Supplementary Information (SI)
Dynamical Regimes of Diffusion Models

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard

In this Appendix we first present more details on the analytical frameworks we used to study the backward dynamics of DMs, and then provide additional information on the numerical experiments. As already stressed in the main text, we need specific methods to study the limit of large dimension and large number of data. Such methods have been developed in statistical physics to study dynamics and thermodynamics of a large number of degrees of freedom. We will refer to original articles and, when needed, provide a short-introduction to the methods.

A Landau-type expansion for estimating the speciation time

The speciation transition is similar to the symmetry breaking phenomenon [2, 17] taking place at thermodynamic phase transitions, for instance when a ferromagnetic system develops a non-zero magnetization at low temperature. One way to study this phenomenon is to construct perturbatively the free energy as a function of the field [4]. In our case, we proceed in a similar way by obtaining $\log P_t(\mathbf{x})$ in a perturbative expansion in e^{-t} valid at large times. This approach is justified since speciation takes place at large times.

We rewrite the probability to be at $\mathbf{x}$ at time t as

$$\begin{aligned} P_t(\mathbf{x}) &= \int \mathrm{d}\mathbf{a}\; P_0(\mathbf{a}) \frac{1}{\sqrt{2\pi\Delta_t}^d} \exp\left(-\frac{1}{2}\frac{(\mathbf{x}-\mathbf{a}e^{-t})^2}{\Delta_t}\right) \\ &= \frac{1}{\sqrt{2\pi\Delta_t}^d} \exp\left(-\frac{1}{2}\frac{\mathbf{x}^2}{\Delta_t} + g(\mathbf{x})\right) \end{aligned} \tag{1}$$

where $g(\mathbf{x})$, defined as

$$g(\mathbf{x}) = \log \int \mathrm{d}\mathbf{a}\; P_0(\mathbf{a}) \exp\left(-\frac{1}{2}\frac{\mathbf{a}^2 e^{-2t}}{\Delta_t}\right) \exp\left(\frac{e^{-t}\mathbf{x}\cdot\mathbf{a}}{\Delta_t}\right) \tag{2}$$

can be interpreted in a field-theoretic (or probabilistic) approach, as a generative function for connected correlations of the variables $\mathbf{a}$ [21]. Expanding this function at large times, one finds

$$g(\mathbf{x}) = \frac{e^{-t}}{\Delta_t}\sum_{i=1}^{d} x_i\langle a_i\rangle + \frac{1}{2}\frac{e^{-2t}}{\Delta_t^2}\sum_{i,j=1}^{d} x_i x_j[\langle a_i a_j\rangle - \langle a_i\rangle\langle a_j\rangle] + O\left((xe^{-t})^3\right) \tag{3}$$

where the brackets $\langle . \rangle$ denote the expectation value with respect to the effective distribution $P_0(\mathbf{a})e^{-\mathbf{a}^2 e^{-2t}/(2\Delta_t)}$. Using this expansion, one finds at large times:

$$\log P_t(\mathbf{x}) = C + \frac{e^{-t}}{\Delta_t}\sum_{i=1}^{d} x_i\langle a_i\rangle - \frac{1}{2\Delta_t}\sum_{i,j=1}^{d} x_i M_{ij} x_j + O\left((xe^{-t})^3\right) \tag{4}$$

where C is an $\mathbf{x}$-independent term and

$$M_{ij} = \delta_{ij} - e^{-2t}[\langle a_i a_j\rangle - \langle a_i\rangle\langle a_j\rangle] \tag{5}$$

The curvature of $\log P_t(\mathbf{x})$ depends on the spectrum of the matrix M. At large times M is close to the identity, all its eigenvalues are positive. The shape changes qualitatively at the largest time t_S such that the largest eigenvalue of M crosses zero. The speciation time is characterized by a change of curvature of the effective potential $-\log P_t(\mathbf{x})$.

Notice that the effective measure used for computing the correlations of a_i variables appearing in the definition of M can be substituted at large times by the original measure $P_0(\mathbf{a})$. Therefore $M \simeq \mathbb{I} - e^{-2t}\hat{C}_0$ where $\hat{C}_0$ is the covariance of the distribution P_0, which can be estimated as the covariance of the data. This leads to the (second) general criterion discussed in the main text for the speciation transition.

B Gaussian mixtures: speciation time and asymptotic stochastic processes in the high-dimensional limit

We consider the case when $P_0(\mathbf{a})$ is the superposition of two Gaussian clusters of equal weight, with means $\pm\mathbf{m}$, and the same variance σ^2. We ensure that the two Gaussians are well separated in the limit of large dimensions, by assuming that $|\mathbf{m}|^2 = d\tilde{\mu}^2$, with σ and $\tilde{\mu}$ of order 1. In the following we present the detail of the analysis of the backward dynamics in the limit of large dimension. In the study of regimes I and II, we shall assume that $P_t^e(\mathbf{x})$ coincides with its population counterpart $P_t(\mathbf{x})$, and use the latter. We justify this assumption in the analysis of regime III, by showing that for $\alpha = \frac{\log n}{d}$ finite, the speciation time, and hence regime I and II, take place before the collapse, i.e. when $P_t^e(\mathbf{x})$ coincides with its population counterpart $P_t(\mathbf{x})$.

B.1 Spectral estimate of t_S

The speciation time computed from the eigenvalue criterion of the previous section is easily computed in this case. The matrix M is given by $M_{ij} = (1 - \sigma^2 e^{-2t})\delta_{ij} - e^{-2t} m_i m_j$ and its largest eigenvalue is $(1 - \sigma^2 e^{-2t} - d\tilde{\mu}^2 e^{-2t})$. We get therefore in the large d limit $t_S = \frac{1}{2}\log(d\tilde{\mu}^2)$ which up to subleading corrections identifies the speciation timescale as

$$t_S = \frac{1}{2}\log(d).$$

B.2 Asymptotic stochastic process in regime I and symmetry breaking

In the large dimensional limit, we can provide a full analytic study of the dynamics in regime I, i.e. at the beginning of the backward process following [2]).

