A Theoretical Framework For Overfitting In Energy-based Modeling

Giovanni Catania, Aurélien Decelle, Cyril Furtlehner, Beatriz Seoane

Universidad Complutense Madrid (ES)

International Conference on Machine Learning, 2025

Motivation

Study the impact of the amount data in the training of Energy-Based Models (EBMs) and how overfitting emerges when data is limited

In generative models, overfitting occurs when the model “memorizes” the training data instead of learning the underlying data distribution.

- poor diversity in generated samples, lack of variability present in real data

- model learns specific noise-dominated information in the data

Energy based models (EBMs) in generative AI

EBMs encode the empirical distribution of a dataset into a Boltzmann distribution with a given Energy function

Data — Model

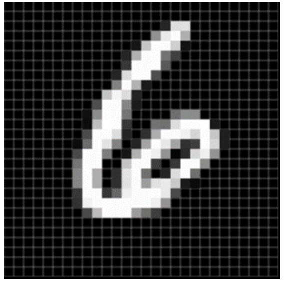
$$p_{\text{data}}(\boldsymbol{x}) \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{\mathcal{Z}_{\boldsymbol{\theta}}} e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}$$

Boltzmann distribution

rooted in statistical physics

Dataset (e.g. MNIST)

Datum $\boldsymbol{x}$



$\boldsymbol{\theta}$: vector of parameters to be trained

Training: log-likelihood maximization

Used for generative purposes and for interpretability of the effective model

Theoretical analysis of overfitting in EBMs

- Use a simple model (analytically solvable) for a synthetic experiment:

- Track the quality of the inferred model as a function of the number of samples

↓

Gaussian Model

Gaussian Energy-based Model (GEBM)

True model $\quad \boldsymbol{x} \in \mathbb{R}$

$$E_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{2}\sum_{ij} J_{ij} x_i x_j \qquad \boldsymbol{J} = (\boldsymbol{C}^*)^{-1}$$

Population covariance matrix

Sample $\boldsymbol{M}$ configurations from

$$^*p_{\boldsymbol{J}}(\boldsymbol{x}) = \frac{1}{\mathcal{Z}_{\boldsymbol{J}}} e^{-E_{\boldsymbol{J}}(\boldsymbol{x})}$$

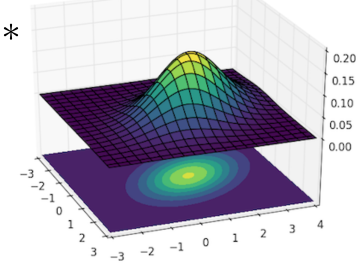
$$\widehat{\boldsymbol{C}}^M = \frac{1}{M}\sum_{\mu=1}^{M} \boldsymbol{x}_\mu \boldsymbol{x}_\mu^T$$

Empirical covariance matrix

Infer back the model from these $\boldsymbol{M}$ samples

Multivariate Gaussian

*



$$\widehat{\boldsymbol{C}}^M \xrightarrow{M \to \infty} \boldsymbol{C}^*$$

$M \equiv$ Number of data

Gaussian Energy-based Model (GEBM)

Why Gaussian Model?
Analytically solvable, both Maximum likelihood estimator and (most importantly) the training dynamics

Max-likelihood (ML) estimator

$$\boldsymbol{J}^{\mathrm{ML}} = \left(\widehat{\boldsymbol{C}}^{M}\right)^{-1}$$

same eigenvector basis

$$\boldsymbol{J}(t) = \sum_{\alpha} J_{\alpha}(t)\, \boldsymbol{u}_{\alpha}\boldsymbol{u}_{\alpha}^{\top}$$

Study training dynamics (likelihood maximization)
by projecting on eigenvector basis
$\rightarrow$ each eigenvalue of $\boldsymbol{J}$ now evolves independently on the others

