0.000	102.4±4.8
	ARD-NMF (oracle ϕ)	1.00	0.983±0.001	97.3±1.3	1.00	0.999±0.000	201.2±8.6
	ARD-NMF (tuned ϕ=0.03)	–	–	–	1.00	0.998±0.000	95.9±0.3

the band at every N except $N = 32$, where four bands hold 31.2. With $8N$ samples at 20 dB it rises with the band at $N \geq 64$, and the two bands differ by at most 0.2 components at $N \leq 48$.

The reported band at 30 dB. The reported band at 30 dB is $[0.99, 0.995)$, the highest band in the table. With $8N$ samples its count is within 3.1 percent of the truth at every N. With $4N$ samples its count is above the truth by 4 to 16 percent at $N = 48$ to 160, and within four percent elsewhere.

The reported band at 20 dB. The reported band at 20 dB is $[0.98, 0.985)$ with $4N$ samples and $[0.975, 0.98)$ with $8N$. With $4N$ samples its count is within two percent of the truth at $N \leq 64$ and above the truth by 8 to 17 percent at $N \geq 96$. With $8N$ samples its count is above the truth by 12 to 26 percent at $N \leq 48$, within three percent at $N = 64$ and 96, and above the truth by 11 to 18 percent at $N \geq 128$. At $N \geq 96$ with $4N$ samples, a lower band holds a count nearer the truth than the reported band does.

Theorem 3(b). The excess at 20 dB does not test Theorem 3(b). The theorem applies to the near-minimizer of the penalized criterion under a known noise law, and the count read at a band is not that minimizer. The excess that doubling the samples leaves (§6.1) is a fact about the solver.

Table 3: Comparison of EB-gMCR with six baseline solvers on $4N$ synthetic samples, $N = 128$ to 256, with the columns of Table 2.

Components	Method	20dB (fit rate: $R^2 \geq 0.98$)			30dB (fit rate: $R^2 \geq 0.99$)		
		Fit	R^2	EC	Fit	R^2	EC
128	NMF	0.00	0.938±0.005	152.0±2.2	0.00	0.962±0.006	151.4±3.0
	Sparse-NMF	0.00	0.501±0.011	102.0±0.0	0.00	0.513±0.014	102.0±0.0
	Bayes-NMF	0.00	0.970±0.002	152.0±2.2	0.66	0.990±0.001	150.7±3.5
	ICA	0.00	0.275±0.065	129.3±13.7	0.00	0.235±0.085	120.0±9.9
	MCR-ALS	0.00	0.703±0.020	116.6±8.8	0.00	0.763±0.035	118.3±7.4
	EB-gMCR	-	0.980±0.002	147.8±14.9	-	0.991±0.000	137.6±3.8
	ARD-NMF (oracle ϕ)	1.00	0.984±0.001	128.7±0.9	1.00	0.999±0.000	248.6±9.9
	ARD-NMF (tuned ϕ=0.03)	–	–	–	1.00	0.998±0.000	127.8±0.4
160	NMF	0.00	0.923±0.005	191.2±2.7	0.00	0.945±0.005	190.8±2.9
	Sparse-NMF	0.00	0.500±0.012	128.0±0.0	0.00	0.513±0.011	128.0±0.0
	Bayes-NMF	0.00	0.967±0.002	191.4±2.2	0.18	0.988±0.002	191.3±2.3
	ICA	0.00	0.309±0.050	151.4±17.0	0.00	0.270±0.068	146.3±11.6
	MCR-ALS	0.00	0.445±0.055	140.0±10.9	0.00	0.480±0.080	142.6±10.9
	EB-gMCR	-	0.979±0.002	187.2±15.1	-	0.991±0.001	169.6±5.3
	ARD-NMF (oracle ϕ)	1.00	0.984±0.000	160.6±0.9	1.00	0.999±0.000	293.8±9.6
	ARD-NMF (tuned ϕ=0.03)	–	–	–	1.00	0.997±0.000	159.7±0.5
192	NMF	0.00	0.906±0.005	229.2±2.7	0.00	0.928±0.005	228.2±4.1
	Sparse-NMF	0.00	0.497±0.010	153.0±0.0	0.00	0.510±0.009	153.0±0.0
	Bayes-NMF	0.00	0.962±0.002	228.9±3.4	0.00	0.985±0.002	229.4±2.4
	ICA	0.00	0.316±0.048	171.4±17.2	0.00	0.285±0.054	170.7±12.7
	MCR-ALS	0.00	-0.029±0.117	161.3±9.8	0.00	-0.002±0.126	164.1±12.2
	EB-gMCR	-	0.979±0.002	213.5±11.6	-	0.989±0.002	191.8±4.6
	ARD-NMF (oracle ϕ)	1.00	0.984±0.000	192.0±1.0	1.00	0.999±0.000	340.9±11.6
	ARD-NMF (tuned ϕ=0.03)	–	–	–	1.00	0.997±0.000	191.5±0.7
224	NMF	0.00	0.893±0.004	266.9±3.3	0.00	0.914±0.004	267.2±2.8
	Sparse-NMF	0.00	0.497±0.009	268.0±0.0	0.00	0.510±0.009	268.0±0.0
	Bayes-NMF	0.00	0.957±0.002	267.7±1.9	0.00	0.981±0.002	267.2±2.8
	ICA	0.00	0.321±0.042	196.0±14.0	0.00	0.292±0.049	191.6±11.6
	MCR-ALS	0.00	-0.676±0.164	185.9±9.2	0.00	-0.617±0.150	186.4±10.2
	EB-gMCR	-	0.979±0.002	251.5±13.0	-	0.989±0.002	224.8±6.3
	ARD-NMF (oracle ϕ)	1.00	0.984±0.000	223.7±0.9	1.00	0.998±0.000	386.1±11.6
	ARD-NMF (tuned ϕ=0.03)	–	–	–	1.00	0.997±0.000	223.1±0.8
256	NMF	0.00	0.884±0.004	304.9±4.8	0.00	0.902±0.004	305.1±4.6
	Sparse-NMF	0.00	0.497±0.009	307.0±0.0	0.00	0.509±0.009	307.0±0.0
	Bayes-NMF	0.00	0.951±0.003	306.0±3.5	0.00	0.977±0.002	306.6±2.2
	ICA	0.00	0.341±0.031	212.8±9.0	0.00	0.307±0.036	214.3±9.2
	MCR-ALS	0.00	-1.473±0.242	212.2±11.4	0.00	-1.420±0.250	213.6±12.7
	EB-gMCR	-	0.978±0.003	286.7±6.4	-	0.990±0.000	248.8±1.1
	ARD-NMF (oracle ϕ)	1.00	0.984±0.000	255.4±1.0	1.00	0.998±0.000	435.2±12.1
	ARD-NMF (tuned ϕ=0.03)	–	–	–	1.00	0.997±0.000	254.6±1.1

