A Theoretical Framework For Overfitting In Energy-based Modeling

Giovanni Catania [1] **Aurélien Decelle** [1 2] **Cyril Furtlehner** [3] **Beatriz Seoane** [1]

Abstract

We investigate the impact of limited data on training pairwise energy-based models for inverse problems aimed at identifying interaction networks. Utilizing the Gaussian model as testbed, we dissect training trajectories across the eigenbasis of the coupling matrix, exploiting the independent evolution of eigenmodes and revealing that the learning timescales are tied to the spectral decomposition of the empirical covariance matrix. We see that optimal points for early stopping arise from the interplay between these timescales and the initial conditions of training. Moreover, we show that finite data corrections can be accurately modeled through asymptotic random matrix theory calculations and provide the counterpart of generalized cross-validation in the energy based model context. Our analytical framework extends to binary-variable maximum-entropy pairwise models with minimal variations. These findings offer strategies to control overfitting in discrete-variable models through empirical shrinkage corrections, improving the management of overfitting in energy-based generative models. Finally, we propose a generalization to arbitrary energy-based models by deriving the neural tangent kernel dynamics of the score function under the score-matching algorithm.

1. Introduction

Controlling overfitting is basic in machine learning, particularly as modern, over-parameterized architectures enhance learning capabilities. To prevent learning noise or irrelevant patterns, numerous empirical solutions have been proposed that modulate the model's implicit bias through its architecture and optimization, employing various implicit regularization mechanisms (Belkin, 2021). Finding the optimal balance between maximizing data utility, preserving generalization, and ensuring the privacy of training data represents a critical trade-off that can be challenging to pinpoint. In supervised learning tasks like classification, overfitting is readily identified using standard practices. Metrics like test-set accuracy, particularly when augmented by cross-validation in data-scarce scenarios, clearly signal overfitting, enabling strategies like early stopping, regularization, and hyperparameter tuning to mitigate it. Furthermore, training and generalization performance in regression and classification tasks are now well understood in certain simplified regimes, such as high-dimensional ridge (Atanasov et al., 2024; Advani et al., 2020; Saxe et al., 2014; Tomasini et al., 2022) or logistic (Mai et al., 2019; Loffredo et al., 2024) regression or numerous more complex setting of non-linear regression in various scaling regime (see for instance (Mei et al., 2018; Arnaboldi et al., 2023; Saad & Solla, 1995) among many other recent works). This gives the possibility to assess some simple indicator like the generalized cross-validation (GCV) (Golub et al., 1979), an exact relation between train/test errors valid for the ridge regression that can be derived using a leave-one-out argument (see e.g. (Furtlehner, 2023)). This methodology is also relevant in deep learning contexts (Wei et al., 2022), particularly in over-parameterized regimes where it aligns with observed stochastic gradient descent behaviors (Patil et al., 2024).

Recent advancements in training and architecture have greatly enhanced the generative capabilities of neural network models across various fields (Bengesi et al., 2024), enabling the creation of photorealistic images, credible speech synthesis, and biologically functional synthetic proteins (Wu et al., 2021). Despite this progress, selecting optimal models from a pool remains challenging due to noisy training data often leading to undetected overfitting. Yet, detecting overfitting in unsupervised learning settings, particularly for generative modeling, is elusive but crucial, especially with sensitive datasets like human genomic data (Yelmen et al., 2021; 2023) and copyrighted content. Unlike supervised learning, unsupervised learning lacks clear overfitting indicators, complicating model development and validation. While some theoretical insights on optimal regularization tuning for simple energy-based models exist (Fanthomme

[1]Departamento de Física Teórica, Universidad Complutense de Madrid, Spain. [2]Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid, Spain [3]Inria-Saclay, Université Paris-Saclay, LISN, Gif-sur-Yvette, France.. Correspondence to: Giovanni Catania <gcatania@ucm.es>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

et al., 2022), practical indicators such as early stopping points during training dynamics remain undefined. Moreover, estimating log-likelihood for model selection poses significant computational challenges (Béreux et al., 2022; 2024). Consequently, there is an urgent need for methods to detect and mitigate overfitting in these contexts.

This paper focuses on energy-based models (EBMs) (Ackley et al., 1985), which encode the empirical distributions of various data types —such as neural recordings (Roudi et al., 2009), images (Du & Mordatch, 2019), and genomic (Yelmen et al., 2021) or proteomic (Morcos et al., 2011) sequences—into a probability framework rooted in Boltzmann's law. By adopting a Bayesian approach, EBMs aim to maximize the likelihood function, enabling the generation of new data that closely resembles the training set and facilitates the extraction of detailed microscopic insights. EBMs range from simple Boltzmann Machines (BMs) and Restricted Boltzmann Machines (RBMs) to more complex architectures like convolutional neural networks, making them versatile in statistical physics for solving inverse problems like deducing Hamiltonian parameters from observed data. The interpretability of simple EBMs enables to uncover underlying rules within datasets: this capability has proven highly effective in fields ranging from neuroscience (Roudi et al., 2009) to bio-molecular structure prediction (Cocco et al., 2018) predominantly through pairwise maximum-entropy models. Recent advancements extend these applications, using complex EBMs to infer high-order interactions (Decelle et al., 2024; 2025; Feinauer et al., 2022; Feinauer & Lucibello, 2022) or constitutive patterns (Tubiana et al., 2019; Decelle et al., 2023), significantly deepening our comprehension of data structures.

This work develops a theoretical framework for understanding and mitigating overfitting in EBMs. We begin with a simple Gaussian model as a fundamental non-trivial example, using it to quantitatively analyze overfitting through synthetic experiments with predefined ground truths. We examine eigenvalue dynamics using artificial covariance matrices that simulate real datasets, exploring how overfitting arises from different learning timescales associated with various eigenmodes of the empirical covariance matrix. We address inaccuracies in learned eigenvalues with corrections based on random matrix theory (RMT), showing that the quality of model generation in EBMs is less affected by the lower modes of the covariance matrix, while the accuracy of inferred couplings is significantly impacted. We demonstrate that regularization techniques like shrinkage corrections are crucial to counteract overfitting, providing a robust framework to refine EBM training by considering finite-sample-size effects. This approach also informs our analysis of more complex models like the BM, underscoring the importance of regularization strategies to enhance model reliability and predictive accuracy.

2. Gaussian Model

The Gaussian Energy-Based Model (GEBM) specifies a multivariate Gaussian distribution for real-valued variables $\boldsymbol{x} \in \mathbb{R}^N$, characterized by 2-body interactions encoded within a symmetric, positive-definite coupling matrix, $\boldsymbol{J} \in \mathbb{R}^{N\times N}$. The GEBM is the simplest model that effectively captures the first and second-order statistics of a set of data. For the purposes of this analysis, we assume 0 means for the data components, thus simplifying the initial model by excluding the learning of external biases. Nevertheless, the theoretical framework presented below can be readily extended to the above to accommodate non-zero means. The probability distribution of a configuration $\boldsymbol{x}$ is then:

$$p\left(\boldsymbol{x}\,|\,\boldsymbol{J}\right)=(2\pi)^{-N/2}\sqrt{\det \boldsymbol{J}}e^{-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{J}\boldsymbol{x}}. \tag{1}$$

It is straightforward to check that the population covariance matrix of such distribution is $\boldsymbol{C}=\mathbb{E}_{\boldsymbol{J}}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]=\boldsymbol{J}^{-1}$, with $\mathbb{E}_{\boldsymbol{J}}\left[\cdot\right]$ denoting the average with respect to (1).

Inference problem. Consider a dataset $\mathcal{D}=\{\boldsymbol{x}^\mu\}_{\mu=1}^M$ with M entries generated with a GEBM model with coupling matrix $\boldsymbol{J}^*$. Our objective is then to find the parameters $\hat{\boldsymbol{J}}$ that best approximate the empirical distribution of the data —formally $p_{\mathcal{D}}(\boldsymbol{x})=M^{-1}\sum_{\mu=1}^M\delta\left(\boldsymbol{x}-\boldsymbol{x}^\mu\right)$—, with the probabilistic model (1). Without prior information about the model parameters $\boldsymbol{J}$, the maximum likelihood (ML) estimator $\hat{\boldsymbol{J}}^{\mathrm{ML},M}$ is calculated as $\hat{\boldsymbol{J}}^{\mathrm{ML},M}=\left(\widehat{\boldsymbol{C}}^M\right)^{-1}$, where $\widehat{\boldsymbol{C}}^M$ is the empirical covariance matrix from M data points, provided it is invertible (MacKay, 2003). Denoting with N the number of data components (data dimensions), this condition requires that $M\geq N$, assuming samples to be independent. Clearly, when $M\to\infty$, $\hat{\boldsymbol{J}}^{\mathrm{ML},M}$ recovers the true set of parameters used to generated the data, $\boldsymbol{J}^*$.

Training dynamics. The GEBM stands out as one of the few high-dimensional inference problems where an analytical expression for the ML estimator is available, independent of both M and N. However, our focus here is on the training dynamics associated with an iterative maximization of the likelihood function through gradient ascent dynamics, as is typical in EBMs. This approach allows us to explore the adaptive process of parameters' estimation over time.

In the GEBM, the log-likelihood (LL) of the parameters $\boldsymbol{J}$ depends only on $\widehat{\boldsymbol{C}}^M$ and it reads $\mathcal{L}(\boldsymbol{J})=-\frac{1}{2}\sum_{i,j}J_{ij}\widehat{C}_{ij}^M+\frac{1}{2}\log\det\boldsymbol{J}$. This quantifies how well $\boldsymbol{J}$ matches the observed data. In a standard gradient ascent algorithm, the update rule for the parameters reads:

$$J_{ij}^{t+1}=J_{ij}^t+\gamma\frac{\partial\mathcal{L}}{\partial J_{ij}}, \tag{2}$$

where γ is the learning rate. Assuming non-symmetric perturbations on the parameters J_{ij}, the gradient in (2) reads:

$$\frac{\partial\mathcal{L}}{\partial J_{ij}}=-\widehat{C}_{ij}^M+\mathbb{E}_{\boldsymbol{J}}\left[x_ix_j\right]=-\widehat{C}_{ij}^M+\left(\boldsymbol{J}^{-1}\right)_{ij}. \tag{3}$$

The second equality comes from the exact expression of 2-point correlations of the Gaussian model in terms of its coupling matrix $\boldsymbol{J}$: this is equivalent to assume that we perfectly sample the model with an infinite amount of configurations at any t. For a more generic EBM, another source of noise should be added due to the finite number of samples (or chains) used to estimate the empirical correlations used to compute the gradient.

To analyze the learning dynamics, we use the spectral decomposition of $\boldsymbol{J}$ and project the gradient onto its eigenbasis, denoted by $\boldsymbol{V} = \{\boldsymbol{v}_\alpha\}_{\alpha=1}^N$. From (3), the gradient projected on modes α and β is expressed as:

$$\left(\frac{\partial \mathcal{L}}{\partial \boldsymbol{J}}\right)_{\alpha\beta} = -\hat{c}^M_{\alpha\beta} + \frac{\delta_{\alpha\beta}}{J_\alpha}, \tag{4}$$

where $\hat{c}^M_{\alpha\beta}$ is the projection of $\widehat{\boldsymbol{C}}^M$. Generally, for any matrix $\boldsymbol{\mathcal{M}}$, we write $m_{\alpha\beta} \overset{\text{def}}{=} \boldsymbol{v}_\alpha^\top \boldsymbol{\mathcal{M}} \boldsymbol{v}_\beta$.

This approach enables us to formulate a set of evolution equations for the eigenvalues $\{J_\alpha\}$ and the rotation of the eigenvectors. By assuming an infinitesimal learning rate, we can transform the discrete-time update equation (2) into a continuous set of differential equations (see Appendix A for further derivation details):

$$\tau \frac{\mathrm{d}J_\alpha}{\mathrm{d}t} = \frac{1}{J_\alpha} - \hat{c}^M_{\alpha\alpha}; \qquad \tau \boldsymbol{v}^\alpha \frac{\mathrm{d}\boldsymbol{v}^\beta}{\mathrm{d}t} = \frac{\hat{c}^M_{\alpha\beta}}{J_\alpha - J_\beta} \text{ for } \alpha \neq \beta, \tag{5}$$

where τ is a timescale set by the learning rate, $\tau = 1/\gamma$. From Eq. (5) we see that eigenvectors of $\boldsymbol{J}$ stop rotating when they align with the eigenvectors of $\widehat{\boldsymbol{C}}^M$ to cancel out the numerator $\hat{c}^M_{\alpha\beta}$. Eq. (5) can be integrated analytically, and $\hat{c}^M_{\alpha\alpha}$ can be replaced by the matrix eigenvalue $\hat{c}^M_\alpha$.

The solution of (5) can be expressed in an explicit (although not closed) form using Lambert W_0 function, namely:

$$J_\alpha(t) = \frac{1}{\hat{c}^M_\alpha} + \frac{1}{\hat{c}^M_\alpha} W_0\left[\mathcal{B}_\alpha e^{-\left(\hat{c}^M_\alpha\right)^2 \frac{t}{\tau}}\right], \tag{6}$$

with the constant $\mathcal{B}_\alpha$ is fixed by the initial condition at $t = 0$. Eq. (6) delineates the evolution of each eigenvalue, which progresses independently once the eigenvectors of $\boldsymbol{J}$ align with those of $\widehat{\boldsymbol{C}}^M$. A crucial aspect of this equation is that the relaxation time it takes for an eigenvalue to reach its steady-state value $J_\alpha^{(\infty)} = \lim_{t\to\infty} J_\alpha(t) = 1/\hat{c}^M_\alpha$ is inversely proportional to the square of the corresponding eigenmode in the covariance matrix: indeed, for $t \to \infty$ Eq. (6) describes an exponential relaxation to the fixed point with a timescale $\propto (\hat{c}^M_\alpha)^{-2}$.

This relationship shows that the evolution of each eigenvalue is closely linked to the significance of the corresponding eigenvector in representing the data, so that stronger modes in the covariance matrix are learned more quickly than weaker ones. The idea that information is learnt progressively starting from strong PCA's directions is closely related to the concept of spectral bias (Rahaman et al., 2019) - although here the decomposition is spectral rather than in a Fourier basis - and it has been characterized theoretically in the case of linear regression (Advani et al., 2020). The interaction between these varying timescales can result in an initial phase where the strongest components of the dataset's PCA are effectively captured, followed by a phase where training begins to adjust noise-dominated directions, potentially leading to overfitting.

We've shown that GEBMs' learning dynamics are governed by the spectral decomposition of $\widehat{\boldsymbol{C}}^M$, with finite-sample effects arising from changes in the spectrum due to finite M. This falls within the realm of random matrix theory (RMT) (Potters & Bouchaud, 2020), as we detail shortly.

Asymptotic RMT analysis. In our simplified setting of GEBMs, the parameters of the model $\boldsymbol{J}(t)$ along the learning trajectory are an explicit function of the empirical covariance matrix $\widehat{\boldsymbol{C}}^M$ upon choosing the same constant $\mathcal{B}_\alpha = \mathcal{B} > -1/e$ in (6) for the initialization ($\mathcal{B} = -1/e$ corresponds to the initialization $\boldsymbol{J}(0) = 0$) and assuming that $\boldsymbol{J}(t)$ is aligned with $\widehat{\boldsymbol{C}}^M$ at $t = 0$. This choice simplifies considerably the analysis, and as explained in Appendix G. Using RMT, all relevant quantities can be derived in closed forms based solely on the *population spectrum* ν and the *aspect ratio* $\rho = M/N$, under the asymptotic proportional scaling where $M, N \to \infty$ with ρ held constant.

We are interested in the train and test energies:

$$E_{\text{train}} = N^{-1}\text{Tr}[\boldsymbol{J}\widehat{\boldsymbol{C}}^M] \text{ and } E_{\text{test}} = N^{-1}\text{Tr}[\boldsymbol{J}\boldsymbol{C}^*],$$

the coupling error $\mathcal{E}_{\text{J}} \overset{\text{def}}{=} N^{-1}\|\boldsymbol{J} - \boldsymbol{J}^*\|_F^2$, with $\|\cdot\|_F^2$, the Frobenius norm, and the LL (train and test)

$$LL_{\text{train,test}} \overset{\text{def}}{=} \frac{1}{2N}\log\det[\boldsymbol{J}] - \frac{1}{2}E_{\text{train,test}} \tag{7}$$

where $\boldsymbol{C}^* \overset{\text{def}}{=} \lim_{M\to\infty} \widehat{\boldsymbol{C}}^M$ is the population matrix. In addition, we will also explore the behavior of the maximizer of the log-likelihood with regularization, i.e. $\mathcal{L}[\boldsymbol{J}] = LL_{\text{train}}[\boldsymbol{J}] - \lambda A(\boldsymbol{J})$ focusing on $A(\boldsymbol{J}) = \text{Tr}[\boldsymbol{J}^2]$ for L_2 ridge regularization, and $A(\boldsymbol{J}) = \text{Tr}[\boldsymbol{J}]$ for $\widetilde{L}_1$ lasso regularization on the spectrum. $\widetilde{L}_1$ is applicable since $\boldsymbol{J}$ is symmetric and remains positive definite throughout the trajectory, ensured by the logarithmic barrier.