Separation of learning timescales

Eigenvalue of

Eigenvalue of $\boldsymbol{C}^M$

$$\frac{\mathrm{d}J_\alpha}{\mathrm{d}t} = -c_\alpha^M + \frac{1}{J_\alpha}$$

Separation of time-scales

Modes corresponding to stronger correlations are learnt faster

$$J_\alpha(t) = \frac{1}{c_\alpha^M} + \frac{1}{c_\alpha^M} W\left(\text{const } e^{-\left(c_\alpha^M\right)^2 t}\right)$$

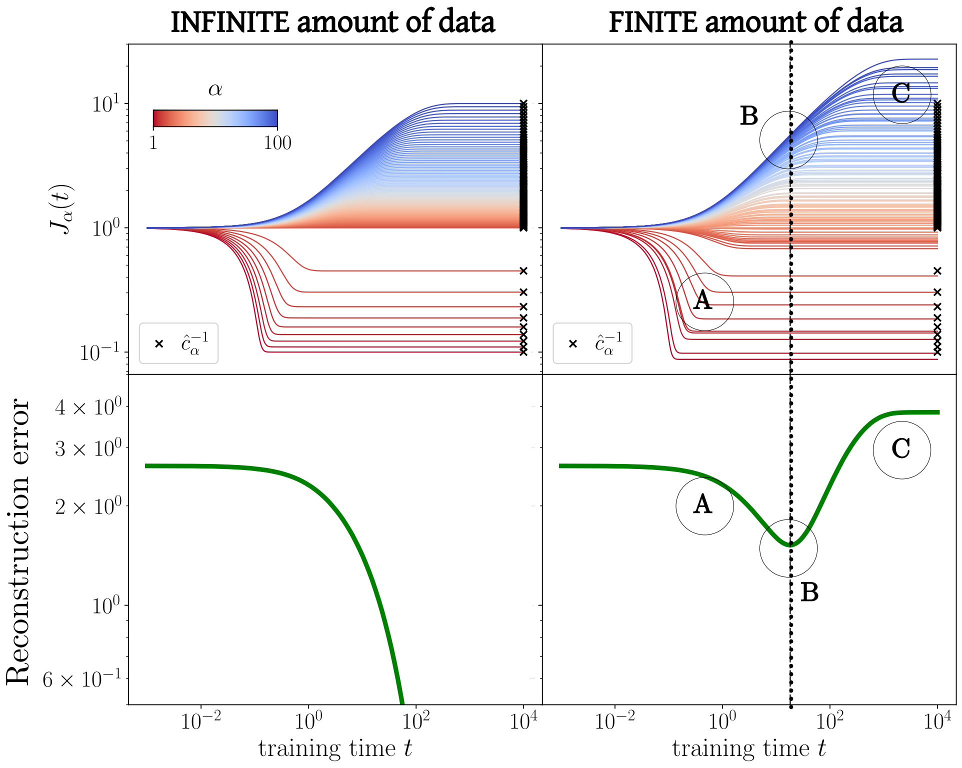
$c_\alpha^M =$ Modes of PCA decomposition

1. Model learns information sequentially way* stronger PCA direction first, then weaker ones

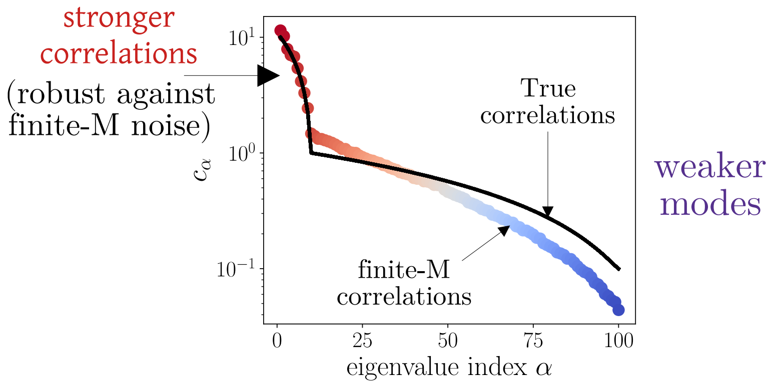
Learning time-scale $\propto \frac{1}{(c_\alpha^M)^2}$