Table 4: Estimated component count of EB-gMCR in every band, the numbers plotted in Figure 4. Columns are the true component count N. Entries are mean $\pm$ one standard deviation over the runs that hold a checkpoint in the band, for $4N$ and $8N$ samples at 20 and 30 dB.

R^2 band	16	32	48	64	96	128	160	192	224	256
Data size = 4N (SNR = 20dB)										
0.97–0.975	15.2±1.6	29.8±0.4	44.4±0.5	58.6±1.1	88.8±3.0	122.6±6.5	156.2±4.3	184.0±3.7	225.6±6.3	260.4±6.6
0.975–0.98	15.8±1.3	30.4±0.5	46.0±0.7	61.4±1.3	92.8±4.9	131.4±6.9	166.0±6.8	192.4±3.0	234.0±6.0	265.4±5.6
0.98–0.985	16.2±1.1	31.4±0.9	48.0±1.2	64.8±2.6	103.8±12.0	147.8±14.9	187.2±15.1	213.5±11.6	251.5±13.0	286.7±6.4
Data size = 4N (SNR = 30dB)										
0.97–0.975	13.6±0.5	27.0±0.7	39.8±0.4	52.4±1.1	79.6±2.1	110.2±1.9	138.6±4.3	170.0±2.0	200.4±3.8	234.4±4.4
0.975–0.98	14.2±0.4	27.6±0.5	40.4±0.5	53.8±0.8	82.0±2.4	113.2±0.8	141.2±3.0	173.2±1.8	202.8±2.5	237.2±2.6
0.98–0.985	14.4±0.5	28.2±0.8	41.8±0.4	55.2±1.1	85.2±2.8	116.0±1.9	145.2±2.2	176.2±0.8	207.0±2.0	239.2±1.9
0.985–0.99	14.8±0.4	29.2±0.8	43.2±0.4	57.2±1.1	88.4±1.9	120.8±1.1	149.2±1.6	180.0±1.0	211.2±1.5	242.8±1.3
0.99–0.995	15.4±0.5	31.4±2.2	50.0±4.2	74.4±5.0	102.4±4.8	137.6±3.8	169.6±5.3	191.8±4.6	224.8±6.3	248.8±1.1
Data size = 8N (SNR = 20dB)										
0.97–0.975	20.4±1.9	40.0±3.7	53.8±2.9	65.0±2.2	93.4±0.9	128.2±2.5	166.0±4.8	200.2±5.5	237.8±10.4	272.4±6.3
0.975–0.98	20.2±1.6	39.8±3.4	53.8±2.9	65.6±1.9	96.4±0.9	142.6±11.0	187.0±10.0	225.0±9.0	260.8±9.9	301.0±6.4
Data size = 8N (SNR = 30dB)										
0.97–0.975	14.4±0.5	31.2±0.8	44.4±1.7	57.0±0.7	85.8±0.8	114.8±1.8	147.2±1.9	178.4±1.7	210.4±3.0	241.6±2.3
0.975–0.98	14.8±0.4	31.2±0.8	44.8±1.3	58.2±0.8	89.2±2.5	117.4±1.5	150.0±1.9	181.6±1.1	213.2±3.6	245.0±1.2
0.98–0.985	15.0±0.0	31.2±0.8	45.4±0.9	59.2±0.8	90.6±2.2	120.6±1.1	153.0±1.9	185.0±1.0	216.4±2.3	248.0±1.0
0.985–0.99	15.4±0.5	31.4±0.5	46.0±0.7	60.6±0.5	92.2±1.6	124.0±1.6	156.0±1.2	187.2±1.3	220.0±1.9	251.6±0.5
0.99–0.995	16.0±0.0	32.4±0.5	47.4±0.5	66.0±2.5	96.2±0.8	128.0±1.6	159.6±1.1	191.8±2.4	223.0±0.8	255.0±1.4