We simply quote here the result of the asymptotic limits (for $\rho > 1$, more details in Appendix G), based on RMT (Marčenko & Pastur, 1967; Ledoit & Péché, 2011). First, the spectral density $\bar{\nu}$ of $\widehat{\boldsymbol{C}}^M$ reads in this limit:

$$\bar{\nu}(x) = \frac{\rho\Lambda_i(x)}{\pi x} = \frac{\rho}{\pi x}\frac{\Gamma_i(x)}{\left[1-\Gamma_r(x)\right]^2 + \Gamma_i(x)^2}, \tag{8}$$

where $\Lambda(z) = \Lambda_r(x) + i\Lambda_i(x)$ and $\Gamma(z) = \Gamma_r(x) \pm i\Gamma_i(x)$ for $z = x + i0^+$, obey the self-consistent equations

$$\Lambda(z) = \frac{1}{1-\Gamma(z)\tau}, \qquad \Gamma(z) = \frac{1}{\rho}\int \frac{\nu(dx)x}{z - \Lambda(z)x},$$

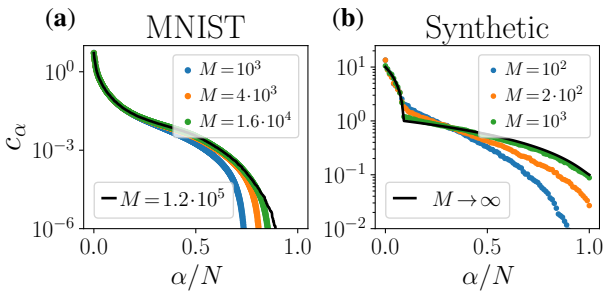

Figure 1. (**a**): Eigenvalue spectra of the empirical covariance matrices for MNIST dataset (Deng, 2012). Black lines show spectra using the full dataset size (M^*), while scatter colored points represent subsets ($M < M^*$). (**b**): Black line shows a synthetic population eigenvalue spectrum based on (10) for $N = 100$, $r = 0.9$, $\beta = 0.9$, $\gamma = 1.1$, $x_1 = 10^{-1}$, $x_2 = 10$; colored points show the eigenvalues from $\widehat{\boldsymbol{C}}^M$ calculated by sampling different M configurations from a GEBM model with $\boldsymbol{J}^* = \boldsymbol{C}^{*-1}$ (Eq. (1)).

in terms of the population spectral density $\nu(dx)$. In turn we obtain

$$E_{\text{train}} = \frac{\rho}{\pi}\int_0^\infty dy j(y)\left[\Lambda_r(y)\Gamma_i(y) + \Lambda_i(y)\Gamma_r(y)\right],$$

$$E_{\text{test}} = \frac{\rho}{\pi}\int_0^\infty dy j(y)\Gamma_i(y), \qquad (\rho \geq 1),$$

while the coupling error takes the form

$$\mathcal{E}_{\text{J}} = \int_0^\infty \frac{\nu(dx)}{x^2} + \int_0^\infty \bar{\nu}(dx) j(x)\left[j(x) - \frac{2}{\rho}\frac{(1-\rho)+2\rho\Lambda_r(x)}{x}\right]$$

where $j(x)$ is one of the analytical functions j_t, j_{L_1} and j_{L_2} corresponding respectively to the time dependent, $\widetilde{L}_1$ and L_2 regularized forms of $\boldsymbol{J}$. Remarkably, for the $\widetilde{L}_1$ regularized coupling matrix we get a deterministic relation between the train and test energies (Appendix G)

$$E_{\text{test}} = \left(1 - \rho^{-1}E_{\text{train}}\right)^{-1} E_{\text{train}}, \tag{9}$$

which is the counterpart of GCV for GEBMs which might be usable in practice for arbitrary EBM (in the same way as GCV can be used for deep regression models), as it allows one to get an estimation of the test LL. This concept remains a topic for future research. Instead, we have focused on strategies for data cleaning and regularization, specifically employing shrinkage techniques (Bun et al., 2017). Using a model of the data defined in the next section, allows us to specify $\nu(x)$ in order to assess these strategies by comparing with the expected optimal performances given by RMT.

3. Modeling realistic data covariances

To effectively study the impact of finite number of data on the learning process of a GEBM in a controlled setting, we need to define a synthetic model that facilitates the analysis of different learning timescales. The first step is to artificially create a *population* covariance matrix $\boldsymbol{C}^*$, from which a ground truth coupling matrix $\boldsymbol{J}^*$ is constructed, through $\boldsymbol{J}^* = \boldsymbol{C}^{*-1}$. Using this setup, we generate a multivariate Gaussian distribution and extract M data points from it. These data points are then used to train a new GEBM using the *empirical* covariance matrix, $\widehat{\boldsymbol{C}}^M$, derived from the M of these samples, with the goal of inferring the original model parameters.

As previously discussed, the training dynamics of each mode of $\boldsymbol{J}$ are directly linked to the eigenvalues of $\widehat{\boldsymbol{C}}^M$. To enhance this analysis, we have developed a synthetic model for the spectrum of $\boldsymbol{C}^*$, which influences the spectrum of $\widehat{\boldsymbol{C}}^M$ in scenarios with finite datasets. This model closely mimics the eigenvalue spectra of real datasets, as illustrated in Figure 1-(**a**), which shows the eigenvalue spectrum (in descending order) of covariance matrices from MNIST for several sizes M (more examples are given in Appendix B). Our analysis reveals that the spectrum of $\widehat{\boldsymbol{C}}^M$ remains relatively stable w.r.t. M for a significant number of modes, indicating $\hat{c}_\alpha^M \approx c_\alpha^\infty = c_\alpha^*$. However, smaller eigenvalues fluctuate markedly with M; they tend to be underestimated as M decreases, suggesting $\hat{c}_\alpha^M < c_\alpha^*$ for small c_α^*, and are slightly overestimated for the larger eigenvalues. This behavior is rigorously characterized using RMT tools in simplified data models (Baik & Silverstein, 2006; Ledoit & Péché, 2011). Additional insights into the conservation of eigenvectors across modes are detailed in Appendix B.

Inspired by these findings, we will characterize our synthetic population matrix $\boldsymbol{C}^*$ by an eigenvalue spectrum $\{c_\alpha^*\}_{\alpha=1}^N$ generated according to a mixture of power laws. The cumulative distribution is defined as follows:

$$P[\lambda < x] = r\left[\frac{x-x_1}{1-x_1}\right]^\beta \mathbb{1}_x^{(x_1,1)} + \left[r + (1-r)\left(\frac{x-1}{x_2-1}\right)^\gamma\right]\mathbb{1}_x^{(1,x_2)} \tag{10}$$

where $\mathbb{1}_x^{(a,b)}$ denotes the indicator function in the interval (a, b). This setup distinguishes between "strong" modes with $c_\alpha^* > 1$ and "weak" modes with $c_\alpha^* < 1$, with their prevalence controlled by parameter r. Parameters β and γ represent the power-law exponents for these two categories, with x_1 and x_2 the respective lower and upper cutoffs. Model (10) is chosen to $i)$ mimic the eigenvalue distribution of a realistic dataset's covariance matrix, though our numerical results are robust to specific spectral details, and $ii)$ to extract asymptotic quantities in the continuous-density limit $N, M \to \infty$ (with $\rho = M/N$ finite) using RMT.

Figure 1-(**b**) shows an example spectrum of population eigenvalues c_α^* from Eq. (10) with $N = 100$ (black line), alongside empirical estimates of finite-data eigenvalues $\hat{c}_\alpha^M$ (scatter points) for various M values, demonstrating that strong modes remain stable despite finite-M noise, while weak modes are consistently underestimated. The full matrix $\boldsymbol{C}^*$ is finally assembled by projecting the diagonal matrix of these eigenvalues onto a random orthogonal matrix $\boldsymbol{U}^* = \{\boldsymbol{u}_\alpha^*\}_{\alpha=1}^N$, resulting in $\boldsymbol{C}^* = \sum_\alpha c_\alpha^* \boldsymbol{u}_\alpha^* \boldsymbol{u}_\alpha^{*\top}$.

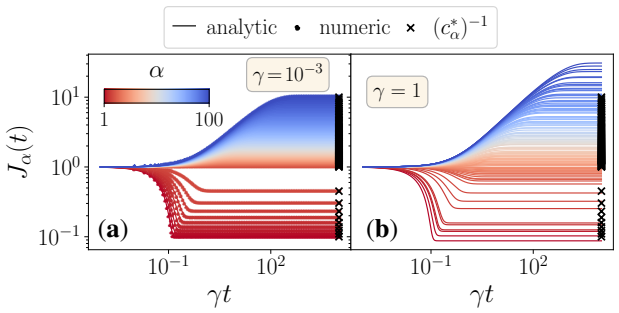

Figure 2. Training dynamics of the GEBM from a population matrix $\boldsymbol{C}^*$ (in (**a**), with system size and parameters matching those in Fig. 1-(**b**)), and from an empirical covariance matrix $\widehat{\boldsymbol{C}}^M$ (generated from $\boldsymbol{C}^*$ through (2), with $\rho = 2.11$ (in (**b**)). (**a**)-(**b**) display the analytic evolution of eigenvalues J_α toward the steady-state (lines), and a comparison with numerical training (points, in (**a**)). In all cases the initial condition is an identity matrix.

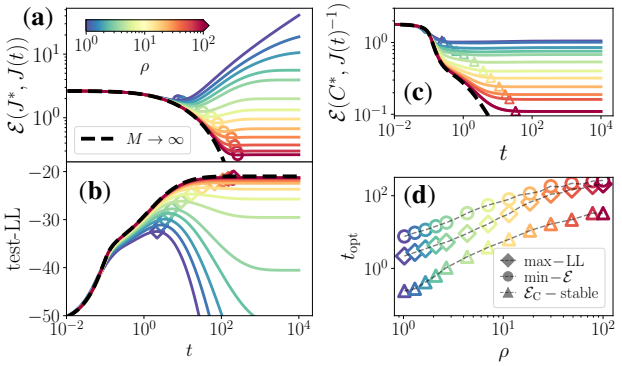

Figure 3. Results for GEBM training with finite data. (**a**)-(**b**)-(**c**) display respectively the reconstruction error $\mathcal{E}_{\rm J}$, the test-LL and the generation error $\mathcal{E}_{\rm C}$, all plotted vs time, for various sample sizes M (indicated by a color gradient from blue to red for increasing $\rho = M/N$. Dashed black lines refer to a training from $\boldsymbol{C}^*$ (i.e $M \to \infty$). (**d**): comparison between time of minimum reconstructon (circles), maximum test LL (diamonds) and time at which the generation error converges to its steady-state value. These quantities are also shown in the related panels for better clarity.

4. Training Dynamics on Synthetic Data

The introduced synthetic model enables analysis of training dynamics in two scenarios: i) an ideal setting with an infinite amount of samples using $\boldsymbol{C}^*$ as the data covariance matrix, and ii) a more realistic situation with a finite dataset, represented by the empirical covariance matrix $\widehat{\boldsymbol{C}}^M$.

Training Dynamics with Infinite Data. For clarity, we begin by training our GEBM using the population covariance matrix $\boldsymbol{C}^*$. Fig. 2-(**a**) illustrates the evolution of the coupling matrix eigenvalues J_α, comparing analytical solutions from Eq. (6) and numerical iterative training using Eq. (2), both starting from the same initial condition ($J_\alpha(0) = 1$). The analytical and numerical results match perfectly, demonstrating the expected time-scale separation: eigenvalues J_α corresponding to stronger covariance modes converge faster to their fixed point $J_\alpha^{(\infty)} = J_\alpha^* = 1/c_\alpha^*$, which are the smallest in the coupling matrix, while weaker modes converge slower. Starting from an initial condition $\boldsymbol{J}(0)$ that does not commute with $\boldsymbol{C}^*$, the coupling matrix must initially align its eigenvectors with those of $\boldsymbol{C}^*$, a process detailed in Appendix C and guided by Eq. (5). Following this alignment, the eigenvalues evolve independently according to Eq. (12), supporting our analytical approach. Notably, training directly from the population matrix achieves perfect reconstruction of the original model, thereby avoiding any discrepancies or generation errors as expected.

Impact of Finite Datasets: Interplay of Initialization and Time Scales Favoring Early Stopping Strategies. We explore the training dynamics using finite-data estimates of the population covariance matrix, $\widehat{\boldsymbol{C}}^M$. With any finite M, the GEBM trained with $\widehat{\boldsymbol{C}}^M$ will show discrepancies from the true model $\boldsymbol{J}^*$. We track these discrepancies by computing the reconstruction error $\mathcal{E}_{\rm J}$ between $\boldsymbol{J}^*$ and the trained $\boldsymbol{J}(t)$ defined in the previous section. Fig 3-(**a**) illustrates the error's evolution over training time. Beginning from an identity matrix, at low $\rho = M/N$ values, the error displays marked non-monotonic behavior, peaking at a specific $t_{\rm min}(\rho)$ before stabilizing at the training's fixed point. At higher ρ, the error decreases monotonically until stabilization, following a trend consistent with the $M \to \infty$ scenario (i.e. using $\boldsymbol{C}^*$, in black dashed line). This behavior, also noted in complex EBMs (Decelle et al., 2024; Agoritsas et al., 2023), underscores the GEBM's utility as a simple model yet capturing complex phenomena in EBMs.

This analysis shows that with limited data, there is an optimal training duration beyond which model inference accuracy declines, highlighting a sweet point for early stopping. However, detecting this point without ground truth is challenging: it does not coincide with the peak of test LL (as in (**b**)), a phenomenon also noted in RBMs (Decelle et al., 2024). Moreover, the generation's quality, given error between $\boldsymbol{C}(t) = \boldsymbol{J}(t)^{-1}$ and the population matrix $\boldsymbol{C}^*$ in (**c**) (computed as $\mathcal{E}_{\rm C} \stackrel{\rm def}{=} \|\boldsymbol{C}^* - \boldsymbol{C}(t)\|_{\rm F}$), stabilizes well before $t_{\rm min}(\rho)$ and remains flat afterwards. This suggests that the generation quality of the GEBM isn't solely dependent on the model itself, as evidenced by consistent generation errors at both the minimum-error point $t_{\rm min}$ and the training's fixed point, indicating this metric fails to capture the deterioration of model parameters over time. These optimal times are shown against ρ in Fig 3-(**d**). Additional evaluation metrics are discussed in Appendix F.

Fig. 2-(**b**) illustrates the evolution of eigenvalue J_α over time for a sample size with $\rho = 2.11$. Initially, the stronger

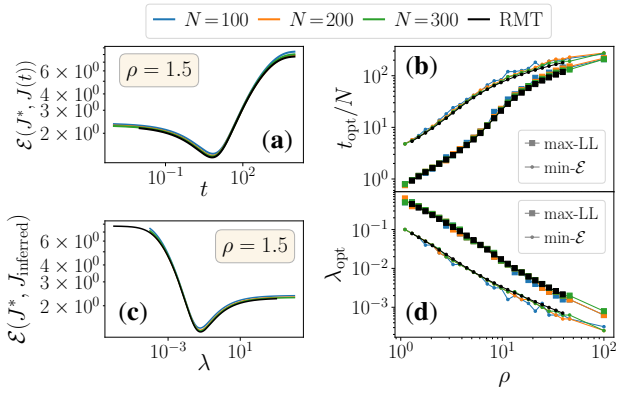

Figure 4. (**a**)-(**c**): for $\rho = M/N = 1.5$ we plot the reconstruction error during training (in (**a**), vs t) and the final reconstruction obtained using a L_2-norm regularization (in (**c**), vs λ). (**b**): training time achieving optimal reconstruction error (points) and time of maximum test LL (squares), plotted vs ρ. (**d**): optimal value of regularization prior λ vs ρ, again selecting the optimum w.r.t. reconstruction error and w.r.t. the test LL. All panels show comparisons between numerical results for various N (colored lines) against asymptotic results from RMT (black line).

modes (deep red curves), which are less influenced by low-M induced noise, quickly stabilize, aligning their eigenvalues $\hat{c}_\alpha^M$ close to the population values c_α^*, marked by black crosses. This early alignment to very small and relatively accurate values makes the error curve ($\mathcal{E}_J$) for finite M closely resemble that of the $M \to \infty$ scenario. However, after a period around $t_{\min}(\rho)$, the model starts encoding the weaker modes (blue lines), which are systematically understated relative to the population, causing $J_\alpha^{(\infty)} = 1/\hat{c}_\alpha^M$ to significantly exceed the ground-truth $J_\alpha^* = 1/c_\alpha^*$. If training begins from small J_α values, there's a critical point where the eigenvalues temporarily align more closely with their ground-truth than at the fixed point, effectively creating an optimal time t_α^* where $J_\alpha(t_\alpha^*) \approx J_\alpha^*$. This alignment markedly decreases discrepancies between the trained model's eigenvalues and those of the true model, highlighting the significance of initial conditions in training dynamics. Yet, the specific initial values of $J_\alpha(0)$ are less critical, as long as they are substantially smaller than $1/\hat{c}_\alpha^M$ for the weaker modes of $\widehat{\boldsymbol{C}}^M$ (see Appendix D for further details).

Now, we can also explain the stable generation performance of the GEBM, shown in Fig. 3-(**c**), using scale separation arguments. Generation error mainly depends on the strongest $\hat{c}_\alpha^M$ values, which are learned early on, whereas the overall model quality is controlled by the weakest $\hat{c}_\alpha^M$ (where $J_\alpha^{(\infty)} = 1/\hat{c}_\alpha^M$), which minimally affects generation error due to their small value. While this phenomenon appears unique to the GEBM, a similar effect is observed in binary pairwise EBMs (cf. Sec. 6).

Asymptotic analysis. Our findings so far have been established by numerically integrating the gradient ascent dynamics (i.e. using Eq. (2) with a slow learning rate), or with the analytical expression for the eigenvalue evolution (Eq. (6)). In both cases, we utilized empirical covariance matrices extracted from a finite number of samples M, sampled from the distribution (1) with finite N. These results are almost insensible to the choice of the population spectrum as discussed in Appendix F.

We demonstrate that the phenomena of overfitting and finite-M corrections can be accurately modeled using RMT to predict the $N, M \to \infty$ limit, thereby removing the need for empirical data. Detailed methodologies are provided in Appendix G. For a constant $\rho = M/N = 1.5$, Fig. 4-(**a**) compares the reconstruction error of $\boldsymbol{J}(t)$ over the training period for various N values (colored lines) against the asymptotic RMT prediction (dashed black lines), showing strong consistency as N increases. This agreement extends to the evolution of the test LL (not shown) and the timing of the minimum error and peak test LL as functions of ρ (see Fig. 4-(**b**)). Notably, the optimal stopping times for the two estimators do not coincide, yet finite (N, M) trainings align precisely with the asymptotic predictions.

5. Protocols to mitigate overfitting

In the GEBM, non-monotonic behavior stems from adjustments to the eigenvalues of $\widehat{\boldsymbol{C}}^M$ compared to the population covariance matrix. In fact, one can easily check that replacing the population eigenvectors while retaining M-dependent eigenvectors to form an optimally corrected matrix, $\widehat{\boldsymbol{C}}^M_{\text{val-pop}} \overset{\text{def}}{=} \sum_\alpha c_\alpha^* \boldsymbol{u}_\alpha^M \boldsymbol{u}_\alpha^{M\top}$, almost eliminates the non-monotonic effects on model quality and overfitting, as shown in Fig. 5 **(a)** (green), and in other datasets or ρ values we see the bump completely disappear. While effective, this approach is useless for real experiments where the population matrix is unknown. Nonetheless, this idealized scenario informs the design of protocols aimed at minimizing overfitting and reducing reliance on uncontrollable early-stopping strategies. We now explore common strategies to mitigate overfitting within our framework, focusing on regularization and shrinkage corrections. We also introduce a versatile downsampling-guided mode-fitting scheme that allows circumvent the traditional limitations of RMT strategies, and design corrections that should be valid beyond GEBMs.

Regularization. In machine learning, regularization priors are standard for preventing overfitting. In the GEBM, they constrain the growth of eigenvalues J_α, avoiding suboptimal fixed points affected by mode fluctuations in $\widehat{\boldsymbol{C}}^M$. For training dynamics, L_2 regularization is applied to the coupling matrix $\boldsymbol{J}$, and similar outcomes are achieved with projected L_1-regularization on $\boldsymbol{J}$'s eigenbasis, (a protocol that facilitating asymptotic RMT analysis). The impact of

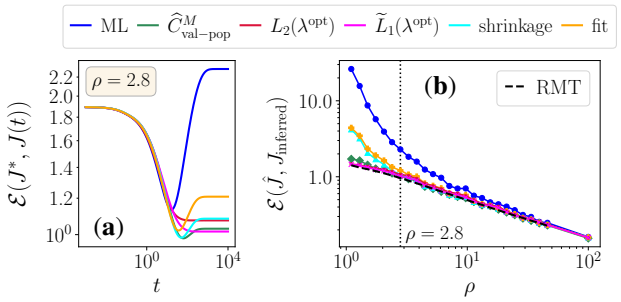

Figure 5. Effect of data-correction protocols on the training a GEBM (in (**a**), for $\rho = 2.8$) and on the final model's quality as a function of ρ (in (**b**)): comparison of the reconstruction error $\mathcal{E}_J$ between training from an empirical covariance matrix $\widehat{\boldsymbol{C}}^M$ (blue), optimal L_2-regularization (w.r.t. reconstruction, red), shrinkage formula (cyan). The settings are the same as in Fig.3.