As shown in Sect. C of this Appendix, when one studies the dynamics at times $t > t_C$ the empirical distribution at time t, $P_t^e(\mathbf{x})$ is well approximated by $P_t(\mathbf{x})$ which is the convolution of the initial distribution P_0 (a mixture of Gaussians centered at $\pm\mathbf{m}$) and the diffusion kernel proportional to $e^{-(\mathbf{x}-\mathbf{a}e^{-t})^2/2}$. The explicit expression is thus

$$P_t(\mathbf{x}) = \frac{1}{2\sqrt{2\pi\Gamma_t}^d}\left[e^{-(\mathbf{x}-\mathbf{m}e^{-t})^2/(2\Gamma_t)} + e^{-(\mathbf{x}+\mathbf{m}e^{-t})^2/(2\Gamma_t)}\right] \tag{6}$$

where $\Gamma_t = \sigma^2 e^{-2t} + \Delta_t$ goes to 1 at large times. The log of this probability is

$$\log P_t(\mathbf{x}) = -\frac{\mathbf{x}^2}{2\Gamma_t} + \log\cosh\left(\mathbf{x}\cdot\mathbf{m}\,\frac{e^{-t}}{\Gamma_t}\right) \tag{7}$$

and hence the score reads

$$S_i(\mathbf{x}) = -\frac{x_i}{\Gamma_t} + m_i\frac{e^{-t}}{\Gamma_t}\tanh\left(\mathbf{x}\cdot\mathbf{m}\,\frac{e^{-t}}{\Gamma_t}\right) \tag{8}$$

At the beginning of the backward process (for $t \gg t_S$), the x_i are i.i.d. Gaussian variables with unit variance. In consequence the overlap with the centers of the Gaussian model, $\mathbf{m}\cdot\mathbf{x}(t)$, scales as $\sqrt{d}$ in the large dimensional limit. Strictly speaking, regime I is defined as the scaling regime (or the time-window) in which $\mathbf{m}\cdot\mathbf{x}(t)$ keeps this scaling with d. Introducing the quantity $q(t) = \frac{1}{\sqrt{d}}\mathbf{m}\cdot\mathbf{x}(t)$ and using the notation $t_S = (1/2)\log d$, one can therefore write the backward equation on each x_i as:

$$-\mathrm{d}x_i = x_i + 2\left(-\frac{x_i}{\Gamma_t} + m_i\frac{e^{-t}}{\Gamma_t}\tanh\left(q(t)\frac{e^{-(t-t_S)}}{\Gamma_t}\right)\right)\mathrm{d}t + \mathrm{d}\eta_i(t) \tag{9}$$

where $\mathrm{d}\eta_i(t)$ is square root of two times the Brownian motion. This shows that in regime I, on the timescale t_S, each x_i satisfies a Langevin equation in which the interactions with all the other variables is through the fluctuating rescaled overlap $q(t)$. It is a kind of dynamical mean-field equation [3].

In order to obtain the equation on $q(t)$, one can sum the equation on all x_is. One therefore finds that this projection of the trajectory point $\mathbf{x}(t)$ on the direction linking the centers of the two Gaussians in the mixture satisfies the closed backward stochastic differential equation cited in the main text:

$$-\mathrm{d}q = -\frac{\partial V(q,t)}{\partial q}\mathrm{d}t + \mathrm{d}\eta(t) \tag{10}$$

where $\mathrm{d}\eta(t)$ is square root of two times the Brownian motion, and the potential $V(q,t)$ reads

$$V(q,t) = \frac{1}{2}q^2 - 2\tilde{\mu}^2 \log\cosh\left(qe^{-t}\sqrt{d}\right) \tag{11}$$

Clearly this potential is quadratic at times $t \gg (1/2)\log d$, and it develops a double well structure when $t \ll (1/2)\log d$. Equation (10) is a Langevin equation in a potential $V(q,t)$. As explained in the main text, the resulting dynamics leads to the symmetry breaking in which trajectories commit to one of the two classes. The scaling variable controlling such phenomenon is $e^{-\sqrt{t}}d$, as also found by the other means explained in the main text and this Appendix.

Note that the time where the curvature at $q=0$ vanishes, $\frac{\partial^2 V}{\partial q^2}(q=0,t^*)=0$, is

$$t^* = \frac{1}{2}\log(2d\tilde{\mu}^2). \tag{12}$$

In the main text we identify the speciation time with $t_S = \frac{1}{2}\log(d\tilde{\mu}^2)$, or equivalently $e^{-2t_S}d\tilde{\mu}^2 = 1$ but any other choice of constant different from one, e.g. $1/2$ as above, would be equivalent in the large d limit (it gives subleading correction to t_S in the large d limit).

B.3 Analytic computation of the speciation time defined from cloning

We now consider the criterion for speciation used in numerical experiments, namely the probability $\phi(t)$ that two trajectories, cloned at time t, end at time 0 in the same class. For gaussian mixtures, this can be computed as follows.

$$P(t) = \left[\int_{\mathbf{x}_1\cdot\mathbf{m}>0,\mathbf{x}_2\cdot\mathbf{m}>0} + \int_{\mathbf{x}_1\cdot\mathbf{m}<0,\mathbf{x}_2\cdot\mathbf{m}<0}\right]\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{x}_2\mathrm{d}\mathbf{y}\ \left(P(\mathbf{x}_1,0|\mathbf{y},t)P(\mathbf{x}_2,0|\mathbf{y},t)P(\mathbf{y},t)\right) \tag{13}$$

where $P(\mathbf{x}_1,0|\mathbf{y},t)$ is the probability that the backward process ends in $\mathbf{x}_1$ knowing that it was in $\mathbf{y}$ at time t. This conditional probability can be rewritten as $P(\mathbf{x}_1,0|\mathbf{y},t) = P(\mathbf{x}_1,0,\mathbf{y},t)/P(\mathbf{y},t)$. We can now use the forward process to compute all the probabilities involved, using $P(\mathbf{x}_1,0,\mathbf{y},t) = P(\mathbf{y},t|\mathbf{x}_1,0)P(\mathbf{x}_1,0)$:

$$P(\mathbf{x}_1,0,\mathbf{y},t) = P_0(\mathbf{x}_1)\frac{e^{-\frac{(\mathbf{y}-\mathbf{x}_1e^{-t})^2}{2\Delta_t}}}{\sqrt{2\pi\Delta_t}^d}$$

hence leading to:

$$P(\mathbf{x}_1,0|\mathbf{y},t) = P_0(\mathbf{x}_1)\frac{e^{-\frac{(\mathbf{y}-\mathbf{x}_1e^{-t})^2}{2\Delta_t}}}{\sqrt{2\pi\Delta_t}^d}\frac{1}{P(\mathbf{y},t)} \tag{14}$$