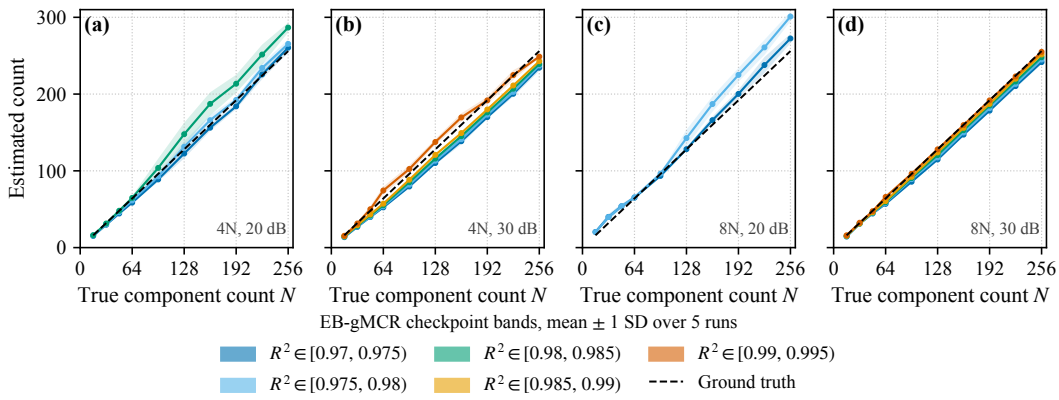


Figure 4: EB-gMCR on the synthetic mixtures: estimated against true component count in every R^2 band, for $4N$ and $8N$ samples at 20 and 30 dB. Mean $\pm$ one standard deviation over the runs that hold a checkpoint in the band.

K.3 ARD-NMF on the Synthetic Grid

The three settings. ARD-NMF is the one baseline that chooses its count. The count is set by the noise dispersion of §2, written ϕ. ARD-NMF is run from the same initial rank of 1024 as the EB-gMCR pool, on $4N$ samples, in three settings of ϕ. The default $\phi = 1$ returns no component for $N \leq 48$ and, on average, fewer than three at larger N. That setting is omitted from the tables. The oracle setting puts ϕ at the true noise variance of the draw, 0.027 at 20 dB and 0.003 at 30 dB. The tuned setting puts $\phi = 0.03$, chosen at 30 dB so that the returned count matches the truth. Figure 5 plots the returned count over the true count against N. Table 5 gives the ratio and the R^2 at four values of N. The rows of Tables 2 and 3 give every value.

Table 5: ARD-NMF on the synthetic grid, $4N$ samples, one hundred runs per cell. EC/N is the returned count over the true count, and R^2 is the mean reconstruction R^2. Oracle: ϕ at the true noise variance. Tuned: $\phi = 0.03$.

Noise	Setting	$N = 16$	$N = 64$	$N = 128$	$N = 256$
20 dB	oracle, EC/N	1.15	1.04	1.01	1.00
	oracle, R^2	0.983	0.983	0.984	0.984
30 dB	oracle, EC/N	2.87	2.34	1.94	1.70
	oracle, R^2	0.999	0.999	0.999	0.998
30 dB	tuned, EC/N	1.01	1.00	1.00	0.99
	tuned, R^2	0.998	0.998	0.998	0.997

The counts. At 20 dB, ARD-NMF with ϕ at the true noise variance returns the true count to within fifteen percent at $N = 16$, four percent at $N \geq 64$, and one percent at $N \geq 128$. At 30 dB the same setting returns 1.7 to 2.9 times the true count, while its R^2 is at or above 0.998 at every N. The ϕ that returns the true count at 30 dB is ten times the true noise variance of that draw and is near the true noise variance at 20 dB.

What the two settings show. Setting ϕ to the noise variance of the data returns the true count at one noise level and not at the other. Both settings at 30 dB have an R^2 of 0.997 or higher, so the fit of the output does not tell which of the two returned the true count.

Relation to the certificate. Both settings fit the data, and the usage condition of Theorem 2 has no counterpart in ARD-NMF. The R^2 band of EB-gMCR (§5.3) is the fit condition of the certificate, checked on every checkpoint. ARD-NMF on the real datasets, with ϕ swept, is in Appendix L.

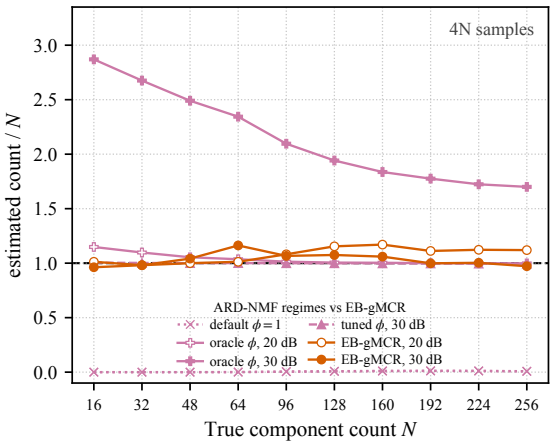


Figure 5: ARD-NMF on the synthetic grid: estimated over true component count against N on $4N$ samples, at the default ϕ, at ϕ equal to the true noise variance, and at the tuned ϕ, with EB-gMCR for reference.

K.4 Runtime

Table 6 gives wall-clock times at 20 dB and $N = 256$.

What is timed. One EB-gMCR run needs no count, and its time is the time of the run. A baseline needs a count, so it is timed on a scan of the counts $i, 2i, \ldots, 256$, where i is the interval of the table. The entry is the total time of the scan.

The times. One EB-gMCR run takes two to four hours. The three NMF baselines complete a scan in 0.01 to 0.68 hours, MCR-ALS in 0.39 to 4.28 hours, and ICA in 1.06 to 15.8 hours. EB-gMCR is slower than the NMF baselines at every interval. It is faster than ICA at intervals of 4 and 2 with $4N$ samples and at intervals of 16 and below with $8N$, and faster than MCR-ALS at an interval of 2. ARD-NMF is a single fit and not a scan, with a median of 74 s per fit at $N = 256$ with $4N$ samples. EB-gMCR uses one set of hyperparameters (Table 1) in every setting.