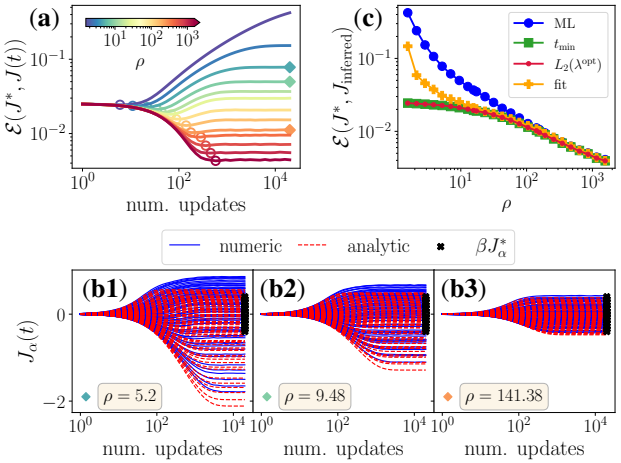

Figure 6. Results on the BM for the inverse Ising problem. (**a**)-(**b**): Training dynamics. (**a**): Reconstruction error (Frobenius norm) vs time (number of updates), for different values of dataset's size. (**b1**)-(**b2**)-(**b3**): evolution of eigenmodes J_α during training for 3 values of ρ: comparison between numerics (blue lines) and analytic curve (red). Black crosses indicate the true models' eigenvalues $\beta\hat{J}_\alpha$. (**c**): effect between data-correction strategies on the inferred model. Comparison between optimal L_2-regularization (red), standard ML-training (blue), modes-fitting (orange) and best reconstruction computed at $t_{\min}$ (green).

regularization at the fixed point is studied for finite N and in the $N \to \infty$ limit via RMT. Fig. 4 (**c**) shows the reconstruction error as a function of λ, with empirical results aligning closely with RMT predictions. An optimal λ_{opt} minimizes the error but does not match to the value that maximizes the test LL (see (**d**)), complicating λ_{opt}'s identification without knowing the population parameters, akin to identifying optimal early stopping. Further details on regularized training and RMT are provided in Appendices H.1 and G. Red line in Fig. 5-(**a**) illustrates the error over time for a L_2 regularized training using the optimal parameter λ^{opt}.

Shrinkage correction protocols are pivotal in statistical learning and signal processing for estimating covariance matrices, particularly when the sample size is small relative to data dimensionality (Bun et al., 2017). Some of these protocols use rotationally invariant estimators (RIEs) to adjust eigenvalues distorted by sampling noise (Ledoit & Wolf, 2004; 2020), while preserving eigenvectors, ensuring corrections are independent of the coordinate system. Based on RMT, RIEs align the eigenvalues of finite-sample covariance matrices to minimize the deviation of the covariance matrix from the population one. Using the optimal RIE from (Bun et al., 2017), we correct our $\widehat{\boldsymbol{C}}^M$ matrices, and use them for training our GEBMs. Depicted in light blue in Fig. 5-(**a**), this new training shows significant improvements in model inference quality, although some non-monotonic behaviors persist. The main drawback of this approach is that it is specific to the GEBM case.

Polynomial fit of eigenmodes. To overcome the limitation of RIEs, we introduce a simple strategy to correct empirically the eigenvalues of $\widehat{\boldsymbol{C}}^M$: the idea is to downsample our dataset to obtain $\widehat{\boldsymbol{C}}^m$ with $m < M$ and use the corresponding eigenvalues $\hat{c}_\alpha^m$ values to extrapolate the $m \to \infty$ limit from a linear fit in $1/m$ (as expected from (Baik & Silverstein, 2006)). Additional details are given in Appendix H.2.

We then use the extrapolated eigenvalues to clean our covariance matrix and run a new training. The evolution of the reconstruction error is shown in Fig. 5-(**a**) (yellow).

Comparison of strategies. The effectiveness of various strategies to counteract overfitting is depicted in Fig. 5-(**b**), presenting the reconstruction error across different ρ values. Notably, the optimal L_2 regularization, the $\widehat{\boldsymbol{C}}^M_{\text{val-pop}}$ strategy, and the performance at the optimal early stopping point derived from RMT all follow similar trajectories, with a $1/\sqrt{\rho}$ scaling for large ρ as expected. While the high performance of these strategies stems from knowing the true model to optimize parameters—not usable in practice—we demonstrate that similar performance can be achieved with RMT-based shrinkage corrections or empirical polynomial fits. These methods do not require prior knowledge of the model, making them especially suitable for real-world inference applications.

6. Boltzmann Machine for inverse Ising

We extend our analysis to the Boltzmann Machine (BM) or the so-called *inverse Ising problem* (Nguyen et al., 2017), adapting our approach to binary variables $\boldsymbol{x} = \{\pm 1\}^N$. This model is able to capture multimodal distributions through its pairwise energy function $E(\boldsymbol{x}) = -\sum_{i<j} J_{ij} x_i x_j - \sum_i h_i x_i$, with parameters $\boldsymbol{\theta} = (\boldsymbol{J}, \boldsymbol{h})$. Due to the lack

of closed-form solutions for the correlation functions and likelihood in BMs, we employ a mean-field approximation suitable e.g. at high temperatures. This approximation allows for an analytic, albeit not exact, expression linking the model's correlation matrix $\boldsymbol{C}$ to the coupling matrix $\boldsymbol{J}$ as $\boldsymbol{C} = (\mathbb{I}_N - \boldsymbol{J})^{-1}$, facilitating an analytical exploration of the training dynamics shown in (Agoritsas et al., 2023) and further elaborated in Appendix I.

Similar to GEBMs, an analytical description of spectral dynamics can be applied to BMs, though with certain limitations due to two main factors: a) the ML estimator for the coupling matrix $\boldsymbol{J}(t)$ may not strictly preserve the same eigendecomposition as $\widehat{\boldsymbol{C}}^M$, despite typically observing a nice alignment for the strongest modes; and b) the binary nature of the variables is not accounted for in the diagonals of the covariance matrices. We must remove the diagonal constraint to allow independent evolution of the modes, similar to the GEBM scenario, since a spherical constraint as in (Fanthomme et al., 2022) would introduce mode coupling. This decision is crucial for deriving approximate analytical expressions for training dynamics; without it, the problem becomes intractable in time. However, the proper fixed point alone can still be effectively analyzed using mean field techniques (Kappen & Rodríguez, 1998).

As detailed in Appendix I, we can project the gradient on the spectral basis of $\widehat{\boldsymbol{C}}^M$ and obtain an approximate analytic expression for the evolution of the eigenvalues of $\boldsymbol{J}$:

$$\tau \tfrac{\mathrm{d}J_\alpha}{\mathrm{d}t} \approx \hat{c}_\alpha^M - \tfrac{1}{1-J_\alpha}, \tag{11}$$

$$J_\alpha(t) \approx 1 - \tfrac{1}{\hat{c}_\alpha^M} - \tfrac{1}{\hat{c}_\alpha^M} W_0\left[B_\alpha e^{-\left(\hat{c}_\alpha^M\right)^2 \frac{t}{\tau}}\right], \tag{12}$$

whose fixed point is $J_\alpha^\infty \approx 1 - (\hat{c}_\alpha^M)^{-1}$. This fixed point is shifted due to our unconstrained diagonal, and neither is $\boldsymbol{J}$ traceless, as compared to the complete treatment. However, Eq. (12) still qualitatively captures the training dynamics of BMs, revealing significant differences from the GEBM scenario where J_α and c_α are no longer inversely proportional. Notably, the smallest c_α values are associated with negative J_α values which are not necessarily small in absolute terms, which significantly contribute to the reconstruction of $\boldsymbol{J}$. Additionally, for positive J_α, J_α increases when $\hat{c}_\alpha^M$ does.

Results. We conducted numerical experiments training an Ising-BM on equilibrium data sampled from a 2D Ising model (i.e. defining $\boldsymbol{J}^*$ on a 2D nearest neighbors lattice) with $N = 8 \times 8$ spins at high temperature ($\beta = 0.1$). Figure 6 presents the results: Panel (**a**) shows the reconstruction error between the trained model and the ground truth $\beta \boldsymbol{J}^*$ for different ρ values, revealing a non-monotonic trend at low ρ that mirrors observations made with GEBMs (a behavior which is robust w.r.t. the system size, see Fig. 20 in Appendix I). Panels (**b1**)-(**b3**) track the eigenvalue evolution during training for three ρ values, comparing numerical results (eigenspectrum of $\boldsymbol{J}(t)$) with the analytic curve from Eq. (12). While the trends align qualitatively, particularly in capturing the separation of time scales, the analytic curves consistently underestimate the actual eigenvalue evolution due to the overlooked diagonal constraint in the BM model. Nonetheless, this timescale separation is similar to that observed in the GEBM: stronger covariances ($\hat{c}_\alpha^M > 1 \leftrightarrow J_\alpha > 0$) are learned faster, while weaker covariances ($\hat{c}_\alpha^M < 1 \leftrightarrow J_\alpha < 0$) take longer. This pattern indicates that the training dynamics are dominated by the convergence of weak $\hat{c}_\alpha^M$. In sparse Ising models, this involves learning the negative spectrum of $\boldsymbol{J}$. Unlike in GEBMs where negative eigenvalues do not exist, in BMs, these later-encoded modes significantly impact the overall reconstructed $\boldsymbol{J}$ due to their large absolute eigenvalues, even though they have negligible effect in sampling quality. Accurately inferring sparse Ising models hinges on effectively learning weaker covariances, heavily influenced by finite-data noise. However, training good generative models is considerably faster, see Appendix I.

Strategies similar to those used in GEBMs can be employed to mitigate overfitting in BMs, with comparable outcomes as shown in Fig. 6-(**c**). Shrinkage formulas are not applicable to BMs, yet the empirical polynomial fit correction for the eigenvalues proves still effective at reducing overfitting effects. This is a very good outcome as it is the only non-informative correction (as the identification of $t_{\min}$ or λ_{opt} requires knowing $\boldsymbol{J}^*$).

7. Theoretic extension to generic EBM learning

The analysis of overfitting for simple models such as the GEBM or the BM constitutes a preparatory attempt to address this question in the broader context of EBMs. Let us see now to which extent the analyses of overfitting carried out so far is also relevant for more general EBMs. The line of arguments bears some similarity with the one justifying that high-dimensional linear regressions are relevant to analyze deep learning (Belkin et al., 2018; Hastie et al., 2022). To this end let us consider the score-matching algorithm (Hyvärinen & Dayan, 2005) as a theoretical proxy for the analyses of overfitting in EBM. Even though this approach might appear sub-optimal in many circumstances we postulate that the mechanisms leading to overfitting have similar origin as in more sophisticated methods. Consider a generic EBM of the form $p(\boldsymbol{x}|\boldsymbol{\theta}) = Z^{-1}(\boldsymbol{\theta})\, e^{-E(\boldsymbol{x}|\boldsymbol{\theta})}$ where $\theta \in \mathbb{R}^P$ is the vector of parameters and having a train set $\mathcal{D} = \{\boldsymbol{x}_i, i = 1, \dots M\}$ of size M. Defining the score function as $\psi(\boldsymbol{x}|\theta) \overset{\text{def}}{=} -\nabla_{\boldsymbol{x}} E(\boldsymbol{x}|\theta)$, the score matching loss is given by $\mathcal{L}_{\mathrm{SM}}(\theta) = \frac{1}{2}\hat{\mathbb{E}}_{\boldsymbol{x}}\left[\left\|\left(\psi(\boldsymbol{x}|\theta) - \nabla \log \hat{p}(\boldsymbol{x})\right\|^2\right]\right.$

which, thanks to a by part integration rewrites as

$$\mathcal{L}_{\rm SM}(\theta) = \hat{\mathbb{E}}_{\boldsymbol{x}}\Big[\frac{1}{2}\big\|\nabla\big(E(\boldsymbol{x}|\theta)\big\|^2 - \Delta E(\boldsymbol{x}|\theta)\Big] + \text{Cst.} \quad (13)$$

Notice first that for the GEBM, this leads to a learning dynamics of the coupling matrix corresponding to $\frac{dJ(t)}{dt} = -\big(\hat{C}J(t) + J(t)\hat{C}\big) + \mathbb{I}$ leading to the solution

$$j_t(x) = \frac{1-e^{-xt}}{x}, \quad (14)$$

when assuming that the initial condition commutes with $\hat{C}$. More generally, the dynamics of the score function is governed by a neural tangent kernel (NTK) (Jacot et al., 2018). We have

$$\frac{d\psi(\boldsymbol{x}|\theta_t)}{dt} = -\hat{\mathbb{E}}_{\boldsymbol{x}'}\Big[K_t(\boldsymbol{x},\boldsymbol{x}')\big(\psi(\boldsymbol{x}'|\theta_t) - \nabla\log\hat{p}(\boldsymbol{x}')\big)\Big] \quad (15)$$

with $K_t(\boldsymbol{x},\boldsymbol{x}') = \partial_{\theta^\top}\psi(\boldsymbol{x}|\theta_t)\partial_\theta\psi(\boldsymbol{x}'|\theta_t)^\top$. Integrating by parts the second term we obtain

$$\frac{d\psi(\boldsymbol{x}|\theta_t)}{dt} = -\hat{\mathbb{E}}_{\boldsymbol{x}'}\Big[K_t(\boldsymbol{x},\boldsymbol{x}')\big(\psi(\boldsymbol{x}'|\theta_t)\Big] + \hat{\phi}_t(\boldsymbol{x}) \quad (16)$$

with $\hat{\phi}_t(\boldsymbol{x}) = -\hat{\mathbb{E}}_{\boldsymbol{x}'}\Big[\partial_{\theta^\top}\psi(\boldsymbol{x}|\theta)\partial_\theta\nabla^\top_{\boldsymbol{x}'}\psi(\boldsymbol{x}'|\theta)\Big] = -\hat{\mathbb{E}}_{\boldsymbol{x}'}\Big[\nabla_{\boldsymbol{x}'}\cdot K_t(\boldsymbol{x},\boldsymbol{x}')\Big]$. As for supervised learning we expect a kernel regime for large enough network's width (Chizat et al., 2019). Then K becomes deterministic, the dynamics is linear, with $\psi(\boldsymbol{x},t)$ an explicit function of the kernel matrix $K(\boldsymbol{x}_s,\boldsymbol{x}_{s'})$ on the training set. Indeed, the NTK dynamics takes place on a reproducing kernel Hilbert space (RKHS) of finite dimension corresponding either to $\mathcal{H}_P \overset{\text{def}}{=} \text{Span}\{\partial_{\theta_q}\psi(\boldsymbol{x}|\theta), q=1,\dots P\}$, or to the $\mathcal{H}_M \overset{\text{def}}{=} \text{Span}\{K(\boldsymbol{x},\boldsymbol{x}_s), s=1,\dots M\}$, depending respectively on whether we are in the under or over-parameterized regime. In the latter case $\hat{K}_{ss'} \overset{\text{def}}{=} \frac{1}{M}K(\boldsymbol{x}_s,\boldsymbol{x}_{s'})$ is full rank and we have

$$\hat{\psi}(t) = -\frac{1-e^{-\hat{K}(t-t_0)}}{\hat{K}}\hat{\phi} + \hat{\psi}(t_0) \quad (17)$$

where $\hat{\psi}(t)$ and $\hat{\phi}$ are respectively the vectors $\{\psi(\boldsymbol{x}_s|\theta_t), s=1,\dots M\}$ and $\{\hat{\phi}(\boldsymbol{x}_s), s=1,\dots M\}$ and assuming $\hat{\psi}(t_0) = 0$. In any case we can consider only the projection of ψ on the RKHS, its transverse part being assumed to be zero at $t = t_0$. As a result, the dynamics takes place in the "empirical" RKHS and we have

$$\psi(\boldsymbol{x}|\theta_t) = \frac{1}{M}\sum_{s=1}^{M} K(\boldsymbol{x},\boldsymbol{x}_s)\beta_s(t) \quad (18)$$

where the vector β is obtained from (17) yielding finally

$$\psi(\boldsymbol{x}|\theta_t) = -\hat{K}(\boldsymbol{x})^\top\frac{j_t(\hat{K})}{\hat{K}}\hat{\phi} + \psi(\boldsymbol{x}|\theta_{t_0}) \quad (19)$$

where $\hat{K}(\boldsymbol{x}) = \{\frac{1}{M}K(\boldsymbol{x},\boldsymbol{x}_s), s=1,\dots M\}$ is the vector of empirical features spanning $\mathcal{H}_E$. Additionally θ_t is directly read off from $\psi(\boldsymbol{x}|\theta_t)$ at first order in θ_t in the lazy regime

$$\psi(\boldsymbol{x}|\theta_t) \approx \nabla_\theta^\top\psi(\boldsymbol{x}|\theta_{t_0})(\theta_t - \theta_{t_0}) \quad (20)$$

Using the parameter-sample duality eventually leads to

$$\theta_t = \theta_{t_0} + \frac{j_t(C^{(M)})}{C^{(M)}}\phi^{(M)} \quad (21)$$

where

$$C^{(M)} = \frac{1}{M}\sum_{s=1}^{M}\nabla_\theta\psi(\boldsymbol{x}_s|\theta)^\top\nabla_{\theta^\top}\psi(\boldsymbol{x}_s|\theta) \quad (22)$$

$$\phi^{(M)} \overset{\text{def}}{=} \frac{1}{M}\sum_{s=1}^{M}\nabla_\theta\psi(\boldsymbol{x}_s|\theta)^\top\hat{\phi}(\boldsymbol{x}_s) \quad (23)$$

In GEBM case we recover (14) by letting $\psi(\boldsymbol{x}|\theta) = \theta\boldsymbol{x}$, $K(\boldsymbol{x},\boldsymbol{x}') = \frac{1}{2}(\boldsymbol{x}\boldsymbol{x}'^\top + \boldsymbol{x}^\top\boldsymbol{x}')$ and $\hat{\phi}(\boldsymbol{x}) = \boldsymbol{x}$, leading to $\phi^{(M)} = C^{(M)}$ with $C^{(M)} = \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{x}_i\boldsymbol{x}_i^t$.