2. Strong/weak PCA modes have very different fluctuations w.r.t. the number of data

Eigenvalues' evolution



$$\mathcal{E}_\mathrm{J} = \left\|\boldsymbol{J}^{\mathrm{true}} - \boldsymbol{J}(t)\right\|$$



A) Eigenvalues of stronger modes are the first to converge
→ error decreases

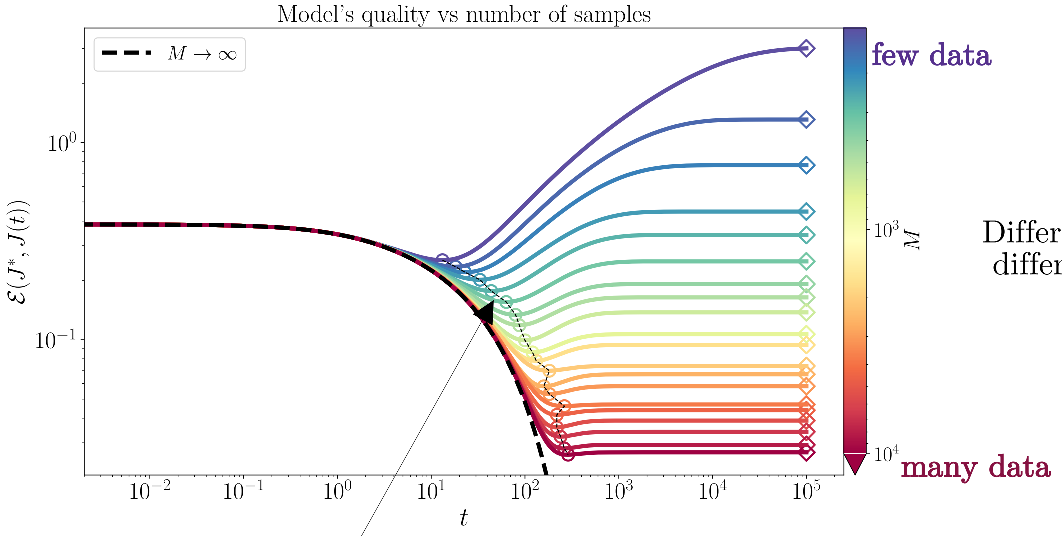
B) weaker modes are starting to be learnt
→ error decreases (up to minimum)
eigenvalues are closer to the true value than to the fixed point!

C) weaker modes converge to the fixed point,
error increases after minimum is reached
→ error dominated by fluctuations of weaker correlations due to the low number of samples

Early stopping points in training dynamics

Non-monotonic behavior (w.r.t. training time) of discrepancy between true and inferred model

Models inferred with few training data are **worse** at fixed point than during training

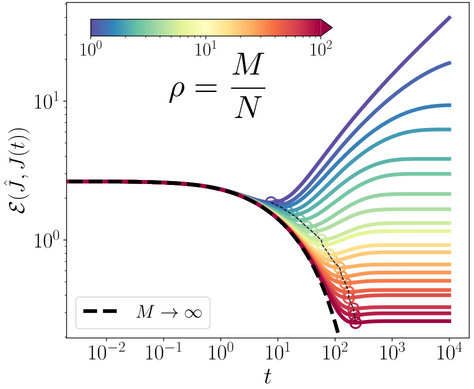


Different trainings for different sample sizes

early-stopping point at which best inference is achieved, after which model starts to overfit

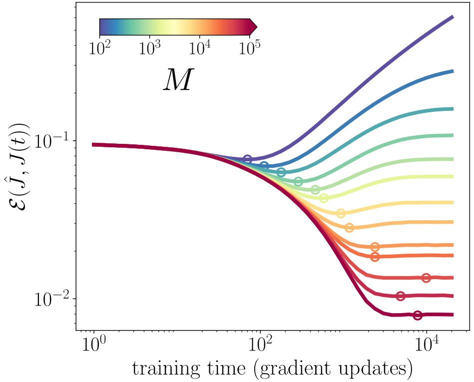
Different EBMs, same phenomenology

Gaussian Model (GEBM)