Hardware. The synthetic benchmarks and the runtime table ran on Ubuntu 24.04 LTS with an Intel i9-14900K CPU and NVIDIA RTX A6000 GPUs. The ARD-NMF timing used an Intel i9-12900K with NVIDIA RTX 3090 GPUs, so the ARD-NMF time is not comparable with the others. The Carbs runs of Appendix L used that second machine as well.

K.5 Ablation

Table 7 removes one term of the loss at a time, at $N = 256$, and measures what each term of §5 does. Each run is read at the highest band it holds, so the full-solver row at $4N$ and 20 dB averages all five runs and differs from the entry of Table 3, which averages the three runs that hold the reported band (§K.1).

Usage penalty. Without the usage penalty the count grows in all four settings, most at $8N$ and 30 dB, where it rises from 255 to 298. At 30 dB the runs without the penalty reach a higher band than the full solver, and the count is read there. No other term of the loss prefers a selection with fewer components.

Energy regularizer. Without the energy regularizer the termination rule is never met, and the run ends at the epoch cap of 10^5. The runtime rises from 3.6 to 6.1 hours with $4N$ samples at 20 dB, and from 2.0 to 12.1 hours with $8N$.

Exploration noise. Without the exploration noise the count is higher by 0.3 to 7.6 percent.

Table 6: Wall-clock time in hours at $N = 256$ and 20 dB. A baseline entry is the total time of a scan of the counts $i, 2i, \ldots, 256$ at interval i. The EB-gMCR entry is the time of one run, mean $\pm$ one standard deviation over five runs. Sp-NMF is Sparse-NMF and Ba-NMF is Bayes-NMF.

Interval	NMF	Sp-NMF	Ba-NMF	ICA	MCR-ALS	EB-gMCR
4N samples (1024)						
32	0.03	0.01	0.07	1.06	0.39	
16	0.06	0.02	0.14	1.75	0.75	
8	0.10	0.03	0.27	3.16	1.41	3.6±0.8
4	0.20	0.06	0.52	6.03	2.71	
2	0.27	0.10	0.68	10.1	4.28	
8N samples (2048)						
32	0.01	0.01	0.02	1.49	0.48	
16	0.01	0.02	0.03	2.71	0.89	
8	0.02	0.03	0.06	5.15	1.81	2.0±0.2
4	0.03	0.05	0.08	7.46	1.64	
2	0.06	0.10	0.17	15.8	3.77	

Ambiguity term. Without the ambiguity term the count changes by at most 5.1 percent, and the R^2 changes by at most 0.002. The usage penalty prefers one slot per component over several slots that each hold part of it. The ambiguity term acts on the remaining case, two slots with the same profile that are used by different samples, which the usage penalty does not distinguish from one slot.

Table 7: Ablation at $N = 256$: one term of the loss removed at a time. Each run is read at the highest band it holds, and an entry is the mean $\pm$ one standard deviation over the five runs. Time is in hours, and runs are capped at 10^5 epochs.

Solver	20dB			30dB		
	R^2	EC	Time	R^2	EC	Time
$4N$ *samples*						
EB-gMCR	0.978±0.003	278.0±13.4	3.6±0.8	0.990±0.000	248.8±1.1	3.3±0.6
w/o exploration noise	0.980±0.001	292.0±8.2	4.7±0.3	0.990±0.000	257.8±6.2	2.9±0.3
w/o usage penalty $\lambda C(x)$	0.980±0.000	284.6±5.0	4.1±0.3	0.995±0.000	268.2±4.5	4.1±0.3
w/o energy regularizer $\|E\|_2^2$	0.979±0.002	282.0±20.3	6.1±0.2	0.990±0.000	251.8±3.4	6.0±0.4
w/o ambiguity term $\mathcal{R}(B)$	0.980±0.000	292.2±13.0	3.8±0.6	0.990±0.000	249.0±1.6	4.5±0.7
$8N$ *samples*						
EB-gMCR	0.975±0.000	301.0±6.4	2.0±0.2	0.991±0.001	255.0±1.4	2.1±0.2
w/o exploration noise	0.976±0.002	323.8±41.2	2.5±0.3	0.991±0.000	255.8±0.4	2.0±0.3
w/o usage penalty $\lambda C(x)$	0.976±0.002	318.6±32.2	2.2±0.1	0.994±0.002	298.2±29.3	2.4±0.2
w/o energy regularizer $\|E\|_2^2$	0.975±0.000	290.4±11.9	12.1±0.1	0.991±0.000	255.6±0.5	12.2±0.1
w/o ambiguity term $\mathcal{R}(B)$	0.975±0.000	305.4±4.2	2.5±0.5	0.991±0.000	254.8±0.8	2.2±0.1

K.6 Per-Sample Support Recovery

The tables above report the component count, which is the size of the pool a run keeps. Theorem 2 says more than that. It is a statement about each sample's support: which components each sample is decomposed into. This subsection checks the supports sample by sample, which is possible here because the data are synthetic and the true supports are known. The short answer is that at most checkpoints the gate selects many more components than a sample contains, and the decoded concentrations tell the extra components apart from the true ones.