We now insert this expression into the equation for $P(t)$ and integrate over $\mathbf{x}_1$ and $\mathbf{x}_2$. The integrals can be done by noticing that this is equivalent to drawing both x_1 and x_2 from one Gaussian with probability $1/4$ and from the other Gaussian with probability $1/4$. One thus finds:

$$\phi(t) = \int \frac{\mathrm{d}\mathbf{y}}{2\sqrt{2\pi(\sigma^2e^{-2t}+\Delta_t)}^d}\ \frac{e^{-2\frac{(\mathbf{y}-\mathbf{m}e^{-t})^2}{2(\sigma^2e^{-2t}+\Delta_t)}} + e^{-2\frac{(\mathbf{y}+\mathbf{m}e^{-t})^2}{2(\sigma^2e^{-2t}+\Delta_t)}}}{e^{-\frac{(\mathbf{y}-\mathbf{m}e^{-t})^2}{2(\sigma^2e^{-2t}+\Delta_t)}} + e^{-\frac{(\mathbf{y}+\mathbf{m}e^{-t})^2}{2(\sigma^2e^{-2t}+\Delta_t)}}} \tag{15}$$

All the integrals on the components orthogonal to $\mathbf{m}$ factorize and can be done: they give one. One ends up with a one-dimensional integral equal to:

$$\phi(t) = \frac{1}{2}\int \mathrm{d}y \frac{G(y,me^{-t},\Gamma_t)^2 + G(y,-me^{-t},\Gamma_t)^2}{G(y,me^{-t},\Gamma_t) + G(y,-me^{-t},\Gamma_t)}\ , \tag{16}$$

where $G(y,u,v)$ is a Gaussian probability density function for the real variable y, with mean u and variance v, $m = |\mathbf{m}| = \tilde{\mu}\sqrt{d}$, $\Gamma_t = \Delta_t + \sigma^2e^{-2t}$. This one-dimensional integral is easily done numerically. For large d, the probability $\phi(t)$ that the two clones end up in the same cluster is a decreasing function of t, going from $\phi(0)=1$ to $\phi(\infty)=1/2$, and the speciation time for this Gaussian mixture model, estimated from $\phi(t_s)=.775$ is of order $t_s \simeq (1/2)\log d$ in the limit of large dimensions.

B.4 Asymptotic stochastic process in regime II

At the end of regime I the overlap of $\mathbf{x}(t)$ with the centers of the Gaussian model diverge: for some trajectories $q(t)$ goes to plus infinity, for others to minus infinity. This corresponds to a change of scaling regime. In fact, beyond regime I, the overlap $\mathbf{x}(t) \cdot \mathbf{m}$ scales proportionnally to d.

After the speciation transition, trajectories are committed to a given center. Conditioning on trajectories which correspond to the center $+\mathbf{m}$, and using that $\mathbf{x}(t) \cdot \mathbf{m} \to +\infty$, the score (8) simplifies to:

$$S_i^+(\mathbf{x}) = -\frac{x_i}{\Gamma_t} + m_i \frac{e^{-t}}{\Gamma_t} \tag{17}$$

This is the score of the backward process of a single Gaussian centered in $+\mathbf{m}$. The resulting equation on x_i is therefore:

$$-\mathrm{d}x_i = \left(-\frac{x_i}{\Gamma_t} + m_i \frac{e^{-t}}{\Gamma_t} \right) \mathrm{d}t + \mathrm{d}\eta_i(t) \tag{18}$$

where again $\mathrm{d}\eta_i(t)$ is square root of two times a Brownian motion. This equation is the one of the backward process of the single Gaussian centered in $+\mathbf{m}$, and therefore guarantees that all trajectories evolving with such equation will generate the single Gaussian centered in $+\mathbf{m}$.

We have focused on the trajectories committed to the center $+\mathbf{m}$. One can proceed analogously for the ones committed to the center $-\mathbf{m}$. The results above still holds with m_i replaced by $-m_i$.

We conclude this section on regime I and II by stressing that we expect that our analysis can be generalized to other models. The Curie-Weiss already studied in [2] is an example. The dynamics in regime II has also been studied in [7] using a model for the score.

C Gaussian Mixtures: collapse time

C.1 Setting and methods

In order to study the structure of $P_t^e(\mathbf{x})$ around a point $\mathbf{x} = \mathbf{a}^1 e^{-t} + \mathbf{z}\sqrt{\Delta_t}$, where $\mathbf{z}$ has i.i.d. Gaussian distributed components with mean zero and variance one, we start from the representation $P_t^e(\mathbf{x}) = [Z_1 + Z_{2\dots n}] / \sqrt{2\pi\Delta_t}^d$ where $Z_1 = e^{-(\mathbf{x}-\mathbf{a}^1 e^{-t})^2/(2\Delta_t)} = e^{-(\mathbf{z}^2)/2}$ and

$$Z_{2\dots n} = \sum_{\mu=2}^{n} e^{-(\mathbf{x}-\mathbf{a}^\mu e^{-t})^2/(2\Delta_t)} \tag{19}$$

We now study $P_t(\mathbf{x})$ in the limit of large d, large n, keeping $\alpha = \log n/d$ fixed. The first piece behaves as $Z_1 \simeq e^{-d/2}$. In the second piece, $Z_{2\dots n}$, we assume, without loss of generality, that the first n_+ data points were sampled from the Gaussian with mean $\mathbf{m}$, and the last $n_- = n - n_+$ ones were sampled from the Gaussian with mean $-\mathbf{m}$. In the large n limit, because of the law of large numbers n_+/n goes to $1/2$. We can write $Z_{2\dots n} = Z_+ + Z_-$,where

$$Z_+ = \sum_{\mu=2}^{n_+} e^{-(\mathbf{x}-\mathbf{a}^\mu e^{-t})^2/(2\Delta_t)} \tag{20}$$

$$Z_- = \sum_{\mu=n_++1}^{n} e^{-(\mathbf{x}-\mathbf{a}^\mu e^{-t})^2/(2\Delta_t)} \tag{21}$$

Z_1, Z_+ and Z_- give the contribution to $P_t(\mathbf{x})$ from, respectively, the data point $\mu = 1$ (which is supposed to have been generated from the gaussian with mean $\mathbf{m}$), the other $n_+ - 1$ data points generated from this same gaussian with mean $\mathbf{m}$, and the n_- data points generated from the gaussian with mean $-\mathbf{m}$. The two quantities $Z_\pm$ are partition functions and we shall show that, for typical values of $\mathbf{x} = \mathbf{a}^1 e^{-t} + \mathbf{z}\sqrt{\Delta_t}$, the free energies $(1/d)\log Z_\pm$ concentrates in the large d limit around a value $\psi_\pm(t)$. Furthermore, $\psi_- < \psi_+$. Therefore in the large d limit the collapse transition is identified as the time t_c such that $\psi_+(t_c) = -1/2$.