8. Discussion

This work presents a theoretical framework to understand overfitting in simple energy-based models, using the eigen-decomposition of the data covariance matrix to analyze training dynamics. We illustrate how the principal components control a timescale separation, where information progressively encoded from the strongest to the weakest data modes. Due to varying impacts of finite-size noise on different components, this results in an early-stopping point dictated by their interplay. Furthermore, we show that finite sample corrections can be very accurately described analytically using asymptotic RMT analyses. This analysis provide us with an analogous of the GCV in the context of EBM, which deserves further empirical investigations. This analysis is exact for Gaussian EBMs and approximate for Ising-BMs at high temperatures, capturing similar phenomena observed in more complex EBMs like RBMs. We discuss data-correction protocols typically used to mitigate overfitting and propose to extend these strategies to more complex models leveraging higher-order data correlations (e.g. by exploiting the SVD decomposition). Further investigations into RMT may clarify how early-stopping points relate to the asymptotic properties of the covariance matrix's spectrum or which should be the proper observables to pinpoint them without a prior knowledge of the data model. Finally, an extension of the theory to EBM via a neural tangent kernel dynamics of the score function deserves further experimental investigations to find relevant hypothesis for the spectrum of population covariance matrices of tangent features.

Acknowledgments

Authors acknowledge financial support by the Comunidad de Madrid and the Complutense University of Madrid through the Atracción de Talento program (Refs. 2019-T1/TIC-13298 & Refs. 2023- 5A/TIC-28934), the project PID2021-125506NA-I00 financed by the "Ministerio de Economía y Competitividad, Agencia Estatal de Investigación" (MICIU/AEI/10.13039/501100011033), the Fondo Europeo de Desarrollo Regional (FEDER, UE) and the French ANR grant Scalp (ANR-24- CE23-1320).

Impact Statement

This paper aims to advance the field of Machine Learning by deepening our understanding of generative models under data scarcity. While our findings may have broad societal implications, we do not identify any that require specific emphasis at this stage.

References

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. A learning algorithm for Boltzmann machines. *Cognitive science*, 9 (1):147–169, 1985.

Advani, M. S., Saxe, A. M., and Sompolinsky, H. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

Agoritsas, E., Catania, G., Decelle, A., and Seoane, B. Explaining the effects of non-convergent MCMC in the training of energy-based models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 322–336. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/agoritsas23a.html.

Arnaboldi, L., Stephan, L., Krzakala, F., and Loureiro, B. From high-dimensional & mean-field dynamics to dimensionless odes: A unifying approach to sgd in two-layers networks. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 1199–1227. PMLR, 2023.

Atanasov, A., Zavatone-Veth, J., and Pehlevan, C. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.

Baik, J. and Silverstein, J. W. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis*, 97(6):1382–1408, 2006.

Belkin, M. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

Belkin, M., Ma, S., and Mandal, S. To understand deep learning we need to understand kernel learning. In *proc. of ICML*, pp. 541–549. PMLR, 2018.

Bengesi, S., El-Sayed, H., Sarker, M. K., Houkpati, Y., Irungu, J., and Oladunni, T. Advancements in generative ai: A comprehensive review of gans, gpt, autoencoders, diffusion model, and transformers. *IEEE Access*, 2024.

Béreux, N., Decelle, A., Furtlehner, C., and Seoane, B. Learning a restricted Boltzmann machine using biased monte carlo sampling. *arXiv preprint arXiv:2206.01310*, 2022.

Béreux, N., Decelle, A., Furtlehner, C., Rosset, L., and Seoane, B. Fast, accurate training and sampling of restricted Boltzmann machines. *arXiv preprint arXiv:2405.15376*, 2024.

Bun, J., Bouchaud, J.-P., and Potters, M. Cleaning large correlation matrices: Tools from random matrix theory. *Physics Reports*, 666:1–109, 2017. ISSN 0370-1573. doi: https://doi.org/10.1016/j.physrep.2016.10.005. URL https://www.sciencedirect.com/science/article/pii/S0370157316303337. Cleaning large correlation matrices: tools from random matrix theory.

Bun, J., Bouchaud, J.-P., and Potters, M. Overlaps between eigenvectors of correlated random matrices. *Phys. Rev. E*, 98:052145, Nov 2018. doi: 10.1103/PhysRevE.98.052145. URL https://link.aps.org/doi/10.1103/PhysRevE.98.052145.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. *In proc. of NeurIPS*, 32, 2019.

Cocco, S., Feinauer, C., Figliuzzi, M., Monasson, R., and Weigt, M. Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics*, 81 (3):032601, 2018.

Consortium, . G. P. et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

Decelle, A., Fissore, G., and Furtlehner, C. Thermodynamics of restricted Boltzmann machines and related learning dynamics. *Journal of Statistical Physics*, 172 (6):1576–1608, 2018. doi: https://doi.org/10.1007/s10955-018-2105-y.

Decelle, A., Seoane, B., and Rosset, L. Unsupervised hierarchical clustering using the learning dynamics of restricted Boltzmann machines. *Physical Review E*, 108(1):014110, 2023.

Decelle, A., Furtlehner, C., Gómez, A. D. J. N., and Seoane, B. Inferring effective couplings with restricted Boltzmann machines. *SciPost Phys.*, 16:095, 2024. doi: 10.21468/SciPostPhys.16.4.095. URL https://scipost.org/10.21468/SciPostPhys.16.4.095.

Decelle, A., de Jesús Navas Gómez, A., and Seoane, B. Inferring high-order couplings with neural networks. 2025. URL https://arxiv.org/abs/2501.06108.

Delon, J., Desolneux, A., and Salmona, A. Gromov–wasserstein distances between gaussian distributions. *Journal of Applied Probability*, 59(4):1178–1198, 2022. doi: 10.1017/jpr.2022.16.

Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Du, Y. and Mordatch, I. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32, 2019.

Fanthomme, A., Rizzato, F., Cocco, S., and Monasson, R. Optimal regularizations for data generation with probabilistic graphical models. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(5):053502, may 2022. doi: 10.1088/1742-5468/ac650c. URL https://dx.doi.org/10.1088/1742-5468/ac650c.

Feinauer, C. and Lucibello, C. Reconstruction of pairwise interactions using energy-based models. In *Mathematical and Scientific Machine Learning*, pp. 291–313. PMLR, 2022.

Feinauer, C., Meynard-Piganeau, B., and Lucibello, C. Interpretable pairwise distillations for generative protein sequence models. *PLOS Computational Biology*, 18(6): e1010219, 2022.

Furtlehner, C. Free dynamics of feature learning processes. *J.Stat.Phys*, 190(3):51, 2023.

Golub, G., Heath, M., and Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Hachem, W., Loubaton, P., and Najim, J. Deterministic equivalents for certain functionals of large random matrices. 2007.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

Hyvärinen, A. and Dayan, P. Estimation of non-normalized statistical models by score matching. *JMLR*, 6(4), 2005.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *In proc. of NeurIPS*, volume 31, 2018.

Kappen, H. J. and Rodríguez, F. B. Efficient Learning in Boltzmann Machines Using Linear Response Theory. *Neural Computation*, 10(5):1137–1156, 07 1998. ISSN 0899-7667. doi: 10.1162/089976698300017386. URL https://doi.org/10.1162/089976698300017386.

Kiwata, H. Estimation of quenched random fields in the inverse ising problem using a diagonal matching method. *Phys. Rev. E*, 89:062135, Jun 2014. doi: 10.1103/PhysRevE.89.062135. URL https://link.aps.org/doi/10.1103/PhysRevE.89.062135.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Ledoit, O. and Péché, S. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1):233–264, 2011.

Ledoit, O. and Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. ISSN 0047-259X. doi: https://doi.org/10.1016/S0047-259X(03)00096-4. URL https://www.sciencedirect.com/science/article/pii/S0047259X03000964.

Ledoit, O. and Wolf, M. Analytical nonlinear shrinkage of large-dimensional covariance matrices. *The Annals of Statistics*, 48(5):3043 – 3065, 2020. doi: 10.1214/19-AOS1921. URL https://doi.org/10.1214/19-AOS1921.

Loffredo, E., Pastore, M., Cocco, S., and Monasson, R. Restoring balance: principled under/oversampling of data for optimal classification. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

MacKay, D. J. C. *Information Theory, Inference, and Learning Algorithms*. Copyright Cambridge University Press, 2003.

Magnus, J. R. and Neudecker, H. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley, second edition, 1999. ISBN 0471986321 9780471986324 047198633X 9780471986331.

Mai, X., Liao, Z., and Couillet, R. A large scale analysis of logistic regression: Asymptotic performance and new insights. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3357–3361. IEEE, 2019.

Marčenko, V. and Pastur, L. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

Nguyen, H. C., Zecchina, R., and Berg, J. Inverse statistical problems: from the inverse Ising problem to data science. *Advances in Physics*, 66(3):197–261, 2017. doi: 10.1080/00018732.2017.1341604. URL https://doi.org/10.1080/00018732.2017.1341604.

Patil, P., Wu, Y., and Tibshirani, R. Failures and successes of cross-validation for early-stopped gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pp. 2260–2268. PMLR, 2024.

Potters, M. and Bouchaud, J.-P. *A first course in random matrix theory: for physicists, engineers and data scientists*. Cambridge University Press, 2020.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., and Courville, A. On the spectral bias of neural networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/rahaman19a.html.

Ricci-Tersenghi, F. The bethe approximation for solving the inverse Ising problem: a comparison with other inference methods. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08015, aug 2012. doi: 10.1088/1742-5468/2012/08/P08015. URL https://dx.doi.org/10.1088/1742-5468/2012/08/P08015.

Roudi, Y., Aurell, E., and Hertz, J. A. Statistical physics of pairwise probability models. *Frontiers in computational neuroscience*, 3:652, 2009.

Saad, D. and Solla, S. Dynamics of on-line gradient descent learning for multilayer neural networks. *Advances in neural information processing systems*, 8, 1995.

Saxe, A. M., McClelland, J. L., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In Bengio, Y. and LeCun, Y. (eds.), *ICLR*, 2014. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2014.html#SaxeMG13.

Suzuki, M. and Kubo, R. Dynamics of the Ising model near the critical point. i. *Journal of the Physical Society of Japan*, 24(1):51–60, 1968. doi: 10.1143/JPSJ.24.51.

Tomasini, U. M., Sclocchi, A., and Wyart, M. Failure and success of the spectral bias prediction for Laplace kernel ridge regression: the case of low-dimensional data. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21548–21583. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/tomasini22a.html.

Tubiana, J., Cocco, S., and Monasson, R. Learning protein constitutive motifs from sequence data. *Elife*, 8:e39397, 2019.

Wei, A., Hu, W., and Steinhardt, J. More than a toy: Random matrix models predict how real-world neural representations generalize. In *International Conference on Machine Learning*, pp. 23549–23588. PMLR, 2022.

Wu, Z., Johnston, K. E., Arnold, F. H., and Yang, K. K. Protein sequence design with deep generative models. *Current opinion in chemical biology*, 65:18–27, 2021.

Yasuda, M. and Tanaka, K. Susceptibility propagation by using diagonal consistency. *Phys. Rev. E*, 87:012134, Jan 2013. doi: 10.1103/PhysRevE.87.012134. URL https://link.aps.org/doi/10.1103/PhysRevE.87.012134.

Yelmen, B., Decelle, A., Ongaro, L., Marnetto, D., Tallec, C., Montinaro, F., Furtlehner, C., Pagani, L., and Jay, F. Creating artificial human genomes using generative neural networks. *PLoS genetics*, 17(2):e1009303, 2021.

Yelmen, B., Decelle, A., Boulos, L. L., Szatkownik, A., Furtlehner, C., Charpiat, G., and Jay, F. Deep convolutional and conditional neural networks for large-scale genomic data generation. *PLoS Computational Biology*, 19(10):e1011584, 2023.

A. Derivation of projected gradient equations

Starting from the log-likelihood's derivative w.r.t. a parameter J_{ij}, we can assume that in the limit of an infinitely small learning rate $\gamma \to 0$ we can replace the discrete-time update equation for the parameter (2) into a differential equation for the evolution of each parameter J_{ij}:

$$J_{ij}\left(t+1\right) = J_{ij}\left(t\right) + \gamma \left.\frac{\partial \mathcal{L}}{\partial J_{ij}}\right|_{\boldsymbol{J}(t)} \longrightarrow \frac{1}{\gamma}\frac{dJ_{ij}}{dt} = \left.\frac{\partial \mathcal{L}}{\partial J_{ij}}\right|_{\boldsymbol{J}(t)}. \tag{24}$$

We now decompose the rhs of the above expression in terms of time-evolution of eigenvalues and eigenvectors of $\boldsymbol{J}$ at time t. Given the eigendecomposition $J_{ij} = \sum_\gamma v_i^\gamma J_\gamma v_j^\gamma$, we have

$$\frac{dJ_{ij}}{dt} = \frac{d}{dt}\sum_\gamma v_i^\gamma J_\gamma v_j^\gamma = \sum_\gamma \left(\frac{dv_i^\gamma}{dt} J_\gamma v_j^\gamma + v_i^\gamma \frac{dJ_\gamma}{dt} v_j^\gamma + v_i^\gamma J_\gamma \frac{dv_j^\gamma}{dt}\right). \tag{25}$$

We now project this on the eigenbasis of the eigenvectors of $\boldsymbol{J}$, which after simple algebraic manipulations leads to

$$\sum_{ij} v_i^\alpha \frac{dJ_{ij}}{dt} v_j^\beta = \delta_{\alpha\beta}\frac{dJ_\alpha}{dt} + \left(1-\delta_{\alpha\beta}\right)\left(\sum_i v_i^\alpha \frac{dv_i^\beta}{dt} J_\beta + \sum_j \frac{dv_j^\alpha}{dt} J_\alpha v_j^\beta\right) \tag{26}$$

$$= \delta_{\alpha\beta}\frac{dJ_\alpha}{dt} + \left(1-\delta_{\alpha\beta}\right)\left(J_\beta - J_\alpha\right)\sum_i v_i^\alpha \frac{dv_i^\beta}{dt}, \tag{27}$$

where we used the property $\mathrm{d}\left(\boldsymbol{u}^\alpha \cdot \boldsymbol{u}^\beta\right) = 0$ because they are vectors of an orthonormal basis. Finally, combining Eqs. (4) and (27) separating the contributions for $\alpha = \beta$ and $\alpha \neq \beta$ we get to Eq. (5) in the main text.
A final note on the log-likelihoods' gradient: in the first expression (3) we have assumed that perturbations are not symmetric, that is when taking the derivative w.r.t. J_{ij} for $i \neq j$ we assume that $J_{ij} \neq J_{ji}$. Assuming instead *symmetric perturbations* one would get a slight different form of the log-likelihood's gradient w.r.t. (3), given by:

$$\frac{\partial \mathcal{L}}{\partial J_{ij}} = \Lambda_{ij}\left[-\widehat{C}_{ij}^M + \left(\boldsymbol{J}^{-1}\right)_{ij}\right], \tag{28}$$

with $\Lambda_{ij} = 1 - \delta_{ij}/2$. From the point of view of the training fixed point this is not an issue, the ML estimator is exactly the same in both cases. However, the modified gradient (28) leads to a slight different dynamics: in particular, it is not anymore true that the dynamics can be exactly decomposed into a separate evolution for the different eigenvalues of $\boldsymbol{J}$ by following the above steps. One could either include symmetric constraints on the eigendecomposition when computing the projected gradient (a more cumbersome process from a mathematical point of view, see e.g. (Magnus & Neudecker, 1999)) or simply double the learning rate on the diagonal terms $i = j$ to compensate for the factor Λ_{ij}. Nevertheless, the difference between the analytic evolution (i.e. Eq. (6), obtained assuming non-symmetric perturbation) and a numerical training performed using the gradient (28) are almost coincident as seen from Figure 7. None of the results presented in the manuscript is affected by such a difference in the gradient computation.

B. Finite-M fluctuations of eigenbasis

We detail additional eigenvalue spectra of covariance matrices from various datasets in Fig. 8, complementing those in Fig. 1 of the main text. The spectra for CIFAR-10 (Krizhevsky et al., 2009) and the Human Genome Dataset (Consortium et al., 2015) are displayed in (**a**) and (**b**), respectively. Panel (**c**) illustrates the empirical covariance matrix from equilibrium configurations sampled from a 2D Ising model with periodic boundaries at high temperature ($\beta = 0.1$), i.e. in a paramagnetic phase. Panel (**d**) presents another synthetic spectrum, generated through a mixture of power-laws from Eq. (10), but using a different set of parameters than those used for Fig. 1-(**b**) (and used for the trainings in Figs. 2 and 3). This new spectrum was used for the figures involving comparisons with RMT (Figs. 4, 5), for numerical stability issues with the integration of the RMT equations. Nonetheless, all the results presented in the main text about the training dynamics of the GEBM can be perfectly reproduced on a wide range of the parameters defining the population eigenvalues, so that the qualitative picture that emerges from our analysis is extremely robust with respect to specific details of the spectrum.

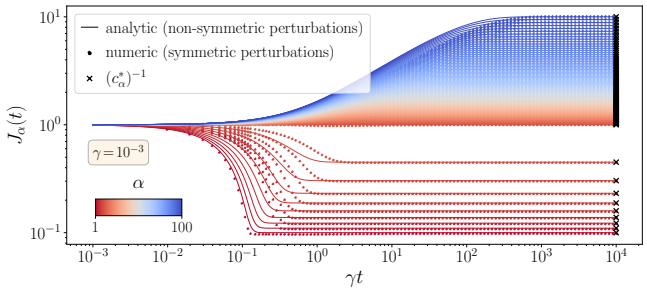

Figure 7. Difference in the eigenvalues' evolution in the training of a GEBM when imposing symmetry or allowing asymmetry in the perturbation of J_{ij}. The points correspond to numerical results obtained by enforcing symmetry on J_{ij} after each update during training (i.e. using Eq. (28)), while the lines represent analytical expressions derived for the case of non-symmetric perturbations (i.e. Eq. (6)). The setting is the same as Figure 2 in the main text.