The runs. Both tables here come from thirty further runs made for the per-sample analysis, under the benchmark's configuration with the same generator, hyperparameters, bands, and termination rule. The analysis needs each learned slot matched to a true component, a step that is more reliable at small component counts, where a run is also short. The runs use $N = 16$, 32, and 64 at 20 and 30 dB, with $4N$ samples and five new seeds each, so their counts differ slightly from those of

Table 2. Every run saves the gate's energies, the per-sample selections, and twenty repeated decodes at each retained checkpoint (the repeated decodes are used in Appendix G), and all thirty runs are reported. Every table entry is a mean $\pm$ one standard deviation over the five seeds. Table 8 reports the checkpoint band of Table 2, and Table 9 reports every retained band.

The columns. A learned slot is matched to a true component by the Hungarian assignment on profile cosine, over the slots active in at least one sample. The columns are then as follows.

- *Usage ratio.* The total number of components the gate selects over all samples, divided by the true total. A ratio of one is usage-minimal. This column needs no matching.
- *Min cos and misses.* Min cos is the smallest cosine between a learned profile and its matched true profile. Misses is the number of true components left without a slot because fewer slots were active than components. A component matched to a poor slot is not counted as a miss. Min cos reports how poor the worst match is, and a low min cos marks a slot whose profile lies between two true profiles.
- *Exact and no-miss.* Exact is the fraction of samples whose gate support equals the true support after the matching. No-miss is the fraction whose gate support contains the true support.
- *Isolating pairs (ceiling).* The number of components that admit an isolating pair, two samples whose supports share exactly that component, evaluated on the gate's supports, out of the N true components. The number in parentheses, the ceiling, is the same count on the true supports.
- *Usage and exact at $\hat{c} > 0.5$.* Usage ratio and exact, recomputed after each gate support is restricted to the slots whose decoded concentration exceeds a floor of 0.5. The floor was fixed in advance, below the smallest true concentration of one, and applying it lowers no-miss by at most 0.012 in any setting. The paragraph on the certificate below says why this is the support the certificate evaluates.
- *Spurious $\hat{c}$* (Table 9). The median decoded concentration on selected slots outside the true support, that is, on the extra components.
- *Epoch and n* (Table 9). Epoch is the fraction of the run at which the checkpoint was taken, and n is the number of the five runs that retained the band.

The profiles are recovered at the smaller counts. At $N = 16$ at both noise levels and at $N = 32$ at 20 dB, min cos is 0.98 or higher and at most 0.2 components are missed on average. In the other three settings min cos is between 0.80 and 0.90, and up to 1.8 components are missed.

At the reported checkpoints the gate selects too many components. In four of the six settings of Table 8 the gate selects seven to eleven times the true usage, and exact is zero. Additions are far more frequent than omissions. The true support is inside the selection on every sample, or nearly every sample, at $N = 16$ and $N = 32$ at 20 dB, where no-miss is 1.00 and 0.99, and on 89% of the samples at $N = 32$ at 30 dB and $N = 64$ at 20 dB. In the two remaining settings, $N = 16$ and $N = 64$ at 30 dB, the reported checkpoint is an early one, taken at 0.19 and 0.12 of the run (Table 9). The usage ratio there is 1.8 and 1.1, and no-miss is lower, 0.93 and 0.74.

The concentrations separate the extra components. At the later checkpoints the extra components carry decoded concentrations of 0.01 to 0.09, against true concentrations of at least one (Table 9, last column). Applying the floor of 0.5 removes them. At the reported checkpoints the usage ratio is then 0.96 to 1.10 and the support is exact on 68 to 89% of the samples (Table 8, last two columns). At the early checkpoints the extra components are fewer and larger, with concentrations of 0.10 to 1.75, and the floor does not remove all of them. Across all bands the exact fraction under the floor ranges from 14 to 96%.

What the certificate counts as a support. The certificate does not read a support from the gate alone. Condition (C8) of the formal certificate (Appendix C) requires every used concentration to be at least a floor ϑ, so the support the certificate evaluates is the set of selected slots whose concentration is at least ϑ. The columns at $\hat{c} > 0.5$ are that reading with $\vartheta = 0.5$. Under it the returned solutions are close to usage-minimal, and what over-selects is the gate when it is read without the concentrations.

Table 8: Per-sample support recovery of EB-gMCR on the synthetic mixtures at the reported band (pool 1024, $4N$ samples, five seeds). Columns are defined in the text.

N	SNR	usage ratio	exact	no-miss	min cos	misses	isolating pairs (ceiling)	usage, $\hat{c} > 0.5$	exact, $\hat{c} > 0.5$
16	20	7.025±0.828	0.000±0.000	1.000±0.000	0.978±0.010	0.0±0.0	0.0±0.0 / 16 (5.4)	1.104±0.062	0.756±0.140
16	30	1.837±0.607	0.144±0.196	0.931±0.041	0.991±0.008	0.2±0.4	0.6±1.3 / 16 (5.4)	1.038±0.089	0.831±0.187
32	20	10.581±0.948	0.000±0.000	0.991±0.007	0.980±0.007	0.0±0.0	0.0±0.0 / 32 (8.6)	1.042±0.043	0.889±0.097
32	30	9.849±3.795	0.000±0.000	0.886±0.045	0.904±0.058	1.8±1.1	0.0±0.0 / 32 (8.6)	1.102±0.046	0.775±0.132
64	20	11.491±7.872	0.000±0.000	0.885±0.097	0.877±0.118	0.4±0.5	0.0±0.0 / 64 (19.4)	1.062±0.069	0.780±0.191
64	30	1.110±0.210	0.393±0.339	0.742±0.028	0.800±0.171	0.2±0.4	10.6±9.8 / 64 (19.4)	0.962±0.061	0.675±0.078

Usage is low early in a run and high late. Table 9 follows each run through its retained checkpoints. In these runs the highest R^2 bands are reached early. Checkpoints taken within the first fifth of a run have usage ratios of 1.1 to 2.4, so the gate is near usage-minimal there, and no-miss between 0.69 and 0.97, so some true components are still missing from the selections. Later checkpoints have usage ratios of 4 to 19, and their no-miss is 0.89 to 1.00 at 20 dB with $N \leq 32$ and 0.71 to 0.91 in the other settings. Read from the gate alone, no retained checkpoint has a usage ratio near one together with no-miss near one. Read under the floor, the later checkpoints are exact on 56 to 89% of the samples, and the best checkpoint in the table, $N = 16$ at 30 dB in the highest band, is exact on 96%.