As already pointed out in the main text, in order to obtain $Z_\pm$, standard concentration methods, e.g. central limit theorem, do not apply as each term of the sum corresponds to the exponential of a random variable scaling as $\log n$ [1]. $Z_\pm$ correspond to the partition function of a system with $n-1$ independent 'random energy levels'. This is some elaboration of the 'Random Energy Model' which was introduced originally in [6] as a simple model of

glass transition, and studied in the limit we are interested in, i.e. n, d going to infinity with $\alpha = \frac{\log n}{d}$ fixed. This model has played a central role in the physics of glassy systems [15]. It was first solved with methods of theoretical physics [6]. It was then thoroughly studied by probabilistic rigorous methods [18].

For a mathematical introduction that make connections with physics and computer science see the book [14]. For our purposes, it is also useful the recent work [12] where a similar problem was addressed in the context of dense associative memories.

C.2 Computation of ψ_+

Given a time t, Z_+ is the partition function of a system with $n_+ - 1$ independent 'energy levels' $E^\mu = (\mathbf{x} - \mathbf{a}^\mu e^{-t})^2/(2\Delta_t)$ at thermal equilibrium at a temperature 1. All these energies are independent and identically distributed, and they are of order d. Writing $\varepsilon^\mu = E^\mu/d$, we denote by $\rho_t(\varepsilon)$ the probability density function from which these reduced energies are drawn independently. This density $\rho_t(\varepsilon)$ satisfies a large deviation principle with a rate function $f_t(\varepsilon)$:

$$\rho_t(\varepsilon) = e^{-df_t(\varepsilon)} \tag{22}$$

We define $g_t(\lambda)$ the Legendre transform $g_t(\lambda) = \max_\varepsilon[-f_t(\varepsilon) - \lambda\varepsilon]$ and use

$$e^{dg_t(\lambda)} = \int \mathrm{d}\varepsilon \; e^{-d[f_t(\varepsilon)+\lambda\varepsilon]} = \mathbb{E}_+ e^{-\lambda(\mathbf{x}-\mathbf{a}e^{-t})^2/2\Delta_t} \tag{23}$$

where $\mathbb{E}_+$ is an expectation on $\mathbf{a}$ drawn from the gaussian measure of mean $\mathbf{m}$ and variance σ^2. The right-hand side of (23) is the convolution of two gaussians which is easily computed. Using the fact that $(\mathbf{x} - \mathbf{m}e^{-t})^2/d$ concentrates at large d around $\sigma_t^2 + \Delta_t$, where $\sigma_t^2 = \sigma^2 e^{-2t}$, we get:

$$g_t(\lambda) = \frac{1}{2}\log\frac{\Delta_t}{\Delta_t + \lambda\sigma_t^2} - \frac{1}{2}\frac{\lambda(\Delta_t + \sigma_t^2)}{\Delta_t + \lambda\sigma_t^2} \tag{24}$$

The Legendre transform can be done analytically. the maximum of $g_t(\lambda) + \lambda\varepsilon$ is obtained at

$$\lambda^*(\varepsilon) = \frac{\sigma_t^2 - 4\Delta_t\varepsilon + \sqrt{\sigma_t^4 + 8\Delta_t^2\varepsilon + 8\Delta_t\sigma_t^2\varepsilon}}{4\sigma_t^2\varepsilon} \tag{25}$$

The inverse transform is obtained from the maximum of $-f_t(\varepsilon) - \lambda\varepsilon$ which is obtained at

$$\varepsilon^*(\lambda) = -\frac{\mathrm{d}g_t}{\mathrm{d}\lambda} = \frac{\Delta_t^2 + 2\Delta_t\sigma_t^2 + \lambda\sigma_t^2}{2(\Delta_t + \lambda\sigma_t^2)^2} \tag{26}$$

and the rate function is

$$f_t(\varepsilon) = \frac{\Delta_t}{2\sigma_t^2(\sigma_t^2 + A)}\left[-8\Delta_t\varepsilon + (1-6\varepsilon)\sigma_t^2 + (1+2\varepsilon A)\right] - \frac{1}{2}\log\frac{4\Delta_t\varepsilon}{\sigma_t^2 + A} \tag{27}$$

where

$$A = \sqrt{\sigma_t^4 + 8\varepsilon\Delta_t(\Delta_t + \sigma_t^2)} \tag{28}$$

The partition function is $Z_+ = \sum_{\mu=2}^{n_+} e^{-d\varepsilon^\mu}$. The annealed approximation to this partition function is $Z_+^{ann} = n_+ \int \mathrm{d}\varepsilon \; e^{-d(\varepsilon + f_t(\varepsilon))}$. Using Laplace's method, we see that this integral is dominated by $\varepsilon = \varepsilon^*(\lambda = 1) = 1/2$, and gives

$$\psi_+^{ann} = \lim_{d\to\infty}\frac{1}{d}\log Z_+^{ann} = \alpha - \frac{1}{2} - f_t\left(\frac{1}{2}\right) = \alpha + g_t(1) = \alpha + \frac{1}{2}\log\frac{\Delta_t}{\Delta_t + \sigma_t^2} - \frac{1}{2} \tag{29}$$

The methods for analyzing random energy models are standard (see [6, 14]). For a presentation close to the developments which we use here, see [12], in particular its Appendix A. Using either with first and second moment methods, or with replicas, we find that the annealed free energy is exact, namely $\psi_+ = \psi_+^{ann}$ when $t < t_{\mathrm{cond}}$, but at $t > t_{\mathrm{cond}}$ a condensation phenomenon occurs, the partition function is dominated by the levels with lowest energy, ε_{min} given by the smallest root of $\alpha = \hat{f}_t(\varepsilon)$. The condensation time t_{cond} is thus characterized by the fact that