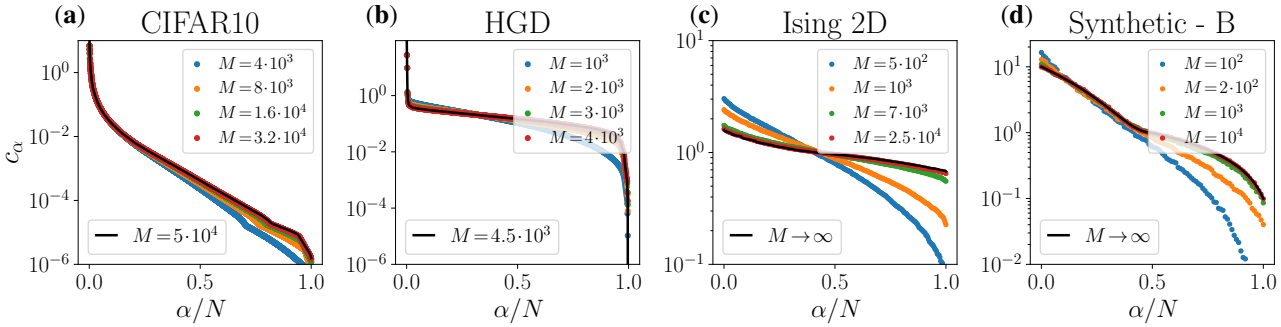

Figure 8. (**a**)-(**b**)-(**c**): Eigenvalue spectra of the empirical covariance matrix of real datasets, respectively CIFAR-10 (in (**a**)) , Human Genome Dataset (in (**b**)), and a dataset made of equilibirum configurations of a 2-d Ising model of size $N = 16^2$ at $\beta = 0.1$ (in (**c**)) . Black lines represent the spectrum computed with the full set of available data (of size M^*), while scatter colored points show the result for a subset of data $M < M^*$. (**d**): the black line shows a synthetic population eigenvalue spectrum generated according to (10), with $N = 100$, and $r = 0.5$, $\beta = 1.0$, $\gamma = 0.5$, $x_1 = 10^{-1}$, $x_2 = 10$; colored points display the eigenvalues of the empirical covariance matrix $\widehat{\boldsymbol{C}}^M$ computed by sampling M configurations from a GEBM with $\boldsymbol{J}^* = \boldsymbol{C}^{*-1}$ (from (1)) for different values of M.

Fig. 9 illustrates how the eigenbasis of the covariance matrix for real datasets remains consistent against downsampling. Starting with the eigenbasis decomposition of the covariance matrix for the largest available dataset M^*—considered our closest approximation to the population matrix $\widehat{\boldsymbol{C}}^{M^*} \approx \widehat{\boldsymbol{C}}^{\infty} = \boldsymbol{C}^*$—we denote its eigenvector matrix as $\boldsymbol{U}^* = \{\boldsymbol{u}^*_\alpha\}_{\alpha=1}^N$. These eigenvectors are arranged columnwise and sorted in descending order by their corresponding eigenvalue. For each reduced sample size $M < M^*$, we perform a similar decomposition on the resultant empirical covariance matrix $\widehat{\boldsymbol{C}}^M$, with its basis represented as $\boldsymbol{U}^M = \left\{\boldsymbol{u}^M_\alpha\right\}_{\alpha=1}^N$. To evaluate the preservation of eigenvectors, we calculate the norm of the matrix product between a projection operator $\boldsymbol{P}^n$ —defined as $\boldsymbol{P}^n = \boldsymbol{U}^*_{1:n}$ (incorporating the first n eigenvectors of $\boldsymbol{C}^*$)— and the α-th eigenvector of $\widehat{\boldsymbol{C}}^M$, $\boldsymbol{u}^M_\alpha$. This measurement determines whether $\boldsymbol{u}^M_\alpha$ falls within the subspace spanned by the first n eigenvectors of $\boldsymbol{C}^*$, thereby helping to mitigate eigenvector oscillations due to exchanges between the ordering of the associated eigenvalues. Fig. 9 shows the norm $\left\|\boldsymbol{P}^{n\top} \cdot \boldsymbol{u}^M_\alpha\right\|$ plotted versus n and for each value of α, for the same 4 datasets of Fig. 8. We can observe that for high values of M most eigenvectors are well preserved: this means that most eigenvectors of $\widehat{\boldsymbol{C}}^M$ are contained in the subspace spanned by $\boldsymbol{C}^*$, as the norm of such a matrix product raises sharply to 1 when $n \approx \alpha$. On the other hand, when M is lowered (lower panels on each subplot) the conservation starts to deteriorate, especially in the middle-lower part of the spectrum. Interestingly, we observe that the most conserved directions (at least in (**a**)-(**b**)) are both the strongest covariance modes and the lowest ones, a phenomenon already highlighted in (Bun et al., 2018).

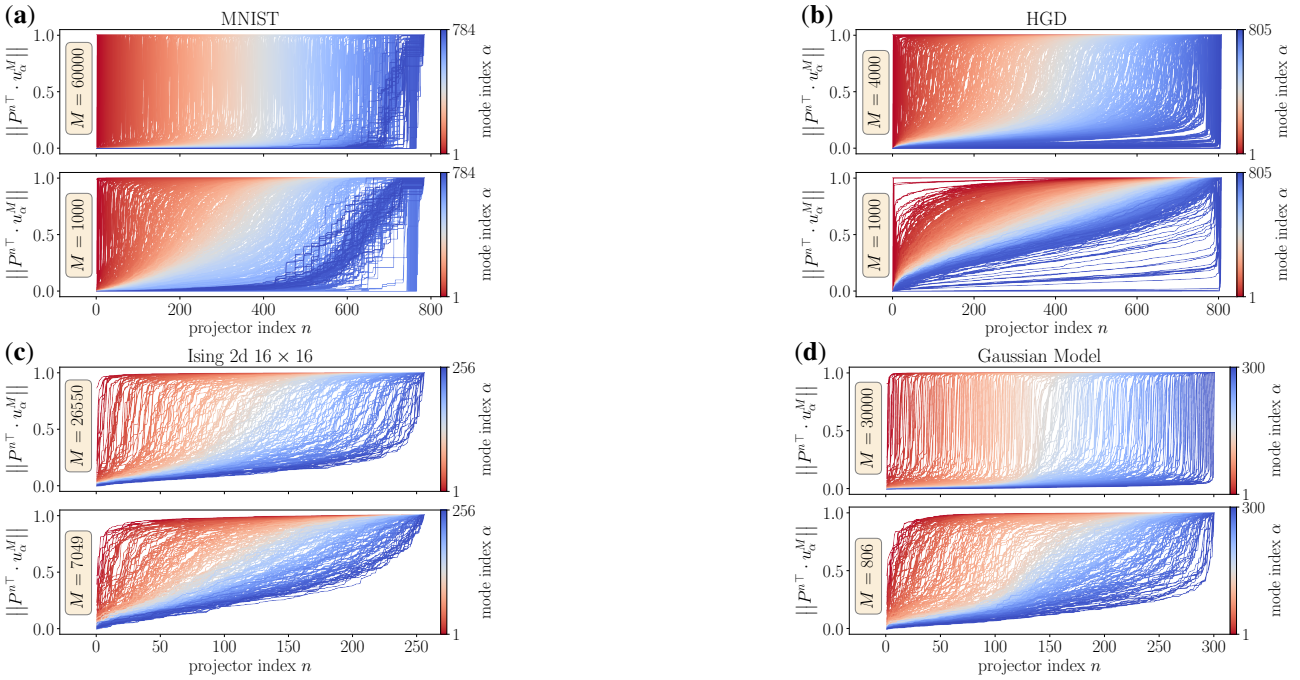

Figure 9. Finite-M fluctuations of eigenvectors in the covariance matrix of datasets. The four panels show the norm of the matrix product between the n-th projection operator $\boldsymbol{P}^n$, containing the first n eigenvectors of the population matrix $\boldsymbol{C}^*$ (for a real dataset, we just take the covariance matrix with the full available data M^*) and the α-th eigenvector of the covariance matrix $\widehat{\boldsymbol{C}}^M$ with $M < M^*$. (**a**)-(**b**)-(**c**) respectively refer to MNIST dataset (Deng, 2012), Human Genome dataset, and equilibrium configurations drawn from a 2d Ising model (same setting as in Fig. 8-(**c**). (**d**) refers to a synthetic Gaussian Model generated as discussed in the main text with the same settings as in Fig. 1. All panels show the results for two values of M, a larger one at the top and a lower one at the bottom. Results are plotted w.r.t. the projector index n and each line correspond to a different α.

C. Training dynamics in GEBM with non-commutative initialization

This section provides a brief follow-up to what discussed in the first part of Section 4, concerning the training dynamics of a GEBM. For simplicity we focus here only on the infinite-sample scenario (i.e. when training from $\boldsymbol{C}^*$), although the same reasoning holds also for finite data. We are also interested in describing the training dynamics for a generic initialization of the matrix $\boldsymbol{J}$, which in general will not commute with $\boldsymbol{C}^*$. In this scenario, the model has also to learn the eigenvectors of $\boldsymbol{C}^*$. Fig. 8 shows the evolution of the coupling matrix eigenvalues J_α according to Eq. (6) (shown with solid lines), in comparison to a numerical training done iteratively maximizing the likelihood as in Eq. (2) (points). The initial condition here is a matrix $\boldsymbol{J}(0)$ constructed from a random population of modes $J_\alpha(0) \sim U[0,1]$ and projected on to a random orthogonal matrix. In this way, $\boldsymbol{J}(0)$ and $\boldsymbol{C}^*$ do not commute. At the beginning of the training, there is indeed a discrepancy between theory and simulations, because of the wrong assumption of independence of eigenvalues. Once eigenvectors align, the evolution proceeds independently for each eigenvalue and perfectly follows Eq. (6).

Note that the initial oscillations of the mode-to-mode eigenvector overlap in Fig. 10-(**b**) is due to the fact that eigenvalue learning is non monotonic at the beginning, so that there is an initial exchange in the ordering of the eigenvectors. Nonetheless, after an initial transient all the eigenvectors align to their counterparts in the covariance matrix. This alignment process and with a much faster timescale w.r.t. the learning of eigenvalues (especially the ones associated to weaker covariances): for this reason, the assumption leading to our analytic description about independency on the eigenvalues' evolution remains justified for practical purposes. This reasoning about eigenvector alignment holds for any input covariance matrix: what determines the non trivial dynamics of the reconstruction error (explained in Sec. 4) is fully determined by the noise in the eigenvalues.

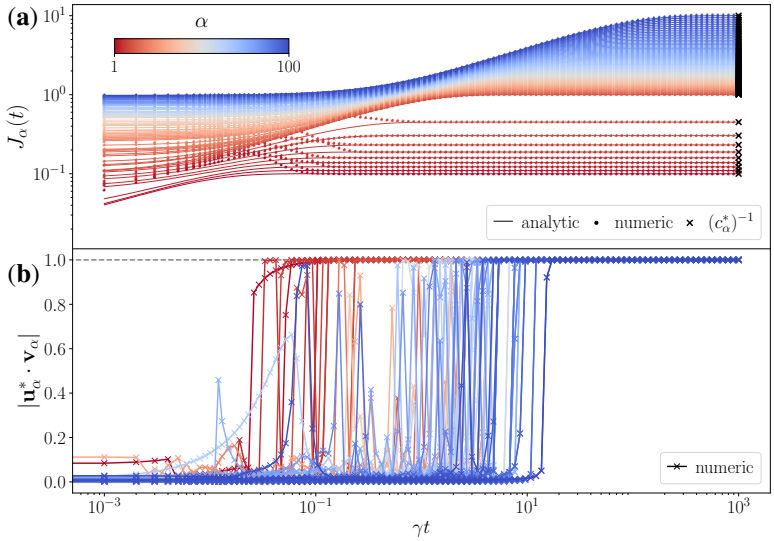

Figure 10. Training dynamics of the GEBM from a population matrix $\boldsymbol{C}^*$. The system size and the parameters defining $\boldsymbol{C}^*$ are the same as in Fig. 8-(**d**). (**a**): Evolution of eigenvalues, comparison between analytic solution (full line) and numerical training (points). The initial condition $\boldsymbol{J}(0)$ is constructed from a random distribution of modes and projecting it back on to a random orthogonal matrix which differs from the eigenbasis of $\boldsymbol{C}^*$, so that the two matrices do not commute. (**b**): Alignment of eigenvectors, computed as the mode-to-mode overlap between eigenvectors of the population matrix $\boldsymbol{u}^*_\alpha$ and eigenvectors of $\boldsymbol{J}$, i.e. $\boldsymbol{v}_\alpha$. Red-ish (resp. blue-ish) colors correspond to strong (resp. weak) covariances c^*_α. The learning rate is set to $\gamma = 10^{-3}$.

D. Robustness of results w.r.t. initialization at finite M

We show in Fig. 11 some analogous results w.r.t. Fig. 2-(**b**) for the training dynamics of the GEBM in the case of finite M (here we set $\rho = M/N = 2.11$) by varying the initialization. In this case, we do not care about eigenvectors' alignment as in the previous section: we only consider different initializations for the eigenmodes of the coupling matrix, i.e. $\{J_\alpha\}_{\alpha=1}^N$. We can observe how the non-monotonic behavior of the reconstruction error (plotted in the bottom rows for each initialization) is robust against different standard, uninformative and small initializations, and it keeps appearing as long as $J_\alpha(0) < 1/\hat{c}^M_\alpha$ for the majority of the eigenvalues (especially the ones corresponding to weak covariances). We can also observe how, increasing the initial conditions to higher values than the fixed point (i.e. moving from columns (**1**)-(**2**)-(**3**) to the right most ones (**4**)-(**5**)) the non-monotonic behavior disappears, indicating that the early-stopping break-point no longer exists and that there is now a way to mitigate the overfitting effects with this strategy. However, it is common practice to start with small values. Moreover, in more complex EBMs where sampling is required to estimate the correlations of the model in the LL gradient (e.g., BMs or RBMs), it may be a very bad idea to assume extreme initializations (i.e. far from an uninformative initialization where the parameters of the model are small): this could lead to ergodicity problems in sampling, as the model may get stuck in spin-glass-like phases, a phenomenon that has been well studied in several EBMs (see e.g. (Decelle et al., 2018)).

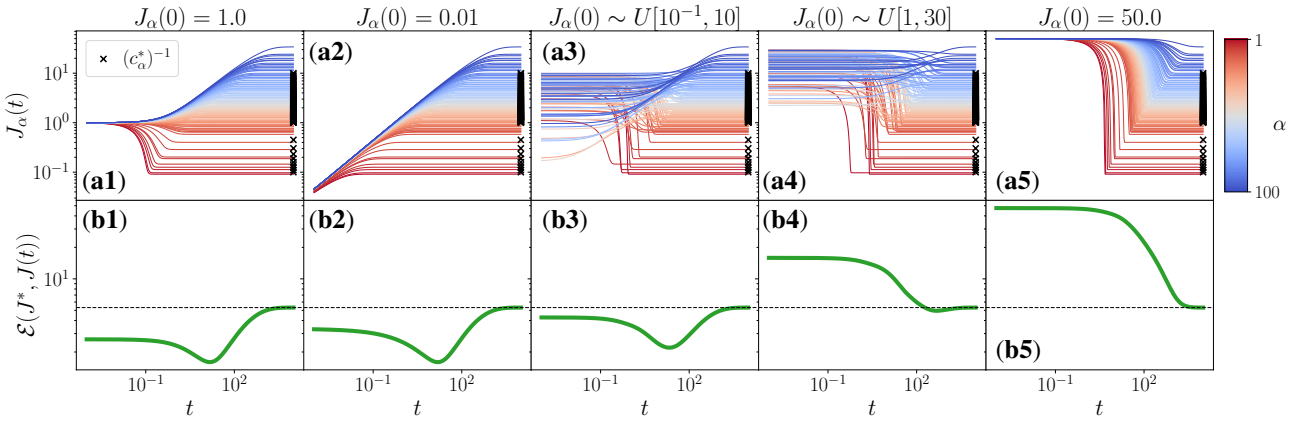

Figure 11. Results on the training dynamics GEBM at finite amount of data by varying the initial conditions. Panels (**a**)'s (top row) show the eigenvalues' evolution (according to Eq. (12)), while panels (**b**)'s (bottom row) show the corresponding reconstruction error $\mathcal{E}_J$ w.r.t the ground truth model $\boldsymbol{J}^*$; all quantities are shown versus time. Each column corresponds instead to a different initial condition. From left to right: an identity-like initialization ($J_\alpha(0) = 1$) in column (**1**), as in Fig. 2-(**b**); a small-coupling initialization in ($J_\alpha(0) = 10^{-2}$), in column (**2**); two random initialization of modes (resp. in the boundaries $[10^{-1}; 10]$ and $[1; 30]$ in column (**3**)-(**4**)); a constant initialization to very high values larger than the fixed point, i.e. ($J_\alpha(0) = 50 > 1/\hat{c}_\alpha^M$) in column (**5**). All trainings are performed analytically, with an empirical covariance matrix $\widehat{\boldsymbol{C}}^M$ generated with the same settings as in Fig. 2 with $\rho = M/N = 2.11$.

E. Additional generation quality metrics

In this section we consider an additional metric to compute the discrepancy between the trained model and the true one, namely the Wasserstrein distance (Delon et al., 2022). Figure 12 shows the same data as in Figure 3 of the main text, now including the evolution of the Wasserstrein distance between the trained model and the true one w.r.t. training time. Also this quantity shows a non-monotonic behavior in t especially for low M, with a clear early-stopping point. In the rightmost panel, we compare the locations of the minima of each error estimator (and the maximum of the log-likelihood) as functions of ρ. We observe that the time point corresponding to the minimum Wasserstein distance follows a trend very similar to that of the maximum log-likelihood.

F. GEBM analysis with various eigenvalue spectra

This section presents additional results on the GEBM, analogous to Fig. 3 in the main text, obtained using alternative spectra for the population covariance matrix. To assess the robustness of our findings with respect to spectral choice, we replicate the analysis using both synthetic and empirical spectra.

First, in Fig. 13, we consider a synthetic spectrum distinct from Eq. (10): for $N = 100$, we generate 10 dominant modes with amplitudes uniformly distributed in $[2, 10]$, and a bulk of $N - k = 90$ noisy modes with amplitudes in $[10^{-1}, 1]$. The qualitative behavior, including overfitting effects, remains unchanged.

Next, we repeat the analysis using empirical spectra: specifically, the eigenvalues of the sample covariance matrices from the MNIST and Human Genome Dataset (HGD), shown in Figs. 14 and 15, are used as population spectra. The resulting dynamics, analogous to Fig. 12, again show no qualitative deviations. In all cases, the non-monotonic temporal behavior of key metrics and the early-stopping times are preserved.