Most components have no isolating pair in these data. The isolating-pairs column has two readings. On the gate's supports the count is zero wherever the gate over-selects, because a support with extra components cannot share exactly one component with another support. That part is the solver's. The ceiling is computed on the true supports and is a property of the sampled data. At $4N$ samples with random supports of one to four components, 27 to 34% of the components admit a pair. The condition of Conjecture R is therefore not met by these data for any solver, and the exact column describes these runs and does not test the conjecture.

The direction of the gate's error matches Theorem 1. Theorem 1 describes a usage penalty below its λ-window: components that fit noise are kept, and true components are not dropped. The later checkpoints show that direction, many additions and few omissions. The solver's schedule sets λ from the energy scale and not from the window, and the penalty is divided over the 1024 slots of the pool (§5.2).

What the theory does not describe. The theory says nothing about how the pool forms before the anchor of Proposition 2, and the change from low usage to high usage belongs to that part of a run. What is missing is a schedule that keeps the usage penalty inside the λ-window while the pool completes. Theorem 1 gives the location of the window in terms of the noise level. On synthetic data, the usage ratio of Table 8 tells whether such a schedule has succeeded.

L Real-Data Evaluation Details

This appendix gives the details behind §6.2. It describes the two datasets and how each method was run, defines the metrics of Table 10 and reads the table, and then covers the certificate on Carbs, the reuse experiment, and ARD-NMF on the two datasets.

Datasets. Carbs has $M = 21$ Raman spectra with $d = 1401$ coordinates. Each is a mixture of three carbohydrates, with reference profiles and reference concentrations. Every one of the 21 mixtures contains all three components, and the smallest concentration in any mixture is about two percent of that mixture's total. NIR has $M = 166$ spectra with $d = 235$ coordinates. They are first-derivative spectra and take negative values. The concentrations of two analytes are known and the component profiles are not, so the true component count is taken to be two.

How EB-gMCR was run. EB-gMCR starts from a pool of 128 candidate components, with the constants of Table 1. The R^2 bands have width 0.005 and begin at 0.95. The epoch cap is 5×10^5 on Carbs and 3×10^6 on NIR, and the Carbs run ends at its cap (Appendix H.3). On NIR the absolute value on the pool is switched off, so the profiles can take negative values. The absolute value on the concentration network is kept, so the concentrations remain non-negative as Eq. 2 requires.

Table 9: The runs of Table 8 decoded at every band that holds a checkpoint (n runs hold the band). Columns are defined in the text.