$\varepsilon_{min} = 1/2$, which gives:

$$\alpha = f_{t_{\rm cond}}(1/2) \tag{30}$$

So the final expression for the free energy of this random energy model is:

$$\begin{aligned}\psi_+ = \lim_{d\to\infty} \frac{1}{d}\log Z_+ &= \alpha + \frac{1}{2}\log\frac{\Delta_t}{\Delta_t+\sigma_t^2} - \frac{1}{2} \quad \text{if } t < t_{\rm cond} \\ &= -\frac{1}{2} \quad \text{if } t > t_{\rm cond}\end{aligned} \tag{31}$$

and its condensation transition is

$$t_{\rm cond} = \frac{1}{2}\log\left[1 + \frac{\sigma^2}{n^{2/d}-1}\right] \tag{32}$$

C.3 Computation of ψ_-

Given t, the partition function Z_- is again a partition function of a random energy model, which can be studied following exactly the same steps as for Z_+. The computation of $g_t(\lambda)$ is slightly different. Now, $(\mathbf{x} - \mathbf{m}e^{-t})^2/d$ concentrates at large d around $4(\mathbf{m}^2/d)e^{-2t} + \sigma_t^2 + \Delta_t$+ This gives:

$$g_t(\lambda) = \frac{1}{2}\log\frac{\Delta_t}{\Delta_t+\lambda\sigma_t^2} - \frac{1}{2}\frac{\lambda[M_t+\Delta_t+\sigma_t^2]}{\Delta_t+\lambda\sigma_t^2} \tag{33}$$

where $M_t = 4(\mathbf{m}^2/d)e^{-2t}$ We now list the changes in the subsequent formulations, keeping the same notations as in the previous section. The maximum of $g_t(\lambda) + \lambda\varepsilon$ is obtained at

$$\lambda^*(\varepsilon) = \frac{\sigma_t^2 - 4\Delta_t\varepsilon + \sqrt{\sigma_t^4 + 8\Delta_t^2\varepsilon + 8\Delta_t\sigma_t^2\varepsilon + 8\Delta_t M_t\varepsilon}}{4\sigma_t^2\varepsilon} \tag{34}$$

The inverse transform is obtained from the maximum of $-f_t(\varepsilon) - \lambda\varepsilon$ which is at

$$\varepsilon^*(\lambda) = -\frac{dg_t}{d\lambda} = \frac{\Delta_t^2 + 2\Delta_t\sigma_t^2 + \lambda\sigma_t^2 + M_t\Delta_t}{2(\Delta_t+\lambda\sigma_t^2)^2} \tag{35}$$

and the rate function is

$$f_t(\varepsilon) = \frac{\Delta_t\left[-8\Delta_t\varepsilon + (1-6\varepsilon)\sigma_t^2 + (1+2\varepsilon A)\right] + M_t\left[\sigma_t^2 - 8\varepsilon\Delta_t + A\right]}{2\sigma_t^2(\sigma_t^2+A)} - \frac{1}{2}\log\frac{4\Delta_t\varepsilon}{\sigma_t^2+A} \tag{36}$$

where

$$A = \sqrt{\sigma_t^4 + 8\varepsilon\Delta_t(\Delta_t+\sigma_t^2) + 8\Delta_t M_t\varepsilon} \tag{37}$$

The annealed approximation to the partition function is $Z_-^{ann} = n_- \int \mathrm{d}\varepsilon\, e^{-d(\varepsilon + f_t(\varepsilon))}$. This integral is dominated by

$$\varepsilon = \varepsilon^*(\lambda=1) = \frac{1}{2}\left[1 + \frac{\Delta_t M_t}{(\Delta_t+\sigma_t^2)^2}\right] \tag{38}$$

and gives

$$\psi_-^{ann} = \lim_{d\to\infty}\frac{1}{d}\log Z_-^{ann} = \alpha + g_t(1) = \alpha + \frac{1}{2}\log\frac{\Delta_t}{\Delta_t+\sigma_t^2} - \frac{1}{2} - \frac{1}{2}\frac{M_t}{\Delta_t+\sigma_t^2} \tag{39}$$

The annealed free energy is exact when $t < t_{\rm cond}$. The condensation takes place at $t > t_{\rm cond}$, then the partition function is dominated by the levels with lowest energy, ε_{min} given by the smallest root of $\alpha = \hat{f}_t(\varepsilon)$. The

condensation time $t_{\rm cond}$ is thus characterized by the fact that $\varepsilon_{min} = \varepsilon^*(\lambda = 1)$, which gives:

$$\alpha = f_{t_{\rm cond}}\left(\varepsilon = \frac{1}{2}\left[1 + \frac{\Delta_t M_t}{(\Delta_t + \sigma_t^2)^2}\right]\right) \tag{40}$$

So the final expression for the free energy of this random energy model is:

$$\begin{aligned}\psi_- = \lim_{d\to\infty}\frac{1}{d}\log Z_- &= \alpha + \frac{1}{2}\log\frac{\Delta_t}{\Delta_t+\sigma_t^2} - \frac{1}{2} - \frac{1}{2}\frac{M_t}{\Delta_t+\sigma_t^2} \quad \text{if } t < t_{\rm cond} \\ &= -\frac{1}{2}\left[1 + \frac{\Delta_t M_t}{(\Delta_t+\sigma_t^2)^2}\right] \quad \text{if } t > t_{\rm cond}\end{aligned} \tag{41}$$

and its condensation transition is

$$t_{\rm cond} = \frac{1}{2}\log\left[1 + \frac{\sigma^2}{n^{2/d} - 1}\right] \tag{42}$$

It is easy to see that at any time $\psi_- < \psi_+$.

C.4 Final expression

Going back to the representation $P_t^e(\mathbf{x}) = \left[Z_1 + Z_{2\dots n}\right]/\sqrt{2\pi\Delta_t}^d$, where $Z_1 = e^{-(\mathbf{x}-\mathbf{a}^1 e^{-t})^2/(2\Delta_t)} = e^{-(\mathbf{z}^2)/2}$, it is made of three contributions:

1. Z_1 which behaves at large d as $e^{-d/2}$
2. Z_+ which behaves at large d as $e^{-d\psi_+}$
3. Z_- which behaves at large d as $e^{-d\psi_-}$

In the limit of large dimensions $d \to \infty$, the contribution from Z_- is irrelevant because $\psi_- < \psi_+$. The comparison of the two other terms depends on whether ψ_+ is larger or smaller than $-1/2$. We thus find two time regimes separated by a collapse transition time t_C.