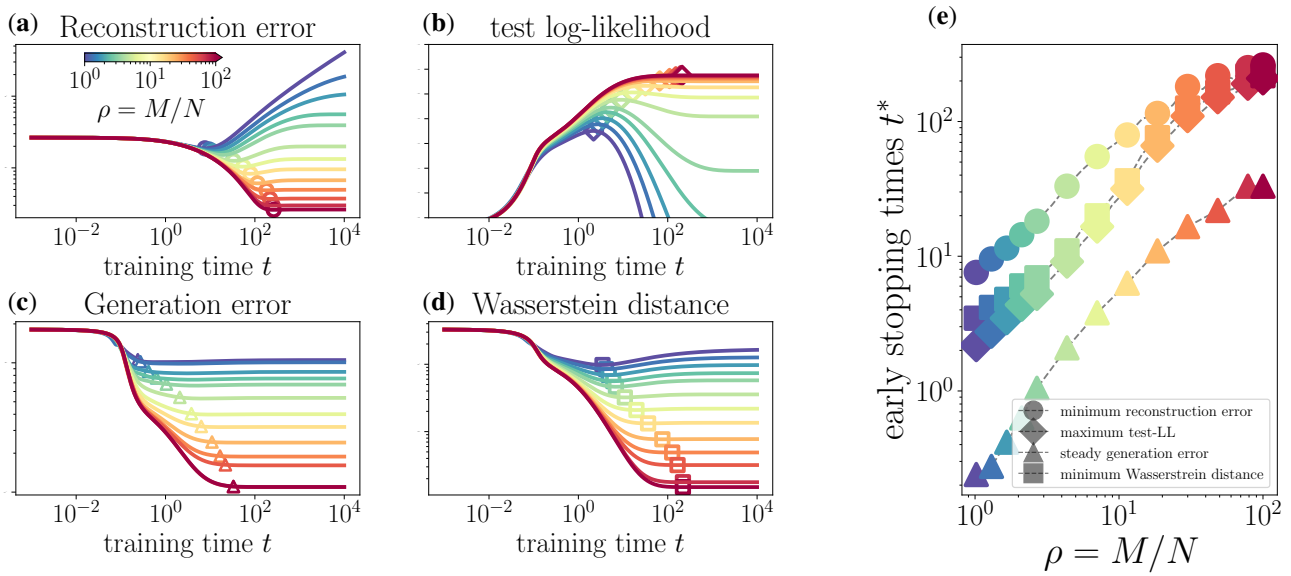

Figure 12. We compare the results shown in Fig. 3, obtained using various generation quality measures, with the corresponding curves computed using the Wasserstein distance. (**a**)-(**b**)-(**c**)-(**d**) display respectively the reconstruction error $\mathcal{E}_J$, the test-LL, the Wasserstein distance and the generation error $\mathcal{E}_C$, all plotted vs time, for various sample sizes M (indicated by a color gradient from blue to red for increasing $\rho = M/N$. Dashed black lines refer to a training from $\boldsymbol{C}^*$ (i.e. $M \to \infty$). (**e**): comparison between time of minimum reconstruction (circles), maximum test LL (diamonds), minimum Wasserstein distance (squares) and time at which the generation error converges to its steady-state value. These quantities are also shown in the related panels for better clarity. Apart on panel (**d**), this figure contains the same information and quantities as Fig. 3.

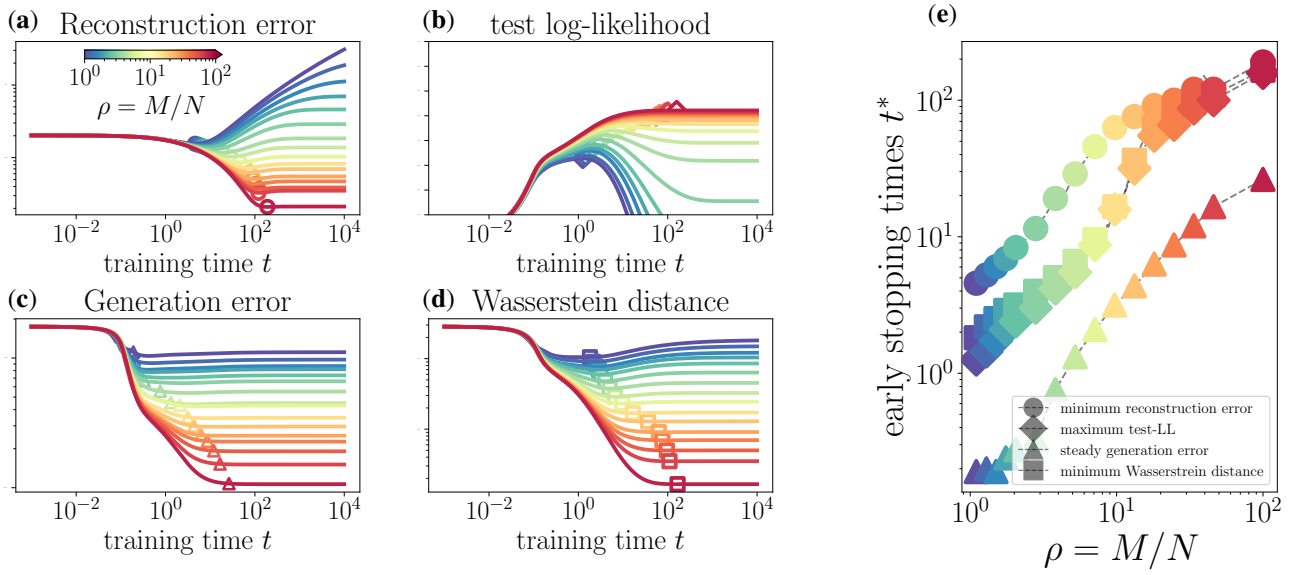

Figure 13. Same plots as in Fig. 12, this time obtained by training a GEBM starting from a synthetic population covariance matrix spectrum of dimension $N = 100$, with a set of 10 dominant modes with amplitudes uniformly distributed in the interval $[2, 10]$, and a bulk of $N - k = 90$ noisy modes with amplitudes uniformly distributed between 10^{-1} and 1.

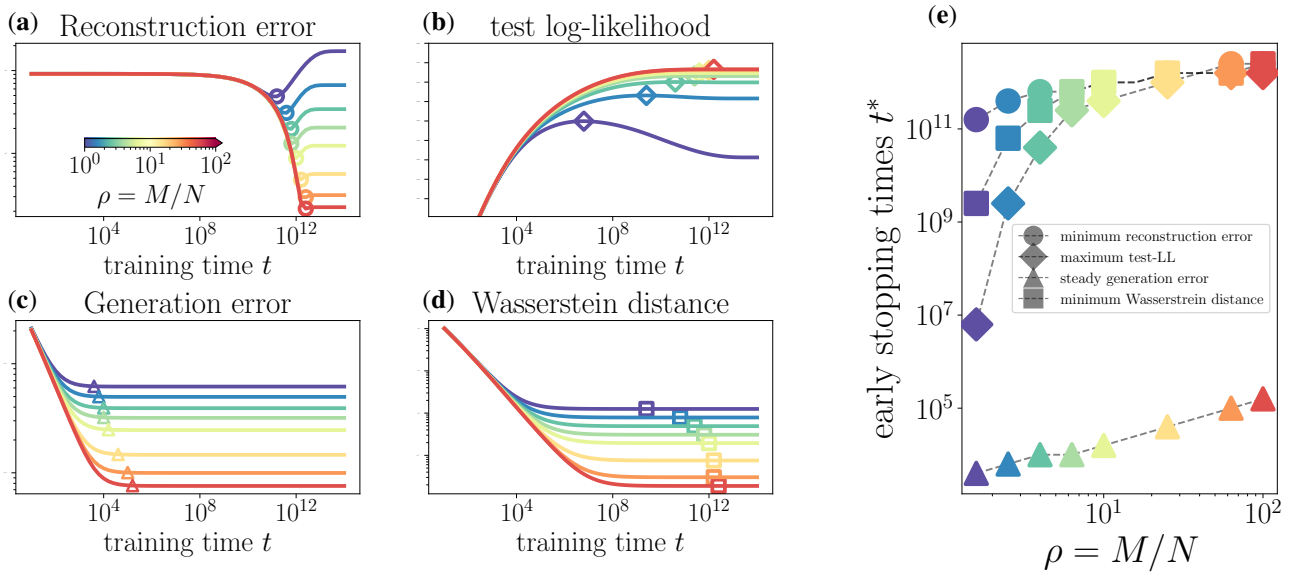

Figure 14. Same plots as in Fig. 12, this time obtained by training a GEBM starting from the eigenvalue spectrum of the empirical covariance matrix computed from the MNIST dataset, with a cutoff at 10^{-6} to filter out weak modes.

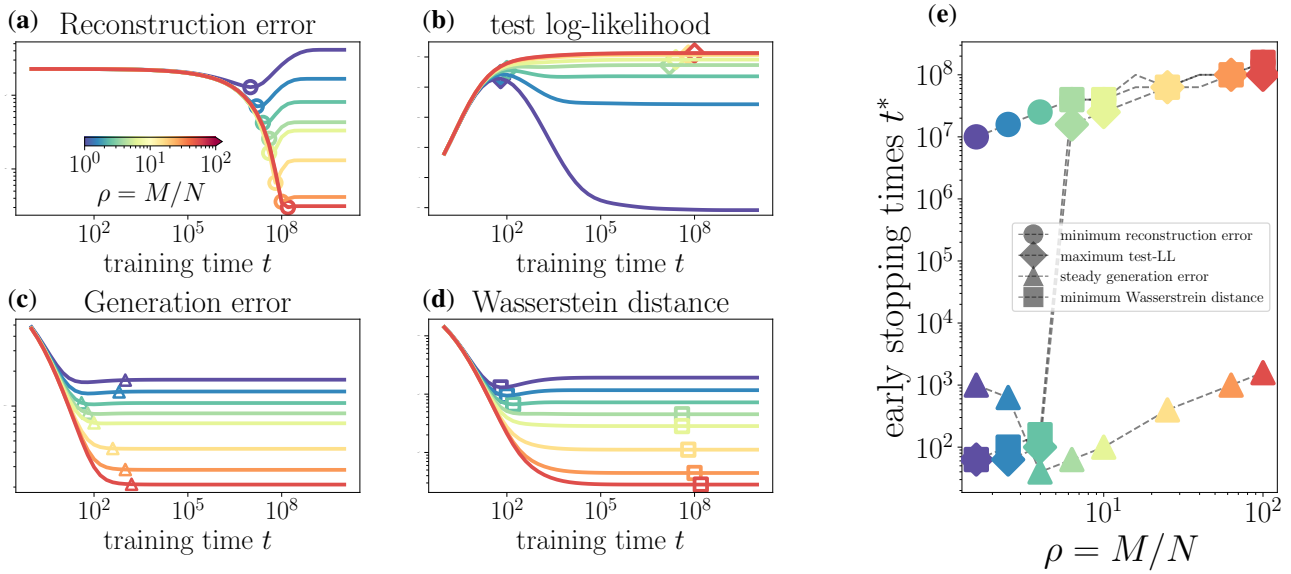

Figure 15. Same plots as in Fig. 12, this time obtained by training a GEBM starting from the eigenvalue spectrum of the empirical covariance matrix computed from the HGD dataset, with a cutoff at 10^{-12} to filter out weak modes.

G. Asymptotic analysis through Random Matrix Theory

G.1. General case

Various quantities appearing in the core of the manuscript are explicit function of the empirical covariance matrix $\widehat{\boldsymbol{C}}^M$ and as such are amenable to asymptotic analysis thanks to random matrix theory (RMT). These quantities are respectively the train, test energy (associated to the EBM), the coupling error and the LL (train and test). For the sake of clarity, we repeat them here:

$$E_{\text{train}} = \frac{1}{N}\text{Tr}[\boldsymbol{J}\widehat{\boldsymbol{C}}^M], \tag{29}$$

$$E_{\text{test}} = \frac{1}{N}\text{Tr}[\boldsymbol{J}\boldsymbol{C}^*], \tag{30}$$

$$\mathcal{E}_{\text{J}} \stackrel{\text{def}}{=} \frac{1}{N}\|\boldsymbol{J} - \boldsymbol{J}^*\|_F^2 \tag{31}$$

$$LL_{\text{train,test}} \stackrel{\text{def}}{=} \frac{1}{2N}\log\det[\boldsymbol{J}] - \frac{1}{2}E_{\text{train,test}}, \tag{32}$$

where $\boldsymbol{C}^* \stackrel{\text{def}}{=} \lim_{M\to\infty}\widehat{\boldsymbol{C}}^M$ is the *population matrix*, $\|\cdot\|_F^2$ the Frobenius norm while $\boldsymbol{J}$ is the estimation of the coupling matrix from the train samples $\boldsymbol{x}$ assumed to be of the form $\boldsymbol{x} = \boldsymbol{F}\boldsymbol{z}$, with $\mathbb{E}(\boldsymbol{z}\boldsymbol{z}^\top) = \mathbb{I}$, $\boldsymbol{F}\boldsymbol{F}^\top = \widehat{\boldsymbol{C}}^M$, $\tau = \|\boldsymbol{x}\|$ distributed w.r.t. some density $\sigma(\tau)$. Depending on the setting (dynamical, spectral $\widetilde{L}_1$ or L_2) $\boldsymbol{J}$ may appear in three different explicit functional form j_t, $j_\alpha^{(\widetilde{L}_1)}$ and $j_\alpha^{(\text{L}_2)}$ of $\widehat{\boldsymbol{C}}^M$. We have

$$j_t(x) = \frac{1}{x}\left(1 + W_0[-e^{-x^2t-1}]\right), \qquad \text{training dynamics,} \tag{33}$$

$$j_\alpha^{(\widetilde{L}_1)}, = \frac{\alpha}{1+\alpha x}, \qquad (L_1 \text{ (spectral) regularization}), \tag{34}$$

$$j_\alpha^{(\text{L}_2)}, = \frac{\alpha}{2}\left(\sqrt{x^2 + \frac{4}{\alpha}} - x\right), \qquad (L_2 \text{ regularization}). \tag{35}$$

A derivation of Eqs. (34)-(35) is given in Appendix H.1. The j_t corresponds to the situation where all eigenvalues J_α have the initial condition $J_\alpha(0) = 0$ and follow the time evolution of Eq. (6). Let us call generically j the functions given above.

Using the resolvant

$$\mathbf{G}^{(M)}(z) \stackrel{\text{def}}{=} \frac{1}{z\mathbb{I} - \widehat{\boldsymbol{C}}^M}\text{ç},$$

we can express the various quantities of interest with help of Cauchy integrals

$$E_{\text{train}} = \frac{1}{2i\pi}\oint_{\mathcal{C}} dz j(z)\text{Tr}\big[\mathbf{G}^{(M)}(z)\boldsymbol{C}^{(M)}\big],$$

$$E_{\text{test}} = \frac{1}{2i\pi}\oint_{\mathcal{C}} dz j(z)\text{Tr}\big[\mathbf{G}^{(M)}(z)\boldsymbol{C}^*\big],$$

$$\mathcal{E}_{\text{J}} = \text{Tr}\big[\boldsymbol{C}^{*-2}\big] + \frac{1}{2i\pi}\oint_{\mathcal{C}} dz\Big(j^2(z)\text{Tr}\big[\mathbf{G}^{(M)}(z)\big] - 2j(z)\text{Tr}\big[\mathbf{G}^{(M)}(z)\boldsymbol{C}^{*-1}\big]\Big),$$

$$LL_{\text{train,test}} = \frac{1}{2i\pi}\oint_{\mathcal{C}} dz\frac{\log[j(z)]}{2}\text{Tr}\big[\mathbf{G}^{(M)}(z)\big] - \frac{E_{\text{train,test}}}{2},$$

where $\mathcal{C}$ is a contour of integration around the real axis. Next, from RMT, in the proportional asymptotic limit $M, N \to \infty$ with fixed $M/N = \rho$, $\mathbf{G}^{(M)}(z)$ has a deterministic equivalent (Hachem et al., 2007) $\mathbf{G}$, defined as

$$\mathbf{G}(z) = \frac{1}{z\mathbb{I} - \Lambda(z)\boldsymbol{C}^*},$$

with $\boldsymbol{C}^*$ the population matrix and $\Lambda(z)$ is given implicitly by the following self-consistent equations (Marčenko & Pastur, 1967)

$$\Lambda(z) = \int \frac{\sigma(\tau)}{1 - \Gamma(z)\tau}, \tag{36}$$

$$\Gamma(z) = \frac{1}{\rho} \int \frac{\nu(dx)x}{z - \Lambda(z)x}, \tag{37}$$

with $\nu(dx)$ the spectral density of the population matrix and where

$$\Gamma \overset{\text{def}}{=} \lim_{\substack{N,M \to \infty \\ \frac{M}{N} = \rho}} \frac{\alpha}{M} \text{Tr}\big[\mathbf{G}^{(M)} \boldsymbol{C}^*\big]. \tag{38}$$

For sake of clarity we do not consider the Marchenko-Pastur equations in full-generality and actually assume the fluctuation of $\boldsymbol{z}$ to be negligible i.e. we take $\sigma(\tau) = \delta(\tau - 1)$. Letting $\bar{\nu}(dx)$ the asymptotic limit of the empirical spectrum in the proportional regime, its Stieltjes transform is given by the trace of the resolvent:

$$g(z) \overset{\text{def}}{=} \int \frac{\bar{\nu}(dx)}{z - x}.$$

Then the bulk spectrum is given by the Stieltjes transform

$$g(y + i\epsilon) = g_r(y) + i\pi \frac{\epsilon}{|\epsilon|} \bar{\nu}(y),$$

which rewrites (disregarding the pole at $z = 0$ for $\rho < 1$)

$$\bar{\nu}(y) = \frac{\rho \Lambda_i(y)}{\pi y} = \frac{\rho}{\pi y} \frac{\Gamma_i(y)}{\big[1 - \Gamma_r(y)\big]^2 + \Gamma_i(y)^2}. \tag{39}$$

Along the contour we integrate over $z = y + i\epsilon$ with ϵ infinitesimal. In the limit $\epsilon \to 0$, both Λ and Γ may acquire a finite imaginary part which we write as

$$\lim_{\epsilon \to 0^\pm} \Lambda(z) = \Lambda_r(y) \pm \Lambda_i(y)$$

$$\lim_{\epsilon \to 0^\pm} \Gamma(z) = \Gamma_r(y) \pm \Gamma_i(y).$$

In terms of these quantities we obtain the following equations for the train and test energies:

$$E_{\text{train}} = \frac{\rho}{\pi} \int_0^\infty dy j(y) \big[\Lambda_r(y)\Gamma_i(y) + \Lambda_i(y)\Gamma_r(y)\big],$$

$$E_{\text{test}} = \frac{\rho}{\pi} \int_0^\infty dy j(y)\Gamma_i(y) + \mathbb{1}_{\{\rho<1\}} j(0)c(\rho),$$

where $c(\rho)$ given implicitly by

$$\int \frac{\nu(dx)x}{x + c(\rho)} = \rho. \tag{40}$$

E_{train} may also be written

$$E_{\text{train}} = \int_0^\infty dy y j(y) \bar{\nu}(y),$$

with $\bar{\nu}(y)$ given in (39).

The coupling error takes the form for any $\rho > 0$

$$\mathcal{E}_{\text{J}} = \int_0^\infty \frac{\nu(dx)}{x^2} + \int_0^\infty \bar{\nu}(dy) j^2(y) - \frac{2}{\rho} \int_0^\infty \frac{\bar{\nu}(dy)}{y} \Big[(1 - \rho) + 2\rho\Lambda_r(y)\Big] j(y)$$

$$+ \mathbb{1}_{\{\rho<1\}} \Big[(1 - \rho) j^2(0) + 2j(0) \Big(\frac{1 - \rho}{c(\rho)} - \int_0^\infty \frac{\nu(dx)}{x} \Big) \Big],$$

but in practice we consider only the under-parameterized regime corresponding to $\rho > 1$.