N	SNR	band	n	epoch	usage ratio	no-miss	min cos	exact	exact, $\hat{c} > 0.5$	spurious $\hat{c}$
16	20	0.97–0.975	5	0.53±0.27	6.9±0.6	0.97±0.04	0.94±0.07	0.00±0.00	0.73±0.17	0.06±0.03
16	20	0.975–0.98	5	0.52±0.27	6.8±0.8	0.98±0.04	0.96±0.03	0.00±0.00	0.79±0.10	0.04±0.02
16	20	0.98–0.985	5	0.47±0.19	7.0±0.8	1.00±0.00	0.98±0.01	0.00±0.00	0.76±0.14	0.03±0.01
16	20	0.985–0.99	5	0.91±0.10	8.3±0.6	1.00±0.00	0.97±0.02	0.00±0.00	0.74±0.11	0.01±0.00
16	20	0.99–0.995	4	0.08±0.02	2.0±0.3	0.77±0.07	0.93±0.03	0.05±0.05	0.14±0.03	1.33±0.95
16	30	0.97–0.975	5	0.77±0.24	5.6±0.3	0.79±0.04	0.96±0.02	0.00±0.00	0.61±0.12	0.05±0.04
16	30	0.975–0.98	5	0.46±0.18	4.3±1.9	0.82±0.07	0.97±0.01	0.00±0.00	0.70±0.12	0.06±0.04
16	30	0.98–0.985	5	0.64±0.31	4.6±1.7	0.87±0.04	0.95±0.07	0.00±0.01	0.66±0.15	0.07±0.07
16	30	0.985–0.99	5	0.41±0.31	3.7±2.4	0.91±0.05	0.98±0.01	0.17±0.33	0.78±0.12	0.09±0.07
16	30	0.99–0.995	5	0.19±0.08	1.8±0.6	0.93±0.04	0.99±0.01	0.14±0.20	0.83±0.19	0.10±0.05
16	30	0.995–1.0	5	0.20±0.04	2.1±0.5	0.97±0.03	1.00±0.00	0.03±0.07	0.96±0.04	0.10±0.03
32	20	0.97–0.975	5	0.45±0.15	9.6±1.7	0.91±0.04	0.91±0.09	0.00±0.00	0.78±0.08	0.03±0.01
32	20	0.975–0.98	5	0.58±0.19	10.6±0.7	0.96±0.01	0.93±0.09	0.00±0.00	0.84±0.12	0.02±0.01
32	20	0.98–0.985	5	0.55±0.17	10.6±0.9	0.99±0.01	0.98±0.01	0.00±0.00	0.89±0.10	0.02±0.01
32	20	0.985–0.99	5	0.48±0.40	9.6±5.9	0.89±0.15	0.97±0.01	0.00±0.00	0.56±0.20	0.05±0.06
32	30	0.97–0.975	5	0.55±0.22	9.0±1.3	0.74±0.03	0.94±0.02	0.00±0.00	0.60±0.08	0.04±0.02
32	30	0.975–0.98	5	0.53±0.18	9.8±1.2	0.78±0.02	0.93±0.03	0.00±0.00	0.63±0.10	0.03±0.01
32	30	0.98–0.985	5	0.65±0.27	10.6±0.9	0.82±0.04	0.93±0.02	0.00±0.00	0.69±0.08	0.02±0.01
32	30	0.985–0.99	5	0.74±0.23	11.3±0.6	0.88±0.03	0.88±0.12	0.00±0.00	0.69±0.14	0.02±0.00
32	30	0.99–0.995	5	0.69±0.29	9.8±3.8	0.89±0.04	0.90±0.06	0.00±0.00	0.78±0.13	0.04±0.04
32	30	0.995–1.0	4	0.14±0.04	2.4±0.5	0.78±0.07	0.99±0.01	0.00±0.00	0.62±0.18	0.16±0.06
64	20	0.97–0.975	5	0.78±0.21	17.4±4.4	0.86±0.01	0.88±0.08	0.00±0.00	0.79±0.01	0.01±0.00
64	20	0.975–0.98	5	0.65±0.22	15.7±6.2	0.90±0.02	0.86±0.11	0.00±0.00	0.84±0.04	0.01±0.01
64	20	0.98–0.985	5	0.41±0.32	11.5±7.9	0.89±0.10	0.88±0.12	0.00±0.00	0.78±0.19	0.03±0.02
64	20	0.985–0.99	3	0.36±0.48	11.4±13.3	0.83±0.14	0.96±0.02	0.00±0.00	0.57±0.20	0.06±0.05
64	30	0.97–0.975	5	0.54±0.23	12.3±4.6	0.71±0.03	0.91±0.01	0.00±0.00	0.56±0.03	0.02±0.01
64	30	0.975–0.98	5	0.71±0.26	15.2±5.6	0.76±0.02	0.89±0.03	0.00±0.00	0.63±0.03	0.02±0.01
64	30	0.98–0.985	5	0.65±0.13	15.9±6.5	0.79±0.01	0.92±0.03	0.00±0.00	0.73±0.02	0.02±0.01
64	30	0.985–0.99	5	0.79±0.28	19.0±8.7	0.83±0.02	0.92±0.04	0.00±0.00	0.74±0.06	0.02±0.03
64	30	0.99–0.995	5	0.12±0.07	1.1±0.2	0.74±0.03	0.80±0.17	0.39±0.34	0.68±0.08	0.65±0.64
64	30	0.995–1.0	1	0.09±0.00	1.1±0.0	0.69±0.00	0.97±0.00	0.45±0.00	0.50±0.00	1.75±0.00

How the baselines were run. Each baseline that takes the count as input is run at every count from 1 to 10, and Figure 2(a, b) plots its R^2 at each count. Scanning the count from low to high and reading the reconstruction at each count is common practice for choosing the count in MCR, although other criteria have been proposed (Smith et al., 2019). The NMF baselines require non-negative data. On NIR they are run through a sign-splitting transform: each spectrum is split into its positive part and its negative part, the two parts are stacked and factorized, and the two halves of each recovered profile are subtracted. ARD-NMF is run over a sweep of its noise dispersion ϕ.

The end state on NIR. On NIR no method reaches an R^2 above about 0.91 with two components, which is below the lowest band. Every checkpoint kept in a band therefore has three or more components. The two-component solution reported for EB-gMCR is the end state of the run (Appendix H.2).

Metrics. Table 10 scores every method at the true component count. EB-gMCR is scored at its three-component checkpoint on Carbs, in the band $[0.960, 0.965)$, and at its end state on NIR. $R^2(D)$ is the reconstruction R^2. The recovered components are first matched to the references, by the cosine of the profiles on Carbs and by the correlation of the concentrations on NIR, which has no reference profiles. Each matched profile and each matched concentration column is then rescaled by a positive least-squares factor. Cos.(S) is the mean cosine between matched profiles. RMSE(S) and RMSE(C) are the root-mean-square errors of the rescaled profiles and concentrations, in the units of the references. Higher is better for $R^2(D)$ and Cos.(S), and lower is better for the two errors. NIR has no profile columns for the reason given above.

What the table shows. On Carbs, EB-gMCR has the same $R^2(D)$ as MCR-ALS, 0.964, the highest in the table. Its profiles have the same cosine, 0.996, and its errors are close to those of MCR-ALS, 0.032 against 0.026 in the concentrations and 0.628 against 0.610 in the profiles. On NIR, EB-gMCR has the highest $R^2(D)$, 0.910. Its concentration error, 24.49, is second to that of ICA, whose $R^2(D)$ is negative. ARD-NMF is scored at the ϕ with the best $R^2(D)$ among the values that return the true count, namely $\phi = 0.016$ on Carbs and $\phi = 0.08$ on NIR. On Carbs it has the