- at large times, $t > t_C$, $\psi_+ > -(1/2)$ and the probability $P_t^e(\mathbf{x})$ is dominated by the term Z_+. This is the regime which is not collapsed.
- at small times, $t < t_C$, $\psi_+ < -(1/2)$ and the probability $P_t^e(\mathbf{x})$ is dominated by the term $Z_1 = e^{-(\mathbf{x}-\mathbf{a}^1 e^{-t})^2/(2\Delta_t)}$. When used in the backward diffusion, this gives a score that attracts $\mathbf{x}$ towards $\mathbf{a}^1$ at short times.

We have thus shown that, in the backward process, trajectories which are at time t at typical points of the form $\mathbf{x} = \mathbf{a}^1 e^{-t} + \mathbf{z}\sqrt{\Delta_t}$ are attracted towards $\mathbf{a}^1$ if $t < t_C$. This identifies t_C as the collapse time.

Looking at the explicit expression (31) of ψ_+, we see that the collapse time t_C is identical to the condensation time of the random energy model, $t_{\rm cond}$. This result, derived exactly in this case of gaussian mixtures, is actually a general result, as discussed in the main paper. It could also be shown using the REM approach, provided the distribution of energies satisfies a large deviation principle.

D Details on the numerical experiments

D.1 Gaussian mixtures

Speciation experiment. In the case of a mixture of two Gaussian clusters centered on $\pm\mathbf{m} \in \mathbb{R}^d$ with variance σ^2, the score function $\mathcal{F}(\mathbf{x}, t)$ can be analytically expressed as

$$\mathcal{F}(\mathbf{x}(t), t) = \mathbf{m}\frac{e^{-t}}{\Gamma_t}\tanh\left(\frac{e^{-t}}{\Gamma_t}\mathbf{x}(t)\cdot\mathbf{m}\right) - \frac{\mathbf{x}(t)}{\Gamma_t}, \tag{43}$$

where $\Gamma_t = \Delta_t + \sigma^2 e^{-2t}$, with $\Delta_t = 1 - e^{-2t}$. We can then discretize the stochastic differential equation associated to the backward process as

$$\mathbf{x}(t+1) = \mathbf{x}(t) + \eta\left[\mathbf{x}(t) + 2\mathcal{F}(\mathbf{x}(t), t)\right] + \boldsymbol{\xi}\sqrt{2\eta}, \tag{44}$$

where $\boldsymbol{\xi} \sim \mathcal{N}(0, I)$, $\mathbf{x}_T \sim \mathcal{N}(0, I)$ and $\eta = T/L$, with $T = 10$ the time horizon and $L = 1000$ the number of discrete steps. Two clones $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ share the same trajectory until a given time below which they have independant noise realizations $\boldsymbol{\xi}$. At the end of the backward diffusion, one can then recover the class of the clone i by projecting it onto $\mathbf{m}$ as $\text{sign}\,(\mathbf{m} \cdot \mathbf{x}_i(0))$. In the experiment corresponding to Fig. 2 of the main text, we use $\mathbf{m} = [1, \dots, 1]$, $\sigma^2 = 1$, and each point is obtained by averaging over $10\,000$ initial conditions.

Collapse experiment. The main component of the collapse experiment is the excess entropy density $f(t)$ which vanishes for $t \leq t_C$. When dealing with Gaussian mixtures, $f(t)$ can be derived exactly and we additionally establish a strategy to approximate it numerically as follows. At a given time t, we draw n' samples $\mathbf{x}^{(\nu)}(t) = e^{-t}a^{(\mu)} + \sqrt{1 - e^{-2t}}\boldsymbol{\xi}^{(\nu)}$, where μ is chosen uniformly in $\{1, \dots, n\}$. The entropy $s(t) = -\int \mathrm{d}\mathbf{x}\, P_t(\mathbf{x}) \log P_t(\mathbf{x})$ can then be approximated as the empirical average over the n' samples, leading to the empirical estimate $f^e(t)$ of the excess entropy, using

$$s^e(t) = -\frac{1}{n'd} \sum_{\nu=1}^{n'} \log P_t^e\left(\mathbf{x}_t^{(\nu)}\right), \tag{45}$$

where $P_t^e(\mathbf{x})$ is given as usual in terms of the original dataset by

$$P_t^e(\mathbf{x}) = \frac{1}{n} \sum_{\mu=1}^{n} \frac{1}{\sqrt{2\pi\Delta_t}^d} \exp\left(-\frac{(\mathbf{x} - \mathbf{a}^\mu e^{-t})^2}{2\Delta_t}\right). \tag{46}$$

This approximation is valid for $t > t_C$, where $P_t(\mathbf{x}) \approx P_t^e(\mathbf{x})$. As shown in Fig. 2 of the main text where we fix $n = 20\,000$ and $n' = 250\,000$ for several d, this estimate fits quite remarkably the analytical curve. Such a procedure therefore allows for the numerical derivation of the collapse t_C from the dataset only, under the assumption that one uses the exact empirical score.