Boltzmann Machine Inverse Ising

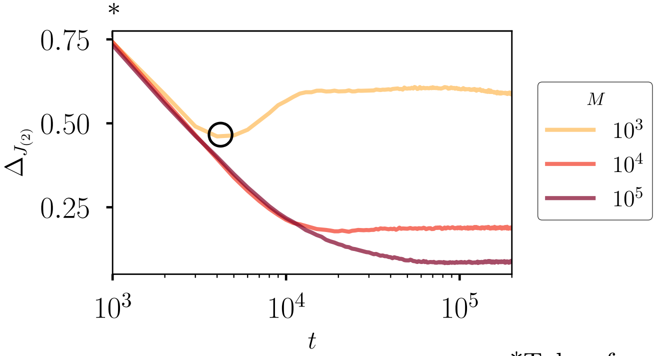
Dataset: equilibrium configurations from 2D Ising model (high-Temp)



Similar analysis of training dynamics can be carried out analytically, using Mean-Field approximation

Restricted Boltzmann Machine

Dataset: equilibrium configurations from 1D Ising model at high T

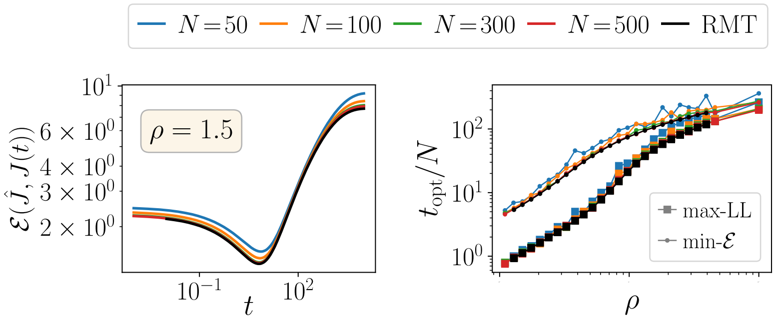


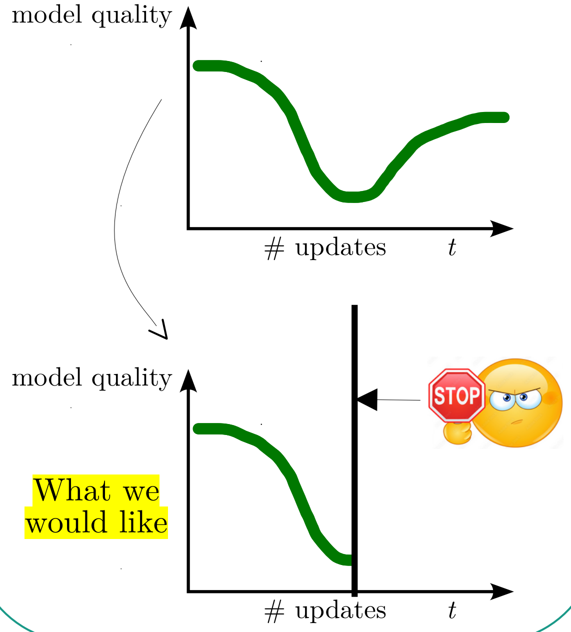
*Taken from Decelle, Furtlehner, Navas Gómez, Seoane, SciPost Physics 16(4)095 (2024)

Random matrix theory analysis

Asymptotic analysis through Random Matrix theory (RMT) to analyze finite-samples fluctuations in the training dynamics
<u>Exact on GEBM</u>