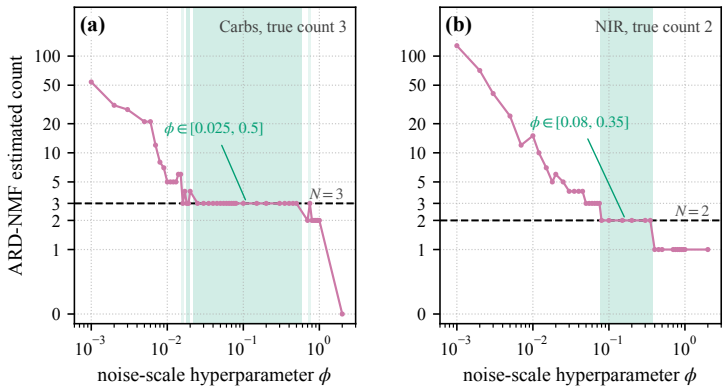


Figure 6: ARD-NMF on the real datasets: returned count against the noise dispersion ϕ on (a) Carbs, true count three, and (b) NIR, true count two. The shading marks the values of ϕ that return the true count.

lowest profile error, a concentration error equal to that of MCR-ALS, and an $R^2(D)$ of 0.951. On NIR its $R^2(D)$ is 0.902.

Notes on Figure 2. EB-gMCR is plotted at the counts that its checkpoints and its end state returned, and ARD-NMF at the ϕ with the best R^2 for each count. Series outside the axis range are omitted: all methods at counts of two or fewer on Carbs and at a count of one on NIR, ICA on Carbs, and Sparse-NMF, ICA, and MCR-ALS on NIR.

The certificate on Carbs. Two hypotheses of Theorem 2 are not available on Carbs. The first concerns the size of the pool. The pool has more slots than there are samples, $128 > M = 21$. A decomposition that gives every sample a slot of its own fits every sample, and its total usage is 21, below the usage of the truth, which is $21 \times 3 = 63$. The usage condition does not exclude that decomposition. Assumption (C7) of Appendix C does, by requiring that no candidate set represent a sample's span with fewer directions than the truth, and (C7) is not checked here. The second concerns the isolating pairs. Every Carbs mixture contains all three components, so no two supports share exactly one component, and the pinning step of Theorem 2 has no pair to use. Theorem 2 therefore does not certify the Carbs result. What supports the identity of each slot with a component is the match to the reference profiles, a cosine of 0.996 in Table 10, and the reuse experiment.

Reuse on the Carbs simplex. The protocol is in Appendix H.5, and Figure 7 plots the error of the frozen model against the noise floor at each of the 861 mixtures. Of the 861 grid points, 120 have at least one concentration equal to zero. The frozen model keeps three components on these mixtures as well, and the absent component decodes to a nonzero concentration. The ratio of that concentration to the smallest present concentration in the same sample has a median of 0.22 and a 90th percentile of 0.52, while the nMSE stays below 0.01. The per-sample support is read from the gate, which selects more components than these mixtures contain, as it does on the synthetic data (Appendix K.6). A second run under the same protocol, with the checkpoint chosen by the same rule, reproduces the result of Figure 2(c). An earlier run at the same training R^2 reached an nMSE below 0.01 on 75% of the simplex. The reuse away from the training mixtures varies between runs, and the certificate, which is evaluated on the training mixtures, does not distinguish those runs.

ARD-NMF on the real datasets. Figure 6 plots the count ARD-NMF returns against its noise dispersion ϕ. The values of ϕ that return the true count are about 0.025 to 0.5 on Carbs, with isolated values outside that range, and 0.08 to 0.35 on NIR. The ϕ of Table 10 on Carbs, 0.016, is one of the isolated values. Outside those ranges the output still reconstructs well and returns another count, and a single run does not tell whether its ϕ is inside the range. The same sweep on the synthetic mixtures is in Appendix K.3.

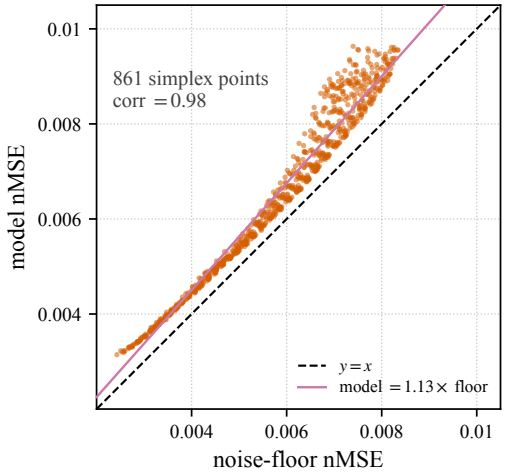


Figure 7: Frozen-model reuse on the Carbs simplex: model nMSE against the noise-floor nMSE at each of the 861 unseen mixtures, with the line $y = x$ and the fitted constant ratio.

Table 10: Results on the real datasets at the true component count. ARD-NMF is scored at the ϕ with the best $R^2(D)$ among the values that return the true count. Columns are defined in the text.

	Carbs (three components)				NIR (two components)	
Solver	$R^2(D)$	RMSE(C)	Cos.(S)	RMSE(S)	$R^2(D)$	RMSE(C)
NMF	0.963	0.070	0.992	0.929	0.876	32.46
Sparse-NMF	0.963	0.068	0.992	0.924	-131.4	42.32
Bayes-NMF	0.962	0.072	0.992	0.931	0.888	34.06
ICA	-6.206	0.338	0.462	6.564	-66.99	23.47
MCR-ALS	0.964	0.026	0.996	0.610	-25.17	27.63
ARD-NMF (tuned ϕ)	0.951	0.026	0.997	0.490	0.902	31.72
EB-gMCR	0.964	0.032	0.996	0.628	0.910	24.49