D.2 Realistic datasets

Denoising Diffusion Probabilistic Models. In the second part of the paper, we learn the score by training a Denoising Diffusion Probabilistic Model (DDPM) in discrete time, as introduced by [9]. In this context, the forward process has a variance schedule $\{\beta'_t\}_{t'=1}^{L}$, where L is the time horizon given as a number of steps, fixed to 1000 in our experiments. In our case, the variance is evolving linearly from $\beta_1 = 10^{-4}$ to $\beta_{1000} = 2 \times 10^{-2}$. A sample at timestep t', denoted $\mathbf{x}(t')$ can therefore be expressed readily from its initial state, $\mathbf{x}(0) = \mathbf{a}$, as

$$\mathbf{x}(t') = \sqrt{\overline{\alpha}(t')}\mathbf{a} + \sqrt{1 - \overline{\alpha}(t')}\boldsymbol{\xi}(t'), \tag{47}$$

where $\overline{\alpha}(t') = \prod_{s=1}^{t'}(1 - \beta_s)$ and $\boldsymbol{\xi}$ is a standard and centered Gaussian noise. This equation in fact corresponds to the discretization of the Ornstein-Uhlenbeck Eq. (1) given in the main text under the following reparameterization of the timestep t',

$$t = -\frac{1}{2}\log\left(\overline{\alpha}(t')\right), \tag{48}$$

where t is the time as defined in the main text. We then train a neural network to learn $\boldsymbol{\xi}_\theta\left(\mathbf{x}(t'), t'\right)$, the noise at time t' by iteratively optimizing the loss function

$$\mathcal{L}(\theta) = \|\boldsymbol{\xi}(t') - \boldsymbol{\xi}_\theta\left(\mathbf{x}(t'), t'\right)\|^2. \tag{49}$$

Once learned, the denoiser can then be used to generate a new sample as

$$\mathbf{x}(t'-1) = \frac{1}{\sqrt{1-\beta_{t'}}}\mathbf{x}(t') - \frac{\beta_{t'}}{\sqrt{(1-\beta_{t'})\left(1-\overline{\alpha}_{t'}\right)}}\boldsymbol{\xi}_\theta\left(\mathbf{x}(t'), t'\right) + \sqrt{\beta_{t'}}\mathbf{z}, \tag{50}$$

where t' runs from T to one, $\mathbf{x}(T) \sim \mathcal{N}(0, I)$, and $\mathbf{z} \sim \mathcal{N}(0, I)\delta(t' > 1)$, which, in the continuous limit, is equivalent to the score-based diffusion studied in the main text [13], where $-\boldsymbol{\xi}_\theta\left(\mathbf{x}(t'), t'\right)/\sqrt{1 - \overline{\alpha}_{t'}}$ corresponds to the score.

Datasets and preprocessing. Our experiments are based on five datasets: MNIST [11], CIFAR [10], ImageNet16/32 [5], and LSUN [20]. For each dataset, we focus only on two well-disctinct and balanced classes that are:

zero and seven for MNIST ($n = 10\,000$), horses and cars for CIFAR ($n = 3000$), lesser pandas and seashores for ImageNet16 and 32 ($n = 2000$), and churches and conference rooms for LSUN ($n = 40\,000$ or $n = 200$). In the case of MNIST, the images are zero-padded from their original size to 32×32. For LSUN, images are center-cropped at size 256×256 before being resized to 64×64. All the datasets but MNIST have three color channels that are kept for the training, and the data are centered in each channel. The different datasets therefore allow to span several values of d ranging from 1024 (for MNIST) to $12\,288$ for LSUN. The values of Λ given in the main text corresponding to the largest eigenvalue of the covariance matrix are computed from the first channel only.

Architecture and training. For all the datasets, the denoiser $\boldsymbol{\xi}_\theta\left(\mathbf{x}(t'), t'\right)$ has a U-Net architecture similar to previous applications of DDPMs [9, 19]. More precisely, it implements four resolution levels, each having two convolutional residual blocks with group normalization, and self attention is applied to the two intermediary levels. The time is added into each block through a standard sinusoidal position embedding, resulting in a total of 25.7 million parameters. The models are then trained using ADAM optimizer with a fixed learning rate of 10^{-4} for 350k steps with batches of size 128, except for LSUN for which we reduced it to 64. All the n images were used to train the models, without any data augmentation. It is worth mentioning that the results exposed in the main text do not depend on the formal architecture of the denoiser, nor the optimization scheme, as long as the network is able to learn a score close to the true one (or equivalently, a sufficiently good denoiser). In Figs. 1 to 5, we show several samples generated from our six trained models.

Speciation experiment. To estimate numerically the speciation time from the probability $\phi(t)$ that the two clones end up in the same class, we need to classify the generated samples. To do so, we train the PyTorch [16] implementation of the ResNet-18 architecture [8] with pre-trained weights on ImageNet. One model is trained for each of the dataset using all the n samples. The resulting classifiers yield 95% test accuracy at worst on LSUN, and more than 99% for the other datasets.

Collapse experiment. To study the collapse, we devise two methods. First, we exploit the cloned trajectories but we now focus on the indices of the nearest neighbors in the training set at the end of the backward process. Using a small number of training data ($n = 2000$ for ImageNet and $n = 200$ for LSUN) ensures the backward dynamics to collapse onto one of the training datapoint. When $t < t_C$, the two clones have the same nearest neighbor, with a very small Euclidean distance, showing that the collapse has indeed taken place. This approach allows us to compute $\phi_C(t)$, the probability that the two clones collapse on the same datapoint. We illustrate this phenomenon on the LSUN dataset trained with either $n = 200$ or $n = 40\,000$ in Fig. 6. For a random batch of generated samples, we show the four nearest neighbors $\{a^{\mu_1}, a^{\mu_2}, a^{\mu_3}, a^{\mu_4}\}$ (ordered by increasing distance) in the training set. In the case of a model trained with a small number of datapoints, the generated images collapse onto μ_1, as shown in Fig. 6a. The model being fixed, one solution to avoid the collapse is to increase n, resulting in the generation of new samples, as illustrated in Fig. 6b. In a second approach, we exploit the indices of nearest neighbors in the training set during the backward process. At any times $t < t_C$, the index of the nearest neighbor $\mu_1(t)$ is therefore fixed. At $t = 0$, the distance d_{μ_1} to the nearest neighbor is vanishing, as seen before. The average of this distribution for $n = 4000$ generated samples yields the estimate $\hat{t}_C$. The entropy appearing in $f^e(t)$ is computed the same way as in Gaussian mixtures (see Sect. D.1), making use of the training sets and the forward process, with $n' = 400\,000$.