$N \to \infty$
$M \to \infty$
$\rho = \dfrac{M}{N}$ finite

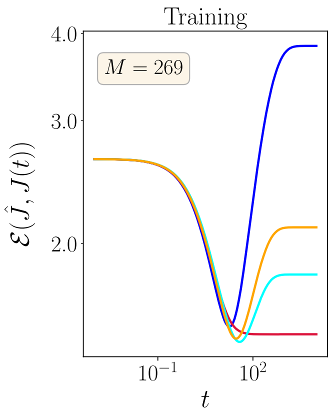


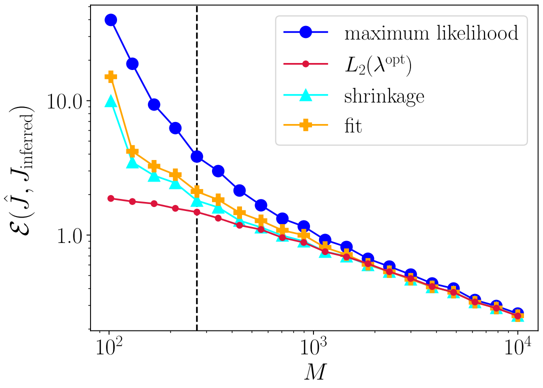


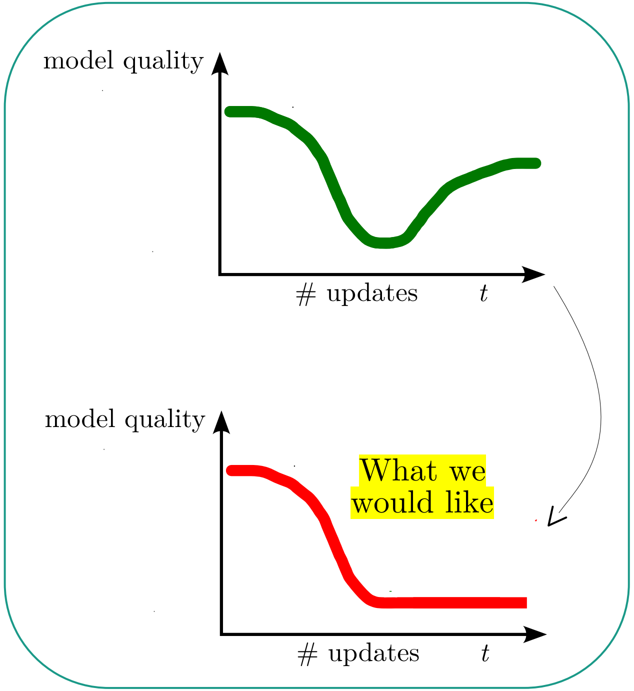
Protocols to mitigate overfitting

- regularization priors

- shrinkage correction protocols
- downsampling-based modes fitting







A Fanthomme, Rizzato, Cocco, Monasson JStat-Mech (2022) 053502
Bun, Bouchaud, Potters, Physics Reports (2017) 666/1-109

Extensions to more complex EBMs

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{\mathcal{Z}_{\boldsymbol{\theta}}} e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}$$

Study of overfitting in arbitrary complex EBMs can be done using the score-matching algorithm

<u>Score function</u>

$$\psi(\boldsymbol{x}|\boldsymbol{\theta}) = -\nabla_{\boldsymbol{x}} E(\boldsymbol{x}|\boldsymbol{\theta})$$

Learning dynamics of score function governed by a Neural Tangent Kernel

$$\frac{d\psi(x|\theta_t)}{dt} = -\hat{\mathbb{E}}_{x'}\left[K_t(x,x')\psi(x'|\theta_t)\right] + \hat{\phi}_t(x)$$

Similar learning dynamics to GEBM w.r.t. empirical covariance of latent feature in the tangent space $\rightarrow$ can lead to similar mechanism that justify the onset of overfitting

Summary

- Introduction of a novel theoretical framework to study overfitting in EBMs
- **Interplay between learning timescales** associated to different PCA directions (with different finite-sample fluctuations) can result in overfitting.
- Analysis on GEBM, asymptotics through RMT
- Theoretical extension on Boltzmann Machine (high-T), **extension to generic EBM in the context of NTK of the score function dynamics**
- sets the stage for
  a) early-stopping point determination through RMT
  b) extension of data-correction protocols to non-pairwise EBMs

arXiv: 2501.19158

ICML POSTER ID: 45237

with

Aurélien Decelle
Universidad Politecnica Madrid (ES)

Cyril Furtlehner
INRIA, Université Paris Saclay (FR)

Beatriz Seoane
Universidad Complutense Madrid (ES)

Thank you