G.2. Special case of spectral L_1 Regularization

The case corresponding to the form (34) can be treated more directly without use of Cauchy integrals. In that case, considering instead the resolvant

$$\mathbf{G}^{(M)} = \frac{1}{\mathbb{I} + \alpha \widehat{\boldsymbol{C}}^M},$$

with the inverse penalty

$$\alpha = \frac{1}{\lambda}$$

introduced here for convenience. This leads to the following form of the various quantities of interest

$$E_{\text{train}} = \frac{\alpha}{N} \text{Tr}[\mathbf{G}^{(M)} \widehat{\boldsymbol{C}}^M] \tag{41}$$

$$E_{\text{test}} = \frac{\alpha}{N} \text{Tr}[\mathbf{G}^{(M)} \boldsymbol{C}^*]. \tag{42}$$

$$E_{\text{couplings}} \stackrel{\text{def}}{=} \frac{1}{N} \|\alpha \mathbf{G}^{(M)} - \boldsymbol{J}\|_F^2 \tag{43}$$

$$LL_{\text{train,test}} = \frac{1}{2N} \text{Tr}[\log \alpha \mathbf{G}^{(M)}] - \frac{1}{2} E_{\text{train,test}}. \tag{44}$$

In the scaling limit we again have a deterministic equivalent (Hachem et al., 2007) of the resolvent of the form

$$\mathbf{G} = \frac{1}{\mathbb{I} + \Lambda \boldsymbol{C}^*}$$

where the fixed point equations now read ($\sigma(\tau) = \delta(\tau - 1)$)

$$\Gamma = \frac{\alpha}{\rho} \int \nu(dx) \frac{x}{1 + \Lambda x} \tag{45}$$

$$\Lambda = \frac{\alpha}{1 + \Gamma} \tag{46}$$

with again Γ given by (38). The expression for $E_{\text{train,test}}$ are straightforward in the scaling limit:

$$E_{\text{train}} = 1 - \int \frac{\nu(dx)}{1 + \Lambda x}$$

$$E_{\text{test}} = \frac{\Gamma}{\rho}.$$

Remarkably, thanks to a leave-one out argument there is a deterministic relationship between the train and test energy. For $s \in \mathcal{I}_{\text{train}}$, we have a leave-one out relation of the form

$$\mathbf{G}^{(M)} \boldsymbol{x}_s = \frac{\mathbf{G}^{(M)}_{\setminus s} \boldsymbol{x}_s}{1 + \frac{\alpha}{M} \boldsymbol{x}_s^t \mathbf{G}^{(M)}_{\setminus s} \boldsymbol{x}_s}$$

where $\mathbf{G}^{(M)}_{\setminus s}$ is the resolvant obtained after removing s from the train set. Assuming the samples to be of the form $\boldsymbol{x}_s = F \boldsymbol{z}_s$ with $FF^t = C$ with $\boldsymbol{z}_s = \mathcal{N}(0, M\mathbb{I})$ such that $\mathbb{E}(\boldsymbol{z}_s \boldsymbol{z}_s^t) = \mathbb{I}$, then for large N we have the concentration property

$$\boldsymbol{x}_s^t \mathbf{G}^{(M)}_{\setminus s} \boldsymbol{x}_s = \frac{1}{M} \text{Tr}\big[\mathbf{G}^{(M)} C\big] + \mathcal{O}\big(\frac{1}{\sqrt{M}}\big).$$

As a result for large N, M we immediately obtain

$$E_{\text{train}} = \frac{E_{\text{test}}}{1 + \frac{1}{\rho} E_{\text{test}}},$$

which can be reverted as

$$E_{\text{test}} = \frac{E_{\text{train}}}{1 - \frac{1}{\rho} E_{\text{train}}}. \tag{47}$$

Concerning the error on the couplings, we have

$$\begin{aligned} E_{\text{J}} &= \frac{1}{N} \text{Tr}\big[\big(\alpha \mathbf{G}^{(M)} - \boldsymbol{J}^*\big)^2\big] \\ &= \frac{1}{N}\Big(\alpha^2 \text{Tr}\big[\mathbf{G}^{(M)^2}\big] + \text{Tr}\big[\boldsymbol{J}^{*2}\big] - 2\alpha \text{Tr}\big[\mathbf{G}^{(M)} \boldsymbol{J}^*\big]\Big) \end{aligned}$$

From Ledoit-Péchet the last term simply reads (up to $\mathcal{O}\big(1/\sqrt{M}\big)$ corrections):

$$\frac{1}{M} \text{Tr}\big[\mathbf{G}^{(M)} \boldsymbol{J}^*\big] = \frac{1}{M} \text{Tr}\Big[\frac{\boldsymbol{J}^*}{\mathbb{I} + \Lambda \boldsymbol{C}^*}\Big] = \frac{1}{M} \text{Tr}\Big[\frac{\boldsymbol{J}^{*2}}{\boldsymbol{J}^* + \Lambda}\Big]$$

For the first term we use the following identity:

$$\mathbf{G}^{(M)^2} = \mathbf{G}^{(M)}) + \alpha \frac{d}{d\alpha} \mathbf{G}^{(M)}$$

As a result, asymptotically we have:

$$\begin{aligned} \text{Tr}\Big[\mathbf{G}^{(M)^2}\Big] &= \text{Tr}\Big[\frac{1}{\mathbb{I} + \Lambda \boldsymbol{C}^*}\Big] + \alpha \frac{d}{d\alpha} \text{Tr}\Big[\frac{1}{\mathbb{I} + \Lambda \boldsymbol{C}^*}\Big] \\ &= \big(1 - \Lambda'(\alpha)\big) \text{Tr}\Big[\frac{1}{\mathbb{I} + \Lambda \boldsymbol{C}^*}\Big] + \Lambda'(\alpha) \text{Tr}\Big[\frac{1}{(\mathbb{I} + \Lambda \boldsymbol{C}^*)^2}\Big] \end{aligned}$$

For this we need to compute $\Lambda'(\alpha)$ which can be done from the self-consistent equation (46,45):

$$\Lambda'(\alpha) = \frac{\Lambda^2}{\alpha^2} \frac{\rho}{\rho - Q[\Lambda]}$$

with

$$Q[\Lambda] = \frac{1}{M} \text{Tr}\Big[\frac{\Lambda^2 \boldsymbol{C}^{*2}}{(\mathbb{I} + \Lambda \boldsymbol{C}^*)^2}\Big]$$

Ultimately we obtain:

$$\mathcal{E}_J = \frac{1}{M} \text{Tr}\Big[\Big(\frac{\alpha}{1 + \Lambda \boldsymbol{C}^*} - \frac{1}{\boldsymbol{C}^*}\Big)^2\Big] + \frac{\alpha^2 (1 - \Lambda')}{M} \text{Tr}\Big[\frac{\Lambda \boldsymbol{C}^*}{(1 + \Lambda \boldsymbol{C}^*)^2}\Big],$$

So we have

$$\mathcal{E}_J = \int \nu(dx) \big[\frac{\alpha}{1 + \Lambda x} - \frac{1}{x}\big]^2 + \alpha^2 (1 - \Lambda') \int \nu(dx) \frac{\Lambda x}{(1 + \Lambda x)^2}$$

Finally, concerning $LL_{\text{train,test}}$ we don't see how to avoid the Cauchy integral, but the train-test relationship (47) has an important consequence, because it allows us to get a very precise estimation of the test likelihood when M becomes large:

$$LL_{\text{test}}(J) = \frac{1}{2} \log \det(J) - \frac{E_{\text{train}}}{1 - \frac{1}{\rho} E_{\text{train}}}$$

as long as J is the function (34) of $C^{(M)}$. For general EBM models we have a LL of the form

$$LL_{\text{train,test}}[J] = -\log Z[J] - E_{\text{train,test}}[J]$$

so by analogy with GCV, it is not excluded that we can use this train-test relation in practice.

H. Details on data-correction protocols

In this section, we analyze the effect of different ways to improve the estimation of the covariance matrix's eigenvalues in order to avoid or diminish the effect of overfitting during the training dynamics.

H.1. Training dynamics with regularization prior for finite N

We first discuss what happens to the training in the presence of a regularization. We employ two regularization protocols: a standard L_2-norm, and a projected L_1-norm. The choice of the second regularization is justified because it allows to have a maximum-a-posteriori coupling matrix which commutes with the original covariance matrix $\widehat{\boldsymbol{C}}^M$, as it happens in the absence of regularization, thus facilitating the asymptotic analysis through RMT discussed in Appendix G.

H.1.1. L_2 REGULARIZATION

The log-posterior now reads

$$\frac{1}{M}\log p\left(\boldsymbol{J}\mid\mathcal{D}\right)=\mathcal{L}_{\mathcal{D}}\left(\boldsymbol{J}\right)-\frac{\lambda}{4}\mathrm{Tr}\left(\boldsymbol{J}^2\right) \tag{48}$$

where λ is the reguavlarization strength. The derivative w.r.t. the parameters now reads

$$\frac{1}{M}\frac{\partial\log p\left(\boldsymbol{J}\mid\mathcal{D}\right)}{\partial J_{ij}}=\left[-\hat{C}_{ij}^M+\left(\boldsymbol{J}^{-1}\right)_{ij}-\lambda J_{ij}\right] \tag{49}$$

Notice that the new term commutes with the second one ($\boldsymbol{J}$ and $\boldsymbol{J}^{-1}$ are diagonal in the same basis), so even in this case the maximum-a-posteriori matrix $\widehat{\boldsymbol{J}}^{\mathrm{MAP}}$ will share the same basis as $\widehat{\boldsymbol{C}}^M$, as it happens in the absence of regularization. Therefore, we can apply the same reasoning discussed in the main text and project the log-posterior's gradient on the basis of $\boldsymbol{J}$. The evolution equation of each eigenvalues reads:

$$\tau\frac{\mathrm{d}J_\alpha}{\mathrm{d}t}=\frac{1}{J_\alpha}-\hat{c}_\alpha^M-\lambda J_\alpha, \tag{50}$$

Although there exist no closed expression for the full time-dependent solution of Eq. (50), it is possible at least to compute analytically its fixed point:

$$J_\alpha^{(\infty)-L_2}\left(\lambda\right)=\frac{1}{2\lambda}\left[-\hat{c}_\alpha^M+\sqrt{\left(\hat{c}_\alpha^M\right)^2+4\lambda}\right] \tag{51}$$

The full coupling matrix corresponding to the above fixed point is finally computed projecting back Eqs. (51) onto the eigenbasis of $\widehat{\boldsymbol{C}}^M$, that is $\boldsymbol{J}^{(\infty)-L_2}\left(\lambda\right)=\sum_\alpha J_\alpha^{(\infty)-L_2}\left(\lambda\right)\boldsymbol{u}_\alpha^M\boldsymbol{u}_\alpha^{M\top}$.

H.1.2. SPECTRAL $\widetilde{L}_1$-NORM

This regularization schemes utilizes a L_1-norm but on the projected basis of the coupling matrix $\boldsymbol{J}$. This construction still allows to employ a similar formula to Eq. (12) to describe the evolution of eigenvalues, each one independently on the others. The original differential equation describing the evolution of J_α is now modified as

$$\frac{1}{\gamma}\frac{\mathrm{d}J_\alpha}{\mathrm{d}t}=\frac{1}{J_\alpha}-\hat{c}_\alpha^M-\lambda, \tag{52}$$

whose fixed point reads

$$J_\alpha^{(\infty)-\widetilde{L}_1}\left(\lambda\right)=\frac{1}{\hat{c}_\alpha^M+\lambda} \tag{53}$$

Note that equations (51)- (53) just derived are the same ones as Eqs. (35)-(34) in Appendix 6, respectively.

H.1.3. RESULTS ON THE EFFECT OF REGULARIZATION

We can check the performances of either type of regularization by looking at the training fixed point, and at how it modifies the quality of the inferred model. In Fig. 16-(**a**), we show the reconstruction error $\mathcal{E}_{\mathrm{J}}$ computed between the ground truth and the inferred model in the presence of a regularization prior with strength λ, both for the L_2-norm (solid lines) and for the spectral-L_1 norm (dashed lines), for a given value of number of samples: here we have $\rho=M/N=1.5$. Comparison is

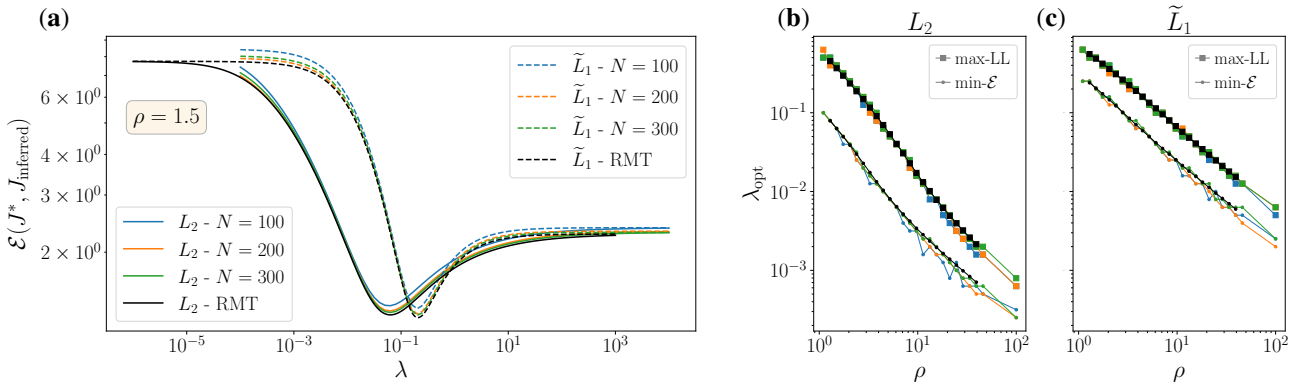

Figure 16. Effect of the regularization priors on the inferred model's quality. (**a**): the plot shows the reconstruction error $\mathcal{E}_J$ computed between the ground truth and the inferred model in the presence of a regularization prior with strength λ. Solid lines refer to the L_2 norm, while dashed lines to the spectral-L_1 norm, both discussed in Section H.1. For each prior, we compare finite-size results (colored lines) and RMT asymptotic estimations. (**b**)-(**c**): we show the values of the regularization strength λ_{opt} that achieves optimal reconstruction of the model (lines with scatter points), and the optimal value of the regularization that maximizes the test log-likelihood (lines with scatter diamonds). Panels (**b**) (resp. (**c**)) refers to the optimal values when using the L_2-norm (resp. the spectral L_1-norm). Note that (**b**)-(**c**) are on the same y-scale. Settings are the same as in Fig. 4.

shown between finite-size trainings (colored lines) and RMT estimation (black lines). All the quantities are plotted versus the regularization strength λ: we can observe how there is clear non-monotonic behavior with a minimum developing at a certain λ_{opt}. As one might expect the optima value differs between the two regularization priors, i.e. $\lambda_{\text{opt}}^{L_2} \neq \lambda_{\text{opt}}^{\widetilde{L}_1}$. Results on both priors are shown in the same panel to highlight that the two regularization schemes have qualitatively the same effect on the final reconstruction error: that is, the quality of the inferred model at the optimal value is the same for both regularizations. Panels (**b**)-(**c**) show instead the optimal value of λ computed either by minimizing the reconstruction error (shown with points) or by maximizing the test log-likelihood (diamonds). Panel (**b**) is actually a repetition of Fig. 4-(**d**) and refers to the L_2-prior, while (**c**) refers to the L_1 spectral prior. Again, we can observe a similar behavior of the two norms, with the only difference that $\lambda_{\text{opt}}^{L_2} < \lambda_{\text{opt}}^{\widetilde{L}_1}$ independently on the chosen criterion. Finally, all the optimal values go to 0 when $\rho \to \infty$, as expected.

What is the effect of the regularization on the training dynamics? Considering that the standard training has a non-monotonic behavior w.r.t. the training time, we would expect that, since the regularization strongly improves on the models' quality w.r.t. the standard case (at least at the optimal optimizing regularization strength λ_{opt}), such a non-monotonic behavior is diminished. This is indeed the case, as shown by Fig. 17-(**a**), displaying different training curves for different regularization strengths: the closest the regularization to its optimal value (highlighted in red), the smoother the model's quality is w.r.t. training time. At the optimal point the model's quality is completely non-monotonic and approaches the fixed point at the same reconstruction error as the minimum w.r.t. time.

Actually, due to the simplicity of the GEBM it is even possible to interpret the L_2 regularization as a shrinkage correction protocol. Consider indeed the training fixed point given by Eq. (50). We stress again that the maximum-posterior matrix $\boldsymbol{J}$ has the same basis decomposition as the empirical covariance matrix, because the regularization term commutes with the other two terms in Eq. (50). By the dualism between covariance matrix and coupling matrix in the Gaussian EBM, we can think at the reciprocal values of Eq. (50) as eigenvalues of a corrected covariance matrix w.r.t. $\widehat{\boldsymbol{C}}^M$, depending on λ. We can therefore define another eigenvalue-corrected covariance matrix, by using the analytic fixed point on the training dynamics obtained through the regularization:

$$\widehat{\boldsymbol{C}}^M_{\text{val}-L_2(\lambda)} = \sum_\alpha \frac{1}{J_\alpha^{(\infty)-L_2}(\lambda)} \boldsymbol{u}_\alpha^M \boldsymbol{u}_\alpha^{M\top} \tag{54}$$

By definition, the training fixed point obtained using i) a training dynamics with the un-touched empirical covariance matrix $\widehat{\boldsymbol{C}}^M$ plus the regularization term or ii) a regularization-free dynamics using matrix (54) are the same. Fig. 17-(**a**) shows indeed how the regularization modifies the eigenvalues of $\widehat{\boldsymbol{C}}^M$ when interpreting the fixed point of the training dynamics

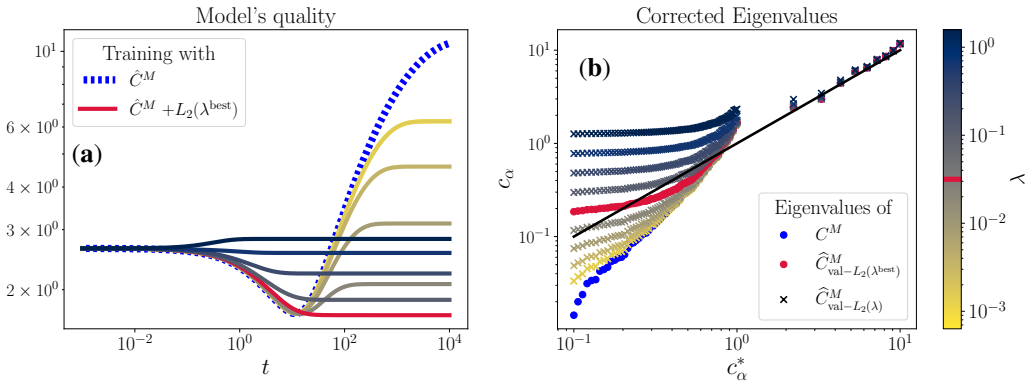

Figure 17. Effect of different L_2-norm regularization strengths λ on the GEBM's learning dynamics. Panel (**a**) shows the reconstruction error vs time. The dotted blue line corresponds to the standard training over $\widehat{\boldsymbol{C}}^M$. All the other full lines correspond to a training with a certain value of regularization strength λ, obtained by numerically solving Eq. (50) for all modes. The regularization strength λ increases from yellowish colors to blueish (see the colorbar at the right). The curve corresponding to the optimal regularization that minimizes the reconstruction error after training is highlighted in red. Panel (**b**): plot of the equivalent covariances modes corrected by the regularization. For each curve, we scatter plot these values Eq. (54) against the population eigenvalues. Here we set $\rho = 1.66$.

as a shrinkage correction. Each set of points shows the quantities $1/J_\alpha^{(\infty)-L_2}(\lambda)$ (i.e. the eigenvalues of (54)) vs the population ones c^*_α. It is intuitive to notice that the optimal regularization (red points) is the one that makes such corrected eigenvalues as close as possible to the population ones. This entire reasoning holds analogously with the spectral L_1 norm.