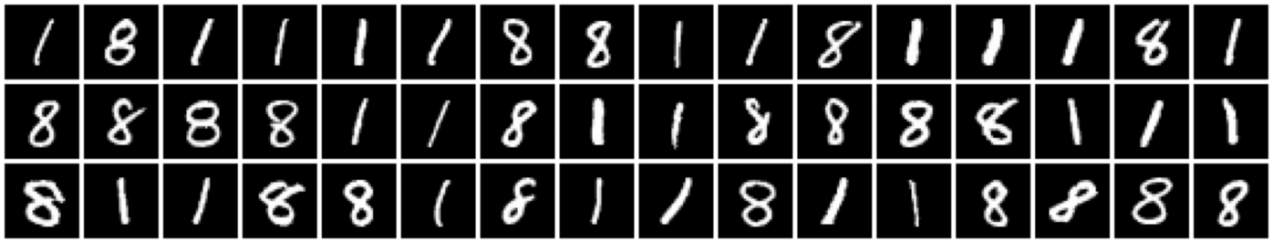
Supplementary References

[1] Gérard Ben Arous, Leonid V Bogachev, and Stanislav A Molchanov. Limit theorems for sums of random exponentials. *Probability theory and related fields*, 132:579–612, 2005.

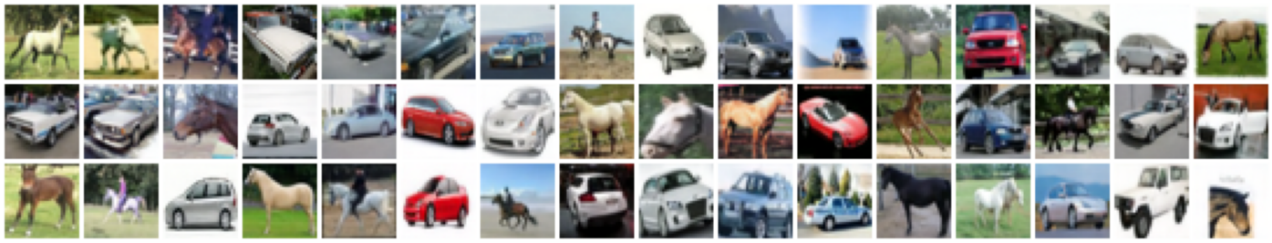
[2] Giulio Biroli and Marc Mézard. Generative diffusion in very large dimensions. *J. Stat. Mech.*, page 093402, 2023.

[3] Tony Bonnaire, Davide Ghio, Kamesh Krishnamurthy, Francesca Mignacco, Atsushi Yamamura, and Giulio Biroli. High-dimensional non-convex landscapes and gradient descent dynamics. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(10):104004, 2024.

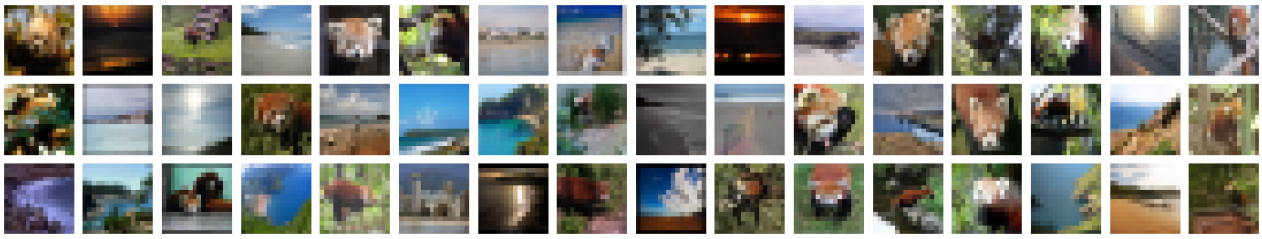
[4] Paul M Chaikin, Tom C Lubensky, and Thomas A Witten. *Principles of condensed matter physics*, volume 10. Cambridge university press Cambridge, 1995.



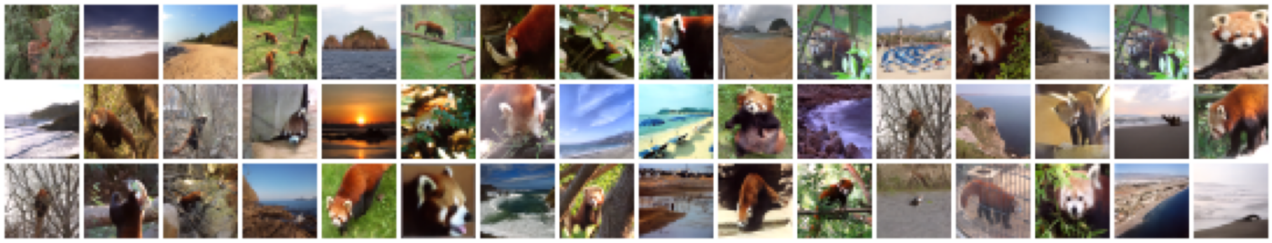
Supplementary Figure 1: MNIST one and eight generated samples with $n = 10\,000$.



Supplementary Figure 2: CIFAR horses and cars generated samples with $n = 3000$.



Supplementary Figure 3: ImageNet16 lesser pandas and seashores generated samples with $n = 2000$.



Supplementary Figure 4: ImageNet32 lesser pandas and seashores generated samples with $n = 2000$.

(a) **(b)**

Supplementary Figure 5: LSUN churches and conference rooms generated samples with (a) $n = 200$ or (b) $n = 40\,000$.

Supplementary Figure 6: Illustration of the collapse on the LSUN dataset trained with (a) $n = 200$ **or (b)** $n = 40\,000$ **datapoints.** Each row corresponds to a generated sample shown in the left-most column. The other columns display the first four nearest neighbors in the training set, denoted as a^{μ_1}, a^{μ_2}, a^{μ_3}, and a^{μ_4}. Below each of them is indicated the L_2 norm d to the generated sample.

[5] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

[6] Bernard Derrida. Random-energy model: An exactly solvable model of disordered systems. *Physical Review B*, 24(5):2613, 1981.

[7] Davide Ghio, Yatin Dandi, Florent Krzakala, and Lenka Zdeborová. Sampling with flows, diffusion, and autoregressive neural networks from a spin-glass perspective. *Proceedings of the National Academy of Sciences*, 121(27):e2311810121, 2024.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

[10] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Canadian Institute for Advanced Research*, 2009.

[11] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[12] Carlo Lucibello and Marc Mézard. The exponential capacity of dense associative memories. *Phys.Rev.Lett*, 132: 077301, 2024.

[13] Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

[14] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[15] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[17] Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[18] David Ruelle. A mathematical reformulation of derrida's rem and grem. *Communications in Mathematical Physics*, 108:225–239, 1987.

[19] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[20] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2016.

[21] Jean Zinn-Justin. *Quantum field theory and critical phenomena*, volume 171. Oxford university press, 2021.