H.2. Empirical shrinkage correction through modes fitting

A simple way to perform a heuristic shrinkage correction is to down-sample the empirical covariance matrix and estimate the asymptotic eigenvalues through a fitting procedure. Starting from the available dataset with M samples - whose covariance matrix is $\widehat{\boldsymbol{C}}^M$ - we can randomly extract subsets of $N < m < M$ samples and estimate the eigenvalues of the size-reduced covariance matrices $\widehat{\boldsymbol{C}}^m$. Every time a down-sampling procedure of this kind is performed, both the eigenvalues and the eigenvectors will be different from the original $\widehat{\boldsymbol{C}}^M$; however, we here suppose to account for the eigenvalues, keeping the basis fixed to the one of $\widehat{\boldsymbol{C}}^M$. After applying this computation to different values of $N < m < M$, for each eigenvalue α we can fit the resulting data $\{\hat{c}^m_\alpha\}_{m\in(N;M]}$ according to

$$\hat{c}_\alpha(m) = \frac{1}{m^\nu} A^M_\alpha + B^M_\alpha \tag{55}$$

with ν being an exponent of choice. The coefficients B^M_α will represent the asymptotic estimate of the α-th eigenvalue of the population matrix, c^*_α, corresponding to the $m \to \infty$ extrapolation. After fitting each mode separately, we can construct an eigenvalue-corrected covariance matrix as

$$\widehat{\boldsymbol{C}}_{\text{val}-\text{fit}} = \sum_\alpha B^M_\alpha \boldsymbol{u}^M_\alpha \boldsymbol{u}^{M\top}_\alpha \tag{56}$$

This procedure can in principle be generalized e.g by using a combination of powers in the fitting function, although in this work we only restricted to the functional form (56); secondly, the estimation of the fitting coefficients can be improved by collecting mean values of eigenvalues $\{\hat{c}^m_\alpha\}_{m\in(N;M]}$ by evaluating the down-sampled covariance matrix $\widehat{\boldsymbol{C}}^m$ multiple times with different subsets of the M original data. In the experiments presented in the main text for the GEBM (orange line in Fig. 5-(**b**)), we used a simple linear fitting in $1/m$ ($\nu = 1$) and 10 random resampling for each value of m, from which the mean eigenvalue is extracted to perform the fit. This linear scaling (in $1/m$) for the finite-size fluctuation of the eigenvalues in the GEBM is also justified by theoretical evidence (see e.g. (Ledoit & Péché, 2011)). In the experiments for the Ising-BM instead, (orange line in Fig. 6-(**c**)) we find that best reconstruction is achieved with $\nu = 1/2$, while a linear fitting has always very bad performances. Also in this case we used 10 resampling steps. An example of such a fitting procedure is shown in Figure 18 for both the GEBM (in (**a**)) and for the BM (in (**b**)), in each case for a given value of M.

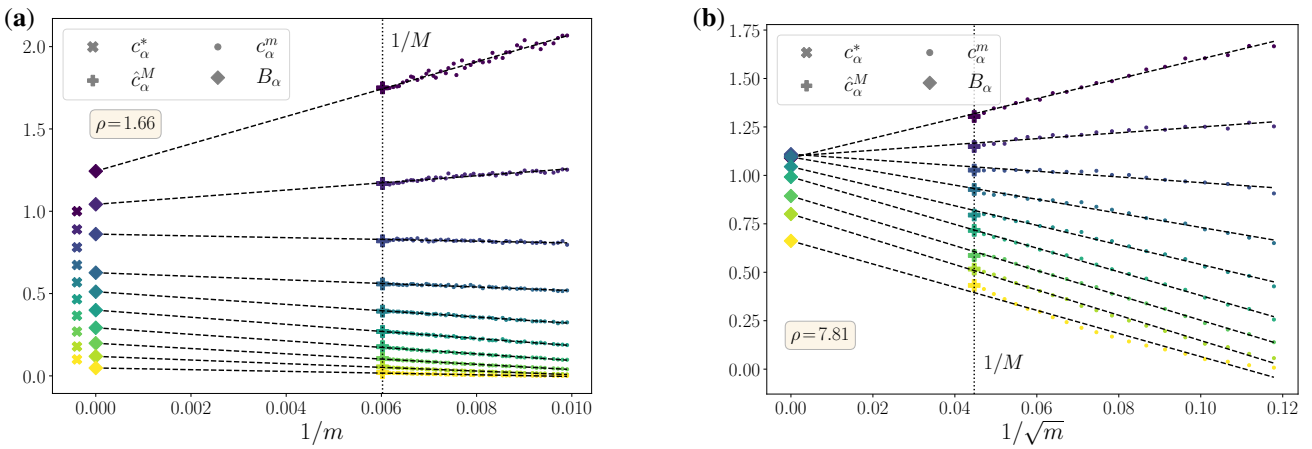

Figure 18. **Examples of eigenmode fitting procedures for the GEBM and the Ising-BM.** Each panel illustrates the procedure used to fit the eigenmodes of the covariance matrix $\widehat{\boldsymbol{C}}^M$ by downsampling to $m < M$ samples, in order to extrapolate their behavior as $m \to \infty$, following Eq. (55). Panel (**a**) refers to the GEBM used in the main text (e.g., Fig. 5). For a fixed value of M such that $\rho = M/N = 1.66$, we show a subset of eigenvalues $\{\hat{c}^M_\alpha\}$ (denoted by + markers), along with their downsampled counterparts $\{c^m_\alpha\}$ for several values of $m < M$ (small circles). These downsampled eigenvalues are obtained by randomly selecting m samples from the full dataset and computing the eigenvalues of the resulting covariance matrix, averaged over 10 independent instances. Dashed black lines correspond to fits using Eq. (55) with $\nu = 1$, and colored diamonds indicate the extrapolated intercepts B^M_α, i.e., the estimated eigenvalues at $m \to \infty$. Crosses mark the population eigenvalues $\{c^*_\alpha\}$, showing that the fitted extrapolations are significantly closer to the true population spectrum than the empirical eigenvalues obtained from M samples. Panel (**b**) shows the analogous procedure for the Ising-BM, using $\rho = 7.81$. In this case, the best fits are obtained with $\nu = 1/2$, and the horizontal axis is accordingly rescaled as $1/\sqrt{m}$.

I. Eigendecomposition of training dynamics on Boltzmann Machine

We consider a Ising-like Boltzmann Machine for the inference of binary-valued data. The probability of a configuration $\boldsymbol{x}$ where $x_i \in [-1, 1]$ at given parameters is expressed as

$$p\left(\boldsymbol{x} \mid \boldsymbol{J}, \boldsymbol{h}\right) = \frac{1}{Z} e^{\sum_{i<j} J_{ij} x_i x_j + \sum_i x_i h_i}. \tag{57}$$

We suppose to generate equilibrium configurations from a known model with $\boldsymbol{\theta}^* = (\boldsymbol{J}^*, \boldsymbol{h}^*)$ (eventually these parameters are rescaled by an external factor β that plays the role of an external inverse temperature), and we want to infer back the original model through a likelihood maximization procedure. The LL of a certain set of parameters $\boldsymbol{\theta} = (\boldsymbol{J}, \boldsymbol{h})$ is given by

$$\mathcal{L}_{\mathcal{D}}\left(\boldsymbol{J}, \boldsymbol{h}\right) = \frac{1}{M} \sum_{\mu=1}^{M} \log p\left(\boldsymbol{x}^\mu \mid \boldsymbol{J}, \boldsymbol{h}\right) = \sum_{i<j} J_{ij} \mathbb{E}_{\mathcal{D}}\left[x_i x_j\right] + \sum_i h_i \mathbb{E}_{\mathcal{D}}\left[x_i\right] - \log Z \tag{58}$$

where $\mathbb{E}_{\mathcal{D}}\left[\cdot\right]$ denotes the average w.r.t. the dataset $\mathcal{D} = \{\boldsymbol{x}_\mu\}_{\mu=1}^M$. In what follows, for the analytic treatment of the training dynamics in the ML procedure we neglect the problem of learning the external fields h_i. This assumption is consistent with a scenario where the data have null magnetizations. Therefore, the gradient of the LL w.r.t. the couplings J_{ij} reads

$$\frac{\partial \mathcal{L}_{\mathcal{D}}}{\partial J_{ij}} = \mathbb{E}_{\mathcal{D}}\left[x_i x_j\right] - \mathbb{E}_{\boldsymbol{J}}\left[x_i x_j\right] = \widehat{\boldsymbol{C}}^M - \mathbb{E}_{\boldsymbol{J}}\left[x_i x_j\right], \tag{59}$$

Then, the couplings are updated as

$$J_{ij}(t+1) \leftarrow J_{ij}(t) + \gamma \frac{\partial \mathcal{L}_{\mathcal{D}}(t)}{\partial J_{ij}}, \tag{60}$$

where γ is the learning rate. From here, we can assume an ideal training with an infinitesimal learning rate to recast the evolution equation of the matrix $\boldsymbol{J}$ in time in the following matrix form

$$\tau \frac{dJ_{ij}}{dt} = \left.\frac{\partial \mathcal{L}_{\mathcal{D}}}{\partial J_{ij}}\right|_{J(t)} \qquad \Longrightarrow \qquad \tau \frac{d\boldsymbol{J}}{dt} = \widehat{\boldsymbol{C}}^M - \left\langle \boldsymbol{x}\boldsymbol{x}^\top \right\rangle_{\boldsymbol{J}}. \tag{61}$$

where $\mathbb{E}_{\boldsymbol{J}}[\cdot]$ denotes the average w.r.t. to model (57), and $\tau = 1/\gamma$. Note that, since we assumed to neglect local magnetizations, the r.h.s (61) computed in its diagonal entries is 0. This is consistent with the fact that self-couplings J_{ii} do not evolve in time, because they correspond to constant energy terms in the energy function. In order to make the models' correlation $\mathbb{E}_{\boldsymbol{J}}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]$ analytically treatable, we implement now a mean-field approximation. We can exploit the following exact expression (self-consistent) for the correlator (Suzuki & Kubo, 1968), which we further expand for high temperatures

$$\mathbb{E}_{\boldsymbol{J}}[x_i x_j] = \delta_{ij} + (1-\delta_{ij})\mathbb{E}_{\boldsymbol{J}}\left[x_i \tanh \sum_k J_{jk}x_k\right] \approx \delta_{ij} + (1-\delta_{ij})\sum_k J_{jk}\mathbb{E}_{\boldsymbol{J}}[x_i x_k], \tag{62}$$

Then, in matrix form ($C_{ij} = \mathbb{E}_{\boldsymbol{J}}[x_i x_j]$) we get

$$\boldsymbol{C} = \mathbb{I}_N + \boldsymbol{J}\boldsymbol{C} - \text{diag}\left[\boldsymbol{J}\boldsymbol{C}\right] \tag{63}$$

where $\mathbb{I}_N$ is the identity matrix of size N, and the operator diag $[\mathcal{M}]$ extracts the diagonal part of a matrix $\mathcal{M}$. The last term is introduced to correct the wrong estimation of the diagonal entries of $\boldsymbol{C}$, which should be equal to 1. The problem is that the above equation does not admit a simple analytical solution for the model's correlation matrix $\boldsymbol{C}$, which was the original goal. Instead, is typical implemented in the literature is the following expression of the linear response correlations

$$\boldsymbol{C} = f(\boldsymbol{J}) = (\mathbb{I}_N - \boldsymbol{J})^{-1} \tag{64}$$

which gives a reliable estimate of the correlation matrix and is at the core of well-studied mean-field like expression for the inferred couplings (Kappen & Rodríguez, 1998; Ricci-Tersenghi, 2012). However, the diagonal entries of (64) are not in general equal to 1. Normally, this is not an issue because one is interested in correlations for $i \neq j$ (i.e. off-diagonal entries). However, in our approach such a diagonal mismatch creates a non-null gradient on the diagonal entries of $\boldsymbol{J}$. Indeed, by plugging Eq. (64) into the LL's gradient, we get

$$\tau\frac{d\boldsymbol{J}}{dt} = \widehat{\boldsymbol{C}}^M - (\mathbb{I}_N - \boldsymbol{J})^{-1} \tag{65}$$

Now, the diagonal part of the r.h.s. of Eq. (65) not null anymore. In order to circumvent this additional issue, a possible solution could be to modify the gradient by remove the diagonal terms - which would results in an nonphysical evolution of the self-couplings - "by hand" :

$$\tau\frac{d\boldsymbol{J}}{dt} = \widehat{\boldsymbol{C}}^M - (\mathbb{I}_N - \boldsymbol{J})^{-1} + \text{diag}\left[\mathbb{I}_N - (\mathbb{I}_N - \boldsymbol{J})^{-1}\right] \tag{66}$$

The last term in the above expression correctly fixes the diagonal problem for the matrix $\boldsymbol{J}$ in the gradient. This strategy is similar to what carried out in (Fanthomme et al., 2022) where the authors impose a add a spherical constraint on the gradient in the form of a Lagrange multiplier. However, this leads to a complicated expression even for the training fixed point because the evolution of all the eigenvalues is now coupled. For this reason, since here we are interested in the dynamics of training, adding this constraint would result into a system of coupled differential equations for the eigenvalues, which has computationally the same complexity of the original problem, so there would be no gain in that. The simplest strategy is therefore to use the approximate expression for the correlator and avoid adding the constraint, so to use the gradient (65) as it is. Although it might seem a crude approximation, it still allows us to decompose the dynamics in the same way as for the GEBM. Before going on, it is worth noticing that the diagonal matching problem is at the core of some refined mean-field like approximations for binary (Ising-like) maximum-entropy models which exploit e.g. iterative diagonal consistency tricks (see e.g (Yasuda & Tanaka, 2013; Kiwata, 2014)). Therefore, as explained in Sec. 4 for the GEBM we can project the log-likelihood's gradient onto the eigenbasis of $\boldsymbol{J}$, leading to an expression for the rotation of eigenvectors (same one as in (5)) and another one for the evolution of its eigenvalues, which is given below:

$$\tau\frac{\mathrm{d}J_\alpha}{\mathrm{d}t} = \hat{c}^M_{\alpha\alpha} - \frac{1}{\mathbb{I}_N - J_\alpha} \tag{67}$$

which is the same equation shown in the main text (Eq. (11)). As explained in the main text, the solution of the above equation describes an independent evolution of eigenvalues which is not quantitative accurate with respect to the numerical results, but still captures the qualitative trend: in particular, it perfectly describes the separation of timescales in terms of PCA's modes during the training dynamics, an effect which is observed also in the numerical results (see panels (**b**) in Fig. 6).

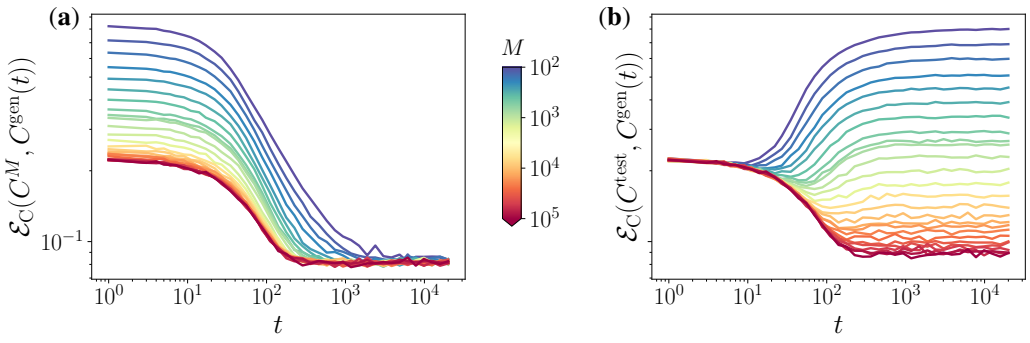

Figure 19. Supplementary results on the Boltzmann Machine for the inverse Ising problem. The model, dataset and training setting are the same as in Fig. 6. (**a**): we show the error between the covariance matrix of generated configurations from the model (along the training trajectory) and the covariance matrix of the training set $\widehat{\boldsymbol{C}}^M$, so $\mathcal{E}_{\widehat{\boldsymbol{C}}^M} = \left\| \widehat{\boldsymbol{C}}^M - \boldsymbol{C}^{\text{gen}}(t) \right\|_{\text{F}}$. (**b**): plot of the generation error computed w.r.t. a test set, again between the covariance matrices, i.e. $\mathcal{E}_{\boldsymbol{C}^{\text{test}}} = \left\| \boldsymbol{C}^{\text{test}} - \boldsymbol{C}^{\text{gen}}(t) \right\|_{\text{F}}$. Both quantities are plotted versus training time (number of updates) for different values of M shown in the colorbar. The learning rate is set to $\gamma = 10^{-2}$.

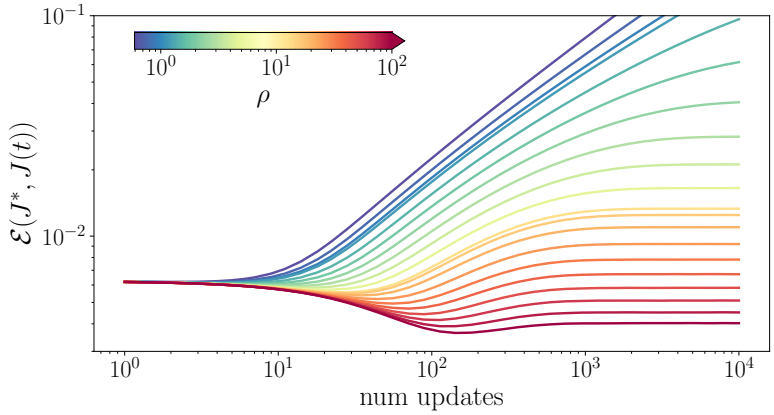

Figure 20. Supplementary results on the Boltzmann Machine for the inverse Ising problem. The model, dataset and training setting are the same as in Fig. 6, except for the system size which here is equal to a lattice size of $L = 32$, so that $N = L^2 = 1024$. We show the Frobenius norm of the error between the true model and the trained one versus training time (number of updates) for different values of $\rho = M/N$ shown in the colorbar. The learning rate is set to $\gamma = 10^{-2}$.